INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

KTH Computer Science and Communication

A Hidden Markov Model for Cancer Progression

Miguel Antunes Dias Alfaiate Simões

Dissertação para a obtenção do Grau de Mestre em*

Engenharia Electrotécnica e de Computadores

Júri

Presidente:	Prof. Carlos Silvestre
Orientador†:	Prof. Luís Borges de Almeida
Vogal:	Prof. João Paulo Neto

Julho de 2010

*Tese feita no Stockholm Bioinformatics Center, Kungliga Tekniska högskolan, Suécia
†Co-orientador: Prof. Jens Lagergren

I am very grateful for the cooperation of everybody who helped me on this thesis and throughout my academic life, especially my supervisor Jens Lagergren for his support and for offering me the opportunity of developing this interesting project. I have also learnt a great deal from Professor Luís Borges de Almeida over the years and gratefully acknowledge my debt to him.

This work was performed at the Stockholm Bioinformatics Center (SBC) in the context of the ERASMUS programme and I would like to express my gratitude to the SBC and especially to Hossein Farahani, since this project would not have been possible without his help.

I must also express my thanks to my friends in Portugal and in Stockholm, who made this one of the most special years of my life.

Finally, I would also like to thank my family for their support in my life decisions and for the education they provided me.

Abstract

Cancers are characterized by chromosomal aberrations which, particularly in solid tumors, appear in complex patterns of progression. The characterization of the timing and the order of the genetic mutations that drive tumor progression is difficult due to this high level of complexity.

For medical treatment reasons, it is important to understand how these patterns develop and several models have been proposed. The first ones were of a more descriptive nature, but since then several attempts have been made to obtain mathematical models. In this work, a generative probabilistic model based on Hidden Markov Models (HMMs) is presented and provides a natural framework for the inclusion of unobserved or missing data. The number of parameters is also reduced and the inference algorithm used to estimate them is the Expectation–Maximization (EM) algorithm (earlier algorithms have been rather straightforward heuristics).

The tests performed in a synthetic data generator also developed in this work (which tries to recreate the behavior of real cancers) seem to show that the algorithm infers the hidden parameters with high accuracy.

Keywords: cancer, chromosomal aberration, probabilistic model, Hidden Markov Models (HMM), Expectation–Maximization (EM) algorithm.

Resumo

O cancro caracteriza-se por aberrações cromossómicas que, particularmente em tumores sólidos, progridem em complicados padrões. Fazer a descrição temporal das mutações genéticas que levam ao aparecimento de um cancro é difícil devido ao seu elevado nível de complexidade.

Devido a razões de tratamento médico, é importante perceber como estes padrões se desenvolvem e diversos modelos têm sido propostos. Os primeiros eram de uma natureza mais descritiva mas, mais tarde, diversas tentativas foram feitas para obter modelos matemáticos. Neste trabalho, um modelo generativo baseado em Modelos de Markov Escondidos (HMM) é apresentado, providenciando a possibilidade de incluir dados não observados ou ausentes. O número de parâmetros é também reduzido e o algoritmo de inferência usado foi o algoritmo Esperança–Maximização (EM) (os algoritmos anteriores caracterizavam-se por serem mais heurísticos).

Os testes realizados num gerador sintético de dados também desenvolvido neste trabalho (que tenta recriar o comportamento de cancros reais), parecem mostrar que o algoritmo infere os parâmetros bastante bem.

Palavras-chave: cancro, mutações genéticas, modelo probabilístico, Modelos de Markov Escondidos (HMM), Algoritmo Esperança–Maximização (EM).

Contents

List of Figures

List of Tables

List of Abbreviations

DAG	Directed Acyclic Graph
EM	Expectation-Maximization
HMM	Hidden Markov Model
NIPT	Number of cytogenetic Imbalances Per Tumor
TO	Time of Occurence

Chapter 1

Introduction

Being one of the deadliest diseases in the world, cancer remains one of the most studied topics in science. However, an effective treatment is still quite far away, mainly because cancer is not just a single disease. There are many different cancers that originate from different defects and a treatment that is effective for one may be completely ineffective for other.

On the other hand, particularly in solid tumors, the chromosome changes that drive the origin of cancer, albeit being in a large number, are generally not tumor-specific. The way cancer progresses is characterized by accumulating mutations, so by studying how this process develops, it becomes possible to predict and to devise better medication and treatments.

Several models have been suggested in the literature. The first ones were of a more descriptive nature, but since then several attempts have been made to obtain mathematical models. Given that the way cancer occurs and develops is highly complex, a model for cancer progression is still an open problem. In this report, a new model is proposed, built on a previous one by Hjelm *et al.* [1]. This new model takes missing data from erroneous measures into account and provides a more rigorous method of detecting the parameters which characterize how the different mutations progress.

Overview of this report

- **Chapter 2**

  An introduction to the biological aspects of cancer progression is presented.

- **Chapter 3**

  A literature review and a discussion of previous results are discussed, motivating the choices made about the model chosen to represent cancer progression.

- **Chapter 4**

  This chapter presents an introduction to Hidden Markov Models (HMMs) and a brief description of the Expectation-Maximization (EM) algorithm.

- **Chapter 5**

  The model used and developed in this project is discussed, as well as the innovations introduced.

- **Chapter 6**

  The description of the theory that combines the model presented with our new optimization methods, as well as its computer implementation, is made here.

- **Chapter 7**

  The synthetic data generator is introduced and several simulations are performed to test the method. One improvement regarding lack of data in some tests is also studied.

- **Chapter 8**

  Here, the project in general and the results obtained are discussed, with some ideas for future work.

Part I

Background

Chapter 2

Biological Background

2.1 Cancer: an Introduction

According to the World Health Organization, cancer accounted for 13% of all deaths worldwide in 2004 (7.4 millions). These deaths are projected to continue rising, with an estimated 12 million deaths in 2030 ([2]).

One of the biggest concerns with cancer is that it is not a single disease but in fact a class of diseases, which means that treatments effective in one type might not be effective in controlling another one; it can also affect any part of the body. Cancer starts when several cells lose their regulatory mechanisms controlling growth. **Neoplasm** (from the Greek *new growth*) is another name given to cancer (considered to be more accurate), referring to the deregulated multiplication of these cells. In some cases, these cells will eventually form a **tumor**. If the abnormal cells grow beyond their usual boundaries and invade other parts of the body, the tumor is said to be **malignant** and this process is called **metastasis**. Contrary to a **benign** cancer, this one is said to be life threating (with lung cancer to be the most common one). For the cells to suddenly lose control of their growth, the genetic material must suffer some sort of mutation, which can be caused by several agents that interact with a person's genetic material: carcinogens like tobacco smoke can account for 30% of cancers, while another 30% might be diet related ([3]). Some genetic abnormalities that potentially induce cancer can also be inherited, which might explain why some cancers can occur repeatedly in the same family.

A cancer is usually triggered by a **mutagen** (an agent which modifies the genetic material of an organism) but the exposure to more mutagens is required for the cancer to progress. As such, one can say that cancer progression is a multi-step process to transform normal human cells into malignant cancers. The idea that cancer progresses through different stages was proposed in [4], in which several characteristics of the stage progression were analyzed. One of those was that "progression follows one of alternative paths of development", referring to the fact that there can be more than one way for the cancer to progress. Research over the past decades has revealed that a small number of molecular, biochemical and cellular capabilities are

shared by most and perhaps all different types of human cancer ([5]).

Since there are so many different types of this disease, it is of the utmost importance to try to understand the mechanism that lies behind all of them. There are two classes of genetic aberrations that lead to cancer. **Oncogenes** are genes that promote cancer. For example, consider an oncogene whose original function is to code for the proteins that control cell division. Once this oncogene mutates, the division process is disrupted and the cell becomes cancerous, with uncontrolled growth and division. The opposite of these genes are the **tumor suppression** ones, which are inactivated in a cancer. These genes are, for example, involved in the detection of abnormal genetic material and induce **apoptosis** (cell death) if they detect any error. When they are not working properly, genetic errors are accumulated from one cell generation to another. Simply put, an oncogene increases the probability of a cancer to occur when there is an increase of expression of its protein. The opposite is true for tumor suppressor genes.

The aforementioned defects can be grouped in seven categories which will be analyzed briefly in the following section. Research shows ([3] and [5]) that probably all these conditions must be met in order for the cancer to develop, which possibly explains why cancer takes many years to develop and is more common in older people.

2.2 Defects Associated to Cancer

2.2.1 Abnormal Signaling Pathways

Cell growth and division are induced by signals sent from the exterior of the cell. Defects in this process induce the cell to constantly multiply. One prominent example is the *ras* gene which codes for a protein involved in cell division. In normal conditions, this protein has the ability of being disabled. However, when mutated, this ability is not available anymore and, thus, cell division is continuous. According to [3], defects in this gene are present in 20 %–30 % of all human cancers.

Another type of defect can be in the receptor of the signal, which can be overly activated by a normal signal.

2.2.2 Insensitivity to Growth-inhibitory Signals

This is an example of tumor suppression genes: a cell usually receives signals that inhibit the growth of a cell. After mutations on the receptors that are supposed to respond to these signals, the cell can become insensitive to growth-inhibitory signals.

2.2.3 Abnormalities in Cell Cycle Regulation

During growth and multiplication, a cell has different phases in this process that must be accomplished. These are related to the necessary operations to divide a

cell into two daughter cells and the transition from one phase to another is signaled by proteins that can also be disrupted.

2.2.4 Evasion of Programmed Cell Death

The apoptosis process can either be induced by the cell itself (as described earlier) or by external signals (for example, by the immune system when searching for damaged cells). In order for the cancer to progress, these mechanisms must be bypassed.

2.2.5 Limitless Replicative Potential

When a cell replicates it also replicates its DNA, which has a limited number of times that can be replicated. This also operates as a way to reduce errors in cell division and is yet another process that must be corrupted in order for cancer to develop. Telomeres are the name given to a region in the end of a chromosome which is shortened every time the cell replicates. Eventually, this region is too short and the cell is unable to replicate from now on. For the cell to be immortal, thanks to a mutation in the telemore which makes it to never get smaller, this cell life limitation disappears.

2.2.6 Angiogenesis

Angiogenesis refers to the sustained capacity to develop new blood cells. After several cell divisions, the cancer is now big enough for it to need several resources to continue growing, like oxygen and carbohydrates. This resources can be provided by new blood vessels. The growth of these new blood vessels is a process present when a new tissue is formed and that is not supposed to be available to cancer cells. They have to, somehow, develop this capacity in order to progress, which will also damage the surrounding tissue.

2.2.7 Tissue Invasion and Metastasis

So far, the tumor is said to be benign since it is localized in a distinct part of the body. In order to be death threatening, it must be able to form secondary tumors in other parts of the body. This is done by cancerous cells that travel through the blood circulation system.

All the defects referred seem to make cancer a highly unprobable disease, since there are so many process in a cell that monitor cell division and the existence of genetic aberrations. Yet it still has a quite substantial frequency in the human population, which suggests that cells must acquire some sort of increased mutability for the cancer progression to proceed. In [6], the existence of a mutator phenotype which would increase the mutation rate in a cell was suggested.

2.3 Cancer as a Multi-step Progress

Tumor progression can be described by a multi-step process by means of the accumulation of complex chromosome alterations for several types of cancers ([7] and references therein). As such, it is important to relate DNA mutations to the processes described in the last section. Linking the molecular events that are beneath the progression of cancers is an important area of research which faces several challenges ([8]). For example, an aberration in the genetic material of the cell usually leads the cell to more changes, which makes the distinction between the first and last events difficult to make. As it is usually hard to have samples of the same cancer over time, the order of these events must be estimated by methods able to infer this progression by means of a single sample. Another problem is that two cancers, in spite of looking clinically homogeneous, can be very different genetically. The last problem is that it is difficult to collect samples from tumors in a large scale, making analysis necessary to be done in a not so large number of samples (which is usually done by means of cross-sectional data[1]).

Chromosomal abnormalities in cancer were first discovered in a form of leukemia by Nowell *et al.* (1976, [9]), in which a model for cancer progression based on this abnormalities was suggested. Here, it is proposed that cancer start with a single cell, suffering from then on several aberrations, in which the majority of them are not viable. Compared to solid tumors, leukemia proves to be easier to study. The reason for this is that, in solid tumors, once a set of critical genetic alterations is developed (a set of primary disease-causing events), the cancer cell starts to accumulate apparently random alterations, making it hard to identify which ones are the first and which ones did not induce cancer ([10]).

A pioneering work in this field was the one of Vogelstein *et al.* (1988; [11]). Studying colorectal tumors, which are easy to study since they have a comprehensive database of samples, they tried to establish a connection between genetic aberrations and the **tumorigenesis** process (see next chapter for more information about this model). After that, more complex models have arisen which also debated the question of how to detect genetic aberrations. For example, Desper *et al.* (1999; [10]) have used a laboratory technique called *comparative genomic hybridization*, commonly called CGH. When comparing a normal cell with a cancerous one, it is possible to realize if there are chromosomal regions that have significant gains (or losses) of genetic material. This would mean the region affected can have an impact on tumor progression. They then propose a model for cancer progression based on this set of aberrations.

The investigation of these sets of data from tumors have been used to discover patterns and study the way a cancer progresses. Several models have been proposed which analyze this multi-step progress and a brief literature review of these methods will be the focus of the next chapter.

[1]Cross-sectional data is the name given to data collected from several individuals without regarding time differences between them.

Chapter 3

Literature Review

As seen in Chapter 2, research in cancer genetics has identified several steps that can be related to a tumor formation and progression. In [11], the model of the colon cancer was the first work which tried to establish a link between genetic aberrations and changes in the cancer itself (like the beginning of the metastasis process). They also realized that these aberrations did not need to occur in an invariant order and that it was their accumulation that seemed to be more important; cancer progression seemed to be characterized by the combined effect of several aberrations (see also [5]). Later, it was suggested in [12] that linear models similar to this one could not represent oncogenesis exactly for other type of cancers.

The purpose of this branch of cancer research is to find which genetic mutations tend to occur early or later and how they influence each other. There are two types of approaches that have been followed: the first is formed by the so-called **narrative** models, in which cancer progression is described in a more qualitative, handmade way. As an example, see [13].

The other approach is done in a more quantitative, mathematical way. The mathematical approach of cancer began to be studied in the 1950s when trying to explain incidence curves that relate cancer and age (see [14] and references therein). This led to the idea that cancer progression might be dependent by a set of probabilistic events and could be modeled as such.

After the already referred first model, several attempts have been made to analyze cancer progression. Desper *et al.* expanded this first chain model to a tree one in [10], proposing what they later called a *branching tree*. This is a tree with a root that represents an healthy cell and other nodes representing events (set of aberrations). The edges between then would have a probability assigned to them and this would signal that the occurrence of the first one would influence the occurrence of the other. Events occurring earlier will be closer to the root and those events which group together in a specific branch can signal a tumor subgroup. The method used to reconstruct the tree with the cancer data was a maximum-weight branching algorithm. In [8], the *distance-based trees* were implemented: they are basically the same as the previous ones, with the difference that now only the leafs

are aberrations and the middle nodes are considered to be hidden states. The reason for this is to have the flexibility of having events that might not be represented in the data (for example, an aberration occurred and was not detected for some reason). The method to reconstruct the tree (fitting a *distance matrix* to a tree) is also different, as this model tries to take advantage of the several methods already developed in the study of phylogenetical trees.

The goal of these models is to represent the first-order dependencies between one pair of events. This is obvious with a tree with one root and two leafs, but for larger trees this dependency is harder to model, since it is not possible to represent all the pairwise relationships exactly. Another problem with these two last models is that they cannot represent pathway convergence: having two different aberrations that, together, may lead to another one is not possible to model, for example.

Beerenwinkel *et al.* ([15]) proposed, modeling HIV evolution, a mixture of trees which can be used to solve these two problems. In this model, one of the trees is modeled to be a star in which all events are connected to the root with the same probability. This star represents independence of the nodes, as the events are connected to the root and not between themselves. This model has also been used, for example, to try to predict the time of death, for example, by means of a *genetic progression score* in [7]. Since it is important to measure how far a cancer has already progressed (for therapy reasons), this score gives an indication of this. The method uses an EM-like algorithm which was not proven to deliver local optima.

However, with these innovations, it is still impossible to have an aberration to occur as a child if not all its parents aberrations had also occurred – the so-called **monotonicity** problem. A model with Directed Acyclic Graphs (DAGs) has been proposed by [16] which, in terms of edges and vertices, works basically in the same way as the trees already discussed. Connected to the root (healthy cell), every single aberration is represented by its node. In the second layer, two aberrations can get together in the same node, having two edges that indicate the probability of going from one node to the other. In the subsequent layers, the process is the same. Note that this method can also be seen as a Markov Model (Markov Models will be discussed in the next chapter).

Since the number of parameters to estimate in the DAG starts to become very large, several assumptions are made to reduce them. In this model, it was considered that only an aberration could influence another if the later was the next one to occur. [17] extended this in order to include later time events also. They do this by introducing pairwise dependency between a limited number of aberrations which is not confined to the next event as before. Hjelm *et al.*, in [1], introduced the Network Aberration Model in which modules of events are also possible (in order to reduce the parameter complexity). These modules cluster variables that share some properties; for example, two nodes that have the same pairwise dependency to a third node, could be grouped in a module.

Another attempt to solve the monotonicity problem was made in [18] with the use of hidden-variable trees, where, associated to each vertex there is an observable and a hidden variable. The later indicates that tumor progression has reached a

vertex and the observable variable represents the detection of an event. This method uses a structural EM-algorithm to find these variables.

The models discussed so far are generative models in the sense that the model proposed can be used to generate synthetic data. Another class of generative methods uses conjunctive Bayesian networks which also allow to have multiple parent nodes (see [19] and references therein).

Other methods have been proposed which rely on tools that try to analyze data correlation. Principal component analysis has been made in several cases as [20], in which they also devise a way to try to see when an aberration occurs by means of a simple statistical method (which will be analyzed on chapter 6.1).

Chapter 4

Theoretical Background

4.1 Hidden Markov Models

4.1.1 Introduction

Characterizing the behavior of real-world objects and processes in terms of models is a problem of great interest in several fields of engineering. Processes usually produce observable output—a signal—and these models are theoretical descriptions of them.

Modeling a process serves several objectives: for example, knowing a theoretical description of a process allows us to predict its behavior as well as how to make it provide a desired output. To better describe an object's behavior, two classes of models are available: **deterministic** models usually take advantage of the fact that the output signal has some known properties that are deterministic (like being a sine wave). This means that there is no randomness involved and, by means of the model's parameters and its previous states, it is always possible to determine its current state. It is then only necessary to estimate the parameters that characterize this process (for example, the frequency of the sine wave). **Statistical** models describe a process in terms of its random variables, i.e., its statistical properties. As an example, consider the mean that describes a process assumed to be Gaussian. Hidden Markov Models (HMM) are included in this class.

HMM theory has been applied extensively in certain fields as speech recognition. This chapter will be partly based on [21], a review about HMMs and their application to this field. Since the focus here is more on biological research, [22] will also be used. Note that not all aspects of HMMs will be discussed here, as the focus will be on the ones that matter to the rest of this work.

Before introducing Hidden Markov Models, Markov models of the non-hidden variety should also be reviewed.

4.1.2 Markov Chain Models

Considering a system which can be described by a set of N states, $\{q_t\} = q_1, q_2, ..., q_N$; these states usually represent a set of parameters and the system is said to be in a

particular state when having the characteristics of this state. It is possible to change to another state or to continue in the same one and, for each of these transitions, there is a probability associated.

$$\begin{tikzcd} A \arrow[r, "ab"] \arrow[d, "ad"'] & B \arrow[d, "bc"] \\ D \arrow[r, "dc"'] & C \end{tikzcd}$$

This example consider 4 different states with $(q_1 = A, q_2 = B, ...)$ and each edge represents the transition between them. The sequence of different states can be finite or infinite (for example, A, B, C) and we will denote the present state as x_t. Usually, the index t is thought as a time index. To be able to describe a system like this, it is necessary to account for the current state as well as the previous ones. These models enjoy the so-called Markov property; this means that the probabilistic behavior of a Markov chain of states depends only on the dependencies between successive states, i.e., between x_1 and x_2, x_2 and x_3, etc... We call this first-order dependency; if a state depended on the last two ones, this would be a second-order dependency, and so on...

The conditional probability that describes the system's evolution up to the present time is given by

$$\Pr(x_t = q_t | x_{t-1} = q_{t-1}, ..., x_1 = q_1) \tag{4.1}$$

which, according to the Markov property, is equal to

$$\Pr(x_t = q_t | x_{t-1} = q_{t-1}) \tag{4.2}$$

This corresponds, then, to the transition probability (from state to state) already mentioned and is usually denoted as $a_{q_t, q_{t-1}}$. As it is possible to see, the present depends only on the immediate past. Note that $a_{q_t, q_{t-1}}$ can also depend on time (for example, the transition from one state to the other has different probabilities if we are in the beginning of the experience or in the end), but here will be assumed to be independent. As such, this Markov Chain is said to be *time-homogeneous*.

Considering a generic transition from a state q to q' in which $q, q' \in \{q_t\}$, the following properties apply (which are the usual in a stochastic process):

$$a_{q,q'} \geq 0 \tag{4.3}$$

$$\sum_{q'=q_1}^{q_N} a_{q,q'} = 1 \tag{4.4}$$

One can view a Markov chain as a path in time through its state space [23]. For a sequence $\{q_t, q_{t-1}, ..., q_1\}$, its probability is given by

$$\begin{aligned}\Pr(x_t = q_t, &x_{t-1} = q_{t-1}, ..., x_1 = q_1) \\ &= \Pr(x_t = q_t | x_{t-1} = q_{t-1}, ..., x_1 = q_1) \\ &\times \Pr(x_{t-1} = q_{t-1}, ..., x_1 = q_1) \end{aligned} \tag{4.5}$$

since, by the definition of conditional probability, $\Pr(A, B) = P(A|B)P(B)$. Applying recursively the same procedure:

$$\begin{aligned}\Pr(x_t = q_t, &x_{t-1} = q_{t-1}, ..., x_1 = q_1) \\ &= \Pr(x_t = q_t | x_{t-1} = q_{t-1}, ..., x_1 = q_1) \\ &\times \Pr(x_{t-1} = q_{t-1} | x_{t-2} = q_{t-2}, ..., x_1 = q_1) ... \Pr(x_1 = q_1) \end{aligned} \tag{4.6}$$

and, applying again the Markov property,

$$\begin{aligned}\Pr(x_t = q_t, &x_{t-1} = q_{t-1}, ..., x_1 = q_1) \\ &= \Pr(x_t = q_t | x_{t-1} = q_{t-1}) \Pr(x_{t-1} = q_{t-1} | x_{t-2} = q_{t-2}) ... \Pr(x_1 = q_1) \\ &= a_{q_{t-1}, q_t} a_{q_{t-2}, q_{t-1}} ... \Pr(x_1 = q_1) \end{aligned} \tag{4.7}$$

Note that $\Pr(q_1)$ refers to the initial distribution of the Markov chain and the initial probabilities of each state are usually denoted as π_q. Note also that the initial distribution and the transition probabilities—which are usually organized in a $N * N$ matrix, $A = \{a_{q,q'}\}$, with the probabilities from one state to all the others explicit in each row—are enough to fully characterize a Markov chain.

4.1.3 HMMs

Until now, it was assumed that in each state there was always an observable event that allowed to see in which state the system was—the sequence of states $\{x_t, x_{t-1}, ..., x_1\}$. This cannot always be true for several reasons (due to observation erros, for example). As such, we should now assume that, in each state, there is a stochastic process that models this observation event. The resulting model, which is called a Hidden Markov Model is, then, a pair of process: one to model the states' progression (which is hidden) and other to model the observation process. The only way to access the first process is through the stochastic process of observations produced by the second one.

The classical example to introduce HMMs is the so-called Occasionally Dishonest Casino. In this example, consider that two (or more) dice are available. One is a fair die and the other is weighted. This means that the probabilistic distribution that describe the numbers observed after rolling the fair die can be modeled by a uniform distribution as opposed to the weighted one, in which one of the numbers can be more common than the others. In this experience, it is possible to observe the outcome of these dice, but it is not possible to know which one was rolled.

This situation can be modeled with a HMM considering that each die is a state (the aforementioned set of parameters—in this case, the distribution parameters of these two dice) which has as outcome a number. This can be represented as shown in figure 4.1.

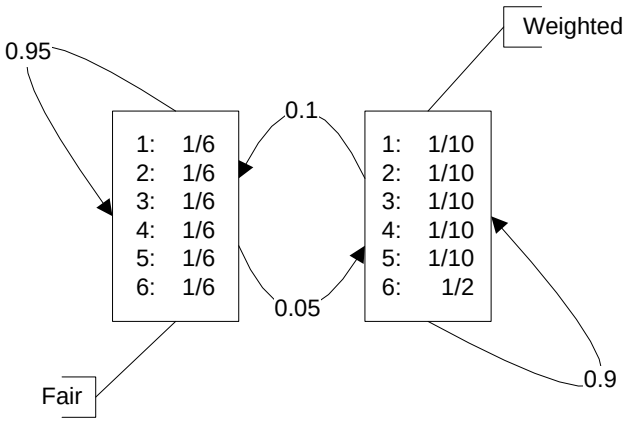


Figure 4.1: Two dice (one fair and one weighted) and their probability distributions

As it is not possible to know which die was rolled, we say that the state sequence is hidden, as opposed to the previous situation with ordinary Markov chains. The probabilities that describe each dice are often called *emission probabilities* and here will be represented by e.

To set ideas, consider a pair of processes $(\{z_t\}, \{x_t\})$ in which $\{z_t\}$ is a Markov chain (corresponding to the hidden process) and $\{x_t\}$ is a sequence of *letters* in an alphabet $\mathcal{A}$ which represents the possible outcomes of the system. These letters are emitted by the states visited during the experience. The transition probabilities, a, and the initial probabilities, π_q, work in a similar fashion to the regular Markov chains but now we also have the emission probabilities which can also be organized in a matrix, with its elements mapping each state to each letter in the alphabet $\mathcal{A}$, i.e., q $\rightarrow \mathcal{A}$, or:

$$e_q(i) \triangleq \Pr[\text{ state } q \text{ emits letter } i \text{ }], \tag{4.8}$$

considering again that q represents a state and $i \in \mathcal{A}$. With N still referring to the number of different states, we can define M as being the number of letters in $\mathcal{A}$. As such, the matrix that represents these probabilities, $E = \{e_q(i)\}$ has dimension $N * M$.

Now, when evaluating a path probability like in (4.1) we have to consider both

the transition as well as the emitted value in each state:

$$\begin{aligned}\Pr(x_t = i_t, &x_{t-1} = i_{t-1}, ..., x_1 = i_1; z_t = q_t, z_{t-1} = q_{t-1}, ..., z_1 = q_1) \\ &= a_{q_{t-1},q_t} e_{q_t}(i_t) a_{q_{t-2},q_{t-1}} e_{q_{t-1}}(i_{t-1}) ... \Pr(z_1 = q_1) e_{q_1}(i_1) \\ &= \pi_{q_1} e_{q_1}(i_1) \prod_{j=2}^{t} a_{q_{j-1},q_j} e_{q_j}(i_j) \end{aligned} \tag{4.9}$$

It is interesting to note that HMMs can also be seen as generator of observations besides being also a model for a system.

Rabiner in [21] presents three basic problems that are can be addressed by means of a HMM:

1. Given an observation sequence O and a HMM, compute the probability $\Pr(O|A, E, \{\pi_q\}_q)$ of this observation sequence in an efficient way.

2. Given an observation sequence O and a HMM, find the optimal sequence of states that better explains O.

3. Given an observation sequence O, find a model that explains the sequence, i.e., maximizes $\Pr(O|A, E, \{\pi_q\}_q)$.

The last problem is, predictably, the hardest and also the one that shall be discussed in 6.2.

4.2 Expectation-Maximization Algorithm

4.2.1 Introduction

The purpose of using the Expectation-Maximization (EM) algorithm in this thesis is to find the parameters of the HMM used to describe cancer progression. This section intends to introduce this algorithm, based on [24] and [25].

Consider that a process with a known probability distribution, p, is observed. The goal is to find the distribution's parameters, θ', that fit the observed data, $x = x_1, x_2, ..., x_K$ with K being the size of the dataset drawn from this distribution. As an example, if the distribution supposed to fit the data was a Gaussian distribution, then the parameters to estimate would be the mean and the variance.

$$\Pr(x|\theta') = \prod_{j=1}^{K} \Pr(x_j|\theta'). \tag{4.10}$$

Equation (4.10) is also called the **likelihood function**, $L(x|\theta')$, since it measures how likely the distribution's parameters describe the data. This dataset is assumed to be fixed and our goal is to maximize the parameters according to:

$$\theta'^* = \arg\max_{\theta'} L(x|\theta')$$

Computationally, it is common to have some low values for this probabilities and, as such, the log-likelihood ($l = \log[L(x|\theta']$) is usually the function that is optimized, since they will have the same maxima. This problem of finding the **Maximum Likelihood** estimate can be solved in different ways and the EM algorithm is one of them. This iterative procedure also provides a way to deal with missing or incomplete data from the observation set (sometimes, due to errors or from the system's design itself, not all the data is available). This situation is of particular interest in this project (as discussed in 5.5).

The hidden data is usually denoted by z, and x is now called incomplete data. Rewriting now the log-likelihood function considering z and x:

$$\Pr(x|\theta') = \sum_z \Pr(x, z|\theta') \tag{4.11}$$

which is the marginal distribution of $\Pr(x|\theta')$.

4.2.2 General Derivation of the EM Algorithm

As already said, the EM algorithm is an iterative process. This implies that the parameters being estimated will change in each iteration and, as such, we will denote θ as the parameters estimated in the last iteration and θ' as the new ones.

Considering that we want to solve the optimization problem discussed in the last section, we have:

$$\begin{aligned}
l &= \log[(\Pr(x|\theta')] \\
&= \log\Big(\sum_z \Pr(x, z|\theta')\Big) \\
&= \log\left(\sum_z \Pr(z|x,\theta)\frac{\Pr(x, z|\theta')}{\Pr(z|x,\theta)}\right) && (4.12)\\
&\geq \sum_z \Pr(z|x,\theta)\log\left(\frac{\Pr(x, z|\theta')}{\Pr(z|x,\theta)}\right) && (4.13)\\
&= \sum_z \Pr(z|x,\theta)\log\Big(\Pr(x, z|\theta')\Big) - \sum_z \Pr(z|x,\theta)\log\Big(\Pr(z|x,\theta)\Big) \\
&= Q(\theta',\theta) - R(\theta,\theta)
\end{aligned}$$

where Q and R are defined as:

$$Q(\theta',\theta) = \sum_z \Pr(z|x,\theta')\log\Big(\Pr(x, z|\theta')\Big) \tag{4.14}$$

$$R(\theta',\theta) = \sum_z \Pr(z|x,\theta')\log\Big(\Pr(z|x,\theta')\Big). \tag{4.15}$$

Some comments about these last steps should be made here. Note that, from (4.12) to (4.13), Jensen's inequality was used (see appendix A). The term $\Pr(z|x,\theta)$

is the conditional distribution of the hidden data and is dependent on both the observed data and on the parameters from the last iteration. The computation of this probability depends, on the problem being studied and will be discussed later in 6.3.

Moreover, we have that

$$\log[(\Pr(x|\theta)] = Q(\theta, \theta) - R(\theta, \theta) \tag{4.16}$$

and, if

$$Q(\theta', \theta) > Q(\theta, \theta) \Rightarrow \log[(\Pr(x|\theta')] > \log[(\Pr(x|\theta)]. \tag{4.17}$$

This means that finding a new Q term that is bigger than the last one will imply that the new parameters will fit the observed data in a better way, improving the likelihood of these parameters to be able to describe the dataset. The reason for introducing Q is that it is easier to maximize, as shall be clear later in section 6.2.

Note also that 4.14 can be expressed as an expected value:

$$Q(\theta', \theta) = E_z\Big[\log \Pr(x, z|\theta')\Big|x, \theta'\Big]. \tag{4.18}$$

The computation of this term is, then, called the E-step and the M-step of the algorithm consists in finding the terms that maximize Q:

$$\theta'^* = \arg\max_{\theta'} Q(\theta', \theta).$$

These two steps are the two components of one iteration and should be repeated as necessary, i.e., until some stop criterion (as very small changes from one iteration to the other).

4.2.3 Convergence of the EM Algorithm

The convergence of this algorithm would not be discussed here, as it is out of scope of this work. However, the log-likelihood has been proven to increase in each iteration making the algorithm to converge to a local maximum ([26]). Unfortunately, in most problems of interest, the function that is being optimized is very complex and can have many local maxima and we are not guaranteed to achieve the best result possible. One way to fight this is trying several initial conditions.

So far, it is not clear how this algorithm can be transformed into a sequence of computing steps required to carry out a single E- or M-step. This depends greatly on the application desired and, in this work, we will discuss how to apply it to HMMs in 6.2.

4.3 Learning the Parameters of a HMM through the EM Algorithm

The Baum-Welch algorithm (or forward-backward algorithm) was developed in order to efficiently compute the parameters estimation of an HMM by means of the EM algorithm. They are described in [22], [24] and [23]; since its implementation depends greatly in the process analyzed, this topic will be discussed later in 6.3.

Part II

Model Developed and Results

Chapter 5

Cancer Progression Model

5.1 Introduction

The model that was developed in this project will partly be based on the one described by Hjelm *et al.* in [1]. The reasons for this are also discussed in 3.

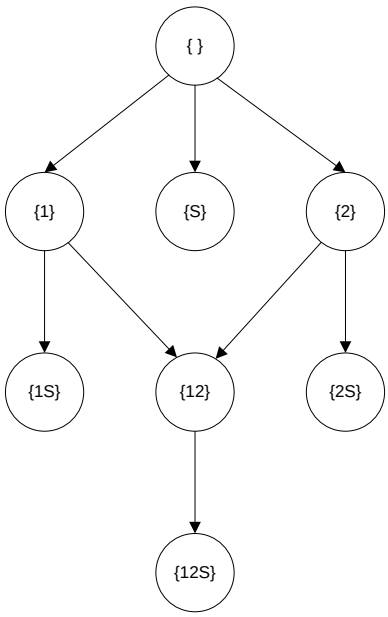


Figure 5.1: DAG representing the Markov Chain used to describe cancer progression. S marks a stop state.

To fix ideas, the model will be composed of a Directed Acyclic Graph with each node representing a state (as in a Markov Chain *state*). Each state, q, will be a set of aberrations that had already occurred and the root will be considerer an healthy state (without aberrations). This means that cancer progression is described by

successively going down this graph, i.e., starting with the root and then moving to one of the child states that denotes the occurrence of an aberration. From one state to the other only one aberration j can occur: in other words, a transition marks a mutation. The reasons for using a graph instead of a tree is to allow for two states to converge in to another one.

This model is also a generative model in the sense that it can be used to generate datasets. It also has a clear graphical representation, with each node representing an aberration state and each arrow representing the aforementioned transitions between them. This would make the use of Markov Chains' theory a possibility, since, when considering a state, only the previous one is enough to describe the system in that time instant. On the other hand, since it is possible to omit the observation of an aberration, i.e., sometimes it is impossible to see that a transition has occurred, Hidden Markov Chains shall be used.

To represent the way different aberrations interact with each other, Hjelm *et al.* also developed what they called a Network Aberration Model (NAM). This is used to compute the transition probabilities between states.

The model will be discussed in the next five sections. The first one will present some notations and definitions, followed by the analysis of cancer progression as a Markov Model and as a HMM. The section after these ones will study an extension of this model and the last one will discuss the NAM.

5.2 Definitions

Different types of genetic mutations (like chromosomal breakpoints or copy number changes for various segments) will be considered here simply as *an aberration*. In each dataset, n different aberrations can occur and $[n] = \{1, ..., n\}$ represents this set of aberrations. Since a cancer is comprised of aberrations, each sample that is studied will be a subset of $[n]$. A sample will be denoted by $D = \{d_1, ..., d_k\}$ with each d representing the aberrations occurred. After an event that will be called *stop state* and denoted q_z, the cancer is considered to be detected. This detection event simulates the cancer detection by a doctor. Note that it is not possible to know the mutations' order of occurrence (among the $k!$ possibilities), since, when testing the cancer in a laboratory, we only have access to the aberrations occurred.

5.3 Markov Model for Cancer progression

As already partly described, each node in the graph will represent a state in the Markov Chain, which accounts for a set of events that had already occurred.

Let $\{X(t) : t \geq 0\}$ be a random process with $X(t)$ representing the set of events that has occurred at time t. As already stated, starting with an healthy cell, there are no aberrations that have occurred: $X(0) = \emptyset$. With the consecutive accumulation of them, we have that $X(t_1) \subset X(t_2)$ with $t_2 > t_1$. Recall also that after a stop state it is not possible to have any other events occurring.

It is not possible to have an aberration more than one time in D. This means that, between two states, the aberrations that can occur are given by $q^C = ([n]\backslash q) \cup \{q_z\}$ considering that q denotes the events that had already occurred.

As discussed in 4.1.2, the path probability of a Markov chain is given by

$$\Pr\{\emptyset, d_1\} * \Pr\{d_1, d_2\} * ... * \Pr\{d_{k_1}, d_k\} = \Pr\{\emptyset, d_1\} * \Pr\{d_1, d_2\} * ... * \Pr\{d_{k_1}, q_z\}. \quad (5.1)$$

The transition probability between two states (q and q') only depends on them, since we are dealing with a first-order Markov process.

The reasons for considering this process a Markov Model were stated by Hjelm *et al.* and the following assumptions were made:

1. The occurrence of a mutation can be thought of as a failure of an engineered system or component. Since these events are usually independent of each other and can occur continuously, they are usually modeled as a Poisson process. This means that the rate for an aberration to occur is given by a parameter which is also called intensity, Λ; it will be considered to be constant in this work—making this Poisson process homogeneous; the exponential distribution is normally used when describing the time length between two events like these ones. The time until an aberration $j \in q^C$ occurs in state q, denoted as T_j^q, is, then, exponentially distributed with intensity $\Lambda_j(q)$. Note also that the process is memoryless:

$$\Pr(T_j^q > t + s | T_j^q > t) = \Pr(T_j^q > s). \quad (5.2)$$

2. The times until different aberrations occur in state q (i.e., all the T_j^q) are considered to be independent. This assumption derives from the fact that we can consider T_j and T_i to be independent from each other before transitioning to a new state q'. Even if one influences the other, this cannot be felt before they occur.

Both assumptions give that this process can be described as a Markov Model with transition probability given by

$$a_{qq'} = \Pr_{q, q \cup \{j\}} = \frac{\Lambda_j(q)}{\sum_{(k \in [n] \backslash q) \cup \{q_z\}} \Lambda_k(q)}. \quad (5.3)$$

Considering figure 5.1, which shows a Markov chain for two aberrations, we have, for example:

$$\Pr\{1\}\{1, 2\} = \frac{\Lambda_2(\{1\})}{\Lambda_2(\{1\}) + \Lambda_{q_z}(\{1\})}$$

and

$$\Pr\{1\}\{1, q_z\} = \frac{\Lambda_{q_z}(\{1\})}{\Lambda_2(\{1\}) + \Lambda_{q_z}(\{1\})}.$$

5.4 Extension to a Hidden Markov Model

In the work of Hjelm *et al.*, the fact that an aberration could not be detected in the laboratory was not analyzed. Here, we will consider this possibility by means of a HMM. The structure followed will be the same as before but now each transition will have an associated emission probability. For simplicity reasons, it will only be possible to emit or not emit an aberration, i.e., the mutation is either detected in the lab or not.

Each aberration will have its own **emission probability**. This means that any transition between states q and q' when $q' \backslash q = j$ will have an $e(j)$ associated which is only dependent on the aberration (see figure 5.2).

From now on, x will denote an emitted sequence of aberrations (a sample from the dataset) and z will be a path in the HMM through several states.

Considering this, the path probability is now given an identical equation to 4.9.

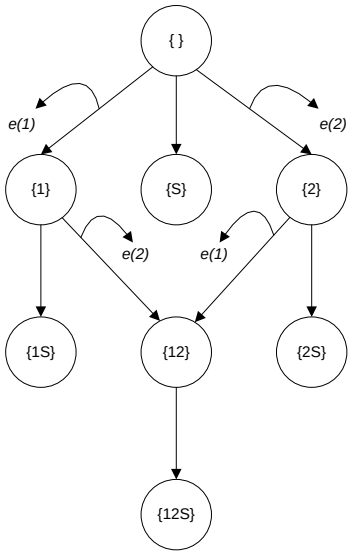


Figure 5.2: DAG representing the HMM used to describe cancer progression. S marks a stop state.

5.5 Emission errors

So far, only the possibility of not emitting an aberration has been considered. Now, the possibility of wrong emissions should also be analyzed. First, one could think that each aberration j could, sometimes, emit another aberration k by mistake. This probability $e_j(k)$ of emitting a wrong aberration would imply a lot of parameters to estimate—$n * (n + 1)$, since it would be possible to emit every aberration from the aberration set $[n]$ or not emit at all—and also much more computations to

simply compute the probability of a given sequence x to fit a path z. For example, admitting that $x = \{1, 2, 3\}$ has been emitted and considering a path $z = \{1\} \cup \{1, 2\} \cup \{1, 2, 3\} \cup \{1, 2, 3, 4\}$ we should have:

$$\begin{aligned}
\Pr(x, z|\theta) = & \, a_{\{\},\{1\}} a_{\{1\},\{1,2\}} a_{\{1,2\},\{1,2,3\}} a_{\{1,2,3\},\{1,2,3,4\}} e_1(1) e_2(2) e_3(3) e_4(0) \\
& + a_{\{\},\{1\}} a_{\{1\},\{1,2\}} a_{\{1,2\},\{1,2,3\}} a_{\{1,2,3\},\{1,2,3,4\}} e_1(2) e_2(1) e_3(3) e_4(0) \\
& + a_{\{\},\{1\}} a_{\{1\},\{1,2\}} a_{\{1,2\},\{1,2,3\}} a_{\{1,2,3\},\{1,2,3,4\}} e_1(1) e_2(3) e_3(2) e_4(0) \\
& + ... \\
& + a_{\{\},\{1\}} a_{\{1\},\{1,2\}} a_{\{1,2\},\{1,2,3\}} a_{\{1,2,3\},\{1,2,3,4\}} e_1(0) e_2(1) e_3(2) e_4(3),
\end{aligned}$$

which has a factorial number of terms. This would also imply that any parameters estimation would be a lot harder to compute. However, a simpler solution might be to consider that when transitioning from one state to another, there is a chance of emitting one aberration (and not the others) and, in the end, some additional aberrations could appear in x. These will be called **error emissions** and, for the reasons already discussed, should be independent of the path z. Now, we will have for each aberration a probability of being **emitted**, $e(j)$, and a probability of being an **error**, $\varepsilon(j)$. This errors will be considered after a cancer is discovered, i.e., after going down the graph through the normal evolution of a HMM process.

In a sense, the emission probabilities can be related to the idea of false-negatives: aberrations that occurred but were not detected. The opposite is true for these error probabilities: they can be regarded as the probability of a false-positive to occur.

5.6 Network Aberration Model

As already discussed in 3, the parameter complexity of the Markov chain model can be reduced from an exponential number of parameters in n to only a quadratic number of them by introducing appropriate assumptions. This constitutes what Hjelm *et al.* called the Network Aberration Model.

The focus here is in the interdependencies between two aberrations and the NAM can be defined as a triple $M = (\lambda, \delta, \psi)$ as follows (note that these three types of parameters will be called *transition parameters* throughout this work):

- $\lambda = \{\lambda_1, ..., \lambda_n\}$ is a set of aberration intensities. This marks how probable is for an aberration to occur by itself, i.e., in the starting state with an healthy cell. Note that this values have only a relative value, i.e., a high intensity value just means that it is more probable to occur than the others mutations. The way of how this can be translated to the regular transition probabilities $a_q q'$ will be clear later (6.4 and 6.5).

- $\delta = \{\delta_{ij} : 1 \leq i, j \leq n, i \neq j\}$ denotes the pairwise dependencies between two mutations. δ_{ij} represents how the intensity for aberration j changes after

aberration i occurs. When evaluating the different probabilities for each aberration to occur in the next state, this parameter accounts for the influence of the already occurred aberrations.

Considering that $i \neq j \in q^C \backslash \{q_z\}$, i.e., j is a possible next aberration but not a stop state, then the intensity is now given by

$$\Lambda_j(q, i) = \Lambda_j(q) * \delta_{ij}. \tag{5.4}$$

It will also be assumed that this interdependence is independent of the time; this would imply that the intensity of the new aberration j will be influenced by all the aberrations already occurred regardless of when they occurred:

$$\Lambda_j(q) = \lambda_j \prod_{i \in q} \delta_{ij}. \tag{5.5}$$

- $\psi = \{\psi_1, ..., \psi_n\}$ is a set of stop intensities. As already stated, the stop state event is an event that marks that the cancer was discovered. Here it is assumed that this intensity only depends on the number of aberrations that had already occurred. This means that a high number of mutations imply a higher probability of the cancer to be discovered ([27]). The intensity for these events is given by

$$\Lambda_{q_z}(q) = \psi_{|q|}. \tag{5.6}$$

with $|q|$ denoting the number of occurred aberrations. Note that two states with the same number of aberrations do not necessarily have the same probability of being discovered as cancers in the next state, since this depends on the other possible next states.

These transition parameters are positive numbers that mark how probable the next aberration (or stop state) is to develop compared to the other ones. It is assumed that the pairwise dependencies can only increase or leave the probability of the following aberrations unchanged. Since this probability cannot be decreased, $\delta_{ij} \geq 1$ for all $1 \leq i, j \leq n$ such that $i \neq j$. It follows that, to keep an aberration unchanged (i.e., these two mutations are independent of each other), $\delta_{ij} = \delta_{ji}$ must be equal to 1. Note also that, in order to avoid situations where one aberration influences another and vice-versa, it is assumed that the corresponding opposite pair is equal to one ($\delta_{ij} > 1 \Rightarrow \delta_{ji} = 1$ and vice-versa).

The matrix λ can also be represented using a graph (see 7.2). This dependency graph (DG) is composed of different nodes, each representing a different aberration. They are connected through arrows that denote the pairwise dependencies between then when they are not independent ($\delta_{ij} > 1$). Note that, albeit graphically similar,

the DG is not the same as the HMM considered before. This DG is just a representation of the λ and contemplates all the possible pairwise influences. The HMM denotes the transitions that can in fact occur.

It is clear now that the NAM will have $2n + n^2$ parameters. This number is much lower than just using the transition probabilities $a_{qq'}$ between states. Since we have n aberrations, this would imply a maximum number of 2^n states and, to connect all of them, the number of edges (and, at the same time, the number of parameters to estimate) would be

$$\binom{2^n}{2} = \frac{2^n!}{2!(2^n-2)!} = \frac{2^n * (2^n - 1)}{2} = 2^{n-1} * (2^n - 1)$$

which starts to be prohibitive for many aberrations.

Chapter 6

Structure and Parameters Estimation

6.1 Structure Estimation

The structure referred here is the HMM that sets how the cancer can evolve. This DAG configuration was already studied by Höglund *et al.* who have develop a somehow heuristic method to extract it from a dataset of cancer and will not be the focus of this work.

This temporal analysis technique, suggested in [20], is based on the assumption that the number of chromosome rearrangements is proportional to the grade of the malignancy.

The number of cytogenetic imbalances per tumor (NIPT) will, thus, reflect somehow the age of the tumor. From this, they derive a statistical measure for the time of occurrence of each aberration which is based on the assumption that mutations appearing early in tumor progression will probably be more common in different samples of the same cancer, being seen in both simple and complex tumors (here, simple and complex refers to the NIPT).

They proceed by plotting the NIPT distribution for tumors in which a given aberration was detected. If this plot shows a tendency to having many tumors with low NIPT, this would signal that this aberration is probably near the root (the healthy state) of the graph. Due to the fact that, as indicated by real data, the majority of this distributions are usually uniform (and then, making it impossible to see if they are early or late aberrations) this method has some problems. The way they use to quantify this time of occurrence (TO) of an aberration is given by the mode of each distribution; the mean is discarded since some distributions are also skewed.

See 7.2 for a deeper analysis of this method. Since it only tells when an aberration occurred and not how they relate to each other (and identify possible pathways), they then proceed to analyze the datasets with principal component analysis techniques.

6.2 Parameters Estimation

Since the EM algorithm is being followed here, one must optimize the Q term:

$$Q(\theta;\theta') = \sum_{x\in D}\sum_{z} \Pr(z|x,\theta)\log\Pr(x,z|\theta'), \tag{6.1}$$

in which x denotes an emitted sequence of aberrations and z is a path in the HMM through several states. D is a set of data (a set of sequence of aberrations that lead to a cancer). θ represents the parameters that must be estimated, with θ referring to the old parameters (the ones from the last iteration of the algorithm) and θ' to the new ones.

Each state is denoted as q and a transition between states that are connected through an edge will imply that a new aberration j has occurred, i.e., $q' = q \cup \{j\}$; $[n]$ denotes the set of possible aberrations. The probability that a transition will occur is given by $a_{qq'}$ and the probability that an emission for this transition will in fact occur is denoted by $e(j)$. It is considered that this emission probability is only dependent on the aberration that had occurred and, as such, independent of what happened before (in other words, independent of the state q.)

Recall that to compute the transition probability, we have:

$$a_{qq'} = \frac{\Lambda_j(q)}{\sum_{k\in[n]\setminus q}\Lambda_k(q)}. \tag{6.2}$$

Since we know the structure that describes how the different aberrations relate to each other, we can assume that the denominator is just computed for the aberrations that are present in the node's children. Assuming that all the aberrations present in the children nodes are Q':

$$a_{qq'} = \frac{\Lambda_j(q)}{\sum_{k\in Q'\setminus q}\Lambda_k(q)}. \tag{6.3}$$

This Λ represents the intensity distribution of an aberration or a stop state and is computed in different ways for both:

$$\Lambda_j(q) = \begin{cases} \lambda_j \prod_{i\in q}\delta_{ij} & \text{if } j \text{ is an aberration,} \\ \psi_{|q|} & \text{if } j \text{ is a stop state.} \end{cases}$$

As such, $a_{qq'}$ will have two different ways to be computed, whether the next state is an aberration or a stop state:

$$a_{qq'} = \frac{\lambda_j \prod_{i\in q}\delta_{ij}}{\sum_{k\in Q'\setminus q}\lambda_k \prod_{i\in q}\delta_{ik} + \psi_{|q|}}, \tag{6.4}$$

and

$$a_{qq'} = \frac{\psi_{|q|}}{\sum_{k\in Q'\setminus q}\lambda_k \prod_{i\in q}\delta_{ik} + \psi_{|q|}}, \tag{6.5}$$

Defining q_z as the stop state of a path z, the probability $\Pr(x, z|\theta')$ is given by

$$\Pr(x, z|\theta') = \prod_{q,q' \in z} a'_{qq'} \prod_{j \in q_z \cap x} e'(j) \prod_{j \in q_z \setminus x} (1 - e'(j)). \tag{6.6}$$

Here, $j \in q_z \cap x$ refers to the observed aberrations and $j \in q_z \setminus x$ to the ones that were not emitted. This means that for each path z in the DAG, the transition probabilities between the states that compose a path must be multiplied, since we are dealing with Markov chains. The probability that an emission has (or not) occurred also depends on the path and the observed sequence.

$$Q(\theta; \theta') = \sum_{x \in D} \{ \sum_{z} \Pr(z|x, \theta) [\sum_{q,q' \in z} \log a'_{qq'} + \sum_{j \in q_z \cap x} \log e'(j) + \sum_{j \in q_z \setminus x} \log(1 - e'(j))] \}. \tag{6.7}$$

Considering that the emission probability is independent of the states, (6.7) can be rewritten considering that in different paths the aberration j has the same probability of being emitted:

$$Q(\theta; \theta') = \sum_{x \in D} \{ \sum_{z} \Pr(z|x, \theta) [\sum_{q,q' \in z} \log a'_{qq'} + \sum_{j \in q_z \cap x} \sum_{\substack{q,q' \in z: \\ q' \setminus q = j}} \log e'(j) + \sum_{j \in q_z \setminus x} \sum_{\substack{q,q' \in z: \\ q' \setminus q = j}} \log(1 - e'(j))] \}.$$

Inverting the order of the z and q, q' sum in order to be able to apply faster methods of computation, we have:

$$\begin{aligned} Q(\theta; \theta') = \sum_{x \in D} \{ & \sum_{q,q'} \sum_{\substack{z: \\ q,q' \in z}} \Pr(z|x, \theta) \log a'_{qq'} \\ & + \sum_{j} \sum_{\substack{q,q': \\ q' \setminus q = j}} \sum_{\substack{z: \\ q,q' \in z \\ j \in q_z \cap x}} \Pr(z|x, \theta) \log e'(j) \\ & + \sum_{j} \sum_{\substack{q,q': \\ q' \setminus q = j}} \sum_{\substack{z: \\ q,q' \in z \\ j \in q_z \setminus x}} \Pr(z|x, \theta) \log(1 - e'(j)) \} \end{aligned} \tag{6.8}$$

and using, for example, (6.4) to rewrite (6.8) and moving the sum in $x \in D$:

$$\begin{aligned} Q(\theta; \theta') = & \sum_{j} \sum_{x \in D} \sum_{\substack{q,q': \\ q' \setminus q = j}} \sum_{\substack{z: \\ q,q' \in z \\ j \in q_z \cap x}} \Pr(z|x, \theta) \log e'(j) \\ & + \sum_{j} \sum_{x \in D} \sum_{\substack{q,q': \\ q' \setminus q = j}} \sum_{\substack{z: \\ q,q' \in z \\ j \in q_z \setminus x}} \Pr(z|x, \theta) \log(1 - e'(j)) \\ & + \sum_{x \in D} \sum_{q,q'} \sum_{\substack{z: \\ q,q' \in z}} \Pr(z|x, \theta) [\log \lambda'_j + \sum_{i \in q} \log \delta'_{ij} - \log(\sum_{k \in [n] \setminus q} \lambda'_k \prod_{i \in q} \delta'_{ik} + \psi'_{|q|})]. \end{aligned} \tag{6.9}$$

6.3 Computing the $\Pr(z|x,\theta)$ term using the forward-backward algorithm

Defining $d^x(j)$ as:

$$d^x(j) = \begin{cases} e(j) & \text{if } j \text{ was emitted in } x, \\ 1 - e(j) & \text{otherwise.} \end{cases} \tag{6.10}$$

We have

$$\Pr(x|\theta) = \sum_z \Pr(x, z|\theta) \tag{6.11}$$

and

$$\begin{aligned} \Pr(x, z|\theta) &= \frac{\Pr(z, x, \theta)}{\Pr(\theta)} = \frac{\Pr(z, x, \theta)}{\Pr(x, \theta)} \frac{\Pr(x, \theta)}{\Pr(\theta)} \\ &= \Pr(z|x, \theta) \frac{\Pr(x, \theta)}{\Pr(\theta)} = \Pr(z|x, \theta)\Pr(x|\theta). \end{aligned}$$

Finally,

$$\Pr(z|x, \theta) = \frac{\Pr(x, z|\theta)}{\Pr(x|\theta)}. \tag{6.12}$$

We can compute these probability values using the DAG and the θ term. In [24], for example, the Baum-Welch algorithm (or forward-backward algorithm) is explained. The procedures described there will be introduced here, since they will be necessary in order to computer these probabilities in an efficient way, i.e., recursively.

Defining f_q as forward probability:

$$\begin{aligned} f_q^x &= \Pr\ [\text{ reaching } q \text{ and having generated } q \cap x\ ] \\ &= \sum_{q^-:<q^-,q>\in E} f_{q^-} a_{q^-q} d^x(j) \end{aligned}$$

in which q^- refers to the previous states and E is a set containing all edges of the DAG. This means that all the paths that pass through this node are considered to compute f_q^x. The backward probability, b_q will be defined in the opposite way, as:

$$\begin{aligned} b_q^x &= \Pr\ [\text{ starting at } q \text{ and reaching } x \setminus q\ ] \\ &= \sum_{q^+:<q,q^+>\in E} d^x(j) a_{qq^+} b_{q^+} \end{aligned}$$

with q^+ referring to the next states. The sum of paths probabilities in which the transition is made between two states q and q' is given by:

$$\Pr(x, z|\theta) = f_q a_{qq'} d^x(j) b_{q'} \tag{6.13}$$

Consider figure 6.1 for a graphical representation of this algorithm.

The computation of f_q and b_q for each state shall be made following these points:

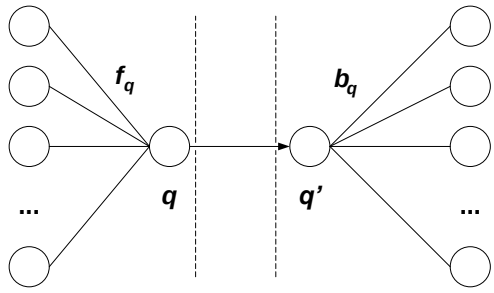


Figure 6.1: Illustration of the sequence of operations required to compute both the forward and backward probabilities

1. Starting with the root, f_q should be considered to be 1, since it is the first and only state.

2. For the next state, the transition probability should be computed using (6.3). The value for f_q for this node is then easily computed.

3. Picking another state, f_q should be computed in the same way. Note that a node can have more than one parent. All these values should be stored since they will be used later.

4. b_q can be computed considering that $b_{q_z} = 1$ and then going backwards.

5. Note that each node needs only to be picked once; so, special care should be taken just to consider the nodes of the DAG and not simply transitioning from parents to children and vice-versa.

Note also that computing $\Pr(x|\theta)$ is equivalent to computing the last value of f_q for all the paths and then adding the result, i.e., adding all the f_{q_z}.

$$\Pr(x|\theta) = f_{q_z} b_{q_z} \tag{6.14}$$

Considering this, we can use this probabilities to compute $\Pr(z|x, \theta)$.

6.4 Emission Errors

Considering now that it is possible to emit erroneous aberrations (as described in 5.5) after going down the graph, $\Pr(x, z|\theta')$ would now be:

$$\Pr(x, z|\theta') = \prod_{q,q' \in z} a'_{qq'} \prod_{j \in q_z \cap x} e'(j) \prod_{j \in q_z \setminus x} (1 - e'(j)) \prod_{j \in x \setminus q_z} \varepsilon'(j) \prod_{j \in [n] \setminus [x \cup q_z]} (1 - \varepsilon'(j)), \tag{6.15}$$

in which $j \in x \backslash q_z$ refers to the aberrations that were emitted as errors and $j \in [n] \backslash [x \cup q_z]$ to the rest of the aberrations that were not emitted. Equation (6.8) is now given by

$$\begin{aligned} Q(\theta;\theta') = & \sum_{x \in D} \sum_{q,q'} \sum_{\substack{z \\ q,q' \in z}} \Pr(z|x,\theta) \log a'_{qq'} \\ & + \sum_{j} \sum_{x \in D} \sum_{\substack{q,q' \\ q' \backslash q = j}} \sum_{\substack{z: \\ q,q' \in z \\ j \in q_z \cap x}} \Pr(z|x,\theta) \log e'(j) \\ & + \sum_{j} \sum_{x \in D} \sum_{\substack{q,q' \\ q' \backslash q = j}} \sum_{\substack{z: \\ q,q' \in z \\ j \in q_z \backslash x}} \Pr(z|x,\theta) \log[1 - e'(j)] \\ & + \sum_{j} \sum_{x \in D} \sum_{q_z} \sum_{\substack{z: \\ q_z \in z \\ j \in x \backslash q_z}} \Pr(z|x,\theta) \log \varepsilon'(j) \\ & + \sum_{j} \sum_{x \in D} \sum_{q_z} \sum_{\substack{z: \\ q'_z \in z \\ j \in [n] \backslash [x \cup q_z]}} \Pr(z|x,\theta) \log[1 - \varepsilon'(j)], \end{aligned} \tag{6.16}$$

For these last two members, there is no need to compute $\Pr(z|x,\theta)$ for each q, q' but we still need its value for each path z. On the other hand, the second and third term should take into account these error probabilities in the final transition q, q'. The sum of paths probabilities in which the transition is made between two states q and q' is now given by:

$$\sum_{\substack{z \\ q,q' \in z}} \Pr(x,z|\theta) = f_q a_{qq'} e(j) b_{q'} \prod_{j \in x \backslash q_z} \varepsilon'(j) \prod_{j \in [n] \backslash [x \cup q_z]} (1 - \varepsilon'(j)), \tag{6.17}$$

considering that an emission j was observed. The definition of b_{q_z} can now be considered to be the final two terms of this last equation:

$$b_{q_z} = \prod_{j \in x \backslash q_z} \varepsilon'(j) \prod_{j \in [n] \backslash [x \cup q_z]} (1 - \varepsilon'(j)) \tag{6.18}$$

and (6.17) becomes

$$\sum_{\substack{z \\ q,q' \in z}} \Pr(x,z|\theta) = f_q a_{qq'} e(j) b_{q'}. \tag{6.19}$$

$\Pr(x|\theta)$ is now

$$\Pr(x|\theta) = \sum_{q_z} f_{q_z} b_{q_z}. \tag{6.20}$$

6.5 Computer Implementation and Optimization of e and ε

The purpose of this computer implementation is to estimate the transition parameters (λ, δ and ψ), as well as the emission probabilities (e) and error probabilities (ε). The optimization of the last two sets of variables can be done in a easier way than the transition parameters and, as such, would be the ones analyzed here. Recall that

$$\prod_{i=1}^{n} x_i^{a_i} \text{ where } \sum_{i=1}^{n} x_i = 1 \ , \ 0 \leq x_i \leq 1 \tag{6.21}$$

is maximized by $x_i = \frac{a_i}{\sum_{i=1}^{n} a_i}$. This property can be applied in (6.16) to estimate both the emission and errors probabilities since $e'(j) + (1 - e'(j)) = 1$ and $\varepsilon'(j) + (1 - \varepsilon'(j)) = 1$. As such, we have

$$e'(j) = \frac{\sum_{x \in D} \sum_{\substack{q,q' \\ q' \setminus q = j}} \sum_{\substack{z \\ q,q' \in z \\ j \in q_z \cap x}} \Pr(z|x,\theta)}{\sum_{x \in D} \sum_{\substack{q,q' \\ q' \setminus q = j}} \sum_{\substack{z \\ q,q' \in z \\ j \in q_z \cap x}} \Pr(z|x,\theta) + \sum_{x \in D} \sum_{\substack{q,q' \\ q' \setminus q = j}} \sum_{\substack{z \\ q,q' \in z \\ j \in q_z \setminus x}} \Pr(z|x,\theta)} \tag{6.22}$$

and

$$\varepsilon'(j) = \frac{\sum_{x \in D} \sum_{q_z} \sum_{\substack{z \\ q_z \in z \\ j \in x \setminus q_z}} \Pr(z|x,\theta)}{\sum_{x \in D} \sum_{q_z} \sum_{\substack{z \\ q_z \in z \\ j \in x \setminus q_z}} \Pr(z|x,\theta) + \sum_{x \in D} \sum_{q_z} \sum_{\substack{z \\ q'_z \in z \\ j \in [n] \setminus [x \cup q_z]}} \Pr(z|x,\theta)}. \tag{6.23}$$

Before starting, the cancer sequences should be organized in some way. Since there would probably be several sequences that are repeated along the training set, all these probabilities can be computed only once and then the results multiplied according to the number of times they appear. The EM algorithm can be described as follows:

1. Start by giving initial values to the parameters being generated.

2. Start by picking the first sequence of aberrations and count how many times it is repeated along the cancer set.

3. For this particular sequence, compute all the f_q for each node as described earlier. This should be done layer by layer, since the results of a node are always dependent on its parent nodes. So, for each node, the new aberrations that are present on its children should be considered in order to compute the transition probabilities (using the estimated transition parameters). They are later multiplied by the emission probabilities (taking into account if the aberration was or not emitted, since they have different probability values). This should be done by every node and it is a reasonable idea to store this values in order to be used later when computing b_q.

4. Starting with the last layer of the graph, all the b_q should be recursively computed using b_{q_z}—see (6.18)—and the results that were stored in the previous step.

5. The computation of the Q term can be divided in three parts:

   a) The part related to the transitions parameters, $Q_a(\theta;\theta')$, shall considerer all the edges of the graph. This is done by considering every node and, for each, its outgoing edges. Taking into account both (6.19) and (6.12), we have:

   $$Q_a(\theta;\theta') = \sum_{q,q'} \frac{f_q a_{qq'} e(j) b_{q'}}{\Pr(x|\theta)} \log a'_{qq'} = \sum_{q,q'} \frac{f_q a_{qq'} e(j) b_{q'}}{\sum_{q_z} f_{q_z} b_{q_z}} \log a'_{qq'} \quad (6.24)$$

   b) $Q_e(\theta;\theta')$ denotes the second and third term of (6.16). They are computed in a similar fashion as $Q_a(\theta;\theta')$, but taking into account that only some edges can be considered. For this purpose, when constructing the structure that relates the different aberrations, it should be possible to list all the edges that correspond to every aberration—when considering the edges that correspond to an aberration j, it is then assured that it would be also in q_z. So, it would be easy to select $j \in q_z \cap x$ and $j \in q_z \backslash x$.

   c) $Q_\varepsilon(\theta;\theta')$ corresponds to the last two terms of the Q term and we have:

   $$Q_\varepsilon(\theta;\theta') = \sum_{j \in x \backslash q_z} \sum_{q_z} \frac{f_{q_z} b_{q_z}}{\sum_{q_z} f_{q_z} b_{q_z}} \log \varepsilon'(j) + \sum_{j \in [n] \backslash [x \cup q_z]} \sum_{q_z} \frac{f_{q_z} b_{q_z}}{\sum_{q_z} f_{q_z} b_{q_z}} (1 - \varepsilon'(j)) \quad (6.25)$$

   To select the right aberrations and final states, it would be reasonable to have a list connecting each aberration with the final states that do not contain this aberration. For the first member, only aberrations that are in x should be considered and for the second term the ones that correspond to $n \backslash x$.

6. The next aberration sequence should be analyzed and the results should be added to the previous ones.

7. In the end, the parameters that are being estimated should be computed as in (6.22) and (6.23) and all this steps should be repeated using this new values.

Chapter 7

Simulations, Results and Discussion

The simulations and tests were performed in MATLAB. This solution proved to be good to check if the algorithms work, but was too slow when performing some tests.

7.1 Data Generator

This section will deal with how synthetic data can be generated in order to simulate cancer progression. An example generated by this method will be discussed in 7.3.1.

Following the ideas described previously, this generator will consist of three parts. The fist one will generate the transition parameters, the second a DAG and the third part will use the previous ones in order to generate a sequence of aberrations.

7.1.1 Transition Parameters Generation

Each aberration will have its own randomly generated parameters: intensity (λ), emission probability (e) and error probability (ε). The stop intensities (ψ) are also randomly generated, having in mind that their values must increase with the number of aberrations already occurred. The way aberrations relate to each other, i.e., the dependency graph of the Network Aberration Model which graphically represents the matrix δ, is also randomly generated; note that situations when an aberration directly influence another aberration and vice-versa were discarded.

7.1.2 DAG Generation

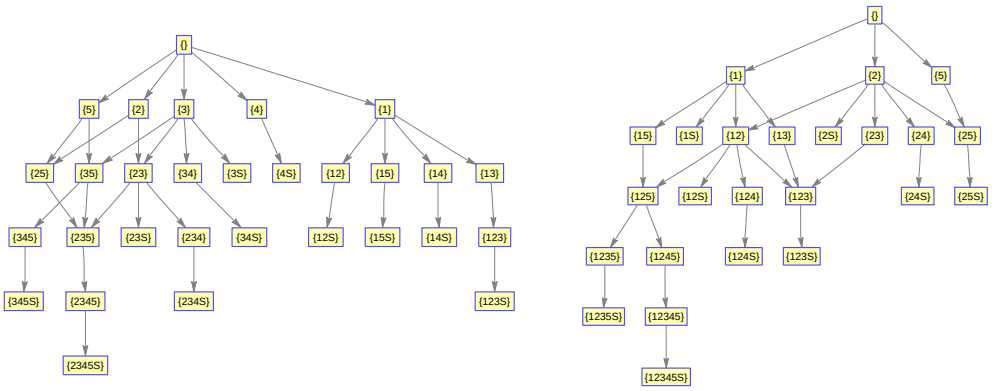
The second script will generate a Directed Acyclic Graph which denotes how the aberrations evolve to form a cancer. The way this graph is constructed tries to capture the most probable paths followed during cancer progression (considering the transition parameters). For this, starting with the root as a healthy state (no aberrations), it will find which possible child node is the most probable one and then connect it to the root. In the next step, it will randomly pick one of the already generated nodes and do the same. This operations should be done

a sufficient amount of times to populate the graph. Also, nodes that have a low probability of occurrence will not be able to progress with more aberrations. This "low-probability"is given by $\bar{f_q} * T$, in which $\bar{f_q}$ denotes the mean of the f_q for all the q nodes generated so far and T is a threshold value. This algorithm is described in more detail in Algorithm 1, in which R refers to the maximum number of iterations. This number of iterations is connected with the number of nodes that the final tree will have; recall also that $[n]$ was previously defined as the set of possible aberrations.

Algorithm 1: DAG generation algorithm

```
Input: λ, δ, ψ, T, R
while iteration number < R do
    pick one node randomly
    if f_q(node) > f̄_q * T then
        if the node is not a stop state then
            Q ← aberrations that had already occurred for this node
            C ← aberrations present in the children nodes of this node
            Q_c ← [n]\(Q ∪ C) (set of possible aberrations to occur)
            compute the probabilities of occurrence for every aberration
            present in Q_c
            generate one child, that must correspond to the most probable
            aberration
            if the child node is a stop state then
                connect the parent node to the child one
            else
                if the child node already belongs to the DAG, i.e., if a similar
                node was generated before through another parent then
                    connect the parent node to the original "new"node
                else
                    connect the parent node to the child one
                end
            end
        end
    end
end
for all the leafs in the DAG that are not a stop state do
    generate the corresponding stop state (new node with the
    "stop"aberration)
end
```

Although not explicit in this algorithm, the possibility of the root to generate a stop state was discarded, as this would indicate that a cancer was discovered but no aberrations were detected. Note also that the last loop is necessary since it can happen that a node has already such a low probability that it cannot generate



(a) DAG 1 (b) DAG 2

Figure 7.1: Two different DAGs

children. As it is not possible to have a branch without a stop state in the end, the last loop is added.

7.1.3 Aberration Sequence Generation

After the generation of the DAG, the cancers themselves can start to be generated; they are represented by a sequence of aberrations. The first step is to pick the root and then, accordingly to the probabilities of its children nodes, to jump to the next one randomly. According to the model for cancer progression followed in this work, one transition between nodes implies one aberration. Then, it is necessary to take into consideration if it should be emitted or not (again, randomly, but taking into account the different emission probabilities). When the aberration is considered not to be emitted, it does not appear in the final sequence of aberrations. Again, according to the model, it is possible that, in the end (after going down the tree), each aberration can be emitted again as an error (according to its error emission probability)—see 5.5.

7.2 Structure Estimation

After running the data generator described in the last section, the algorithm described in 6.1 will be tested. This method tries to find the time of occurrence of an aberration based on cross-sectional data. After generating two different DAGs (figure 7.1), the test was performed. The results are given in table 7.1 and refer to datasets of 500, 1000 and 2000 cancers, all with the same result. The *real* results that are in this table are mere numbers that pretend to give an indication of when an aberration occurred compared to each other.

Table 7.1: TO for each DAG

i	TO for DAG 1		TO for DAG 2	
	estimated	real	estimated	real
1	2	4	3	2
2	4	4	2	2
3	4	4	3	3
4	4	4	5	3
5	4	4	2	2

7.3 Parameters Estimation

Throughout all this section, the structure is assumed to be know (or estimated before).

Since the parameters estimation turns up to be quite complex, here it will be analyzed in different parts. The emission probabilities will be the first to be studied, considering that the error and transition probabilities are assumed to be known. After this, the same approach will be taken for the error probabilities.

7.3.1 Emission Probabilities

This section will deal with how well the emission probabilities are estimated by this method. We will start with an example and then proceed to a compared study of several runs of the algorithm with different conditions (like the number of cancer samples available, for example).

Consider that several cancers (with a maximum of 4 mutations) were generated based on the transition and emission probabilities shown in tables 7.2 and 7.3.

Table 7.2: Randomly generated start and stop intensities arrays and emission probabilities

i	λ_i	ψ_i	$e(i)$
0		0	
1	4	2	0.7156
2	1	2	0.7360
3	2	2	0.8635
4	3	3	0.7487

Note that table 7.3 can be also represented by its DG as seen in figure 7.2. The DAG generated is represented in figure 7.3.

Table 7.3: Randomly generated pairwise dependencies between aberrations

δ			
1	1	5	3
2	1	1	1
1	3	1	1
1	5	1	1

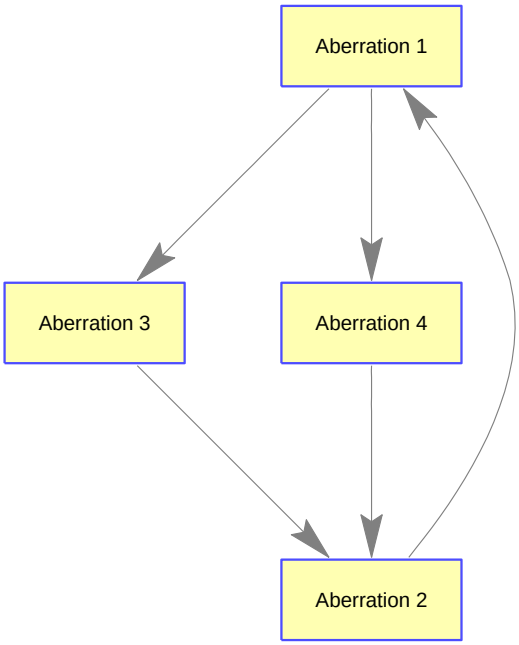


Figure 7.2: Dependency graph

In this case, 1000 cancer samples were generated. Since we are excluding additional aberrations that could appear as false-positives, we only need to consider that some aberrations were not emitted.

In order to estimate these emission probabilities, the steps referred in 6.5 should be followed. The estimated values (after 20 iterations) were computed using different initial steps (table 7.4).

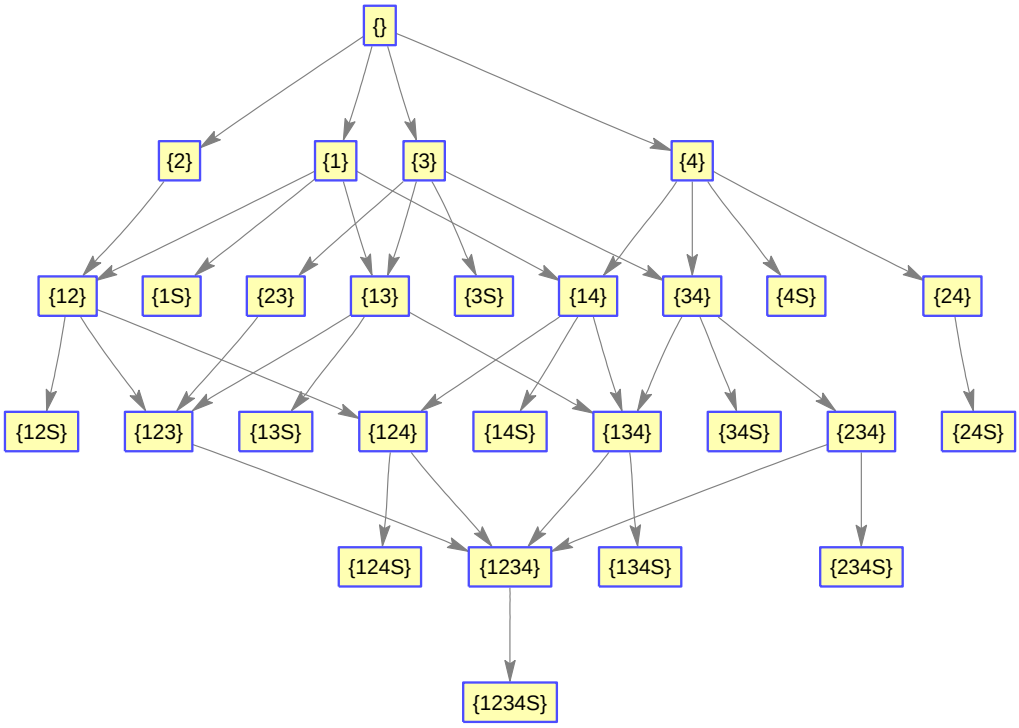


Figure 7.3: DAG representing the Markov Chain used to describe cancer progression

The evolution of the likelihood

$$l = \log[(\Pr(x|\theta')] = \sum_{q,q' \in z} \log a'_{qq'} + \sum_{j \in q_z \cap x} \log e'(j) + \sum_{j \in q_z \setminus x} \log(1 - e'(j)) \qquad (7.1)$$

(with $a_{qq'}$ given by 6.3) is shown in Figure 7.4 .

The evolution of the Q-term is plotted in Figure 7.5 and in Figure 7.6 is plotted the 2-norm between the estimated value and the real one for all iterations.

It is interesting to see that the estimated values are practically independent of the values chosen as initial. However, since the emission probabilities are assumed to be relatively high (i.e., an aberration is supposed to be detected the majority of times), it is reasonable to assume that a good starting value will be next to one (for example, $e'_0(j) = 0.9$, which is the value that will be used from now on).

Proceeding to a more general evaluation of this method, consider the same test performed 3 times but for a different number of samples. This test will include the generation of a new cancer structure (DAG) and new samples. The error is given, like before, by $||\frac{e' - e_{\text{real}}}{e_{\text{real}}}||$ (see table 7.5).

From these results and others that are not shown here, it was possible to conclude that the number of cancer samples does not have a big impact in how well the

Table 7.4: Estimated emission probabilities

$e_0(j)$, $\forall_j$	e'				$\|\|e' - e_{\text{real}}\|\|$	$\|\|\frac{e'-e_{\text{real}}}{e_{\text{real}}}\|\|$
0.1	0.6398	0.6879	0.7825	0.7572	0.1211	6.4879%
0.2	0.6398	0.6879	0.7825	0.7572	0.1210	6.4864%
0.3	0.6398	0.6879	0.7826	0.7572	0.1210	6.4853%
0.4	0.6398	0.6879	0.7826	0.7571	0.1210	6.4843%
0.5	0.6398	0.6879	0.7826	0.7571	0.1210	6.4833%
0.6	0.6398	0.6879	0.7826	0.7571	0.1210	6.4822%
0.7	0.6398	0.6879	0.7827	0.7571	0.1209	6.4807%
0.8	0.6398	0.6879	0.7827	0.7571	0.1209	6.4783%
0.9	0.6399	0.6879	0.7828	0.7571	0.1208	6.4735%
e_{real}	0.7156	0.7360	0.8635	0.7487		

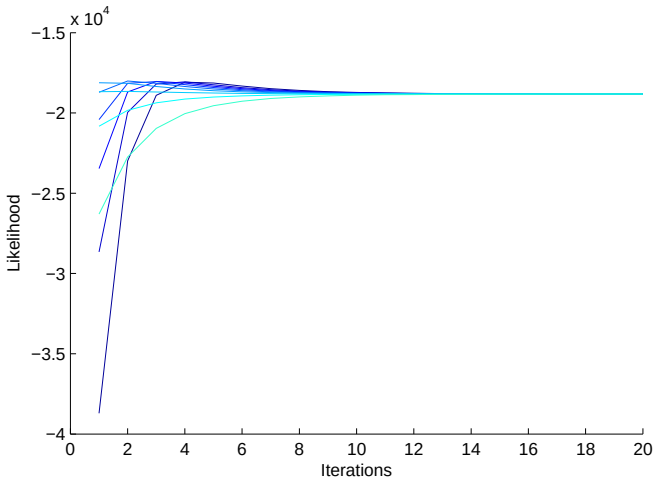


Figure 7.4: Progression of the likelihood for the different initial values. Lighter colors mark the lowest initial values

parameters are estimated for more than 500 of them (which is the number of samples expected in a real database). However, some results are distinctly wrong. Consider the first test for a dataset of 2000 cancers. The estimated and real values, as well as the intensities for each aberrations are presented in 7.6. As it is possible to see, the value with the worst estimation is the one with the lowest intensity aberration, i.e., the lowest probable aberration to occur. This will lead to very few samples (or none at all) that can used to estimate the correct values.

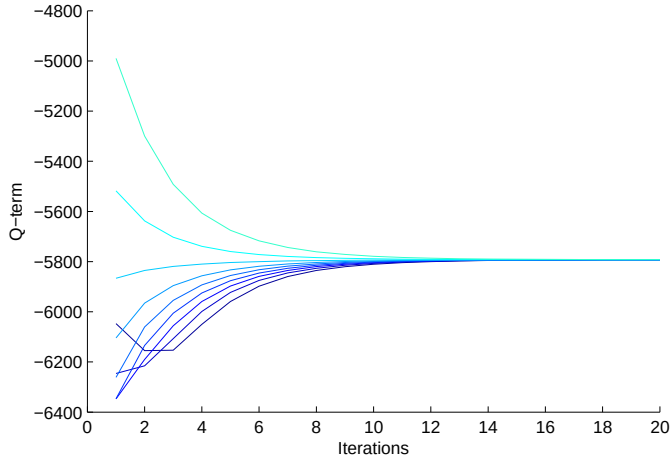


Figure 7.5: Progression of the Q-term for the different initial values. Lighter colors mark the lowest initial values

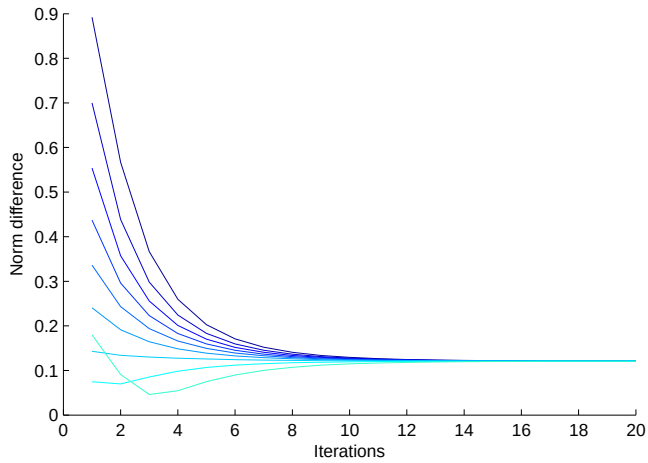


Figure 7.6: $||e' - e_{\text{real}}||$ for the different initial values. Lighter colors mark the lowest initial values

Pseudocounts

The easiest way to deal with this problem is to add pseudocounts to the observed samples [22].

One can do this by adding a constant to the estimate of the parameter being studied. Its purpose is to work as fake counts when the aberration is underrepresented in the sampling. On the other hand, this regularizing effect should be

Table 7.5: Estimated emission probabilities for different number of samples

number of samples	errors (%)		
500	31.4422	0.0544	5.0935
1000	7.9443	9.8783	1.0253
2000	24.7247	3.4226	0.5679
5000	0.1687	3.1517	0.1065

insignificant when a large amount of data is available. Defining A and B as:

$$A = \sum_{x \in D} \sum_{\substack{q,q' \\ q' \setminus q = j}} \sum_{\substack{z \\ q,q' \in z \\ j \in q_z \cap x}} \Pr(z|x,\theta) \tag{7.2}$$

$$B = \sum_{x \in D} \sum_{\substack{q,q' \\ q' \setminus q = j}} \sum_{\substack{z \\ q,q' \in z \\ j \in q_z \setminus x}} \Pr(z|x,\theta) \tag{7.3}$$

our estimation of $e'(j)$ 6.5 is now given by

$$e'(j) = \frac{A + p}{A + B + p} \tag{7.4}$$

with p being the regularizing term that we added in order to act as pseudocount. It is defined as

$$p = \frac{A + B}{n^2}. \tag{7.5}$$

Note that p is just a constant that needs to be smaller than the numerator of 7.4. When A is *small* for one aberration (compared to the other ones), p will work as fake samples.

Table 7.6: Comparison of the estimated values using or nor pseudocounts

	$i = 1$	$i = 2$	$i = 3$	$i = 4$	error (%)
λ	6	8	5	9	
e_{real}	0.9479	0.7461	0.8013	0.7808	
e'	0.7095	0.6814	0.3507	0.7373	24.7247
e'_{pseudo}	1.0000	0.9133	0.9974	0.8959	15.5616

In table 7.6, a comparison of the two methods is given. Note that the error is reduced, specially for the aberration with the lowest intensity value. However, the other aberrations also suffer an increase on its own values of intensity. As such, this method should be applied with care (for example, when one aberration has a very low emission probability, which is something not to be expected).

7.3.2 Error Probabilities

A similar structure to the one followed in the last section will be given here. Considering that the emission and transition probabilities are assumed to be known and generating a new cancer structure and dataset, the results of the estimation are given in table 7.7 and figure 7.7. Note that, now, when generating the samples from the DAG, it is necessary to consider the false-positives already discussed before instead of the false-negatives. Again, the number of samples is 1000; note also that what is plotted now as Q-term$_\varepsilon$ refers to the part of the Q-term that is changed by changing ε and not all the Q-term (i.e., since we are assuming that the transition and emission probabilities are fixed, they will just change Q-term by adding a constant value).

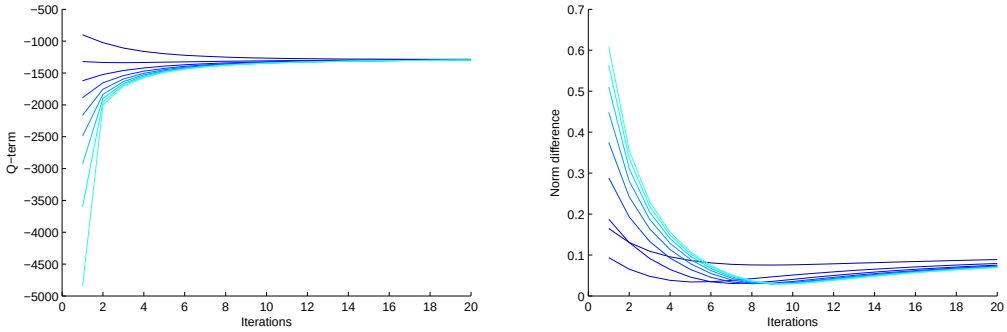
Table 7.7: Estimated error probabilities

$\varepsilon_0(j)$, $\forall_j$	e'				$\|\|\varepsilon' - \varepsilon_{\text{real}}\|\|$	$\|\|\frac{\varepsilon' - \varepsilon_{\text{real}}}{\varepsilon_{\text{real}}}\|\|$
0.1	0.3256	0.3505	0.0398	0.0649	0.2893	16.3919%
0.2	0.3253	0.3521	0.0496	0.0691	0.2802	14.7359%
0.3	0.3252	0.3527	0.0537	0.0708	0.2766	14.0613%
0.4	0.3251	0.3530	0.0558	0.0717	0.2746	13.7008%
0.5	0.3251	0.3532	0.0571	0.0722	0.2734	13.4785%
0.6	0.3250	0.3534	0.0580	0.0726	0.2726	13.3287%
0.7	0.3250	0.3535	0.0586	0.0729	0.2721	13.2215%
0.8	0.3250	0.3535	0.0591	0.0731	0.2716	13.1412%
0.9	0.3250	0.3536	0.0595	0.0733	0.2713	13.0791%
$\varepsilon_{\text{real}}$	0.2057	0.2899	0.2903	0.1190		

It is possible to see that the initial value starts to be more influential on the final result. It is interesting, on the other hand, to note that higher initial values perform better than lower ones. This is perhaps unexpected, since the error probabilities are themselves low. The explanation for this is connected with the function that we are minimizing; since it is highly complex and non-linear, nothing assures that starting in a point close to the optimal will lead faster to this one.

As before, we should compare how well this method performs for a different number of samples: see table 7.8; the initial value used was 0.1, used to test the algorithm (note that it is not supposed to work so well as with a higher value).

Considering the results obtained, it seems that working with a minimum of 2000 samples will be required for this estimation. With 0.9 as initializing value a minimum of 500 samples seems to work (as was the case for e).



(a) Q-term$_\varepsilon$ (b) $||\varepsilon' - \varepsilon_{\text{real}}||$

Figure 7.7: Lighter colors mark the lowest initial values

Table 7.8: Estimated emission probabilities for different number of samples

number of samples	errors (%)		
500	132.3922	22.9550	48.2251
1000	9.5298	13.9690	8.8385
2000	0.3395	2.8859	6.4910
5000	11.4220	7.7792	9.1856

7.3.3 Emission and Error Probabilities

Here will be discussed how well the estimation of the these two type of parameters is performed.

Considering that, as before, a new cancer DAG and samples were generated, the mean of the errors in the estimation of these parameters (for 4, 6 and 8 aberrations and 5 tests) are presented in table 7.9. Note that a new DAG is generated for each run. As it is possible to see from this table, since the numerical values given tend to oscillate, more tests should have been performed in order to be able to infer more reliable conclusions from this data.

7.3.4 Problems with this Method

One of the problems detected when working with this method was when dealing with sparsely populated DAGs.

Considering the example given in figure 7.8, the estimated values for when only the emission probabilities are being studied are $e' = (0.8422, 0.6565, 0.1617, 0.9354)$ which are clearly different from the real ones ($e = 0.9680, 0.7926, 0.6836, 0.9518$). This test was done for 500 runs and using the pseudocounts technique.

Table 7.9: Comparison of the errors in estimating the parameters (4,6 and 8 aberrations)

Number of samples	Errors (%)					
	4 aberr.		6 aberr.		8 aberr.	
	e	ε	e	ε	e	ε
500	2.3463	9.5257	2.8662	9.8275	1.4472	17.9473
1000	1.9711	12.1050	1.0230	6.1388	1.1051	3.4415
2000	0.3994	3.6644	1.9024	5.6091	1.2312	11.5465
5000	2.1404	5.1805	1.4946	7.3688	1.9888	9.0453

When considering the aberrations intensity, $\lambda = (9, 10, 5, 9)$, it is possible to see that it is the aberration with the lowest intensity that also has the worst estimation. One possible explanation for this is that, for this kind of DAGs, the *information* available to estimate the parameters (like how they relate with each other) is not so comprehensive as with not so sparsely populated DAGs. This would imply that not-so-probable aberrations would be even more underrepresented.

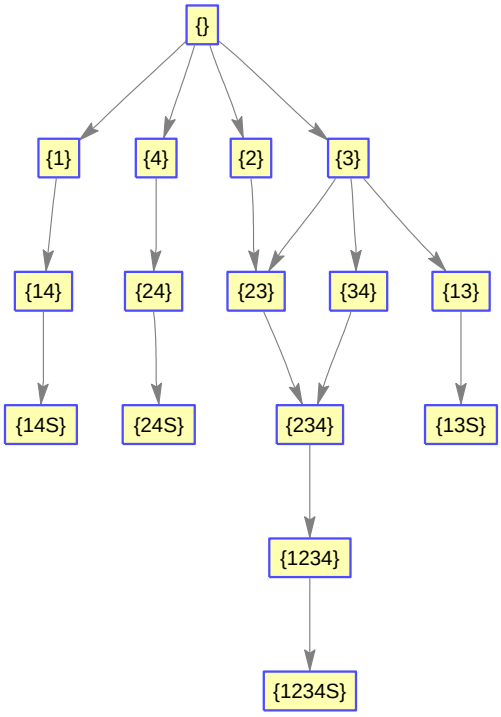


Figure 7.8: Sparsely populated DAG

Chapter 8

Conclusions and Future Work

This report presents a solid theoretical construction of how to model cancer progression and how to deal with the high number of parameters to estimate. Starting from biological studies of how cancer progresses and trying to model this evolution in a mathematical and precise way, the method followed here adapts Hidden Markov Models' theory to the study of how the different aberrations that trigger a cancer relate to each other. This model is different from past models in the sense that it takes into account hidden data – in other words, data that is not always observed.

The algorithm developed in this project (namely, the application of the EM algorithm to infer the hidden parameters, which is also a novelty introduced here) was tested using a synthetic data generator based on the cancer model. This algorithm seems to perform very well, which makes us believe that it can be useful when analyzing cancer samples from real-life examples. This holds as long as our Hidden Markov model is a suitable representation of the way a real cancer evolves. Unfortunately, this is not possible to test since we do not have access to which aberrations were the first (and when they occurred) in a cancer sample before it is detected by a medical doctor. This is also one of the reasons why this problem is hard to solve. On the other hand, since the model is based on biological studies, it is possible to expect it to represent this progression in a meaningful way. The analysis made in this report is important since the study of cancer progression allows, for example, scientists to devise better treatments for the possible upcoming aberrations in a given patient.

There are, however, some problems with this method, especially when the graph that represents cancer progression is too sparse, meaning there is lack of information. New methods should be devised in order to solve this problem, for example, changing the method used to reduce the number of parameters, i.e., the transition parameters λ, δ and ψ. This seems to be more problematic in less frequent aberrations.

Since the algorithm was implemented in MATLAB, the performance was significantly slower than expected. In retrospect, it seems that coding in C++ would have been a better choice. This and the problem referred in the last paragraph are also the reasons why a rigorous estimation of the transition parameters was not

possible, a question that should also be addressed in a future work.

After solving these two problems, a study with real-life cancers should be very interesting, possibly allowing medical doctors to increase their chances when selecting cancer treatments.

Appendix A

Jensen's Inequality

The proof provided here is based on [25].

Consider f to be a convex function defined on an interval I, with $x_1, x_2, ..., x_n \in I$ and $\lambda_1, \lambda_2, ..., \lambda_n \geq 0$ with $\sum_{i=1}^{n} \lambda_i = 1$,

$$f(\sum_{i=1}^{n} \lambda_i x_i) \leq \sum_{i=1}^{n} \lambda_i f(x_i).$$

For $n = 1$ this is trivial. For $n = 2$, we have:

$$\begin{aligned} f(\lambda_1 x_1 + \lambda_2 x_2) &\leq \lambda_1 f(x_1) + \lambda_2 f(x_2) \\ = f(\lambda_1 x_1 + (1-\lambda_1) x_2) &\leq \lambda_1 f(x_1) + (1-\lambda_1) f(x_2), \end{aligned} \tag{A.1}$$

which corresponds to the definition of convexity.

Proceeding, by induction,

$$\begin{aligned} f(\sum_{i=1}^{n+1} \lambda_i x_i) &= f(\sum_{i=1}^{n} \lambda_i x_i + \lambda_{n+1} f(x_{n+1})) \\ &= f[(1-\lambda_{n+1}) \frac{1}{1-\lambda_{n+1}} \sum_{i=1}^{n} \lambda_i x_i + \lambda_{n+1} f(x_{n+1})] \\ &\leq \lambda_{n+1} f(x_{n+1}) + (1-\lambda_{n+1}) f\Big(\frac{1}{1-\lambda_{n+1}} \sum_{i=1}^{n} \lambda_i x_i\Big) \\ &= \lambda_{n+1} f(x_{n+1}) + (1-\lambda_{n+1}) f\Big(\sum_{i=1}^{n} \frac{\lambda_i}{1-\lambda_{n+1}} x_i\Big) \\ &\leq \lambda_{n+1} f(x_{n+1}) + (1-\lambda_{n+1}) \sum_{i=1}^{n} \frac{\lambda_i}{1-\lambda_{n+1}} f(x_i) \\ &= \lambda_{n+1} f(x_{n+1}) + \sum_{i=1}^{n} \lambda_i f(x_i) \\ &= \sum_{i=1}^{n+1} \lambda_i f(x_i), \end{aligned}$$

which proves that our claim is true for all natural numbers. Considering that $-\ln(x)$ is convex (this can be proven by A.1), we have:

$$\ln \sum_{i=1}^{n} \lambda_i f(x_i) \geq \sum_{i=1}^{n+1} \lambda_i \ln(x_i)$$

which is the result used in 4.2.2.

Bibliography

[1] M. Hjelm, M. Höglund, and J. Lagergren, "New probabilistic network models and algorithms for oncogenesis," *Journal of Computational Biology*, vol. 13, no. 4, pp. 853–865, 2006.

[2] W. H. Organization, "Cancer," 2009. Retrieved 2010-05-05.

[3] G. L. Patrick, *An Introduction to Medicinal Chemistry, Fourth Edition*. Oxford Univ Press, 2009.

[4] L. Foulds, "The experimental study of tumor progression: a review," *Cancer research*, vol. 14, no. 5, p. 327, 1954.

[5] D. Hanahan and R. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, no. 1, pp. 57–70, 2000.

[6] L. Loeb, "Mutator phenotype may be required for multistage carcinogenesis," *Cancer Research*, vol. 51, pp. 3075–3079, 1991.

[7] J. Rahnenfuhrer, N. Beerenwinkel, W. Schulz, C. Hartmann, A. Von Deimling, B. Wullich, and T. Lengauer, "Estimating cancer survival and clinical outcome based on genetic tumor progression scores," *Bioinformatics*, vol. 21, no. 10, p. 2438, 2005.

[8] R. Desper, F. Jiang, O. Kallioniemi, H. Moch, C. Papadimitriou, and A. Schäffer, "Distance-based reconstruction of tree models for oncogenesis," *Journal of Computational Biology*, vol. 7, no. 6, pp. 789–803, 2000.

[9] P. Nowell, "The clonal evolution of tumor cell populations," *Science*, vol. 194, no. 4260, p. 23, 1976.

[10] R. Desper, F. Jiang, O. Kallioniemi, H. Moch, C. Papadimitriou, and A. Schäffer, "Inferring tree models for oncogenesis from comparative genome hybridization data," *Journal of Computational Biology*, vol. 6, no. 1, pp. 37–51, 1999.

[11] E. Fearon and B. Vogelstein, "A genetic model for colorectal tumorigenesis," *Cell*, vol. 61, no. 5, pp. 759–767, 1990.

[12] T. Kuukasjarvi, R. Karhu, M. Tanner, M. Kahkonen, A. Schäffer, N. Nupponen, S. Pennanen, A. Kallioniemi, O. Kallioniemi, and J. Isola, "Genetic heterogeneity and clonal evolution underlying development of asynchronous metastasis in human breast cancer," *Cancer research*, vol. 57, no. 8, p. 1597, 1997.

[13] P. T. Simpson, J. S. Reis-Filho, T. Gale, and S. R. Lakhani, "Molecular evolution of breast cancer," *Journal of pathology*, vol. 205, 2, pp. 248–254, 2005.

[14] F. Michor, Y. Iwasa, and M. A. Nowak, "Dynamics of cancer progression," *Nat Rev Cancer*, vol. 4, pp. 197–205, 2004.

[15] N. Beerenwinkel, J. Rahnenführer, M. Däumer, D. Hoffmann, R. Kaiser, J. Selbig, and T. Lengauer, "Learning multiple evolutionary pathways from cross-sectional data," *Journal of Computational Biology*, vol. 12, no. 6, pp. 584–598, 2005.

[16] R. Simon, R. Desper, C. Papadimitriou, A. Peng, D. Alberts, R. Taetle, J. Trent, and A. Schäffer, "Chromosome abnormalities in ovarian adenocarcinoma: III. Using breakpoint data to infer and test mathematical models for oncogenesis," *Genes, Chromosomes and Cancer*, vol. 28, no. 1, pp. 106–120, 2000.

[17] M. Radmacher, R. Simon, R. Desper, R. Taetle, A. Schäffer, and M. Nelson, "Graph models of oncogenesis with an application to melanoma," *Journal of Theoretical Biology*, vol. 212, no. 4, pp. 535–548, 2001.

[18] A. Tofigh, *Using trees to capture reticulate evolution.* PhD thesis, KTH School of Computer Science and Communication. Stockholm, Sweden, 2009.

[19] M. Gerstung, M. Baudis, H. Moch, and N. Beerenwinkel, "Quantifying cancer progression with conjunctive Bayesian networks," *Bioinformatics*, vol. 25, no. 21, p. 2809, 2009.

[20] M. Höglund, D. Gisselsson, T. Säll, and F. Mitelman, "Coping with complexity multivariate analysis of tumor karyotypes," *Cancer genetics and cytogenetics*, vol. 135, no. 2, pp. 103–109, 2002.

[21] L. Rabiner, "A tutorial on hidden Markov models and selected applications inspeech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[22] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis.* Cambridge Univ. Press, 1998.

[23] D. Ocone, "Discrete and probabilistic models in biology," 2009. Retrieved 2010.05.21.

[24] J. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," *International Computer Science Institute*, vol. 4, 1998.

[25] S. Borman, "The Expectation Maximization Algorithm A short tutorial," *Unpublished paper available at http://www.seanborman.com/publications*, 2004.

[26] A. Dempster, N. Laird, D. Rubin, *et al.*, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[27] S. Heim and F. Mitelman, *Cancer cytogenetics.* Wiley-Blackwell, 2009.

[28] N. Beerenwinkel, N. Eriksson, and B. Sturmfels, "Evolution on distributive lattices," *Journal of theoretical biology*, vol. 242, no. 2, pp. 409–420, 2006.

[29] J. Liu, J. Mohammed, J. Carter, S. Ranka, T. Kahveci, and M. Baudis, "Distance-based clustering of CGH data," *Bioinformatics*, vol. 22, no. 16, p. 1971, 2006.