Analysing Microarry Data Using Homogeneous And Inhomogeneous Hidden Markov Models

Diploma Thesis

Diploma study path bioinformatics

Institute of Computer Science
Martin-Luther-University Halle-Wittenberg

presented by

Michael Seifert

supervised by

Dr. Alexander Schliep - Max Planck Institute for Molecular Genetics
Prof. Dr. Stefan Posch - Martin-Luther-University Halle-Wittenberg

Halle 2006

Abstract

Chromosomal imbalances and gene expression alterations play a central role in different types of cancer. Microarray experiments are common techniques to generate data sets which contain such information.

This diploma thesis has two goals. The first goal is to develop *inhomogeneous Hidden Markov Models* to detect chromosomal imbalances and gene expression alterations and the second goal is to test the performance of this novel approach on breast cancer data.

To achieve these goals we have extended the mathematical theory of standard *Hidden Markov Models* to obtain *inhomogeneous Hidden Markov Models* with a few easily interpretable parameters. The improved quality of our novel approach in comparison with the standard *Hidden Markov Models* is the result of using the chromosomal locations of genes and the microarray measurements of theses genes as input data for our *Hidden Markov Models*. To our knowledge the simultaneous usage of both, the chromosomal locations and the microarray measurements, is a novel strategy. Previously described methods use the chromosomal locations only to interpolate the data.

The fact that our *inhomogeneous Hidden Markov Models* are able to find known candidate genes for over-expression and in literature described losses and gains of DNA segments in breast cancer data represents the good quality of this approach.

Contents

1 Introduction **3**
1.1 Molecular Basics Of Cancer 4
1.2 Breast Cancer 5
1.3 Microarrays 6
1.3.1 Measuring Gene Expression With Microarrays 6
1.3.2 Measuring DNA Copy Number Changes With Microarrays 6

2 Hidden Markov Models **8**
2.1 Basics of Hidden Markov Models 8
2.2 Hidden Markov Models With Continuous Emissions 9
2.3 Hidden Markov Models With Continuous Emissions And Transition Classes . . . 10
2.4 Specific Definitions For Hidden Markov Models 11
2.5 Basic Questions For Continuous Hidden Markov Models 13

3 Hidden Markov Models With Coupled Transition Matrices **15**
3.1 Coupling Of Transition Matrices 15
3.2 Definition Of Hidden Markov Models With Coupled Transition Classes 17
3.3 Estimation Of Transition Matrices 18
3.4 Convergence Of The Baum-Welch Algorithm 23
3.5 Analytical Solution For One Transition Class 25
3.6 Analytical Solution For Two Transition Classes 26

4 Emission Densities For Hidden Markov Models **28**
4.1 EM Algorithm 28
4.1.1 Mixture Of Univariate Normal Distributions 29
4.2 Clustering Mixture Components 32
4.3 Example Analysis 33

5 Modelling Microarray Data **37**
5.1 Creating *HMM*s For Microarray Data 37
5.1.1 Modelling Start Distributions And Transition Matrices 38
5.1.2 Modelling Transition Class Switching Functions 39
5.2 Annotating Microarray Profiles 40
5.3 The Role Of Chromosome 17 In Breast Cancer 42
5.4 Breast Cancer Data Set 43

6 Analysing Microarray Data **49**
6.1 Analysing Gene Expression Profiles 49
6.1.1 Standard *HMM*s 49
6.1.2 *HMM*s With Transition Classes 67

6.2 Analysing *ArrayCGH* Profiles 85

6.2.1 *HMM*s With Transition Classes 85

6.3 Comparing Gene Expression And *ArrayCGH* Profiles 101

7 Discussion Of The *HMM* Approach 103

7.1 Summary And Discussion Of The *HMM* Approach 103

7.2 Performance Of The *HMM* Approach On Breast Cancer Data 104

7.3 Outlook 106

Bibliography 107

Chapter 1

Introduction

Tumours are driven by an accumulation of mutations resulting in altered gene expression levels. Altered gene expression patterns modify the normal cell growth and survival. In many types of cancer are mutations as losses or gains of chromosomal regions frequently observed. One or more mechanisms which normally control the genome stability or the cell cycle must be affected in order to turn a normal cell into a tumour cell.

Two types of data can be measured with microarray experiments. That is, we are able to determine the gene expression rates and the DNA copy number status of a tumour sample in comparison with a sample of normal cells. Mutations can cause chromosomal imbalances which include any change of the chromosomal structure or the number of chromosomes. In general a diploid genome has two copies of each gene and therefore the copy number of each gene is equal to two. A chromosomal imbalance as the loss or the gain of a gene leads to an altered copy number. Most of the mutations affect large chromosomal regions and lead to insertions, deletions and amplifications of DNA segments. The influences of gains and losses of genes on the expression levels of these genes are shown by Pollack *et al.* [23]. The amplification of a gene can cause a higher expression level of this gene and the loss of a gene can induce a lower expression level. These relationships between gene copy numbers and gene expression levels are often observed in many different types of cancer, but not each altered expression level of a gene is the result of a changed copy number of this gene because the biological mechanisms can be more complex.

Nevertheless, it is possible to recognise proximity effects of genes in microarray data. Genes which are located close to each other on the chromosome are more likely amplified or deleted together as genes which have a greater chromosomal distance. Our aim is to develop a method which is able to model the proximity effects of genes in microarray data. Using this method the microarray data of a chromosome is divided into different regions. That is, we can observe regions of increased, unchanged and decreased DNA copy number status for DNA copy number data and under-expressed, identically expressed and over-expressed regions for gene expression data.

In particular, it is desirable to make use of the proximity effects of genes and to translate these effects into dependencies of microarray data. All currently available methods to find chromosomal imbalances in DNA copy number data do not consider these effects [16]. To our knowledge *MACAT - microarray chromosome analysis tool* [25] is the only approach which links differential gene expression to the chromosomal locations of genes by interpolating the data with distance-dependent kernel functions.

The natural framework of our novel method is an *Hidden Markov Model (HMM)* approach using the chromosomal distance between adjacent genes in microarray profiles to model the proximity effects of these genes. In order to obtain the basics of our *HMM* approach we modify the inhomogeneous *HMM*s which were introduced by Knab [14] to get an inhomogeneous *HMM*

with a few easily interpretable parameters. Our novel concept uses a predefined number of coupled transition matrices whereby one of these transition matrices is individually selected for each gene in a microarray profile by a distance-dependent switching function. The initial model parameters of our *HMM* are determined on the basis of the given microarray data. That is, we estimate a mixture model of normal distributions using the well known *Expectation Maximisation algorithm.* Afterwards the mixture components of this mixture model are clustered to the states of our *HMM* as emission functions. The start distribution of our *HMM* is set to the vector of prior probabilities of the clusters which represent the emission functions. A basic transition matrix with an equilibrium distribution equal to the start distribution and a vector of scaling parameters are also assigned to our *HMM*. The scaling parameters are used to generate the coupled transition matrices by scaling the state durations of the basic transition matrix and therewith a more realistic model of proximity effects can be described.
This diploma thesis occupies with the mathematical theory of our novel *HMM*s and the performance of our novel *HMM*s in comparison with the homogeneous standard *HMM*s on breast cancer microarray data. Finally, it is to emphasise that our developed method is able to work with gene expression and DNA copy number data.

Structure of the diploma thesis

1. In Chapter 1 molecular basics of cancer, actual facts about breast cancer and microarray experiments to measure gene expression rates and DNA copy number changes are introduced.

2. In Chapter 2 the basics of homogeneous and inhomogeneous *HMM*s are presented.

3. In Chapter 3 the concept of coupled transition matrices is introduced and afterwards our novel *HMM*s are defined and prepared for the training with microarray data.

4. In Chapter 4 mixture models for microarray data are estimated as basics for the agglomerative clustering algorithm which assigns the mixture components as emission functions to the *HMM*.

5. In Chapter 5 we show how to create and use *HMM*s to analyse microarray data, we introduce the role of chromosome 17 in breast cancer and we present our breast cancer data set.

6. In Chapter 6 the performance of *HMM*s on our breast cancer data set is shown and the results are compared with published literature.

7. In Chapter 7 we summarise and discuss the results of Chapter 6 and we give an outlook to future research work in relation to our novel *HMM*s.

This chapter continues with some basics about cancer and microarray experiments.

1.1 Molecular Basics Of Cancer

About ten million cells divide in a human body per minute. Naturally the cell division takes places without problems in strictly controlled patterns. When mutations in a cell occur, then this cell can show abnormal behaviour. In most cases these mutations do not have consequences for the organism, because the affected cell normally dies off and the surrounding cells countervail this loss. The most important exception of this observation are mutations in genes which regulate the cellular proliferation. Changes in such genes can cause the uncontrolled proliferation of a single cell, which is then called a *cancerous cell*, and these changes can lead to the development of *cancer*. Cancer is a disease which is characterised by uncontrolled cell division and it often occurs that some of the cancerous cells have the ability to invade in other tissues either through invasion into adjacent tissue or through implantation into distant sites. A *tumour* develops from a single cancerous cell and consits of a huge number of such cells. These development is an individual evolutionary process with the aim to achieve a better independent cell proliferation. It is estimated that five or six mutations in a cell are necessary to show the phenotype of a cancerous cell. Two groups of genes are known to play a central role in the development of a tumour. The one group consists of *oncogenes* and the other contains the *tumour suppressor genes*. An oncogene is a mutated gene which increases the malignancy of a tumour cell. Such genes encode for proteins which often work in signal cascades to activate the cell division and therefore characteristic mutations can lead to increased activity or increased expression rates of genes in a signal cascade. A tumour suppressor gene is a gene that reduces the probability that a cell will turn into a cancerous cell. Such genes encode for proteins which can stop the cell cycle and therefore a mutation or deletion of such a gene will increase the probability of the formation of a tumour. In general both alleles that code for a particular protein must be affected before an effect is manifested, but exceptions are also known.

Pathologists classify tumours on the basis of histological parameters as tumour size, tumour grade and nodal status. This classification system is useful, but it can be improved using molecular markers.

The analysis of microarray data as gene expression data and DNA copy number data will play the central role in this diploma thesis where we develop an approach to detect molecular markers.

1.2 Breast Cancer

In the year 2000 twenty-two percent of all cancer among women were breast cancer and fifteen percent of cancer deaths in women were caused by this disease. It is estimated that three hundred twenty thousand women come down with breast cancer in Europe each year. Several risk factors for breast cancer as age, reproductive factors, familial history of breast cancer and others are intensively studied in literature. The works of Campbell [2] and Dumitrescu *et al.* [5] offer a good introduction to statistics and risk factors in connection with breast cancer. The *National Cancer Institute* has estimated the average chance for a woman of being diagnosed with breast cancer for the whole population under consideration of the age. The results are shown in the Table 1.1 and it is clearly to see that the probability for coming down with breast cancer increases with the age.

Age	Chance
30 - 39	1 in 229
40 - 49	1 in 68
50 - 59	1 in 37
60 - 69	1 in 26

Table 1.1: A woman's chance of being diagnosed with breast cancer. These probabilities were estimated by the *National Cancer Institute* (www.cancer.gov).

Various changes in DNA regions and gene expression patterns in breast cancer cells in comparison with normal cells have been found with the help of microarray technologies.

Nevertheless, there is a great requirement for efficient analysis methods on the basis of more realistic models to discover such alterations. Therefore we will develop an approach and test the performance of this novel strategy on microarray data of breast cancer cells.

1.3 Microarrays

For a better understanding of processes in cancerous cells it is necessary to have techniques to analyse the differences between these cells and normal cells. Today microarray technologies are standard in most of the labs. These technologies provide a good framework for the genome-wide search for oncogenes and tumour suppressor genes. Let us first consider how microarrays are used to measure expression levels of genes and afterwards we introduce how microarrays are used to detect DNA copy number changes.

1.3.1 Measuring Gene Expression With Microarrays

Microarrays for gene expression analysis can be divided into *cDNA microarrays* and *oligonucleotide microarrays*. The basis of cDNA microarrays are on the microarray fixed DNA molecules which are derived from RNA transcripts. The oligonucleotide microarrays are manufactured either in a photolithographic process which directly synthesises oligonucleotides on the glass slide or oligonucleotides are deposited onto a glass slide. The molecules fixed on the microarray are called *probes*. The information about the expression levels of genes in a cell is contained in the *transcriptome* which consists of all RNA molecules in a cell at a particular time. The expression level of an individual gene is given by the number of RNA molecules for this gene in the transcriptome. Normally, a RNA pool is generated using the transcriptomes of many cells with the same expression status in the optimal case. The RNAs for genes of interest are rewritten in cDNAs and thereby labelled with a fluorescent dye. The rewritten cDNAs of a gene are called *targets*. The usage of different dyes for the labelling of targets allows the parallel measurement of the relative expression status of two different tissue samples. This measurement is done by competitive hybridisation of the target molecules from the two tissue samples with the specifically designed probes on the microarray. Afterwards the microarray undergoes a wash step to remove unspecific attachments. It is possible to determine the relative intensity for each gene in the two different labelled tissue samples, because the probes on a microarray are specifically designed for the targets. The normalisation of the relative intensity for each gene in the two tissue samples makes the gene expression levels of these two tissue samples comparable.

An introduction to cDNA microarrays and oligonucleotide microarrays is given in the review of Garnis *et al.* [7]. The technical aspects of cDNA microarrays are described in the review of Duggan *et al.* [4].

1.3.2 Measuring DNA Copy Number Changes With Microarrays

Microarray techniques to measure DNA copy number changes are summarised in the class of *Array Comparative Genomic Hybridisation* (*ArrayCGH*) approaches. *ArrayCGH* was developed to improve the quality of *Comparative Genomic Hybridisation* (*CGH*) and is now a widespread technique to detect chromosomal imbalances in tumour tissues.

The *CGH* technique was first described by Kallioniemi *et al.* 1992 [11] and technically improved by Kallioniemi *et al.* 1994 [12]. This technique is used to detect segmental DNA copy number changes. Tumour DNA and an identical amount of control DNA are differentially labelled by nick translation. Afterwards these two samples are mixed together for the following denaturation. The resulting sample is hybridised with denatured normal lymphocyte metaphase chromosomes on a slide. After several hours of hybridisation the intensities along the chromosome can be determined for the tumour DNA and the control DNA. *CGH* was the first efficient approach to scan a genome for variations in DNA copy number. A disadvantage of this approach is the low resolution which is nowadays between three and five million base pairs for the minimal detectable segment alteration and therefore it is difficult to determine alterations for specific genes.

The *ArrayCGH* approach was developed for the mapping of chromosomal imbalances at a higher resolution in comparison with *CGH*. Pollack *et al.* [23] have shown the usage of cDNA microarrays for analysing probes which are derived from genomic DNA. The problem of this approach is the suboptimal hybridisation of the targets to the probes on the microarray, because the genomic targets have introns which are absent in the cDNA probes. Another technique is *Bacterial Artificial Chromosome ArrayCGH* which also allows the detection of segmental copy number changes at a higher resolution as *CGH*. This technique is similar to *CGH* except that it uses segments of human DNA as hybridisation probes instead of prepared metaphase chromosomes. In the review of Gebhart [8] the *CGH* technique is explained in more detail and the data which was created by this approach in the last years is discussed. The review of Garnis *et al.* [7] gives a good introduction to *CGH* and *ArrayCGH*.

Chapter 2

Hidden Markov Models

The majority of papers on Hidden Markov Models (*HMM*s) belong to the speech recognition literature where *HMM*s were applied first in the early 1970s. A general tutorial to this topic is the review by Rabiner [24] which covers also the history of *HMM*s. Many problems in biological sequence analysis have the same structure as speech recognition problems. Nowadays *HMM*s are an accepted framework for analysing DNA and protein sequences.

The following topics are contained in this chapter:

1. A short introduction to the basics of *HMM*s is given in the Section 2.1.
2. The homogeneous *HMM*s with continuous emissions are introduced in the Section 2.2.
3. The inhomogeneous *HMM*s with continuous emissions and transition classes are introduced in the Section 2.3.
4. The theoretical background of *HMM*s with continuous emissions and transition classes is presented in the Section 2.4.
5. The basic questions for *HMM*s are given in the Section 2.5.

2.1 Basics of Hidden Markov Models

An *HMM* describes a stochastic process. This process contains two coupled sub-processes: a *Markov chain* and an *Emission Process.* The Markov chain consists of a finite set of states which are traversed in discrete time steps. During this traversion each state which is passed through emits an output signal which is called emission. From the standpoint of the observer the Markov chain is in a black box and therefore not visible. The Markov chain inside this black box is called *hidden Markov chain.* The only thing the observer perceives are the output signals from the black box. An *HMM* is plotted as a directed graph where the states of the hidden Markov chain are represented by the nodes of this graph and the directed edges comply to the transitions between the states of the Markov chain. Every edge is labelled with the transition probability from the outgoing state to the incoming state of this edge. Edges with a transition probability of zero are not shown. A stochastic function for emissions and a start probability are assigned to each state of the *HMM*. In general the following assumptions for an *HMM* are made:

1. The transition probability to move from a state to an other state depends only on that two states and not on the previously traversed states of the Markov chain. This is called the *first-order Markov property.*

2. The emission probabilities of a state depend only on this state.

3. The Markov chain is *time-homogeneous*. That is, the transition probability does not depend on the point in time when a transition is done.

*HMM*s can be divided into two basic classes. In the one class are the *HMM*s with discrete emissions and the other class contains the *HMM*s with continuous emissions. For the analysis of microarray data we are only interested in *HMM*s with continuous emissions and therefore we describe this model class in more detail.

2.2 Hidden Markov Models With Continuous Emissions

First we present some basic knowledge about the probability density functions for emissions (*PDFE*) and afterwards we define *HMM*s with continuous emissions. Reestimation[1] formulas for the parameters of the *PDFE* are known for normal distributions and mixtures of normal distributions [14]. In practice often mixtures of univariate normal distributions are used, and it is known that each univariate density function can be approximated by this function class. The parameter vector of a mixture which consists of N_i normal distributions is

$$\vec{E}_i := ((\mu_1^{(i)}, \sigma_1^{(i)}, \alpha_1^{(i)}), \ldots, (\mu_{N_i}^{(i)}, \sigma_{N_i}^{(i)}, \alpha_{N_i}^{(i)})), \text{ where} \tag{2.1}$$

$$1 = \sum_{k=1}^{N_i} \alpha_k^{(i)} \text{ and } \alpha_k^{(i)} \geq 0.$$

Here, $\mu_k^{(i)} \in \mathbb{R}$ is the mean, $\sigma_k^{(i)} \in \mathbb{R}^+$ is the standard deviation and $\alpha_k^{(i)} \in (0,1)$ is the weight of the k-th mixture component. The density $\mathbb{M}(x|\vec{E}_i)$ of an emission x under a mixture model of univariate normal distributions with the parameter vector $\vec{E}_i$ is defined as

$$\mathbb{M}(x|\vec{E}_i) := \sum_{k=1}^{N_i} \alpha_k^{(i)} \cdot \frac{1}{\sigma_k^{(i)}\sqrt{2\pi}} e^{-\frac{(x-\mu_k^{(i)})^2}{2(\sigma_k^{(i)})^2}}.$$

Later we will denote with i a state in an *HMM*. An emission sequence $\mathcal{O}$ of length T for an *HMM* with univariate continuous emissions is defined as

$$\mathcal{O} := \mathcal{O}_1, \ldots, \mathcal{O}_T \tag{2.2}$$

for each $\mathcal{O}_t \in \mathbb{R}$ with $t \in \{1, \ldots, T\}$.

Definition 2.2.1. *An Hidden Markov Model λ with continuous emissions and N states is a four tuple $(\vec{\pi}, \mathcal{N}, A, \vec{E})$ whose components are:*

1. *The start distribution $\vec{\pi} = (\pi_1, \ldots, \pi_N)$, where π_i is the probability to start an hidden Markov chain in state i.*

2. *The set of states $\mathcal{N} = \{1, \ldots, N\}$, where $i \in \mathcal{N}$ is a state of λ.*

3. *The stochastic transition matrix A, where the entry a_{ij} represents the probability to go from state i to state j.*

4. *The parameter vector $\vec{E} = (\vec{E}_1, \ldots, \vec{E}_N)$ of the probability density functions for emissions, where $\vec{E}_i$, which is defined in (2.1), represents the parameter vector of state i.*

[1] reestimation: iterative update and improvement [24]

An internal state sequence $\mathcal{Q}$ of length T is denoted by

$$\mathcal{Q} := \mathcal{Q}_1, \ldots, \mathcal{Q}_T \tag{2.3}$$

for each $\mathcal{Q}_t \in \mathcal{N}$ with $t \in \{1, \ldots, T\}$. The emission $\mathcal{O}_t$ is emitted by the mixture $\mathbb{M}(x|\vec{E}_{\mathcal{Q}_t})$ of state $\mathcal{Q}_t$. The general definitions for the start probability π_i, the transition probability a_{ij} and the emission $\mathcal{O}_t$ are:

$$\begin{aligned}
\forall i \in \mathcal{N} : \quad & \pi_i := P[\mathcal{Q}_1 = i|\lambda], \\
\forall i, j \in \mathcal{N} : \forall t \in \mathbb{N} : \quad & a_{ij} := P[\mathcal{Q}_{t+1} = j|\mathcal{Q}_t = i, \lambda], \\
\forall t \in \mathbb{N} : \forall j \in \mathcal{N} : \forall x \in \mathbb{R} : \quad & \mathbb{M}(x|\vec{E}_{\mathcal{Q}_t}) := P[\mathcal{O}_t = x|\mathcal{Q}_t = j, \lambda].
\end{aligned}$$

2.3 Hidden Markov Models With Continuous Emissions And Transition Classes

The basis of *HMM*s with continuous emissions and transition classes are *HMM*s with continuous emissions. The concept to use one transition matrix which is also called transition class is extended to use a set of transition matrices $\mathcal{A} := \{A_1, \ldots, A_L\}$. We assume that the probability for a transition from a state to an other state at time step t depends on the for the time step t predefined transition class which has to be element of $\mathcal{A}$. This is modelled by mapping additional information, like the chromosomal distance between consecutive genes in an emission sequence, to one of the predefined transition classes. Hence, for an emission sequence $\mathcal{O}$ we must know a transition class sequence $\mathcal{C}_{\mathcal{O}}$ which is defined by

$$\mathcal{C}_{\mathcal{O}} := c_1, \ldots, c_T \tag{2.4}$$

for each $c_t \in \{1, \ldots, L\}$ with $t \in \{1, \ldots, T\}$. The *HMM* with continuous emissions and transition classes always needs the transition class sequence $\mathcal{C}_{\mathcal{O}}$ to generate or to analyse an emission sequence $\mathcal{O}$. In both cases the transition class sequence $\mathcal{C}_{\mathcal{O}}$ is external previous knowledge which is added to the *HMM*.
The *HMM* with continuous emissions and transition classes is *time-inhomogeneous*, because the transition probabilities depend on the point in time when a transition is done.

Definition 2.3.1. *An Hidden Markov Model λ with continuous emissions, L transition classes and N states is a four tuple $(\vec{\pi}, \mathcal{N}, \mathcal{A}, \vec{E})$ whose components are:*

1. *The start distribution $\vec{\pi} = (\pi_1, \ldots, \pi_N)$, where π_i is the probability to start an hidden Markov chain in state i.*

2. *The set of states $\mathcal{N} = \{1, \ldots, N\}$, where $i \in \mathcal{N}$ is a state of λ.*

3. *The set of transition classes $\mathcal{A} = \{A_1, \ldots, A_L\}$ consisting of stochastic transition matrices, where $a_{ij}(A_l)$ represents the probability to go from state i to state j in the transition class A_l.*

4. *The parameter vector $\vec{E} = (\vec{E}_1, \ldots, \vec{E}_N)$ of the probability density functions for emissions, where $\vec{E}_i$, which is defined in (2.1), represents the parameter vector of state i.*

5. *The implicit transition class sequence $\mathcal{C}_{\mathcal{O}}$ which is a variable part of the model parameters and used by the HMM to generate or to analyse an emission sequence.*

For the internal state sequence $\mathcal{Q}$, the emission sequence $\mathcal{O}$, the start probability π_i and the mixture $\mathbb{M}(x|\vec{E}_i)$ for state i the same definitions as for *HMM*s with continuous emissions are used. The transition probability for a transition from state i to state j at time point t depends on the predefined transition class c_t,

$$\forall t \in \mathbb{N} : \forall i, j \in \mathcal{N} : \quad a_{ij}(c_t) := a_{ij}(A_{c_t}) := P[\mathcal{Q}_{t+1} = j | \mathcal{Q}_t = i, \mathcal{C}_{\mathcal{O}}, \lambda].$$

In the following section we introduce necessary notations which we will need later for the modification of *HMM*s with continuous emissions and transition classes.

2.4 Specific Definitions For Hidden Markov Models

The following definitions are for *HMM*s with continuous emissions and transition classes. These definitions can be used to answer basic questions for given emission sequences or they are the basis for modifications of *HMM*s.

Definition 2.4.1. *The density of an emission sequence $\mathcal{O}$ (2.2) under a given transition class sequence $\mathcal{C}_{\mathcal{O}}$ (2.4) and a given HMM λ (2.3.1) is*

$$\mathbb{D}[\mathcal{O}|\mathcal{C}_{\mathcal{O}}, \lambda] = \sum_{\mathcal{Q} \in \mathcal{N}^T} \pi_{\mathcal{Q}_1} \mathbb{M}(\mathcal{O}_1 | \vec{E}_{\mathcal{Q}_1}) \prod_{t=2}^{T} a_{\mathcal{Q}_{t-1}\mathcal{Q}_t}(c_{t-1}) \mathbb{M}(\mathcal{O}_t | \vec{E}_{\mathcal{Q}_t}).$$

Definition 2.4.2. *The density of an emission sequence $\mathcal{O}$ (2.2) and a state sequence $\mathcal{Q}$ (2.3) under a given transition class sequence $\mathcal{C}_{\mathcal{O}}$ (2.4) and a given HMM λ (2.3.1) is*

$$\mathbb{D}[\mathcal{O}, \mathcal{Q}|\mathcal{C}_{\mathcal{O}}, \lambda] = \pi_{\mathcal{Q}_1} \mathbb{M}(\mathcal{O}_1 | \vec{E}_{\mathcal{Q}_1}) \prod_{t=2}^{T} a_{\mathcal{Q}_{t-1}\mathcal{Q}_t}(c_{t-1}) \mathbb{M}(\mathcal{O}_t | \vec{E}_{\mathcal{Q}_t}).$$

Using the logarithm on the expression $\mathbb{D}[\mathcal{O}, \mathcal{Q}|\mathcal{C}_{\mathcal{O}}, \lambda]$ transforms this expression into three independent parameter-terms

$$\log(\mathbb{D}[\mathcal{O}, \mathcal{Q}|\mathcal{C}_{\mathcal{O}}, \lambda]) = \log(\pi_{\mathcal{Q}_1}) + \sum_{t=1}^{T-1} \log(a_{\mathcal{Q}_t\mathcal{Q}_{t+1}}(c_t)) + \sum_{t=1}^{T} \log(\mathbb{M}(\mathcal{O}_t | \vec{E}_{\mathcal{Q}_t})).$$

Definition 2.4.3. *For two HMMs λ and λ^* as in Definition 2.3.1 the Q-function $\mathbb{Q}(\lambda^*|\lambda)$ is defined as*

$$\mathbb{Q}(\lambda^*|\lambda) := \sum_{\mathcal{Q} \in \mathcal{N}^T} \mathbb{D}[\mathcal{O}, \mathcal{Q}|\mathcal{C}_{\mathcal{O}}, \lambda] \log(\mathbb{D}[\mathcal{O}, \mathcal{Q}|\mathcal{C}_{\mathcal{O}}, \lambda^*]).$$

In the Definition 2.4.2 we have an expression for $\mathbb{D}[\mathcal{O}, \mathcal{Q}|\mathcal{C}_{\mathcal{O}}, \lambda^*]$ which splits the Q-function into *three independent parts:*

$$\begin{aligned}
\mathbb{Q}(\lambda^*|\lambda) &= \sum_{\mathcal{Q}\in\mathcal{N}^T} \mathbb{D}[\mathcal{O},\mathcal{Q}|\mathcal{C}_\mathcal{O},\lambda]\log(\pi^*_{\mathcal{Q}_1}) + \sum_{\mathcal{Q}\in\mathcal{N}^T} \mathbb{D}[\mathcal{O},\mathcal{Q}|\mathcal{C}_\mathcal{O},\lambda]\sum_{t=1}^{T-1}\log(a^*_{\mathcal{Q}_t\mathcal{Q}_{t+1}}(c_t)) \\
&\quad + \sum_{\mathcal{Q}\in\mathcal{N}^T} \mathbb{D}[\mathcal{O},\mathcal{Q}|\mathcal{C}_\mathcal{O},\lambda]\sum_{t=1}^{T}\log(\mathbb{M}(\mathcal{O}_t|\vec{E}^*_{\mathcal{Q}_t})) \\
&= \mathbb{Q}_S(\vec{\pi}^*|\lambda) + \mathbb{Q}_A(\mathcal{A}|\lambda) + \mathbb{Q}_E(\vec{E}^*|\lambda), \text{ where} \qquad (2.5)\\
\mathbb{Q}_S(\vec{\pi}^*|\lambda) &:= \sum_{\mathcal{Q}\in\mathcal{N}^T} \mathbb{D}[\mathcal{O},\mathcal{Q}|\mathcal{C}_\mathcal{O},\lambda]\log(\pi^*_{\mathcal{Q}_1}), \qquad (2.6)\\
\mathbb{Q}_A(\mathcal{A}^*|\lambda) &:= \sum_{\mathcal{Q}\in\mathcal{N}^T} \mathbb{D}[\mathcal{O},\mathcal{Q}|\mathcal{C}_\mathcal{O},\lambda]\sum_{t=1}^{T-1}\log(a^*_{\mathcal{Q}_t\mathcal{Q}_{t+1}}(c_t)) \text{ and} \qquad (2.7)\\
\mathbb{Q}_E(\vec{E}^*|\lambda) &:= \sum_{\mathcal{Q}\in\mathcal{N}^T} \mathbb{D}[\mathcal{O},\mathcal{Q}|\mathcal{C}_\mathcal{O},\lambda]\sum_{t=1}^{T}\log(\mathbb{M}(\mathcal{O}_t|\vec{E}^*_{\mathcal{Q}_t})). \qquad (2.8)
\end{aligned}$$

Definition 2.4.4. *The Forward-Variable $\alpha_t(i)$ represents the density for observing $\mathcal{O}_1^t = \mathcal{O}_1,\ldots,\mathcal{O}_t \sqsubseteq \mathcal{O}$ (2.2) and being in state i at time step t for a given transition class sequence $\mathcal{C}_\mathcal{O}$ (2.4) and an HMM λ (2.3.1),*

$$\alpha_t(i) := \mathbb{D}[\mathcal{O}_1,\ldots,\mathcal{O}_t,\mathcal{Q}_t = i|\mathcal{C}_\mathcal{O},\lambda].$$

Definition 2.4.5. *The Backward-Variable $\beta_t(i)$ represents the density for observing $\mathcal{O}_{t+1}^T = \mathcal{O}_{t+1},\ldots,\mathcal{O}_T \sqsubseteq \mathcal{O}$ (2.2) given the HMM λ (2.3.1) which was in state i at time step t and the transition class sequence $\mathcal{C}_\mathcal{O}$ (2.4),*

$$\beta_t(i) := \mathbb{D}[\mathcal{O}_{t+1},\ldots,\mathcal{O}_T|\mathcal{Q}_t = i,\mathcal{C}_\mathcal{O},\lambda].$$

Definition 2.4.6. *The probability $\varepsilon_t(i,j)$ for going from state i to state j at time step t given an emission sequence $\mathcal{O}$ (2.2), a transition class sequence $\mathcal{C}_\mathcal{O}$ (2.4) and an HMM λ (2.3.1) is defined as*

$$\varepsilon_t(i,j) := \mathbb{P}[\mathcal{Q}_t = i,\mathcal{Q}_{t+1} = j|\mathcal{O},\mathcal{C}_\mathcal{O},\lambda].$$

It is possible to compute $\varepsilon_t(i,j)$ with the help of the Forward-Variable $\alpha_t(i)$ and the Backward-Variable $\beta_{t+1}(j)$,

$$\varepsilon_t(i,j) = \frac{\alpha_t(i)\cdot a_{ij}(c_t)\cdot \mathbb{M}(\mathcal{O}_{t+1}|\vec{E}_j)\cdot\beta_{t+1}(j)}{\mathbb{D}[\mathcal{O}|\lambda]}.$$

Definition 2.4.7. *The probability $\gamma_t(i)$ for being in state i at time step t given an emission sequence $\mathcal{O}$ (2.2), a transition class sequence $\mathcal{C}_\mathcal{O}$ (2.4) and an HMM λ (2.3.1) is*

$$\gamma_t(i) := \mathbb{P}[\mathcal{Q}_t = i|\mathcal{O},\mathcal{C}_\mathcal{O},\lambda].$$

It is possible to compute $\gamma_t(i)$ with the help of the Forward-Variable $\alpha_t(i)$ and the Backward-Variable $\beta_t(i)$ and therefore we obtain

$$\gamma_t(i) = \frac{\alpha_t(i)\cdot\beta_t(i)}{\mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O},\lambda]}.$$

In the next chapter we consider the function $\mathbb{Q}_A(\mathcal{A}^*|\lambda)$ (2.7) in more detail and therefore we show how to transform this function. Recall the definition of $\mathbb{Q}_A(\mathcal{A}^*|\lambda)$ which is

$$\mathbb{Q}_A(\mathcal{A}^*|\lambda) = \sum_{\mathcal{Q}\in\mathcal{N}^T} \mathbb{D}[\mathcal{O},\mathcal{Q}|\mathcal{C}_{\mathcal{O}},\lambda] \sum_{t=1}^{T-1} \log(a^*_{\mathcal{Q}_t\mathcal{Q}_{t+1}}(c_t)).$$

First we sum over all N^2 possibilities to go from state $\mathcal{Q}_t = i$ to state $\mathcal{Q}_{t+1} = j$ at time step t and so we obtain

$$\mathbb{Q}_A(\mathcal{A}^*|\lambda) = \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T-1} \log(a^*_{ij}(c_t)) \sum_{\substack{\mathcal{Q}\in\mathcal{N}^T \\ \mathcal{Q}_t=i\wedge\mathcal{Q}_{t+1}=j}} \mathbb{D}[\mathcal{O},\mathcal{Q}|\mathcal{C}_{\mathcal{O}},\lambda].$$

In the next step we rewrite the sum over all state sequences $\mathcal{Q}$ with the help of a simplification and the Definition 2.4.6

$$\begin{aligned}\sum_{\substack{\mathcal{Q}\in\mathcal{N}^T \\ \mathcal{Q}_t=i\wedge\mathcal{Q}_{t+1}=j}} \mathbb{D}[\mathcal{O},\mathcal{Q}|\mathcal{C}_{\mathcal{O}},\lambda] &= \mathbb{D}[\mathcal{O},\mathcal{Q}_t=i,\mathcal{Q}_{t+1}=j|\mathcal{C}_{\mathcal{O}},\lambda] \\ &= \mathbb{D}[\mathcal{O}|\mathcal{C}_{\mathcal{O}},\lambda]\mathbb{P}[\mathcal{Q}_t=i,\mathcal{Q}_{t+1}=j|\mathcal{O},\mathcal{C}_{\mathcal{O}},\lambda] \\ &= \mathbb{D}[\mathcal{O}|\mathcal{C}_{\mathcal{O}},\lambda]\varepsilon_t(i,j).\end{aligned}$$

As result we obtain

$$\mathbb{Q}_A(\mathcal{A}^*|\lambda) = \mathbb{D}[\mathcal{O}|\mathcal{C}_{\mathcal{O}},\lambda]\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T-1} \varepsilon_t(i,j)\log(a^*_{ij}(c_t)). \tag{2.9}$$

Similar transformations of the functions $\mathbb{Q}_S(\vec{\pi}^*|\lambda)$ (2.6) and $\mathbb{Q}_E(\vec{E}^*|\lambda)$ (2.8) can be done.

2.5 Basic Questions For Continuous Hidden Markov Models

The application of *HMM*s in the modelling of emission sequences needs efficient methods to solve the following basic questions:

1. What is the density $\mathbb{D}[\mathcal{O}|\lambda]$ for an emission sequence $\mathcal{O}$ given an *HMM* λ? The *Forward algorithm* is an efficient method to solve this problem.

2. What state sequence $\mathcal{Q}$ is the most probable state sequence given an emission sequence $\mathcal{O}$ and an *HMM* λ? So we want to compute the *Viterbi Path* $\mathcal{Q}^*$ which is defined by

$$\mathcal{Q}^* = \underset{\mathcal{Q}}{\text{argmax}}\ \mathbb{P}[\mathcal{Q}|\mathcal{O},\lambda].$$

   This problem is efficiently solved by the *Viterbi algorithm*.

3. How is an emission sequence $\mathcal{O}$ used to optimise the parameters of a given *HMM* λ so that the density $\mathbb{D}[\mathcal{O}|\lambda]$ is maximised? The well known *Baum-Welch algorithm* is used to solve this problem.

The solutions of these three problems are discussed for example by Durbin *et al.* [6], Knab [14] or Rabiner [24]. The *Forward algorithm*, the *Viterbi algorithm* and the *Baum-Welch algorithm* can be easily adapted to the *HMM*s with continuous emissions and transition classes. We can

solve the basic questions for this class of *HMM*s on the basis of the Definitions for the *Forward-Variables* (2.4.4) and the *Backward-Variables* (2.4.5).
The *General Hidden Markov Model library (GHMM)* (www.ghmm.org) includes the basic algorithms for *HMM*s with continuous emissions and for *HMM*s with continuous emissions and transition classes. We will use the *GHMM* as basis of the analysis of microarray data and we will develop a modification for *HMM*s with continuous emissions and transition classes which could be integrated in a new release of the *GHMM*.

Chapter 3

Hidden Markov Models With Coupled Transition Matrices

The inhomogeneous *HMM*s with transition classes extend the standard homogeneous *HMM*s. More realistic models of biological phenomena can be created with the help of transition classes. Nevertheless, the number of model parameters increases with every additional transition class. To have the advantages of transition classes and to reduce the number of model parameters we develop an *HMM* with coupled transition classes in this chapter.

The following topics are contained in this chapter:

1. The basics of coupled transition matrices are given in the Section 3.1.
2. The definition of *HMM*s with coupled transition classes is given in the Section 3.2.
3. The optimisation of transition parameters for *HMM*s with coupled transition classes is shown in the Section 3.3.
4. An overview of the *Baum-Welch algorithm* and a general proof of the convergence are presented in the Section 3.4.
5. Analytical solutions for the reestimation formulas of the transition parameters which proves the equivalence of *HMM*s with one coupled transition class and standard *HMM*s are derived in the Section 3.5.
6. Analytical solutions for the reestimation formulas of the transition parameters for *HMM*s with two coupled transition classes are given in the Section 3.6.

3.1 Coupling Of Transition Matrices

Given a stochastic transition matrix A_1 and a scaling vector $\vec{S} = (S_1 := 1, S_2, \ldots, S_L)$ with the scaling parameters $S_1 < S_2 < \ldots < S_L$, $S_l \in \mathbb{R}^+$, we want to create a mapping which generates transition matrices on the basis of A_1 and $\vec{S}$. Later we need the transition probabilities of A_1 and therefore we define

$$A_1 := \begin{pmatrix} a_{11}(1) & & \ldots & & a_{1N}(1) \\ \vdots & \ddots & & & \vdots \\ a_{i1}(1) & \ldots & a_{ii}(1) & \ldots & a_{iN}(1) \\ \vdots & & & \ddots & \vdots \\ a_{N1}(1) & & \ldots & & a_{NN}(1) \end{pmatrix}. \tag{3.1}$$

The expected state duration $d_i^{(1)}$ for state i of A_1 is given by

$$d_i^{(1)} := \frac{1}{1 - a_{ii}(1)}. \tag{3.2}$$

The Definition (3.2) follows directly from the mean of the geometric distribution for staying in state i. The vector of all expected state durations for A_1 is

$$\vec{d}^{(1)} := (d_1^{(1)}, \ldots, d_N^{(1)}).$$

The scaling parameter S_l is used to calculate the expected state durations of the transition matrix A_l and therefore we obtain

$$\vec{d}^{(l)} := S_l \cdot \vec{d}^{(1)}. \tag{3.3}$$

When we look at the definition of the scaling parameters, it follows the vector $\vec{d}^{(1)}$ is the one with the lowest expected state durations and the vector $\vec{d}^{(L)}$ is the one with the highest expected state durations. With the help of the Definition (3.3) and the generalisation of the Definition (3.2) we can compute a diagonal element $a_{ii}(l)$ with $i \in \{1, \ldots, N\}$ of A_l,

$$d_i^{(l)} = \frac{1}{1 - a_{ii}(l)} \Leftrightarrow a_{ii}(l) = 1 - \frac{1}{d_i^{(l)}} \Leftrightarrow a_{ii}(l) = 1 - \frac{1}{S_l d_i^{(1)}}, \text{ and here}$$

$$a_{ii}(l) = \frac{a_{ii}(1) - 1 + S_l}{S_l}. \tag{3.4}$$

The non-diagonal elements $a_{ij}(l)$ of A_l, with $i, j \in \{1, \ldots, N\}$ and $i \neq j$, are computed by multiplying the non-diagonal elements $a_{ij}(1)$ of A_1 with a scaling factor $m_i(l)$. We know that A_1 is a stochastic matrix and so each row i of A_1 fulfils the following condition

$$1 = \sum_{j=1}^{N} a_{ij}(1).$$

Now we assume that the parameter $a_{ii}(1)$ is substituted by $a_{ii}(l) \in (0, 1)$. We are searching for a scaling factor $m_i(l) \in \mathbb{R}^+$ of the non-diagonal elements so that the following equation is fulfilled

$$1 = a_{ii}(l) + m_i(l) \sum_{\substack{j=1 \\ j \neq i}}^{N} a_{ij}(1) \Leftrightarrow 1 = a_{ii}(l) + m_i(l)(1 - a_{ii}(1)) \Leftrightarrow m_i(l) = \frac{1 - a_{ii}(l)}{1 - a_{ii}(1)}.$$

We use the Equation (3.4) to obtain the final expression

$$m_i(l) = \frac{1}{S_l},$$

and from the definition of the scaling parameters follows that $1 \geq m_i(l) > 0$. With the help of the scaling factor $m_i(l)$ the non-diagonal elements $a_{ij}(l)$ of the transition matrix A_l can be calculated in the following manner

$$a_{ij}(l) = m_i(l) a_{ij}(1) = \frac{a_{ij}(1)}{S_l}. \tag{3.5}$$

The Equations (3.4) and (3.5) are the basis to couple the transition matrix A_l to the basic transition matrix A_1 and so we define

$$A_l := \begin{pmatrix} \frac{a_{11}(1)-1+S_l}{S_l} & & \cdots & & \frac{a_{1N}(1)}{S_l} \\ \vdots & \ddots & & & \vdots \\ \frac{a_{i1}(1)}{S_l} & \cdots & \frac{a_{ii}(1)-1+S_l}{S_l} & \cdots & \frac{a_{iN}(1)}{S_l} \\ \vdots & & & \ddots & \vdots \\ \frac{a_{N1}(1)}{S_l} & & \cdots & & \frac{a_{NN}(1)-1+S_l}{S_l} \end{pmatrix}. \tag{3.6}$$

The matrix A_l is a stochastic matrix on the basis of the stochastic transition matrix A_1 and the predefined scaling parameters $\vec{S}$

$$\begin{aligned} \forall i \in \{1,\ldots,N\}: \quad 1 &= \frac{a_{ii}(1)-1+S_l}{S_l} + \sum_{\substack{j=1 \\ j\neq i}}^{N} \frac{a_{ij}(1)}{S_l} \\ &= \frac{a_{ii}(1)-1+S_l}{S_l} + \frac{1-a_{ii}(1)}{S_l} = 1. \end{aligned}$$

With the help of the transition matrix A_l, which is given in (3.6), we can define the coupling function

$$\mathbb{COUPLE}(A_1, \vec{S}) := (A_1, \ldots, A_L). \tag{3.7}$$

3.2 Definition Of Hidden Markov Models With Coupled Transition Classes

The *HMM*s with coupled transition classes are defined on the basis of the *HMM*s with continuous emissions and transition classes. The transition matrices are coupled with the help of the coupling function (3.7). The standard *HMM* with continuous emissions, N states and L transition classes has LN^2 transition parameters. The extended *HMM* with continuous emissions, N states and L coupled transition classes has also LN^2 transition parameters, but $(L-1)N^2$ of these parameters are coupled to the basis transition matrix and that reduces the number of free transition parameters to N^2.

Definition 3.2.1. *An Hidden Markov Model λ with N states, continuous emissions and L coupled transition classes is a five tuple $(\vec{\pi}, \mathcal{N}, A_1, \vec{S}, \vec{E})$ whose components are:*

1. *The start distribution $\vec{\pi} = (\pi_1, \ldots, \pi_N)$, where π_i is the probability to start an hidden Markov chain in state i.*

2. *The set of states $\mathcal{N} = \{1, \ldots, N\}$, where $i \in \mathcal{N}$ is a state of λ.*

3. *The basic stochastic transition matrix A_1 which is defined in (3.1).*

4. *The vector of scaling parameters $\vec{S} = (S_1 := 1, S_2, \ldots, S_L)$, where S_l is the scaling parameter for the transition matrix A_l. The scaling parameters have to fulfil $S_1 < S_2 < \ldots < S_L$ for $S_1 = 1$ and $S_l \in \mathbb{R}^+$ for all $l \in \{2, \ldots, L\}$.*

5. *The vector of parameters $\vec{E} = (\vec{E}_1, \ldots, \vec{E}_N)$ of the probability density functions for emissions, where $\vec{E}_i$, which is defined in (2.1), represents the parameter vector of state i.*

6. *The implicit generated L transition classes $\mathcal{A} = \mathbb{COUPLE}(A_1, \vec{S})$. The coupling function $\mathbb{COUPLE}$ is defined in (3.7).*

7. *The implicit transition class sequence $\mathcal{C}_{\mathcal{O}}$ (2.4) which is a variable part of the model parameters and used by the HMM to generate or to analyse an emission sequence.*

3.3 Estimation Of Transition Matrices

The basis of the estimation of transition parameters is the optimisation of the function $\mathbb{Q}_A(\mathcal{A}^*|\lambda)$. The *HMM*s in the Definition 3.2.1 use coupled transition matrices which depend on the transition matrix A_1 and the scaling vector $\vec{S}$. Our aim is to maximise the function $\mathbb{Q}_A(\mathcal{A}^*|\lambda)$ which is henceforth written as $\mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda)$ to refer to the coupling of the transition matrices.
Recall the function

$$\mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda) = \mathbb{D}[O|\mathcal{C}_{\mathcal{O}}, \lambda] \left(\sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T-1} \varepsilon_t(i,j) \log(a^*_{ij}(c_t)) \right)$$

which is also defined in (2.9). First we split $\mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda)$ into diagonal elements $a^*_{ii}(c_t)$, with $i \in \mathcal{N}$, and non-diagonal elements $a^*_{ij}(c_t)$, with $i, j \in \mathcal{N}$ and $i \neq j$, and so we obtain

$$\mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda) = \mathbb{D}[O|\mathcal{C}_{\mathcal{O}}, \lambda] \left(\sum_{i=1}^{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} \sum_{t=1}^{T-1} \varepsilon_t(i,j) \log(a^*_{ij}(c_t)) + \sum_{i=1}^{N} \sum_{t=1}^{T-1} \varepsilon_t(i,i) \log(a^*_{ii}(c_t)) \right). \quad (3.8)$$

In the next step the diagonal elements $a^*_{ii}(c_t)$ and the non-diagonal elements $a^*_{ij}(c_t)$ are replaced by their Transformations (3.4) and (3.5) and we add also the constraints for the parameter estimation

$$\begin{aligned} \mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda) = \mathbb{D}[O|\mathcal{C}_{\mathcal{O}}, \lambda] & \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} \sum_{t=1}^{T-1} \varepsilon_t(i,j) \log\left(\frac{a^*_{ij}(1)}{S_{c_t}}\right) \\ & + \mathbb{D}[O|\mathcal{C}_{\mathcal{O}}, \lambda] \sum_{i=1}^{N} \sum_{t=1}^{T-1} \varepsilon_t(i,i) \log\left(\frac{a^*_{ii}(1) - 1 + S_{c_t}}{S_{c_t}}\right) \\ & - \sum_{i=1}^{N} \lambda_i \left(\left(\sum_{j=1}^{N} a^*_{ij}(1) \right) - 1 \right). \end{aligned}$$

Now we are able to compute the partial derivatives of the non-diagonal elements $a^*_{ij}(1)$ and the diagonal elements $a^*_{ii}(1)$.

$$\forall i, j \in \mathcal{N} \wedge i \neq j : \quad \frac{\partial \mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda)}{\partial a^*_{ij}(1)} = \mathbb{D}[O|\mathcal{C}_{\mathcal{O}}, \lambda] \left(\sum_{t=1}^{T-1} \varepsilon_t(i,j) \frac{1}{a^*_{ij}(1)} \right) - \lambda_i \quad (3.9)$$

$$\forall i \in \mathcal{N} : \quad \frac{\partial \mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda)}{\partial a^*_{ii}(1)} = \mathbb{D}[O|\mathcal{C}_{\mathcal{O}}, \lambda] \left(\sum_{t=1}^{T-1} \varepsilon_t(i,i) \frac{1}{a^*_{ii}(1) - 1 + S_{c_t}} \right) - \lambda_i \quad (3.10)$$

We set the Derivative (3.9) equal to zero and bring the parameter $a^*_{ij}(1)$ to the left site to get the zero of this derivative.

$$\forall i, j \in \mathcal{N} \wedge i \neq j : \quad a^*_{ij}(1) = \frac{\mathbb{D}[O|\mathcal{C}_{\mathcal{O}}, \lambda]}{\lambda_i} \sum_{t=1}^{T-1} \varepsilon_t(i,j) \quad (3.11)$$

In the next step we use the constraint to determine an expression for the Lagrange multiplier λ_i. We have to pay attention that we only know values for the non-diagonal elements at the moment.

$$1 = \sum_{j=1}^{N} a^*_{ij}(1) = \left(\sum_{\substack{j=1 \\ j \neq i}}^{N} a^*_{ij}(1) \right) + a^*_{ii}(1) = \left(\sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{\mathbb{D}[O|\mathcal{C}_{\mathcal{O}}, \lambda]}{\lambda_i} \sum_{t=1}^{T-1} \varepsilon_t(i,j) \right) + a^*_{ii}(1)$$

$$\lambda_i = \frac{\mathbb{D}[O|\mathcal{C}_{\mathcal{O}}, \lambda]}{1 - a^*_{ii}(1)} \left(\sum_{t=1}^{T-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \varepsilon_t(i,j) \right) \tag{3.12}$$

The zero a^*_{ij} of the Derivative (3.9) can be rewritten by substituting λ_i in the Equation (3.11) by the Equation (3.12). As result we obtain an expression which depends on the diagonal element $a^*_{ii}(1)$

$$\forall i, j \in \mathcal{N} \wedge i \neq j: \quad a^*_{ij}(1) = \frac{(1 - a^*_{ii}(1)) \sum_{t=1}^{T-1} \varepsilon_t(i,j)}{\sum_{t=1}^{T-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \varepsilon_t(i,j)}. \tag{3.13}$$

We set the Derivative (3.10) equal to zero to obtain an expression which will give us the possibility to determine the zero $a^*_{ii}(1)$ of this derivative. However, it is not possible to bring $a^*_{ii}(1)$ to the left side because $a^*_{ii}(1)$ is bound in the denominator which depends on the parameter t of the sum. So we get the following expression

$$\forall i \in \mathcal{N}: \quad \lambda_i = \mathbb{D}[O|\mathcal{C}_{\mathcal{O}}, \lambda] \sum_{t=1}^{T-1} \varepsilon_t(i,i) \frac{1}{a^*_{ii}(1) - 1 + S_{c_t}}. \tag{3.14}$$

The Lagrange multiplier λ_i is known from the Equation (3.12) and therefore we substitude λ_i in Equation (3.14) and obtain the following equation after dividing by $\mathbb{D}[O|\mathcal{C}_{\mathcal{O}}, \lambda]$ and multiplying with $1 - a^*_{ii}(1)$.

$$\forall i \in \mathcal{N}: \quad \sum_{t=1}^{T-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \varepsilon_t(i,j) = \sum_{t=1}^{T-1} \varepsilon_t(i,i) \frac{1 - a^*_{ii}(1)}{a^*_{ii}(1) - 1 + S_{c_t}} \tag{3.15}$$

The Equation (3.15) is the starting point to determine the zeros of the diagonal elements. For one ($L = 1$) or two ($L = 2$) transition classes the analytical solution can be found as we will see later. Cases with more than two ($L > 2$) transition classes are too complex and so we have to use a numerical procedure to find a solution. However, we want to know how many solutions of the Equation (3.15) exist and whether we get such a solution in every case. To find answers we transform the Equation (3.15).

$$\sum_{t=1}^{T-1}\sum_{\substack{j=1\\j\neq i}}^{N}\varepsilon_t(i,j) = -\sum_{t=1}^{T-1}\varepsilon_t(i,i)\frac{1-a_{ii}^*(1)}{1-a_{ii}^*(1)-S_{c_t}}$$

$$\sum_{t=1}^{T-1}\sum_{\substack{j=1\\j\neq i}}^{N}\varepsilon_t(i,j) = -\sum_{t=1}^{T-1}\left(1+\frac{S_{c_t}}{1-a_{ii}^*(1)-S_{c_t}}\right)\varepsilon_t(i,i)$$

$$\sum_{t=1}^{T-1}\gamma_t(i) = -\sum_{t=1}^{T-1}\frac{S_{c_t}}{1-a_{ii}^*(1)-S_{c_t}}\varepsilon_t(i,i)$$

$$\sum_{t=1}^{T-1}\gamma_t(i) = \sum_{t=1}^{T-1}\frac{S_{c_t}}{a_{ii}^*(1)-1+S_{c_t}}\varepsilon_t(i,i) \tag{3.16}$$

$$0 = \underbrace{\sum_{t=1}^{T-1}\frac{S_{c_t}}{a_{ii}^*(1)-1+S_{c_t}}\varepsilon_t(i,i) - \sum_{t=1}^{T-1}\gamma_t(i)}_{=:f_i(a_{ii}^*(1))}$$

We obtain a zero $\hat{a}_{ii}^*(1)$ when the Equation $f_i(a_{ii}^*(1)) = 0$ can be solved. It is necessary to understand the behaviour of $f_i(a_{ii}^*(1))$. We choose a transition matrix A_1 as defined in (3.1) where each transition probability $a_{ij}(1)$ for $i,j \in \mathcal{N}$ is in the interval $(0,1)$ and therefore each $\varepsilon_t(i,j)$ for $i,j \in \mathcal{N}$ is also in the interval $(0,1)$. We start to analyse the first term $f_1^i(a_{ii}^*(1))$ of $f_i(a_{ii}^*(1))$ which is

$$f_1^i(a_{ii}^*(1)) := \sum_{t=1}^{T-1}\frac{S_{c_t}}{a_{ii}^*(1)-1+S_{c_t}}\varepsilon_t(i,i).$$

The function $f_1^i(a_{ii}^*(1))$ consists of a sum of hyperbolas. Each of these hyperbolas has a point of discontinuity for $a_{ii}^*(1) = 1 - S_{c_t}$, but if we look at the definition of the scaling factor S_{c_t} we see that such a point of discontinuity is always less than zero for $c_t > 1$ and equal to zero for $c_t = 1$. These points of discontinuity are excluded because we choose every $a_{ii}^*(1) \in (0,1)$ and therewith the denominator $a_{ii}^*(1) - 1 + S_{c_t}$ of such a hyperbola is always greater than zero.

Proposition 3.3.1. *The function $f_1^i(a_{ii}^*(1))$ is strictly monotonic decreasing in the interval $(0,1)$.*

Proof: We have to prove the following:

$$\forall a_{ii}^*(1) \in (0,1) : \exists \varepsilon \in \mathbb{R}^+ : \quad a_{ii}^*(1)+\varepsilon \in (0,1) \wedge f_1^i(a_{ii}^*(1)) > f_1^i(a_{ii}^*(1)+\varepsilon).$$

To show this we consider the denominator $a_{ii}^*(1) - 1 + S_{c_t}$ of the t-th hyperbola of $f_1^i(a_{ii}^*(1))$ and use the previous knowledge that $a_{ii}^*(1) + \varepsilon > a_{ii}^*(1)$:

$$a_{ii}^*(1)+\varepsilon-1+S_{c_t} > a_{ii}^*(1)-1+S_{c_t}$$

$$\sum_{t=1}^{T-1}\frac{S_{c_t}}{a_{ii}^*(1)-1+S_{c_t}}\varepsilon_t(i,i) > \sum_{t=1}^{T-1}\frac{S_{c_t}}{a_{ii}^*(1)+\varepsilon-1+S_{c_t}}\varepsilon_t(i,i)$$

$$f_1^i(a_{ii}^*(1)) > f_1^i(a_{ii}^*(1)+\varepsilon).$$

□

Proposition 3.3.2. *The limit of $f_1^i(a_{ii}^*(1))$ for $a_{ii}^*(1) \to 0^+$ is ∞ if a $t' \in \{1, \ldots, T\}$ exists which fulfils $c_{t'} = 1$.*

Proof: We assume that the transition class sequence $\mathcal{C}_{\mathcal{O}}$ (2.4) contains a $c_{t'} = 1$. Now we have to recall that $S_1 = 1$ and therefore the denominator $a_{ii}^*(1) - 1 + S_{c_{t'}}$ is zero for $a_{ii}^*(1) = 0$.

$$\lim_{a_{ii}^*(1) \to 0^+} f_1^i(a_{ii}^*(1)) = \lim_{a_{ii}^*(1) \to 0^+} \sum_{t=1}^{T-1} \frac{S_{c_t}}{a_{ii}^*(1) - 1 + S_{c_t}} \varepsilon_t(i,i) \to \infty$$

□

Proposition 3.3.3. *The limit of $f_1^i(a_{ii}^*(1))$ for $a_{ii}^*(1) \to 1^-$ is $\sum_{t=1}^{T-1} \varepsilon_t(i,i)$.*

Proof:

$$\lim_{a_{ii}^*(1) \to 1^-} f_1^i(a_{ii}^*(1)) = \sum_{t=1}^{T-1} \frac{S_{c_t}}{1 - 1 + S_{c_t}} \varepsilon_t(i,i) = \sum_{t=1}^{T-1} \varepsilon_t(i,i) := \mathbb{U}$$

□

Now that we have analysed the behaviour of $f_1^i(a_{ii}^*(1))$ we give a short summary of the results.

- The result of Proposition 3.3.1 is that $f_1^i(a_{ii}^*(1))$ is strictly monotonic decreasing in the interval $(0, 1)$.
- The Proposition 3.3.2 shows that $f_1^i(a_{ii}^*(1))$ can be infinite if at least one $c_t = 1$ in $\mathcal{C}_{\mathcal{O}}$ exists and therefore no upper bound for $f_1^i(a_{ii}^*(1))$ can be found.
- The Proposition 3.3.3 gives us a lower bound of the function $f_1^i(a_{ii}^*(1))$ in the interval $(0, 1)$.
- If at least one $c_t = 1$ in $\mathcal{C}_{\mathcal{O}}$ exists, then it follows from the Propositions 3.3.2, 3.3.3 and 3.3.1 that for each $y \in (\mathbb{U}, \infty)$ only one $a_{ii}^*(1) \in (0, 1)$ exists which fulfils the condition $f_1^i(a_{ii}^*(1)) = y$.

These results will help us to study the behaviour of $f_i(a_{ii}^*(1))$. The second term f_2^i of $f_i(a_{ii}^*(1))$ is a constant which is always greater than zero because each $\varepsilon_t(i,j)$ is greater than zero.

$$f_2^i := \sum_{t=1}^{T-1} \gamma_t(i) = \sum_{t=1}^{T-1} \sum_{j=1}^{N} \varepsilon_t(i,j) > \sum_{t=1}^{T-1} \varepsilon_t(i,i) > 0$$

This implies that the constant f_2^i is always in the interval $(\mathbb{U}, \infty)$. The function $f_i(a_{ii}^*(1))$ is composed of $f_1^i(a_{ii}^*(1))$ and f_2^i and can be written as

$$f_i(a_{ii}^*(1)) = f_1^i(a_{ii}^*(1)) - f_2^i.$$

We assume to have at least one $c_t = 1$ in $\mathcal{C}_{\mathcal{O}}$. In this case we know from the Propositions 3.3.2, 3.3.3 and 3.3.1 that we can exactly find one $\hat{a}_{ii}^*(1) \in (0, 1)$ which fulfils the following condition

$$f_1^i(\hat{a}_{ii}^*(1)) = f_2^i.$$

This $\hat{a}_{ii}^*(1)$ is the only zero of $f_i(\hat{a}_{ii}^*(1))$. We can determine the $\hat{a}_{ii}^*(1)$ for each $f_i(\hat{a}_{ii}^*(1))$ with $i \in \mathcal{N}$ using the same proceeding.
It is interesting to see what happens when we assume that no $c_t = 1$ in $\mathcal{C}_{\mathcal{O}}$ (2.4) exists. In such

a case the upper bound of $f_1^i(a^*_{ii}(1))$ is $\sum_{t=1}^{T-1} \frac{S_{c_t}}{S_{c_t}-1} \varepsilon_t(i,i)$ and we cannot ensure that this sum is greater or equal than f_2^i and so there can be cases where no zero in $(0,1)$ exists.
From now on we only consider the case where we can find the zero for $f_i(a^*_{ii}(1))$. It follows from the Equation (3.13) that each non-diagonal element $\hat{a}^*_{ij}(1)$ depends on the diagonal element $\hat{a}^*_{ii}(1)$. The values of the diagonal elements $\hat{a}^*_{ii}(l)$ and the non-diagonal elements $\hat{a}^*_{ij}(l)$ are calculated by the Equations (3.4) and (3.5). The parameter space $\mathcal{D}$ of the zeros for the diagonal elements $\hat{a}^*_{ii}(1)$ is

$$\mathcal{D} = \{(a^*_{11}(1), \ldots, a^*_{NN}(1)) \mid \forall : 1 \leq i \leq N : a^*_{ii}(1) \in (0,1)\}.$$

We have seen how to determine the $\vec{d} \in \mathcal{D}$ which fulfils $f_1(a^*_{11}(1)) = 0, \ \ldots, \ f_N(a^*_{NN}(1)) = 0$ and we know how we can calculate the critical point $\hat{\mathcal{A}}^*_{\vec{S}}$ of $\mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda)$ (3.8). This critical point is a stochastic transition matrix. To analyse the behaviour of $\mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda)$ at the critical point $\hat{\mathcal{A}}^*_{\vec{S}}$ we consider the Hessian matrix.

$$\begin{aligned} & \forall i, j \in \mathcal{N} \wedge i \neq j : \\ & \frac{\partial \mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda)}{\partial a^*_{ij}(1) \partial a^*_{ij}(1)} = -\mathbb{D}[O|\mathcal{C}_{\mathcal{O}}, \lambda] \sum_{t=1}^{T-1} \varepsilon_t(i,j) \frac{1}{(a^*_{ij}(1))^2} := \delta ij \end{aligned} \tag{3.17}$$

$$\begin{aligned} & \forall i, j, n, m \in \mathcal{N} \wedge i \neq j \wedge n \neq i \wedge m \neq j : \\ & \frac{\partial \mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda)}{\partial a^*_{ij}(1) \partial a^*_{nm}(1)} = 0 \end{aligned}$$

$$\begin{aligned} & \forall i \in \mathcal{N} : \\ & \frac{\partial \mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda)}{\partial a^*_{ii}(1) \partial a^*_{ii}(1)} = -\mathbb{D}[O|\mathcal{C}_{\mathcal{O}}, \lambda] \sum_{t=1}^{T-1} \varepsilon_t(i,i) \frac{1}{(a^*_{ii}(1) - 1 + S_{c_t})^2} := \delta_{ii} \end{aligned} \tag{3.18}$$

$$\begin{aligned} & \forall i, n, m \in \mathcal{N} \wedge n \neq i \wedge m \neq i : \\ & \frac{\partial \mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda)}{\partial a^*_{ii}(1) \partial a^*_{nm}(1)} = 0 \end{aligned}$$

So the resulting Hessian matrix $\mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda)''$ has the following structure.

$$\begin{aligned} \mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda)'' &= \begin{pmatrix} \frac{\partial \mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda)}{\partial a^*_{11}(1) \partial a^*_{11}(1)} & \frac{\partial \mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda)}{\partial a^*_{11}(1) \partial a^*_{12}(1)} & \cdots & \frac{\partial \mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda)}{\partial a^*_{11}(1) \partial a^*_{NN}(1)} \\ \vdots & \vdots & & \vdots \\ \frac{\partial \mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda)}{\partial a^*_{ij}(1) \partial a^*_{11}(1)} & \frac{\partial \mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda)}{\partial a^*_{ij}(1) \partial a^*_{12}(1)} & \cdots & \frac{\partial \mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda)}{\partial a^*_{ij}(1) \partial a^*_{NN}(1)} \\ \vdots & \vdots & & \vdots \\ \frac{\partial \mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda)}{\partial a^*_{NN}(1) \partial a^*_{11}(1)} & \frac{\partial \mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda)}{\partial a^*_{NN}(1) \partial a^*_{12}(1)} & \cdots & \frac{\partial \mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda)}{\partial a^*_{NN}(1) \partial a^*_{NN}(1)} \end{pmatrix} \\ &= \begin{pmatrix} \delta_{11} & 0 & 0 & \ldots & 0 \\ 0 & \delta_{12} & 0 & \ldots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & 0 & \ldots & \delta_{NN} \end{pmatrix} \end{aligned} \tag{3.19}$$

This matrix is symmetric and all δ_{ij} are less than zero as the Functions (3.17) and (3.18) show. Hence, this matrix is negative definite and therefore the critical point $\hat{\mathcal{A}}^*_{\vec{S}}$ maximises $\mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda)$.

This result plays an important role because we will be able to show the convergence of the *Baum-Welch algorithm*.
Before we analyse the convergence of the *Baum-Welch algorithm* we prepare the function $f_i(a^*_{ii}(1))$ for Newton's method. The derivative of $f_i(a^*_{ii}(1))$ is

$$f'_i(a^*_{ii}(1)) = -\sum_{t=1}^{T-1} \frac{S_{c_t}}{(a^*_{ii}(1) - 1 + S_{c_t})^2} \varepsilon_t(i,i)$$

and therewith we can use Newton's method to determine the zero for each $f_i(a^*_{ii}(1))$ with

$$a^*_{ii}(1)^{(t+1)} = a^*_{ii}(1)^{(t)} - \frac{f_i(a^*_{ii}(1)^{(t)})}{f'_i(a^*_{ii}(1)^{(t)})}.$$

3.4 Convergence Of The Baum-Welch Algorithm

The *Baum-Welch algorithm* is a widely used method to optimally adapt model parameters to observed training data and therefore good models for real processes can be created. An analytical way to determine solutions for the model parameters which maximise the probability of the training data is unknown. However, we can use the *Baum-Welch algorithm* to train our *HMM*s. Good introductions for the *Baum-Welch algorithm* are given by Rabiner [24], Knab [14] or Durbin *et al.* [6]. We only recapitulate the general steps of this algorithm.

Baum-Welch algorithm

- Initialisation: Choose initial model parameters.
- Recurrence: Calculate the *Forward-Variables* and the *Backward-Variables* for the training data and determine the new model parameters using known reestimation formulas or numerical procedures.
- Termination: Stop if the change in the likelihood is less than a predefined threshold value or when the maximum number of iterations is exceeded.

The final result of the *Baum-Welch algorithm* is called a *maximum likelihood estimate* of the *HMM*, and the *Baum-Welch algorithm* stops in general in a local maxima. There is no optimal way known how to estimate the model parameters and in most problems the optimisation surface is very complex and has many local maxima [24].
Now that we have an overview of the *Baum-Welch algorithm* we want to prove the convergence of this algorithm. The proof which we develop here can be seen as a general proof of the convergence of the *Baum-Welch algorithm* for *HMM*s with transition classes. The same strategy works also for the standard *HMM*s. The basis of this proof is that we have determined model parameters which improve the likelihood of the training data under the given *HMM*. That is, we have calculated improved start probabilities, improved transition probabilities and improved emission parameters by independent maximisation of the functions $\mathbb{Q}_S(\vec{\pi}^*|\lambda)$ (2.6), $\mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda)$ (2.7) and $\mathbb{Q}_E(\vec{E}^*|\lambda)$ (2.8).
Reestimation formulas for the model parameters are known for the standard *HMM*s in the Definition 2.2.1 and the *HMM*s with transition classes in the Definition 2.3.1. In more detail Rabiner [24], Knab [14] and Durbin *et al.* [6] have introduced reestimation formulas for the standard *HMM*s and Knab [14] has determined reestimation formulas for *HMM*s with transition classes.
We have introduced *HMM*s with continuous emissions and coupled transition classes in the Definition 3.2.1 and therefore we have only modified the function $\mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda)$ (2.7). That means

we can use the reestimation formulas for the start probabilities and the emission functions which have been determined by Knab [14]. In the Section 3.3 we have shown that we always obtain a stochastic matrix $\hat{A}^*_{\vec{S}}$ which maximises $\mathbb{Q}_A(A^*_{\vec{S}}|\lambda)$ (3.8) if at least one $c_t = 1$ in the transition class sequence $\mathcal{C}_\mathcal{O}$ (2.4). So we have everything to prove the convergence of the Baum-Welch algorithm.

Structure of the convergence proof

1. We assume that λ is our current *HMM* for the training data and that λ^* is the newly estimated *HMM* which improves the likelihood of the training data in comparison with λ.
2. In Proposition 3.4.1 we use the results of the independent maximisation of $\mathbb{Q}_S(\vec{\pi}|\lambda)$ (2.6), $\mathbb{Q}_A(\mathcal{A}_{\vec{S}}|\lambda)$ (2.7) and $\mathbb{Q}_E(\vec{E}|\lambda)$ (2.8) to prove that the estimated model parameters for the *HMM* λ^* increase $\mathbb{Q}(\lambda^*|\lambda)$ (2.5) in comparison with $\mathbb{Q}(\lambda|\lambda)$ (2.5) of the given *HMM* λ.
3. In Proposition 3.4.2 we start with the results of the Proposition 3.4.1 and show that the likelihood for an emission sequence $\mathcal{O}$ (2.2) is increased under the newly estimated *HMM* λ^*.

Proposition 3.4.1. *If* $\mathbb{Q}_S(\vec{\pi}^*|\lambda) \geq \mathbb{Q}_S(\vec{\pi}|\lambda)$, $\mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda) \geq \mathbb{Q}_A(\mathcal{A}_{\vec{S}}|\lambda)$ *and* $\mathbb{Q}_E(\vec{E}^*|\lambda) \geq \mathbb{Q}_E(\vec{E}|\lambda)$ *for the two HMMs* $\lambda^* = (\vec{\pi}^*, \mathcal{N}, A^*_1, \vec{S}, \vec{E}^*)$ *and* $\lambda = (\vec{\pi}, \mathcal{N}, A_1, \vec{S}, \vec{E})$*, then* $\mathbb{Q}(\lambda^*|\lambda) \geq \mathbb{Q}(\lambda|\lambda)$.

Proof: Let us consider the Q-functions (2.5) for the two HMMs λ *and* λ^* *and so we have*

$$\mathbb{Q}(\lambda^*|\lambda) = \mathbb{Q}_S(\vec{\pi}^*|\lambda) + \mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda) + \mathbb{Q}_E(\vec{E}^*|\lambda) \text{ and}$$

$$\mathbb{Q}(\lambda|\lambda) = \mathbb{Q}_S(\vec{\pi}|\lambda) + \mathbb{Q}_A(\mathcal{A}_{\vec{S}}|\lambda) + \mathbb{Q}_E(\vec{E}|\lambda).$$

Now we can compute the difference between $\mathcal{Q}(\lambda^*|\lambda)$ *and* $\mathcal{Q}(\lambda|\lambda)$ *and obtain*

$$\mathbb{Q}(\lambda^*|\lambda) - \mathbb{Q}(\lambda|\lambda) = \mathbb{Q}_S(\vec{\pi}^*|\lambda) - \mathbb{Q}_S(\vec{\pi}|\lambda) + \mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda) - \mathbb{Q}_A(\mathcal{A}_{\vec{S}}|\lambda) + \mathbb{Q}_E(\vec{E}^*|\lambda) - \mathbb{Q}_E(\vec{E}|\lambda).$$

We know that $\mathbb{Q}_S(\vec{\pi}^*|\lambda) \geq \mathbb{Q}_S(\vec{\pi}|\lambda)$, $\mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda) \geq \mathbb{Q}_A(\mathcal{A}_{\vec{S}}|\lambda)$ *and* $\mathbb{Q}_E(\vec{E}^*|\lambda) \geq \mathbb{Q}_E(\vec{E}|\lambda)$ *and so the following conditions*

$$\mathbb{Q}_S(\vec{\pi}^*|\lambda) - \mathbb{Q}_S(\vec{\pi}|\lambda) \geq 0, \quad \mathbb{Q}_A(\mathcal{A}^*_{\vec{S}}|\lambda) - \mathbb{Q}_A(\mathcal{A}_{\vec{S}}|\lambda) \geq 0 \quad \text{and} \quad \mathbb{Q}_E(\vec{E}^*|\lambda) - \mathbb{Q}_E(\vec{E}|\lambda) \quad \geq 0$$

are fulfilled. The result is that $\mathcal{Q}(\lambda^*|\lambda) - \mathcal{Q}(\lambda|\lambda)$ *can be transformed to* $\mathbb{Q}(\lambda^*|\lambda) - \mathbb{Q}(\lambda|\lambda) \geq 0$ *and therefore we obtain* $\mathbb{Q}(\lambda^*|\lambda) \geq \mathbb{Q}(\lambda|\lambda)$.

□

Proposition 3.4.2. *If* $\mathbb{Q}(\lambda^*|\lambda) \geq \mathbb{Q}(\lambda|\lambda)$*, then* $\mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O}, \lambda^*] \geq \mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O}, \lambda]$.

Proof: We know that $\mathbb{Q}(\lambda^*|\lambda) \geq \mathbb{Q}(\lambda|\lambda)$*. Let us look at the definition of the Q-function (Definition 2.4.3):*

$$\mathbb{Q}(\lambda^*|\lambda) = \sum_{Q \in \mathcal{Q}} \mathbb{D}[\mathcal{O}, Q|\mathcal{C}_\mathcal{O}, \lambda] \log(\mathbb{D}[\mathcal{O}, Q|\mathcal{C}_\mathcal{O}, \lambda^*]).$$

To prove the proposition we use the logsum-inequality which is

$$\forall a_l, b_l \in \mathbb{R}^+ : \sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \sum_{i=1}^{n} a_i \log \frac{\sum_{j=1}^{n} a_j}{\sum_{j=1}^{n} b_j}$$

and therewith we give the following proof.

$$\mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O},\lambda]\log\left(\frac{\mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O},\lambda]}{\mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O},\lambda^*]}\right) = \sum_{Q\in\mathcal{Q}}\mathbb{D}[\mathcal{O},Q|\mathcal{C}_\mathcal{O},\lambda]\log\left(\frac{\sum_{Q\in\mathcal{Q}}\mathbb{D}[\mathcal{O},Q|\mathcal{C}_\mathcal{O},\lambda]}{\sum_{Q\in\mathcal{Q}}\mathbb{D}[\mathcal{O},Q|\mathcal{C}_\mathcal{O},\lambda^*]}\right)$$

$$\mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O},\lambda]\log\left(\frac{\mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O},\lambda]}{\mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O},\lambda^*]}\right) \leq \sum_{Q\in\mathcal{Q}}\mathbb{D}[\mathcal{O},Q|\mathcal{C}_\mathcal{O},\lambda]\log\left(\frac{\mathbb{D}[\mathcal{O},Q|\mathcal{C}_\mathcal{O},\lambda]}{\mathbb{D}[\mathcal{O},Q|\mathcal{C}_\mathcal{O},\lambda^*]}\right)$$

$$-\mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O},\lambda]\log\left(\frac{\mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O},\lambda^*]}{\mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O},\lambda]}\right) \leq -\sum_{Q\in\mathcal{Q}}\mathbb{D}[\mathcal{O},Q|\mathcal{C}_\mathcal{O},\lambda]\log\left(\frac{\mathbb{D}[\mathcal{O},Q|\mathcal{C}_\mathcal{O},\lambda^*]}{\mathbb{D}[\mathcal{O},Q|\mathcal{C}_\mathcal{O},\lambda]}\right)$$

$$\mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O},\lambda]\log\left(\frac{\mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O},\lambda^*]}{\mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O},\lambda]}\right) \geq \sum_{Q\in\mathcal{Q}}\mathbb{D}[\mathcal{O},Q|\mathcal{C}_\mathcal{O},\lambda]\log\left(\frac{\mathbb{D}[\mathcal{O},Q|\mathcal{C}_\mathcal{O},\lambda^*]}{\mathbb{D}[\mathcal{O},Q|\mathcal{C}_\mathcal{O},\lambda]}\right)$$

$$\begin{aligned}\mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O},\lambda]\log\left(\frac{\mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O},\lambda^*]}{\mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O},\lambda]}\right) &\geq \sum_{Q\in\mathcal{Q}}\mathbb{D}[\mathcal{O},Q|\mathcal{C}_\mathcal{O},\lambda]\log(\mathbb{D}[\mathcal{O},Q|\mathcal{C}_\mathcal{O},\lambda^*]) \\ &\quad -\sum_{Q\in\mathcal{Q}}\mathbb{D}[\mathcal{O},Q|\mathcal{C}_\mathcal{O},\lambda]\log(\mathbb{D}[\mathcal{O},Q|\mathcal{C}_\mathcal{O},\lambda])\end{aligned}$$

$$\log\left(\frac{\mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O},\lambda^*]}{\mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O},\lambda]}\right) \geq \frac{\mathbb{Q}(\lambda^*|\lambda)-\mathbb{Q}(\lambda|\lambda)}{D[\mathcal{O}|\mathcal{C}_\mathcal{O},\lambda]}$$

We know that $\mathbb{Q}(\lambda^*|\lambda)-\mathbb{Q}(\lambda|\lambda)\geq 0$ *and therefore the following equations are fulfilled.*

$$\begin{aligned}\log\left(\frac{\mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O},\lambda^*]}{\mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O},\lambda]}\right) &\geq 0 \\ \log(\mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O},\lambda^*]) &\geq \log(\mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O},\lambda]) \\ \mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O},\lambda^*] &\geq \mathbb{D}[\mathcal{O}|\mathcal{C}_\mathcal{O},\lambda]\end{aligned}$$

□

With the result of the Proposition 3.4.2 we have given a general proof of the convergence of the *Baum-Welch algorithm*. Whenever we modify model components of *HMM*s we can go back to this proof if the preconditions can be fulfilled.

3.5 Analytical Solution For One Transition Class

Let us consider an *HMM* with one coupled transition class. For such an *HMM* each c_t in the transition class sequence $\mathcal{C}_\mathcal{O}$ is one. In this case the reestimation formulas for the diagonal elements $a^*_{ii}(1)$ and the non-diagonal elements $a^*_{ij}(1)$ have to be equal to the reestimation formulas of an *HMM* without transition classes. Let us first prove that the reestimation formulas for the diagonal elements are identical. Therefore we have to solve the Equation (3.16). Recall, the scaling factor S_1 is one and so we obtain

$$\sum_{t=1}^{T-1}\gamma_t(i) = \sum_{t=1}^{T-1}\frac{1}{a^*_{ii}(1)}\varepsilon_t(i,i) \text{ and this can be transformed to}$$

$$a^*_{ii}(1) = \frac{\sum_{t=1}^{T-1}\varepsilon_t(i,i)}{\sum_{t=1}^{T-1}\gamma_t(i)}. \tag{3.20}$$

This reestimation formula of the diagonal element $a^*_{ii}(1)$ is equal to the reestimation formula of the diagonal element $a^*_{ii}(1)$ in an *HMM* without transition classes. We have already mentioned that the reestimation formulas for standard *HMM*s are discussed by Rabiner [24], Knab [14] and Durbin *et al.* [6].

Now we have the possibility to determine the reestimation formula of the non-diagonal element $a_{ij}(1)^*$ by substituting $a^*_{ii}(1)$ in Equation (3.13) by the expression of $a_{ij}(1)^*$ in the Equation (3.20).

$$a^*_{ij}(1) = \frac{\sum_{t=1}^{T-1} \gamma_t(i) \sum_{t=1}^{T-1} \varepsilon_t(i,j) - \sum_{t=1}^{T-1} \varepsilon_t(i,i) \sum_{t=1}^{T-1} \varepsilon_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i) \sum_{t=1}^{T-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \varepsilon_t(i,j)}$$

$$= \frac{\sum_{t=1}^{T-1} \varepsilon_t(i,j) \left(\sum_{t=1}^{T-1} \gamma_t(i) - \sum_{t=1}^{T} \varepsilon_t(i,i) \right)}{\sum_{t=1}^{T-1} \gamma_t(i) \sum_{t=1}^{T-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \varepsilon_t(i,j)} = \frac{\sum_{t=1}^{T-1} \varepsilon_t(i,j) \sum_{t=1}^{T-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \varepsilon_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i) \sum_{t=1}^{T-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \varepsilon_t(i,j)}$$

$$a^*_{ij}(1) = \frac{\sum_{t=1}^{T-1} \varepsilon_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

The reestimation formula of the non-diagonal element $a^*_{ij}(1)$ is equal to the reestimation formula of the non-diagonal element $a^*_{ij}(1)$ in an *HMM* without transition classes.

3.6 Analytical Solution For Two Transition Classes

We have to solve the Equation (3.16) to get the analytical solutions of the reestimation formulas for an *HMM* with two coupled transition classes. Recall, the scaling factor S_1 is one. First we make some basic definitions

$$\mathcal{K}^i_1 := \sum_{t=1}^{T-1} \gamma_t(i), \qquad \mathcal{K}^i_2 := \sum_{\substack{t=1 \\ c_t=1}}^{T-1} \varepsilon_t(i,i) \quad \text{and} \quad \mathcal{K}^i_3 := \sum_{\substack{t=1 \\ c_t=2}}^{T-1} \varepsilon_t(i,i).$$

The lower and the upper bounds for these constants are

$$0 < \mathcal{K}^i_1 < T-1, \qquad 0 < \mathcal{K}^i_2 < T-1 \quad \text{and} \quad 0 < \mathcal{K}^i_3 < T-1,$$

because each $\varepsilon_t(i,j)$ for $i,j \in \mathcal{N}$ is in the interval $(0,1)$.
Now we start to determine the analytical solution for the Equation (3.16).

$$
\begin{aligned}
\sum_{t=1}^{T-1} \gamma_t(i) &= \sum_{t=1}^{T-1} \frac{S_{c_t}}{a_{ii}^*(1) - 1 + S_{c_t}} \varepsilon_t(i,i) \\
\sum_{t=1}^{T-1} \gamma_t(i) &= \sum_{\substack{t=1 \\ c_t=1}}^{T-1} \frac{1}{a_{ii}^*(1)} \varepsilon_t(i,i) + \sum_{\substack{t=1 \\ c_t=2}}^{T-1} \frac{S_2}{a_{ii}^*(1) - 1 + S_2} \varepsilon_t(i,i) \\
\sum_{t=1}^{T-1} \gamma_t(i) &= \frac{1}{a_{ii}^*(1)} \sum_{\substack{t=1 \\ c_t=1}}^{T-1} \varepsilon_t(i,i) + \frac{S_2}{a_{ii}^*(1) - 1 + S_2} \sum_{\substack{t=1 \\ c_t=2}}^{T-1} \varepsilon_t(i,i) \\
\mathcal{K}_1^i &= \frac{1}{a_{ii}^*(1)} \mathcal{K}_2^i + \frac{S_2}{a_{ii}^*(1) - 1 + S_2} \mathcal{K}_3^i \\
0 &= (a_{ii}^*(1))^2 + \underbrace{\frac{(S_2 - 1)\mathcal{K}_1^i - \mathcal{K}_2^i - S_2 \mathcal{K}_3^i}{\mathcal{K}_1^i}}_{p} a_{ii}^*(1) + \underbrace{\frac{(1 - S_2)\mathcal{K}_2^i}{\mathcal{K}_1^i}}_{q}
\end{aligned}
$$

The result is a quadratic equation which has the two solutions

$$
a_{ii}^*(1)_1 = -\frac{p}{2} + \sqrt{\frac{p^2}{4} - q} \quad \text{and} \quad a_{ii}^*(1)_2 = -\frac{p}{2} - \sqrt{\frac{p^2}{4} - q}.
$$

The term q is always less than zero because $S_2 > 1$, $\mathcal{K}_2^i > 0$ and $\mathcal{K}_1^i > 0$. The following inequality is fulfilled for each $q < 0$

$$
\frac{p}{2} < \sqrt{\frac{p^2}{4} - q}.
$$

So we can exclude the second solution $a_{ii}^*(1)_2$ because $a_{ii}^*(1)_2$ is always less than zero. Another observation on the basis of this inequality is that the first solution $a_{ii}^*(1)_1$ is always greater than zero.

Now we show that $a_{ii}^*(1)_1$ is less than one. Let us assume that $\mathcal{K}_2^i = T - 1$ then $\mathcal{K}_2^i$ is equal to $\mathcal{K}_1^i$, $\mathcal{K}_3^i$ is equal to zero, p is equal to $S_2 - 2$ and q reaches its minimal value $1 - S_2$. The result is that $a_{ii}^*(1)_1$ is equal to one and this is the maximal value which $a_{ii}^*(1)_1$ can reach if we would allow the upper bound of $\mathcal{K}_2^i$.

The non-diagonal elements $a_{ij}^*(1)_1$ are determined by the Equation (3.13) and the diagonal element $a_{ii}^*(2)_1$ and the non-diagonal elements $a_{ij}^*(2)_1$ are calculated by the Equations (3.4) and (3.5).

Chapter 4

Emission Densities For Hidden Markov Models

The definitions of *HMM*s with continuous emissions in the Definitions 2.2.1, 2.3.1 and 3.2.1 request that an emission density is assigned to every state of the *HMM*. We have decided to model the microarray data with a mixture of normal distributions. The mixture is estimated by the well known expectation maximisation algorithm (*EM algorithm*). This is followed by the assignment of each mixture component to a state of the *HMM*. Now we describe our proceeding in more detail, and for the *EM algorithm* we follow the tutorial of Bilmes [1].

The following topics are contained in this chapter:

1. A general introduction of the *EM algorithm* and the application of this algorithm for the estimation of a mixture model of univariate normal distributions are presented in the Section 4.1.

2. An agglomerative clustering algorithm to assign mixture components to the states of an *HMM* is developed in the Section 4.2.

3. Two examples for the agglomerative clustering on microarray data are shown in the Section 4.3.

4.1 EM Algorithm

The *EM algorithm* is a widely spread method for parameter estimation in incomplete data problems. The parameter estimation is computed iteratively and the monotonic convergence of this process is guaranteed. A well known application of the *EM algorithm* is the estimation of mixture models. An incomplete data set is obtained if some components of the data can and others cannot be measured. In the case of a mixture model the output signals of the system are known, but the internal mixture components which have created these output signals are unknown.

Let us consider an observation sequence $x = (x_1, \ldots, x_N)$ which is generated by some distribution and modelled as an instance of the random vector $\mathbb{X} = (\mathbb{X}_1, \ldots, \mathbb{X}_N)$. We assume that a complete data set $\mathbb{Z} = (\mathbb{X}, \mathbb{Y})$ exists and we also suppose the following joint density function

$$\mathbb{D}[\mathbb{Z}|\Theta] = \mathbb{D}[\mathbb{X}, \mathbb{Y}|\Theta] = \mathbb{D}[\mathbb{Y}|\mathbb{X}, \Theta]\mathbb{D}[\mathbb{X}|\Theta].$$

This joint density comes from the marginal density function $\mathbb{D}[\mathbb{X}|\Theta]$ and the assumptions of hidden variables or relationships between the missing and the observed values. We can define the likelihood function $\mathcal{L}[\Theta|\mathbb{X},\mathbb{Y}]$ with this density function and therefore we assume

$$\mathcal{L}[\Theta|\mathbb{X},\mathbb{Y}] = \mathbb{D}[\mathbb{X},\mathbb{Y}|\Theta].$$

The likelihood is also a random variable since the missing information is unknown and random. That is, we can think of $\mathcal{L}[\Theta|\mathbb{X},\mathbb{Y}]$ as a function where $\mathbb{X}$ and Θ are constant and $\mathbb{Y}$ is a random variable. In the first step of the *EM algorithm*, the *E-step*, the expectation value of the log-likelihood $\log(\mathcal{L}[\Theta|\mathbb{X},\mathbb{Y}])$ is computed with respect to the unknown data $\mathbb{Y}$ given the observed data $\mathbb{X} = x$ and the current parameter estimates Θ_t. That is, we define

$$\begin{aligned}\mathcal{L}[\Theta|\mathbb{X},\Theta_t] &= \mathbb{E}[\log(\mathcal{L}[\Theta|\mathbb{X},\mathbb{Y}])|\mathbb{X},\Theta_t] \\ &= \int_y \log(\mathbb{D}[\mathbb{X}=x,\mathbb{Y}=y|\Theta])\mathbb{D}[\mathbb{Y}=y|\mathbb{X}=x,\Theta_t].\end{aligned}$$

The current parameter estimates Θ_t are used to evaluate $\mathcal{L}[\Theta|\mathbb{X},\Theta_t]$ and Θ contains the new parameters that we optimise to increase the expectation value. The expression $\mathbb{D}[\mathbb{Y}|\mathbb{X},\Theta_t]$ is the marginal distribution of the missing data $\mathbb{Y}$.

In the second step of the *EM algorithm*, the *M-step*, the computed expectation value $\mathcal{L}[\Theta|\mathbb{X},\Theta_t]$ of the *E-step* is maximised. That is, we find

$$\Theta_{t+1} = \underset{\Theta}{\operatorname{argmax}}\ \mathcal{L}[\Theta|\mathbb{X},\Theta_t].$$

The *M-step* and the *E-step* are repeated until a local maximum of the likelihood function is reached, but this is not shown here.

In the following subsection we use the *EM algorithm* to find the maximum likelihood parameters of a mixture model which consists of univariate normal distributions.

4.1.1 Mixture Of Univariate Normal Distributions

The parameter estimation of a mixture model is one of the most widely used applications of the *EM algorithm*. We assume the following probabilistic model

$$\mathbb{D}[X = x|\Theta] = \sum_{i=1}^{M} \alpha_i \mathcal{N}(x|\mu_i,\sigma_i)$$

with $\Theta = ((\alpha_1,\mu_1,\sigma_1),\ldots,(\alpha_M,\mu_M,\sigma_M))$. The parameter $\alpha_i > 0$ is the weight, the parameter μ_i the mean and the parameter σ_i is the standard deviation of the i-th mixture component. The sum over all mixture component weights is one. In other words, we assume to have a model with M mixture components which are mixed together with respect to their mixture weight.

We model the observed data $x = (x_1,\ldots,x_N)$ as a random vector $\mathbb{X} = (\mathbb{X}_1,\ldots,\mathbb{X}_N)$ which consists of independent, identically distributed random variables $\mathbb{X}_i \sim X$. The likelihood of $\mathbb{X}$ under the assumed model is given by

$$\mathcal{L}[\Theta|\mathbb{X}=x] = \prod_{i=1}^{N} \mathbb{D}[\mathbb{X}_i = x_i|\Theta] = \prod_{i=1}^{N} \mathbb{D}[X = x_i|\Theta] = \prod_{i=1}^{N}\sum_{j=1}^{M} \alpha_j \mathcal{N}(x_i|\mu_j,\sigma_j)$$

and therefore the log-likelihood is

$$\log(\mathcal{L}[\Theta|\mathbb{X}=x]) = \sum_{i=1}^{N} \log\left(\sum_{j=1}^{M} \alpha_j \mathcal{N}(x_i|\mu_j,\sigma_j)\right).$$

The log-likelihood is difficult to optimise because of the logarithm of a sum. Let us consider $\mathbb{X}$ as incomplete and posit the existence of unobserved data $y = (y_1, \dots, y_N)$ which is modelled as a random vector $\mathbb{Y} = (\mathbb{Y}_1, \dots, \mathbb{Y}_N)$ consisting of independent, identically distributed random variables $\mathbb{Y}_i \sim Y$. The component $y_i \in \{1, \dots, M\}$ informs us which mixture component has generated the observation x_i. If we know the values of $\mathbb{Y}$, the likelihood becomes

$$\begin{aligned}\mathcal{L}[\Theta|\mathbb{X} = x, \mathbb{Y} = y] &= \prod_{i=1}^{N} \mathbb{D}[\mathbb{X}_i = x_i, \mathbb{Y}_i = y_i|\Theta] = \prod_{i=1}^{N} \mathbb{D}[X = x_i, Y = y_i|\Theta] \\ &= \prod_{i=1}^{N} \mathbb{D}[Y = y_i|\Theta]\mathbb{D}[X = x_i|Y = y_i, \Theta] \\ &= \prod_{i=1}^{N} \alpha_{y_i}\mathcal{N}(x_i|\mu_{y_i}, \sigma_{y_i})\end{aligned}$$

and therefore the log-likelihood is

$$\log(\mathcal{L}[\Theta|\mathbb{X} = x, \mathbb{Y} = y]) = \sum_{i=1}^{N} \log(\alpha_{y_i}\mathcal{N}(x_i|\mu_{y_i}, \sigma_{y_i}))$$

which is easily to optimise. Of course we do not know the values of $\mathbb{Y}$, but we assume that $\mathbb{Y}$ is a random vector and therefore we can use the *EM algorithm*.

First we must derive an expression for the distribution of the unobserved data $\mathbb{Y}$. We guess the initial parameters $\Theta_t = ((\alpha_1^t, \mu_1^t, \sigma_1^t), \dots, (\alpha_M^t, \mu_M^t, \sigma_M^t))$ of the likelihood $\mathcal{L}[\Theta_t|\mathbb{X} = x, \mathbb{Y} = y]$. So we can easily compute $\mathcal{N}(x_i|\mu_{y_j}, \sigma_{y_j})$ for each i and j. The weight α_i can be interpreted as prior probability of a mixture component. Therefore we use *Bayes's rule* to compute

$$\begin{aligned}\mathbb{D}[Y = y_i|X = x_i, \Theta_t] &= \frac{\mathbb{D}[Y = y_i|\Theta_t]\mathbb{D}[X = x_i|Y = y_i, \Theta_t]}{\mathbb{D}[X = x_i|\Theta_t]} \\ &= \frac{\alpha_{y_i}\mathcal{N}(x_i|\mu_{y_i}, \sigma_{y_i})}{\sum_{j=1}^{M} \alpha_{y_j}\mathcal{N}(x_i|\mu_{y_j}, \sigma_{y_j})}\end{aligned}$$

and

$$\begin{aligned}\mathbb{D}[\mathbb{Y} = y|\mathbb{X} = x, \Theta_t] &= \prod_{i=1}^{N} \mathbb{D}[\mathbb{Y}_i = y_i|\mathbb{X}_i = x_i, \Theta_t] \\ &= \prod_{i=1}^{N} \mathbb{D}[Y = y_i|X = x_i, \Theta_t].\end{aligned}$$

We have obtained the desired marginal density by assuming the existence of hidden variables $\mathbb{Y}_i$ and guessing the initial parameters Θ_t of their distribution. The expectation value

$\mathcal{L}[\Theta|\mathbb{X}=x,\Theta_t]$ takes the form

$$\begin{aligned}
\mathcal{L}[\Theta|\mathbb{X}=x,\Theta_t] &= \sum_{y} \log(\mathcal{L}[\Theta|\mathbb{X}=x,\mathbb{Y}=y])\mathbb{D}[\mathbb{Y}=y|\mathbb{X}=x,\Theta_t] \\
&= \sum_{y}\sum_{i=1}^{N} \log(\alpha_{y_i}\mathcal{N}(x_i|\mu_{y_i},\sigma_{y_i})) \prod_{j=1}^{N} \mathbb{D}[Y=y_j|X=x_j,\Theta_t] \\
&= \sum_{y_1=1}^{M} \dots \sum_{y_N=1}^{M}\sum_{i=1}^{N} \log(\alpha_{y_i}\mathcal{N}(x_i|\mu_{y_i},\sigma_{y_i})) \prod_{j=1}^{N} \mathbb{D}[Y=y_j|X=x_j,\Theta_t] \\
&= \sum_{y_1=1}^{M} \dots \sum_{y_N=1}^{M}\sum_{i=1}^{N}\sum_{l=1}^{M} \delta(l,y_i)\log(\alpha_{l}\mathcal{N}(x_i|\mu_{l},\sigma_{l})) \prod_{j=1}^{N} \mathbb{D}[Y=y_j|X=x_j,\Theta_t] \\
&= \sum_{l=1}^{M}\sum_{i=1}^{N} \log(\alpha_{l}\mathcal{N}(x_i|\mu_{l},\sigma_{l})) \sum_{y_1=1}^{M} \dots \sum_{y_N=1}^{M} \delta(l,y_i) \prod_{j=1}^{N} \mathbb{D}[Y=y_j|X=x_j,\Theta_t].
\end{aligned}$$

This can be simplified if we consider the last sums for a fixed i and l

$$\begin{aligned}
&\sum_{y_1=1}^{M} \dots \sum_{y_N=1}^{M} \delta(l,y_i) \prod_{j=1}^{N} \mathbb{D}[Y=y_j|X=x_j,\Theta_t] \\
&= \left(\sum_{y_1=1}^{M} \dots \sum_{y_{i-1}=1}^{M}\sum_{y_{i+1}=1}^{M} \dots \sum_{y_N=1}^{M} \prod_{\substack{j=1\\ j\neq i}}^{N} \mathbb{D}[Y=y_j|X=x_j,\Theta_t]\right) \mathbb{D}[Y=l|X=x_i,\Theta_t] \\
&= \prod_{\substack{j=1\\ j\neq i}}^{N} \left(\sum_{y_j=1}^{M} \mathbb{D}[Y=y_j|X=x_j,\Theta_t]\right) \mathbb{D}[Y=l|X=x_i,\Theta_t] \\
&= \mathbb{D}[Y=l|X=x_i,\Theta_t].
\end{aligned}$$

The tricks behind this simplification are to exclude cleverly and to use

$$\sum_{k=1}^{M} \mathbb{D}[Y=k|X=x_j,\Theta_t] = 1.$$

Now it is possible to rewrite $\mathcal{L}[\Theta|\mathbb{X}=x,\Theta_t]$ and so we obtain

$$\begin{aligned}
\mathcal{L}[\Theta|\mathbb{X}=x,\Theta_t] &= \sum_{l=1}^{M}\sum_{i=1}^{N} \log(\alpha_{l}\mathcal{N}(x_i|\mu_{l},\sigma_{l}))\mathbb{D}[Y=l|X=x_i,\Theta_t] \\
&= \sum_{l=1}^{M}\sum_{i=1}^{N} \log(\alpha_{l})\mathbb{D}[Y=l|X=x_i,\Theta_t] \\
&\quad + \sum_{l=1}^{M}\sum_{i=1}^{N} \log(\mathcal{N}(x_i|\mu_{l},\sigma_{l}))\mathbb{D}[Y=l|X=x_i,\Theta_t].
\end{aligned}$$

This function can be maximised to get the model parameters Θ^{t+1}.

$$\forall l \in \mathcal{M}: \quad \alpha_l^{t+1} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{D}[Y = l | X = x_i, \Theta_t] \qquad \mu_l^{t+1} = \frac{\sum_{i=1}^{N} x_i \mathbb{D}[Y = l | X = x_i, \Theta_t]}{\sum_{i=1}^{N} \mathbb{D}[Y = l | X = x_i, \Theta_t]}$$

$$\sigma_l^{t+1} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu_l^{t+1})^2 \mathbb{D}[Y = l | X = x_i, \Theta_t]}{\sum_{i=1}^{N} \mathbb{D}[Y = l | X = x_i, \Theta_t]}}.$$

These new model parameters are taken for the next iteration of the *EM algorithm*. The likelihood increases in each iteration and so always a local maximum is reached asymptotically for $t \to \infty$, but this is not shown here.
To estimate a mixture model for microarray data as it is shown in the Figures 4.1 and 4.2 we generate mixture models for a different number of mixture components M. To compare these models we use the *Akaike* (*AIC*) and the *Bayesian information criterion* (*BIC*) and choose some candidate models.

4.2 Clustering Mixture Components

To motivate this section we consider the upper left images of the Figures 4.1 and 4.2. Here we see two histograms which are nearly symmetric around zero. That is the normal situation when we consider the log ratios of microarray data. The upper left image of the Figure 4.1 shows the log ratios of gene expression data and we are interested in under-, identically and over-expressed genes. Suppose that we have a mixture model for the data in the Figure 4.1 with $M \geq 3$ mixture components. Now our aim is to divide the mixture components into an under-, an identically and an over-expressed cluster. So each cluster should represent a mixture model and should describe the properties of the dedicated gene class. If we suppose to have a histogram of *ArrayCGH* log ratios we are interested in DNA regions which have a decreased, an unchanged or an increased copy number. These DNA regions can also be modelled by three clusters as we have motivated for the gene expression data.
Let $\Theta = ((\alpha_1, \mu_1, \sigma_1), \ldots, (\alpha_M, \mu_M, \sigma_M))$ be the vector of model parameters for an univariate mixture model of normal distributions (Section 4.1.1). The mean μ_i of the i-th normal distribution represents the centre of this normal distribution and therefore μ_i is an ideal attribute to cluster the mixture components.

Agglomerative Clustering

- *Initialisation*
  Each mixture component is modelled as a cluster.

$$\forall i \in \mathcal{M}: \quad \mathbb{C}_i := \{(\alpha_i, \mu_i, \sigma_i)\}$$
$$\mathbb{C} := \{\mathbb{C}_1, \ldots, \mathbb{C}_M\}$$

- *Iteration*
  The two most similar clusters form a new cluster.

```
while $|\mathbb{C}| > 3$
    //determine the two most similar clusters
    $(\mathbb{C}_i^*, \mathbb{C}_j^*) := \operatorname*{argmin}_{\substack{1 \le i,j \le |\mathbb{C}| \\ i \neq j}} \mathcal{D}(\mathbb{C}_i, \mathbb{C}_j)$

    //remove the two most similar clusters and add their union
    $\mathbb{C} = (\mathbb{C} \setminus \{\mathbb{C}_i^*, \mathbb{C}_j^*\}) \cup \{\mathbb{C}_i^* \cup \mathbb{C}_j^*\}$
```

  Here we use single linkage to determine the distance $\mathcal{D}(\mathbb{C}_i, \mathbb{C}_j)$ between the two clusters $\mathbb{C}_i$ and $\mathbb{C}_j$ and define

$$\mathcal{D}(\mathbb{C}_i, \mathbb{C}_j) = \min_{\substack{(\alpha_1,\mu_1,\sigma_1) \in \mathbb{C}_i \\ (\alpha_2,\mu_2,\sigma_2) \in \mathbb{C}_j}} |\mu_1 - \mu_2|.$$

- *Result*
  We obtain the three clusters

$$\mathbb{C} := \{\mathbb{C}_1, \mathbb{C}_2, \mathbb{C}_3\}.$$

  The weight $\mathbb{W}_i$ of a cluster $\mathbb{C}_i$ is calculated by summing up all weights α_j in this cluster and so we get

$$\mathbb{W}_i = \sum_{(\alpha_j,\mu_j,\sigma_j) \in \mathbb{C}_i} \alpha_j.$$

  The weight of a cluster i is used to normalise each weight α_j in the cluster $\mathbb{C}_i$ by dividing α_j by $\mathbb{W}_i$. Therefore the sum of all weights in a cluster is one and therewith the cluster is a mixture model. Afterwards we determine the mean value $\mathbb{M}_i$ of each cluster $\mathbb{C}_i$ with respect to the mean μ_j and obtain

$$\mathbb{M}_i = \frac{1}{|\mathbb{C}_i|} \sum_{(\alpha_j,\mu_j,\sigma_j) \in \mathbb{C}_i} \mu_j.$$

  Now we have an attribute for each cluster that gives us the possibility to determine the class of a cluster. We sort the three cluster means in ascending order and obtain the following permutation $\mathbb{M}_{\pi_1} \le \mathbb{M}_{\pi_2} \le \mathbb{M}_{\pi_3}$. This is the basis to assign the following cluster classes

  1. $\mathbb{C}_{\pi_1}$ is the under-expressed,
  2. $\mathbb{C}_{\pi_2}$ is the identically expressed and
  3. $\mathbb{C}_{\pi_3}$ is the over-expressed cluster.

4.3 Example Analysis

Here we want to show how the methods of the last two sections work in practice. Let us consider the histogram of preprocessed gene expression data in the upper left image of the Figure 4.1. It shows a nearly symmetric distribution of log ratios around zero for test versus reference. The log ratios which are significantly less than zero can give hints for under-expressed and log ratios which are significantly greater than zero can give hints for over-expressed genes in the test set. Now we want to estimate a mixture model for this histogram using the *EM algorithm*. One of the best models is shown in the upper right image of the Figure 4.1. It consists of three mixture components and therefore the agglomerative clustering algorithm has only to determine the classes of the three components. So we obtain the red coloured under-expressed, the grey coloured identically expressed and the green coloured over-expressed cluster. In the lower image of the Figure 4.1 each probability density function of a cluster is shown. Here we see the unweighted clusters of the upper right image of the Figure 4.1. When we use a three-state *HMM* as in the Figure 5.1 to model microarray data, then each probability density function is assigned to a state of the *HMM* as a state emission function.

In the Figure 4.2 the same graphics are shown for the preprocessed *ArrayCGH* data for chromosome 1 of the breast cancer data set form Pollack *et al.* [23]. The nearly symmetric histogram of the log ratios is shown in the upper left image of the Figure 4.2. The upper right image of the Figure 4.2 contains the associated mixture model with five mixture components. The agglomerative clustering algorithm has assigned three mixture components to the grey cluster which represents the unchanged DNA copy number status. The green cluster which represents the increased DNA copy number status consists of one mixture component and the red cluster which represents the decreased DNA copy number status consists also of one mixture component. In summary, the mixture model was chosen to demonstrate the good performance of the agglomerative clustering algorithm. The densities around zero are slightly too high, but other regions look better. As above the probability density functions in the lower subfigure of Figure 4.2 can be assigned to a state of the three-state *HMM* as shown in the Figure 5.1.

In general we search for a mixture model which has a small or nearly no overlap between the red and the green cluster and this mixture model has also to approximate the histogram of the given microarray data very well. Such models have shown the best performance with a three-state *HMM* as in the Figures 5.1 or 5.2.

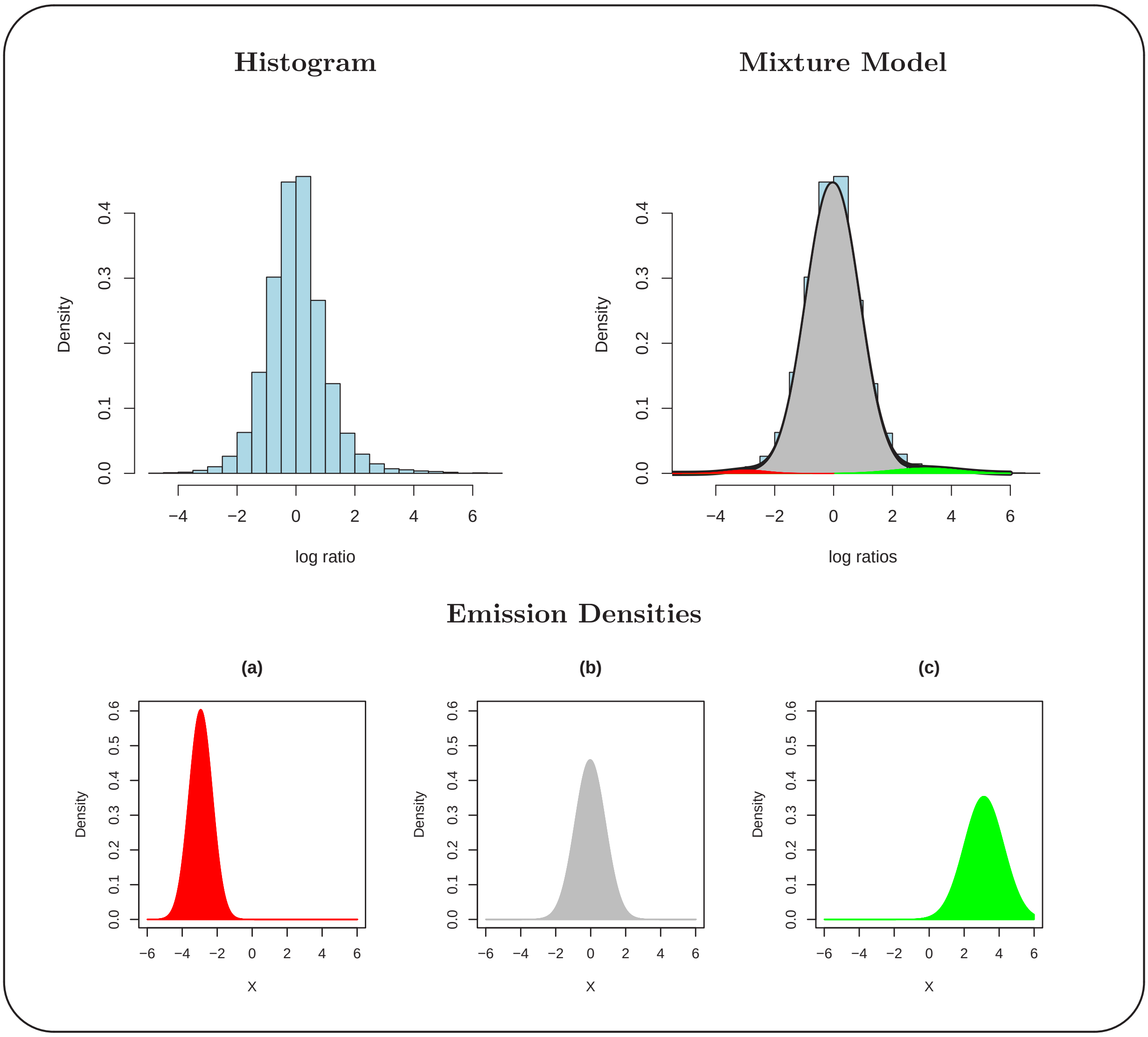


Figure 4.1: Analysis of the gene expression values for chromosome 17 of the breast cancer data set from Pollack *et al.* [23]. **Histogram**: Histogram of the log ratios for test versus normal. **Mixture Model**: The associated mixture model for the histogram. The black curve is the probability density function of the mixture model. The red area shows the proportion of the under-expressed, grey area shows the proportion of the identically expressed and the green area shows the proportion of the over-expressed cluster. **Emission Densities**: (a) shows the probability density function of the cluster $\mathbb{C}_{\pi_1}$ (under-expressed). (b) shows the probability density function of the cluster $\mathbb{C}_{\pi_2}$ (identically expressed). (c) shows the probability density function of the cluster $\mathbb{C}_{\pi_3}$ (over-expressed).

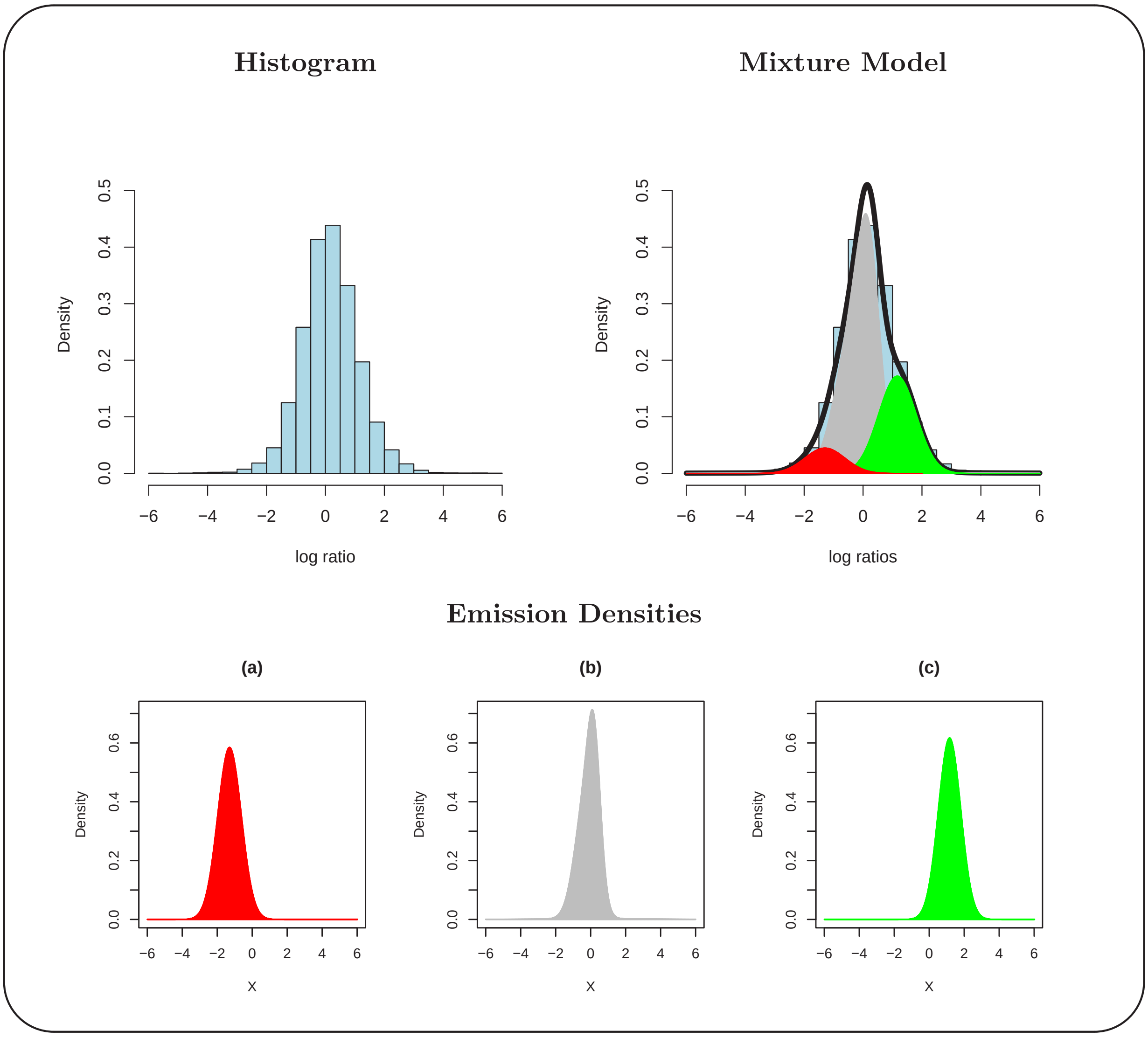


Figure 4.2: Analysis of the *ArrayCGH* values for chromosome 1 of the breast cancer data set from Pollack *et al.* [23]. **Histogram**: Histogram of the log ratios for test versus normal. **Mixture Model**: The associated mixture model for the histogram. The black curve is the probability density function of the mixture model. The red area shows the proportion of the decreased, grey area shows the proportion of the unchanged and the green area shows the proportion of the increased copy number cluster. **Emission Densities**: (a) shows the probability density function of the cluster $\mathbb{C}_{\pi_1}$ (decreased copy number). (b) shows the probability density function of the cluster $\mathbb{C}_{\pi_2}$ (unchanged copy number). (c) shows the probability density function of the cluster $\mathbb{C}_{\pi_3}$ (increased copy number).

Chapter 5

Modelling Microarray Data

The huge progress in life sciences is reflected in the amounts of data which are created by microarray experiments. These amounts of data cannot be analysed without the help of efficient bioinformatics strategies. Microarray data can be seen as a sequence of measurements. *HMM*s are an ideal framework to analyse sequences of measurements. This chapter has two main parts. The first part shows how to create *HMM*s to annotate profiles of microarray data and the second part introduces the role of chromosome 17 in breast cancer.

The following topics are contained in this chapter:

1. How to create *HMM*s to analyse microarray data is described in the Section 5.1.
2. How microarray profiles are annotated and some quality criterions for the annotation results are explained in the Section 5.2.
3. Selected breast cancer publications are introduced in the Section 5.3.
4. A general overview of our breast cancer data set is given in the Section 5.4.

5.1 Creating *HMM*s For Microarray Data

The basic components of homogeneous *HMM*s in the Definition (2.2.1) and the inhomogeneous *HMM*s with coupled transition matrices in the Definition (3.2.1) are a set of states, a start distribution, emission density functions and a transition matrix. The individual selection of these components give us the possibility to create *HMM*s to analyse microarray data.
Let us recall the agglomerative clustering in the Section 4.2 which we use to divide the microarray data into three clusters. So it is obvious that we propose an *HMM* with three states and one transition matrix as our basic model. This basic model is shown in the Figure 5.1. The colour of an emission density function represents the assigned cluster and therewith we refer to the results of the agglomerative clustering.
An extension of the basic model is shown in the Figure 5.2. Here we use two transition matrices to model microarray data and therefore it is necessary to have functions which can map additional information as the distance between neighbour genes into a transition class sequence $\mathcal{C}_{\mathcal{O}}$ (2.4).
The following general overview describes how an *HMM* for the analysis of microarray data is created.

Creating an HMM to analyse microarray data

1. Create a mixture model using mixture estimation (Section 4.1.1).

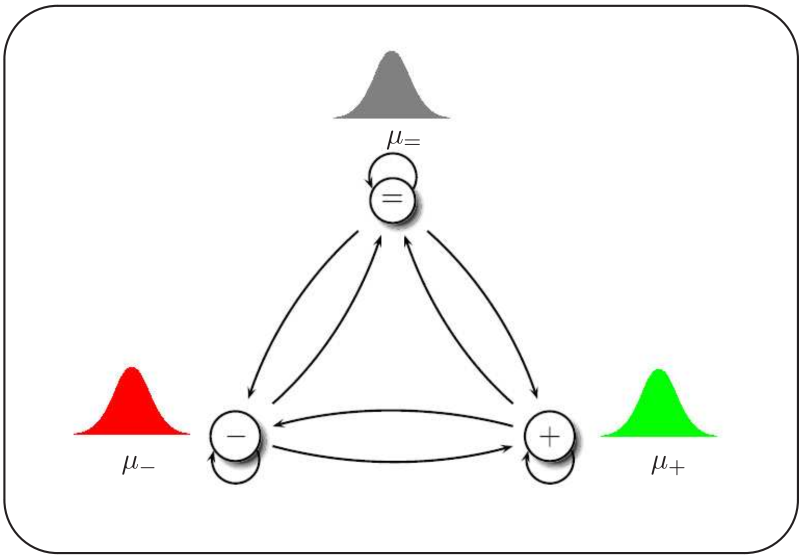


Figure 5.1: Three State Model: *HMM* with continuous emissions and one transition class. If gene expression profiles are modelled, then the state − represents under-expressed, the state = represents identically expressed and the state + represents over-expressed regions in these profiles. If *ArrayCGH* profiles are modelled, then the state − represents decreased, the state = represents unchanged and the state + represents increased copy number regions in these profiles.

2. Create a clustered mixture model using agglomerative clustering for the mixture model above (Section 4.2).
3. Choose a three state *HMM* either with one or more than one transition matrix (Figures 5.1 and 5.2).
4. Assign the emission density funcitions from the second step to the *HMM*. Choose the initial start distribution and the initial transition matrix. If we select an *HMM* with coupled transition matrices, then the scaling parameters and a transition class switching function are required.
5. Train the *HMM* with the microarray data using the *Baum-Welch algorithm*.

In the next section we introduce how to choose the start distribution and how to create transition matrices.

5.1.1 Modelling Start Distributions And Transition Matrices

It is important for the training of the *HMM* to choose good initial model parameters because otherwise the risk is higher that the *Baum-Welch algorithm* stops in a bad local maxima. Therefore the initial parameters have to include information about the modelled microarray profiles. Let us consider the results of the mixture estimation on the basis of microarray data after we have used the agglomerative clustering. We obtain the cluster weights $\mathbb{W}_-$, $\mathbb{W}_=$ and $\mathbb{W}_+$ which contain information how great the probability for a microarray measurement is to be created by one of these clusters. It follows that the expected number of microarray measurements which are created by the cluster i is $n\mathbb{W}_i$ for n microarray measurements. These weights can be used to set the *Markov chain* which underlies the *HMM* to these expectation values. We do this by modelling the equilibrium distribution of the *Markov chain* and therefore we choose the start distribution $\vec{\pi} = (\mathbb{W}_=, \mathbb{W}_+, \mathbb{W}_-)$ and model a transition matrix $\mathcal{A}(\vec{\pi}, \mathbb{S}, n)$ with the equilibrium distribution equal to $\vec{\pi}$. The equilibrium distribution does not completely determine the

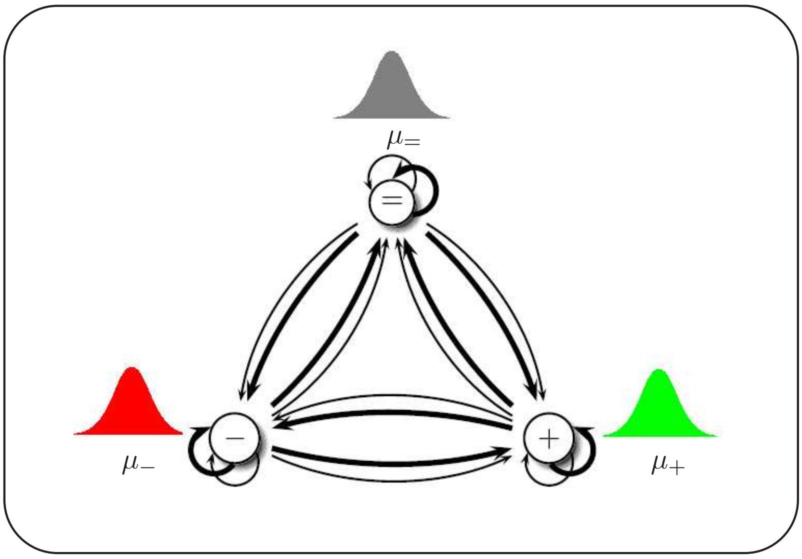


Figure 5.2: Three State Model: *HMM* with continuous emissions and two transition classes. The first class is represented by normal arrows and the second by thicker arrows. If gene expression profiles are modelled, then the state − represents under-expressed, the state = represents identically expressed and the state + represents over-expressed regions in these profiles. If *ArrayCGH* profiles are modelled, then the state − represents decreased, the state = represents unchanged and the state + represents increased copy number regions in these profiles.

transition matrix and so we can model different transition matrices of the following type, for $\mathbb{S} \in (0, n \cdot \min\{\mathbb{W}_-, \mathbb{W}_=, \mathbb{W}_+\})$,

$$\mathcal{A}(\vec{\pi}, \mathbb{S}, n) = \begin{pmatrix} a_{==} & a_{=+} & a_{=-} \\ a_{+=} & a_{++} & a_{+-} \\ a_{-=} & a_{-+} & a_{--} \end{pmatrix} = \begin{pmatrix} 1 - \frac{\mathbb{S}}{n\mathbb{W}_=} & \frac{\mathbb{S}}{2n\mathbb{W}_=} & \frac{\mathbb{S}}{2n\mathbb{W}_=} \\ \frac{\mathbb{S}}{2n\mathbb{W}_+} & 1 - \frac{\mathbb{S}}{n\mathbb{W}_+} & \frac{\mathbb{S}}{2n\mathbb{W}_+} \\ \frac{\mathbb{S}}{2n\mathbb{W}_-} & \frac{\mathbb{S}}{2n\mathbb{W}_-} & 1 - \frac{\mathbb{S}}{n\mathbb{W}_-} \end{pmatrix}$$

which all have the predefined equilibrium distribution $\vec{\pi}$. This can be easily proven by multiplying $\mathcal{A}(\vec{\pi}, \mathbb{S}, n)$ with $\vec{\pi}$. The probability to change the current state $i \in \{-, =, +\}$ is $1 - a_{ii}$. The transition probability to go from a state i to another state j is modelled by a uniform distribution. The self-transition probabilities can be scaled by the parameter $\mathbb{S}$ and therefore we can influence the state durations. The parameter n is set to the mean length of all microarray profiles.

These transition matrices are the basis of all our *HMM*s. Let us assume we have assigned the emission density functions of the agglomerative clustering and an initial transition matrix $\mathcal{A}(\vec{\pi}, \mathbb{S}, n)$ to an *HMM* with continuous emissions as in the Definition 2.2.1. Now we are able to approximate the distribution of the underlying microarray data if we use the cluster weights as initial start distribution.

5.1.2 Modelling Transition Class Switching Functions

The extensions of the homogeneous *HMM*s are inhomogeneous *HMM*s which can model relations between microarray measurements. We have extended the standard *HMM*s to be able to consider such relations.

When we use the standard *HMM*s it is not possible to model effects which can be caused by the distance between adjacent genes. In this case we have to work with the simplification that all adjacent genes have the same distance. The distance between adjacent genes can be modelled with the extended *HMM*s in the Definitions 2.3.1 and 3.2.1 and therefore we can include the possibility of mutations.

For instance, it could make sense to assume that a mutation between two adjacent genes occurs more often the greater the distance between these two genes is. The result of such a mutation could be decoupled expression levels of these two genes in affected cells instead of coupled expression levels of these genes in normal cells. Such simple models can be easily adapted to other problems as copy number changes or the correlation of gene expression between adjacent genes. Later we motivate the usage of transition classes with the help of special mutations which frequently occur in breast cancer.

The effects of additional information are modelled by an *HMM* with L transition matrices. To work with such an *HMM* we need a transition class switching function to generate a transition class sequence $\mathcal{C}_{\mathcal{O}}$ (2.4). We assume that we have a sequence of additional information

$$\mathcal{I}_{\mathcal{O}} = \mathcal{I}_1^{\mathcal{O}}, \ldots, \mathcal{I}_T^{\mathcal{O}}$$

for each emission sequence $\mathcal{O}$ (2.2). Now we can define the transition class switching function

$$\mathbb{SWITCH} : \mathcal{I}_{\mathcal{O}} \mapsto \mathcal{C}_{\mathcal{O}}.$$

For example let us model the distance between adjacent genes with the help of two transition matrices. The following transition class switching function maps all adjacent gene distances $\mathcal{I}_t^{\mathcal{O}} \in \mathbb{R}^+$ to one of the two transition classes using a threshold value $\mathcal{T} \in \mathbb{R}^+$ and so we define

$$\mathbb{SWITCH}(\mathcal{I}_t^{\mathcal{O}}) = \begin{cases} 1, & \mathcal{I}_t^{\mathcal{O}} > \mathcal{T} \\ 2, & \mathcal{I}_t^{\mathcal{O}} \leq \mathcal{T}. \end{cases} \tag{5.1}$$

The *HMM* has to use the transition matrix $A_{c_t} = \mathbb{SWITCH}(\mathcal{I}_t^{\mathcal{O}})$ at time step t and therewith the complete transition class sequence $\mathcal{C}_{\mathcal{O}}$ can be calculated by the same proceeding. We can use this defined transition class switching function for the *HMM* with two transition matrices which is shown in the Figure 5.2.

This short example should illustrate how additional information is used to generate a transition class sequence. To proceed in that way is flexible and allows us to model different relations between data.

5.2 Annotating Microarray Profiles

A microarray profile is annotated by an *HMM* using the *Viterbi algorithm* which is introduced in the Section 2.5. So we obtain for each gene expression profile $\mathcal{O}$ (2.2) the *Viterbi Path* $\mathcal{Q}_{\mathcal{O}}^*$ (2.3) which is the most probable state path through the *HMM*. To get a better impression of the performance of the *HMM* we use the state posterior as a quality criterion for the *Viterbi Path* $\mathcal{Q}_{\mathcal{O}}^*$. The state posterior $\gamma_t(i)$ is already defined in the Definition 2.4.7. Recall that it represents the probability for being in state i at time step t given the emission sequence $\mathcal{O}$. We will use the following quality criteria:

1. The state posterior $\gamma_t(\mathcal{Q}_t^*)$ of each state in the *Viterbi Path* $\mathcal{Q}_{\mathcal{O}}^* = \mathcal{Q}_1^*, \ldots, \mathcal{Q}_T^*$.
2. The state posterior $\gamma_t(=)$ for being in the state $=$ for each time step $t \in \{1, \ldots, T\}$.
3. The state posterior $\gamma_t(+)$ for being in the state $+$ for each time step $t \in \{1, \ldots, T\}$.
4. The state posterior $\gamma_t(-)$ for being in the state $-$ for each time step $t \in \{1, \ldots, T\}$.

Each of these criteria is used to display a state posterior profile for all $t \in \{1, \dots, T\}$ when we consider the annotation results in more detail. The profiles of $\gamma_t(=)$, $\gamma_t(+)$ and $\gamma_t(-)$ should give us the possibility to recognise segments in a microarray profile where the measurements show the same behaviour. That is, we can find segments in these state posterior profiles where the state posterior is significantly greater than zero. We can also obtain hints for segments in a microarray profile which have been annotated in another way by the *HMM*. Cases for this behaviour could be that the *HMM* is not able to model a good segment structure or that measurements are marginal cases in the annotation process of the *HMM*. The profile of $\gamma_t(\mathcal{Q}_t^*)$ should mostly consist of state posteriors which are only slightly less than one, but it is also possible that this profile contains regions where the state posteriors are significantly less than one. The *HMM* is relatively unsure how to annotate such regions in these cases. Smaller values of $\gamma_t(\mathcal{Q}_t^*)$ can occur for microarray data values which are marginal cases in the annotation process.

The state posteriors $\gamma_t(=)$, $\gamma_t(+)$ and $\gamma_t(-)$ allow us also to determine the absolute proportion $\mathbb{SSP}(.)$ of microarray measurements which are seen as identically expressed, over-expressed or under-expressed in a microarray profile $\mathcal{O}$ independent of the most probable annotation which is given by the *Viterbi Path* $\mathcal{Q}_{\mathcal{O}}^*$. Therefore we calculate the sums over the state posteriors

$$\mathbb{SSP}(=) = \sum_{t=1}^{T} \gamma_t(=), \quad \mathbb{SSP}(+) = \sum_{t=1}^{T} \gamma_t(+) \quad \text{and} \quad \mathbb{SSP}(-) = \sum_{t=1}^{T} \gamma_t(-).$$

We can also determine how many of each of the absolute proportions $\mathbb{SSP}(=)$, $\mathbb{SSP}(+)$ and $\mathbb{SSP}(-)$ is included in the *Viterbi Path* $\mathcal{Q}_{\mathcal{O}}^*$. This is done by summing over the suitable annotation values and so we obtain

$$\mathbb{SV}(=) = \sum_{t=1}^{T} \gamma_t(=)\delta(\mathcal{Q}_t^*, =), \quad \mathbb{SV}(+) = \sum_{t=1}^{T} \gamma_t(+)\delta(\mathcal{Q}_t^*, +) \quad \text{and} \quad \mathbb{SV}(-) = \sum_{t=1}^{T} \gamma_t(-)\delta(\mathcal{Q}_t^*, -)$$

with the help of the function $\delta(i, j)$ which is one if i is equal to j or zero in the other case. The structure of the *Viterbi Path* always fulfils $\mathbb{SV}(i) \leq \mathbb{SSP}(i)$ and this is the basis to compute the relative proportion of $\mathbb{SSP}(i)$ in the *Viterbi Path*. That is, we determine

$$\mathbb{S}(=) = \frac{\mathbb{SV}(=)}{\mathbb{SSP}(=)}, \quad \mathbb{S}(+) = \frac{\mathbb{SV}(+)}{\mathbb{SSP}(+)} \quad \text{and} \quad \mathbb{S}(-) = \frac{\mathbb{SV}(-)}{\mathbb{SSP}(-)} \tag{5.2}$$

and therewith we get an impression how many of the basic state posterior profiles $\gamma_t(=)$, $\gamma_t(+)$ and $\gamma_t(-)$ for all $t \in \{1, \dots, T\}$ is contained in the *Viterbi Path*. If we think of a state posterior profile where for each time step t exactly a $j \in \{=, +, -\}$ exists which fulfils $\gamma_t(j) = 1$, then all relative proportions $\mathbb{S}(=)$, $\mathbb{S}(+)$ and $\mathbb{S}(-)$ are one (if each state is at least used one time). This means that the *HMM* is always sure how to annotate the microarray profile $\mathcal{O}$. Such a state posterior profile is unrealistic and not possible for the *HMM*s that we use and therefore $\mathbb{S}(=)$, $\mathbb{S}(+)$ and $\mathbb{S}(-)$ are less than one.

We finally consider the quality of the *Viterbi Path* by summing up the state posteriors of this path and dividing by the length T of the microarray data $\mathcal{O}$ and therewith we obtain

$$\mathbb{S} = \frac{1}{T} \sum_{t=1}^{T} \gamma_t(\mathcal{Q}_t^*). \tag{5.3}$$

If we recall an unrealistic state posterior profile, then $\mathbb{S}$ is equal to one because no other annotation of the microarray data exists.

In summary, all the introduced quality criteria above can help us to describe the quality of a microarray profile. Nevertheless, it is necessary to choose good initial model parameters to generate reliable annotations.

5.3 The Role Of Chromosome 17 In Breast Cancer

The chromosome 17 is one of the smallest human chromosomes with the highest gene density and it is known that this chromosome is often rearranged in breast cancer and therefore losses and gains of DNA regions are intensively studied [21]. A lot of publications as for example Orsetti *et al.* 1999 [20], Kauraniemi *et al.* 2001 [13], Monni *et al.* 2001 [17], Hyman *et al.* 2002 [9], Clark *et al.* 2002 [3], Pollack *et al.* 2002 [23], Willis *et al.* 2003 [27], Nugoli *et al.* 2003 [19] and Orsetti *et al.* 2004 [21] have analysed chromosome 17 in breast cancer cells with different techniques. We summarise the main results which will help us to analyse the results of our own analysis of the microarray data for chromosome 17 from Pollack *et al.* [23].

Selected Breast Cancer Publications

- Orsetti *et al.* 1999 [20] did allelotyping studies and molecular cytogenetics and found at least four regions of allelic imbalances for chromosome 17 in breast cancer. That is the chromosome 17 can get lost, the regions 17q and 17q22-q24 can contain gains of DNA segments and the regions 17q11-q21 or 17q25-qter or both together can be affected by losses of DNA segments. They found that gains of DNA segments are most commonly observed in the region 17q23-q24.

- Kauraniemi *et al.* 2001 [13] found that the amplification of DNA segments in the region 17q12 leads to simultaneous elevation of expression levels of several genes. They identified for example the genes *CDC6*, *GRB7*, *MLN51*, *MLN64*, *ZNF144* and *ERBB2* to be always over-expressed when their DNA is amplified. The gene *ERBB2* is a well-known oncogene which contributes to a poor clinical outcome when it is amplified. They also found the genes *GRB7*, *MLN64*, *TRAF4* and *PPARBP* to be coamplified with *ERBB2* and that this observation might also have some clinical impact on breast cancer.

- Monni *et al.* 2001 [17] used a combination of molecular, genomic and microarray technologies to analyse the region 17q23 in six breast cancer cell lines. They were able to define two common regions of amplification within the region 17q23. Their cDNA microarray studies have shown that several over-expressed genes exist and that the genes *RPS6KB1*, *MUL*, *APPBP2* and *TRAP240* are located in the two defined regions above. They also identified that the gene expression patterns varied from a cell line to another. The structure of the 17q23 amplicon in the *MCF7* breast cancer cell line was described to consist of two separate highly amplified regions which are flanked by a region of low-level amplification. The 17q23 amplicon of the breast cancer cell line *BT474* was described to consist of a single large segment which is amplified and overlaps both of the regions that were identified in the *MCF7* cell line.

- Hyman *et al.* 2002 [9] made a high-resolution comparative genomic hybridisation analysis on cDNA microarrays for breast cancer to compare DNA copy number and gene expression. They found that high- and low-level copy number changes have substantial impacts on gene expression and a novel amplicon at 17q21.3 which leads to over-expression of the genes *HOXB7* and *HOXB2* is described.

- Clark *et al.* 2002 [3] used comparative genomic hybridisation and cDNA microarrays to identify candidate oncogenes in breast cancer cell lines. They found the amplicons 20q13, 17q11-21 and 17q22-23 for the cell line *BT474*. The amplification of the region 17q22-23 was seen with the over-expression of *HBOA* and *TRAP100* was observed as over-expressed in the amplified the region 17q11-21.

- Willis *et al.* 2003 [27] performed a high-resolution expression array analysis and a comparative genomic hybridisation analysis of genes mapped to the entire region 17q12-23 to identify novel candidate oncogenes. They identified significantly over-expressed genes when gains of DNA segments in the region 17q12-23 had occurred. Several of these genes were previously identified oncogenes. They concluded that chromosome 17 contains at least two frequent amplifications in the regions 17q12-21 and 17q23 which contain the candidate oncogenes *RPS6KB1*, *APPB2*, *MUL*, *TRAP240* and *TBX2*.

- Nugoli *et al.* 2003 [19] made comparative genomic hybridisation and gene expression profiles of *MCF7* breast cancer sublines. They found that the sublines contain important differences in DNA copy number alterations. Gains of 17q22-24 and losses of 17p11-13 were observed.

- Orsetti *et al.* 2004 [21] performed *ArrayCGH* and cDNA analyses. They were able to subdivide the chromosome 17 into thirteen consensus segments, the four regions 17p, 17q11.2, 17q21 and 17q24 which mainly contain DNA losses, the six regions one at 17q12 and five at 17q22-25 which mainly contain DNA gains and the three regions 17q21.3, 17q22 and 17q25 which either show gains or losses.

- Pollack *et al.* 2002 [23] performed genome wide *ArrayCGH* and cDNA microarray analyses to determine the influences of variation in gene copy number on the gene expression in breast cancer cells. They found that highly amplified genes can show moderately or highly elevated expression. So they were able to conclude that the gene expression is influenced across a wide range of DNA copy number alterations and that widespread DNA copy number alterations can lead directly to global deregulation of gene expression and therefore induce the development or the progression of breast cancer.

5.4 Breast Cancer Data Set

We use the *ArrayCGH* and the gene expression data from Pollack *et al.* [23] to analyse chromosomal imbalances on chromosome 17 in breast cancer cells.

Normalisation and data set creation

First we have inspected the data of Pollack *et al.* [23] and thereby we have found that the gene expression data was not normalised as described by Pollack *et al.* [23]. Therefore we have normalised each experiment in the gene expression data set to mean zero and standard deviation one. The *ArrayCGH* data of Pollack *et al.* [23] looks better than the gene expression data and therefore we use these data without an additional normalisation. After that we have used the *ArrayCGH* and the gene expression data to create our own data set for chromosome 17. In addition we have added the newest available chromosomal location of each gene using the *Entrez Gene* at the *NCBI*. Our data set contains forty-one experiments with two hundred sixty-five genes for both data classes. A measurement for a gene g in our data set represents

$$\log_2 \frac{\text{Intensity of } g \text{ in the test sample}}{\text{Intensity of } g \text{ in the reference sample}}.$$

Correlation between gene expression data and correlation between *ArrayCGH* data

The Figure 5.3 shows the correlations between the experiments in our data set. The correlation matrix for the gene expression experiments contains only some weak positive correlations except the cell lines. We see another behaviour for the correlations between *ArrayCGH* experiments where most of the correlations are positive. In general we can expect that the *ArrayCGH* profiles look more similar than the gene expression profiles. A profile of microarray data for an experiment contains all measurements sorted by the chromosomal locations of the measured genes.

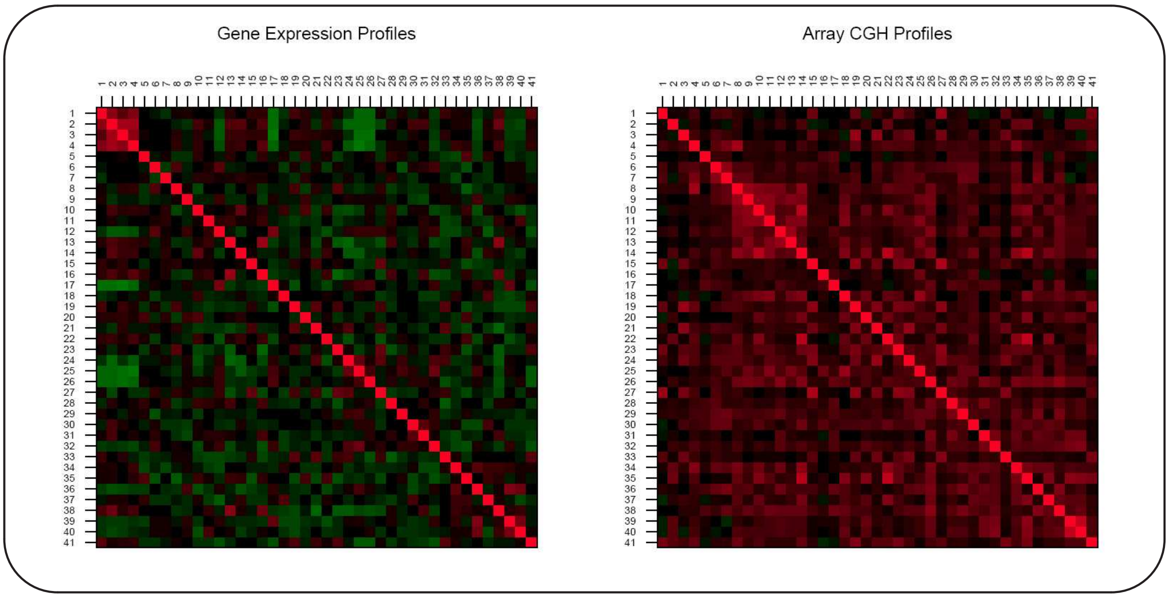


Figure 5.3: *Left*: The correlation matrix of all gene expression profiles of chromosome 17 in our breast cancer data set. *Right*: The correlation matrix of all *ArrayCGH* profiles of chromosome 17 in our breast cancer data set. In general the green fields show negative correlation between profiles, black fields show no correlation and red fields show positive correlation. The first four profiles are the cell lines.

Overview of gene expression data and *ArrayCGH* data

Let us consider the overviews in the Figures 5.4 and 5.5 to get a more detailed impression of chromosome 17 in our breast cancer data set. The Figure 5.4 represents an overview of the gene expression profiles in our data set. On the basis of the upper graphic we can see in nearly every band of the chromosome 17 that log ratios can be found which differ from other log ratios at the same chromosomal location. Such extreme log ratios are candidates for over- or under-expression. The middle graphic shows the mean and the median log ratios of all genes over all experiments mapped to the chromosomal location. The values of the mean log ratio per gene are close to zero and the median values contain information where most of the log ratios for a gene over all experiments lie. The lower graphic represents the variance of the log ratios per gene over all experiments. Regions with high variance refer to chromosomal locations which have been observed with different expression levels over all experiments. In summary, we should expect that most of the gene expression profiles are individual and this fortifies the results which are shown in the correlation matrix of the gene expression profiles in the Figure 5.3.

The same graphics are shown in the Figure 5.5 for all *ArrayCGH* profiles in our data set. Extreme log ratios are mostly found in the regions 17p13.1, 17q11.2 and 17q12-q25. These regions represent candidates which are affected by losses or gains of DNA segments. The mean values and the median values lie close to each other and therefore the profiles should be more homogeneous as the gene expression profiles. The variance values are mostly close to zero and support the statement of homogeneous *ArrayCGH* profiles.

The Table 5.1 shows the number of measured data values per band for an experiment in our breast cancer data set of chromosome 17. The bands 17q21.1, 17q22, 17q23.1, 17q24.1, 17q24.3

and 17q25.2 have at most six measured data values and therefore general statements about the gene expression or DNA alterations in such a band are not possible because a lot of genes which are not considered are located their. The region 17q11.1 contains no data value.

Candidate genes for over-expression

The Table 5.2 contains candidate genes for over-expression which are mentioned by Pollack *et al.* [23]. The genes from the Table 5.2 are located in the characteristic regions which we have described above.

Sensitivity of the ***ArrayCGH*** data

Pollack *et al.* [23] have tried to test the sensitivity of their *ArrayCGH* approach by analysing the cell lines 45-X0, 46-XX, 47-XXX, 48-XXXX and 49-XXXXX. Each of these cell lines contains a different number of X chromosomes. They reported that the mean fluorescence ratio of such a cell line to the normal cell line 46-XX is linearly proportional to the slightly underestimated copy number ratios. We have analysed this observation for a better understanding of the background. We found that the mean fluorescence ratios for all cell lines contain information about the number of X chromosomes in these cell lines, but the histograms of the fluorescence ratios of the cell lines overlap. This is the result of the high variances of the fluorescence ratios of a cell line and therefore it should not be possible to determine the degree of a loss or a gain of a segment without previous knowledge.

We have determined the ranges of the *ArrayCGH* and the gene expression log ratios in our data set. The range for the *ArrayCGH* data is from -2.56 to 3.73 and the range for the gene expression data is from -4.63 to 6.88. The hybridisation signals for *ArrayCGH* data are smaller than for gene expression data. If we assume that one copy of a gene is lost in a breast cancer cell, then we will observe a log ratio of $\log_2 \frac{1}{2} = -1$. Or let us assume that two copies of a gene are gained in a breast cancer cell so we would expect a log ratio of $\log_2 \frac{4}{2} = 1$. Both cases are assumed ideal cases and we should not be able to determine the quantity of copy number changes for genes because of the high variances in the control experiments with different numbers of X chromosomes.

In summary, the analysis of *ArrayCGH* profiles should be done carefully when we test the performance of our *HMM*s on these profiles.

Band	p13.3	p13.2	p13.1	p12	p11.2
$\sum$	9	17	21	8	17

Band	q11.1	q11.2	q12	q21.1	q21.2
$\sum$	0	24	29	1	20

Band	q21.31	q21.32	q21.33	q22	q23.1
$\sum$	24	15	12	4	1

Band	q23.2	q23.3	q24.1	q24.2	q24.3
$\sum$	12	9	2	6	2

Band	q25.1	q25.2	q25.3		
$\sum$	22	1	9		

Table 5.1: Number of measured data values per band for an experiment in our breast cancer data set for chromosome 17.

Gene	**Band**	**Gene**	**Band**	**Gene**	**Band**	**Gene**	**Band**
KIAA0524	q11.2	CACNB1	q12	JUP	q21.2	TRAP240	q23.2
UNC119	q11.2	RPL19	q12	HOXB5	q21.32	ICAM2	q23.3
SDF2	q11.2	MLN64	q12	NDP52	q21.32	PECAM1	q23.3
TRAF4	q11.2	ERBB2	q12	NGFR	q21.33	ABCA5	q24.2
FLJ10700	q11.2	GRB7	q12	HBOA	q21.33	SLC9A3R1	q25.1
TIAF1	q11.2	NR1D1	q21.1	DLX4	q21.33	AD023	q25.1
KIAA1321	q11.2	CDC6	q21.2	ABCC3	q21.33	GRB2	q25.1
SCYA3	q12	TOP2A	q21.2	RAD51C	q23.2	ITGB4	q25.1
SCYA4	q12	SMARCE1	q21.2	RPS6KB1	q23.2	HCNGP	q25.1
MLLT6	q12	KRT20	q21.2	APPBP2	q23.2	BIRC5	q25.3
ZNF144	q12	KRTHA4	q21.2	PPM1D	q23.2	LGALS3BP	q25.3
PIP5K2B	q12	KRT19	q21.2	TBX2	q23.2		

Table 5.2: Table of candidate genes for over-expression on chromosome 17. The genes are taken from the results of Pollack *et al.* [23].

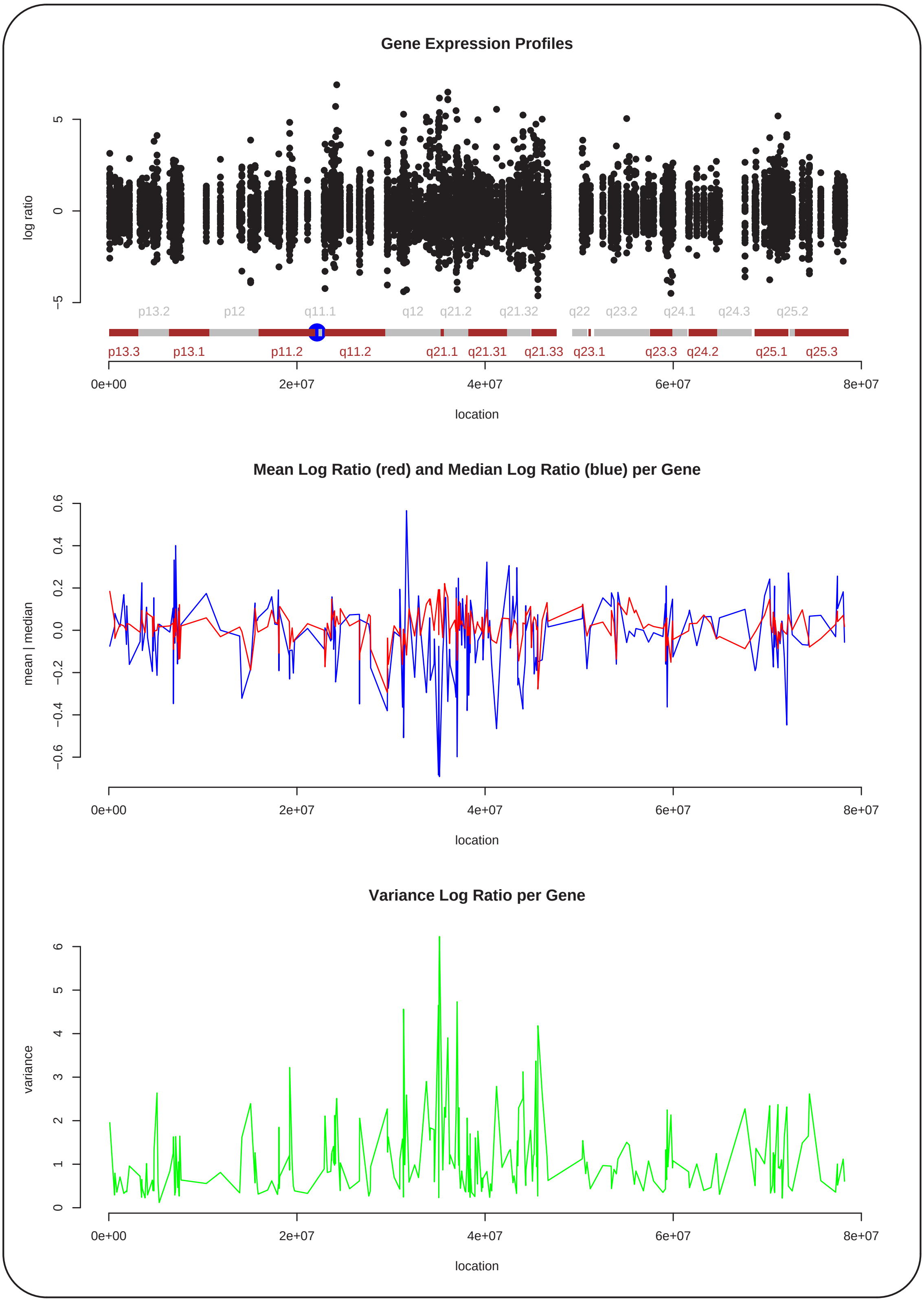


Figure 5.4: Overview of our gene expression profiles for chromosome 17. The upper graphic shows the log ratios of all experiments mapped to their chromosomal location. The middle graphic represents the mean and the median log ratio per gene over all experiments mapped to the chromosomal locations of the genes. The lower graphic shows the variance of a log ratio per gene over all experiments mapped to the chromosomal locations of the genes.

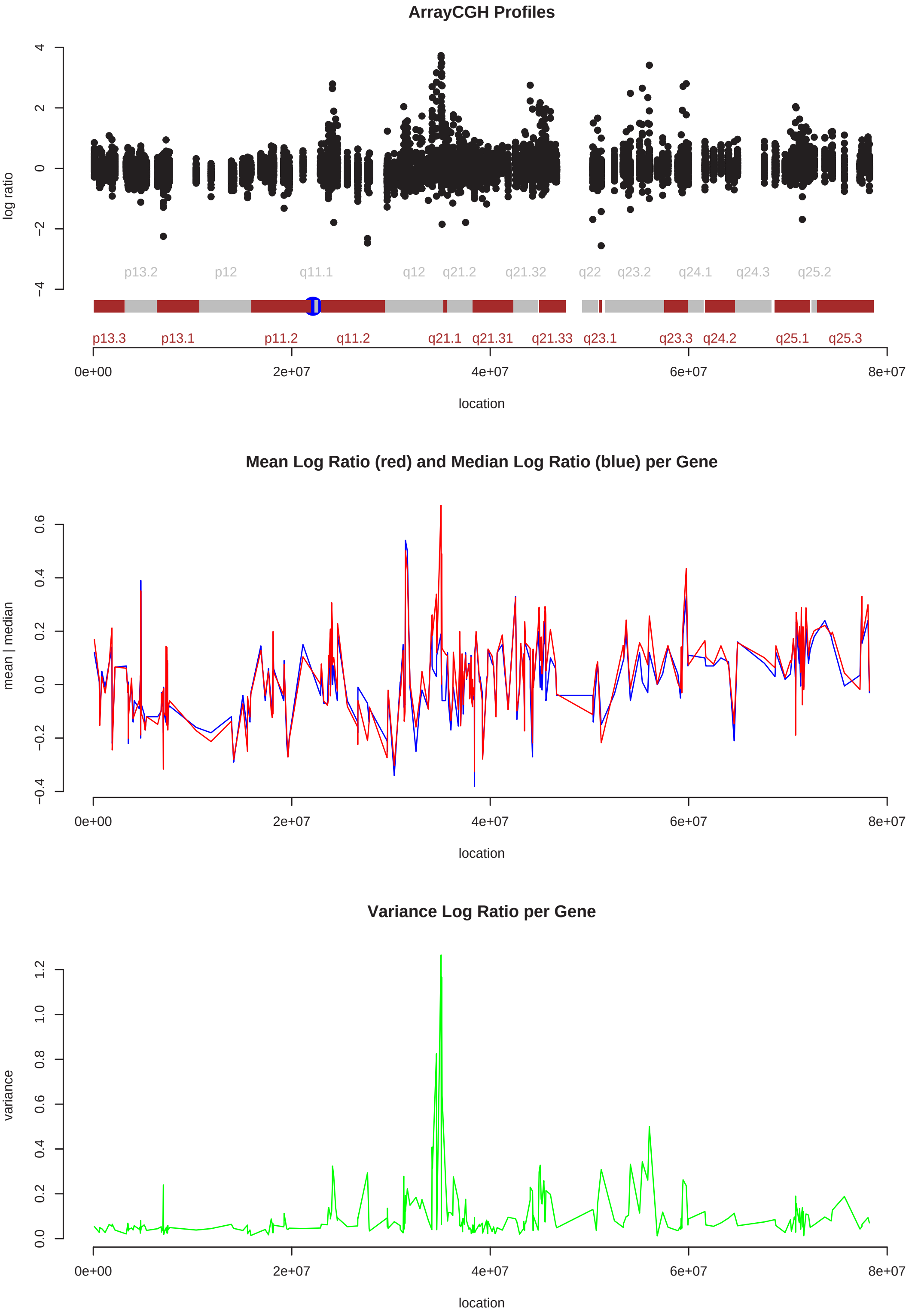


Figure 5.5: Overview of our *ArrayCGH* profiles for chromosome 17. The upper graphic shows the log ratios of all experiments mapped to their chromosomal location. The middle graphic represents the mean and the median log ratio per gene over all experiments mapped to the chromosomal locations of the genes. The lower graphic shows the variance of a log ratio per gene over all experiments mapped to the chromosomal locations of the genes.

Chapter 6

Analysing Microarray Data

A microarray experiment produces an huge amount of data and therefore efficient and realistic models to analyse these data are required. To get an impression how good our developed *HMM* approach works on real microarray data for gene expression and *ArrayCGH* experiments the comparison of the annotation results with published results is a necessary step. In this chapter we present detailed information about the performance of our method on breast cancer microarray data.

The following topics are contained in this chapter:

1. In the Section 6.1 we analyse the performance of the standard *HMM*s and the extended *HMM*s on our breast cancer gene expression data set of chromosome 17.

2. The performance of the the extended *HMM*s on our breast cancer *ArrayCGH* data set of chromosome 17 is analysed in the Section 6.2.

3. The influences of DNA copy number changes on the gene expression levels are described in the Section 6.3.

6.1 Analysing Gene Expression Profiles

We start to analyse the gene expression profiles of our data set of chromosome 17. In the first part of the analysis we use the standard *HMM*s of the Definition 2.2.1 and in the second part we show how the performance is improved using the *HMM*s with coupled transition matrices (extended *HMM*s) which are introduced in the Definition 3.2.1. The results of the analysis with the extended *HMM* are directly compared with the results of the standard *HMM*.

6.1.1 Standard *HMM*s

The standard *HMM*s are created as we have described in the Section 5.1. We have tested the standard *HMM*s with the same estimated mixture model and different initial transition matrices. The annotation quality of these *HMM*s were nearly equal. For that reason we present the representative results which have been created by a single *HMM*.

Over-expression

The Table 6.5 shows an overview of genes which have been annotated as over-expressed in all gene expression profiles. The segmentation of these genes to their bands allows us to find regions which have been frequently annotated as over-expressed. In general this table confirms that the gene expression profiles of all experiments differ from each other and this fact is consistent with the gene expression correlation matrix in the Figure 5.3 and the overview of all gene expression profiles in the Figure 5.4.

Less of the genes on the p-arm of chromosome 17 have been annotated as over-expressed as on the q-arm. Let us consider the annotation results for the p-arm and the q-arm over all gene expression profiles. Ten genes have been annotated as over-expressed in the region 17p13.3-17p12 and twenty-four of such annotations are located in the chromosomal band 17p11.2. The q-arm contains the eight regions 17q11.2, 17q12, 17q21.2, 17q21.32, 17q21.33, 17q23.2, 17q25.1 and 17q25.3 where more than twenty genes have been annotated as over-expressed and thereby more than forty of these annotations are located in each of the regions 17q11.2, 17q12, 17q21.2, 17q21.33 and 17q25.1.

This general overview includes chromosomal regions which have been mentioned in literature. Kauraniemi *et al.* [13] and Willis *et al.* [27] have found that the band 17q12 can be over-expressed in breast cancer. The region 17q23 has been seen as over-expressed by Monni *et al.* [17], and Hyman *et al.* [9] have observed over-expressed genes in 17q21.3. The analyses of Pollack *et al.* [23] and Orsetti *et al.* [20] confirm that the main reason for most of the over-expressed genes are amplifications of DNA segments where these genes are located. Later when we analyse the *ArrayCGH* data we will see which bands are mainly affected by DNA amplifications.

Now we consider which genes have been annotated as over-expressed. The Table 5.2 contains genes which are known to be over-expressed in some types of breast cancer. All these genes have been annotated as over-expressed by the standard *HMM*, but the gene *ABCA5* represents an annotation error as we will see later. In total one hundred sixty-seven genes of the two hundred sixty-five genes in the data set have been annotated as over-expressed over all gene expression profiles. This is about sixty-three percent of all genes. The annotation attributes *Nr*, *Max*, *Min* and *Mean* of the candidate genes for over-expression from Table 5.2 the are shown in the Table 6.1. These attributes give us a good overview how significant the annotations of the candidate genes are. The value *Nr* of a gene *g* counts in how many gene expression profiles the gene *g* has been annotated as over-expressed. The attribute *Max* of a gene *g* contains the maximum log ratio of the gene *g* and the attribute *Min* of a gene *g* represents the minimum log ratio of the gene *g* for all gene expression profiles where the gene *g* has been annotated as over-expressed. The value of the attribute *Mean* of gene *g* shows the mean log ratio of gene *g* for all gene expression profiles where the gene *g* has been annotated as over-expressed. In the ideal case the value of *Nr* counts several gene expression profiles and the values of *Max*, *Min* and *Mean* are significantly greater than zero. In general the attributes of the Table 6.1 fulfil these criterions. An exception is the gene *ABCA5* which has not correctly been annotated. The attributes *Max*, *Min* and *Mean* of this gene are significantly less than zero and therefore these attribute values refer to under-expression. This annotation error has occurred because *ABCA5* is located in a region with a high expression level and therewith the standard *HMM* has problems to separate such a region into over-expressed and under-expressed segments. Form time to time we can observe these problems as negative *Min* values in the Table 6.1 show. These cases are counted in the attribute *E* of the Table 6.5.

The Table 6.2 contains genes that have been annotated as over-expressed in addition to the Table 5.2. To get a more general view on the additional candidates we use a threshold value which is the minimal number of profiles where a gene must have been annotated as over-expressed. We choose the threshold value three for the p-arm and four for the q-arm and therewith we make a compromise to the low correlations between the gene expression profiles. We use the *Entrez Gene* at the *NCBI* to get a better impression what functions some of these candidate genes have.

- *MAPK7* encodes a protein which is a member of the MAP kinase family. The mitogen-activated protein kinase 7 *MAPK7* is specifically activated by the mitogen-activated protein kinase kinase 5 *MAP2K5*. *MAPK7* is involved in the downstream signalling processes of various receptor molecules including receptor type kinases and G protein-coupled receptors. In response to extracellular signals *MAPK7* translocates to cell nucleus where it regulates the gene expression by phosphorylating and activating different transcription factors.

- *MFAP4* encodes a protein with similarity to a bovine microfibril-associated protein. This protein has binding specificities for both collagen and carbohydrate. It is thought to be an extracellular matrix protein which is involved in cell adhesion or intercellular interactions.
- *SCYA14* is one of several cytokine genes clustered on the q-arm of chromosome 17. These cytokines are secreted proteins which are characterised by two adjacent cysteines. The cytokine which is encoded by this gene induces changes in intracellular calcium concentration and enzyme release in monocytes. The gene *SCYA3L1* encodes a cytokine, too.
- *KRT17* encodes the type I intermediate filament chain keratin 17. Rijn *et al.* [26] have found that the expression of *KRT17* is associated with a poor clinical outcome of breast cancer.
- *FLJ22041* encodes a protein which belongs to the FKBP-type peptidyl-prolyl cis/trans isomerase family. This protein is located in endoplasmic reticulum and acts as molecular chaperones.
- *HOXB6* is a member of the Antp homeobox family and encodes a protein with a homeobox DNA-binding domain. It is included in a cluster of homeobox B genes which are located on chromosome 17. The encoded protein functions as a sequence-specific transcription factor that is involved in development and has been localised to both the nucleus and cytoplasm. Altered expression of this gene or a change in the subcellular localisation of its protein is associated with some cases of acute myeloid leukaemia and colorectal cancer.
- *COL1A1* encodes the major component of type I collagen, the fibrillar collagen found in most connective tissues, and the only component of the collagen found in cartilage. Mutations in this gene are associated with osteogenesis imperfecta, Ehlers-Danlos syndrome, and idiopathic osteoporosis. Reciprocal translocations between chromosomes 17 and 22, where this gene and the gene for platelet-derived growth factor beta are located, are associated with a particular type of skin tumour called dermatofibrosarcoma protuberans, resulting from unregulated expression of the growth factor.

The candidate genes *HOXB6* has been observed to play a role in cancer and the attributes *Max*, *Min* and *Mean* could indicate that this gene could have an important role in specific types of breast cancer. *KRT17* has an already known role in the outcome of breast cancer as Rijn *et al.* [26] have found. Also genes that act in signalling processes can be targets for mutations that can cause cancer and when we consider our additional candidate genes then *MAPK7* could be such a candidate gene.
The Table 6.2 contains also genes as *PNMT* and *SC65* or others with questionable attribute values *Max*, *Min* and *Mean* for over-expressed genes. As we have explained above such annotations can happen when the affected genes are located in a segment with a higher expression level. The standard *HMM* has problems to find the start or the end of such segments and therefore segments with higher expression levels which are located close to each other can be annotated as one segment with higher expression and therewith genes with lower expression levels can get wrong annotations. Later we will observe this behaviour in the graphics of selected gene expression profiles.

Gene	Band	Nr	Max	Min	Mean	Gene	Band	Nr	Max	Min	Mean
KIAA0524	q11.2	3	2.11	0.55	1.46	JUP	q21.2	5	2.77	-0.87	1.48
UNC119	q11.2	3	2.9	2.48	2.71	HOXB5	q21.32	5	4.28	1.98	3.06
SDF2	q11.2	2	3.16	2.49	2.83	NDP52	q21.32	2	1.81	1.61	1.71
TRAF4	q11.2	3	3.41	2.63	2.91	NGFR	q21.33	3	2.5	-1.98	0.47
FLJ10700	q11.2	3	5.7	1.98	3.91	HBOA	q21.33	4	3.97	0.67	2.42
TIAF1	q11.2	3	4.34	1.88	2.99	DLX4	q21.33	7	4.74	-0.02	2.79
KIAA1321	q11.2	3	3.61	1.73	2.43	ABCC3	q21.33	5	5.01	-1.18	2.83
SCYA3	q12	6	2.64	0.38	1.68	RAD51C	q23.2	2	2.55	2.43	2.49
SCYA4	q12	6	2.6	-0.95	1.23	RPS6KB1	q23.2	6	2.98	1.81	2.36
MLLT6	q12	5	4.89	0.34	2.54	APPBP2	q23.2	5	2.18	1.12	1.63
ZNF144	q12	5	3.67	0.03	2.32	PPM1D	q23.2	5	2.83	0.61	1.77
PIP5K2B	q12	4	4.34	2.94	3.58	TBX2	q23.2	3	1.22	0.52	0.79
CACNB1	q12	5	3.85	-0.22	2.04	TRAP240	q23.2	3	2.86	1.32	2.06
RPL19	q12	6	2.06	0.21	0.77	ICAM2	q23.3	2	2	1.39	1.69
MLN64	q12	10	4.9	1.98	3.62	PECAM1	q23.3	2	2.4	1.7	2.05
ERBB2	q12	10	5.07	2.4	3.64	ABCA5	q24.2	1	-2.08	-2.08	-2.08
GRB7	q12	10	6.16	2.43	4.22	SLC9A3R1	q25.1	2	4	2.81	3.4
NR1D1	q21.1	8	2.28	-1.1	0.71	AD023	q25.1	2	2.78	2.5	2.64
CDC6	q21.2	9	4.54	-0.96	1.82	GRB2	q25.1	3	2.32	1.73	2.1
TOP2A	q21.2	7	5.18	1.02	1.94	ITGB4	q25.1	5	5.18	0.17	2.72
SMARCE1	q21.2	6	6.48	1.06	4.06	HCNGP	q25.1	5	2.32	0.07	1.29
KRT20	q21.2	2	1.74	1.1	1.42	BIRC5	q25.3	3	2.64	1.68	2.22
KRTHA4	q21.2	2	1.44	0.32	0.88	LGALS3BP	q25.3	4	2.93	1.41	2.29
KRT19	q21.2	7	3.29	-2.38	1.2						

Table 6.1: Over-expressed candidate genes from Table 5.2 which have been annotated as over-expressed by the standard *HMM*. Max represents the highest log ratio which was measured for the Nr experiments where the gene has been annotated as over-expressed. Min represents the lowest log ratio which was measured for the Nr experiments where the gene has been annotated as over-expressed. Mean is the mean log ratio of the measured log ratios for the Nr experiments where the gene has been annotated as over-expressed.

Gene	Band	Nr	Max	Min	Mean	Gene	Band	Nr	Max	Min	Mean
MAPK7	p11.2	4	3.06	0.25	1.61	KRT17	q21.2	10	5	-1.01	2.2
MFAP4	p11.2	6	4.83	-2.71	2.28	SC65	q21.2	5	1.55	0.53	1.19
ALDH10	p11.2	3	2.86	-0.55	1.47	FLJ22041	q21.2	5	2.64	-1.64	1.32
ALDOC	q11.2	4	2.92	-0.21	1.26	HOXB6	q21.32	5	5.23	1.93	3.58
SCYA14	q12	6	5.28	-2.17	2.4	UGTREL1	q21.33	4	4.12	0.2	1.97
SCYA18	q12	4	3.95	1.8	2.94	COL1A1	q21.33	5	3.69	-4.63	-0.39
SCYA3L1	q12	7	2.46	-2.21	0.66	CLTC	q23.2	4	5.04	0.37	2.56
TCF2	q12	4	3.92	0.05	1.54	CEP4	q25.1	4	3.28	1.36	2.43
NAP4	q12	5	5.12	2.19	3.64	H3F3B	q25.1	4	2.7	0.42	1.46
PNMT	q12	10	0.7	-0.52	0	WBP2	q25.1	4	3.23	-0.34	1.59
KRT13	q21.2	6	5.47	-1.76	1.36	ACOX1	q25.1	4	3.53	0.04	1.67

Table 6.2: Additional genes on chromosome 17 which have been annotated as over-expressed by the standard *HMM*. For the p-arm a threshold value of Nr ≥ 3 has been used. The threshold value for the q-arm has been set to Nr ≥ 4. Max represents the highest log ratio which was measured for the Nr experiments where the gene has been annotated as over-expressed. Min represents the lowest log ratio which was measured for the Nr experiments where the gene has been annotated as over-expressed. Mean is the mean log ratio of the measured log ratios for the Nr experiments where the gene has been annotated as over-expressed.

Under-expression

Let us now consider in what regions of chromosome 17 the genes are located which have been annotated as under-expressed. The Table 6.6 represents an overview of genes which have been annotated as under-expressed in all gene expression profiles. The p-arm shows only twenty genes which have been annotated as under-expressed over all gene expression profiles. Each of the bands 17p13.2, 17p13.1 and 17p11.2 contains five of these candidates and every band of the p-arm has at least two of these annotations. In general it seems that the p-arm does not play the main role in the development of breast cancer. The situation on the q-arm is different. Here the bands 17q11.2, 17q12, 17q21.2, 17q21.32, 17q21.33, 17q25.1 and 17q25.3 are mainly affected by genes which have been annotated as under-expressed. Each of the regions 17q12, 17q21.2 and 17q21.33 contains more than nineteen of these genes. Candidate genes for under-expression that play a more general role in breast cancer could be located in these regions. The bands which have been affected by under-expression are also known in literature to be candidates for losses of DNA segments. Orsetti *et al.* 1999 [20] found that the regions 17q11-q21 and 17q25 can contain losses of DNA segments. A newer study of Orsetti *et al.* 2004 [21] refined these results. They have found that the regions 17p, 17q11.2 and 17q21 are mainly affected by DNA losses and that the bands 17q21.3, 17q22 and 17q25 can contain losses or gains of DNA segments.

When genes are located in a DNA segment which has been lost then this can cause the under-expression of these genes if the other copy on the homologous chromosome is not able to compensate this loss. This could have happened with some of the genes which have been annotated as under-expressed. But it is also thinkable that regulation mechanisms which control the gene expression have been affected by mutations and therefore we can also observe genes as under-expressed. The analysis of *ArrayCGH* profiles could give more security for individual genes.

As we mentioned at the end of the Section 5.3 the bands 17q21.1, 17q22, 17q23.1, 17q24.1, 17q24.3 and 17q25.2 contain too less genes to make a general statement about the gene expression in these regions. From the Table 5.1 we know that twelve gene expression values have been measured per profile for the band 17q23.2 when we ignore missing data values. On the basis of the Tables 6.5 and 6.6 we know that the band 17q23.2 is mainly affected by over-expression. Only one gene in this region has been annotated as under-expressed and twenty-nine genes have been annotated as over-expressed over all gene expression profiles.

Now we consider which genes have been annotated as under-expressed. The basis of our data set is the study of Pollack *et al.* [23]. In this study no under-expressed genes have explicitly been mentioned, but the figures of the microarray data show that they exist. Orsetti *et al.* [21] have mainly found under-expressed genes in the regions 17q11 and 17q21, but they have used other cell lines and cell probes in comparison with Pollack *et al.* [23] and therefore the results are not directly comparable. We compare the ninety-two genes which have been annotated as under-expressed with the over-expressed candidate genes of the Table 5.2 to get a better impression what candidate genes are also affected by under-expression. The genes *CACNB1*, *TRAP240*, *UNC119*, *RPL19*, *SDF2*, *NDP52*, *ABCA5*, *TIAF1*, *AD023*, *KIAA1321*, *RAD51C*, *RPS6KB1*, *MLLT6*, *KRT20*, *APPBP2*, *ZNF144*, *KRTHA4*, *PPM1D*, *PIP5K2B* and *TBX2* are the twenty candidate genes for over-expression which have never been annotated as under-expressed. All other candidate genes for over-expression of the Table 5.2 are listed in the Table 6.3. The attributes *Max*, *Min* and *Mean* of the genes *SMARCE1* and *JUP* indicate more to identical gene expression than to under-expression. In particular, the genes *ERBB2*, *GRB7*, *SCYA3* and *LGALS3BP* have been observed as under-expressed more than four times. Perou *et al.* [22] have characterised variations in gene expression patterns in human breast tumours and a result has been that they have found two breast cancer tumour groups, the one group expresses *ERBB2* on a high level and the other failed to express *ERBB2*. Security for our results is given by the fact that Perou *et al.* [22] have used some of the tumour cell lines as Pollack *et al.* [23] have done.

Gene	Band	Nr	Max	Min	Mean	Gene	Band	Nr	Max	Min	Mean
KIAA0524	q11.2	1	-1.93	-1.93	-1.93	HOXB5	q21.32	3	-1.5	-2.78	-2.3
TRAF4	q11.2	1	-1.72	-1.72	-1.72	NGFR	q21.33	2	-2.53	-2.7	-2.61
FLJ10700	q11.2	1	-2.53	-2.53	-2.53	HBOA	q21.33	2	-1.12	-1.61	-1.37
SCYA3	q12	6	-1.33	-2.82	-1.9	DLX4	q21.33	3	-1.97	-3.15	-2.6
SCYA4	q12	4	-0.36	-1.76	-1.36	ABCC3	q21.33	4	-1.42	-2.48	-1.83
MLN64	q12	2	-2.36	-3.1	-2.73	ICAM2	q23.3	3	-0.55	-2.45	-1.67
ERBB2	q12	7	-2.07	-3.13	-2.44	PECAM1	q23.3	3	-3.31	-4.5	-3.9
GRB7	q12	7	-1.7	-2.42	-2.11	SLC9A3R1	q25.1	2	-2.47	-3.76	-3.12
NR1D1	q21.1	2	-0.36	-1.82	-1.09	GRB2	q25.1	2	-1.53	-2.01	-1.77
CDC6	q21.2	2	-2.27	-2.82	-2.55	ITGB4	q25.1	3	-1.92	-2.69	-2.32
TOP2A	q21.2	2	-1.86	-2.93	-2.4	HCNGP	q25.1	1	-2.66	-2.66	-2.66
SMARCE1	q21.2	1	-0.38	-0.38	-0.38	BIRC5	q25.3	1	-2.59	-2.59	-2.59
KRT19	q21.2	3	-0.45	-3.36	-2.35	LGALS3BP	q25.3	5	-1.92	-3.43	-2.72
JUP	q21.2	1	-0.91	-0.91	-0.91						

Table 6.3: Over-expressed candidate genes from Table 5.2 which have also been annotated as under-expressed by the standard *HMM*. Max represents the highest log ratio which was measured for the Nr experiments where the gene has been annotated as under-expressed. Min represents the lowest log ratio which was measured for the Nr experiments where the gene has been annotated as under-expressed. Mean is the mean log ratio of the measured log ratios for the Nr experiments where the gene has been annotated as under-expressed.

Let us now consider the genes in the Table 6.4 which have additionally been annotated as under-expressed. These genes are only a subset of the ninety-two different genes that have been annotated as under-expressed over all gene expression profiles. The attributes *Max*, *Min* and *Mean* are in general significantly less than zero and therefore the annotation should be reliable. As above we use the *Entrez Gene* at the *NCBI* to get a better impression what functions some of these candidate genes have.

- *KPNA2* encodes a protein which is involved in the nuclear transport of proteins and could also play a role in recombination processes.

- *CLDN7* encodes a protein which is involved in the formation of tight junctions between epithelial cells. Kominsky *et al.* [15] have found that the loss of *CLDN7* correlates with the histological grade of in situ and invasive ductal carcinomas of the breast. They have also described the potential role of *CLDN7* in the progression and ability of breast cancer cells to disseminate. The expression of *CLDN7* is lower in invasive ductal carcinomas of the breast than in normal breast epithelium. This is exactly what we have found for some of the gene expression profiles.

- *LGALS9* encodes a galectin which is implicated in modulating cell-cell and cell-matrix interactions. Irie *et al.* [10] have found that *LGALS9* is a possible prognostic factor with antimetastatic potential in breast cancer. They have observed that tumours with a low expression level of *LGALS9* do not form tight clusters during the in vitro proliferation.

- The genes *SCYA2*, *SCYA7*, *SCYA11*, *SCYA13*, *SCYA14*, *SCYA18* and *SCYA3L1* are cytokines which are clustered on the q-arm of chromosome 17. These cytokines are secreted proteins which are involved in immunoregulatory and inflammatory processes.

- *SOX9* encodes a protein which interacts with chromatin and activates the transcription via regulation of chromatin modification.

Gene	Band	Nr	Max	Min	Mean	Gene	Band	Nr	Max	Min	Mean
KPNA2	p13.3	2	-2.09	-2.58	-2.33	SCYA13	q12	6	-1.53	-2.07	-1.75
RAB5EP	p13.2	3	-2.39	-2.64	-2.54	SCYA14	q12	7	-1.45	-4.41	-2.55
CLDN7	p13.1	2	-2.6	-2.71	-2.65	SCYA18	q12	6	-0.34	-2.85	-1.9
PMP22	p12	2	-3.81	-3.92	-3.86	SCYA3L1	q12	3	-2.18	-4.3	-3.09
MFAP4	p11.2	2	-1.88	-2.41	-2.15	KRT17	q21.2	6	-2.18	-4.29	-3.1
ALDH10	p11.2	2	-1.37	-1.88	-1.63	FLJ22041	q21.2	4	-1.95	-2.72	-2.3
LGALS9	q11.2	4	-2.38	-4.23	-3.12	COPZ2	q21.32	4	-1.34	-2.85	-2.25
EVI2B	q11.2	4	-1.68	-2.73	-2.07	SKAP55	q21.32	3	-2.91	-3.15	-3.03
EVI2A	q11.2	4	-1.83	-3.36	-2.63	COL1A1	q21.33	7	-2.06	-4.26	-3.05
SCYA2	q12	5	-1.55	-4.04	-2.65	CD79B	q23.3	3	-1.53	-3.78	-2.59
SCYA7	q12	3	-1.21	-2.38	-1.92	SOX9	q24.3	4	-2.46	-3.6	-2.94
SCYA11	q12	3	-1.22	-1.83	-1.63	TIMP2	q25.3	4	-1.56	-2.65	-2.19

Table 6.4: Additional genes on chromosome 17 which have been annotated as under-expressed by the standard *HMM*. For the p-arm a threshold value of Nr ≥ 2 has been used. The threshold value for the q-arm has been set to Nr ≥ 3. Max represents the highest log ratio which was measured for the Nr experiments where the gene has been annotated as under-expressed. Min represents the lowest log ratio which was measured for the Nr experiments where the gene has been annotated as under-expressed. Mean is the mean log ratio of the measured log ratios for the Nr experiments where the gene has been annotated as under-expressed.

- *TIMP2* encodes a protein which is a member of the TIMP gene family. This protein is an inhibitor of matrix metalloproteinases and supresses directly the proliferation of endothelial cells. The gene product of *TIMP2* is involved in the maintenance of a tissue. Nakopoulou *et al.* [18] have found that *TIMP2* is involved in the degradation of extracellular matrix which leads to the invasion of cancer into the surrounding matrix. The basis of this study have been breast cancer cells.
- The function of *COL1A1* have been mentioned above as we discussed the results of genes which have been annotated as over-expressed.

In summary, for the genes *CLDN7*, *LGALS9* and *TIMP2* publications are available which describe the role of these genes in breast cancer when low expression levels have been observed. The standard *HMM*s are able to find such candidate genes.

Selected gene expression profiles

The individually selected gene expression profiles in the Figures 6.1, 6.2, 6.3, 6.4 and 6.5 are now described in detail. Thereby we will discuss the quality of the annotations and so we will obtain a better impression of strengths and weaknesses of the standard *HMM* approach.

- The Figure 6.1 shows the annotations of the cell lines *BT474* and *SKBR3*. In the profile of *BT474* are the bands 17p13.1, 17q11.2, 17q12, 17q21.2, 17q21.32 and 17q25.3 affected by genes which have been annotated as under-expressed. All these genes have log ratios which are significantly less than zero and the state posteriors $sp(-) := \gamma_t(-)$ for these genes are significantly greater than zero and so these annotations should be reliable. The genes which have been annotated as over-expressed are located in the three segments 17q12-21.2, 17q21.32-22 and 17q23.2. The obviously higher expression levels in these regions can also be seen in the high state posterior $sp(+) := \gamma_t(+)$ within these regions. Some annotation problems have occurred in the over-expressed segments. Genes with log ratios around zero or significantly less than zero have been annotated as over-expressed. We have explained these problems above in more detail. The state posterior $sp(+)$ has only a little smaller

values for such annotation problems as for the rest of the over-expressed genes in such a segment. In general, the standard *HMM* has some problems to model the start and the end of segments.
The cell line *SKBR3* contains genes in the bands 17p12, 17q11.2, 17q12, 17q21.33 and 17q23.3 which have been annotated as under-expressed. The log ratios of these genes are significantly less than zero and the state posterior $sp(-)$ shows peaks for these genes. This cell line contains one gene in the band 17p13.3 and a segment in the region 17q12-21.2 which both have been annotated as over-expressed. This segment overlaps with the over-expressed segment 17q12-21.2 of the cell line *BT474* and includes also some genes with log ratios less than zero. The mainly high state posterior values sp for the *Viterbi Path* show that these annotations are significant and that the standard *HMM* has computed a *Viterbi Path* which has a higher probability than other annotation paths.

- The Figure 6.2 contains the annotations of the primary breast tumours *NORWAY 14* and *NORWAY 26*. The *NORWAY 14* profile shows two regions which have been annotated as under-expressed, the one in the band 17q21.31 and the other in the band 17q25.3. The segments which have been annotated as over-expressed are smaller and more widespread than the over-expressed segments in the cell lines above. Nevertheless, the state posterior profiles $sp(-)$ and $sp(+)$ of the states $-$ and $+$ show significant peaks in such segments. The over-expressed segments are located in the bands 17q11.2, 17q12, 17q21.2, 17q21.31 and 17q21.32. Genes which have log ratios which are significantly less than zero have not been annotated as over-expressed. The state posterior profile sp of the *Viterbi Path* illustrates the good quality of this annotation.
The primary tumour *NORWAY 26* contains only one gene in the band 17q23.3 which has been annotated as under-expressed. The state posterior profile $sp(-)$ of the state $-$ has also peaks in the bands 17q11.2 and 17q25 and therefore it is thinkable that the affected genes in these bands could also be under-expressed. This example motivates how it is possible to find additional candidates by comparing the annotation with the state posterior profiles. The *NORWAY 26* profile contains one segment in the band 17q12 which has been annotated as over-expressed. The state posterior profile $sp(+)$ of the state $+$ contains also peaks in the bands 17p13.1, 17q11.2, 17q22, 17q23.3 and 17q25.1. In the state posterior profile sp of the *Viterbi Path* are such regions represented by smaller probabilities. In summary, such regions are contained in nearly every gene expression profile and in most of these cases the log ratios of gene expression values are not significant enough to get other annotations for these genes.

- The Figure 6.3 shows the annotations of the primary breast tumours *NORWAY 47* and *NORWAY 53*. The profile *NORWAY 47* does not contain genes which have been annotated as under-expressed, but the state posterior profile $sp(-)$ of the state $-$ has peaks in the bands 17p13.3, 17p13.1, 17p11.2, 17q12, 17q21.32 and 17q21.33. Thus, the insecurity of the annotation process is visible in the state posterior profile sp of the *Viterbi Path*. The segments 17q11.2, 17q12-21.2, 17q24.3-25.1 and 17q25.1 which have been annotated as over-expressed show significantly high expression levels. The segment in the region 17q12-21.2 contains a subsegment with a lower expression level than the rest of this segment. This subsegment is clearly visible in the state posterior profile sp of the *Viterbi Path* and therewith it is imaginable that this subsegment could be identically expressed.
The profile *NORWAY 53* shows one segment in the band 17q12 which has been annotated as under-expressed and three segments which have been annotated as over-expressed are located in the regions 17p11.2-q11.2, 17q12-21.2 and 17q25.3. The state posterior profile $sp(+)$ of the state $+$ shows also a peak in the region 17q23.3-24.1. The segment 17p11.2-

q11.2 contains a subsegment with a lower expression level. We have observed the same behaviour in the profile of *NORWAY 47*.

- The Figure 6.4 represents the primary breast tumours *NORWAY 100* and *STANFORD 2*. The *NORWAY 100* profile shows two segments in the regions 17q12 and 17q12-21.2 which have been annotated as under-expressed. The segment in 17q12-21.2 contains some genes with higher expression levels as we can observe for other genes in this segment. These genes with higher expression levels are clearly visible on the basis of the peaks in the state posterior profiles *sp* and $sp(-)$ of the *Viterbi Path* and the state $-$. Three segments in the bands 17q12, 17q21.2 and 17q21.32 have been annotated as over-expressed. These segments are small and have significantly higher expression levels.
  The profile of the primary breast tumour *STANFORD 2* shows one segment in the band 17p13.2 which has been annotated as under-expressed and two segments which have been annotated as over-expressed are located in the regions 17q11.2 and 17q12-21.2. The standard *HMM* has problems to end the first over-expressed segment and the second segment contains subsegments of lower expression which lead to peaks in the state posterior profiles. The state posterior profile $sp(-)$ of the state $-$ shows peaks in the bands 17q11.2, 17q12, 17q23.2 and 17q25.3 and the state posterior profile $sp(+)$ of the state $+$ contains peaks in the bands 17q21.2 and 17q21.33.

- The Figure 6.5 contains the primary breast tumours *STANFORD 24* and *STANFORD A*. The profile *STANFORD 24* shows seven segments with significantly low expression levels in the bands 17p11.2, 17q11.2, 17q12, 17q21.33, 17q25.1 and 17q25.3 which have been annotated as under-expressed. This profile does not contain over-expressed annotations, but peaks in the state posterior profile $sp(+)$ of the state $+$ are located in the bands 17p13.1, 17p11.2 and 17q21.1 and therefore we could expect over-expressed candidates in these regions.
  The profile *STANFORD A* shows two genes in the bands 17q11.2 and 17q21.2 which have been annotated as under-expressed and five segments which have been annotated as over-expressed are located in the regions 17q11.2, 17q12, 17q21.2, 17q21.32-33 and 17q23.2-23.3. The under-expressed and the over-expressed regions have characteristic expression levels and the whole profile has a good annotation quality.

Summary of the annotation process

Now we consider the summary of the whole annotation process of our breast cancer gene expression data set. The Figure 6.6 contains the four subfigures *Unchanged Segments*, *Over-expressed Segments*, *Under-expressed Segments* and *Gene Counts*.
The subfigure *Unchanged Segments* shows the locations and the absolute frequencies of segments which have been annotated as identically expressed. The inhomogeneity of gene expression profiles can be seen in the variability of the lengths of identically expressed segments. These lengths are in general greater than the lengths of the segments in the subfigures *Over-expressed Segments* and *Under-expressed Segments*. The darker hexagons indicate that some identically expressed segments with nearly the same locations exist in the gene expression profiles.
The subfigure *Over-expressed Segments* represents the locations and the absolute frequencies of segments which have been annotated as over-expressed. These segments are smaller and show lower variability in comparison with identically expressed segments.
The subfigure *Under-expressed Segments* shows the locations and the absolute frequencies of segments which have been annotated as under-expressed. These segments behave like the over-expressed segments, but in general their lengths are smaller. We have already seen this in the Figures 6.1, 6.2, 6.3, 6.4 and 6.5 of the gene expression profiles. The darker hexagons in the

subfigures *Over-expressed Segments* and *Under-expressed Segments* are the locations where basic candidate genes for different gene expression in subtypes of breast cancer can be found.
The subfigure *Gene Counts* represents the absolute frequencies of under-expressed and over-expressed annotations of a gene in the breast cancer data set. The under-expressed frequencies are shown in the upper subfigure and the over-expressed frequencies in the subfigure below. The overview clarifies that most of the genes which have been annotated as over-expressed have also been annotated as under-expressed. The counts are in general reliable because the log ratios of over-expressed and under-expressed annotations are significantly different from zero. In more detail, the under-expressed annotations do not contain log ratios greater than zero, but the over-expressed annotations include fifty-five log ratios which are less than zero. An overview of annotation errors is given in the column E of the Table 6.5. So we have to say that some of the genes which have been annotated as over-expressed can have too high counts. Nevertheless, we see that the region 17q22-24.2 is mainly affected by over-expression and only some genes in this region have been annotated as under-expressed. The expression level of this region could follow the results of Orsetti *et al.* [21] and [20] and Nugoli *et al.* [19] which we have described above.

	17p					17q																			
Experiment	13.3	13.2	13.1	12	11.2	11.1	11.2	12	21.1	21.2	21.31	21.32	21.33	22	23.1	23.2	23.3	24.1	24.2	24.3	25.1	25.2	25.3	∑	E
BT474								14	1	2		4	12	2		5								40	9
MCF7																5								5	0
SKBR3	1							4	1	13														19	4
T47D										17														17	3
NORWAY 7							2																	2	0
NORWAY 10																								0	0
NORWAY 11							1																	1	0
NORWAY 12											1													1	0
NORWAY 14							1	9		1	1	2												14	0
NORWAY 15	1																						9	10	1
NORWAY 16										10	2													12	0
NORWAY 17																								0	0
NORWAY 18			1	1	3																			5	1
NORWAY 19		1											3								3		4	11	1
NORWAY 26								12																12	1
NORWAY 27		1																						1	0
NORWAY 39								4																4	0
NORWAY 41								7		1	1										1			10	0
NORWAY 47							16	19	1	9										1	13			59	7
NORWAY 48								4				2												6	0
NORWAY 53					12		9	5	1	11													1	39	7
NORWAY 56												2	12	2		5			4	2	22	1	9	59	9
NORWAY 57								11	1	1			1								8			22	0
NORWAY 61					1			5	1	4						2	3							16	1
NORWAY 65					2			1		2		1												6	1
NORWAY 100								3		1		2												6	0
NORWAY 101								10	1	1			2											14	1
NORWAY 102					4					1														5	0
NORWAY 104													1			5	1	2						9	0
NORWAY 109										4		2	1								2			9	1
NORWAY 111				1																				1	0
NORWAY 112					2			1					1											4	0
STANFORD 2							8	4	1	6														19	5
STANFORD 14											1					1					6			8	0
STANFORD 16		2						5									3							10	1
STANFORD 17																								0	0
STANFORD 23							1																	1	0
STANFORD 24																								0	0
STANFORD 35								1					1											2	0
STANFORD 38	1						1					5												7	0
STANFORD A							12	4		3		5	10			6	1							41	2
∑	3	4	1	2	24		51	123	8	87	6	25	44	4		29	8	2	4	3	55	1	23	507	55

Table 6.5: Overview of over-expressed genes per chromosomal band. Candidate genes in a gene expression profile which have been annotated as over-expressed by the standard *HMM* have been assigned to their chromosomal bands. The results are contained in this table. $\sum$ is the row or the column sum of genes which have been annotated as over-expressed. E is the sum of genes in an experiment which have been annotated as over-expressed when their log ratios are less than zero.

	17p					17q																			
Experiment	13.3	13.2	13.1	12	11.2	11.1	11.2	12	21.1	21.2	21.31	21.32	21.33	22	23.1	23.2	23.3	24.1	24.2	24.3	25.1	25.2	25.3	∑	E
BT474			1				2	2		1		1											2	9	0
MCF7							3	7					2				3							15	0
SKBR3				1			3	2					2				1							9	0
T47D				1				2					2											5	0
NORWAY 7										3		2	1											6	0
NORWAY 10							1																	1	0
NORWAY 11																1				1				2	0
NORWAY 12																								0	0
NORWAY 14											1												2	3	0
NORWAY 15		1						3				1												5	0
NORWAY 16								1	1	1							1							4	0
NORWAY 17										1			3								8			12	0
NORWAY 18				1								1	2							1				5	0
NORWAY 19								5																5	0
NORWAY 26																	1							1	0
NORWAY 27													3							1				4	0
NORWAY 39								5					3											8	0
NORWAY 41			1							1		3												5	0
NORWAY 47																								0	0
NORWAY 48							1					2	1								2			6	0
NORWAY 53								2																2	0
NORWAY 56										1		1												2	0
NORWAY 57		2																						2	0
NORWAY 61										1														1	0
NORWAY 65	1							2													1			4	0
NORWAY 100								6	1	4														11	0
NORWAY 101										1		2												3	0
NORWAY 102								4			2	1												7	0
NORWAY 104		1	1				2				1	2								1			3	11	0
NORWAY 109								9			1										1			11	0
NORWAY 111			1		1			5		2							3							12	0
NORWAY 112							2	3																5	0
STANFORD 2		1																						1	0
STANFORD 14																							2	2	0
STANFORD 16										4	1	2												7	0
STANFORD 17	1									2	1								1					5	0
STANFORD 23			1		2			1		1														5	0
STANFORD 24					2		3	4					1								2		2	14	0
STANFORD 35								1										1			2			4	0
STANFORD 38																							1	1	0
STANFORD A							1			2														3	0
∑	2	5	5	3	5		18	64	2	25	7	18	20			1	9	1	1	4	16		12	218	0

Table 6.6: Overview of under-expressed genes per chromosomal band. Candidate genes in a gene expression profile which have been annotated as under-expressed by the standard *HMM* have been assigned to their chromosomal bands. The results are contained in this table. $\sum$ is the row or the column sum of genes which have been annotated as under-expressed. E is the sum of genes in an experiment which have been annotated as under-expressed when their log ratios are greater than zero.

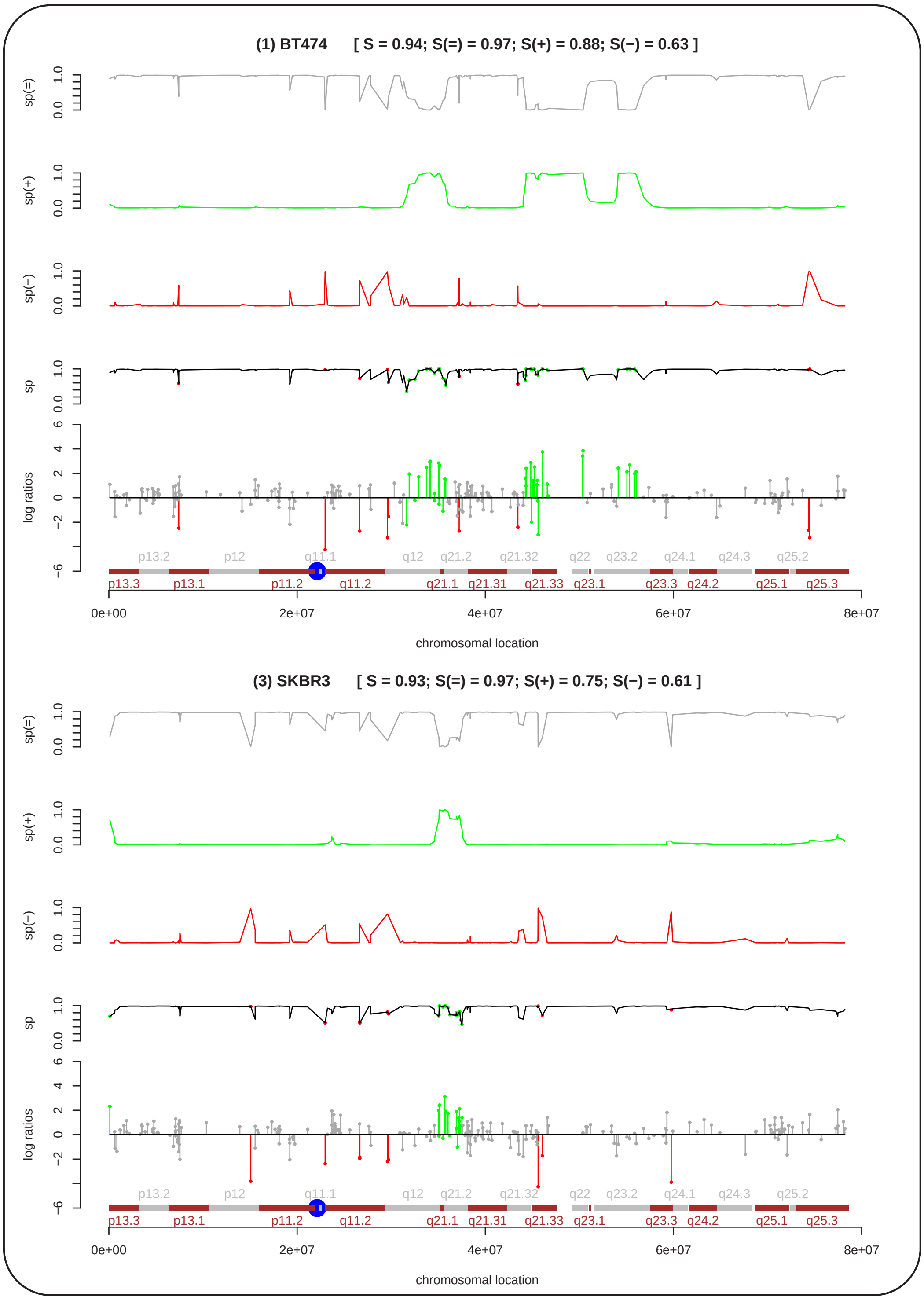


Figure 6.1: Gene expression profiles of chromosome 17 which have been annotated by the standard *HMM*. The first profile shows the *BT474* and the second the *SKBR3* data. In general, the headline of a profile contains the unique profile number (n) which is used in the Figure 5.3, the profile name, the sum S (5.3) of the state posteriors for the *Viterbi Path* and the relative proportion $S(i)$ (5.3), with $i \in \{-, =, +\}$, of the state posterior profile i in the *Viterbi Profile*. Top down the five graphics per experiment show the state posterior profiles for the states $=$, $+$ and $-$, the state posterior profile of the *Viterbi Path* and the annotated gene expression profile. The annotation of a gene is represented by the colour of the log ratio line. A red line represents an as under-expressed, a grey line an as identically expressed and a green line an as over-expressed annotated gene.

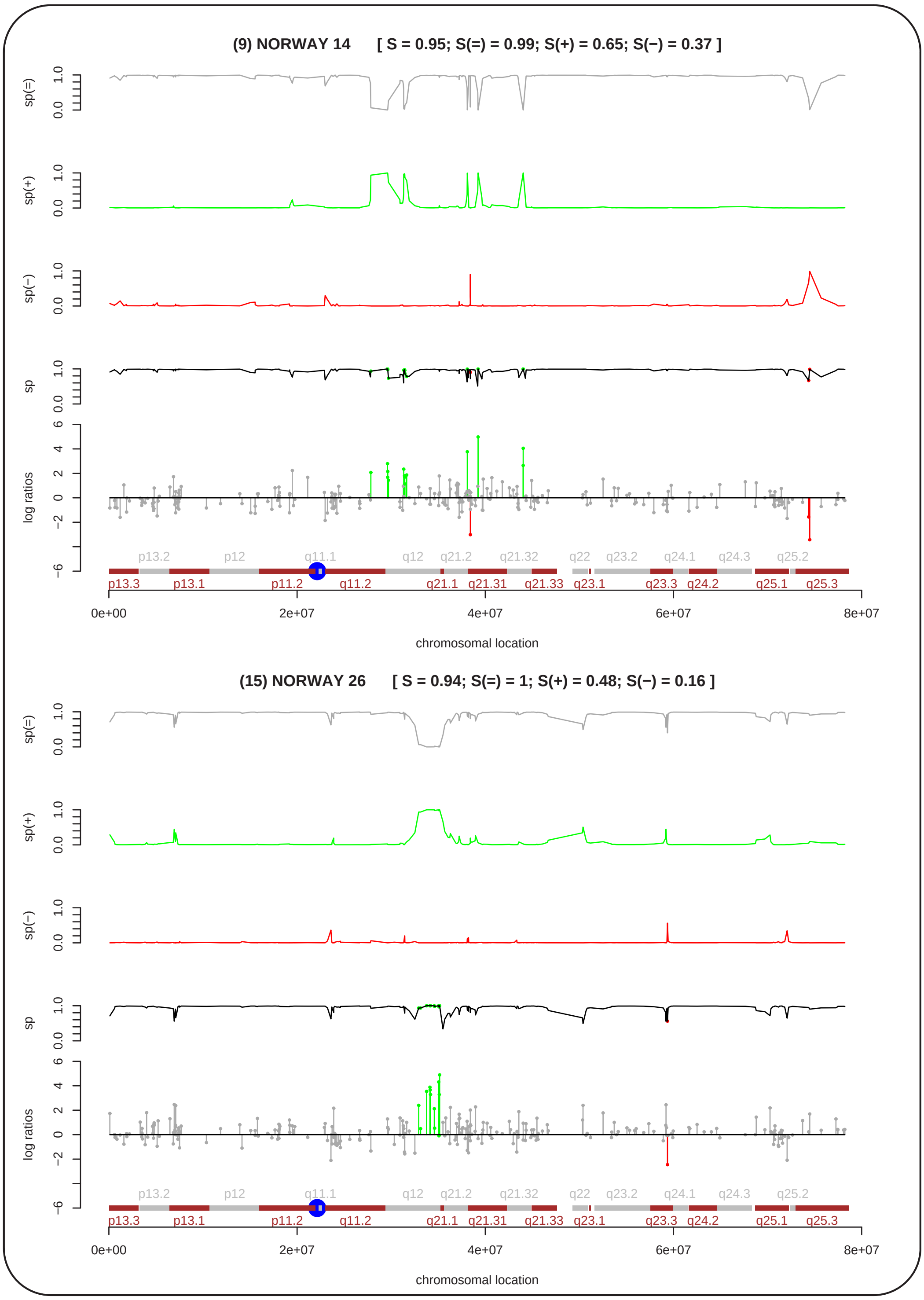


Figure 6.2: Gene expression profiles of chromosome 17 which have been annotated by the standard *HMM*. The first profile shows the *NORWAY 14* and the second the *NORWAY 26* data. In general, the headline of a profile contains the unique profile number (n) which is used in the Figure 5.3, the profile name, the sum S (5.3) of the state posteriors for the *Viterbi Path* and the relative proportion $S(i)$ (5.3), with $i \in \{-, =, +\}$, of the state posterior profile i in the *Viterbi Profile*. Top down the five graphics per experiment show the state posterior profiles for the states $=$, $+$ and $-$, the state posterior profile of the *Viterbi Path* and the annotated gene expression profile. The annotation of a gene is represented by the colour of the log ratio line. A red line represents an as under-expressed, a grey line an as identically expressed and a green line an as over-expressed annotated gene.

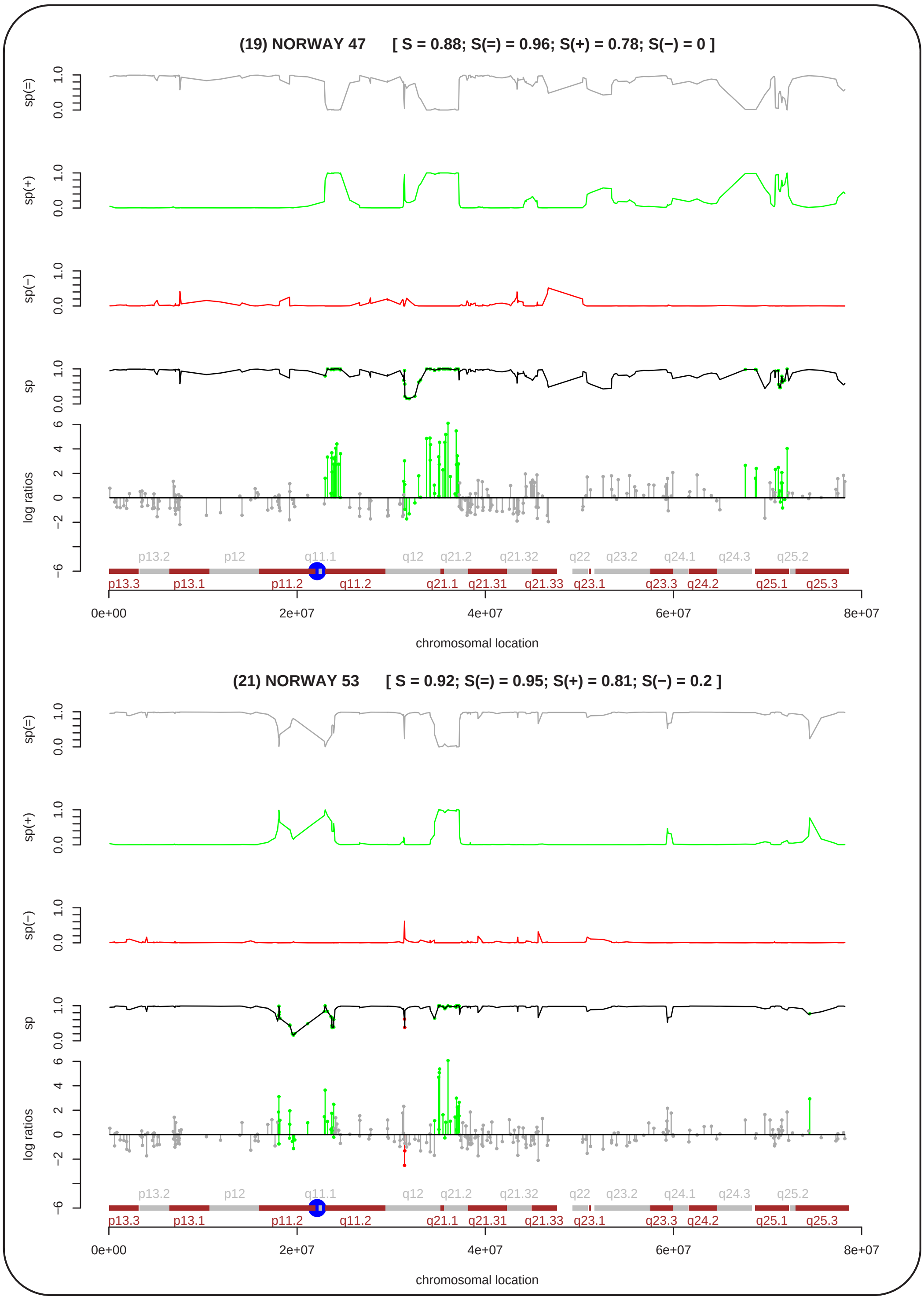


Figure 6.3: Gene expression profiles of chromosome 17 which have been annotated by the standard *HMM*. The first profile shows the *NORWAY 47* and the second the *NORWAY 53* data. In general, the headline of a profile contains the unique profile number (n) which is used in the Figure 5.3, the profile name, the sum S (5.3) of the state posteriors for the *Viterbi Path* and the relative proportion $S(i)$ (5.3), with $i \in \{-, =, +\}$, of the state posterior profile i in the *Viterbi Profile*. Top down the five graphics per experiment show the state posterior profiles for the states $=$, $+$ and $-$, the state posterior profile of the *Viterbi Path* and the annotated gene expression profile. The annotation of a gene is represented by the colour of the log ratio line. A red line represents an as under-expressed, a grey line an as identically expressed and a green line an as over-expressed annotated gene.

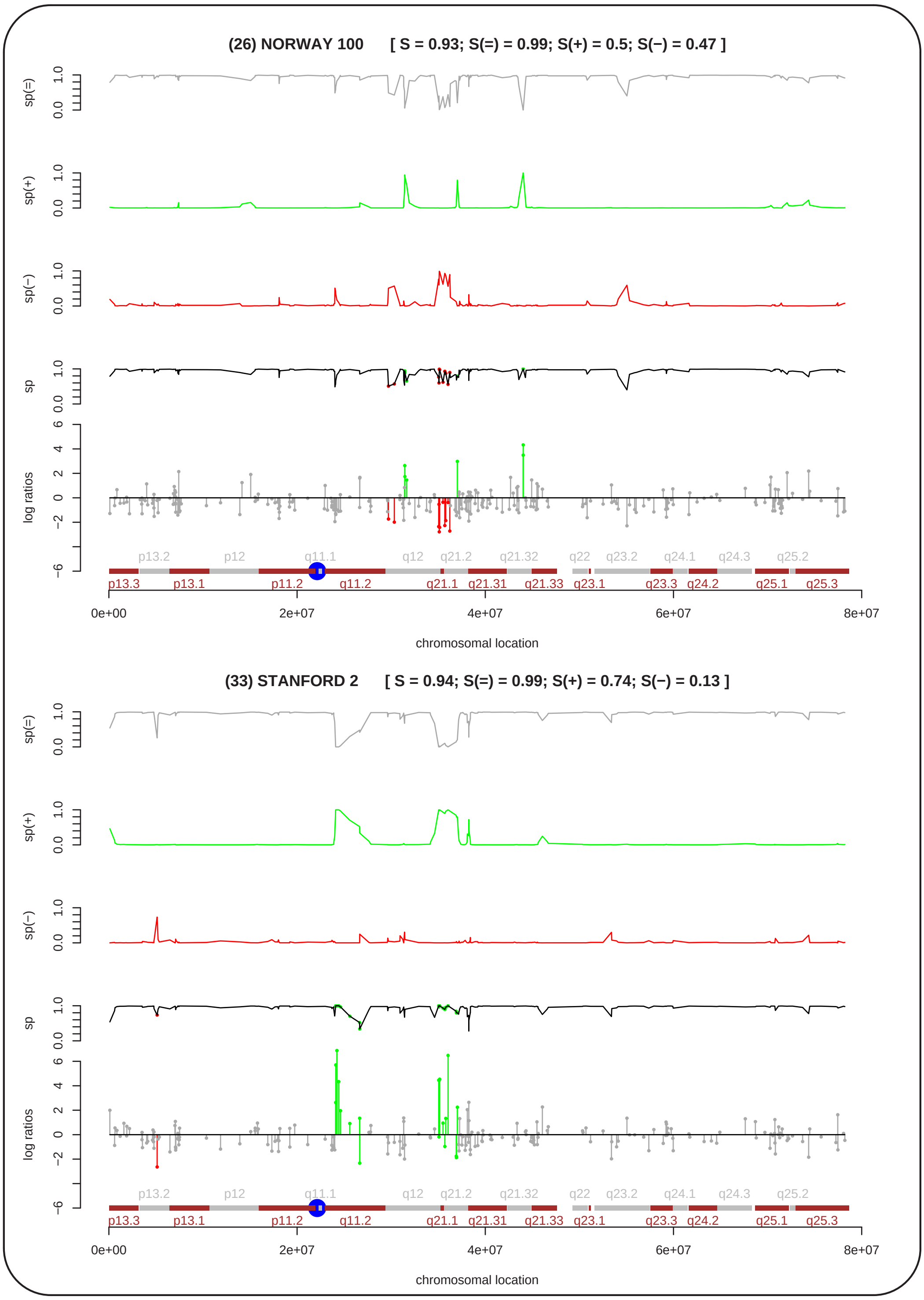


Figure 6.4: Gene expression profiles of chromosome 17 which have been annotated by the standard *HMM*. The first profile shows the *NORWAY 100* and the second the *STANFORD 2* data. In general, the headline of a profile contains the unique profile number (n) which is used in the Figure 5.3, the profile name, the sum S (5.3) of the state posteriors for the *Viterbi Path* and the relative proportion $S(i)$ (5.3), with $i \in \{-, =, +\}$, of the state posterior profile i in the *Viterbi Profile*. Top down the five graphics per experiment show the state posterior profiles for the states $=$, $+$ and $-$, the state posterior profile of the *Viterbi Path* and the annotated gene expression profile. The annotation of a gene is represented by the colour of the log ratio line. A red line represents an as under-expressed, a grey line an as identicallyexpressed and a green line an as over-expressed annotated gene.

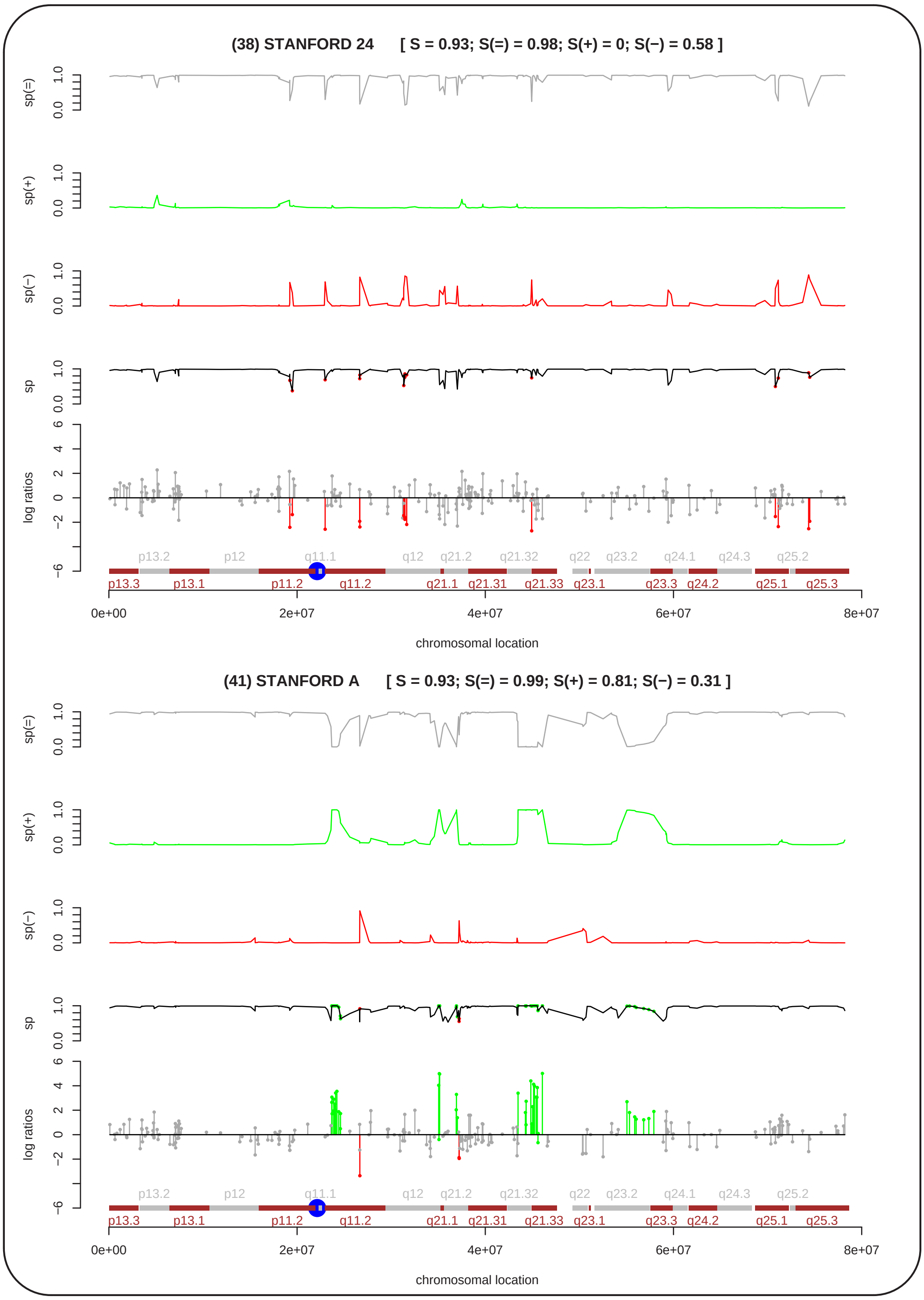


Figure 6.5: Gene expression profiles of chromosome 17 which have been annotated by the standard *HMM*. The first profile shows the *STANFORD 24* and the second the *STANFORD A* data. In general, the headline of a profile contains the unique profile number (n) which is used in the Figure 5.3, the profile name, the sum S (5.3) of the state posteriors for the *Viterbi Path* and the relative proportion $S(i)$ (5.3), with $i \in \{-, =, +\}$, of the state posterior profile i in the *Viterbi Profile*. Top down the five graphics per experiment show the state posterior profiles for the states $=$, $+$ and $-$, the state posterior profile of the *Viterbi Path* and the annotated gene expression profile. The annotation of a gene is represented by the colour of the log ratio line. A red line represents an as under-expressed, a grey line an as identically expressed and a green line an as over-expressed annotated gene.

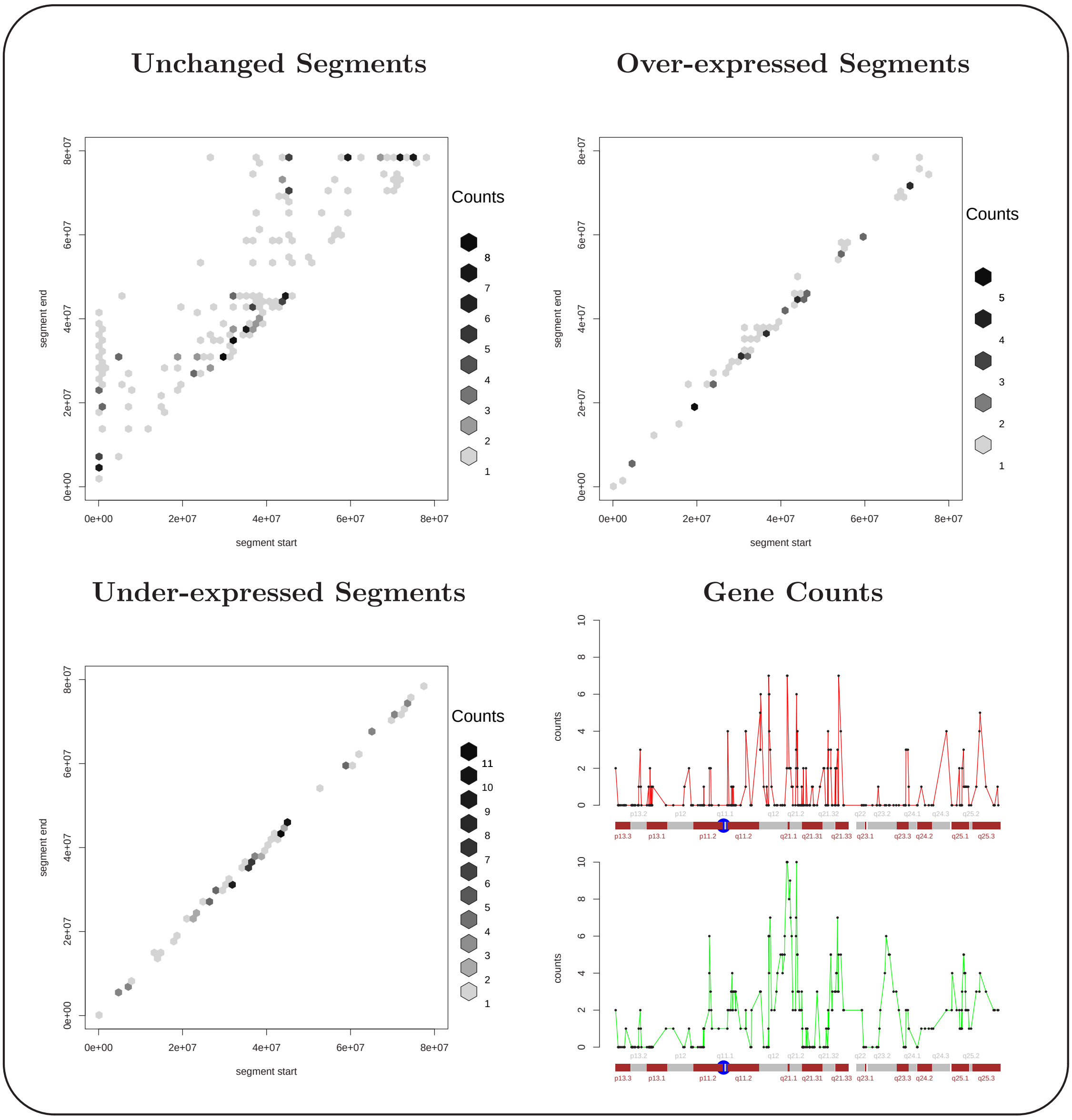


Figure 6.6: Overview of annotations for the standard *HMM*. In general a segment is a sequence of successive genes which have been annotated in the same way. Each segment has a start and an end point on the chromosome. The figure **Unchanged Segments** gives a summary of segments on chromosome 17 which have been annotated as identically expressed. The figure **Over-expressed Segments** shows a summary of segments on chromosome 17 which have been annotated as over-expressed. The figure **Under-expressed Segments** represents a summary of segments on chromosome 17 which have been annotated as under-expressed. The figure **Gene Counts** shows the absolute frequency of over-expressed and under-expressed annotations per gene. The absolute frequencies of under-expressed annotations are shown in the upper subfigure and the absolute frequencies of over-expressed annotations are given in the subfigure below.

6.1.2 *HMM*s With Transition Classes

The *HMM*s with transition classes are created as we have described in the Section 5.1. These *HMM*s extend the standard *HMM* approach to consider proximity effects and thus develop a more realistic model of the effects which can influence the alteration of gene expression levels over regions. The positional proximity of genes on a chromosome and the gene expression levels of these genes can show characteristic patterns. When we consider breast cancer we can easily motivate the usage of the transition class switching approach including the proximity of adjacent genes. All the literature which we have reported at the beginning of the Section 5.3 shows that losses or gains of DNA segments can have important influences on the gene expression levels of genes in these segments.

Modelling of chromosomal imbalances

A natural approach should be able to model over-expression or under-expression in DNA segments respecting the proximity of adjacent genes. So we assume that the chance for two adjacent genes to be both in the same lost or gained segment increases the closer the distance between these two genes is and therefore the probability that these two adjacent genes show the same expression status should be higher. It is realistic to assume that the loss or the gain of a DNA segment does not affect both adjacent genes when the distance between these genes is greater as in the case before. The probability that both genes show the same expression status should be less as in the other case.
We model proximity effects between adjacent genes by mapping the distance between these genes into a set of predefined transition matrices. The mapping is done by a transition class switching function as for example in the Definition (5.1). The transition matrices are coupled with the coupling function of the Definition (3.7). The extended *HMM* requires additional parameters for the transition class switching function and the coupling function. That is, we have to choose threshold values for the distance between adjacent genes which will model the proximity effects, and we have to determine scaling factors for the coupling of the transition matrices which will modify the expected number of genes in a segment. Both together describes our more realistic model from above, because the distance between adjacent genes is used to determine a transition matrix which is scaled by a predefined factor. The choice of good additional parameters without previous knowledge about the underlying biological processes is difficult. The current choice is done by testing the extended *HMM* with different additional parameters and the selection of the *HMM* which shows the best performance. We hope that the progress in cancer research will lead to chromosomal maps where losses and gains of segments are known in detail. Such previous knowledge could be used to determine additional parameters.

Now we represent the annotation results of an *HMM* with two transition classes (extended *HMM*). We use the same estimated mixture model as we have done for the standard *HMM* to have the same initial emission densities. The basic initial transition matrix is equal to the transition matrix of the standard *HMM* and the second is scaled by a factor of two. That is, we use the scaling factors one and two which define the scaling vector $\vec{S} = (1, 2)$. The transition class switching function of the Definition (5.1) is used with a distance threshold value $\mathcal{T} = 30000$ base pairs.

Over-expression

The Table 6.12 represents an overview of genes over all gene expression profiles in our breast cancer data set which have been annotated as over-expressed. The segmentation of this table into chromosomal bands allows the detection of bands which have been frequently annotated by the extended *HMM*. The comparison of these results with the annotation results of the standard *HMM* is also possible. The inhomogeneity of the gene expression profiles in our data set is reflected in the structure of this table. For instance, look at the correlation matrix in the Figure 5.3 or the overview of the gene expression profiles in the Figure 5.4 to recall the quality structure of our breast cancer data set. The general overview of genes which have been annotated as over-expressed includes three hundred fourteen counts for over-expression over all gene expression profiles and only nine of these annotations have log ratios less than zero. In total there are one hundred thirty-two different genes which have been annotated as over-expressed by the extended *HMM*. When we look at the over-expressed annotations of the standard *HMM* we have five hundred seven genes over all profiles which have been annotated as over-expressed and in total one hundred sixty seven different genes have been annotated as over-expressed. The standard *HMM* has done fifty-five over-expressed annotations with log ratios less than zero.

To get a better impression of the annotation quality of the extended *HMM* we analyse the Table 6.12 and compare this table with the Table 6.5. The p-arm of the chromosome 17 contains seventeen genes which have been annotated as over-expressed over all profiles. The chromosomal bands 17p13.3, 17p13.2 and 17p12 contain the equal number of counts for the same gene expression profiles in comparison with the over-expressed annotations of the standard *HMM*. The extended *HMM* has not done over-expressed annotations in the band 17p13.1, but the standard *HMM* has annotated a gene in the profile *NORWAY 18* as over-expressed. Therefore we have inspected both profiles which are not shown here and we have found that the log ratio of this gene is less significant as other log ratios for over-expressed genes. The chromosomal band 17p11.2 contains the first wide differences in the comparison of the extended *HMM* with the standard *HMM*. That is, the extended *HMM* has annotated eight genes over all profiles as over-expressed and the standard *HMM* has found twenty-four. The comparison of these two annotations for the band 17p11.2 shows that the profiles *NORWAY 53* and *NORWAY 102* are the main sources for the differences in the annotation results. The annotation results of the standard *HMM* for the profile *NORWAY 53* are shown in the Figure 6.3. Log ratios with questionable significance can clearly be seen in the over-expressed region of the band 17p11.2. The extended *HMM* has done a better annotation which is shown in the Figure 6.9. The region 17p11.2-q11.2 which has been annotated as over-expressed by the standard *HMM* has been splitted into differently annotated regions by the extended *HMM* so that the significant genes are surrounded by identically expressed regions. The profile *NORWAY 102* seems not to have log ratios that are significant enough to be annotated as over-expressed by the extended *HMM*. Such effects as observed in the profiles *NORWAY 53* and *NORWAY 102* on the p-arm are also seen on the q-arm which we will consider now.

The q-arm contains two hundred seven genes over all gene expression profiles which have been annotated as over-expressed. Each of the chromosomal bands 17q11.2, 17q12, 17q21.2, 17q21.33, 17q23.2 and 17q25.1 has more than nineteen genes which have been annotated as over-expressed. Kauraniemi *et al.* [13] have found over-expression in the region 17q12. Willis *et al.* [27] have observed over-expression in the region 17q12-23. Monni *et al.* [17] have measured over-expression in the region 17q23. The extended *HMM* has annotated these regions as over-expressed in our breast cancer data set and after the inspection of the annotations we can conclude that the extended *HMM* is able to detect these regions.

The Table 6.7 represents the number of over-expressed annotations per band for the q-arm on chromosome 17 for the standard and the extended *HMM*. This table shows that the extended *HMM* has annotated less or the same number of over-expressed genes per band in comparison with the annotation results of the standard *HMM*. We have inspected the annotations of both *HMM*s to find out more about this interesting fact. One result of the analysis of the gene expression profiles with the standard *HMM* is that this approach has problems to model the start and the end of a segment. This has led to questionable annotations for genes whose log

Band	standard *HMM*	extended *HMM*	Band	standard *HMM*	extended *HMM*
q11.2	51	38	q23.2	29	20
q12	123	98	q24.1	2	2
q21.1	8	4	q24.3	3	1
q21.2	87	33	q25.1	55	39
q21.31	6	5	q25.2	1	0
q21.33	44	22	q25.3	23	2
q22	4	4			

Table 6.7: The number of genes per band on the q-arm of chromosome 17 which have been annotated as over-expressed by the standard *HMM* and the extended *HMM* with two transition classes.

ratios do not support these annotations. For instance, look at the genes with negative attribute values in the Tables 6.1 and 6.2. The extended *HMM* is able to model the start and the end of a segment better as the standard *HMM*. We will see this later in more detail when we discuss selected gene expression profiles. Another important observation is that the extended *HMM* has annotated in general only the genes with significant log ratios as over-expressed in comparison with the standard *HMM*.

Now we discuss some gene expression profiles which mainly contribute to the conclusions of the Table 6.7. The basis of this analysis are the Tables 6.12 and 6.5 and selected annotations of both *HMM*s. We concentrate on gene expression profiles whose annotations are shown in the Figures 6.1, 6.3, 6.7 and 6.9. The annotations in the Figures 6.1 and 6.3 have been created by the standard *HMM*. The extended *HMM* has created the annotations in the Figures 6.7 and 6.9.

- The gene expression profile *NORWAY 53* has nine genes in the band 17q11.2 which have been annotated as over-expressed by the standard *HMM* and two such annotations which have been done by the extended *HMM*. When we compare these annotations with the help of the Figures 6.3 and 6.9 we see that the annotation of the extended *HMM* has a better quality because the start and the end of the over-expressed segment is modelled better.

- The breast cancer cell line *BT474* shows fourteen over-expressed annotations in the band 17q12 and twelve such annotations in the band 17q21.33 in the annotation results of the standard *HMM*. The annotation results of the extended *HMM* contain nine over-expressed annotations in the band 17q12 and one in the band 17q21.33. The reason for these different annotations is also the fact that the extended *HMM* has better characterised these over-expressed regions.

- The primary breast tumour *NORWAY 47* has nineteen genes in the band 17q12 and thirteen in the band 17q25.1 which have been annotated as over-expressed by the standard *HMM*. The annotation results of the extended *HMM* show eleven over-expressed annotations in the band 17q12 and five in the band 17q25.1. The extended *HMM* has splitted the over-expressed region 17q12 into three segments. The two external segments have been annotated as over-expressed and these external segments surround an identically expressed segment. The log ratios in these regions indicate such an annotation. Only the genes in the band 17q25.1 which have more significant log ratios have been annotated as over-expressed by the extended *HMM*. Both annotations are shown in the Figures 6.3 and 6.9.

The presented examples generalise the differences in over-expressed annotations between the standard *HMM* and the extended *HMM*.

Let us consider which genes have been annotated as over-expressed by the extended *HMM*.

The Table 5.2 represents genes on chromosome 17 which have been found as over-expressed by Pollack *et al.* [23]. All these genes, excepted *TBX2*, have been annotated as over-expressed by the extended *HMM*. *TBX2* has been found by the standard *HMM* as we can see in the Table 6.1, but the attribute values *Max*, *Min* and *Mean* are not very significant. The overview of the known over-expressed candidate genes which have also been annotated as over-expressed by the extended *HMM* is shown in the Table 6.8. The gene *ABCA5* should be emphasised because the standard *HMM* has done an annotation error as we can see in the Table 6.1, but the extended *HMM* has annotated this gene as over-expressed with significant attribute values. For the genes *SDF2*, *TRAF4*, *FLJ10700*, *TIAF1*, *PIP5K2B*, *ERBB2*, *GBR7*, *KRT20*, *KRTHA4*, *HOXB5*, *NDP52*, *RAD51C*, *SLC9A3R1*, *AD023*, *GRB2* and *ITGB4* nothing changes in the annotation in comparison of the standard *HMM* with the extended *HMM*. All the other genes of the Table 6.8 have improved annotation attributes *Max*, *Min* and *Mean*. This observation confirms the finding that the extended *HMM* annotates in general more significant log ratios as over-expressed. Fifty percent of all two hundred sixty-five genes in our breast cancer data set have been annotated as over-expressed over all gene expression profiles. The genes *MLN64*, *GRB7*, *NR1D1*, *CDC6*, *TOP2A* and *SMARCE1* are located in direct adjacence to the gene *ERBB2* and these genes show more significant attributes *Nr*, *Max*, *Min* and *Mean* in comparison with the standard *HMM*. Pollack *et al.* [23] have found that the genes *MLN64* and *GRB7* are always co-amplified with *ERBB2*. They have reported the possible role of these genes in the pathogenesis and that these genes could be potential useful targets in the treatment of *ERBB2*-positive tumours. The genes *MLLT6*, *ZNF144*, *PIP5K2B* and *TOP2A* have been seen amplified and highly expressed in a subset of the *ERBB2*-coamplified tumours. These genes may contribute to the specific phenotypic features of *ERBB2*-positive tumours. The findings of Pollack *et al.* [23] are consistent with the work of Kauraniemi *et al.* [13], and our annotations in the Table 6.8 match good with both publications.

The Table 6.9 contains additional genes which have been annotated as over-expressed by the extended *HMM*. We have used the same threshold values for the p-arm and the q-arm as in the Table 6.2 which has been created by the standard *HMM*. The Table 6.9 presents less genes in comparison with the Table 6.2. The cause of this is that the extended *HMM*, as described above, needs more significant values for such annotations. All the genes in the Table 6.9 are also contained in the Table 6.2. In more detail, the annotations of the genes *SCYA18*, *PNMT*, *HOXB6* and *CLTC* have not changed when we have used the extended *HMM* in comparison with the annotation results of the standard *HMM*. The other genes *MFPA4*, *SCYA14*, *SCYA3L1* and *KRT17* show better attribute values *Max*, *Min* and *Mean* for the over-expressed annotations. The rest of the genes of the Table 6.2 which are not contained in the Table 6.9 have also been annotated as over-expressed by the extended *HMM*, but these genes are not shown here because these genes have too low values for the attribute *Nr*. The functions of some additional candidate genes for over-expression are explained above where we have analysed the gene expression profiles with the standard *HMM*. Recall that *HOXB6* is known to play a role in acute myeloid leukaemia and colorectal cancer, and that *KRT17* has functions in the outcome of breast cancer [26].

Under-expression

Now we consider which regions on chromosome 17 are mainly affected by under-expression. The Table 6.13 represents an overview of genes which have been annotated as under-expressed by the extended *HMM* over all gene expression profiles. One hundred sixty-seven genes over all profiles have been annotated as under-expressed by the extended *HMM*. The analysis of these genes has identified seventy-nine different genes in the annotation process. The standard *HMM* has annotated two hundred eighteen genes as under-expressed over all profiles and under these genes are ninety-two different genes. The trend that the extended *HMM* requires more significant log

Gene	Band	Nr	Max	Min	Mean	Gene	Band	Nr	Max	Min	Mean
KIAA0524	q11.2	2	2.11	1.71	1.91	KRT19	q21.2	3	3.29	2.71	3
UNC119	q11.2	2	2.9	2.75	2.82	JUP	q21.2	2	2.77	2.28	2.52
SDF2	q11.2	2	3.16	2.49	2.83	HOXB5	q21.32	5	4.28	1.98	3.06
TRAF4	q11.2	3	3.41	2.63	2.91	NDP52	q21.32	2	1.81	1.61	1.71
FLJ10700	q11.2	3	5.7	1.98	3.91	NGFR	q21.33	1	0.88	0.88	0.88
TIAF1	q11.2	3	4.34	1.88	2.99	HBOA	q21.33	2	3.97	2.53	3.25
KIAA1321	q11.2	2	3.61	1.96	2.78	DLX4	q21.33	5	4.74	2.48	3.45
SCYA3	q12	4	2.64	1.69	2.15	ABCC3	q21.33	4	5.01	3.09	3.83
SCYA4	q12	4	2.6	1.13	1.68	RAD51C	q23.2	2	2.55	2.43	2.49
MLLT6	q12	4	4.89	0.34	3.01	RPS6KB1	q23.2	5	2.98	1.81	2.4
ZNF144	q12	4	3.67	1.86	2.9	APPBP2	q23.2	4	2.18	1.45	1.76
PIP5K2B	q12	4	4.34	2.94	3.58	PPM1D	q23.2	3	2.83	2.01	2.32
CACNB1	q12	4	3.85	1.04	2.61	TRAP240	q23.2	1	2.86	2.86	2.86
RPL19	q12	4	2.06	0.21	0.79	ICAM2	q23.3	1	2	2	2
MLN64	q12	9	4.9	2.62	3.8	PECAM1	q23.3	1	2.4	2.4	2.4
ERBB2	q12	10	5.07	2.4	3.64	ABCA5	q24.2	1	2.7	2.7	2.7
GRB7	q12	10	6.16	2.43	4.22	SLC9A3R1	q25.1	2	4	2.81	3.4
NR1D1	q21.1	4	2.28	-0.29	1.04	AD023	q25.1	2	2.78	2.5	2.64
CDC6	q21.2	3	4.54	2.78	3.48	GRB2	q25.1	3	2.32	1.73	2.1
TOP2A	q21.2	3	5.18	1.45	2.85	ITGB4	q25.1	5	5.18	0.17	2.72
SMARCE1	q21.2	5	6.48	1.74	4.66	HCNGP	q25.1	3	2.32	1.44	1.83
KRT20	q21.2	2	1.74	1.1	1.42	BIRC5	q25.3	1	2.64	2.64	2.64
KRTHA4	q21.2	2	1.44	0.32	0.88	LGALS3BP	q25.3	1	2.93	2.93	2.93

Table 6.8: Over-expressed candidate genes from Table 5.2 which have been annotated as over-expressed by the *HMM* with two transition classes. Max represents the highest log ratio which was measured for the Nr experiments where the gene has been annotated as over-expressed. Min represents the lowest log ratio which was measured for the Nr experiments where the gene has been annotated as over-expressed. Mean is the mean log ratio of the measured log ratios for the Nr experiments where the gene has been annotated as over-expressed.

Gene	Band	Nr	Max	Min	Mean	Gene	Band	Nr	Max	Min	Mean
MFAP4	p11.2	3	4.83	3.43	4.17	PNMT	q12	9	0.7	-0.52	0.01
SCYA14	q12	4	5.28	2.34	3.8	KRT17	q21.2	6	5	2.26	3.39
SCYA18	q12	4	3.95	1.8	2.94	HOXB6	q21.32	5	5.23	1.93	3.58
SCYA3L1	q12	4	2.46	1.46	1.86	CLTC	q23.2	4	5.04	0.37	2.56
NAP4	q12	4	5.12	2.5	4						

Table 6.9: Additional genes on chromosome 17 which have been annotated as over-expressed by the *HMM* with two transition classes. For the p-arm a threshold value of Nr ≥ 3 has been used. The threshold value for the q-arm has been set to Nr ≥ 4. Max represents the highest log ratio which was measured for the Nr experiments where the gene has been annotated as over-expressed. Min represents the lowest log ratio which was measured for the Nr experiments where the gene has been annotated as over-expressed. Mean is the mean log ratio of the measured log ratios for the Nr experiments where the gene has been annotated as over-expressed.

ratios as the standard *HMM* to annotate genes as over-expressed or under-expressed is also discerned here on the basis of these numbers. Let us first look at the p-arm of chromosome 17 which has in total nineteen genes which have been annotated as under-expressed over all profiles. The extended *HMM* has annotated the bands 17p13.3, 17p13.2 and 17p13.1 in the same manner as the standard *HMM*. The chromosomal band 17p12 contains two new under-expressed annotations with significant log ratios in the primary breast cancer tumour *STANFORD 17*. The region 17p11.2 includes three changes. That is, the profiles *STANFORD23* and *STANFORD 24* do not show genes which have been annotated as under-expressed in this band and the profile *NORWAY 18* shows a new annotation in this band. The inspection of the profiles *STANFORD 23* and *STANFORD 24* let us assume that the log ratios of the affected genes are marginal cases, because the state posterior $sp(-)$ for the state $-$ shows a good visible peak. The profile *NORWAY 18* has been annotated with a better quality by the extended *HMM* in comparison with the annotation results of the standard *HMM*. The gene which has now been annotated as under-expressed has a log ratio which is significantly less than zero. The standard *HMM* has annotated this gene as over-expressed because this gene is located in front of an over-expressed segment. As we have mentioned several times the standard *HMM* has problems to model the start or the end of segments. The extended *HMM* shows in general much better annotations in such cases.

Let us look at the under-expressed annotations over all gene expression profiles for the q-arm of chromosome 17. The q-arm contains one hundred forty-eight genes which have been annotated as under-expressed. The bands 17q24.1, 17q24.2 and 17q24.3 have been annotated in the same manner by the standard and the extended *HMM*. The extended *HMM* has annotated less genes as under-expressed in the chromosomal regions 17q11.2, 17q12, 17q21.1, 17q21.31, 17q21.32, 17q21.33, 17q23.2, 17q23.3, 17q25.1 and 17q25.3 in comparison with the annotation results of the standard *HMM*. The extended *HMM* has annotated one gene more as under-expressed in the band 17q21.2 as the standard *HMM* has done. The distribution of annotations in this region includes more profiles in the annotation results of the extended *HMM* as in the annotation results of the standard *HMM*. Mainly affected by under-expression are the bands 17q11.2, 17q12, 17q21.2, 17q21.32, 17q21.33 and 17q25.3 which all show more than ten annotations. As we have reported for the annotation results of the standard *HMM* the bands 17q12 and 17q21.2 are also the regions with the most annotations in the annotation results of the extended *HMM*. Orsetti *et al.* 1999 [20] found that the regions 17q11-21 and 17q25 can contain losses of DNA segments. In the year 2004 Orsetti *et al.* [21] refined these results and reported that the regions 17p, 17q11.2 and 17q21 are mainly affected by DNA losses and that the bands 17q21.3, 17q22 and 17q25 can contain losses or gains of DNA segments. Our annotation results which have been created by the extended *HMM* with two transition classes match good with the results of Orsetti *et al.* [21]. We have explained what processes could lead to low expression levels as we have discussed the annotation results of the standard *HMM*.

In the next step of our analysis we consider which genes are mostly affected by under-expression. Pollack *et al.* [23] have not explicitly mentioned under-expressed genes. Orsetti *et al.* [21] have found under-expression in the bands 17q11 and 17q21, but they have used other cell lines and therefore the results are not directly comparable with our results. To get an impression which of the known candidate genes for over-expression of the Table 5.2 have also been annotated as under-expressed by the extended *HMM* we have compared the seventy-nine under-expressed candidates with the genes in the Table 5.2. The results are shown in the Table 6.10. All genes in this table have significant attributes *Max* and *Min* and that is what has conducted to the under-expressed annotation. Let us recall the annotation results of the standard *HMM* which are shown in the Table 6.3. We have compared these results with the results of the extended *HMM* and we have found that the genes *KIAA0524*, *NR1D1*, *JUP* and *HBOA* have never been annotated as under-expressed by the extended *HMM*. The genes *NR1D1*, *JUP* and *HBOA* do not have significant log ratios for such annotations and the log ratio of *KIAA0524* is a marginal case. The extended *HMM* has annotated the over-expressed candidate gene *RAD51C* as under-expressed in contrast to the standard *HMM* and this annotation is significant as the attributes *Max*, *Min* and *Mean* show. In summary, the over-expressed candidate genes *KIAA0524*, *UNC119*, *SDF2*, *TIAF1*, *KIAA1321*, *MLLT6*, *ZNF144*, *PIP5K2B*, *CACNB1*, *RPL19*, *NR1D1*, *KRT20*, *KRTHA4*, *JUP*, *NDP52*, *HBOA*, *RPS6KB1*, *APPBP2*, *PPM1D*, *TBX2*, *TRAP240*, *ABCA5*, *AD023* and *GRB2* have never been annotated as under-expressed by the extended *HMM*.

Gene	Band	Nr	Max	Min	Mean	Gene	Band	Nr	Max	Min	Mean
TRAF4	q11.2	1	-1.72	-1.72	-1.72	NGFR	q21.33	2	-2.53	-2.7	-2.61
FLJ10700	q11.2	1	-2.53	-2.53	-2.53	DLX4	q21.33	2	-2.68	-3.15	-2.91
SCYA3	q12	3	-1.61	-2.82	-2.16	ABCC3	q21.33	1	-2.48	-2.48	-2.48
SCYA4	q12	2	-1.59	-1.76	-1.68	RAD51C	q23.2	1	-2.36	-2.36	-2.36
MLN64	q12	1	-3.1	-3.1	-3.1	ICAM2	q23.3	2	-2.02	-2.45	-2.24
ERBB2	q12	6	-2.07	-3.13	-2.45	PECAM1	q23.3	3	-3.31	-4.5	-3.9
GRB7	q12	6	-1.7	-2.42	-2.12	SLC9A3R1	q25.1	3	-2.27	-3.76	-2.83
CDC6	q21.2	2	-2.27	-2.82	-2.55	ITGB4	q25.1	1	-2.69	-2.69	-2.69
TOP2A	q21.2	2	-1.86	-2.93	-2.4	HCNGP	q25.1	1	-2.66	-2.66	-2.66
SMARCE1	q21.2	1	-2.29	-2.29	-2.29	BIRC5	q25.3	1	-2.59	-2.59	-2.59
KRT19	q21.2	4	-1.88	-3.36	-2.71	LGALS3BP	q25.3	6	-1.92	-3.43	-2.65
HOXB5	q21.32	2	-2.63	-2.78	-2.7						

Table 6.10: Over-expressed candidate genes from Table 5.2 which have also been annotated as under-expressed by the *HMM* with two transition classes. Max represents the highest log ratio which was measured for the Nr experiments where the gene has been annotated as under-expressed. Min represents the lowest log ratio which was measured for the Nr experiments where the gene has been annotated as under-expressed. Mean is the mean log ratio of the measured log ratios for the Nr experiments where the gene has been annotated as under-expressed.

Gene	Band	Nr	Max	Min	Mean	Gene	Band	Nr	Max	Min	Mean
KPNA2	p13.3	2	-2.09	-2.58	-2.33	SCYA18	q12	4	-1.69	-2.85	-2.43
RAB5EP	p13.2	3	-2.39	-2.64	-2.54	SCYA3L1	q12	3	-2.18	-4.3	-3.09
CLDN7	p13.1	2	-2.6	-2.71	-2.65	KRT13	q21.2	3	-1.76	-2.52	-2.16
PMP22	p12	2	-3.81	-3.92	-3.86	KRT17	q21.2	8	-2.18	-4.29	-2.9
LGALS9	q11.2	4	-2.38	-4.23	-3.12	SKAP55	q21.32	3	-2.91	-3.15	-3.03
EVI2B	q11.2	3	-1.68	-2.73	-2.11	COL1A1	q21.33	8	-2.06	-4.63	-3.33
EVI2A	q11.2	3	-2.38	-3.36	-2.89	SOX9	q24.3	4	-2.46	-3.6	-2.94
SCYA2	q12	5	-1.55	-4.04	-2.65	TIMP2	q25.3	3	-2.01	-2.65	-2.4
SCYA14	q12	5	-2.08	-4.41	-2.85						

Table 6.11: Additional genes on chromosome 17 which have been annotated as under-expressed by the *HMM* with two transition classes. For the p-arm a threshold value of Nr ≥ 2 has been used. The threshold value for the q-arm has been set to Nr ≥ 3. Max represents the highest log ratio which was measured for the Nr experiments where the gene has been annotated as under-expressed. Min represents the lowest log ratio which was measured for the Nr experiments where the gene has been annotated as under-expressed. Mean is the mean log ratio of the measured log ratios for the Nr experiments where the gene has been annotated as under-expressed.

The Table 6.11 contains the under-expressed candidate genes which have frequently been annotated as under-expressed over all gene expression profiles. To get general candidate genes for under-expression we have used the same threshold values for the p-arm and the q-arm of chromosome 17 as in the Table 6.4 which was generated by the standard *HMM*. All genes of the Table 6.11 are included in the Table 6.4, excepted *KRT13*. The genes *MFAP4*, *ALDH10*, *SCYA7*, *SCYA11*, *SCYA13*, *FLJ22041*, *COPZ2* and *CD79B* have also been annotated as under-expressed by the extended *HMM*, but the values of the attribute *Nr* for these genes are too low to be contained in the Table 6.11. The reason for this behaviour is that the extended *HMM* has annotated in general more significant log ratios as over-expressed or under-expressed. We have reported this above several times. For some genes in the Table 6.11 we have presented their functions as we have analysed the annotation results of the standard *HMM*. Here we summarise only the main results which show that the extended *HMM* is able to detect known candidate genes for under-expression.

- Kominsky *et al.* [15] have found that the loss of *CLDN7* correlates with the histological grade of in situ and invasive ductal carcinomas of the breast. Invasive ductal carcinomas of the breast show lower expression levels of *CLDN7* than normal breast cells and so *CLDN7* has a potential role in the progression and ability of breast cancer cells to disseminate.

- Irie *et al.* [10] have reported that *LGALS9* is a prognostic factor for antimetastatic potential in breast cancer. Cells with a low expression level of *LGALS9* do not show tight clusters during the in vitro proliferation.

- Nakopoulou *et al.* [18] have observed that *TIMP2* is involved in the maintenance of a tissue and that the under-expression of *TIMP2* leads to the degradation of the extracellular matrix. The degradation of the extracellular matrix is followed by the invasion of cancer into the surrounding matrix. Nakopoulou *et al.* [18] have studied this behaviour on breast cancer cells.

Selected gene expression profiles

The Figures 6.7, 6.8, 6.9, 6.10 and 6.11 show selected gene expression profiles which have been annotated by the extended *HMM*. The annotations for the same profiles have also been created by the standard *HMM*. These annotations are presented in the Figures 6.1, 6.2, 6.3, 6.4 and 6.5. We have described these annotations in detail as we have analysed the performance of the standard *HMM* approach. Now we will compare the results of both *HMM*s.

- The Figures 6.7 and 6.1 represent the annotations of the breast cancer cell lines *BT474* and *SKBR3*. First we consider the *BT474* profile. The extended *HMM* has not done an under-expressed annotation in the region 17p13.1 as the standard *HMM* has done, but both state posterior profiles $sp(-)$ of the state $-$ have a peak in this region. The standard *HMM* has not been able to model a good over-expressed segment structure in the regions 17q12-21.2 and 17q21.32-22. That is, we can see log ratios which are significantly less than zero with over-expressed annotations. The extended *HMM* has improved the segment structure in these regions by dividing the over-expressed segments into over-expressed, under-expressed and identically expressed segments and therefore no genes with log ratios which are significantly less than zero have been annotated as over-expressed. This altered segmentation can also be seen in the state posterior profiles of the extended *HMM*.
  The breast cancer cell line *SKBR3* has been annotated in the same manner by both *HMM*s excepted a segment in the band 17q21.2 which has been annotated as over-expressed by the standard and as identically expressed by the extended *HMM*.

- The Figures 6.8 and 6.2 contain the primary breast tumour profiles *NORWAY 14* and *NORWAY 26*. Both *HMM* approaches show a good annotation quality and therefore the profile *NORWAY 14* contains only a little change in the under-expressed segment which is located in the chromosomal band 17q25.3. The standard *HMM* has annotated two genes

as under-expressed and the extended *HMM* has only annotated the one with the smaller log ratio as under-expressed.
The profile *NORWAY 26* has been annotated in the same manner by both *HMM*s.

- The Figures 6.9 and 6.3 show the profiles of the primary breast tumours *NORWAY 47* and *NORWAY 53*. The standard *HMM* has had problems to achieve a good over-expressed segment structure in the bands 17q12 and 17q25 of the profile *NORWAY 47* and as we can see the extended *HMM* has been able to annotated these regions much better.
  When we consider the annotations of the primary tumour *NORWAY 53* it is clearly to see that the annotation of the extended *HMM* is of better quality in the regions 17p11.2-q11.2 and 17q12-21.2 in comparison with the annotation of the standard *HMM*.

- The profiles *NORWAY 100* and *STANFORD 2* are shown in the Figures 6.10 and 6.4. The over-expressed segments in the profile *NORWAY 100* have been annotated in the same manner by both *HMM*s. The under-expressed segment in the region 17q12-21.2 looks better in the annotation of the extended *HMM*, because log ratios which are close to zero in this segment have been annotated as identically expressed.
  The standard *HMM* has had problems in the annotation of the over-expressed segments in the regions 17q11.2 and 17q12-21.2 of the profile *STANFORD 2*, but the extended *HMM* has been able to improve the annotation performance in these regions.

- The Figures 6.11 and 6.5 show the profiles *STANFORD 24* and *STANFORD A* which have nearly the same annotation quality for both *HMM* approaches.

The comparison of the annotations of the gene expression profiles confirms the fact that the extended *HMM* with two transition classes is able to model the start and the end of a segment in a profile much better as the standard *HMM*. The analysis of gene expression profiles with *HMM*s in consideration of proximity effects between genes represents a good strategy to find candidate genes for under-expression or over-expression.

Summary of the annotation process

We finally consider the summary of the whole annotation process in the Figure 6.12 which contains the subfigures *Unchanged Segments*, *Over-expressed Segments*, *Under-expressed Segments* and *Gene Counts*. The data which is used to create this summary has been created by the extended *HMM*.
The subfigure *Unchanged Segments* gives an overview of the locations and absolute frequencies of segments which have been annotated as identically expressed in the gene expression profiles of our breast cancer data set. The distribution of the hexagons is consistent with the inhomogeneity of the gene expression profiles in our data set. The segments with identical expression levels between tumour and normal tissue are longer than over-expressed or under-expressed segments. The dark hexagons are an evidence that some gene expression profiles could show the same subtype of breast cancer.
The subfigure *Over-expressed Segments* represents where and how frequently over-expressed segments have been observed in the breast cancer data set. This subfigure has improved quality in comparison with the subfigure in the Figure 6.6 for the annotation results of the standard *HMM*. The improved quality is the result of the observation that the extended *HMM* can model the segment structure much better. The dark hexagons show segments which have often been annotated as over-expressed and therefore we should find candidate genes for some breast cancer subtypes in such regions.
The overview of the locations and the absolute frequencies of under-expressed segments is shown in the subfigure *Under-expressed Segments*. The lengths of under-expressed segments are smaller

than the lengths of over-expressed segments. We have also observed this in the Figures 6.7, 6.8, 6.9, 6.10 and 6.11 of the gene expression profiles. It should be possible to find general candidate genes for under-expression in the dark hexagons.

The subfigure *Gene Counts* represents the absolute frequencies of over-expressed and under-expressed annotations on chromosome 17. The frequencies of under-expressed annotations are shown in the upper subfigure and the frequencies of over-expressed annotations are presented in the subfigure below. As we can see genes which have frequently been annotated as over-expressed have also been annotated as under-expressed. We have also observed this behaviour for the subfigure in the Figure 6.6. The *Gene Counts* overview for the extended *HMM* is of better quality as the overview for the standard *HMM* because the extended *HMM* has improved performance to model segments. The region 17q23.2-23.3 has mainly been annotated as over-expressed and this observation could follow the results of Orsetti *et al.* [20], [21], Monni *et al.* [17], Nugoli *et al.* [19] and Clark *et al.* [3] which we have explained in detail in the Section 5.3.

	17p					17q																			
Experiment	13.3	13.2	13.1	12	11.2	11.1	11.2	12	21.1	21.2	21.31	21.32	21.33	22	23.1	23.2	23.3	24.1	24.2	24.3	25.1	25.2	25.3	$\sum$	E
BT474								9				4	1	2		5								21	1
MCF7																5								5	0
SKBR3	1							2	1	3														7	1
T47D																								0	0
NORWAY 7							2																	2	0
NORWAY 10																								0	0
NORWAY 11							1																	1	0
NORWAY 12											1													1	0
NORWAY 14							1	8		1	1	2												13	0
NORWAY 15	1																						1	2	0
NORWAY 16											1													1	0
NORWAY 17																								0	0
NORWAY 18				1	2							1												4	0
NORWAY 19		1																			2			3	0
NORWAY 26								12																12	1
NORWAY 27		1																						1	0
NORWAY 39								4																4	0
NORWAY 41								7		1	1										1			10	0
NORWAY 47							15	11	1	9										1	5			42	0
NORWAY 48								1				2												3	0
NORWAY 53					1		2	4	1	9													1	18	0
NORWAY 56													6	2		3			3		20			34	3
NORWAY 57								7					1								8			16	0
NORWAY 61					1			5	1	3						2	3							15	0
NORWAY 65					2			1		1		1												5	0
NORWAY 100								3		1		2												6	0
NORWAY 101								10					2											12	1
NORWAY 102										1														1	0
NORWAY 104													1			1		2						4	0
NORWAY 109										1		2	1								2			6	0
NORWAY 111				1																				1	0
NORWAY 112					2			1					1											4	0
STANFORD 2							5	4		1														10	1
STANFORD 14											1					1					1			3	0
STANFORD 16		2						4																6	0
STANFORD 17																			1					1	0
STANFORD 23							1																	1	0
STANFORD 24																								0	0
STANFORD 35								1					1											2	0
STANFORD 38	1						1					3												5	0
STANFORD A							10	4		2		5	8			3								32	1
$\sum$	3	4		2	8		38	98	4	33	5	22	22	4		20	3	2	4	1	39		2	314	9

Table 6.12: Overview of over-expressed genes. Candidate genes in a gene expression profile which have been annotated by the *HMM* with two transition classes have been assigned to their chromosomal bands. The results are contained in this table. $\sum$ is the row or the column sum of genes which have been annotated as over-expressed. E is the sum of genes in an experiment which have been annotated as over-expressed when their log ratios are less than zero.

			17p												17q										
Experiment	13.3	13.2	13.1	12	11.2	11.1	11.2	12	21.1	21.2	21.31	21.32	21.33	22	23.1	23.2	23.3	24.1	24.2	24.3	25.1	25.2	25.3	∑	E
BT474							2	1		1			1										2	7	0
MCF7							3	3					1				2							9	0
SKBR3				1			1	2					1				1							6	0
T47D				1				2					2											5	0
NORWAY 7										2		2	1											5	0
NORWAY 10																								0	0
NORWAY 11										1						1				1				3	0
NORWAY 12																								0	0
NORWAY 14											1												1	2	0
NORWAY 15		1						3				1												5	0
NORWAY 16										1							1							2	0
NORWAY 17										1			1								3			5	0
NORWAY 18				1	1							1	2							1				6	0
NORWAY 19								3																3	0
NORWAY 26																	1							1	0
NORWAY 27													1							1				2	0
NORWAY 39								4																4	0
NORWAY 41			1									2				1								4	0
NORWAY 47																								0	0
NORWAY 48							1					1	1								2			5	0
NORWAY 53								1																1	0
NORWAY 56												1	1											2	0
NORWAY 57		2								1														3	0
NORWAY 61										2														2	0
NORWAY 65	1							2		1											1			5	0
NORWAY 100								2		3						1								6	0
NORWAY 101										1		1												2	0
NORWAY 102								4			1	1												6	0
NORWAY 104		1	1				1				1	2								1			2	9	0
NORWAY 109								6			1										1			8	0
NORWAY 111			1		1			5		2							2						1	12	0
NORWAY 112							2	1																3	0
STANFORD 2		1								2														3	0
STANFORD 14																	1						2	3	0
STANFORD 16										1														1	0
STANFORD 17	1			2						2	1								1					7	0
STANFORD 23			2							1														3	0
STANFORD 24							3	3		1			1										2	10	0
STANFORD 35								1		1								1						3	0
STANFORD 38																							1	1	0
STANFORD A							1			2														3	0
∑	2	5	5	5	2		14	43		26	5	12	13			3	8	1	1	4	7		11	167	0

Table 6.13: Overview of under-expressed genes per chromosomal band. Candidate genes in a gene expression profile which have been annotated as under-expressed by the *HMM* with two transition classes have been assigned to their chromosomal bands. The results are contained in this table. $\sum$ is the row or the column sum of genes which have been annotated as under-expressed. E is the sum of genes in an experiment which have been annotated as under-expressed when their log ratios are greater than zero.

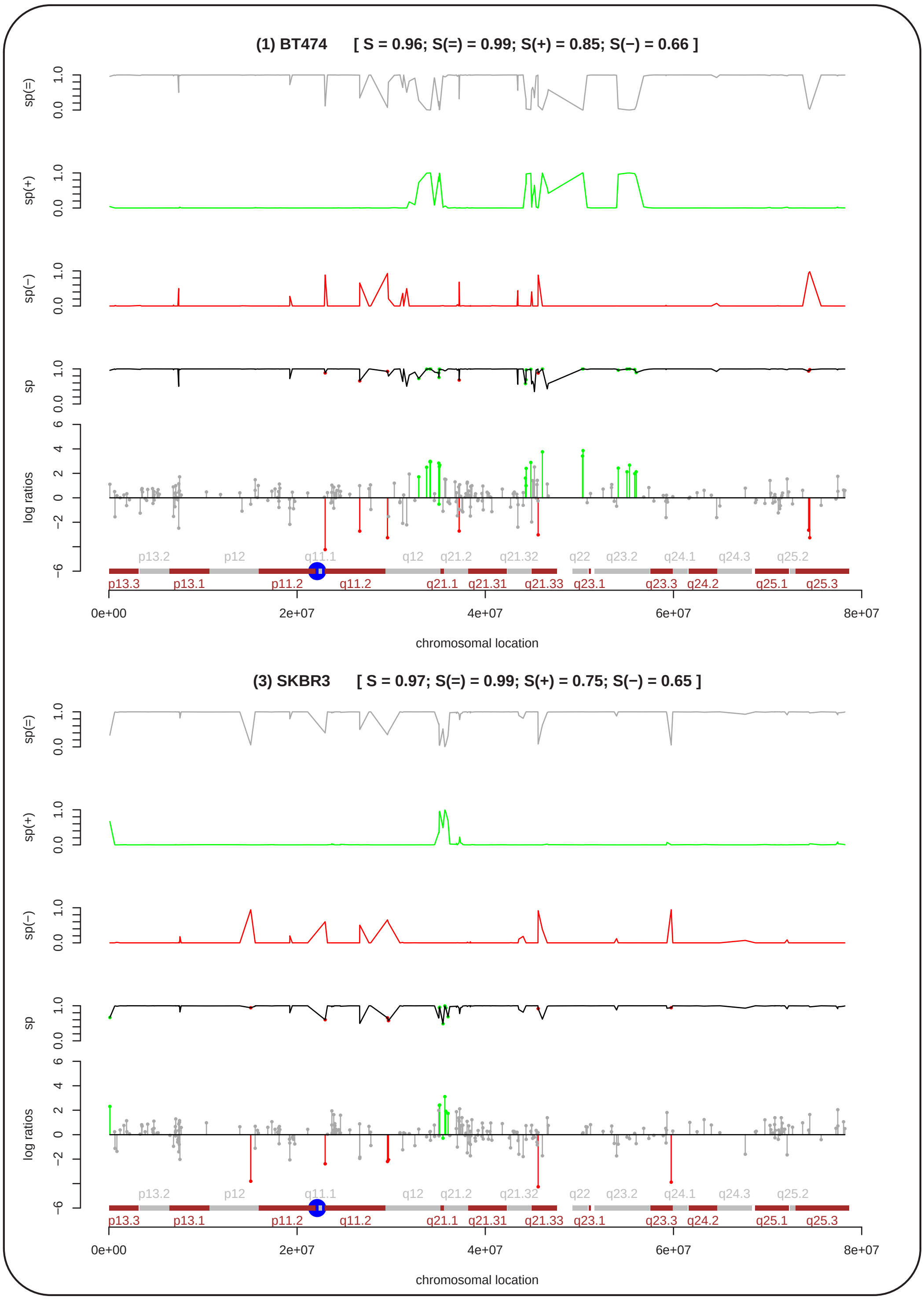


Figure 6.7: Gene expression profiles of chromosome 17 which have been annotated by the extended *HMM* with two transition classes. The first profile shows the *BT474* and the second the *SKBR3* data. In general, the headline of a profile contains the unique profile number (n) which is used in the Figure 5.3, the profile name, the sum S (5.3) of the state posteriors for the *Viterbi Path* and the relative proportion $S(i)$ (5.3), with $i \in \{-, =, +\}$, of the state posterior profile i in the *Viterbi Profile*. Top down the five graphics per experiment show the state posterior profiles for the states $=$, $+$ and $-$, the state posterior profile of the *Viterbi Path* and the annotated gene expression profile. The annotation of a gene is represented by the colour of the log ratio line. A red line represents an as under-expressed, a grey line an as identically expressed and a green line an as over-expressed annotated gene.

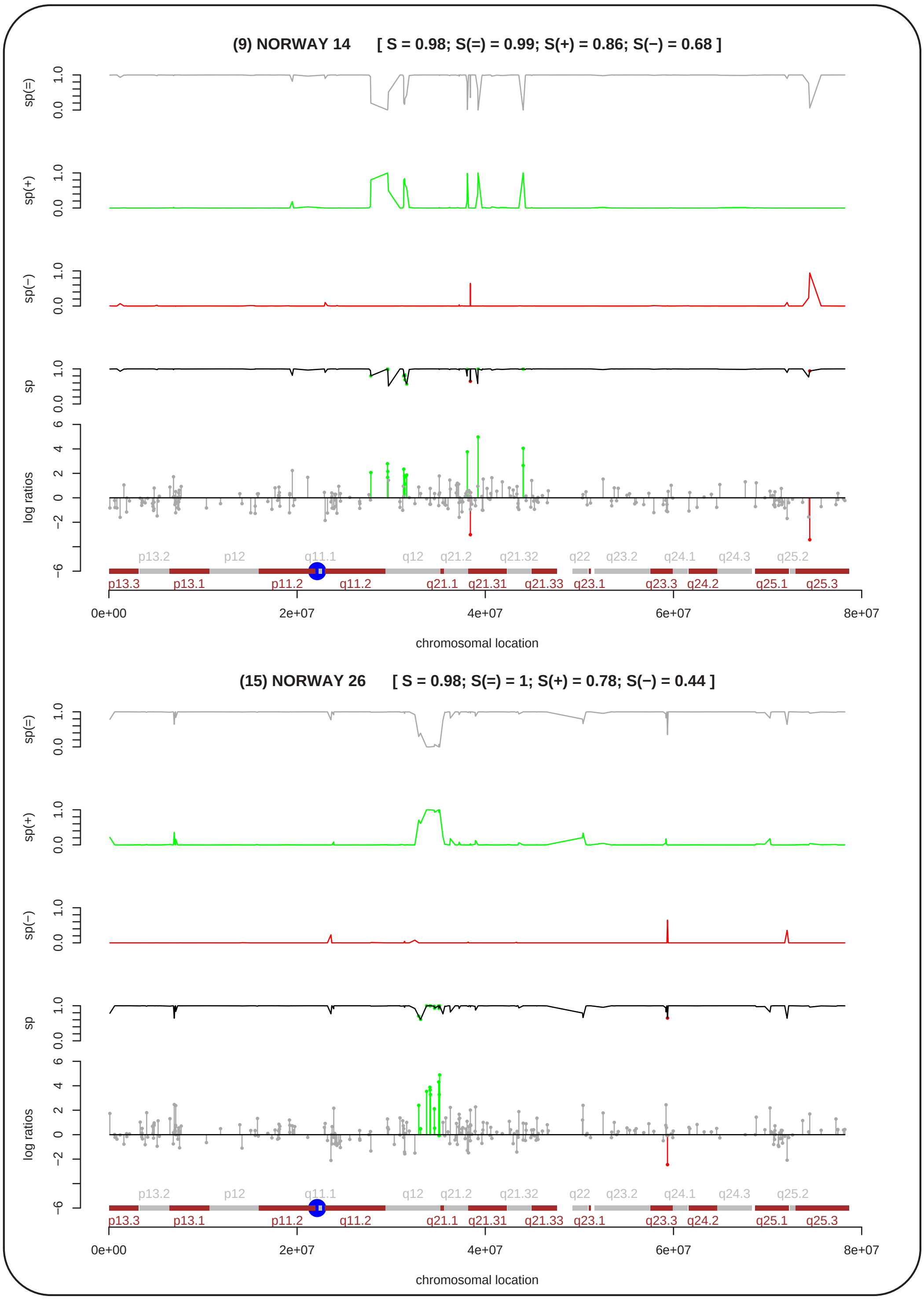
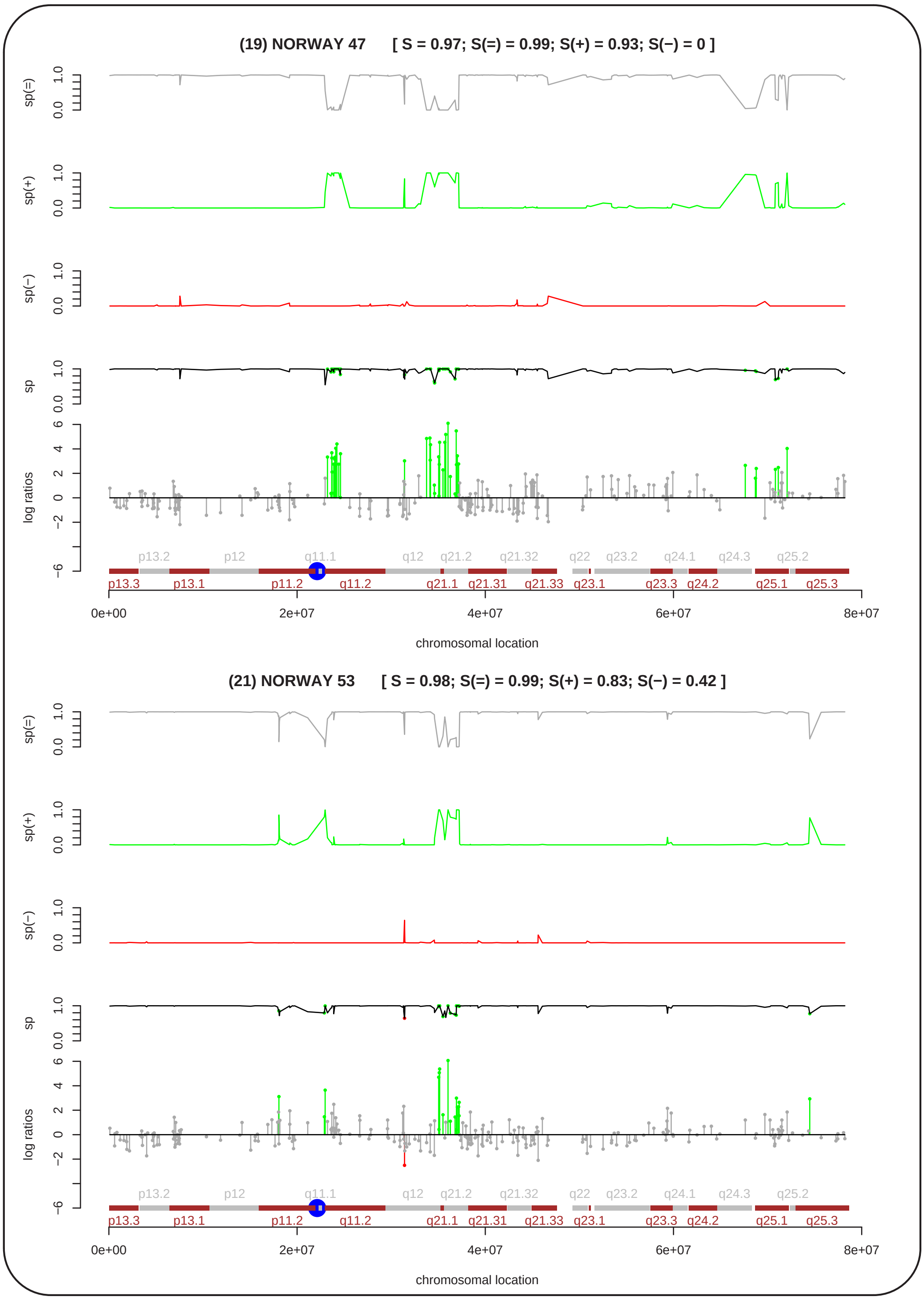


Figure 6.8: Gene expression profiles of chromosome 17 which have been annotated by the extended *HMM* with two transition classes. The first profile shows the *NORWAY 14* and the second the *NORWAY 26* data. In general, the headline of a profile contains the unique profile number (n) which is used in the Figure 5.3, the profile name, the sum S (5.3) of the state posteriors for the *Viterbi Path* and the relative proportion $S(i)$ (5.3), with $i \in \{-, =, +\}$, of the state posterior profile i in the *Viterbi Profile*. Top down the five graphics per experiment show the state posterior profiles for the states $=$, $+$ and $-$, the state posterior profile of the *Viterbi Path* and the annotated gene expression profile. The annotation of a gene is represented by the colour of the log ratio line. A red line represents an as under-expressed, a grey line an as identically expressed and a green line an as over-expressed annotated gene.



(19) NORWAY 47 [S = 0.97; S(=) = 0.99; S(+) = 0.93; S(−) = 0]

(21) NORWAY 53 [S = 0.98; S(=) = 0.99; S(+) = 0.83; S(−) = 0.42]

Figure 6.9: Gene expression profiles of chromosome 17 which have been annotated by the extended *HMM* with two transition classes. The first profile shows the *NORWAY 47* and the second the *NORWAY 53* data. In general, the headline of a profile contains the unique profile number (n) which is used in the Figure 5.3, the profile name, the sum S (5.3) of the state posteriors for the *Viterbi Path* and the relative proportion $S(i)$ (5.3), with $i \in \{-, =, +\}$, of the state posterior profile i in the *Viterbi Profile*. Top down the five graphics per experiment show the state posterior profiles for the states $=$, $+$ and $-$, the state posterior profile of the *Viterbi Path* and the annotated gene expression profile. The annotation of a gene is represented by the colour of the log ratio line. A red line represents an as under-expressed, a grey line an as identically expressed and a green line an as over-expressed annotated gene.

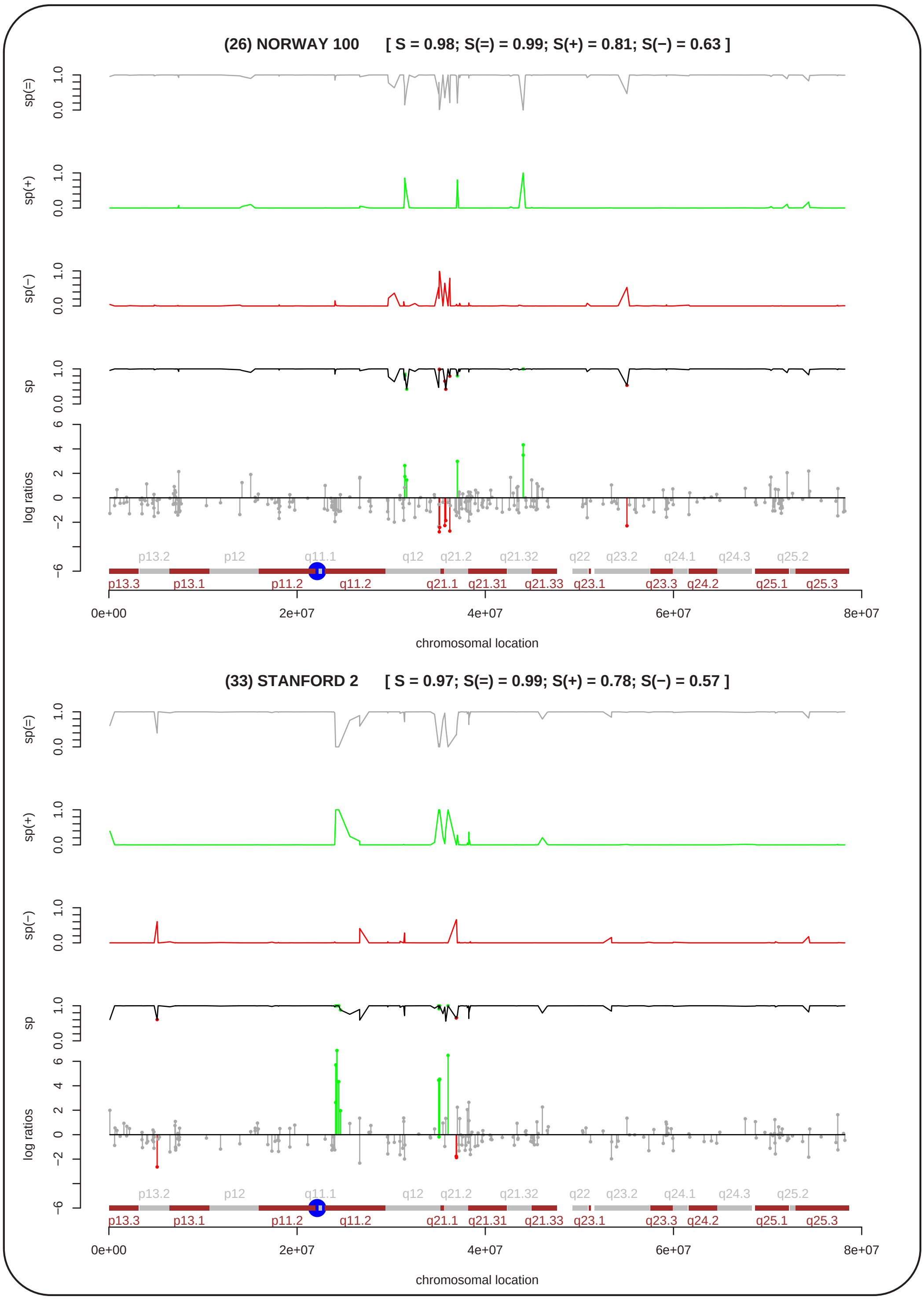


Figure 6.10: Gene expression profiles of chromosome 17 which have been annotated by the extended *HMM* with two transition classes. The first profile shows the *NORWAY 100* and the second the *STANFORD 2* data. In general, the headline of a profile contains the unique profile number (n) which is used in the Figure 5.3, the profile name, the sum S (5.3) of the state posteriors for the *Viterbi Path* and the relative proportion $S(i)$ (5.3), with $i \in \{-, =, +\}$, of the state posterior profile i in the *Viterbi Profile*. Top down the five graphics per experiment show the state posterior profiles for the states $=$, $+$ and $-$, the state posterior profile of the *Viterbi Path* and the annotated gene expression profile. The annotation of a gene is represented by the colour of the log ratio line. A red line represents an as under-expressed, a grey line an as identically expressed and a green line an as over-expressed annotated gene.

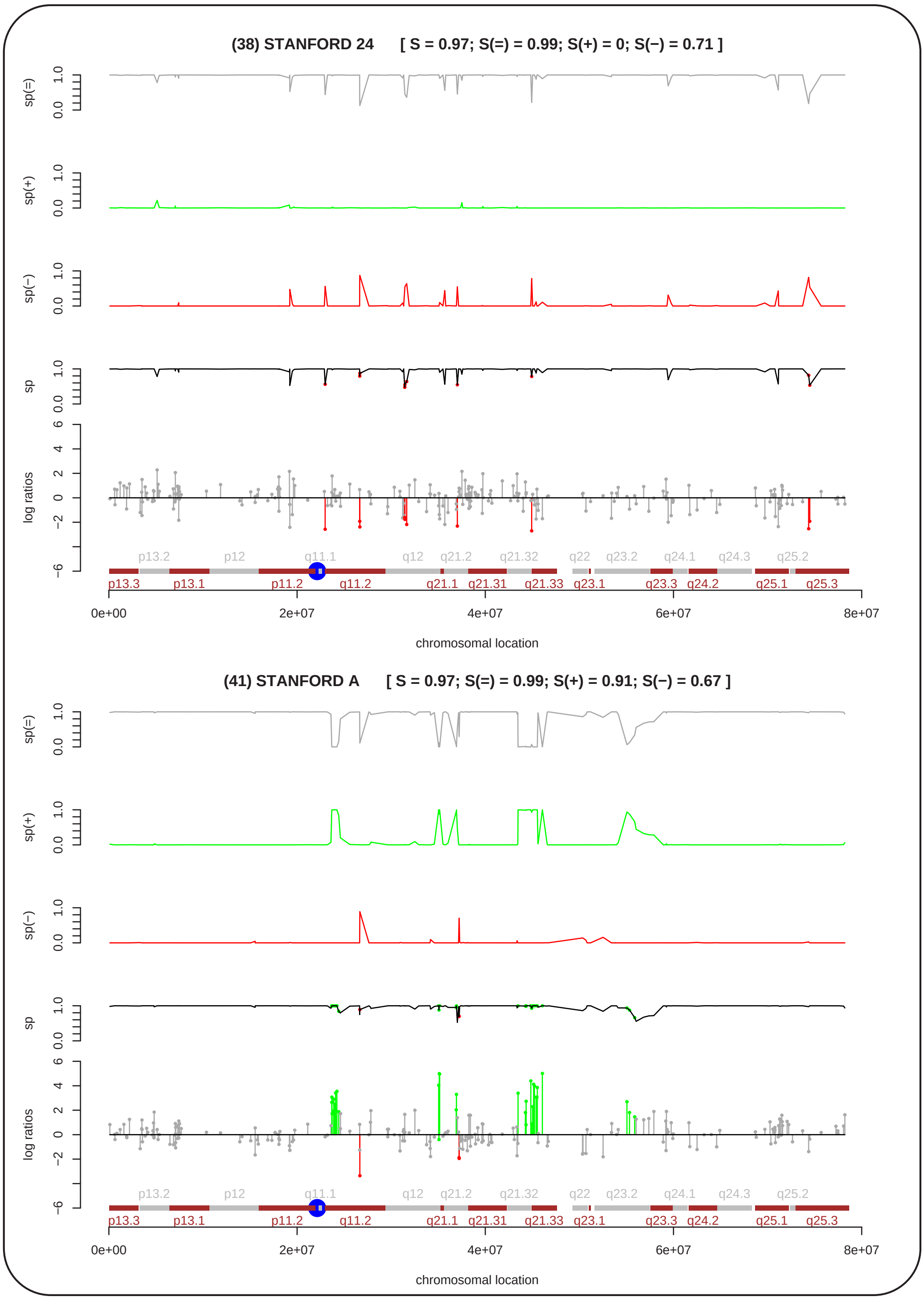
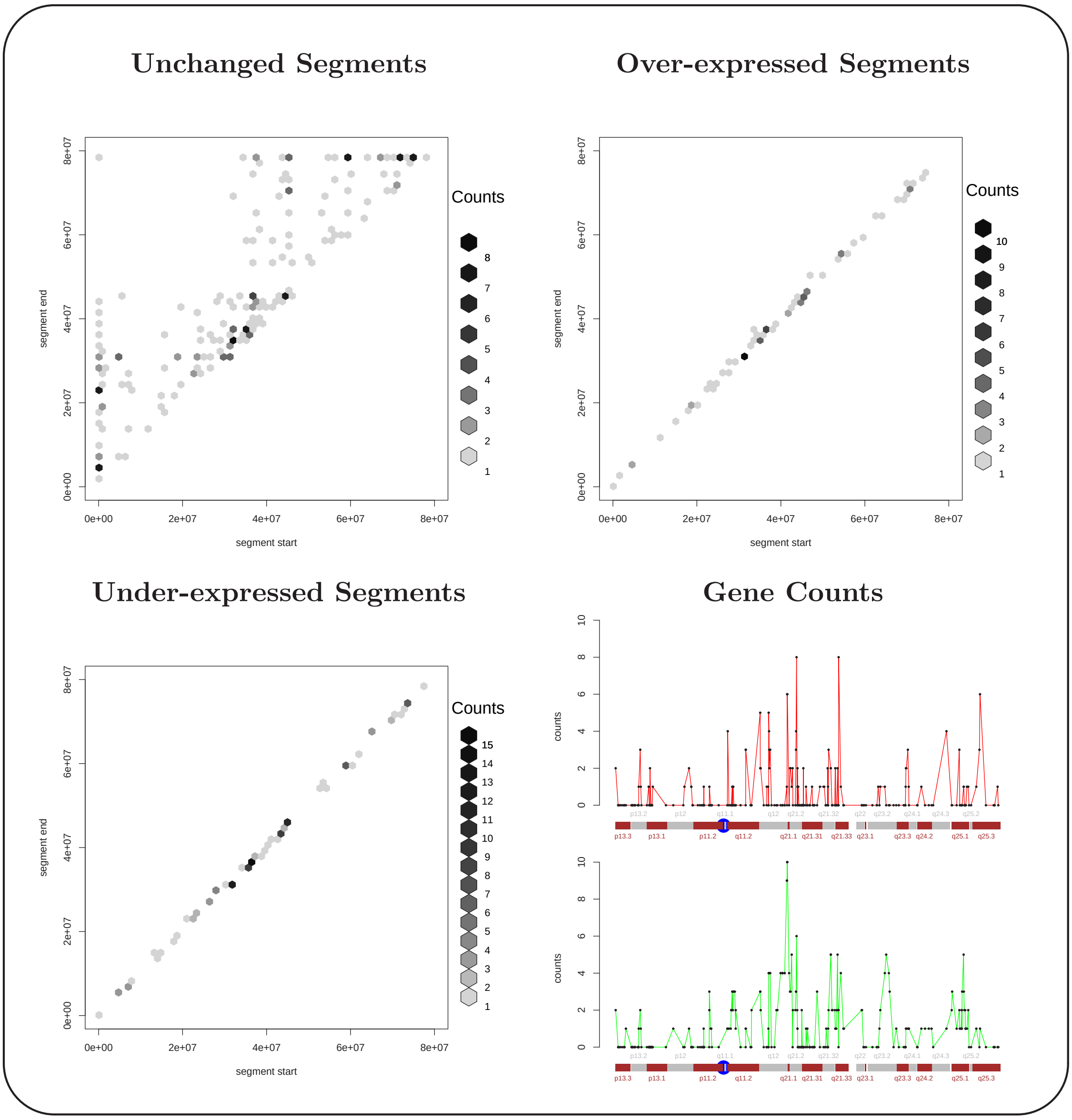


Figure 6.11: Gene expression profiles of chromosome 17 which have been annotated by the extended *HMM* with two transition classes. The first profile shows the *STANFORD 24* and the second the *STANFORD A* data. In general, the headline of a profile contains the unique profile number (n) which is used in the Figure 5.3, the profile name, the sum S (5.3) of the state posteriors for the *Viterbi Path* and the relative proportion $S(i)$ (5.3), with $i \in \{-, =, +\}$, of the state posterior profile i in the *Viterbi Profile*. Top down the five graphics per experiment show the state posterior profiles for the states $=$, $+$ and $-$, the state posterior profile of the *Viterbi Path* and the annotated gene expression profile. The annotation of a gene is represented by the colour of the log ratio line. A red line represents an as under-expressed, a grey line an as identically expressed and a green line an as over-expressed annotated gene.



Unchanged Segments

Over-expressed Segments

Under-expressed Segments

Gene Counts

Figure 6.12: Overview of annotations for the extended *HMM* with two transition classes. In general a segment is a sequence of successive genes which have been annotated in the same way. Each segment has a start and an end point on the chromosome. The figure **Unchanged Segments** gives a summary of segments on chromosome 17 which have been annotated as identically expressed. The figure **Over-expressed Segments** shows a summary of segments on chromosome 17 which have been annotated as over-expressed. The figure **Under-expressed Segments** represents a summary of segments on chromosome 17 which have been annotated as under-expressed. The figure **Gene Counts** represents the absolute frequency of over-expressed and under-expressed annotations per gene. The absolute frequencies of under-expressed annotations are shown in the upper subfigure and the absolute frequencies of over-expressed annotations are given in subfigure below.

6.2 Analysing *ArrayCGH* Profiles

Now we analyse the *ArrayCGH* profiles of our data set of chromosome 17. The analysis of the gene expression profiles have shown that the *HMM*s with coupled transition matrices outperform the standard *HMM*s in the modelling of the segmentation structure of these profiles. On the basis of this knowledge we will only consider the *HMM*s with coupled transition classes in this section.

6.2.1 *HMM*s With Transition Classes

The *HMM*s with transition classes are created as we have described in the Section 5.1. These *HMM*s extend the standard *HMM* approach to consider proximity effects of DNA segments and thus develop a more realistic model of the effects which can influence the alteration of DNA copy numbers on a chromosome. The positional proximity of genes on a chromosome can lead to characteristic gene expression patterns of these genes. These patterns can be observed for breast cancer cells when some DNA segments are lost or gained. It is realistic to assume that DNA segments which are located close to each other have a greater chance to show the same behaviour when a mutation occurs as DNA segments which have a greater distance to each other. Most of the copy number changes in breast cancer are the results of amplifications of DNA segments as the literature which we have reported at the beginning of the Section 5.3 shows. Nevertheless, there are also regions with deleted DNA segments known.

Modelling of chromosomal imbalances

A natural approach should be able to model gains and losses of genes respecting the proximity of adjacent genes. So we assume that the chance for two adjacent genes to be both in the same lost or gained segment increases the closer the distance between these two genes is and therefore the probability that these two adjacent genes show the same DNA copy number status should be higher. It is realistic to assume that the loss or the gain of a DNA segment does not affect both adjacent genes when the distance between these genes is greater as in the case before. The probability that both genes show the same DNA copy number status should be less as in the other case.
The proximity effects between adjacent genes are modelled by mapping the distance between these genes into a set of predefined transition matrices. We have explained this in detail in the Section 6.1.2 as we have motivated the usage of *HMM*s with transition classes to analyse gene expression data.

We have tested different *HMM*s with transition classes on the *ArrayCGH* data and so we have been able to conclude that the quality of the annotation results depends mainly on the used mixture model. The annotation results which we present have been created by an *HMM* with two transition classes (extended *HMM*). We have used a mixture model which consists of three normal distributions and we have also applied the same scaling vector $\vec{S} = (1, 2)$ for the transition classes and the same distance threshold value $\mathcal{T} = 30000$ base pairs for the transition class switching function as we have done in the analysis of the gene expression profiles with the extended *HMM*.

Gains of DNA segments

The Table 6.18 shows an overview of genes in all *ArrayCGH* profiles which have been annotated by the extended *HMM* to have an increased DNA copy number status. The segmentation of the genes to their bands allows us to find regions which have frequently been annotated to have an increased DNA copy number status. In general this table confirms that the *ArrayCGH* profiles of all experiments are more homogeneous as the gene expression profiles and this fact is consistent with the correlation matrices in the Figure 5.3 and the overviews of both data types in the Figures 5.4 and 5.5. All genes which have been annotated to have an increased DNA copy number status are located on the q-arm of chromosome 17 excepted one annotation on the p-arm in the chromosomal band 17p13.3 of the breast cancer cell line *MCF7*. The q-arm contains the ten regions 17q11.2, 17q12, 17q21.1, 17q21.2, 17q21.32, 17q21.33, 17q23.2, 17q23.3, 17q25.1 and 17q25.1 with more than twenty genes of increased DNA copy number status over all *ArrayCGH* profiles and thereby more than forty of these genes are located in each of the regions 17q11.2, 17q12, 17q23.2 and 17q25.1. The mostly affected chromosomal band over all *ArrayCGH* profiles is the band 17q12 with one hundred eleven genes which have been annotated to have an increased DNA copy number status. This general overview includes chromosomal regions which have been mentioned in literature. The band 17q12 has been observed as amplified by Kauraniemi *et al.* [13] and Willis *et al.* [27]. The region 17q21 has been seen as amplified by Hyman *et al.* [9], Clark *et al.* [3], Willis *et al.* [27] and Orsetti *et al.* [21]. The chromosomal band 17q23 has been observed as amplified by Orsetti *et al.* [20], Monni *et al.* [17], Clark *et al.* [3], Willis *et al.* [27] and Nugoli *et al.* [19]. Orsetti *et al.* [21] have found that the chromosomal region 17q25 can be amplified.

Now we consider which genes have been annotated to have an increased DNA copy number status. The Table 5.2 contains genes which are known to be over-expressed in some types of breast cancer. All these genes have been annotated to have an increased DNA copy number status. In total one hundred forty-seven genes of the two hundred sixty-five genes in the *ArrayCGH* data set have been annotated to have an increased DNA copy number status over all *ArrayCGH* profiles. That are fifty-five percent of the genes. The Table 6.14 represents an overview of the annotation attributes *Nr*, *Max*, *Min* and *Mean* of the candidate genes for over-expression of the Table 5.2. The definitions of these attributes are given in the caption of the Table 6.14. With the help of these attributes we have a good overview how significant the annotations are. In general, the attribute values of genes which have been annotated to have an increased DNA copy number status are significant enough to show such annotations. The genes *RPL19*, *NR1D1*, *CDC6* and *NDP52* have negative attribute values for the attribute *Min*. Such annotation errors occur when the affected genes are located in a region with increased DNA copy number which cannot be divided into a better annotation structure by the extended *HMM*. We have already discussed these problems as we have analysed the gene expression profiles. From time to time negative *Min* values have occurred and such cases have been counted in the attribute E of the Table 6.14.

The Table 6.15 contains genes that have been annotated to have an increased DNA copy number status in addition to the Table 5.2. To get a more general view on the additional candidates we use a threshold value which represents the minimal number of profiles where a gene must have been annotated with an increased DNA copy number status. The threshold value for the q-arm has been set to four. The p-arm has been excluded from this table because this arm contains only one annotation. We use the *Entrez Gene* at the NCBI to get a better impression what functions some of these candidate genes have. We have only found two interesting candidate genes which we have not mentioned in our previous analysis of the gene expression profiles.

- *SUPT6H* is a transcription elongation factor that enhances the rate of RNA polymerase II elongation.
- *FLOT2* encodes an integral membrane protein and is associated with melanoma progression.

The amplification of *SUPT6H* could play an important role for the gene expression rate of this gene and of other genes which are affected by an enhanced elongation in their transcription. The

Gene	Band	Nr	Max	Min	Mean	Gene	Band	Nr	Max	Min	Mean
KIAA0524	q11.2	2	1.34	1.08	1.21	JUP	q21.2	2	0.97	0.83	0.9
UNC119	q11.2	3	0.68	0.41	0.54	HOXB5	q21.32	3	2.23	0.77	1.26
SDF2	q11.2	4	1.4	0.49	1	NDP52	q21.32	3	1.96	-0.12	0.81
TRAF4	q11.2	5	2.64	0.53	1.19	NGFR	q21.33	3	2.06	1.83	1.96
FLJ10700	q11.2	5	2.79	0.49	1.22	HBOA	q21.33	3	1.55	0.75	1.15
TIAF1	q11.2	3	1.63	0.44	0.9	DLX4	q21.33	3	1.8	1.08	1.56
KIAA1321	q11.2	5	1.42	0.62	0.88	ABCC3	q21.33	3	1.88	0.75	1.43
SCYA3	q12	7	1.49	0.63	1.01	RAD51C	q23.2	4	2.48	0.7	1.35
SCYA4	q12	7	1.51	0.66	0.96	RPS6KB1	q23.2	5	2.65	0.52	1.38
MLLT6	q12	5	0.94	0.24	0.59	APPBP2	q23.2	4	2.34	0.82	1.38
ZNF144	q12	5	2.7	1.21	1.8	PPM1D	q23.2	4	3.41	1.17	1.98
PIP5K2B	q12	5	2.34	1.01	1.49	TBX2	q23.2	3	0.19	0.06	0.11
CACNB1	q12	7	3.16	1.07	2.12	TRAP240	q23.2	3	1.01	0.48	0.78
RPL19	q12	7	0.6	-0.07	0.34	ICAM2	q23.3	3	2.71	0.51	1.29
MLN64	q12	10	3.73	1.33	2.46	PECAM1	q23.3	4	2.8	0.62	1.51
ERBB2	q12	10	3.48	1.38	2.23	ABCA5	q24.2	1	0.88	0.88	0.88
GRB7	q12	10	2.73	0.1	1.22	SLC9A3R1	q25.1	1	1.21	1.21	1.21
NR1D1	q21.1	8	1.46	-0.07	0.95	AD023	q25.1	2	2.04	1.1	1.57
CDC6	q21.2	4	0.73	-0.17	0.41	GRB2	q25.1	2	2	1.02	1.51
TOP2A	q21.2	3	1.25	0.6	0.93	ITGB4	q25.1	2	0.95	0.83	0.89
SMARCE1	q21.2	3	1.18	0.19	0.55	HCNGP	q25.1	2	1.63	1.12	1.37
KRT20	q21.2	3	1.77	1.44	1.63	BIRC5	q25.3	3	1.02	0.62	0.88
KRTHA4	q21.2	2	1.63	1.11	1.37	LGALS3BP	q25.3	4	1.22	0.77	1.06
KRT19	q21.2	2	0.97	0.73	0.85						

Table 6.14: Over-expressed candidate genes from Table 5.2 which have been annotated by the *HMM* with two transition classes to have an increased DNA copy number status. Max represents the highest log ratio which was measured for the Nr experiments where the gene has been annotated to have an increased DNA copy number status. Min represents the lowest log ratio which was measured for the Nr experiments where the gene has been annotated to have an increased DNA copy number status. Mean is the mean log ratio of the measured log ratios for the Nr experiments where the gene has been annotated to have an increased DNA copy number status.

Gene	Band	Nr	Max	Min	Mean	Gene	Band	Nr	Max	Min	Mean
SUPT6H	q11.2	4	1.29	0.7	1.05	PNUTL2	q23.2	4	0.8	0.38	0.56
FLOT2	q11.2	5	1.89	0.6	1.09	CLTC	q23.2	5	1.49	0.53	0.89
CRYBA1	q11.2	5	0.82	0	0.43	DDX5	q24.1	4	0.44	-0.04	0.29
SCYA3L1	q12	7	1.57	0.65	1.14	SMURF2	q24.1	4	0.81	0.12	0.44
NAP4	q12	4	0.64	0.21	0.42	APOH	q24.2	4	0.89	0.7	0.79
PNMT	q12	10	0.29	-0.2	0.01	SRP68	q25.1	4	1.18	0.58	0.9
SFRS1	q23.2	4	0.61	0.25	0.4	TIMP2	q25.3	4	1.2	0.21	0.61
MPO	q23.2	4	1.21	0.45	0.75	GAA	q25.3	4	1.1	0.25	0.7

Table 6.15: Additional genes on chromosome 17 which have been annotated by the *HMM* with two transition classes to have an increased DNA copy number status. The threshold value for the q-arm has been set to Nr ≥ 4. Max represents the highest log ratio which was measured for the Nr experiments where the gene has been annotated to have an increased DNA copy number status. Min represents the lowest log ratio which was measured for the Nr experiments where the gene has been annotated to have an increased DNA copy number status. Mean is the mean log ratio of the measured log ratios for the Nr experiments where the gene has been annotated to have an increased DNA copy number status.

fact that we have observed *SUPT6H* as over-expressed in the annotation results of the standard and the extended *HMM* supports the possible role of this gene. *SUPT6H* has only been seen as over-expressed in three gene expression profiles and that is why this gene is not contained in the tables of the additional genes which have been annotated as over-expressed.

Losses of DNA segments

Let us now consider in which regions of chromosome 17 losses of DNA segments are located. The Table 6.19 represents an overview of genes over all *ArrayCGH* profiles which have been annotated to have a decreased DNA copy number status. The p-arm shows only nine genes which have been annotated to have a decreased DNA copy number status over all *ArrayCGH* profiles. The number of losses on the p-arm is greater than the number of gains which is one. Both chromosomal bands 17p13.2 and 17p11.2 contain one gene which has been annotated to have a decreased DNA copy number status over all profiles. The region 17p12 has three such annotations and the band 17p13.1 has four losses of DNA segments. When we consider the q-arm we can find thirty-nine genes which have been annotated to have a decreased DNA copy number status. The chromosomal bands 17q11.2 with twelve and 17q12 with nine annotations are mainly affected by losses of DNA segments. The regions 17q21.2, 17q21.31, 17q21.33, 17q22, 17q23.1, 17q23.2 and 17q25.1 contain only some genes which have been annotated to have a decreased DNA copy number status. Most of the annotations are in the breast cancer cell line *SKBR3* and only some annotations are contained in the primary tumours as *NORWAY 27*, *NORWAY 39*, *NORWAY 41* and *NORWAY 65*. It seems that losses of DNA segments do not play an important role in most of the primary breast tumours. The chromosomal bands which have been annotated to have a decreased DNA copy number status are also known in literature to be candidates for losses of DNA segments. Orsetti *et al.* 1999 [20] found that the regions 17q11-q21 and 17q25 can contain losses of DNA segments. A newer study of Orsetti *et al.* [21] refined these results and that is that the bands 17p, 17q11.2 and 17q21 are mainly affected by DNA losses and the bands 17q21.3, 17q22 and 17q25 can contain losses of DNA segments. In summary, the p-arm of chromosome 17 shows only one gain and nine losses of DNA segments and this is confirm with the results of Orsetti *et al.* [21] which mainly show losses on this arm. Now we consider which genes have been annotated to have a decreased DNA copy number status. First we compare the thirty-four different genes which show such an annotation with the over-expressed candidate genes in the Table 5.2 to get an impression which of this candidates can also be affected by losses. The results are shown in the Table 6.16. Only six of the over-expressed candidate genes have been annotated to have a decreased DNA copy number status in one of the *ArrayCGH* profiles. The annotation attributes *Max*, *Min* and *Mean* support these annotations. The candidate genes *GRB7*, *NGFR*, *HBOA* and *DLX4* have been annotated as under-expressed by the standard *HMM* and the candidates *GRB7*, *NGFR*, *DLX4*, *RAD51C* have been annotated as under-expressed by the extended *HMM*. It can be possible that the loss of such a gene could cause the under-expression which we have observed.

Let us now consider the genes which are not included in the Table 5.2, but which have also been annotated to have a decreased DNA copy number status. The additional candidates with such an annotation are listed in the Table 6.17. We have not used a threshold value because of the low number of candidates. Most of the additional candidates show good annotation attributes *Max*, *Min* and *Mean* excepted the gene *CDK5R1* which has attribute values close to zero. As above we have used the *Entrez Gene* at the *NCBI* to get a better impression what functions some of these candidate genes with *Nr* values greater than one have, but we have not found functions which are directly associated with breast cancer or which should necessarily be reported. The genes *EVI2B*, *EVI2A* and *SCYA2* have been annotated as under-expressed by the standard and the extended *HMM* and these genes have also been annotated to have a decreased DNA copy

Gene	Band	Nr	Max	Min	Mean	Gene	Band	Nr	Max	Min	Mean
GRB7	q12	1	-1.85	-1.85	-1.85	DLX4	q21.33	1	-0.79	-0.79	-0.79
NGFR	q21.33	1	-0.62	-0.62	-0.62	RAD51C	q23.2	1	-1.36	-1.36	-1.36
HBOA	q21.33	1	-0.88	-0.88	-0.88	PPM1D	q23.2	1	-1	-1	-1

Table 6.16: Over-expressed candidate genes from Table 5.2 which have also been annotated by the *HMM* with two transition classes to have a decreased DNA copy number status. Max represents the highest log ratio which was measured for the Nr experiments where the gene has been annotated to have a decreased DNA copy number status. Min represents the lowest log ratio which was measured for the Nr experiments where the gene has been annotated to have a decreased DNA copy number status. Mean is the mean log ratio of the measured log ratios for the Nr experiments where the gene has been annotated with a decreased DNA copy number status.

Gene	Band	Nr	Max	Min	Mean	Gene	Band	Nr	Max	Min	Mean
SLC25A11	p13.2	1	-1.12	-1.12	-1.12	SCYA11	q12	1	-0.94	-0.94	-0.94
ACADVL	p13.1	4	-1.12	-2.25	-1.48	SCYA13	q12	1	-0.64	-0.64	-0.64
MAP2K4	p12	1	-0.94	-0.94	-0.94	LIG3	q12	1	-0.86	-0.86	-0.86
COX10	p12	1	-0.74	-0.74	-0.74	NAP4	q12	1	-1.06	-1.06	-1.06
ZNF286	p12	1	-0.97	-0.97	-0.97	KRT10	q21.2	1	-1.15	-1.15	-1.15
MFAP4	p11.2	1	-1.32	-1.32	-1.32	GCN5L2	q21.2	1	-1.79	-1.79	-1.79
TNFAIP1	q11.2	1	-1	-1	-1	IFI35	q21.31	1	-1	-1	-1
FLOT2	q11.2	1	-1.79	-1.79	-1.79	DUSP3	q21.31	1	-1	-1	-1
OMG	q11.2	1	-0.81	-0.81	-0.81	UBTF	q21.31	1	-1.18	-1.18	-1.18
EVI2B	q11.2	2	-0.94	-1.09	-1.02	NSF	q21.31	1	-0.94	-0.94	-0.94
EVI2A	q11.2	2	-0.67	-0.74	-0.71	SPOP	q21.33	1	-0.56	-0.56	-0.56
ZNF207	q11.2	3	-0.4	-2.47	-1.73	UGTREL1	q21.33	1	-0.58	-0.58	-0.58
PSMD11	q11.2	1	-0.81	-0.81	-0.81	TOM1L1	q22	1	-1.69	-1.69	-1.69
CDK5R1	q11.2	1	-0.09	-0.09	-0.09	PCTP	q23.1	2	-1.43	-2.56	-2
SCYA2	q12	3	-0.47	-1.28	-0.94	AKAP1	q23.2	1	-0.74	-0.74	-0.74
SCYA7	q12	1	-0.38	-0.38	-0.38	ACOX1	q25.1	1	-1.69	-1.69	-1.69

Table 6.17: Additional genes on chromosome 17 which have been annotated by the *HMM* with two transition classes to have a decreased DNA copy number status. Max represents the highest log ratio which was measured for the Nr experiments where the gene has been annotated to have a decreased DNA copy number status. Min represents the lowest log ratio which was measured for the Nr experiments where the gene has been annotated to have a decreased DNA copy number status. Mean is the mean log ratio of the measured log ratios for the Nr experiments where the gene has been annotated to have a decreased DNA copy number status.

number status. In summary, the decreased copy numbers of these genes are also visible in the expression levels of these genes.

Selected *ArrayCGH* profiles

Now we describe the selected *ArrayCGH* profiles in the Figures 6.13, 6.14, 6.15, 6.16 and 6.17 in detail. We have taken the profiles of the same breast cancer cell lines and primary tumours as we have done in the analysis of the gene expression profiles and therefore it is possible to see the influences of DNA copy number changes on the gene expression levels of affected genes. But now we concentrate on the quality of the *ArrayCGH* annotations to get an impression of the performance of the extended *HMM*.

- The Figure 6.13 shows the annotations of the cell lines *BT474* and *SKBR3*. The profile *BT474* does not contain genes which have been annotated to have a decreased DNA copy number status. The state posterior profile $sp(-)$ of the state $-$ is close to zero and contains only a small peak in the chromosomal band 17q23.3. Increased copy numbers of genes are located in the two segments 17q12 and 17q21.32-17q23.2 and the state posterior $sp(+)$ is significantly greater than zero in these segments. Clark *et al.* [3] have analysed the DNA copy number status of *BT474* and they have found gains of DNA segments in 17q11-21 and 17q22-23. These results support the good quality of our findings.
  The cell line *SKBR3* contains genes which have been annotated to have a decreased DNA copy number status. These genes are located in the regions 17q11.2-17q12, 17q21.31, 17q21.33 and 17q23.2. The state posterior $sp(-)$ of the affected segments in these regions is significantly greater than zero for most of the genes in these segments. Gains of DNA segments are found in the regions 17q11.2, 17q12-17q21.2 and 17q25.3 and the state posterior $sp(+)$ supports this annotations with high values except smaller values at the end of the segment in 17q11.2.

- The Figure 6.14 represents the annotations of the primary breast tumours *NORWAY 14* and *NORWAY 26*. The *ArrayCGH* profile of *NORWAY 14* does not contain segments with deleted genes, but there is a peak in the state posterior profile $sp(-)$ in the region 17q12. One gene in the whole profile has been annotated to have an increased DNA copy number status and this is supported by a significant peak in the state posterior profile $sp(+)$. In general, most of the log ratios are close to zero as we can also see in the state posterior profile $sp(=)$.
  The primary tumour *NORWAY 26* shows no segments of decreased DNA copy number status and there are no peaks in the state posterior profile $sp(-)$. One segment of increased DNA copy number status is located in the region 17q12-17q21.2 and this segment is clearly visible in the state posterior profile $sp(+)$.

- The Figure 6.15 contains the annotations of the primary breast tumours *NORWAY 47* and *NORWAY 53*. The profile *NORWAY 47* contains one segment of decreased DNA copy number status in the region 17p13.1. This segment can also be seen in the state posterior profile $sp(-)$. Increased DNA copy number status is found in two segments which are located in the chromosomal regions 17q11.2 and 17q12-17q21.2 and these both segments are clearly to see in the state posterior profile $sp(+)$.
  The *ArrayCGH* profile of *NORWAY 53* contains one gene in the chromosomal band 17q12 which has been annotated to have a decreased DNA copy number status. Two peaks can be seen in the state posterior profile $sp(-)$ the one for the gene in 17q12 and the other in the chromosomal band 17p13.1. Two regions of increased DNA copy number status have been annotated by the extended *HMM*. These regions are located close to each other the

one in the region 17q12-21.1 and the other in 17q21.2. Both regions show characteristic patterns in the state posterior profile $sp(+)$.

- The Figure 6.16 represents the annotations of the primary breast tumours *NORWAY 100* and *STANFORD 2*. The profile *NORWAY 100* does not contain segments of decreased DNA copy number status and state posterior profile $sp(-)$ does not contain peaks. All log ratios in this profile are close to zero.
  The profile of the primary breast tumour *STANFORD 2* does not show segments of decreased DNA copy number status. The state posterior profile $sp(-)$ of this annotation does not give hints that such segments exist. The regions 17q11.2 an 17q12-17q21.3 contain segments which have been annotated to have an increased DNA copy number status. These both segments are clearly visible in the state posterior profile of $sp(+)$.

- The Figure 6.17 shows the annotations of the primary breast tumours *STANFORD 24* and *STANFORD A*. The *ArrayCGH* profile of *STANFORD 24* has completely been annotated to have an unchanged DNA copy number status. All log ratios are close to zero except the log ratio of one gene in the region 17q21.31. The log ratio of this gene is less than zero and the state posterior profile $sp(-)$ shows a peak for this gene.
  The profile of *STANFORD A* does not contain segments of decreased DNA copy number status and only two small peaks in the state posterior profile $sp(-)$ which could give hints for such segments are visible in the chromosomal bands 17p13.1 and 17p12. The four regions 17q11.2, 17q12, 17q21.32-17q21.33 and 17q23.2-17q24.2 contain segments which have been annotated to have an increased DNA copy number status. All these segments are clearly visible in the state posterior profile $sp(+)$.

The *ArrayCGH* profiles which we have analysed in more detail give a general overview of all *ArrayCGH* profiles. The annotation performance of the extended *HMM* with two transition classes is good. As we have already observed for the gene expression profiles the extended *HMM* is able to model a good segment structure and this improved modelling leads to a good annotation quality. Nevertheless, it would be a good strategy to test the performance of the extended *HMM* on simulated *ArrayCGH* data or on real data with a known annotation.

Summary of the annotation process

Let us consider the whole annotation process for our breast cancer *ArrayCGH* data set. The Figure 6.18 contains the four subfigures *Unchanged Segments*, *Gained Segments*, *Lost Segments* and *Gene Counts*.

The subfigure *Unchanged Segments* shows the locations and the absolute frequencies of segments which have been annotated to have an unchanged DNA copy number status. Most of the *ArrayCGH* profiles have an individual segment structure, but there are also some common segments with unchanged DNA copy number status which are represented by darker hexagons. The lengths of the unchanged segments are in general longer than the lengths of the segments in the subfigures *Gained Segments* and *Lost Segments*.

The subfigure *Gained Segments* represents the locations and the absolute frequencies of segments which have been annotated to have an increased DNA copy number status. These segments are smaller as the segments with unchanged DNA copy number status. The variability of segments with increased DNA copy number status is low as the subfigure shows. Two segments occur more often than others as the dark hexagons show.

The subfigure *Lost Segments* gives an overview of the locations and absolute frequencies of segments which have been annotated to have a decreased DNA copy number status. These segments behave like the segments with increased DNA copy number status, but in general their lengths are smaller. One segment occurs more often than others as a dark hexagon shows.

The smaller lengths of these segments in comparison with the segments of increased DNA copy number status can be also seen in the *ArrayCGH* profiles of the Figures 6.13 and 6.15.
The subfigure *Gene Counts* represents the absolute frequencies for decreased and for increased DNA copy number status of genes over all *ArrayCGH* profiles in the breast cancer data set. The absolute frequencies for decreased DNA copy number status of genes are shown in the upper subfigure and the absolute frequencies for increased DNA copy number status of genes are shown in the subfigure below. Genes which have frequently been annotated to have a decreased DNA copy number status are located in the regions 17p13.1, 17q11.2, 17q12 and 17q22-17q23.2. These findings are supported by the analysis of Orsetti *et al.* [21] who have observed losses in 17p, 17q11.2 and 17q22. Increased DNA copy number status can be mainly seen for genes in the regions 17q11.2, 17q12 and 17q21.32-17q25.3. These regions are known from the results of Orsetti *et al.* [21], Monni *et al.* [17], Kauraniemi *et al.* [13], Willis *et al.* [27] and Hyman *et al.* [9] which have been described above in detail.

			17p												17q										
Experiment	13.3	13.2	13.1	12	11.2	11.1	11.2	12	21.1	21.2	21.31	21.32	21.33	22	23.1	23.2	23.3	24.1	24.2	24.3	25.1	25.2	25.3	∑	E
BT474								22				6	12	4	1	10								55	4
MCF7	1															11	1	2	5					20	0
SKBR3							14	6	1	16													3	40	3
T47D							7	3																10	0
NORWAY 7																								0	0
NORWAY 10																								0	0
NORWAY 11								3																3	0
NORWAY 12																								0	0
NORWAY 14								1																1	0
NORWAY 15																								0	0
NORWAY 16																								0	0
NORWAY 17								3																3	0
NORWAY 18																								0	0
NORWAY 19																								0	0
NORWAY 26								13	1															14	0
NORWAY 27																								0	0
NORWAY 39																								0	0
NORWAY 41																								0	0
NORWAY 47							16	12	1	13														42	1
NORWAY 48																								0	0
NORWAY 53								4	1	1														6	0
NORWAY 56													9	4	1	12	9	2	6	2	22	1	8	76	8
NORWAY 57								12	1	1		8	3				7	2	1		11		1	47	5
NORWAY 61								4	1							2								7	1
NORWAY 65								4																4	0
NORWAY 100																								0	0
NORWAY 101								11	1															12	0
NORWAY 102																								0	0
NORWAY 104																								0	0
NORWAY 109																					5	1	8	14	0
NORWAY 111																								0	0
NORWAY 112																								0	0
STANFORD 2							6	4	1	3														14	3
STANFORD 14																								0	0
STANFORD 16																								0	0
STANFORD 17								3																3	0
STANFORD 23																					8	1	9	18	1
STANFORD 24																								0	0
STANFORD 35																								0	0
STANFORD 38																								0	0
STANFORD A							12	6				8	10			10	9	2	2					59	3
∑	1						55	111	8	34		22	34	8	2	45	26	8	14	2	46	3	29	448	29

Table 6.18: Overview of amplified genes per chromosomal band. Candidate genes in an *ArrayCGH* profile which have been annotated by the *HMM* with two transition classes to have an increased DNA copy number status have been assigned to their chromosomal bands. The results are contained in this table. $\sum$ is the row or the column sum of genes which have been annotated to have an increased DNA copy number status. E is the sum of genes in an experiment which have been annotated to have an increased DNA copy number status when their log ratios are less than zero.

	17p					17q																			
Experiment	13.3	13.2	13.1	12	11.2	11.1	11.2	12	21.1	21.2	21.31	21.32	21.33	22	23.1	23.2	23.3	24.1	24.2	24.3	25.1	25.2	25.3	$\sum$	E
BT474																								0	0
MCF7																								0	0
SKBR3							6	5			2		5			1								19	0
T47D				2																				2	0
NORWAY 7								1																1	0
NORWAY 10																								0	0
NORWAY 11																								0	0
NORWAY 12																								0	0
NORWAY 14																								0	0
NORWAY 15																								0	0
NORWAY 16																								0	0
NORWAY 17							1																	1	0
NORWAY 18							1																	1	0
NORWAY 19																								0	0
NORWAY 26																								0	0
NORWAY 27							2	1			1			1										5	0
NORWAY 39							2				1										1			4	0
NORWAY 41			1					1							1									3	0
NORWAY 47			1																					1	0
NORWAY 48										1														1	0
NORWAY 53								1																1	0
NORWAY 56																								0	0
NORWAY 57																								0	0
NORWAY 61																								0	0
NORWAY 65			1												1	1								3	0
NORWAY 100																								0	0
NORWAY 101																								0	0
NORWAY 102																								0	0
NORWAY 104																								0	0
NORWAY 109																								0	0
NORWAY 111		1														1								2	0
NORWAY 112																								0	0
STANFORD 2																								0	0
STANFORD 14																								0	0
STANFORD 16																								0	0
STANFORD 17			1																					1	0
STANFORD 23					1																			1	0
STANFORD 24																								0	0
STANFORD 35				1						1														2	0
STANFORD 38																								0	0
STANFORD A																								0	0
$\sum$		1	4	3	1		12	9		2	4		5	1	2	3					1			48	0

Table 6.19: Overview of lost genes per chromosomal band. Candidate genes in an *ArrayCGH* profile which have been annotated by the *HMM* with two transition classes to have a decreased DNA copy number status have been assigned to their chromosomal bands. The results are contained in this table. $\sum$ is the row or the column sum of genes which have been annotated to have a decreased DNA copy number status. E is the sum of genes in an experiment which have been annotated to have a decreased DNA copy number status when their log ratios are greater than zero.

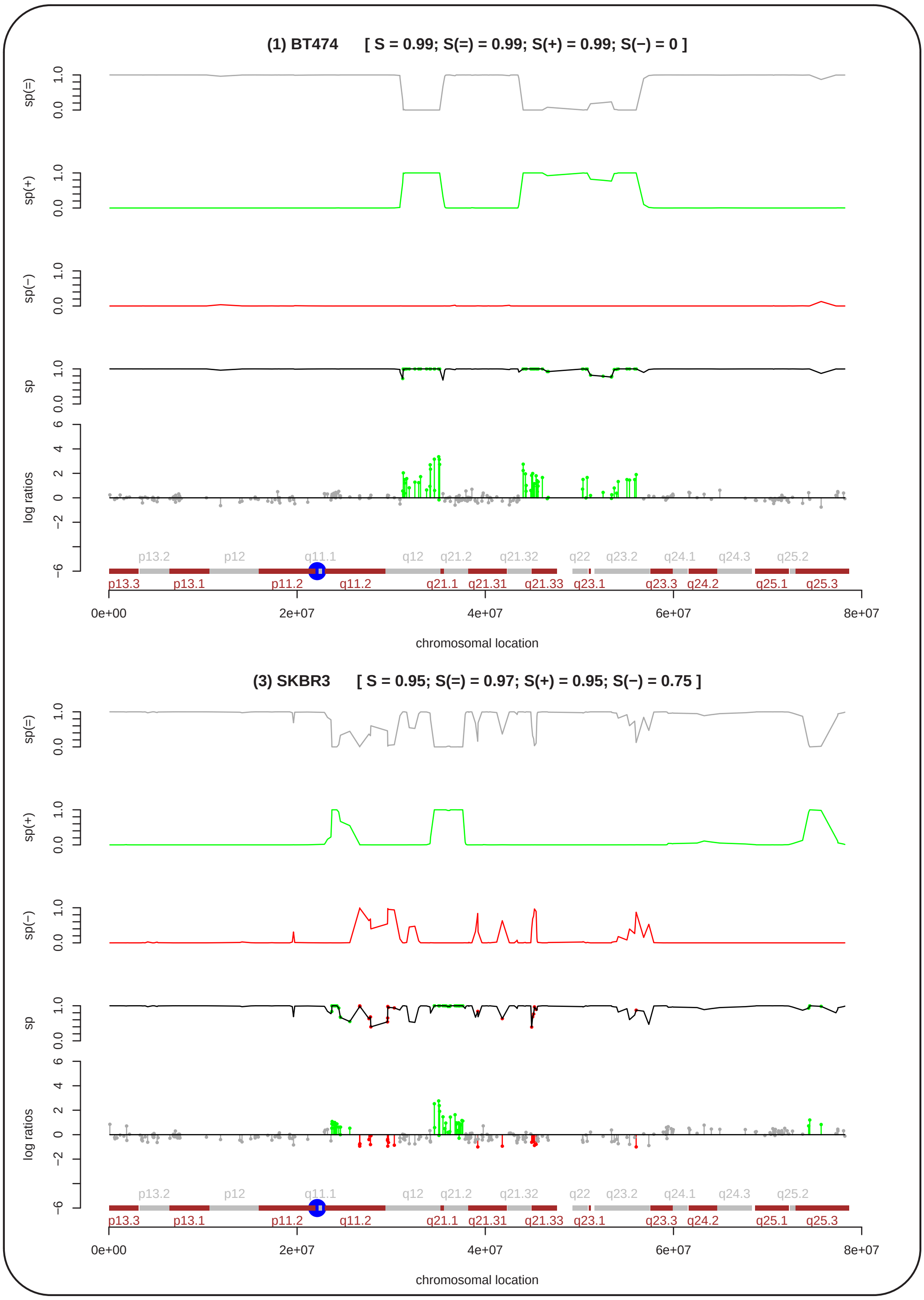


Figure 6.13: *ArrayCGH* profiles of chromosome 17 which have been annotated by the extended *HMM* with two transition classes. The first profile shows the *BT474* and the second the *SKBR3* data. In general, the headline of a profile contains the unique profile number (n) which is used in the Figure 5.3, the profile name, the sum S (5.3) of the state posteriors for the *Viterbi Path* and the relative proportion $S(i)$ (5.3), with $i \in \{-, =, +\}$, of the state posterior profile i in the *Viterbi Profile*. Top down the five graphics per experiment show the state posterior profiles for the states $=$, $+$ and $-$, the state posterior profile of the *Viterbi Path* and the annotated *ArrayCGH* profile. The annotation of a gene is represented by the colour of the log ratio line. A red line represents a decreased, a grey line an unchanged and a green line an increased DNA copy number status of a gene.

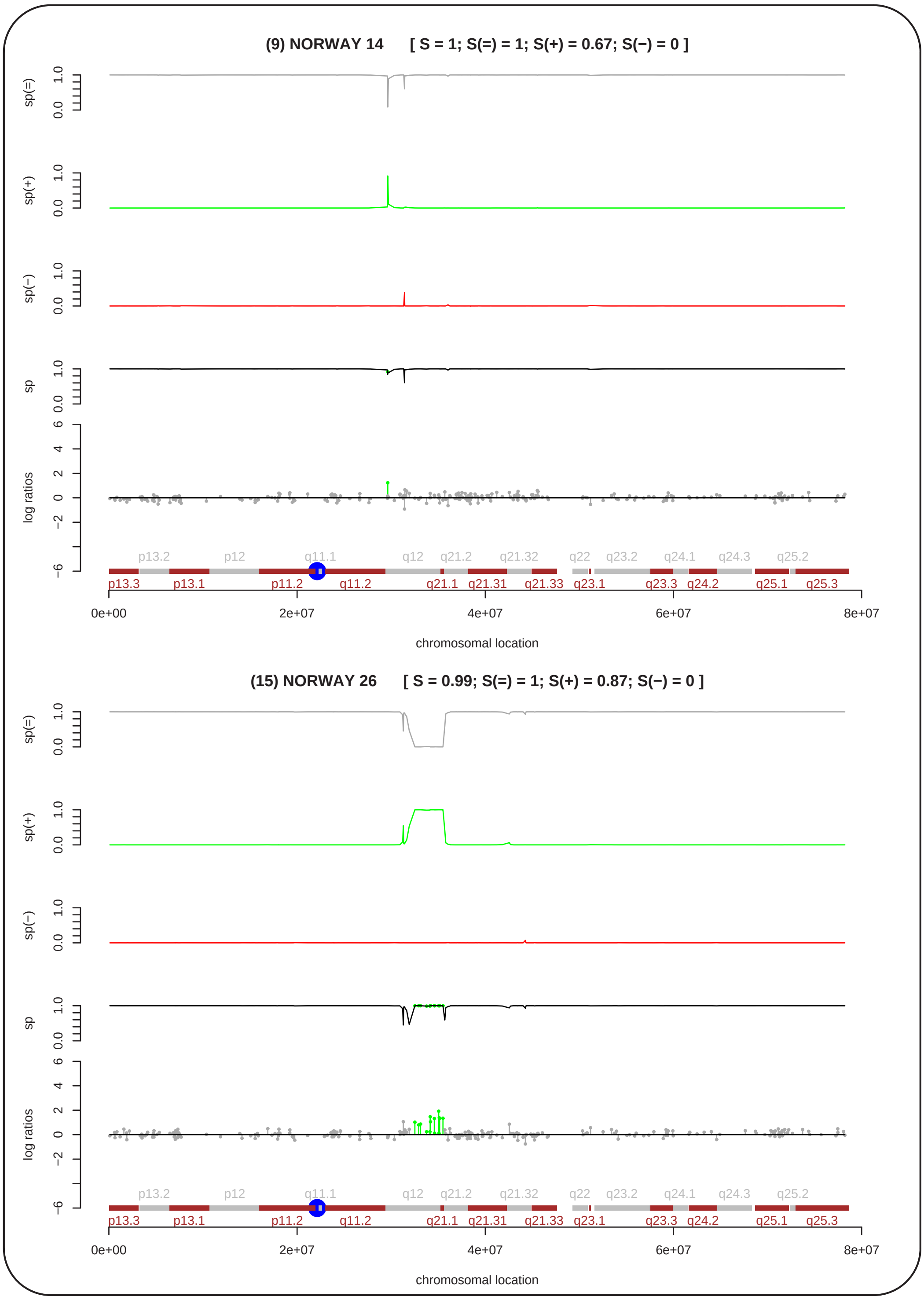


Figure 6.14: *ArrayCGH* profiles of chromosome 17 which have been annotated by the extended *HMM* with two transition classes. The first profile shows the *NORWAY 14* and the second the *NORWAY 26* data. In general, the headline of a profile contains the unique profile number (n) which is used in the Figure 5.3, the profile name, the sum S (5.3) of the state posteriors for the *Viterbi Path* and the relative proportion $S(i)$ (5.3), with $i \in \{-, =, +\}$, of the state posterior profile i in the *Viterbi Profile*. Top down the five graphics per experiment show the state posterior profiles for the states $=$, $+$ and $-$, the state posterior profile of the *Viterbi Path* and the annotated *ArrayCGH* profile. The annotation of a gene is represented by the colour of the log ratio line. A red line represents a decreased, a grey line an unchanged and a green line an increased DNA copy number status of a gene.

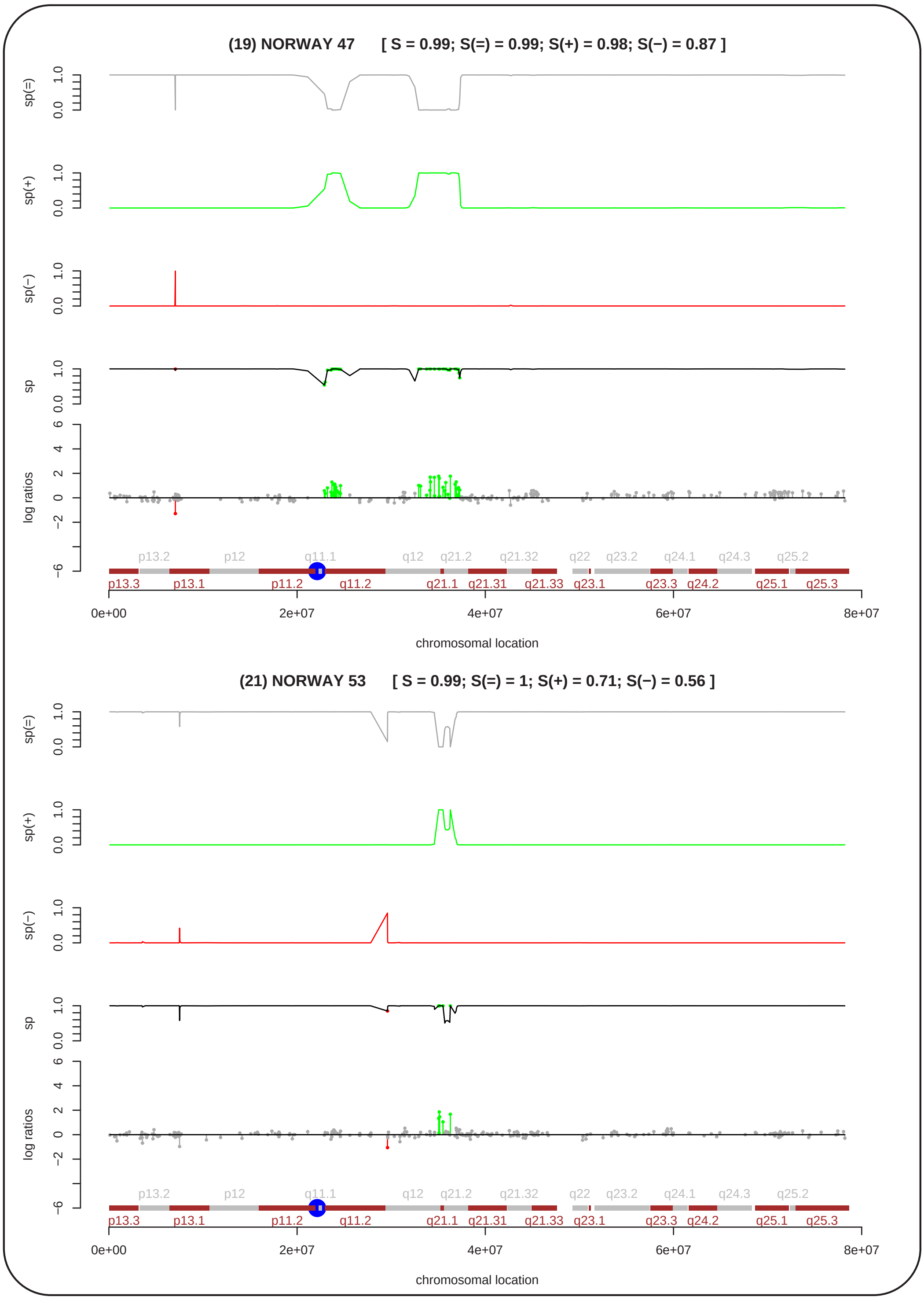


Figure 6.15: *ArrayCGH* profiles of chromosome 17 which have been annotated by the extended *HMM* with two transition classes. The first profile shows the *NORWAY 47* and the second the *NORWAY 53* data. In general, the headline of a profile contains the unique profile number (n) which is used in the Figure 5.3, the profile name, the sum S (5.3) of the state posteriors for the *Viterbi Path* and the relative proportion $S(i)$ (5.3), with $i \in \{-, =, +\}$, of the state posterior profile i in the *Viterbi Profile*. Top down the five graphics per experiment show the state posterior profiles for the states $=$, $+$ and $-$, the state posterior profile of the *Viterbi Path* and the annotated *ArrayCGH* profile. The annotation of a gene is represented by the colour of the log ratio line. A red line represents a decreased, a grey line an unchanged and a green line an increased DNA copy number status of a gene.

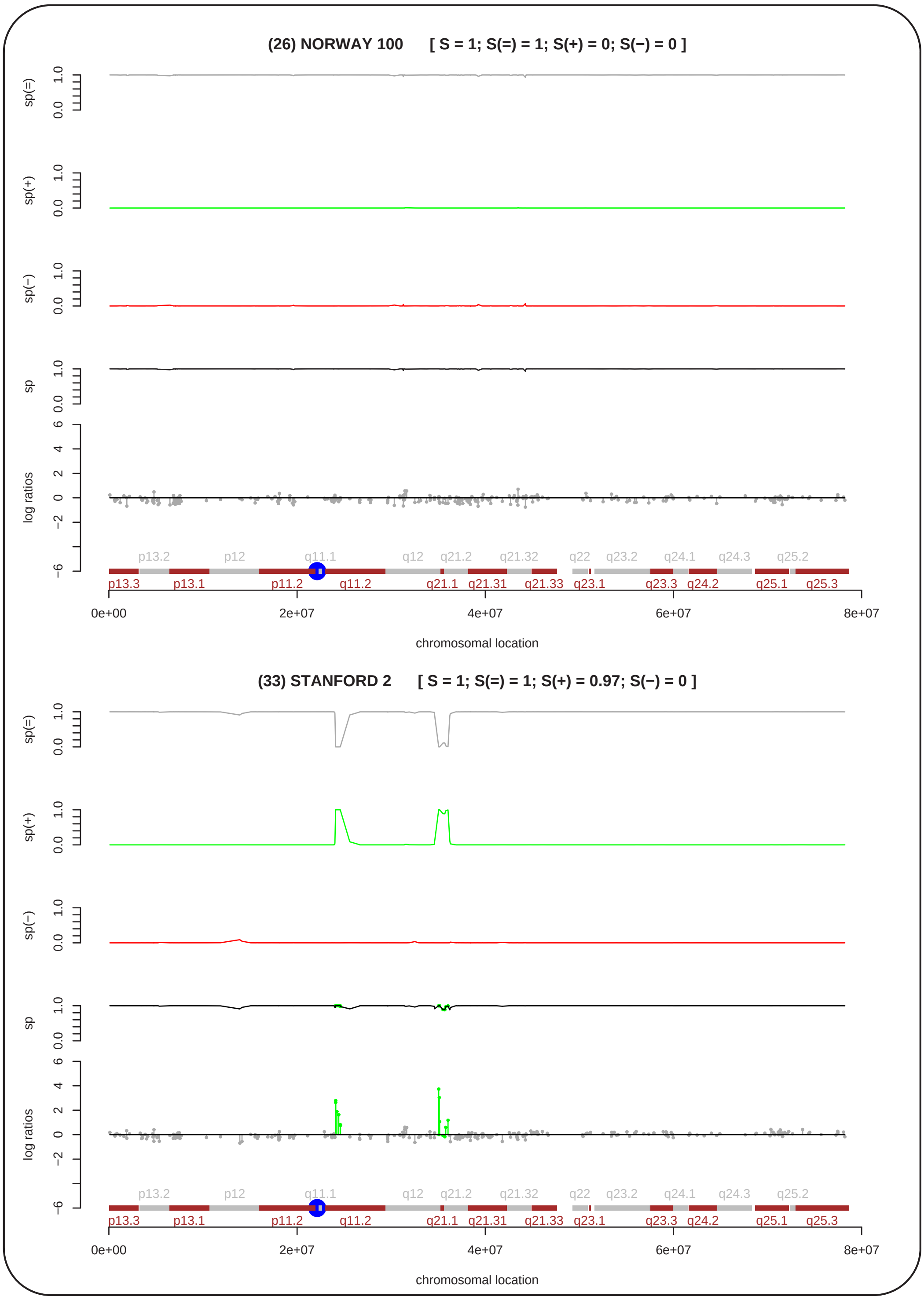


Figure 6.16: *ArrayCGH* profiles of chromosome 17 which have been annotated by the extended *HMM* with two transition classes. The first profile shows the *NORWAY 100* and the second the *STANFORD 2* data. In general, the headline of a profile contains the unique profile number (n) which is used in the Figure 5.3, the profile name, the sum S (5.3) of the state posteriors for the *Viterbi Path* and the relative proportion $S(i)$ (5.3), with $i \in \{-, =, +\}$, of the state posterior profile i in the *Viterbi Profile*. Top down the five graphics per experiment show the state posterior profiles for the states $=$, $+$ and $-$, the state posterior profile of the *Viterbi Path* and the annotated *ArrayCGH* profile. The annotation of a gene is represented by the colour of the log ratio line. A red line represents a decreased, a grey line an unchanged and a green line an increased DNA copy number status of a gene.

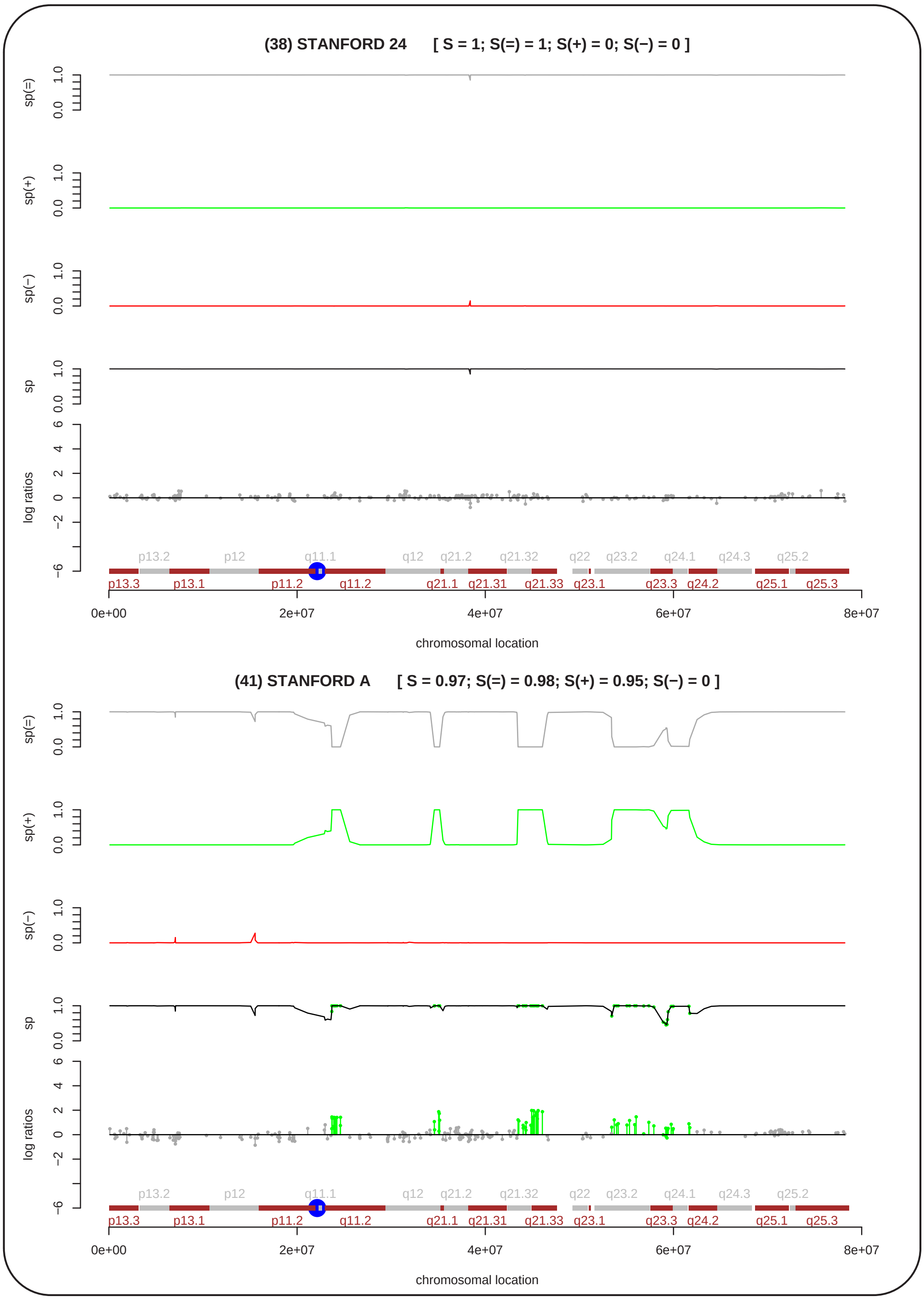


Figure 6.17: *ArrayCGH* profiles of chromosome 17 which have been annotated by the extended *HMM* with two transition classes. The first profile shows the *STANFORD 24* and the second the *STANFORD A* data. In general, the headline of a profile contains the unique profile number (n) which is used in the Figure 5.3, the profile name, the sum S (5.3) of the state posteriors for the *Viterbi Path* and the relative proportion $S(i)$ (5.3), with $i \in \{-, =, +\}$, of the state posterior profile i in the *Viterbi Profile*. Top down the five graphics per experiment show the state posterior profiles for the states $=$, $+$ and $-$, the state posterior profile of the *Viterbi Path* and the annotated *ArrayCGH* profile. The annotation of a gene is represented by the colour of the log ratio line. A red line represents a decreased, a grey line an unchanged and a green line an increased DNA copy number status of a gene.

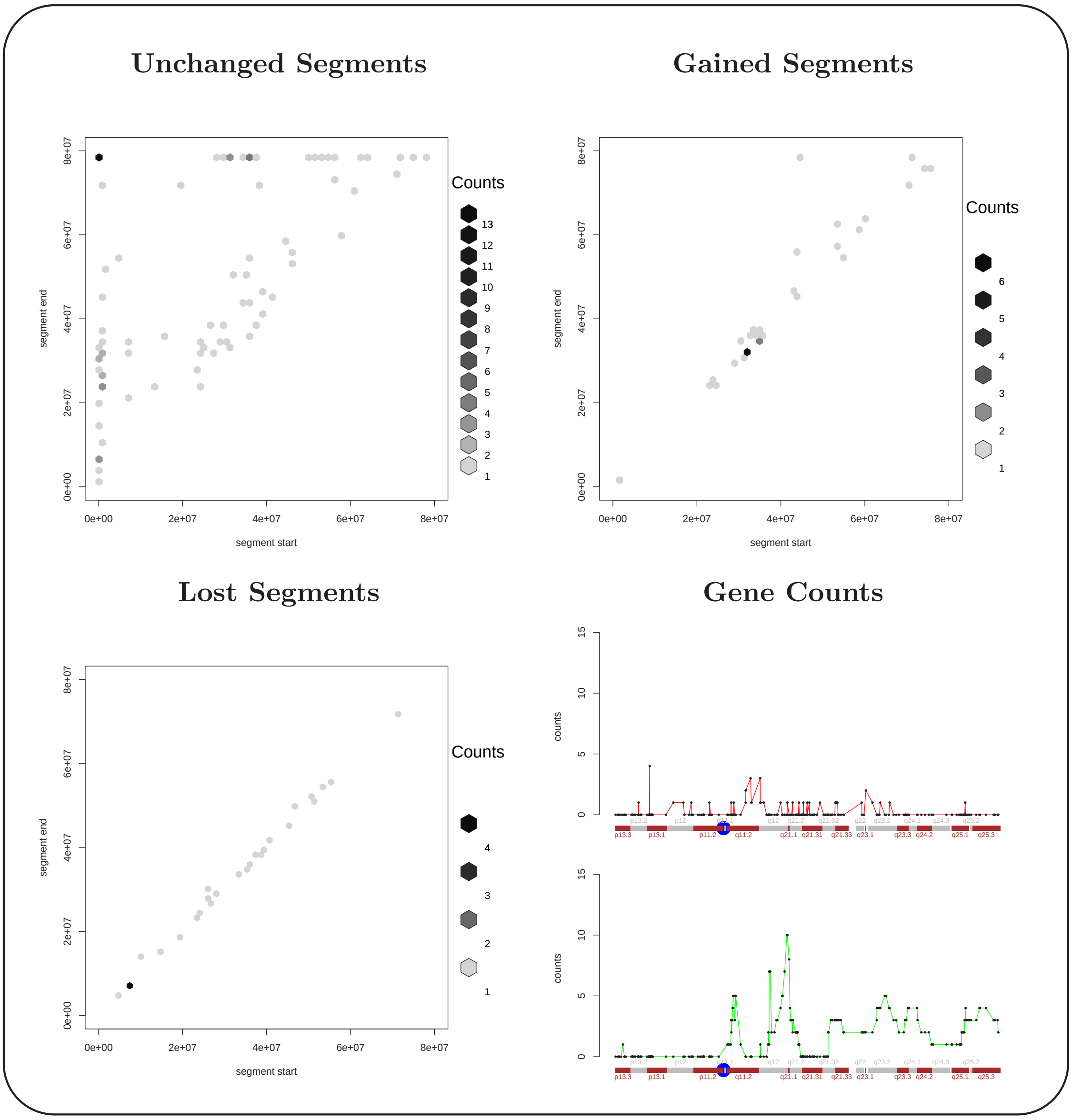


Figure 6.18: Overview of annotations for the extended *HMM* with two transition classes. In general a segment is a sequence of successive genes which have been annotated in the same way. Each segment has a start and an end point on the chromosome. The figure **Unchanged Segments** gives a summary of segments on chromosome 17 which show an unchanged DNA copy number status. The figure **Gained Segments** shows a summary of segments on chromosome 17 which have an increased DNA copy number status. The figure **Lost Segments** represents a summary of segments on chromosome 17 which show a decreased DNA copy number status. The figure **Gene Counts** represents the absolute frequencies for decreased and for increased DNA copy number status of genes over all *ArrayCGH* profiles. The absolute frequencies for decreased DNA copy number status of genes are shown in the upper subfigure and the absolute frequencies for increased DNA copy number status of genes are given in the subfigure below.

6.3 Comparing Gene Expression And *ArrayCGH* Profiles

With the help of an example we will show the direct relation between copy number changes and changes in gene expression levels. The loss of one gene copy in a diploid cell can cause the decreased gene expression of this gene when the other gene copy is not able to compensate this loss. If a gene is amplified, then this can cause the increased gene expression of this gene. There are also a lot of effects thinkable which a loss or a gain of a gene could have on the regulation of the gene expression of other genes.

In our example we consider the *ArrayCGH* and the gene expression profile of the primary breast tumour *STANFORD A*. Both annotated profiles are shown in the Figure 6.19. The effects of gene amplifications are clearly visible in the gene expression levels of the affected genes. The annotations of the profiles are taken from the extended *HMM*s which we have used to analyse the *ArrayCGH* and the gene expression data. The annotation of the *ArrayCGH* profile contains the four regions 17q11.2, 17q12, 17q21.32-17q21.33 and 17q23.2-17q24.2 which have been annotated to have an increased DNA copy number status. The regions 17q11.2, 17q12 and 17q21.32-17q21.33 of increased DNA copy number status have also been annotated as over-expressed in the gene expression profile. The amplification of genes in these regions has led to a significant elevation of the gene expression levels of these genes. The chromosomal region 17q23.2-17q24.2 which has been annotated to have an increased DNA copy number status has only partly been annotated as over-expressed in the gene expression profile. That is, some genes in 17q23.2 have been annotated as over-expressed, but the others have been annotated as identically expressed because the changes in the expression signals are too low to be detected by the extended *HMM*. Nevertheless, the state posterior profile $sp(+)$ of the gene expression profile for *STANFORD A* in the Figure 6.11 shows higher values for some of the genes which have been annotated as identically expressed in the direct neighbourhood of over-expressed genes.

In summary, the comparison of *ArrayCGH* and gene expression profiles is a good starting point to analyse the effects of losses and gains of DNA segments on the gene expression of affected genes. *HMM*s are able to find such regions in both data classes.

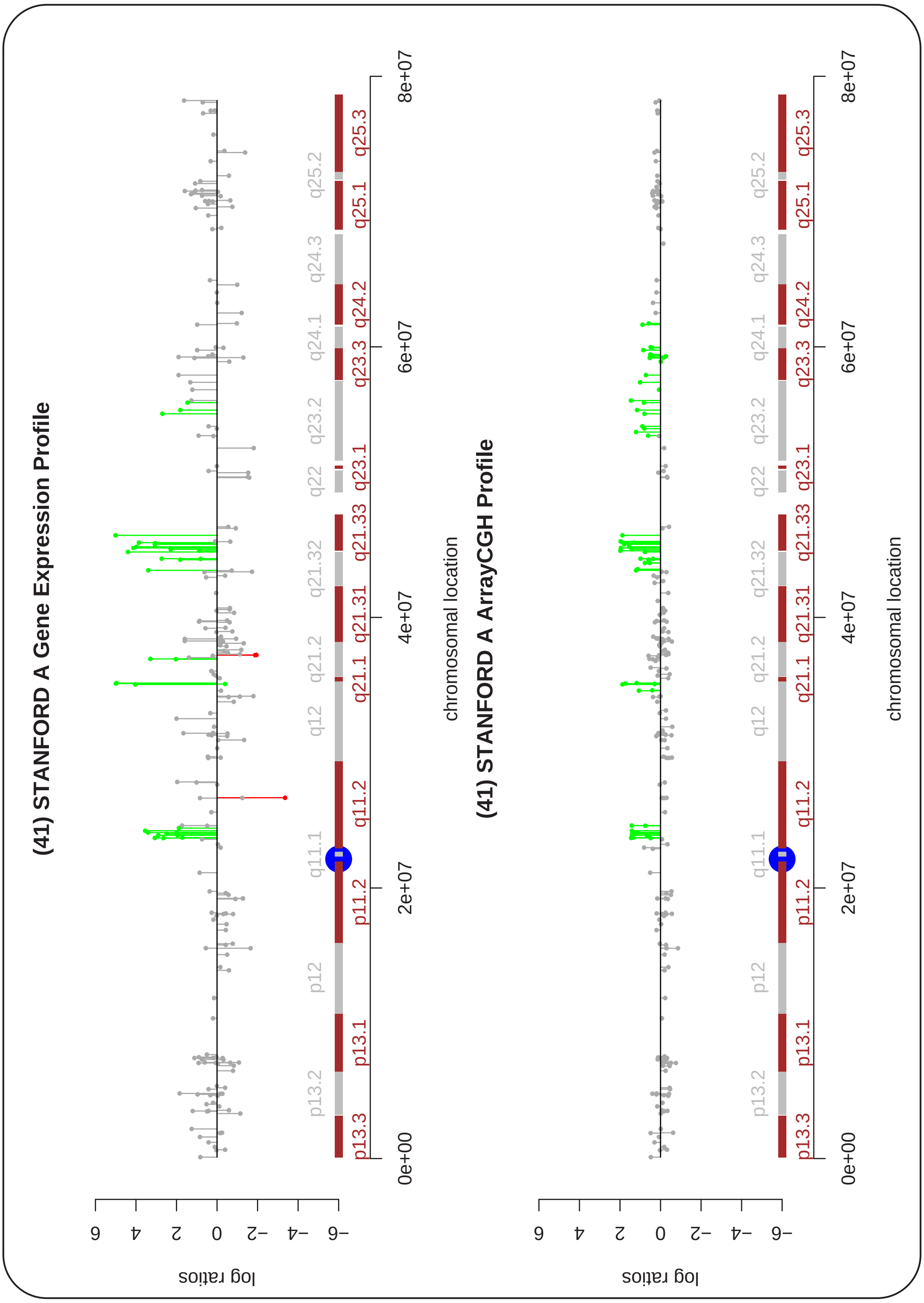


Figure 6.19: The gene expression and the *ArrayCGH* profile of the primary breast tumour *STANFORD A* on chromosome 17 which have been annotated by extended *HMM*s with two transition classes. The annotation of a gene is represented by the colour of the log ratio line. A red line represents an under-expressed (decreased), a grey line an identically expressed (unchanged) and a green line an over-expressed (increased) expression level (DNA copy number status) of a gene.

Chapter 7

Discussion Of The *HMM* Approach

The goal of this chapter is to review the developed *HMM* approach (Section 7.1), to summarise the performance of this approach on breast cancer microarray data (Section 7.2) and to give an outlook of future developments on the basis of this approach (Section 7.3).

7.1 Summary And Discussion Of The *HMM* Approach

In this diploma thesis we have developed a general strategy how to create *HMM*s to analyse microarray data. Our concept is able to work with homogeneous and inhomogeneous *HMM*s and therewith the comparison of both *HMM* classes is possible. The main attention has been focused on the development of a novel approach using the proximity of genes in the analysis of microarray data to create more realistic models for the detection of chromosomal imbalances and gene expression alterations. The theoretical background of our novel *HMM*s with coupled transition matrices has intensively been studied in this diploma thesis and can be used as basis for future modifications.

The losses and gains of DNA segments in breast cancer cells have been an ideal starting point to test the performance of the novel *HMM*s with coupled transition matrices in comparison with the standard *HMM*s.

Let us now review important steps in the development of *HMM*s to analyse microarray data. The first critical step in our strategy is the selection of a fitting mixture model as basis for the emission functions of our proposed three-state *HMM* as for instance shown in the Figures 5.1 and 5.2. The mixture estimation using the *EM algorithm* in combination with *BIC* and *AIC* has shown good performance and is therefore the method of choice. Nevertheless, the selection of a mixture model from the estimated candidates requires some experience. Good candidate models should not show great overlaps between the clustered mixture components. The mixture model in the Figure 4.1 is such a good candidate. The goal is to have initial emission functions which are able to model the data classes in the microarray data that we expect. In summary, we require emission functions which can model the ranges of under-expressed, identically expressed and over-expressed genes. It should be obvious how to adjust this goal to *ArrayCGH* data.

The selected mixture model determines the start distribution of our *HMM*s and therewith the distribution of the microarray measurements can be modelled by these *HMM*s when a transition matrix with an equilibrium distribution equal to the start distribution is used. We have proposed a function class of transition matrices which allows a flexible modelling of the initial transition probabilities. Therewith, it should be possible to include previous knowledge about the frequencies of segment changes in microarray data into the initial transition probabilities of the *HMM*. The training of homogeneous and inhomogeneous *HMM*s has shown that different initial transition matrices have mostly led to nearly equal annotation results. In particular, this has been observed as we have tried to improve the performance of homogeneous *HMM*s.

Our novel developed inhomogeneous *HMM*s with coupled transition matrices make use of the distance between adjacent genes. The transitions between successive measurements of genes are divided into distance classes using a predefined distance-dependent transition class switching function as for instance in the Definition (5.1). Each pair of successive measurements for genes is mapped into a predefined transition matrix which the *HMM* has to use to model the transition between this pair of measurements. All these information are included in the estimation of a basic transition matrix for the *HMM* with coupled transition matrices. Afterwards the basic transition matrix is mapped into the other transition matrices of the *HMM* by scaling the state durations with predefined scaling factors.

However, to determine the number of transition classes, to define distance-dependent transition class switching functions and to choose good scaling parameters are difficult steps if no detailed information about the losses or gains of DNA segments are available. That is why we have tested many different models to get an impression how the performance of our novel *HMM*s with coupled transition matrices is on breast cancer microarray data. Our presented annotation results have been created by *HMM*s with two coupled transition matrices. *HMM*s with three coupled transition matrices have shown nearly the same results. For a better understanding how to choose the parameters in our developed approach a test on simulated data or on real biological data with known annotation should give us more information about the parameter selection.

The annotation of a microarray profile represents the most probable state path through the *HMM* and this *Viterbi Path* is computed by the *Viterbi algorithm*. Each gene in a microarray profile is characterised by its assigned annotation in the *Viterbi Path*. The state posterior of a gene annotation has been used as an indirect quality criterion for the significance of a gene annotation. Nevertheless, we should try to establish a quality criterion in terms of *P-Values*. We have done the first steps to achieve this, but we have not integrated it at this time because there will be a lot of additional work to finish it.

The comparison of the annotation performance between *HMM*s with coupled transition matrices and standard *HMM*s on breast cancer microarray data has revealed the following advantage for the usage of the *HMM*s with coupled transition matrices:

- The annotations of microarray profiles show an improved segmentation structure.

The improved segmentation structure is the basis for a better characterisation of the locations of altered regions in a microarray profile. However, the standard *HMM*s have problems to determine the start and the end of a segment and therefore often questionable starts and ends of segments have been observed. These problems have sometimes led to subsegments with lower log ratios in over-expressed segments. All these problems have rarely been seen in the annotation results which have been created by *HMM*s with coupled transition matrices. We have discussed this observation several times in the Chapter 6.

Now that we have summarised and discussed the theory behind our developed *HMM* approach the main results of the analysis of breast cancer microarray data are considered.

7.2 Performance Of The *HMM* Approach On Breast Cancer Data

In the Chapter 6 we have intensively studied the performance and the results of our developed *HMM* approach on breast cancer microarray data from Pollack *et al.* [23]. We have analysed the supporting information to the study of Pollack *et al.* [23] and so we have been able to create a list of candidate genes for over-expression which is shown in the Table 5.2. All these candidate genes for over-expression have been annotated as over-expressed by a standard homogeneous

HMM and an inhomogeneous *HMM* with two coupled transition matrices (extended *HMM*). The standard *HMM* has done an annotation error for the gene *ABCA5*, but the extended *HMM* has annotated this gene as over-expressed in an other gene expression profile where this gene has a high log ratio. The general view on the annotation attributes of the candidate genes for over-expression shows that these attributes are more characteristic in the annotation results of the extended *HMM*. This can be seen in the associated Tables 6.1 and 6.8. The extended *HMM* is able to improve the segment structures in gene expression profiles in comparison with the standard *HMM* and that has led to these results. Therefore we can report:

- Our *HMM* approach is able to detect in literature known candidate genes for over-expression.

- The extended *HMM* which makes use of the proximity effects between genes has improved the quality of the annotation attributes for over-expressed genes.

In addition to the candidate genes for over-expression from the study of Pollack *et al.* [23] we have found other differentially expressed genes which can play a role in subtypes of breast cancer. An interesting observation is that the works which discovered these genes have been published at a later time as the study of Pollack *et al.* [23]. Let us first review these genes.

- *KRT17* encodes the type I intermediate filament chain keratin 17. Rijn *et al.* [26] have found that the expression of *KRT17* is associated with a poor clinical outcome of breast cancer.

- *CLDN7* encodes a protein which is involved in the formation of tight junctions between epithelial cells. Kominsky *et al.* [15] have found that the loss of *CLDN7* correlates with the histological grade of in situ and invasive ductal carcinomas of the breast. They have also described the potential role of *CLDN7* in the progression and ability of breast cancer cells to disseminate. The expression of *CLDN7* is lower in invasive ductal carcinomas of the breast than in normal breast epithelium.

- *LGALS9* encodes a galectin which is implicated in modulating cell-cell and cell-matrix interactions. Irie *et al.* [10] have found that *LGALS9* is a possible prognostic factor with antimetastatic potential in breast cancer. They have observed that tumours with a low expression level of *LGALS9* do not form tight clusters during the in vitro proliferation.

- *TIMP2* encodes a protein which is a member of the TIMP gene family. This protein is an inhibitor of matrix metalloproteinases and supresses directly the proliferation of endothelial cells. The gene product of *TIMP2* is involved in the maintenance of a tissue. Nakopoulou *et al.* [18] have found that *TIMP2* is involved in the degradation of extracellular matrix which leads to the invasion of cancer into the surrounding matrix. The basis of this study have been breast cancer cells.

The annotation attributes of these genes are significant and, as before, the extended *HMM* shows better results. For the gene *KRT17* this can be seen in the Tables 6.2 and 6.9 and the other genes are shown in the Tables 6.4 and 6.11.
For our method we can report:

- Our *HMM* approach has the potential to detect novel candidate genes for over-expression and under-expression.

Now we summarise the main results for the analysis of breast cancer *ArrayCGH* data using an *HMM* with two coupled transition matrices. The annotation results have intensively been studied in the Chapter 6.

We have analysed the *ArrayCGH* data of chromosome 17. Losses and gains of DNA segments on this chromosome are reported in several publications. The main results of some interesting publications have been presented in the Section 5.3. The fact that these studies have mostly been made on the basis of other or only on a subset of our breast cancer cell lines and breast cancer primary tumours makes it difficult to directly compare the results of these studies with our annotation results. Nevertheless, we have the possibility to look for trends in our *ArrayCGH* data and so we can determine where losses and gains of DNA segments in our data are mainly located.

On the p-arm of chromosome 17 we have mainly observed losses of genes and this follows the study of Orsetti *et al.* [21]. The overview of the annotation results for the *ArrayCGH* data in the Figure 6.18 shows this observation. The situation on the q-arm of chromosome 17 is different. This arm is affected by amplifications and deletions of genes. Each of the chromosomal bands 17q11.2, 17q12, 17q23.2 and 17q25.1 has been annotated more than forty times to have an increased DNA copy number status. Most of the amplified genes are located in the band 17q12 which has been annotated one hundred twelve times to have an increased DNA copy number status. Deletions of genes are mainly located in the regions 17q11.2 and 17q12. Losses of DNA segments play only a role in a subset of our *ArrayCGH* profiles. The annotation results for the q-arm of chromosome 17 are supported by the publications in the Section 5.3. For more information about the annotation results for the *ArrayCGH* data we refer to the Section 6.2. For our developed method we can report:

- Our *HMM* approach is able to detect known regions which are affected by amplifications and deletions of DNA segments.

In summary, *HMM*s with coupled transition classes make use of the chromosomal locations as additional input data. These *HMM*s improve the detection of candidate genes for over-expression and under-expression in comparison with the standard *HMM*s. Known regions of amplifications and deletions can be detected with this novel model class. This observation could lead to a better characterisation of chromosomal imbalances in newer *ArrayCGH* studies which have higher chromosomal resolutions.

7.3 Outlook

We have developed *HMM*s with coupled transition matrices which make use of the chromosomal locations of genes to achieve an improved characterisation and detection of chromosomal imbalances and gene expression alterations. The chromosomal locations of genes are only one source of information which can be included to improve the quality of the annotation results. A future source of information could be detailed information about the locations of chromosomal breakpoints and therewith it should be possible to model specific hotspots of mutations.

The combination of gene expression and DNA copy number data to analyse microarray profiles or the clustering on the *Viterbi Paths* of microarray profiles are possible future developments on the basis of our novel *HMM* approach. The combination of both types of data, the gene expression data and the DNA copy number data, could give us a better understanding of the influences of chromosomal imbalances on the gene expression levels of genes. In the Section 6.3 we have discussed such effects and the Figure 6.19 shows such influences which have been observed for breast cancer microarray data. The Clustering on the *Viterbi Paths* could lead to improvements in the taxonomy of cancer.

Our developed *HMM* approach could play a role in other fields of bioinformatics or computer science where the integration of additional information is required.

Bibliography

[1] J. Bilmes. A Gentle Tutorial on the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, 1998.

[2] Jennifer B. Campbell. Breast cancer-race, ethnicity, and survival: a literature review. *Breast Cancer Research and Treatment*, 74:187–192, 2002.

[3] Jeremy Clark, Sandra Edwards, Megan John, Penny Flohr, Tony Gorden, Karine Maillard, Ian Giddings, Carolanne Brown, Azadeh Bagherzadeh, Colin Campbell, Janet Shipley, Richard Wooster, and Colin S. Cooper. Identification of Amplified and Expressed Genes in Breast Cancer by Comparative Hybridization onto Microarrays of Randomly Selected cDNA Clones. *Genes, Chromosomes and Cancer*, 34:104–114, 2002.

[4] David J. Duggan, Michael Bittner, Yidong Chen, Paul Meltzer, and Jeffrey M. Trent Trent. Expression profiling using cDNA microarrays. *Nature Genetics*, 21:10–14, January 1999.

[5] R. G. Dumitrescu and I. Cotarla. Understanding breast cancer risk - where do we stand in 2005? *Journal of Cellular and Molecular Medicine*, 9(1):208–221, January 2005.

[6] Richard Durbin, Sean Eddy, Anders Krogh, and Graeme Mitchision. *Biological sequence analysis - Probabilistic models of proteins and nucleic acids.* Cambridge University Press, 1998.

[7] Cathie Garnis, Timon P. H. Buys, and Wan L. Lam. Genetic alteration and gene expression modulation during cancer progression. *Molecular Cancer*, 22:3–9, March 2004.

[8] E. Gebhart. Comparative genomic hybridization (CGH): ten years of substantial progress in human solid tumor molecular cytogenetics. *Cytogenetic and Genome Research*, 104:352–358, 2004.

[9] Elizabeth Hyman, Päivikki Kauraniemi, Sampsa Hautaniemi, Maija Wolf, Spyro Mousses, Ester Rozenblum, Markus Ringner, Outi Monni, Abdel Elkahloun, Olli-P. Kallioniemi, and Anne Kallioniemi. Impact of DNA Amplification on Gene Expression Patterns in Breast Cancer. *Cancer Research*, 62:6240–6245, November 2002.

[10] A. Irie, A. Yamauchi, K. Kontani, M. Kihara, D. Liu, Y. Shirato, M. Seki, N. Nishi, T. Nakamura, H. Yokomise, and M. Hirashima. Galectin-9 as a prognostic factor with antimetastatic potential in breast cancer. *Clinical Cancer Research*, 11(8):2962–2928, April 2005.

[11] A. Kallioniemi, O. P. Kallioniemi, D. Sudar, D. Rutovitz, J. W. Gray, F. Waldman, and D. Pinkel. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, 258:818–821, 1992.

[12] O. P. Kallioniemi, A. Kallioniemi, J. Piper, J. Isola, F. Waldman, J. W. Gray, and D. Pinkel. Optimizing comparative genomic hybridization for analysis of dna sequence copy number changes in solid tumors. *Genes Chromosomes Cancer*, 10:231–243, 1994.

[13] Päivikki Kauraniemi, Maarit Bärlund, Outi Monni, and Anne Kallioniemi. New Amplified and Highly Expressed Genes Discovered in the ERBB2 Amplicon in Breast Cancer by cDNA Microarrays. *Cancer Research*, 61:8235–8240, November 2001.

[14] Bernhard Knab. *Erweiterungen von Hidden-Markov-Modellen zur Analyse ökonomischer Zeitreihen.* PhD thesis, Universität Köln, 2000.

[15] S. L. Kominsky, P. Argani, Korz D., E. Evron, V. Raman, E. Garrett, A. Rein, G. Sauter, O. P. Kallioniemi, and S. Sukumar. Loss of the tight junction protein claudin-7 correlates with histological grade in both ductal carcinoma in situ and invasive ductal carcinoma of the breast. *Oncogene*, 22(13):2021–2033, April 2003.

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

[17] Outi Monni, Maarit Bärlund, Spyro Mousses, Juha Kononen, Guido Sauter, Mervi Heiskanen, Paulina Paavola, Kristiina Avela, Yidong Chen, Michael L. Bittner, and Anne Kallioniemi. Comprehensive copy number and gene expression profiling of the 17q23 amplicon in human breast cancer. *PNAS*, 98(10):5711–5716, May 2001.

[18] Lydia Nakopoulou, Ioanna Tsirmpa, Paraskevi Alexandrou, Androniki Louvrou, Constantine Ampela, Sophia Markaki, and Panayiotis S. Davaris. MMP-2 Protein in Invasive Breast Cancer and the Impact of MMP-2/TIMP-2 Phenotype on Overall Survival. *Breast Cancer Research and Treatment*, 77(2):145–155, January 2003.

[19] Melanie Nugoli, Paul Chuchana, Julie Vendrell, Beatrice Orsetti, Lisa Ursule, Catherine Nguyen, Daniel Birnbaum, Emmanuel JP Douzery, Pascale Cohen, and Charles Theillet. Genetic variability in MCF-7 sublines: evidence of rapid genomic and RNA expression profile modifications. *BMC Cancer*, pages 3–13, April 2003.

[20] Beatrice Orsetti, Frank Courjal, Marguerite Cuny, Carmen Rodriguez, and Charles Theillet. 17q21-q25 aberrations in breast cancer: combined allelotyping and cgh analysis reveals 5 regions of allelic imbalance among which two correspond to DNA amplification. *Oncogene*, 18:6262–6270, November 1999.

[21] Beatrice Orsetti, Melanie Nugoli, Nathalie Cervera, Laurence Lasorsa, Paul Chuchana, Lisa Ursule, Catherine Nguyen, Richard Redon, Stanislas du Manoir, Carmen Rodriguez, and Charles Theillet. Genomic and Expression Profiling of Chromosome 17 in Breast Cancer Reveals Complex Patterns of Alterations and Novel Candidate Genes. *Cancer Research*, 64:6453–6460, September 2004.

[22] C. M. Perou, T. Sorlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, Oystein Fluge, Alexander Pergamenschikow, Cherge Williams, Sherley X. Zhu, Per E. Lenning, Anne-Lise Berresen-Dale, Patrick O. Brown, and David Botstein. Molecular Portraits of human breast tumours. *Nature*, 406:747–752, August 2000.

[23] Jonathan R. Pollack, Therese Sorlie, Charles M. Perou, Christian A. Rees, Stefanie S. Jeffrey, Per E. Lonning, Robert Tibshirani, David Botstein, Anne-Lise Borresen-Dale, and

Patrick O. Brown. Microarray analysis reveals a major direct role of Dna copy number alteration in the transcriptional program of human breast tumors. *PNAS*, 99(20):12963–12968, October 2002.

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

[25] Jörn Tödling, Sebastian Schmeier, Matthias Henig, Benjamin Georgi, and Stefan Röpcke. MACAT - microarray chromosome analysis tool. *Bioinformatics*, 21:2112–2113, 2005.

[26] Matt van de Rijn, Charles M. Perou, Rob Tibshirani, Phillippe Haas, Olli Kallioniemi, Juha Kononen, Joachim Torhorst, Guido Sauter, Markus Zuber, Ossi R. Köchli, Frank Mross, Holger Dietrich, Rob Seitz, Doug Ross, David Botstein, and Pat Brown. Expression of Cytokeratins 17 and 5 Identifies a Group of Breast Carcinomas with Poor Clinical Outcome. *The American Journal of Pathology*, 161:1991–1996, December 2002.

[27] Simon Willis, Anne-Marie Hutchins, Fleur Hammet, John Ciciulla, Wee-Kheng Soo, David White, Peter van der Spek, Michael A. Henderson, Kurt Gish, Deon J. Venter, and E. Jane Armes. Detailed Gene Copy Number and RNA Expression Analysis of the 17q12-23 Region in Primary Breast Cancers. *Genes, Chromosomes and Cancer*, 36:382–392, 2003.