Aalto University
School of Science and Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science

Prem Raj Adhikari

Mixture Modelling of Multiresolution 0-1 Data

Master's Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology.

Espoo, November 29, 2010

Supervisor: Professor Samuel Kaski
Instructor: Jaakko Hollmén, D.Sc. (Tech.)

AALTO UNIVERSITY SCHOOL OF SCIENCE AND TECHNOLOGY

ABSTRACT OF THE MASTER'S THESIS

Author:	Prem Raj Adhikari
Name of the Thesis:	Mixture Modelling of Multiresolution 0-1 Data
Date:	November 29, 2010
Number of pages:	xi + 92
Department:	Department of Information and Computer Science
Professorship:	T-61 Computer and Information Science
Supervisor:	Professor Samuel Kaski
Instructor:	Jaakko Hollmén, D.Sc. (Tech.)

Biological systems are complex and measurements in biology are made with high throughput and high resolution techniques often resulting in data in multiple resolutions. Furthermore, ISCN [1] has defined five different resolutions of the chromosome band. Currently, available standard algorithms can only handle data in one resolution at a time. Hence, transformation of the data to the same resolution is inevitable before the data can be fed to the algorithm. Furthermore, comparing the results of an algorithm on data in different resolutions can produce interesting results which aids in determining suitable resolution of data. In addition, experiments in different resolutions can be helpful in determining the appropriate resolution for computational methods.

In this thesis, one method for upsampling and three different methods of downsampling 0-1 data are proposed, implemented and experiments are performed on different resolutions. Suitability of the proposed methods are validated and the results are compared across different resolutions. The proposed methods produce plausible results showing that the significant patterns in the data are retained in the transformed resolution. Thereafter, the mixture models are trained on the data original data and the results are analyzed. However, machine learning methods such as mixture models require high amounts of data to produce plausible results. Therefore, the major aim of the data transformation procedure was the integration of databases. Hence, two different datasets available in two different resolutions were integrated after transforming them to a single resolution and mixture models were trained on them. Trained models can be used to classify cancers and cluster the data. The results on integrated data showed significant improvements compared with the data in the original resolution.

Keywords: mixture models, multiresolution data, 0-1 data, model selection, cross-validation, chromosomal aberration, upsampling, downsampling, cancer genetics.

Acknowledgments

My Master's thesis has been carried out being a part of the Parsimonious Modelling(PM) group in Helsinki Institute for Information Technology (HIIT), Department of Information and Computer Science(ICS) in the Aalto University School of Science and Technology(TKK). First of all, I would like to thank my instructor DSc.(Tech.) Jaakko Hollmén for his enthusiastic engagement in my research and his never ending stream of advice, ideas and support ranging from the minute technical details to the overall research ideas. A share of thanks also goes to the supervisor Prof. Samuel Kaski who invested his precious time in paper work and glancing manuscript. A big share of thanks also goes to the members of the PM group and the whole ICS, AIRC(Adaptive Informatics Research Center) laboratory and ALGODAN (Finnish Center of Excellence for Algorithmic Data Analysis Research) for providing splendid working and research environment.

I would also like to thank each and every professors, teachers and assistants involved in teaching and organizing MACADAMIA programme [2]. Initially, it was difficult but overall it has been a great learning experience. I would especially like to mention the names of Kai Puolamäki and Gemma C Garriga for their advice and suggestions during early part of MSc Degree studies. A big share of thanks also goes to my MACADAMIA mates Yao, Peter, Joel, Agha and Jing, the seniors especially Luis and Dušan and the juniors especially Kranthi and Mudassar.

I would also like to thank my mates, the seniors Aditya and Sandeep as well as Gautam, Tashi, Marko, Tuomas, Timo, Ashis Ji. I would also like to thank other guys in Terassi for making the place a lively place to live in. If it weren't for you guys, I might have graduated earlier but without you, I would never have graduated. Yao and Saurav should have the credit for proof-reading my thesis and helping me with the English grammar and spelling. All the remaining errors are to be blamed on me for the last minute changes. Last but not the least my gratitude and thanks goes to my parents, family and relatives for their everlasting love and support.

Espoo, November 29, 2010
Prem Raj Adhikari

Wherever you go, no matter what the weather, always bring your own sunshine.

— Anthony J. D'Angelo
The College Blue Book

Contents

Abbreviations and Notations

WHO	World Health Organization
ICT	Information and Communication Technologies
CGH	Comparative Genomic Hybridization
aCGH	Array Comparative Genomic Hybridization
BAC	Bacterial Artificial Chromosome
SNP	Single Nucleotide Polymorphisms
DNA	Deoxyribonucleic acid
DM	Data Mining
ML	Machine Learning
bp	Base Pairs
Mbp	Mega-base Pairs
kbp	Kilo-base Pairs
TB	Terabytes
GB	Gigabytes
cDNA	complementary Deoxyribonucleic acid
CNV	Copy Number Variation
ISCN	International System for human Cytogenetic Nomenclature
EM	Expectation Maximization
FMM	Finite Mixture Models
MCMC	Markov Chain Monte Carlo
SVD	Singular Value Decomposition
DAG	Directed Acyclic Graph
DGM	Directed Graphical Model

NCBI	National Center for Biotechnology Information
ANSI	American National Standards Institute
BPCR	Bayesian Piecewise Constant Regression
MAFIA	MAximal Frequent Itemset Algorithm
IQR	Interquartile Range
GHz	Gigahertz (unit of frequency)
CPU	Central Processing Unit
HMM	Hidden Markov Models

List of Figures

List of Tables

Chapter 1

Introduction

> *Science these days has basically turned into a data-management problem.*
>
> — Jimmy Lin
>
> *Associate Professor, University of Maryland*

1.1 Machine Learning in Cancer Research

Cancer

Cancer (Medically: *Malignant Neoplasm*) is a disease characterized by the abnormal and uncontrolled growth of cells; their ability to migrate to other parts of human body and destroy the neighboring cells and tissues [3]. The lack of proper care can be fatal in cancer cases. Consider, for example, some statistics: cancer caused 7.4 million deaths worldwide (13% of the total deaths) in 2004 [4]. In the United States, cancer accounted for 0.56 million deaths (23.1% of all deaths) in 2005 [5]. Finland also has a high number of cancer cases; 26,279 new cancer cases were reported in 2007, by 2015 it is expected to reach 30,000 [6]. It is estimated that more than one-third of the population will develop some form of cancer during their lifetime. Cancers can appear at any age but is more common in the older population. As people have started living longer, the problems with cancer is bound to increase in the near future. As a result of the appalling effect of cancer and their growing rate, cancer is

highly researched through diversified aspects and areas.

Ever since the concept of Evidence based medicine was promulgated in the early 1960s by a Scottish Professor Archie Cochrane [7] in his book "*Effectiveness and Efficiency: Random Reflections on Health Services*", a variety of different engineering tools and techniques have been used in medicine. Several ICT (Information and Communication Technologies) and computational methods such as Telemedicine, Medical imaging, Electronic Patients records have already been deployed with excellent results in hospitals and medical centers. Recently, there has been tremendous improvements in technology especially in hardware and software related to computers. Computers also have increased processing speed and storage space. Furthermore, ultramodern computer architecture, and improved communication and Internet have increased the ease of manipulation and sharing of data and resources among different communities and regions. Thus, data related to diseases, such as cancer, is efficiently stored and readily available.

On the other hand, biological systems are very complex. Technology has not only enabled storage of data, it has also provided means to study the complex biological system. Microarray technology, such as CGH (Comparative Genomic Hybridization) [8] and aCGH (Array Comparative Genomic Hybridization) [9] offer the facilities to study the genomes and the genes in human. CGH is one of the molecular techniques to survey the DNA copy number variation across the whole genome. In CGH experiment, differentially labeled test and reference genomic DNA (Deoxyribonucleic Acid) are cohybridized to normal metaphase chromosomes. Fluorescence ratios along the length of the chromosome provide a cytogenetic representation of DNA copy number variation. However, one major drawback of CGH is the resolution. The mapping resolution is only 20Mbp (Mega-Base Pairs) i.e. the smallest measurable detail is 20Mbp. In addition to that mapping resolution for deletion is 2Mbp. To overcome the problem of CGH, a new microarray technology called aCGH has been developed. aCGH provides higher resolution than CGH. Fluorescence ratios at arrayed DNA elements provide a locus by locus measure of the copy number changes. Furthermore, a type of DNA arrays called BAC arrays (Bacterial Artificial Chromosome) covers the whole genome in an overlapping manner consisting of as many BAC clones as necessary (which is $\approx$32400 for

the human genome) [10]. DNA arrays also includes Oligonucleotide arrays [11] and promoter arrays [12]. Oligonucleotide arrays and cDNA arrays are generally used for gene expression analysis (determining the expression level of each gene). Oligonucleotide arrays also find their application in SNP (Single Nucleotide Polymorphism) analysis. Promoter arrays are often used to identify transcription factor binding sites. Next generation sequencing [13, 14, 15] provides an opportunity for high-throughput sequencing producing data at exponential rates.

These technologies have varying uses including the gene expression analysis, detecting aberrations in genes and chromosomes and have a positive impact on cancer research. Furthermore, completion of the Human Genome Project [16, 17] in 2003 has opened an interesting area of research in computational genomics. The most common aspect of all these techniques is that they produce data in astronomical proportions. For instance, the third generation of DNA sequencers [14, 15] will generate many petabytes[1] of information a year. The introduction and application of these methods in cancer research have led to the accumulation of data at exponential rates. Hence, there is an urgent need to understand complex biological systems from this huge amount of data which involves the analysis of the data exploded by those experiments. This is where a relatively new field of ML (Machine Learning) and DM (Data Mining) is increasingly finding its way in the medical field, especially in the cancer research.

Machine Learning and Data Mining

Machine learning is a branch of artificial intelligence incorporating a myriad of statistical, probabilistic and optimization techniques allowing computers to learn from past examples to help detect and discover meaningful patterns from large, noisy and complex data sets [18, 19, 20]. Machine learning encompasses a variety of methods, including classification, regression, clustering, and pattern discovery with varying applications such as object recognition in computer vision, natural language processing, medical diagnosis, bioinformatics, brain-machine interfaces, classifying DNA sequences, speech and handwriting

[1] 1 petabyte is equal to 1024 TB (terabytes) or 1,048,576 GB (gigabytes).

recognition. The machine learning and data mining, although a relatively new field, its community has already developed a cohort of many fascinating algorithms, interesting ways to handle the concept classes and elegant and clever ways to search through huge databases. The medical field can, therefore, reap the benefit of these methods and adapt these methods for analysis of ever increasing medical data.

Recently, machine learning methods are increasingly used in cancer research because of its versatility, the sheer volume of data generated by the biological experiments, dramatic growth in new scientific questions, and new challenges for learning and inference. The presence of massive population-wide, lifelong, trans-generational, and electronically accessible datasets obligates the use of machine learning and data mining methods in health-care and medicine. Different classification methods are used for cancer diagnosis, clustering for prognosis and tumor class discovery, and feature selection for biomarker identification [21]. The concept of **personalized medicine**[2], which is essentially a set of methods to map diagnostic results to therapies in cancer cases, has led to the application of different novel machine learning methods. Furthermore, a variety of new scientific and clinical problems introduced almost everyday necessitate the development of novel supervised and unsupervised learning methods to use these growing resources in terms of data and knowledge. Cancer genomics is a highly researched area producing significant amount of data and questions for the research.

1.2 Chromosomal Aberrations

It is important to note that cancer is a multifactorial[3] disease as shown in [22]. For example, it is well known that smoking causes cancer, but all cancers are not caused by smoking and all the people who smoke will not develop cancer. However, all the cancer cases incorporate some form of genetic changes in

[2]One United States Senate Bill (proposed law) defines personalized medicine as *the application of genomic and molecular data to better target the delivery of health care, facilitate the discovery and clinical testing of new products, and help determine a person's predisposition to a particular disease or condition.*

[3]Here multifactorial is used to mean there are many factors causing cancer. The majority of the noninfectious diseases are multifactorial.

human beings. Humans have 23 (22, X and Y) pairs of chromosomes. Humans being a diploid organism have two homologous copies of each chromosome, usually, one inherited from the father and the other from the mother. During the complex process of cell division, some abnormalities can occur in the cells and copy number changes from two [23]. These changes are often referred to as CNV (Copy Number Variation). The reasons for such abnormalities have not been identified yet but even the latest studies [24] believe in the abnormality of chromosomes as a cause of cancer. It is, however, important to note that fork stalling and template switching, a replication misstep, has been attributed to such abnormalities [25]. Deletion, often referred to as loss, is the case when the copy number is less than two. Duplication, often referred to as gains, is the case when the copy number is more than two. Amplification is the special case of duplication where the copy number increases more than 5. Chromosomal aberrations such as DNA amplification, deletion and duplication have significant roles in cancer research [23]. Some amplifications have shown more than 100 copies. DNA copy number amplifications have been defined as the hallmarks of cancer [26].

1.3 Multiple Resolutions of Genome

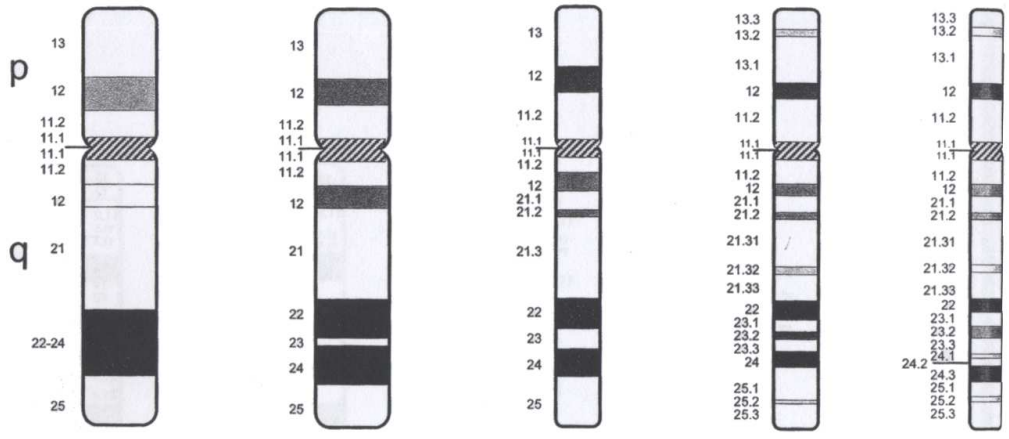

Figure 1.1: G-banding patterns for normal human chromosomes at five different levels of resolution. Source: Shaffer et. al. 2009 [1]. Example case in Chromosome 17.

Biological experiments performed with high throughput and high resolution techniques often produce data in multiple resolutions. Furthermore, ISCN (International System for human Cytogenetic Nomenclature) has defined five different resolutions of the chromosome band: 300, 400, 550, 700 and 850 [1]. In other words, chromosomes are divided into 862 regions in resolution 850 (fine resolution) and 393 regions in resolution 400 (coarse resolution). Figure 1.1 shows the G-banding patterns showing five different resolutions in chromosome 17. For example, chromosome 17 in resolution 300 is divided into 10 parts while in resolution 850, the same chromosome 17 is divided into 24 different parts. Division of the regions is irregular and varied for different regions. Some regions are not divided at all where as some other regions are divided into many different parts. For example, in chromosome 17, the region 17q22 is not divided at all in resolution 400, 550, 700 and 850. However, region 17q21 is divided differently in resolutions 400, 550, 700 and 850. Furthermore, different staining techniques produce chromosome bands in different resolution. However, typically computational algorithms work with only single resolution of the chromosome. Thus, data is available in different resolutions thus necessitating new methods to be devised to work with the multiple resolutions of the data. Currently, the general principle for working in multiple resolution has been to work independently on two different resolutions and get the separate results and at best compare them. The improvement on the above principle is to transform the data to a common representation and apply the machine learning algorithm to the data in the same representation. We implement both the principles in this thesis. Furthermore, the models that directly learn from multiple resolutions of data can be developed, which is left as future work as a perspective post-graduate studies.

Working with multiple resolutions of data is important for the database integration and utilization of all the data and other available resources in multiple resolutions. Furthermore, comparing the results of an algorithm on data in different resolutions can produce interesting results which aid in determining suitable resolution of data. In addition, experiments in different resolutions can be helpful in determining the appropriate method for staining. Furthermore, machine learning and data mining algorithms and methods are in most cases data hungry and require significantly large amount of data for plausible

results. Thus, database integration is important to work with high dimensional data having small number of samples. For example, the validation technique cross-validation used in this thesis has been shown not to work very well with small sized data samples in [27, 28]. Multiresolution data occurs naturally in various fields such as telecommunication industry, image processing; thus working with multiresolution data can be interdisciplinary and signifies the importance of working with multiresolution data.

In this thesis, upsampling, a technique to transform the data from coarse resolution to fine resolution, and downsampling, a technique to transform the data from fine resolution to coarse resolution of chromosome bands, is used to transform the data in different resolutions to a single resolution which are explained in detail in Chapter 3. Then it presents a mixture modelling approach to reveal the structure in the chromosomal aberrations of cancer patients. The use of mixture models is motivated by the fact that cancer is not a single disease but a collective term for a class of diseases with some similarity. As the classes are different, the causes of cancers also differ among different types of cancers. Mixture models usually thrive in modeling such heterogeneous data generated from different classes. These models can be used to develop generic models to combine the samples from different sub-populations[4]. The model based clustering approach is used to optimally divide the data into clusters. Cross-validation technique is used to learn the model i.e. the number of subpopulation that the data supports. The parameters of the mixture models are learned from the data using the Expectation Maximization (EM) algorithm [29, 30]. The chromosomewise modeling generates a probability distribution to express the amplification patterns in cancer for each chromosome. This probability distribution can be used for the classification of different types of cancer. The chromosomal aberrations dataset analyzed in this thesis uses very scarce data as explained in Section 4.2. Thus, we decided to work chromosomewise because of the availability of very few samples of the data to constrain the complexity of the mixture models.

[4]Subpopulation is used here to mean different types of cancers. Each subpopulation represents a type of cancer

1.4 Outline of the Thesis

Chapter 1 introduces the topic of the thesis with motivations for studying cancer using machine learning methods. It also provides brief introduction to the problem of chromosomal aberrations in multiple resolutions. Chapter 2 covers the mixture models, Expectation Maximization (EM) algorithm and other relevant theoretical background required for the work in the thesis. Similarly, Chapter 3 focuses on the methods for upsampling and downsampling of chromosomal aberration data available in multiple resolutions. Chapter 4 discusses the various experiments performed and analyzes the results of experiments. Chapter 5 draws conclusions from experimental results and discusses potential future areas of research.

1.5 Contributions of the Thesis

The major contributions of the thesis are briefly summarized below:

1. Upsampling and downsampling methods to transform the genomic data to different resolution facilitating database integration.
2. The chromosomewise analysis of chromosomal aberrations in multiple resolutions using mixture models of multivariate Bernoulli distributions for the data in the same resolution.
3. Studying the behavior of the mixture models in different resolutions.
4. Investigation of the patterns in the multiple resolutions of data and the trained mixture models.

Chapter 2

Mixture Models and 0-1 data

> *The purpose of models is not to fit the data but to sharpen the questions.*
>
> — Samuel Karlin
>
> 11^{th} *R A Fisher Memorial Lecture (1983)*

Synopsis

This chapter is devoted to the introduction of the mathematical foundation of mixture models, special consideration is on the finite mixture models of multivariate Bernoulli[1] distributions. The chapter also covers EM algorithm [29, 30] and its formulation for the finite mixture models of multivariate Bernoulli distributions [29, 31]. The chapter also provides brief introduction to cross-validation, a method for accessing the results of statistical analysis. Near the end of the chapter, it shortly reviews the literature on the use of finite mixture models of multivariate Bernoulli distributions with a focus on cancer genetics. Part of work discussed in this chapter has been published in [32] and [33].

2.1 0-1 Data

History of collection of information and data is quite long. However, the size of data and information was relatively small. Recently improved technology,

[1]Bernoulli Distribution is named after Swiss scientist Jacob Bernoulli(1654-1705)

increased storage capacity, and more importantly the realization of importance of data has lead to the collection and storage of data. Moreover, as discussed in Chapter 1, recent technologies are producing data at exponential rates. Thus, extracting meaningful information from those data is a matter of extreme urgency. In all the fields of study ranging from biology through astronomy to social sciences, 0-1 data has been of special interest. 0-1 data is a special class of categorical data with only two scales which can be considered as true or false, success or failure. In other words, 0-1 data captures the dichotomy of two classes. 0-1 data have only two classes (categories) and often represented as **0** and **1** or **1** and **-1**. 0-1 data naturally occur in many areas of study: in social science, interview questions relating to marital status, gender, like or dislike, alive or dead can be formulated as 0-1 data. Similarly, in palaeontology 0 can represent absence of fossils and 1 can represent presence of fossils [34]. In universities, the relationship between courses and the students can be represented as 0-1 data where 1 represents that the student has taken the course and 0 represents that the student has not taken the course as discussed and preprocessed in [35]. One of the principal uses of the 0-1 data is in **'Market Basket Data'** which assembles information about whether a customer has bought certain goods or not. One of the popular benchmark dataset RETAIL is a prominent example of a market basket data [36]. Over the years biology and genetics, have been one of the major sources of 0-1 data. For example, 0-1 data can capture the notion of presence or absence of certain characteristics in species. 0-1 data analyzed in this thesis as discussed in Section 4.2 is also a 0-1 data denoting the presence or absence of chromosomal aberrations in chromosome bands.

2.2 Mixture Models

Probabilistic modeling aims to approximate the probability of an event occurring again on the basis of limited instances of observed data. The estimated probability distribution aims to explain the process of data generation. FMM (Finite Mixture Models) are probabilistic models with varying uses such as density estimation, clustering, classification [20, 31, 37]. These models belong

to an interesting and flexible model family for modelling latent (unobserved) variables in complex distributions. Finite Mixture Models have a very long history. Geoffrey McLachlan and David Peel in their book *"Finite Mixture Models"* attribute famous biometrician Karl Pearson for the first use of the mixture models where he fitted two Gaussians with different means (μ_1 and μ_2) and variances (σ_1^2 and σ_2^2) in proportions π_1 and π_2 for some data in 1894 [37]. However, the popularity of mixture models has significantly grown over the past few decades because of the dramatic increase in computing power. Nonetheless, the major share of contribution goes to the mathematical foundation, formulation and understanding of the mixture models. Furthermore, formulation of the EM algorithm [30], which provides a conceptual framework to estimate the maximum likelihood from the incomplete data, in 1977 provided the necessary impetus to the growing use of mixture models. Over the few years, finite mixture models have been extensively used in many application domains including model based clustering, classification, image analysis, and collaborative filtering in analysis of high dimensional data.

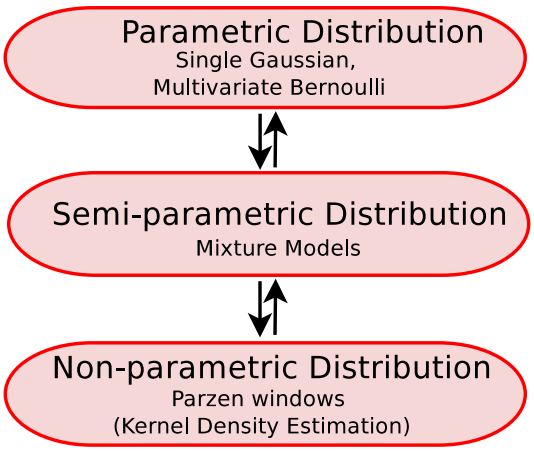

Figure 2.1: Schematic representation of different forms of distributions.

FMM (Finite Mixture Models) models a statistical distribution by a mixture (or weighted sum) of simple distributions such as Gaussian, Poisson and Bernoulli. It decomposes the density function into a set of component density functions. Each of the decomposed density functions defines a specific class of the origination of the data i.e each component density functions represents

a portion of original distribution. This form of representation is not possible with other simple parametric distributions. The basic assumption of FMM is that the different classes[2] in the data originate from the well-known parametric distributions. Except for this assumption of the data source, FMMs are extremely flexible in the choice of the distribution. Any classical parametric distributions such as Normal(Gaussian), Poisson [38], Bernoulli can be chosen as component density function. Unlike the case with one Bernoulli, determining the training sample contributing to a particular component is not possible. Hence, the methods based on mean and covariance matrix are not applicable to the mixture models.

After the choice of distribution, the primary task is then to estimate the parameters of the selected distribution such as mean (μ) and variance (σ^2) for Gaussian distribution; rate of occurrence (λ) for Poisson distribution [38]. It is important to note that each component distribution will be defined by its own set of parameters thus differing itself from the others. This explains the reason why mixture models are called semi-parametric models as depicted in the Figure 2.1. The complexity of mixture models depends on the complexity of the problem being solved, not the size of dataset. In this thesis, 0-1 data is analyzed and the assumption is that it follows the Bernoulli distribution. Bernoulli distribution of a single random variable is parameterized by one parameter θ which denotes the probability of success in a trial with two possible outcomes: success and failure. The learning task is then limited to learning the Bernoulli parameter θ.

Advantages of Mixture Models

Mixture Models have various merits and are often a suitable choice for modelling data. Some of the most useful characteristics of mixture models can be summarized as the following:

- A mixture model learns the structure in the data better than most other methods as the different component distributions capture the dominant patterns present in the data.

[2]The class here is not similar to the class labels.

- Learning mixture models involve well studied statistical inference techniques [37].
- Mixture models are flexible in terms of the choice of the component distributions.
- Mixture Models can generate leptokurtic distributions from mesokurtic ones [39].
- Mixture Models can also generate skewed distributions from symmetric components [39].
- It is suitable for any form of data either discrete or continuous.
- When mixture models are used in clustering, the components represent the clusters thus making it possible to obtain density estimation for each cluster.
- Mixture models also provides the facilities for soft classification [39].

Mixture models are flexible models and have varying uses. Some of the basic areas where mixture models are most prevalent are:

- **Clustering:** Mixture models are at the heart of model based clustering where each component denotes one cluster.
- **Handling Missing Data:** Mixture models have also been extensively used to handle the missing data for building the model [37].
- **Density Estimation:** In Bayesian statistics, mixture models can be used to assign the flexible priors [37].
- **Model Averaging:** Mixture models have often been used to combine different density models [20].
- **Modelling Heterogeneity:** Here in this thesis mixture models have been used to model the heterogeneous cancer cases in different patients.

Mixture of Multivariate Bernoulli Distributions

The major focus of the thesis is concentrated on modelling DNA copy number aberrations which is 0-1 data. Hence, the mixture of multivariate Bernoulli distributions forms the crux of the thesis.

Univariate Bernoulli distribution is a probability distribution with two possible outcomes: success and failure [40]. Consider an example of a single random binary variable, $x \in \{0, 1\}$ where $x = 0$ denotes the failure of an event and $x = 1$ denotes the success of an event or other similar dichotomy such as success or failure of an event and the coin tossing. For example, success of an event may be a student participating a course and failure of an event may be the student not participating in the course [35]. Let the probability of occurrence of $x = 1$ be θ such that $0 \leq \theta \leq 1$. Therefore, the probability of occurrence of $x = 0$ is $1 - \theta$. Thus, $p(x = 1|\theta) = \theta$ and $p(x = 0|\theta) = 1 - \theta$. Accordingly, the probability mass function i.e. probability distribution [40, 20] over x is given by the equation

$$p(x|\theta) = \theta^x(1 - \theta)^{1-x} \tag{2.1}$$

The mean or the expected value of the random binary variable is given by

$$\mathbb{E}[x] = 0 \times p(x = 0|\theta) + 1 \times p(x = 1|\theta) = p(x = 1|\theta) = \theta \tag{2.2}$$

The variance of the random binary variable is defined as the dispersion of random variable. It can be obtained by

$$var(x) = \mathbb{E}(x^2) - \{\mathbb{E}(x)\}^2 \tag{2.3}$$

where

$$\mathbb{E}(x^2) = 0^2 \times p(x = 0|\theta) + 1^2 \times p(x = 1|\theta) = p(x = 1|\theta) = \theta$$

and also

$$\{\mathbb{E}(x)\}^2 = \theta^2$$

Therefore,

$$var(x) = \theta - \theta^2 = \theta(1-\theta) \tag{2.4}$$

The probability $p(x|\theta)$ can be extended to the binary space $\{0,1\}^N$ i.e. to a dataset $\underline{X} = \{\overline{X}_1, \ldots, \overline{X}_d\}$ and $\overline{X}_1 = (X_{11}, X_{12} \ldots X_{1d})$ [20]. Here, $(X_{11}, X_{12} \ldots X_{1d})$ are the observed values of $\underline{X}$. Hence, the probability mass function of the multivariate Bernoulli distribution is given by

$$P(\mathcal{D}|\Theta) = \prod_{i=1}^{\mathrm{d}} p(x_i|\theta) = \prod_{i=1}^{\mathrm{d}} \theta_i^{x_i}(1-\theta_i)^{1-x_i} \tag{2.5}$$

where $\theta \in \mathbb{R}^i$ and $0 \leq \theta_i \leq 1$ for all $1 \leq i \leq \mathrm{d}$ and $x_1, x_2, \ldots x_{\mathrm{d}} = \mathbf{x} \in \{0,1\}^{\mathrm{N}}$

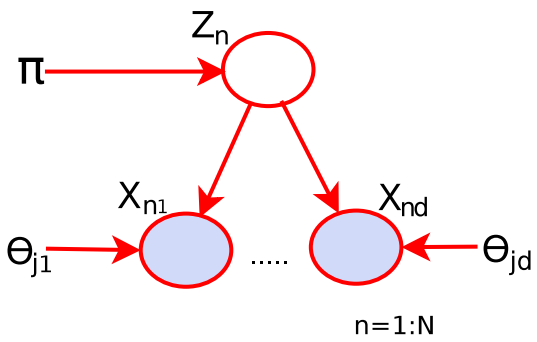

Figure 2.2: A graphical mixture model of mixture of Bernoulli.

This can be represented as the DGM (Directed Graphical Model), which is a type of DAG (Directed Acyclic Graph) [20] as shown in Figure 2.2 which is similar to Naive Bayes classifier except that the class labels Z_n is hidden.

The likelihood function in Equation (2.5) is a function of θ. For independent and identically distributed samples $\overline{X} = \{x_n\}_{n=1}^{\mathrm{N}}$ from $\{0,1\}^{\mathrm{N}}$, the vector $\hat{\theta}$ that maximizes the likelihood function in Equation (2.5) is the estimated value of θ. The joint probability for the N samples of data is given by:

$$\begin{aligned} \ln\left(P(X_1, X_2 \ldots X_N)\right) &= \ln \prod_{i=1}^{\mathrm{N}} p(x_i) \\ \ln \prod_{i=1}^{\mathrm{N}} p(x_i) &= \sum_{i=1}^{\mathrm{N}} \ln p(x_i) \end{aligned} \tag{2.6}$$

Furthermore, maximizing the likelihood function in Equation (2.5) equivalent to maximizing the logarithm of the likelihood. Thus,

$$\ln p(\mathcal{D}|\Theta) = \sum_{i=1}^{\mathrm{d}} ln\ p(x_i|\theta_i) = \sum_{i=1}^{\mathrm{d}} x_i\ ln\ \theta_i + (1 - x_i)(1 - \theta_i) \tag{2.7}$$

From Equation (2.7) it can be seen that the log likelihood function depends on the d samples of x_d through the sum $\sum_{i=1}^{\mathrm{d}} x_n$ which provides adequate statistics about the distribution. Taking the derivative of (2.7) with respect to θ and equating it to zero gives the value of maximum likelihood estimation. The value is given by:

$$\hat{\theta}_{\mathrm{ML}} = \frac{1}{N} \sum_{i=1}^{\mathrm{N}} x_i \tag{2.8}$$

The Definition (2.8) is also known as the sample mean. If the sample $\overline{X}$ contains higher order correlations, the sample covariance matrix will be diagonal. Hence, the maximum likelihood estimator in Equation (2.8) gives unsatisfiable result.

Assuming that the data comes from a mixture of known number of the components, J, finite mixture of multivariate Bernoulli distributions is defined as:

$$p(\mathcal{D}|\Theta) = \sum_{j=1}^{\mathrm{J}} \pi_j P_j(x|\theta_j) \tag{2.9}$$

In Definition (2.9) each P_j is a multivariate Bernoulli Distribution parameterized by θ_j. Hence, the finite mixture model for multivariate Bernoulli distribution can be formulated as:

$$p(\mathcal{D}|\Theta) = \sum_{j=1}^{\mathrm{J}} \pi_j \prod_{i=1}^{\mathrm{d}} \theta_{ji}^{x_i}(1-\theta_{ji})^{1-x_i} \tag{2.10}$$

where π_j are the mixture proportions satisfying the properties such as convex combination such that $\pi_j \geq 0$ and $\sum_{j=1}^{\mathrm{J}} \pi_j = 1$ for all $j = 1, \ldots J$. The model parameters, Θ, is composed of $\theta_1, \theta_2, \theta_3 \ldots \theta_d$ for each component distribution. The combination of J mixtures of multivariate distribution in Equation (2.10) can capture the correlations (the clustering structure) in the sample $\overline{X}$ thus solving the problem of unsatisfiable result in Equation (2.8). Finite mixture of multivariate Bernoulli distributions with number of components equals to J and dimension of dataset $= d$ is parametrized by $\Theta = \{J,\ \{\pi_j, \theta_j\}_{j=1}^{J}\}$ for each component distribution.

Fitting the Bernoulli Mixture Model involves learning the parameters Θ and the number of components J from the given data sample $\overline{X}$. This can be formulated in terms of loglikelihood as:

$$\mathcal{L}(\Theta) = \sum_{n=1}^{\mathrm{N}} log\ P(x_n|\Theta) = \sum_{n=1}^{\mathrm{N}} log \left[\sum_{j=1}^{\mathrm{J}} \pi_j \prod_{i=1}^{\mathrm{d}} \theta_{ji}^{x_{ni}}(1-\theta_{ji})^{1-x_{ni}} \right] \tag{2.11}$$

The Equation (2.11) can be maximized with high number of mixture components i.e. the mixture models gets high likelihood values for the training set. However, large number of mixture components increases model complexity and often results in overfitted model generalizing poorly on future data. On the other hand, smaller number of mixture components results in underfitting. To find the trade-off between the appropriate model complexity and large value of the Equation (2.11) some validation techniques must be used. The basic aim of the thesis is to achieve maximally simple and compact parsimonious models. A parsimonious models are the models having as few parameters as

possible for a given quality of a model. There are different principles for developing parsimonious models such as Ockham's razor [41]. In this thesis, 10-fold cross-validation discussed in Section 2.4 is used for the same purpose. The maximization of the Equation (2.11) can be performed by using EM algorithm discussed in Section 2.3.

One of the major drawbacks of finite mixture models of multivariate Bernoulli distributions is that it belongs to the class of non-identifiable distributions [42]. Thus, there exists distinct parameters $(\alpha\ ,\ \theta)$ and $(\beta\ ,\ \lambda)$ such that they represent same distribution excluding the trivial permutations. The problem of non-identifiability has been extensively studied in literature after it was proved in [42, 43] that these FMMs are non-identifiable. However, studies in [43] have proved that in spite of their non-identifiable nature, they are useful in various applications.

Challenges in Using Mixture Models

In spite of great virtues of mixture models, there are several major challenges in the estimation of mixture models. The mixture models require that the number of components be known *apriori*. Even if the models are known *apriori*, it is often difficult to reliably distinguish different components. In worst case scenario, some of the components may simply converge to the outliers present in the data. It is important to note that selection of the number of mixture components directly influences the performance of the mixture models. Lesser the number of components, the mixture model behaves similar to a simple parametric model and increases the bias. On the contrary, if the mixture model has a large number of components, the model can overfit the data thus producing unreasonable variation. Hence, there is always a trade-off between the two. Secondly, the likelihood function may have multiple local maxima. In order to address these challenges we use 10-fold cross-validation repeated 50 times so that we get the optimal results. Thirdly, the major drawback in using mixture models is the computational complexity of training the mixture models. Normally, training mixture models is computationally expensive when compared with other parametric (such as Poisson distribution [38]) as well as non-parametric (such as k-means [44, 45]) methods.

2.3 Expectation Maximization Algorithm

Different methods have been proposed and implemented to estimate the parameters of the mixture model including EM (Expectation Maximization) [29, 30], MCMC (Markov Chain Monte Carlo) [46], and Spectral Method [47, 48]. MCMC uses Gibbs sampling to sample from posterior distribution. Spectral method, on the other hand, uses SVD (Singular Value Decomposition) [49, 50] on the data. For distributions satisfying specific separation condition, spectral method estimates the mixtures highly similar to the true mixture with high probability [48]. However, in this thesis EM algorithm, is used to estimate the parameters of the mixture model in a cross-validation setting to justify the selection of the number of component distributions.

Given a sample $\overline{X}$, the parameters maximizing Θ and J can not be ascertained analytically. However, EM algorithm can be used to optimize the parameters. The Expectation Maximization (EM) is an iterative algorithm for the computation of maximum likelihood with broad application areas and was first coined by Dempster, Laird and Rubin in [30]. The EM algorithm gets its name because in each iteration of EM algorithm comprises two steps: Expectation Step (E-Step) and Maximization Step (M-Step).

Componentwise differentiation of the Term (2.11) with respect to θ and π results in:

$$\frac{\delta \mathcal{L}}{\delta \pi_j} = \frac{1}{\pi_j} \sum_{n=1}^{\mathrm{N}} P(j|x_n; \pi, \Theta) - N \quad j = 1, \ldots, J \tag{2.12}$$

And also

$$\frac{\delta \mathcal{L}}{\delta \theta_{ji}} = \frac{1}{\theta_{ji}(1-\theta_{ji})} \sum_{n=1}^{\mathrm{N}} P(j|x_n; \pi, \Theta)(x_{ni} - \theta_{ji}) \tag{2.13}$$
$$where \quad j = 1, \ldots, J \; and \; i = 1, \ldots, d$$

The term -N in equation satisfies the constraint $\sum_{j=1}^{J} \pi_j$ introduced in log-likelihood via Lagrange multiplier. Now, From Bayes' theorem the posterior

probability can be calculated as shown below.

$$\begin{aligned} P(j|x_n;\pi,\Theta) &= \frac{p(x_n|j;\pi,\Theta)p(j)}{\sum_{j'=1}^{J} P(x_n|j';\pi\Theta)p(j')} \\ &= \frac{\pi_j \prod_{i=1}^{d} \theta_{ji}^{x_{ni}}(1-\theta_{ji})^{1-x_{ni}}}{\sum_{j'=1}^{J}\prod_{i=1}^{d} \theta_{j'i}^{x_{ni}}(1-\theta_{j'i})^{1-x_{ni}}} \end{aligned} \tag{2.14}$$

Derivation of the EM algorithm is fairly simple and can be referred from the works of Everitt and Hand [31] as well as Wolfe [29]. The basic equations of EM algorithm are:

E-step: E-step computes the posterior probability using the Equation 2.14 for the most recent values of parameters $\theta^{\tau}, \Theta^{\tau}$ at iteration τ i.e. calculate $P(j|x_n;\pi^{\tau},\Theta^{\tau})$

M-step: M-step recomputes the the values of parameters $\theta^{\tau+1}, \Theta^{\tau+1}$ for the next iteration.

$$\begin{aligned} \pi_j^{\tau+1} &= \frac{1}{N}\sum_{n=1}^{N} P(j|x_n;(\pi^{(\tau)}),\Theta^{(\tau)}) \\ \Theta_j^{(\tau+1)} &= \frac{1}{N\pi_j^{(\tau+1)}}\sum_{n=1}^{N} P(j|x_n;(\pi^{(\tau)}),\Theta^{(\tau)})x_n \end{aligned} \tag{2.15}$$

Iterations between E and M step produce a succession of monotonically increasing sequence of values of loglikelihood for the parameters $\tau = 0, 1, 2, 3\ldots$ regardless of the starting point $\{\pi^{(0)}, \Theta^{(0)}\}$. This result is advantageous but also results in the problem of singularities, the possibility of getting an infinite likelihood if a single data point is assigned to one of the mixtures. However, mixture of Bernoulli distribution are not susceptible to the problem of singularities because the likelihood function is bounded by the constraint $0 \leq p(x_n|\theta_j) \leq 1$ except for some trivial cases such as: assume that data is 1 but model is 0, so the likelihood of the model is 0 and if we take the loglikelihood it turns out to be $\log 0 = \infty$. Furthermore, loglikelihood surface is unbounded. Such problems, however, are rare in multivariate case. EM requires that the number of the mixture components in the mixture model be known in advance. Furthermore, EM algorithm is sensitive to the initializations and the

results may differ on the same data for different initializations. Nevertheless, EM algorithm is deterministic with given initializations and a given dataset. EM algorithm can get stuck in local minima and the global optimal results are not often guaranteed. To overcome these problems regularization techniques as discussed in [51] can be used. In spite of these demerits, EM algorithm has been widely used because of its reliability.

One of the important issues to note regarding the non-identifiable problem is that it matters least with respect to this thesis. Our main aim is to maximize the Equation (2.11) considering the trade-off between the model complexity (number of components in the mixture model) and small difference in the maximum likelihood value. If different parameters satisfy the trade-off, choosing any of those parameters will have negligible effect on the final results.

2.4 Cross-validation

The idea of cross-validation, sometimes also called rotation estimation and pioneered by [52] and [53], is fundamental concept in machine learning for assessing the results of the statistical analysis. Various forms of cross-validation techniques have been proposed. The basic definition of k-fold cross-validation states that the training set $\mathcal{T}$ is divided into k exhaustive and exclusive equal sized sub-sets $\mathcal{T}_1,\ \mathcal{T}_2,\ \ldots \mathcal{T}_k$. The main assumption is that the both the training and the validation sets are independent. For each sub-set $\mathcal{T}_i$ where $i \in 1, 2, 3 \ldots k$ the data is trained on the union of all the other subsets and determine the error on the subset $\mathcal{T}_i$. The final error of the algorithm is the average error on all the sub-sets as shown in the Equation 2.16.

$$\varepsilon = \frac{1}{k} \sum_{i=1}^{k} \epsilon_i \tag{2.16}$$

The initial subset of data is called the test set; while union of the remaining subsets is called the training set. The efficiency of k-fold cross-validation largely depends on the choice of k. If the number of k is small, the algorithm is computationally efficient as it requires performing lesser rounds of experi-

ments. Furthermore, the variance of the estimator will be negligible. On the contrary, the bias of the estimator will be significantly larger, larger than the true error (generalization error) on the future data.

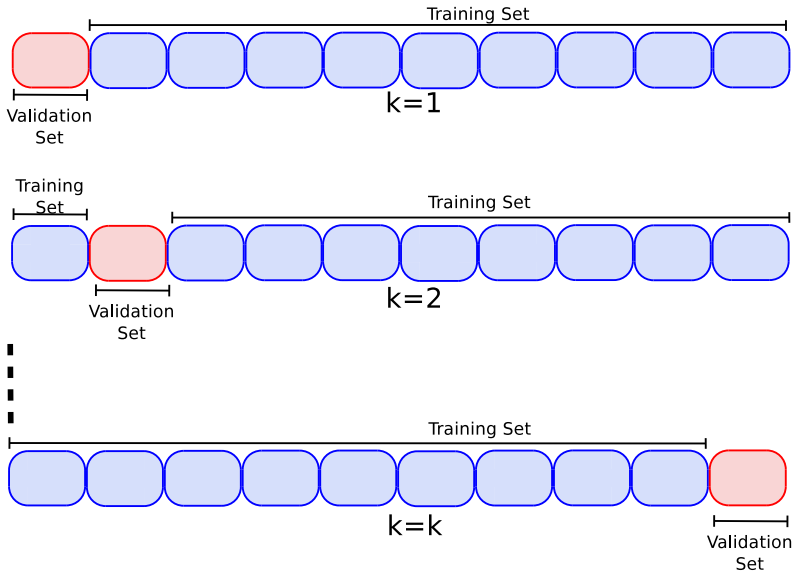

Figure 2.3: Schematic representation of k-fold cross-validation technique showing the division of training data into training and validation set. The figure here shows the case when k=10.

On the other hand, if the number of k is large, the bias of the estimator will be significantly low. With large value of k, the bias is likely to converge to the true error (generalization error) on the future unseen data. On the contrary, the computational time is greatly increased as the number of iterations increases. For example, in simple 10-fold cross-validation the learning algorithm is repeated 10 times with 9/10 of the total data. Additionally, variance of the true error estimate will be significantly larger. In most of the cases, the choice of k depends on the size of dataset. If the size of dataset is larger, smaller number of k is a better option while for smaller datasets larger number of k will be a better option.

The optimal number of k for k-fold cross-validation highly researched area but still an open problem. Although there are some empirical [54] and mathematical results [55] suggesting the optimal value of k, the choice depends on the rule of thumb. Comprehensive studies and experiments on datasets of dif-

ferent sizes have shown that ten is the optimal number of k to get the accurate error [55]. In cross-validation data is randomly divided into k different sets. Hence, different runs of cross-validation with the same learning algorithm on the same data can produce different results. In order to mitigate this problem, different runs (often 10) of the cross-validation procedure is suggested which involves running the learning algorithm 100 times with 9/10 of the total data each time. One 10-fold cross-validation can be seen as a "standard" measure of the performance whereas ten tenfold cross-validations would be a "precise" measure of performance [56]. Furthermore, similar to the problem in hold-out method, cross-validation is also susceptible to "unfortunate split". Thus while partitioning the data into subsets, care should be taken to include different unique samples of data in all rows to each of the subset. The idea of 'stratification' have been suggested as the solution to the problem of unfortunate split thus ensuring that each class is properly represented in both training and the validation sets. It is important to note that different classes are only approximately represented in the proportion present in the training set.

2.5 DNA Copy Number Aberrations Data

Humans, being a diploid organism, have two homologous copies of each chromosome usually, one inherited from the father and the other from the mother. During the complex process of cell division, some abnormalities can occur in the cells and copy number changes from two [23]. Deletion, often referred to as loss, is the case when the copy number is less than two. Duplication, often referred to as gains, is the case when the copy number is more than two. Amplification is a form of chromosomal aberration when the copy number of the chromosome increases more than 5. Amplification is different from duplication because duplication exactly doubles the copy number. For instance, in human the normal copy number is two, so duplication increases the copy number to 4. Higher level amplifications have been known increases the copy number more than hundred fold. Generally, the amplification is developmentally regulated and amplified copies are lost from the cell. However, amplification in many cases manifests itself in larger number throughout the genome [23]. DNA am-

plifications are essentially the hallmarks of cancer. Studies have also shown that copy number amplification results in resistance to certain drugs [23].

CGH (Comparative Genomic Hybridization) [8] is one of the molecular techniques to survey the DNA copy number variation across the whole genome. Differentially labeled test and reference genomic DNA are cohybridized to normal metaphase chromosomes. Fluorescence ratios along the length of the chromosome provide a cytogenetic representation of DNA copy number variation. However, one major drawback of CGH is the resolution. The mapping resolution is only 20Mbp (million base pairs) which is also the average size of the aberrated region. Furthermore, mapping resolution for deletion is 2Mbp. To overcome the problem of CGH, a new microarray technology called aCGH (Array Comparative Genomic Hybridization) [9] has been developed. aCGH provides higher resolution than CGH. Fluorescence ratios at arrayed DNA elements provide a locus by locus measure of the copy number changes. aCGH was initially used to characterize variation in gene expression using cDNA. Using the CGH methods, the chromosome is subbanded to 400 regions, also known as cytogenetic bands. Using different staining techniques, the cytogenetic bands can be visualized and the resolution of the cytogenetic band can be increased to over 800 resolution.

2.6 Review of Literature

The problem of analysis of 0-1 data is a very old problem and has been considered extensively in statistics and machine learning. The mixture model is also a well-studied solution. Recently, mixture models have been a subject of major research. For detailed review regarding mixture models and its applications, the reader is referenced to [37, 57] and the references therein. On the other hand [7] reviews different machine learning methods applied to cancer research. In spite of the great boom of mixture models in the last few decades, comparatively very few instances of research are based on the mixture of multivariate Bernoulli distributions. Nonetheless, the mixture of Bernoulli distribution is found to be suitable in the analysis of the 0-1 data. Thus, this section briefly reviews the research and applications pertaining to the mixture

of multivariate Bernoulli distributions with a special focus on cancer genetics.

The mixture of Bernoulli distributions has found significant application areas when the data is in 0-1 form. For example, in [58], Bernoulli mixture model trained using EM algorithm is used to classify binary images with effective results. In the case of binary image, multiple mixture captures the pixel correlations. Each pixel is assumed to be governed by its associated Bernoulli parameter. One particular application area in which the use of FMM of Bernoulli distributions has excelled is natural language processing. In [59], FMM of Bernoulli distributions is used in text classification. The text classification is used to improve the language modelling for machine translation. The text classification is used as an extension to näive Bayes by modelling the class conditional dependence spreading it over different mixture components. In [60], FMM of Bernoulli distributions has been used in classification. Additionally, the Bernoulli mixture models are used for feature selection and feature extraction including dimensionality reduction from the input data. The combination of the methods implemented in two datasets of varying domains: text mining and hand writing recognition, produces considerable increase in the classification accuracy. Furthermore, the dimensionality reduction of 99.9% is achieved on the sparse 0-1 data. An interesting and early application of Bernoulli mixture models for statistical modelling of teaching styles is explained in [61]. The authors compiled a 38 dimensional 0-1 data set of 1258 samples from a questionnaire consisting of 28 items. The mixture modelling technique was tested on 2 to 22 clusters and 12 clusters was selected as it produced the overall maximum. With this statistical modelling techniques, the authors were able to distinguish different teaching styles.

2.6.1 Mixture Models in Copy Number Analysis

DNA copy number analysis was started in [62] where the authors mainly focused on determining the copy number of the cytogenetic band. Similar works performed are reviewed in [63] to determine the copy number. However, in [62] and [63] the authors did not establish a relation between the copy number and their clinical significance. In the recent past, DNA copy number amplification data collected with bibliomics survey from 838 journal articles published

from 1992 to 2002 was analyzed in [64]. In the work, amplification patterns were determined for 73 different neoplasms and the neoplasms were clustered according to amplification profiles thus identifying the amplification hotspots using independent component analysis. The profiling revealed that human neoplasms formed clusters based on the amplification frequency of the cancer. Continuing the studies in DNA copy number amplification, authors in [22] classified the human cancers based on copy number amplification using probabilistic modelling. Furthermore, the authors extracted the ranges of the amplification in the chromosome and expressed it according to the cytogentic nomenclature. In [26] and [65], the authors modeled the DNA copy number amplification using a mixture of multivariate Bernoulli Distributions. The classification of 73 different neoplasms in [64] were extended to 95 different neoplasm types. Furthermore, in [66], the authors have proposed the enhancement to Bayesian Piecewise Constant Regression(BPCR) called mBPCR changing the segment number estimator and boundary estimator to enhance the fitting procedure. The proposed mBPCR was more accurate in the determination of true breakpoints of amplification. The more recent studies [67] and [68] have mainly focused in cancer specific analysis of DNA copy number. Although the mixture models were used in [26] and [65], they have studied only chromosome 1 data in resolution 400. Chromosome 1 being the largest chromosome, there is significant amount of amplifications [64]. However, single chromosome band and the specific gene responsible for cancer has not been identified. Hence, in this thesis, study was performed on all chromosomes including chromosome 1. Chromosomewise analysis can reveal interesting facts about amplification of specific chromosomes and guarantees efficient computation & ease of analysis. Furthermore, there are several sources of multilevel biological data that comes in multiple resolutions as shown in Figure 1.1 but there seems to be a significant gap in research to study multiple resolution of the data as authors in [64] and in relevant work did not consider the data in multiple resolution. Algorithms and methods that meet the demands such multiresolution data could possess very high clinical significance. Thus, this thesis devises methods able to work with multiple resolutions of genome.

Chapter 3

Sampling Between Different Resolutions

> *For everything you have missed, you have gained something else, and for everything you gain, you lose something else.*
>
> — Ralph Waldo Emerson
>
> *American Poet, Lecturer and Essayist(1803-1882)*

Synopsis

This chapter focuses on the different methods used to upsample data to finer resolutions and downsample data to coarser resolutions. Upsampling, discussed in Section 3.1, transforms the resolution of data from coarse to fine. The three downsampling methods discussed in Sections 3.2.3, 3.2.1, and 3.2.2 transform the data from fine resolution to coarse resolution. Part of the work discussed in this chapter has been published in [32].

Sampling resolutions in cytogenetics is a process of defining the level of precision for the staining techniques to produce the results either global or detailed view. A good metaphor for sampling as given by [69] in terms of speech recognition can be an advertisement recently aired in a Dutch Television where the shot is started with a global view. In this case, a shot was taken from

the orbit satellite and gradually zooming into Europe, the Netherlands, the Dutch North Sea Coast, the Scheveningen beach up to a lady drinking a glass of beer in a terrace. Similar to the advertisement, different staining techniques produce chromosome bands in different resolution. Computational algorithms can be designed to work with only specific resolution of chromosome band. Hence, upsampling or downsampling is necessary before the data can be fed to the algorithm. Furthermore, comparing the results of an algorithm on data in different resolution can produce interesting results which aid in determining suitable resolution of data. In addition, experiments in different resolutions can be helpful in determining the appropriate method for staining.

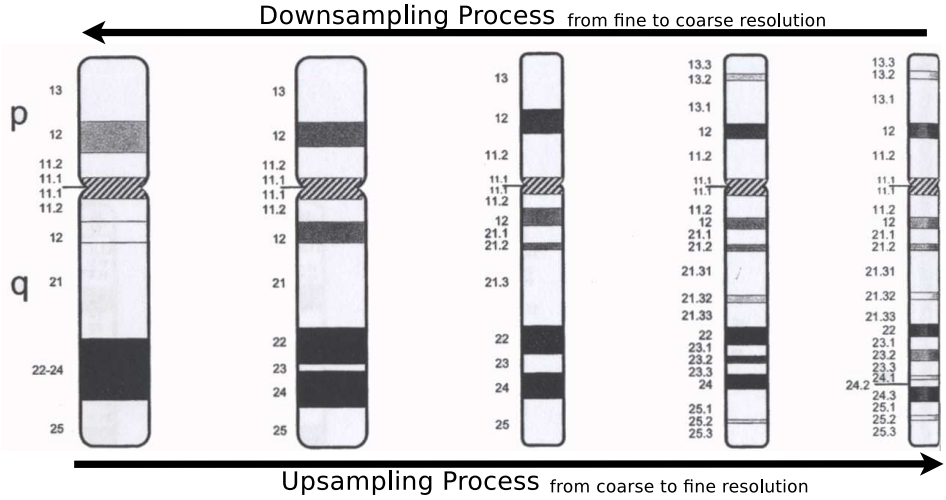

Figure 3.1: Schematic representation of sampling in multiple resolutions where upsampling transforms the data to find resolution while downsampling transforms the data to coarse resolution.

Section 1.3 explained the problem of multiple resolution in chromosome along with the Figure 1.1 which showed the G-banding pattern of Chromosome 17 in five different resolutions. In the context of Figure 1.1, upsampling and downsampling can be seen as the process of data transformations as shown by the arrows in Figure 3.1. Upsampling changes the representation of data from coarse resolution to fine resolution as shown by the arrow pointing to the right in Figure 3.1. Similarly, downsampling changes the representation of the data from fine resolution to coarse resolution as shown by the arrow pointing to the left in Figure 3.1.

3.1 Upsampling

Upsampling, as shown in Figure 3.2, is the process of changing the representation of data to the fine resolution. A simple method was devised to upsample the data from coarse resolution. Upsampling was simple and were implemented using simple transformation tables or lookup tables. Initially, the dataset was in resolution 400 and it was upsampled to three different resolutions 550, 700 and 850. Multiple copies of cytogenetic band in coarser resolution were made to upsample the data to finer resolution. For example, cytogenetic band 1q36.1 in resolution 550 has been divided into three bands $1q36.11$, $1q36.12$ and $1q36.13$ in resolution 850. So, multiple copies of $1q36.1$ was made for all bands $1q36.11$, $1q36.12$ and $1q36.13$ in resolution 850.

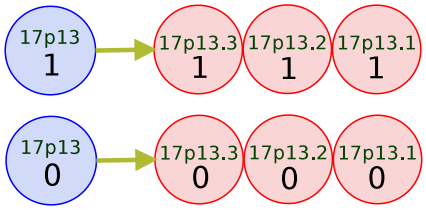

Figure 3.2: Schematic representation of upsampling where duplicate copies of similar cytogenetic bands are made in the finer resolution

Figure 3.2 shows that three copies of similar cytogenetic band in coarser resolution are made to upsample the data to finer resolution. When multiple copies of same cytogenetic band is made finer resolution will have only few unique rows. Hence, when the sample size decreases the complex model in higher dimension can not be trained to convergence thus producing poor results. Implementation of downsampling was performed using simple transformation tables implemented in Perl [70]. Table 3.1 shows an example of table for transformation of data in 400 resolution to 850 resolution for chromosome 17.

Table 3.1 shows that some chromosome bands missing in 400 resolution are observed in resolution 850. Hence, duplicate copies of the similar chromosome band in resolution 400 were made in finer resolution. Duplications are made based on the assumption that if an adjacent area is amplified then the proba-

Chromosome Resolution 400	**Chromosome Resolution 850**
17p13	17p13.3
...	17p13.2
...	17p13.1
17p12	17p12
17p11.2	17p11.2
17p11.1	17p11.1
17q11.1	17q11.1
17q11.2	17q11.2
17q12	17q12
17q21	17q21.1
...	17q21.2
...	17q21.31
...	17q21.32
...	17q21.33
17q22	17q22
17q23	17q23.1
...	17q23.2
...	17q23.3
17q24	17q24.1
...	17q24.2
...	17q24.3
17q25	17q25.1
...	17q25.2
...	17q25.3

Table 3.1: Chromosome bands for resolution 400 & 850 and their transformation

bility of the chromosome band being amplified is high because amplifications typically cover large areas. The transformation table were chromosome specific and resolution specific (i.e. 88 transformation tables in all for different chromosomes).

3.2 Downsampling

Downsampling is the process of changing the representation of the data to the coarser resolution. In both cases of upsampling and downsampling, no attempt is made to infer the structure of the data and no information is added or removed during the process. If the data of the same patients were available in two different resolutions, one of the supervised classification algorithms in machine learning could be used in downsampling dealing the problem as a traditional classification problem. However, such data was not available. Hence, simple but useful methods motivated from biology are used for downsampling. Downsampling methods were implemented in scripts with a script for each chromosome in each resolution. Sections 3.2.1, 3.2.2 and 3.2.3 detail the methods of downsampling.

3.2.1 OR-function Downsampling

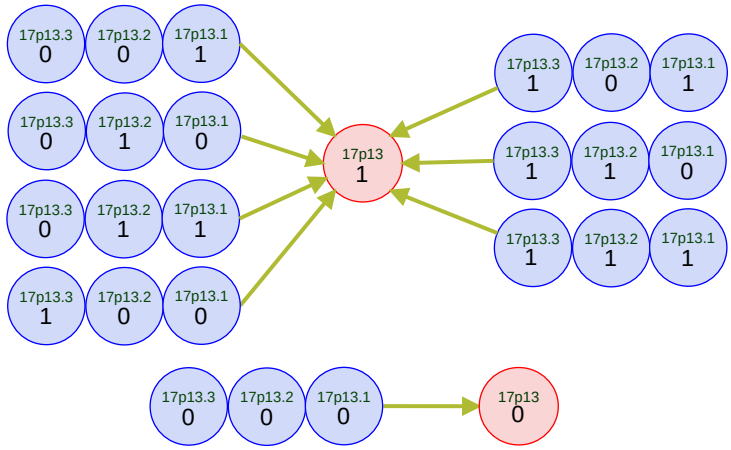

Figure 3.3: Schematic representation of OR-function downsampling procedure. Here the cytogenetic band in coarser resolution is amplified if any of the bands in finer resolution is amplified. Cytogenetic band in coarser resolution is not amplified only when none of the bands in finer resolution is amplified.

In OR-function downsampling method, the cytogenetic band in coarser resolution is not amplified if none of the bands in finer resolution are amplified. The cytogenetic band in coarser resolution is amplified if either of the bands in

finer resolution is amplified. Figure 3.3 depicts the OR-function downsampling method. The OR-function downsampling method is based on simple belief that if the one of the bands in finer resolution is amplified, it signifies the presence of amplification in the band. For the case in the Figure 3.3 downsampling can be considered as a simple 0-1 classification problem in machine learning where input is three dimensional 0-1 variable and output is one dimensional 0-1 variable. The solution is a simple truth table describing the classical OR operation. This method does not consider the length of the cytogenetic bands.

3.2.2 Majority Decision Downsampling

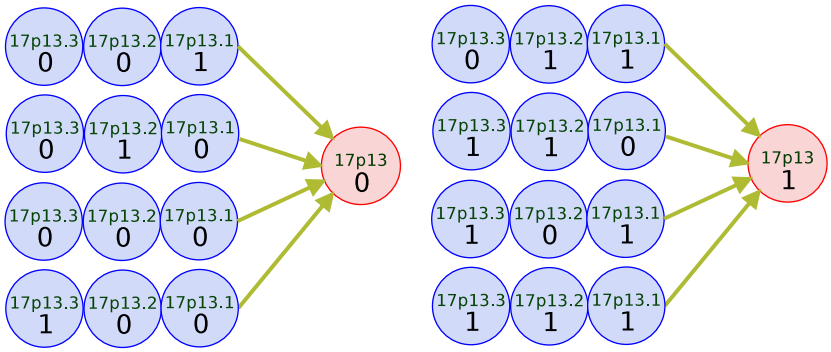

Figure 3.4: Schematic representation of majority decision downsampling procedure. Here the cytogenetic band in coarser resolution is amplified if majority of the bands in finer resolution are amplified, otherwise it not amplified.

In majority decision downsampling method, a cytogenetic band in coarser resolution is amplified if majority of the cytogenetic bands in finer resolution are amplified otherwise the cytogenetic band is not amplified. In case of a tie amplification of two nearest bands one in the left and the other one in the right are taken into consideration iteratively and the amplification pattern of the band is determined using the idea similar to 'golden goal'[1] strategy used in football. In other words, if in any iteration both bands in neighborhood bands

[1]The golden goal is a method used in football to determine the winner which end in a draw after the end of regulation time. Golden goal rules allow the team that scores the first goal during extra time to be declared the winner. The game finishes when a golden goal is scored.

are amplified then the band is amplified and if both the neighbors are unamplified then the band is deemed unamplified. If the amplification of coarser resolution can not be concluded with 'golden goal' strategy then the band in coarser resolution is deemed as amplified. Figure 3.4 shows one of the examples of majority decision in downsampling. There is a shortcoming in this downsampling method because it does not take into consideration the the lengths of the cytogenetic bands. The lengths of cytogentenic bands are considered by length weighted downsampling method discussed in Section 3.2.3.

3.2.3 Length Weighted Downsampling

In length weighted downsampling method, depicted in the Figure 3.5, length of the cytogenetic band is considered. The length of the cytogenetic band varies in each assembly and hence relative lengths were considered. The amplification of cytogenetic band in coarser resolution is determined by the weighted length of cytogenetic band in finer resolution. Each cytogenetic band is weighted according to the relative length of the cytogenetic band. If the total length of amplified region is greater than the total length of unamplified region, the cytogenetic band in coarser resolution is amplified, otherwise the cytogenetic band is unamplified. Here, relative length is considered which gives more accurate measure of the amplification profiles in the cytogenetic band. Absolute lengths of the cytogenetic bands are currently not available and vary with each assembly. Two relative measures were considered in the calculation of the length. From the ideogram dataset available in NCBI [71], the difference between ISCN.top and ISCN.bot were used as relative measures. Similarly, difference between bases-top and bases-bot were also used as the relative measure of the length of each cytogenetic band. The difference in the results produced using the different relative measure of length have also been studied.

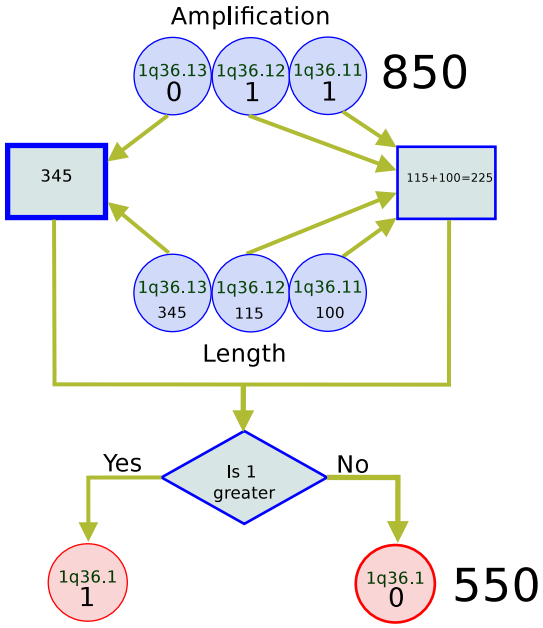

Figure 3.5: Schematic representation of weighted downsampling procedure. Here the cytogenetic band in coarser resolution is amplified if total length of the amplified bands in finer resolution is greater than the total length of unamplified bands, otherwise it not amplified. The figure is an example case in chromosome 1q36.1 where two cytogenetic bands 1q36.11 and 1q36.12 in resolution 850 are amplified and one band 1q36.13 is not amplified. However, total length of unamplified region i.e. band 1q36.13 (345) is greater than total length of the unamplified region i.e. bands 1q36.11 and 1q36.12 (100+115=225). Hence, the band in resolution 550 is unamplified.

Chapter 4

Experiments and Results

> *Knowing is not enough; we must apply.*
> *Willing is not enough; we must do.*
> — Johann Wolfgang von Goethe
> *German Writer(1749-1832)*

Synopsis

This chapter describes the experiments performed on transformation of data between different resolutions and mixture modelling of multivariate Bernoulli distributions on the chromosomal aberrations. The obtained results are analyzed and discussed.

4.1 Software

This thesis uses a ready programme package for mixture models of multivariate Bernoulli distributions. Implementing the mixture models from the beginning and thorough testing would consume significant amount of time. Therefore, the approach in this thesis was to use a ready-made package and analyze the results. This approach provided the time to concentrate the efforts on the machine learning aspects and its application in real world data. Although programming mixture models from the beginning would be very educational and precious programming experience, it takes significant amount time and

diverges the attention from machine learning aspects which was the primary goal of the thesis.

There are several software, both commercial and open-source, available for finite mixture modelling. Few examples are:

- MULTIMIX available in http://www.stats.waikato.ac.nz/Staff/maj/multimix
- MIX (Commerical) available in http://icarus.math.mcmaster.ca/peter/mix/mix.html
- AutoClass available in http://ti.arc.nasa.gov/project/autoclass/
- Clustan available in http://www.clustan.com/
- Snob available in http://www.csse.monash.edu.au/~dld/Snob.html
- Mtreemix available in http://mtreemix.bioinf.mpi-sb.mpg.de/
- PyMix available in http://www.pymix.org/pymix/
- em available in http://www.ar.media.kyoto-u.ac.jp/members/david/softwares/em/
- BernoulliMix available in http://users.ics.tkk.fi/jhollmen/BernoulliMix/
- FlexMix [72] http://www.cran.r-project.org /web/packages/flexmix
- mixtools [73] http://cran.rakanu.com/web/packages/mixtools

Most of the software packages above are open-source but have shortcomings of their own. For example, most of them were designed to work with Gaussian distribution. Since our main aim was to model Multivariate Bernoulli distributions and BernoulliMix provided all the required features and was freely available and hence we converged on BernoulliMix for our modelling purposes. Furthermore, integrating BernoulliMix with other tools such as Matlab [74], Shell Scripting [75], Perl [70] and R [76] is smooth and unconstrained.

4.1.1 BernoulliMix Program Package

BernoulliMix [77] programme package is an open-source programme package for the finite mixture modelling of Multivariate Bernoulli distributions. It is freely available at BernoulliMix Homepage[1] under GPL license. BernoulliMix, implemented is ANSI C, can be used to model the 0-1 data in the probabilistic framework. BernoulliMix has five programs to work with finite mixture models of multivariate Bernoulli Distribution:

- **bmix_init**: To initialize the mixture models with randomly selected parameters sampled from the uniform distribution of selected range.
- **bmix_train**: To train the mixture model from the data using EM algorithm i.e. learn the parameters of the mixture model.
- **bmix_like**: To calculate the likelihood of the data with the mixture model. Likelihood can be calculated either for whole data or each vector separately.
- **bmix_sample**: Mixture models are generative models. bmix_sample provides the facilities to sample the data from the trained mixture model.
- **bmix_cluster**: To cluster the data (associating a component distribution with a cluster) with the mixture model by the maximum posterior rule.

Details about the programme package and its use with examples can be obtained from [77]. The BernoulliMix programme package was used in conjunction with Matlab [74], R [76], Perl [70] and Shell Scripting [75] to garner the results of the experiments.

4.2 DNA Copy Number Aberrations Dataset

The dataset used in the experiments defines DNA copy number aberrations in different chromosomes. The data was collected by the bibliomics survey of

[1]The homepage is http://users.ics.tkk.fi/jhollmen/BernoulliMix/

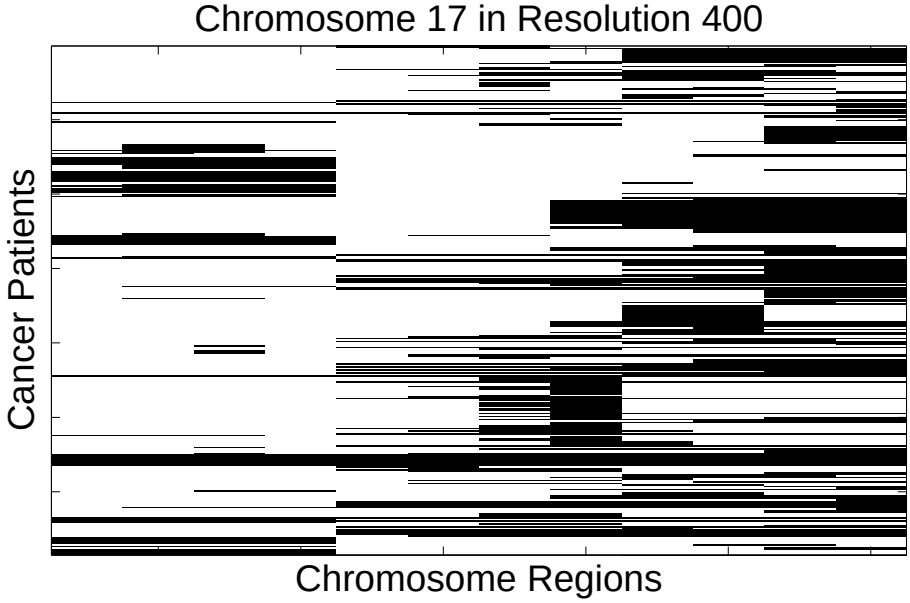

Figure 4.1: DNA copy number aberrations in chromosome 17, resolution 400. $\overline{X} = (X_{ij})$, $X_{ij} \in \{0, 1\}$. Each row represents one sample of the aberrations pattern for a cancer patient and each column represents one of the chromosome bands (regions). In figure dark color denotes the presence of aberrations and the white color denotes the absence of chromosomal aberrations.

838 journal articles during 1992-2002 by hand without using state-of-the-art text mining techniques [22, 65]. The dataset contained the information about the chromosomal aberrations of 4590 cancer patients. Each row describes one sample of the cancer patient while each column identifies one chromosomal band(region). The dataset is a typical 0-1 dataset where aberrated chromosomal regions were marked with 1 while and the value 0 defines that the chromosome band is not aberrated. Chromosomes X and Y were not included in the experiments because of the lack of data. Patients whose chromosomal band had not shown any aberrations for the specific chromosome were not included in the experiments since we are interested in modelling the aberrations, not their absence. Thus different chromosomes had different number of the samples. The chromosomal aberrations dataset analyzed in this thesis uses data containing few samples. Thus, we decided to work chromosomewise because of the availability of very small number of the data samples to constrain the

complexity of the mixture models. The original data dimension for the whole genome ranges from 300 to 850 which will be cumbersome to work with given the very large dimension compared to the number of samples. On the other hand, working with each chromosome will be computationally easier as the largest dimensionality is 63.

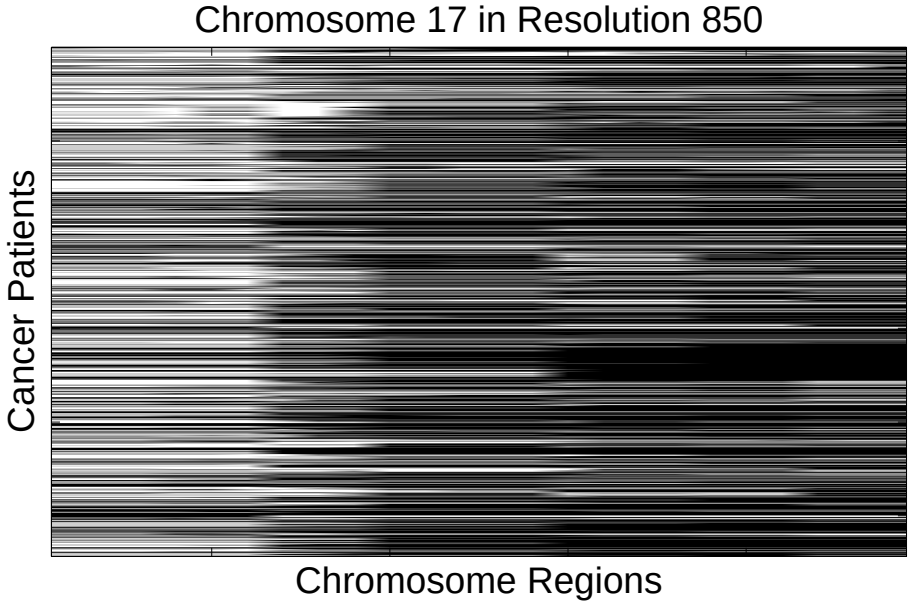

Figure 4.2: DNA copy number aberrations in chromosome 17, resolution 850. $\overline{X} = (X_{ij})$, $X_{ij} \in \{0, 1\}$. Each row represents one sample of the chromosomal aberrations for a cancer patient and each column represents one of the chromosome bands (regions). In figure dark color denotes the presence of aberrations and the white color denotes the absence of chromosomal aberrations.

A shwon in Figures 4.1 and 4.2, copy number aberrations occur very sparsely and are often spatially dependent. The original data was in the resolution 400 i.e. there were 393 chromosomal bands (regions) for the entire genome. The original data was upsampled to resolution 550, 700 and 850 and downsampled to resolution 300 using the methods discussed in Chapter 3. Bands for the specific chromosome were extracted and mixture modelling was preformed on each chromosome. For example: chromosome 1 had 63, 61, 42, 28, and 23 chromosomal bands in resolution 850, 700, 550, 400, and 300 respectively [1]. Similarly, a different set of data was available in resolution 850 from progenetix.net [78].

The data in resolution 850 was different from data in resolution 400. Similar to the data in the resolution 400, the data in resolution 850 was downsampled to resolution 300, 400, 550 and 700. Elementwise AND operation over all the samples in the data results in a zero vector thus necessitating sophisticated machine learning and data mining methods and techniques for classifying and profiling aberrations.

The ISCN (ISCN 2009: An International System for Human Cytogenetic Nomenclature) nomenclature of chromosome, discussed in Appendix A, divides the chromosome into different resolutions shown in Table 4.1.

S.No	Resolution	# Regions	# Regions in Chr 1
1	300	317	23
2	400	393	28
3	550	555	42
4	700	759	61
5	850	862	63

Table 4.1: Number of Chromosome bands(regions) for 5 different resolutions of data studied in the thesis. Included as an example number of bands in Chromosome 1, the largest chromosome.

Thorough study was performed for every chromosome with every resolution using the finite mixture modelling approach.

4.3 Comparison of Downsampling Methods

The downsampling methods, discussed in Chapter 3, were implemented in scripts. There were 110 scripts in all for all transformations, one for each chromosome in 5 different resolutions (# of Chromosomes $\times$ # of Resolutions i.e 22×5). Matlab ® [74] was used for scripting. The individual scripts for downsampling each chromosome takes a file name of the data set in higher resolution as input and first checks for some errors such as mismatch in the number of regions of the chromosome in that specific resolution. Data is then transformed bandwise to lower resolution combining the multiple bands in higher resolution according to the three different methods proposed in Sections 3.2.1, 3.2.2, and 3.2.3. The downsampled data from 850 resolution was

subjected to various tests to access the difference in the results of the downsampling methods.

4.3.1 Property Models

Some simple and efficient property models were defined to compare the results of the three different downsampling procedures.

Column and Row Margins

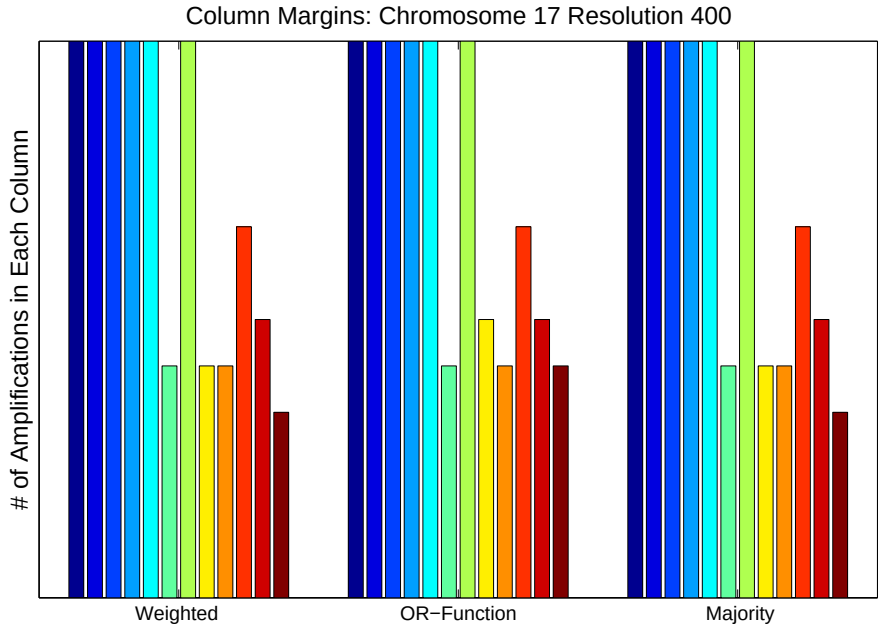

Figure 4.3: Comparison of three different downsampling methods: Example case in chromosome 17 resolution 400. Figure does not show significant difference in the results of the three methods.

The total number of differences in the dataset was studied with respect to each row and column margin produced on downsampling from higher resolution to lower resolution. The total number of differences in each chromosome band and in each cancer patient was computed and compared between three different downsampling methods. The results of the three different downsampling process did not show significant differences with respect to the number of differences in the row and column margins as shown in Figure 4.3 which is

an example result for chromosome 17 in resolution 400. Figure 4.3 shows that results produced by three methods are highly similar. In order to scrutinize the results, mean difference between the number of differences produced by the three methods in various chromosome bands was computed. The results for an example case discussed earlier i.e. chromosome 17 in resolution 400 is shown in the Figure 4.4.

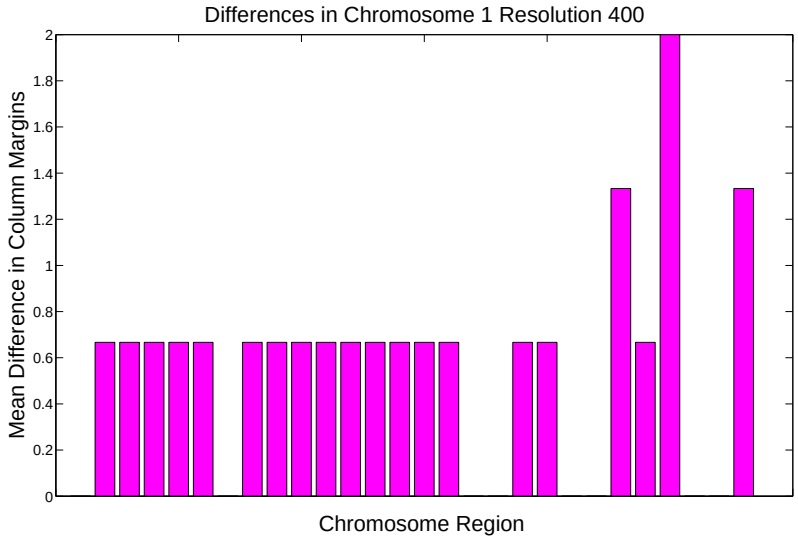

Figure 4.4: Total difference in data produced by three different downsampling methods: Example case in chromosome 1 resolution 400. The figure shows presence of differences in some chromosome regions.

Figure 4.4 suggests that there are differences in the results produced by three downsampling methods albeit rather minute. Three downsampling methods produced no differences in some chromosomes such as chromosome 1, 5, 8, 19, 20, 21 and 22 in resolution 700. In contrast, the methods produced some negligible differences in other chromosomes. Hence with respect to the total number of differences in row and column margins produced in the coarse resolution, the three proposed methods are highly similar.

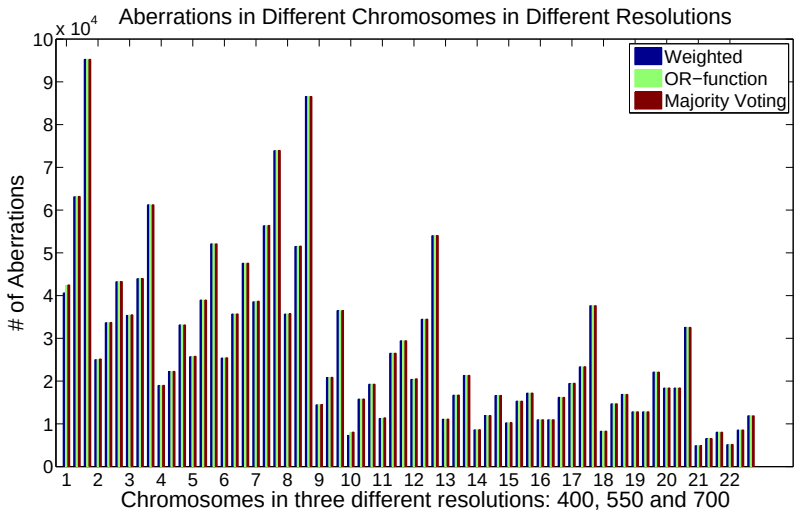

Figure 4.5: Comparison of three different downsampling methods with respect to number of aberrations produced.

Total Number of Differences

Similar to the differences in datasets, we studied the total number of aberrations present in the downsampled data. Total number of aberrations in each chromosome was computed and compared between three different downsampling methods. The results of the three different downsampling methods did not show significant differences with respect to the number of aberrations produced. Figure 4.5 suggests that the three downsampling methods produces similar results. Furthermore, the mean difference between the number of aberrations produced by the three methods in various chromosomes was computed.

Figure 4.6 suggests that there are differences in the results produced by three downsampling methods, albeit very small. However, the differences between the methods are not significant when the number of aberrations are considered, which are significantly high.

4.3.2 Matrix Difference: Frobenius Norm

Property models discussed in Section 4.3.1 demonstrate no significant differences in the downsampling methods. However, the two methods discussed in Section 4.3.1 are susceptible to some errors where the number of chromosomal

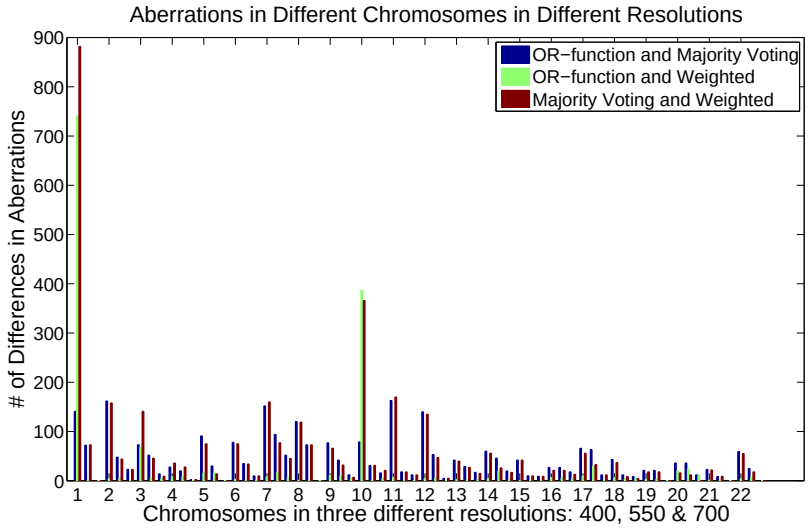

Figure 4.6: Difference in aberrations produced by three different downsampling methods with respect to the number of aberrations produced in the data.

aberrations are same and also number of chromosomal aberrations does not change in different methods. For example, the methods discussed Section 4.3.1 do not show difference between the following two datasets.

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

However, the two datasets above are significantly different. In order to capture these differences, we further analyzed the difference between the different downsampling methods as the difference between the two resulting matrices for different methods using standard matrix difference measures. The distance measure used is the square of the Frobenius norm [79] between two matrices. In 0-1 matrices, Frobenius norm is essentially the number of cells where the two matrices differ.

Figure 4.7 suggests that the three downsampling methods produces fairly similar results. It also suggests that the differences are high in chromosome 1 which is expected because chromosome 1 is the largest chromosome. Differences are also high in lower resolution compared with higher resolution because it is the lower resolution where most of the changes take place. The differ-

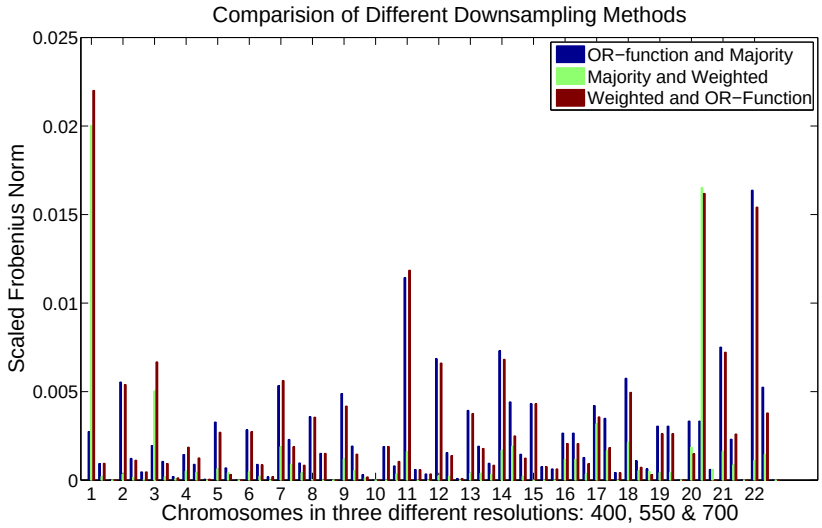

Figure 4.7: Comparison of three different downsampling methods: The difference measure used is scaled Frobenius norm.

ences in the smaller chromosomes especially 20-22 are because of significant variation in the bands combined. Normally, three bands in finer resolution are combined in coarser resolution but in small chromosomes, the number of chromosome bands combined is very different thus making it difficult for weighted and OR-function downsampling method to work. It is to be noted that in the chromosomes where the differences are larger have larger number of differences in number of chromosome bands in different resolutions.

4.3.3 Changes in Aberrations

We also calculated the number of cases where the unaberrated band has changed to the aberrated region in two different methods. Calculating such differences will also help to measure the closeness of different downsampling methods. The number of cases where the unaberrated region (represented by 0 in dataset) changes to the aberrated band (represented by 1 in dataset) is calculated and the results are visualized as shown in the Figure 4.8.

Figure 4.8 exhibits that there are no differences in the number of changing chromosomal aberrations on two methods i.e. the majority decision and the OR-function downsampling. On the other hand, noticeable differences can be

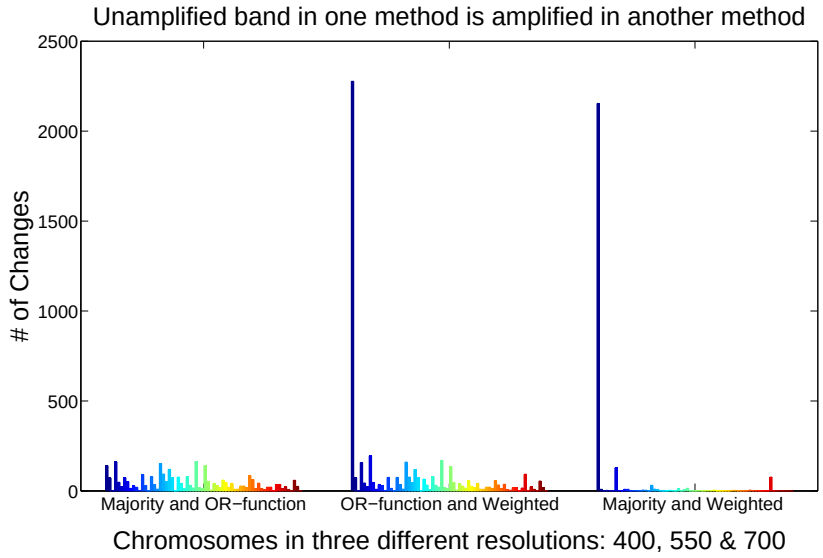

Figure 4.8: Number of unaberrated bands changing to aberrated bands in different downsampling methods. The X-axis varies in small intervals with respect to resolution and larger intervals with respect to the chromosome number. As usual chromosome X and Y were excluded from the experiment.

observed between OR-function and weighted downsampling as well as between majority decision and weighted downsampling. Generally, the OR-function and the majority decision downsampling methods are similar. However, OR-function downsampling is expected to produce more aberrations in the coarse resolution than the majority decision. In any case, these findings highly co-relate with the biological notion that chromosomal aberrations typically cover large areas, thus producing negligible or difference between OR-function and majority decision downsampling methods. On the other hand, weighted down-sampling method is highly effected by length as shown in the Figure 3.5, thus differing from majority decision and OR-function downsampling methods. The length is often not that effective measure because ISCN defined the nomenclature of chromosome based on distinct specific landmarks such that they are distinguished during staining [1].

Similar to Section 4.3.1, the mean of difference between the number of the unaberrated region changing to the aberrated region was also computed and visualized with the results depicted Figure 4.9. Similar to other matrices for

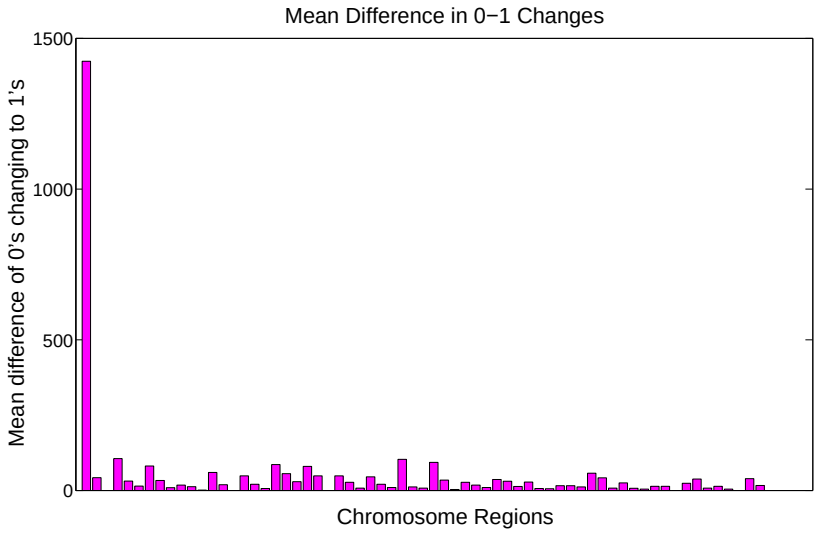

Figure 4.9: Mean of difference in number of unaberrated bands changing to aberrated bands in different downsampling methods. The X-axis varies in small intervals with respect to resolution and in larger intervals with respect to the chromosome number.

defining the similarity/dissimilarity of the results of methods, Figure 4.9 shows negligible differences of cases where the unaberrated band is changed to the aberrated band in two different downsampling methods. After the results from Figure 4.8, it can be inferred that some negligible differences shown in the property models discussed in Section 4.3.1 are the result of weighted downsampling method.

4.3.4 Frequent Itemsets

Given 0-1 data, $\mathcal{D}$ with a set of attributes $\mathcal{I}_1, \mathcal{I}_2 \ldots \mathcal{I}_n$ and a support σ, a frequent set is the set $\mathcal{F}$ of items of $\mathcal{D}$ such that at least a fraction of σ of the rows of $\mathcal{D}$ have 1 in all columns of $\mathcal{F}$ [80, 81]. However, the major problem with frequent itemset is that if an itemset $\{a, b, c\}$ is frequent then their subsets are also frequent because of the anti-monotonicity property of frequent itemsets [82], thus making it unsuitable for comparison and reporting. On the other hand, maximal frequent itemset can be defined as an itemset which is frequent but non of its supersets are frequent [83].

The measure of frequent itemsets also provides a metric for the similarity measure between the sampled data and original data. Furthermore, our major aim was to upsample and downsample the data so that the patterns in the original resolution were retained. Mining maximal frequent itemset in the context of the mixture modelling of multivariate Bernoulli distribution is two fold. It has been shown in [65] that maximal frequent itemset can be used to describe the finite mixture of multivariate Bernoulli distributions compactly and in a language understandable by the domain experts. In [65], the authors implemented a mixture of Bernoulli distributions in clustering 0-1 data to derive frequent itemsets from the cluster-specific data sets and found that the cluster-specific maximal frequent itemset were significantly different from those itemsets extracted globally.

Similar to [65], we used MAFIA (MAximal Frequent Itemset Algorithm) [83] to mine the frequent patterns because other similar algorithms such as Apriori [81] would produce long results which will be difficult to interpret, analyze and report. The frequency or the threshold was chosen as 0.5 motivated by a majority voting protocol. Upsampling is simple and is always guaranteed to retain the frequent itemset although the number of frequent itemset increases with the exactly the same support. Therefore, they have not been reported.

From Table 4.2, we can see that the maximal frequent itemsets are preserved during sampling of resolutions. For example, in OR-function downsampled data in resolution 400 and original data in resolution 850, there is no difference in the maximal frequent itemset because from Table 3.1 used in upsampling, we know that items 7, 8, and 9 in resolution 850 represents items 5, 6 and 7. Items 8 to 14 in 850 are combined to form item 8 in resolution 400. Other itemsets are also formed with similar combinations. Weighted downsampling differs more than other two types of methods but even for weighted downsampling method, the difference is not significant. The results of sampling can be seen more profoundly in integrated datasets where each itemsets in higher resolution can be defined by the frequent itemsets lower resolution. The differences in some cases are only seen because support for those itemsets are less; these differences can be expected because data in lower resolution cannot encompass all the information in higher resolution.

Data Resolution	**Maximal frequent itemsets at threshold(α)=0.5**
Original 400 (A)	{11},{12}
Original 850 (B)	{7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24}
OR-function downsampled from B to 400 (C)	{5, 6, 7, 8, 9, 10, 11, 12}
Weighted downsampled from A to 400 (D)	{7, 8}, {5, 6, 7}, {7, 12}, {7, 11}, {8, 9, 10, 11, 12}
Majority Decision downsampled from B to 400 (E)	{5, 6, 7, 8, 9, 10, 11, 12 }
Combined in 400	{5, 6, 7}, {6,7,8}, {7, 8, 9, 10, 11}, {7, 8, 11, 12}, {8, 9, 10, 11, 12}
Combined in 850	{7, 8, 9}, {8, 9, 10, 11, 12, 13, 14}, {9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21}, {9, 10, 11, 12, 13, 14, 19, 20, 21, 22, 23, 24 }, { 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24}

Table 4.2: Maximal frequent itemsets for data in different resolutions. The support, frequency or threshold (α) used is 0.5. Example case for Chromosome 17.

4.3.5 Motivation for Database Integration

Two sets of original data were available in resolution 400 and 850. Experiments were performed in the original resolution and sampling was performed to sample the data to different resolutions. Data representing whole genome was divided into each chromosome. In each chromosome, the the zero vectors were removed. The cancer patients who did not exbibit chromosomal aberration in a particular chromosome were removed from the data because we were interested in modelling the chromosomal aberrations of cancer patients not their absence.

Furthermore, the sample size of data reduces significantly when the data in resolution 400 is split into each chromosome and samples with all zeros i.e. zero vectors are removed. This phenomenon is captured by Figure 4.10. Additionally, upsampling does not increase the number of unique rows. It also shows that number of samples in upsampled data is significantly less compared

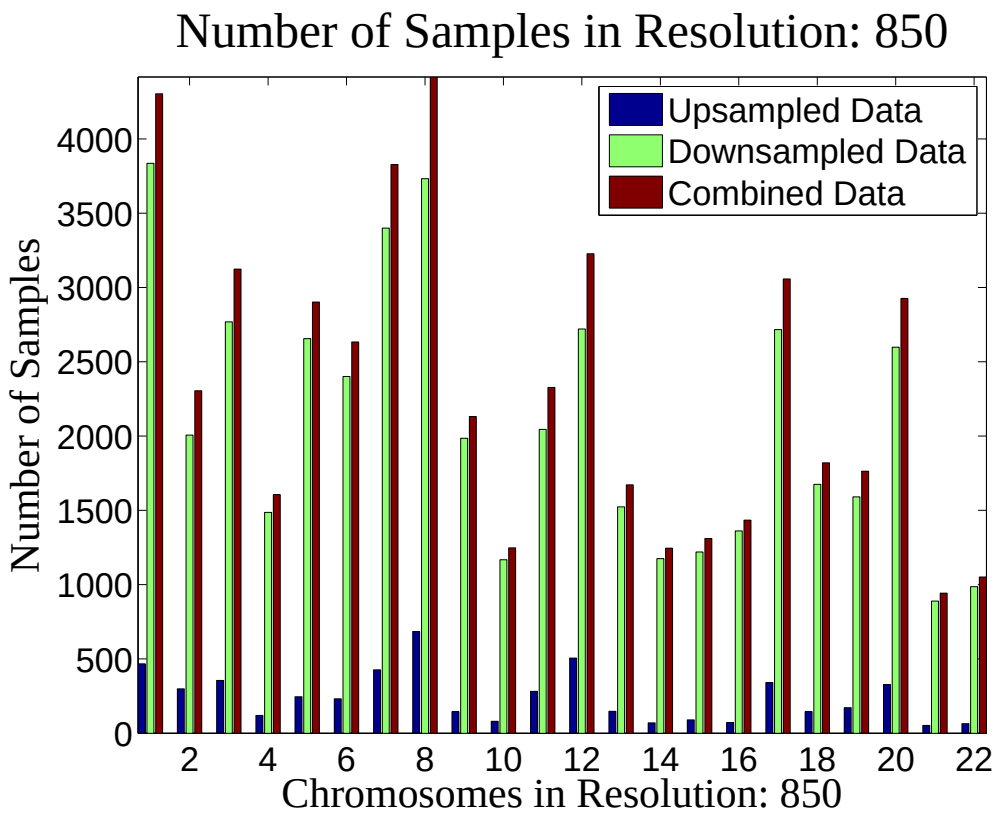

Figure 4.10: Number of samples of data in resolution 850. Figure shows the number of samples for three different datasets used for modelling in this thesis: Upsampled, Downsampled, and Combined.

to the number of samples in downsampled and combined data.

It is important to note that machine learning and data mining algorithms and methods in most cases are data hungry and require significantly large amount of data for plausible results. Thus, database integration is important to work with high dimensional data with small sample size. For example, the validation technique cross-validation used in this thesis has been shown not to work very well with small sized data samples in [27, 28]. A simple example of cross-validation on small sample data is shown in Figure 4.11 as an example case for chromosome 5 in resolution 550. Experiments with different chromosomes have shown that there is a well-defined structure present in data. The details of model selection procedure are discussed in Sections 4.4.1 and 4.4.2.

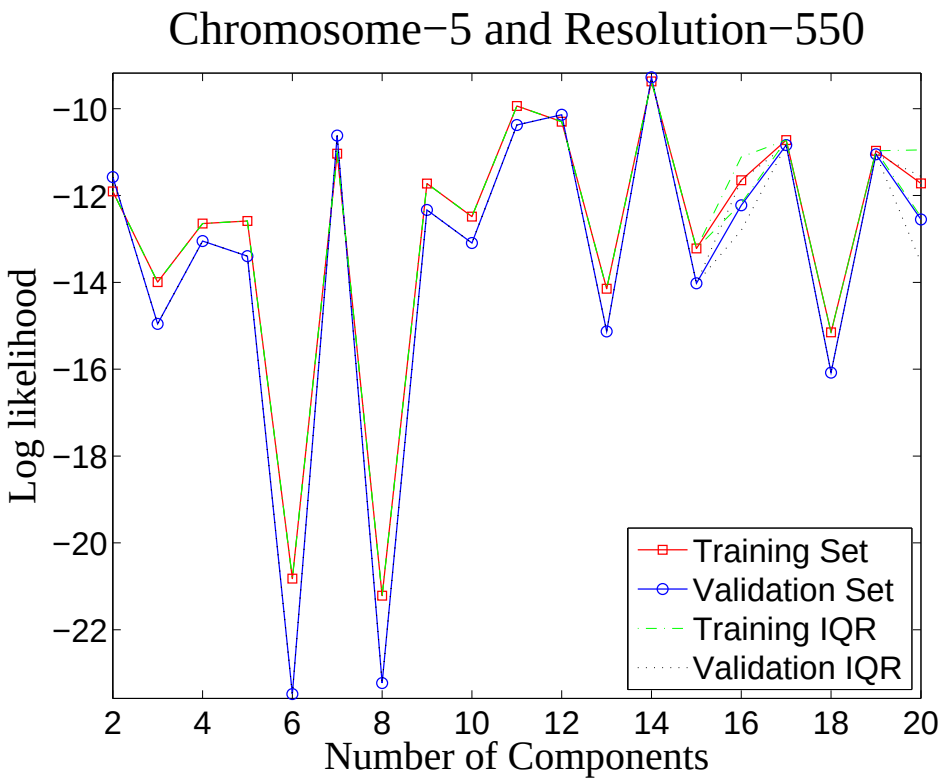

Figure 4.11: Model selection for data in resolution 550. The averaged loglikelihood for training and validation sets in a 10-fold cross-validation setting for different number of components in chromosome 5 & Resolution 550. The interquartile range(IQR) for 50 different training and validation runs have also been plotted. The details of the model selection procedure using cross-validation is discussed in detail in Section 4.4. This example is shown to elaborate that with few samples of data in high dimension (finer resolution) machine learning algorithms such as cross-validation does not work very well.

Same chromosome 5 in the same resolution 550 shows the presence of definite structure in the data when the database is integrated.

Unlike many real valued data, the size of the 0-1 data seems to be significantly large, often large datasets are turned to 0-1 data for the ease of analysis. For instance, consider the size of some of the benchmark datasets: RETAIL [36] is 200000 by 20000; KOSARAK is 100000 by 40000 as described and preprocessed in [35]. One important issue to note is that the main property of

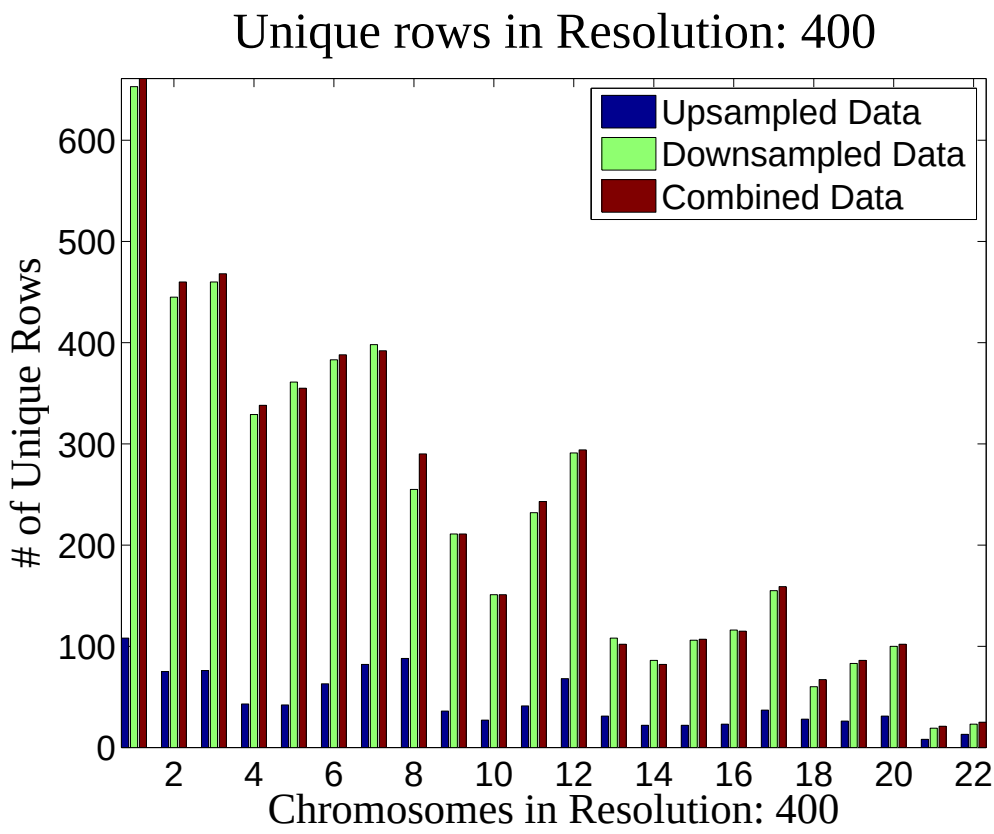

Figure 4.12: Number of unique samples of data in resolution 400. Figure shows the number of unique samples of data for three different datasets used for modelling in this thesis: Upsampled, Downsampled, and Combined.

the 0-1 data is their large dimension. However, dataset at our disposal was relatively small and the problem was further compounded by the presence of few unique rows thus making database integration inevitable.

In general 0-1 datasets, even when the data set is large, ratio of unique rows to the number of samples in the dataset is also approximately 1 i.e. all of the rows in the data are unique. Figure 4.12 shows the number of unique rows for the dataset used in the experiments consisting of 22 chromosomes in 4 different resolutions. Figure 4.12 shows that unlike the other 0-1 dataset, the dataset used in the experiments has very few unique rows. Furthermore, the number of copies of unique rows are not evenly distributed. Additionally, the amplification data is more skewed and sparse. For example, element-wise

AND operation between elements in the same column results in zero vector. Thus, in this setting, with a very few samples of data, cross-validation can always suffer from the problem of "`unfortunate split`"[2]. When database is integrated, the number of samples in the dataset increases also increasing the number of unique samples in the dataset. Thus, experiments were performed after primarily combining the datasets in different resolutions as well as two different resolutions independently in order to compare the results.

4.4 Mixture Modelling of Multivariate Bernoulli Distributions

4.4.1 Model Selection

Model selection is a process of selecting the best model from a set of possible models that optimally fit the data. It is one of the most challenging tasks in machine learning and there are no well defined rules to select the best model and this is an "unsolved" problem in statistics. Often, model selection depends on the use of some prior information, especially about the data, and '*the rule of thumb*'[3]. In other words, the model selection itself can be regarded as "Data Mining". A simple prototyping of models and their statistical analysis can be used to select the model. However, such process will be highly cumbersome. For example, given a machine learning problem, it is very difficult, if not impossible to select the best method from a myriad of the machine learning method such as Support Vector Machines [84, 85], Multilayer Perceptions [19, 86], Extreme Learning Machine [87], among many others. In this thesis, the problem was to analyze copy number aberrations data relevant to cancer. Cancer is not a single disease but a heterogeneous collection of several diseases. We decided to work in the probabilistic context and decided to model the data using a model that possesses clustering capabilities. Furthermore, cancer is

[2]For example, in a classification problem, if certain class is not represented by training set, then the model is not trained to classify it thus producing poor results on the future data.

[3]Definition from Merriam-Webster Online Dictionary: Rule of thumb - a general principle regarded as roughly correct but not intended to be scientifically accurate.

a multifactorial disease. Therefore, mixture modelling was selected to model the copy number aberrations data because they provide an efficient method of modelling the heterogeneous population. Furthermore, since the copy number aberrations data was a high dimensional 0-1 data, the distribution used in the mixture model is the Bernoulli distribution. However, mixture models are too complex in bigger dimension in terms of both time and space complexity. Furthermore, the chromosomal aberrations dataset analyzed in this thesis uses very scarce data as explained in Section 4.2. Thus, we decided to work chromosomewise because of the availability of very few samples of the data to constrain the complexity of the mixture models.

4.4.2 Model Structure Selection

After the selection of the model, the solution of one difficult problem is accomplished but another one awaits which is the problem of model structure selection. The model structure selection is the application of statistical methods for selecting the parameters and hyperparameters of the model. For example, given a machine learning problem, we choose to model it with polynomial curve fitting assuming some prior knowledge that the model is not linear. Even in polynomial curve fitting: the choice among $ax + b$, $ax^2 + bx + c$ and other higher order polynomials is an arduous task. The concept of underfitting, overfitting, bias-variance dilemma (trade-off) are the central issues to be considered in model structure selection. These are very important concepts in machine learning but the thesis does not consider the details of these methods. The details of these concepts can be acquired from Sections 6.1, 9.1 of [20]; Sections 6.8, 6.9 of [88]; and Sections 2.13, 4.13 of [19]. In this thesis, the model selected is mixture models. The hyperparameters of mixture models are the number of mixture components [89]. Therefore, the model structure selection problem in this thesis is restricted to the selection of number of mixture components in the mixture model.

The size of the chromosome in terms of number chromosome bands and also the number of samples varied significantly which are tabulated in Table C.1. Some chromosomes had greater number of bands and some chromosomes had lower number of chromosome bands. Data from different resolutions were in-

dividually subjected to the mixture models. The central problem in this case is model selection which is to determine the number of components in the mixture model. We used the 10-fold cross-validation approach to train the model of different complexities. The exercise was repeated 50 times i.e. for each mixture component, 50 models were trained using training set and their performance was evaluated on the test set. It is often recommended to repeat cross-validation technique a number of times because 10-fold cross-validation can be seen as a "standard" measure of the performance whereas ten 10-fold cross-validations would be a "precise" measure of performance [56]. Since EM algorithm is sensitive to the initializations and the results may differ on the same data for different initializations and it can susceptible local minima and the global optimum results are not often guaranteed [90], 50 different models were trained for each number of components. The number of mixture components was varied from 2 to 20 for all chromosomes in all resolutions. The assumption here is that each chromosome has at least two clusters and more than 20 clusters overfits the data for each chromosome where the maximum dimensionality of the data was 63. Validation set for each model is the one remaining subset of the data which is not used for training. Total likelihood for the training data as well as the validation data is calculated and averaged for each mixture component. We select the model that is able to produce better generalization capabilities (i.e. the number of components for which the likelihood is maximum) taking parsimony into account. In other words, in some cases, models with lesser mixture components are selected instead of models with the larger number of mixture components for which likelihood was higher. We also calculate Interquartile Range(IQR) for training and validation likelihood to analyze the statistical dispersion of the likelihood in different models for the same number of components. Components, for which the variation in IQR is high or which shows more dispersion, are not reliable and hence avoided in most cases. Model selection was performed on all chromosomes as chromosomewise analysis can reveal interesting facts about the aberrations of specific chromosomes and guarantees efficient computation & ease of analysis. Figures 4.13 and 4.14 show model selection procedure for the data in resolution 400 and 850 respectively.

The figure also shows the model selection in case of resolution 400 which

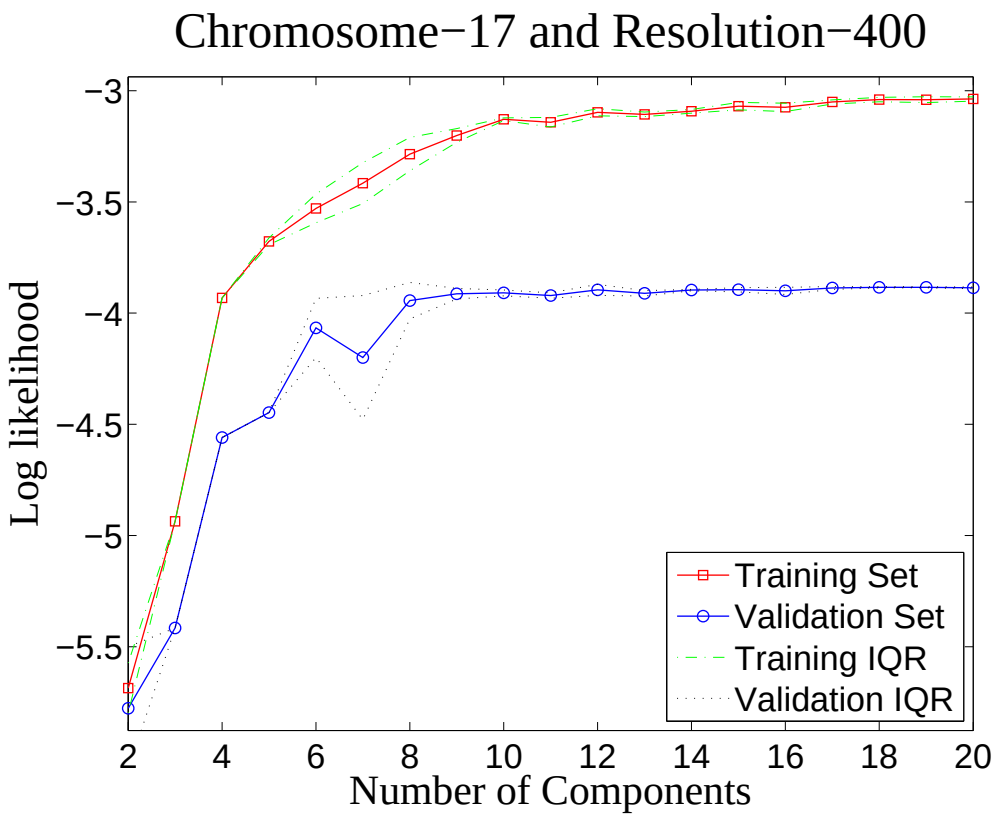

Figure 4.13: Model selection for the the original data in resolution 400. The averaged loglikelihood for training and validation sets in a 10-fold cross-validation setting for different number of components in chromosome 17 & Resolution 400. The interquartile range(IQR) for 50 different training and validation runs have also been plotted. Here, number of components (J) selected is 6.

downsampled from resolution 850. Figure 4.13 shows that the likelihood is smoothly increasing function with respect to the number of components. From Figure 4.13, it can be seen that validation likelihood is maximum when the number of components is 12, but instead of 12 components, 6 components was selected. It is to be noted that sometimes complex models overfit the data and the simple models reduce the time and space complexity. Furthermore, the training and validation likelihood when the number of components is 6 are -3.5293 and -4.0666. In addition, when the number of components is 12, the training and validation likelihood are -3.0972 and -3.8956. Hence, the difference in likelihood is negligible when compared with the efficiency in

terms of time and space complexity. Furthermore, when the number of components is increased, IQR(Inter Quartile Range) shows significant variation. The variation in IQR is because when the number of components is increased, samples can be assigned to different clusters in different runs of the k-fold cross-validation. Additionally, the data in resolution in 400 was upsampled to resolution 850 and similar approach to select the model was followed.

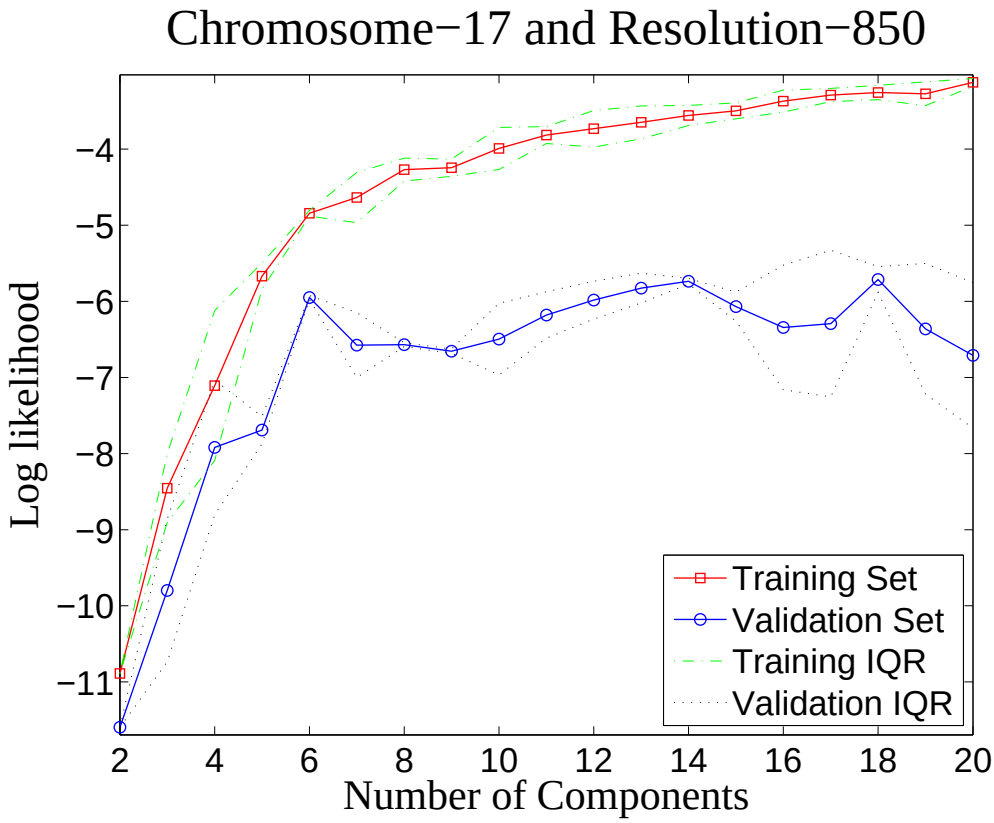

Figure 4.14: The averaged loglikelihood for training and validation sets in a 10-fold Cross-validation setting for different number of components in chromosome 17 and resolution 850. The interquartile range(IQR) for 10 different training and validation runs have also been plotted. Here, number of components (J) selected is 6.

Figure 4.14 also shows that the IQR varies significantly from the mean likelihood. The choice of the number of components is straightforward because Figure 4.14 clearly shows a maximum of validation likelihood when the number of components is 6. Even when the number of components is 6, the variation

in IQR is also low. However, the variation in IQR can be compensated with sufficient training which would produce favorable results. Thus, we train different models and select the best one among them as discussed in Section 4.4.2. The results can be further improved when the size of the dataset is increased which motivates our upsampling and downsampling strategies for database integration.

Parameter Estimation

After the selection of model and its hyperparameters are performed, the parameter estimation is relatively a simple task. Parameter estimation is also often referred to as model fitting, model training or model learning in machine learning literature [20, 18]. Consider, for example, in the above case of polynomial curve fitting, assume that we selected the model is $ax + b$. Now, the value of a and b can be optimized or learned from the data.

In this thesis, after the number of components are selected, the model is trained with all the available data to determine the optimal value of the Bernoulli parameter θ using EM algorithm [29, 31]. In order to achieve the best results while finally selecting the model after selecting the number of components, we further train 50 different models of the same complexity (i.e. the same number of components) to convergence and select the best model in terms of the likelihood produced on the original data. The value of θ specify the probability that a random variable takes the value 0 or 1.

The best of the trained models are used to calculate the likelihood on data as shown Table 4.4. The model was also used to sample the data to be used in validation using resampling approach as discussed in Section 4.4.6. Figures 4.15 and 4.16 are the final models trained to convergence for combined data in resolution 400 and 850 respectively. Similarity of the models can be tracked visually from the model visualization as in Figures 4.15 and 4.16. For example, Component 6 in Figure 4.15 corresponds to Component 1 in 4.16. Similarly, Component 1 in Figure 4.15 corresponds to Component 4 in Figure 4.16.

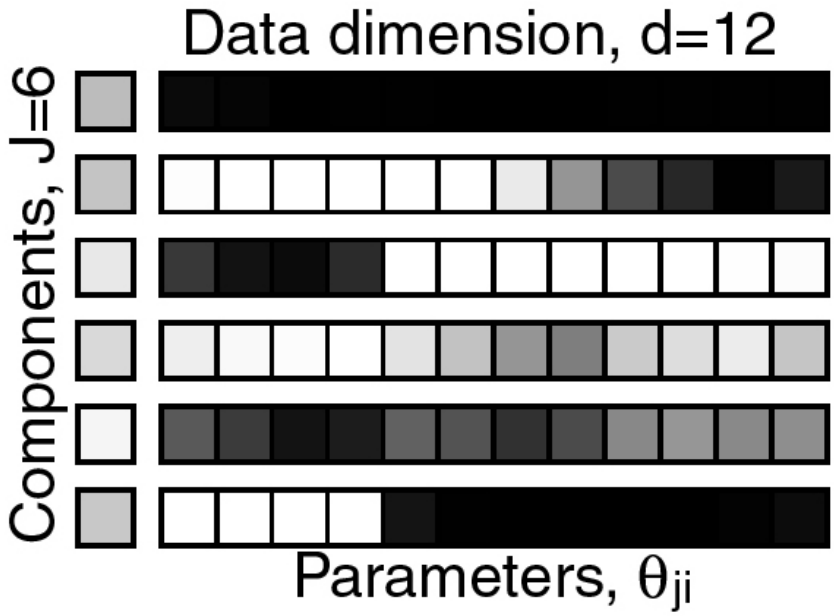

Figure 4.15: Visualization of one of the trained models for chromosome 17 in resolution 400 for combined data. Here the selected number of components is 6 which corresponds to the rows in the model. The first separate column determine the mixing proportions of each mixture component. The remaining 12 columns determines the parameters θ_{ji}. Darker colors denotes the higher values of the parameters.

4.4.3 Computational Complexity

The major drawback in using mixture models is the computational complexity of training the mixture models. Normally, training mixture models are computationally expensive when compared with other parametric (such as Poisson distribution [38]) as well as non-parametric (such as k-means [44, 45]) methods. Similar to other machine learning methods, computational complexity of the mixture model also increases with increasing dimension which is determined by resolution in our case. Thus, computational complexity was also estimated for each resolution for the selected number of components. As shown in the Table 4.3, the computational complexity increases in the fine resolution. To estimate the training time, fifty different models are trained until ten iterations and the mean of the result is taken as final training time. Similarly, likelihood

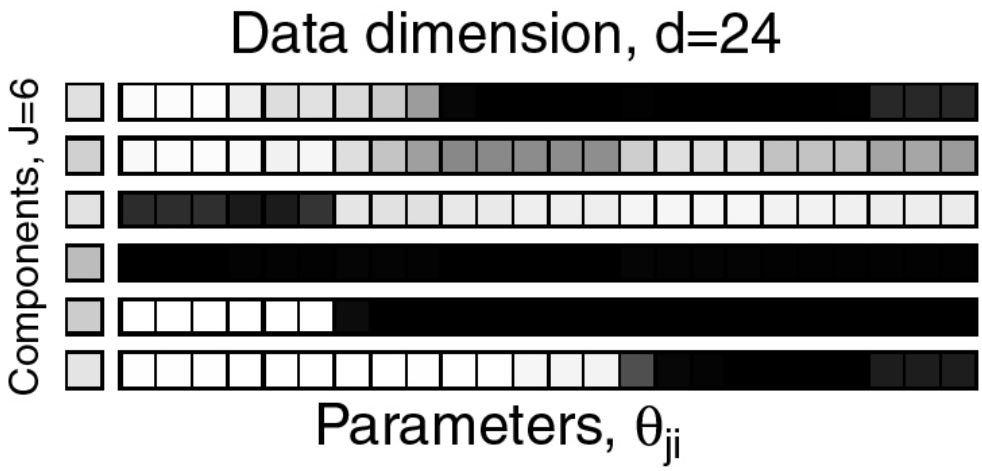

Figure 4.16: Visualization of one of the trained models for chromosome 17 in resolution 850 for combined data. Here the selected number of components is 6 which corresponds to the rows in the model. The first separate column determines the mixing proportions of each mixture component. The remaining 24 columns determine the parameters θ_{ji}. Darker colors denotes the higher values of the parameters.

Chromosome 17			
Data Resolution	**# of Samples**	**Time in Seconds**	
		Training	**Testing**
Original in 400(A)	342	0.25	0.06
Original in 850(B)	2716	0.43	0.30
Downsampled to 400 from B(C)	2716	1.12	0.20
Upsampled to 850 from A(D)	342	2.16	0.08
Combined in 400(A+C)	3058	1.43	0.19
Combined in 850(B+D)	3058	2.51	0.32

Table 4.3: Computational complexity for training and testing of a single mixture model with appropriate number of mixture components as decided in Table 4.4. Experiments are performed on chromosome 17 and time is calculated in seconds. X denotes the number of data samples. The hardware used is Intel Core2Duo 2.00GHz CPU with a memory of 3 GB.

is calculated for fifty different models trained to calculate the training time and the mean of the results is reported. Experiments with resolution 850 required

approximately twice the time required for resolution 400. Furthermore, from Table 4.4, we also know that number of components required is high when the resolution is increased but the likelihood decreases. In addition, the curves are smoother in Figure 4.13 compared with Figure 4.14. This phenomenon is because of the intrinsic problems of working with high dimensional data arising in fine resolution, the phenomenon is often referred to as the '*curse of dimensionality*' [91]. These results suggest that data in lower resolution is preferred but lower resolution does not capture all the available biological information. Thus, there is a trade-off between the two.

4.4.4 Experimental Design

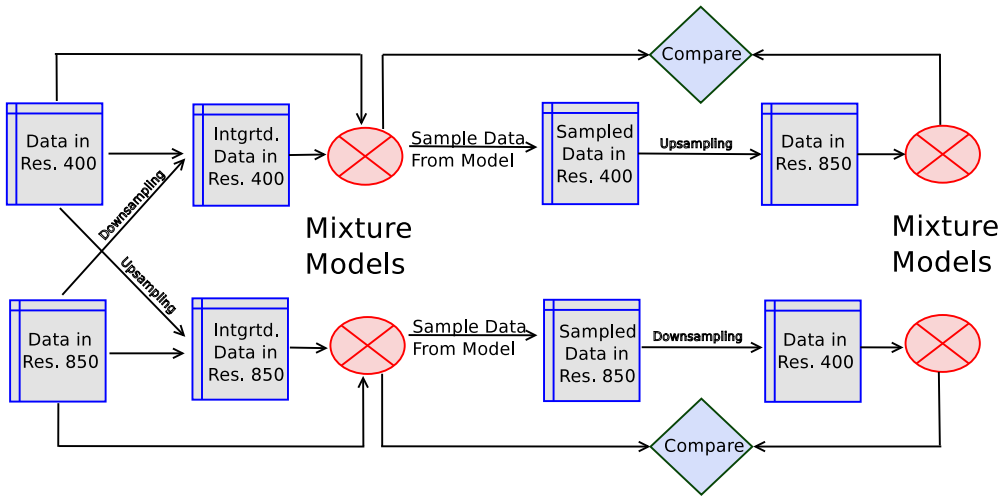

Figure 4.17: The overall experimental procedure in this thesis.

Experimental procedure is as depicted in Figure 4.17 shows that there are two sets of data in two different resolutions: 400 and 850. We use upsampling and downsampling to integrate the data. We then model the data using mixture models. We also model the data individually without integration so that we can compare the results when the database is integrated. We use 10-fold cross-validation repeated fifty times to select the number of components in the mixture model. After selecting the number of components, fifty different models were trained to convergence and best of the trained models in terms of likelihood is taken as the final model for the data. Since the mixture models

are generative models, we sample the data from the trained mixture models. We then than use the same model selection approach to the sampled data after upsampling and downsampling so that we can compare the results. We also compare frequent patterns in the original and the sampled data to evaluate whether our sampling and modelling effort has preserved the overall structure in the data as well as the frequent patterns in the data.

4.4.5 Results

The major aim of upsampling and downsampling was to aid in the integration of databases. The clinical aspects regarding the classification of cancer with mixture models is already established in [22] and [64]. Thus, data in different resolution are integrated after upsampling and downsampling and model selection was performed. Table 4.4 summarizes the results of the experiments on chromosome 17 in different resolutions. To calculate the Likelihood 50 different models were trained to convergence and likelihood of the data was calculated for each model and the mean of the results are reported.

Data Resolution	**Components (J)**	**Likelihood**
Original in 400(A)	6	-3.39
Original in 850(B)	8	-4.53
Downsampled to 400 from B(C)	7	-3.27
Upsampled to 850 from A(D)	8	-4.31
Combined in 400(A+C)	6	-3.48
Combined in 850(B+D)	6	-5.20

Table 4.4: Results of experiments on chromosome 17 in different resolutions showing the number of components required to fit the data along with their respective likelihood. The results for other chromosomes are summarized in Appendix B.

Table 4.4 shows the number of components required to fit the data differs in different resolution. The likelihood of data in fine resolution is lower than the likelihood of the data in the coarse resolution when the number of components are the same. This phenomenon can be attributed to the curse of dimensionality [91]. For example, the dimensionality of data in resolution 400 and 850 differs by 12 in chromosome 17 but likelihood is lesser even when the number

of components is equal. For the original data in resolution 400 and 850, the difference in number of parameters of the model is $6 * (1 + 26) - 6 * (1 + 18) = 48$ which invites significant amount of computational complexity. The increased complexity however does not produce corresponding the increase in the likelihood. With increasing samples, the number of components is not increased because the complexity of mixture models depends on the complexity of the problem being solved, not with the size of dataset. Table 4.5 summarizes the final results of the experiments in all chromosomes.

Upsampled Data						
Resolution	**# of Components**			**Log Likelihood**		
	Mean	*Mode*	*Std. Dev.*	*Mean*	*Mode*	*Std. Dev.*
400	5.1363	4	1.3200	-4.1170	-6.8321	1.3194
550	5.8181	5	1.4354	-5.6478	-12.925	3.3085
700	5.6818	5	1.6442	-9.3383	-21.0159	5.6227
850	6.4091	8	1.8685	-10.2319	-20.7890	6.2510
Downsampled Data						
Resolution	**# of Components**			**Log Likelihood**		
	Mean	*Mode*	*Std. Dev.*	*Mean*	*Mode*	*Std. Dev.*
400	6.1818	7	0.9579	-4.3354	-8.0169	1.7914
550	6.8181	7	1.1396	-5.4993	-11.7850	2.8190
700	6.8181	7	0.9579	-7.2905	-13.4629	3.8663
850	7.0000	6	1.2344	-8.1149	-15.0200	4.0383
Combined Data						
Resolution	**# of Components**			**Log Likelihood**		
	Mean	*Mode*	*Std. Dev.*	*Mean*	*Mode*	*Std. Dev.*
400	6.2272	6	1.1097	-4.3801	-8.1897	1.7546
550	6.6818	6	1.1291	-5.6528	-11.6969	2.8333
700	6.6818	7	1.0413	-7.6022	-13.4560	4.0573
850	6.8181	7	1.1396	-8.3920	-16.5590	4.2943

Table 4.5: Summary of results of experiments on all showing the number of components required to fit the data along with their respective likelihood. Here *Std. Dev.* is the standard deviation (σ.) The details of the results each chromosome are tabulated in Appendix B.

Table 4.5 shows that the number of components selected for the data is

highly co-related with the data resolution: the finer the resolution the higher the number of components required. Increasing resolutions require more number of components and the likelihood of the data also decreases. This phenomenon can be attributed to the curse of dimensionality [91]. The difference in likelihood showing poorer fit to the data is clearly captured by the increasing standard deviation (σ) where in each of the three cases of three different datasets, the standard deviation for the likelihood increases significantly. There is only small differences in selection of number of components where as there is a significant difference in the likelihood of the final model. This behavior can also be attributed to the fact that the models selected in our case were parsimonious models. The models of higher complexity were not selected even if it produced higher validation likelihood for the fear of overfitting and computational & space complexity of complex models. Especially models of complexity greater than ten were discarded. Furthermore, similarity in the number of components also shows that mixture models learns the structure of data relatively well although it is constrained by the increasing dimensionality of the data in finer resolution.

In order to capture the notion of decreasing number the likelihood of data in 22 different chromosomes in 4 different resolutions, we plot the parallel coordinates of the log-likelihood in all three datasets: upsampled, downsampled, and combined. The plots for the three cases are similar, therefore, only the plot for combined data has been shown in Figure 4.18. The trend of decreasing likelihood can be easily captured from Figure 4.18. In few cases, such as chromosome 22 and other small chromosomes[4], the trend in decrease is not significant because the difference in number of chromosome bands (regions) is negligible in the smaller chromosomes.

The computational complexity increases in the finer resolution. Moreover, high-resolution data incorporates significant amount of noise thus producing explosion of redundant patterns thus requiring more number of components to optimally fit the data. The problem with such redundant patterns can be remedied by downsampling the data if the information loss is insignificant

[4]The chromosomes are numbered by their size with only one exception i.e. chromosome 21 is smallest instead of chromosome 22. Thus, chromosome 21 is smallest chromosome while chromosome 1 is the largest of the chromosomes.

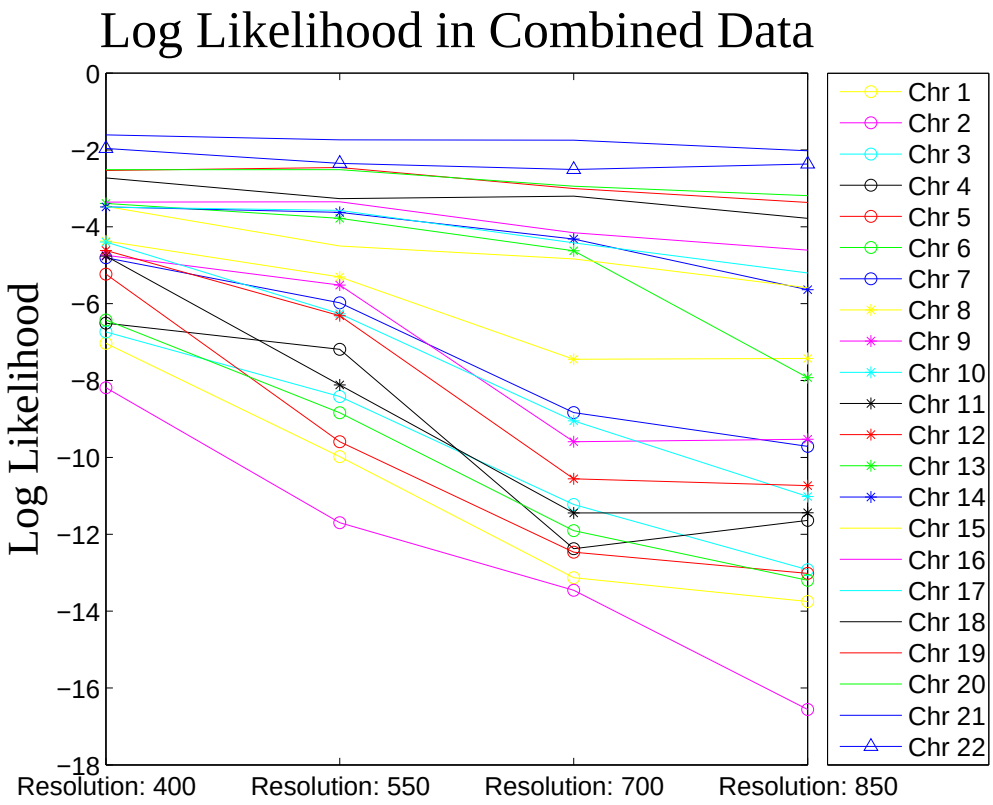

Figure 4.18: Parallel co-ordinates plot for the likelihood of combined data of 22 different chromosomes in 4 different resolutions.

during downsampling.

4.4.6 Validation Using Data Resampling approach

This experiments with the mixture models also show that patterns present in the fine resolution of the data are efficiently and effectively preserved in coarse resolution. Since the mixture models are generative models, we can sample the data from the trained models. Thus, in order to validate the model and determine if it has been able to extract the original structure in the data, we sample the data where the number of samples in the sampled data are equal to the number of samples used in training. We repeat the same model selection procedure as discussed in Section 4.4. It has been shown in [33] that the generative mixture models preserves the statistically significant patterns in multiresolution 0-1 data. From Figure 4.19, we can see that the number of

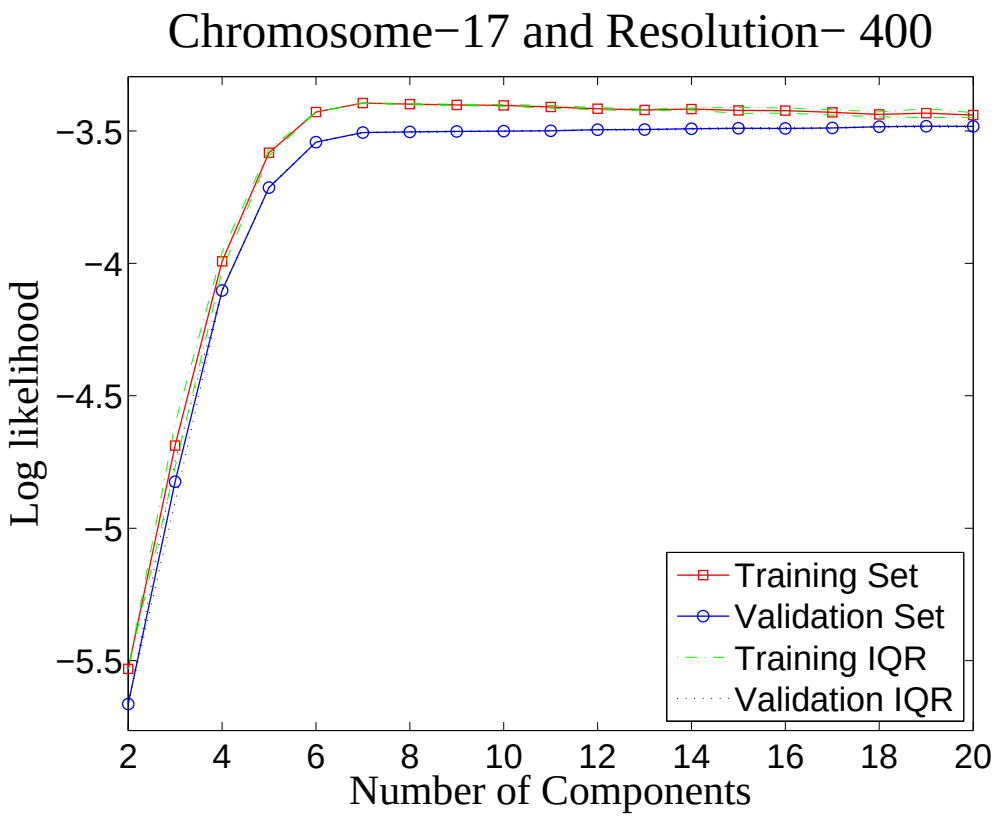

Figure 4.19: The averaged loglikelihood for training and validation sets in a 10-fold Cross-validation setting for different number of components in chromosome 17 and resolution 400 in resampled data from the model of combined data. The interquartile range(IQR) for 10 different training and validation runs have also been plotted. The number of components selected here is 6.

selected components distributions are similar to the original data including the variations in the likelihood with increasing number of components. However, as expected the curve is more smooth. As with all the other experimental procedure, this validation using the data resampling approach was performed in all the chromosomes. There were very few discrepancies which occurred especially in upsampled data. The reason being that there were very few samples of the data in upsampling. We further train the mixture model on the resampled data using the selected number of components. The model trained on the resampled data is also used to calculate the likelihood on the original data.

Data Resolution	J	**Likelihood in**	
		Original	**Resampled**
Original in 400(A)	6	-3.70	-3.32
Original in 850(B)	8	-4.57	-4.66
Downsampled to 400 from B(C)	7	-3.28	-3.26
Upsampled to 850 from A(D)	8	-4.72	-4.30
Combined in 400(A+C)	6	-3.49	-3.49
Combined in 850(B+D)	6	-5.69	-5.61

Table 4.6: Results of experiments on chromosome 17 showing the number of components required to fit the data along with their respective likelihood for the data sampled from the mixture model. J denotes the number of components selected.

An example result reported in Table 4.6 shows that the result is very similar to the original data. The results for other chromosomes were also very similar. The model trained from the resampled data is further used to calculate the likelihood on the original data. The likelihood decreases but the decease is negligible showing that our parsimonious mixture models efficiently captures the overall structure of the data.

Chapter 5

Summary and Conclusion

> *Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning.*
>
> — Sir Winston Churchill
> *After Victory at El Alamein (1942)*

Synopsis

This chapter presents a summary of the work, draws conclusions from experimental results and discusses future areas of research.

5.1 Summary and Conclusions

This thesis studied the problem of multiresolution data in chromosomal aberration. Two datasets were available in different resolutions. In order to work with the multiple resolutions of the data, a upsampling and three different downsampling methods were proposed and their results were studied. The results were plausible and fairly consistent. The resulting data in different resolutions efficiently captures the information of data in different resolutions. Significant patterns and overall structure of the data were effectively preserved during the data transformation process. The major aim of data transformation across different resolutions was to aid in the integration of databases.

Thus, after transformation to different resolutions, data was integrated for the analysis in one resolution.

Mixture models were then applied to the data in different resolutions for all three different types of data: upsampled, downsampled, and combined. We used 10-fold cross validation approach for model selection in mixture models. The analysis of the data was performed chromosomewise in different resolutions. The results suggested that number of components required to fit the data differs across resolutions and increasing resolutions require more number of components. Furthermore, the likelihood of the model on finer resolution is poorer than that of coarse resolution although the data is the same but representation is different. Moreover, the number of components required to the fit the data is increased. The performance of the algorithm in integrated data was better than the ones performed individually in two different resolutions thus showing the importance of our data transformation process.

The trained mixture models can be used in cancer classification and clustering. The clustering results of mixture models possess high clinical significance as shown in [22] and [64]. Furthermore, validation by resampling showed that mixture models trained parsimoniously preserve the original structure of the data. There were only negligible discrepancies on the results of the mixture models on the data sampled from the model. The computational complexity increases with increasing resolution. Experiments with resolution 850 required approximately twice the time required for the resolution 400.

5.2 Future Work

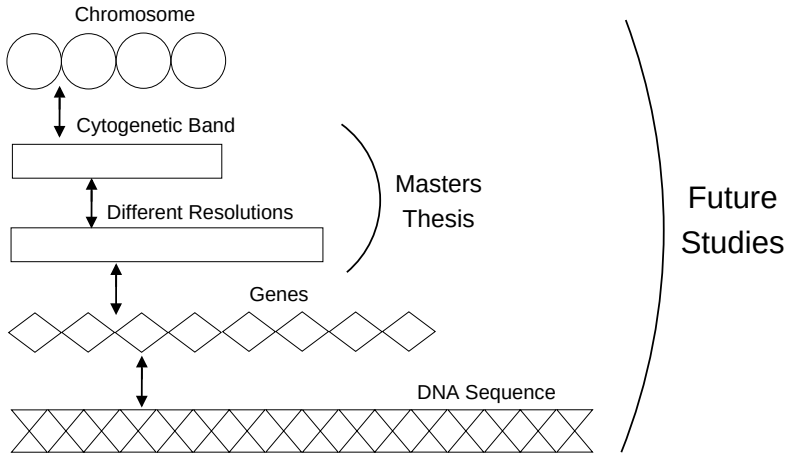

Figure 5.1: Schematic representation of problem studied in the Master's thesis and its seamless extension to the problem to be studied in future.

The multiresolution problem was studied only at chromosome level and the data transformation process was defined only in different resolutions of the chromosome. In the future work, the data transformation process can be defined until the very minute biological details such as genes and DNA sequences. Upsampling technique used in the thesis also needs further investigation and inferencing techniques can be implemented. In further work the probabilistic models, such as mixture models and probabilistic time series models, such as Hidden Markov Models(HMMs) can be extended to cope with data in multiple resolutions.

Bibliography

> *If I have seen further, it is by standing on the shoulders of giants.*
>
> — Isaac Newton
>
> *In a letter to his rival Robert Hooke (1676)*

[1] L. G. Shaffer and N. Tommerup. *ISCN 2005: An International System for Human Cytogenetic Nomenclature(2005) Recommendations of the International Standing Committee on Human Cytogenetic Nomenclature*. Karger, 2005.

[2] T. Raiko, K. Puolamäki, J. Karhunen, J. Hollmén, A. Honkela, H. Mannila, E. Oja, and E. Simula. Macadamia: Master's Programme in Machine Learning and Data Mining. In *Proceedings of Teaching Machine Learning Workshop on Open Problems and New Directions*, 2008.

[3] J. F. Bishop. *Cancer facts : a concise oncology text*. Harwood Academic Publishers, Amsterdam, The Netherlands, 1999.

[4] World Health Organization. Cancers: World Health Organization Factsheet No. 297. Website, February 2009. `http://www.who.int/mediacentre/factsheets/fs297/en/print.html`: Last Accessed: 15 Nov 2010.

[5] American Cancer Society, Inc. Cancer Facts & Figures 2006 No. 500806. Website, 2006. http://www.cancer.org/ Research/ CancerFactsFigures/CancerFactsFigures/index : Last Accessed: 15 Nov 2010.

[6] Cancer Society of Finland. Cancer 2015. Website, 2010. http://www.cancer.fi/english/organisation/publications /cancer_2015 : Last Accessed: 15 Nov 2010.

[7] A. Vellido and P. J. G. Lisboa. Neural Networks and Other Machine Learning Methods in Cancer Research. In *International Work-Conference on Artificial Neural Networks*, pages 964–971, 2007.

[8] A. Kallioniemi, O. P. Kallioniemi, D. Sudar, D. Rutovitz, J. W. Gray, F. Waldman, and D. Pinkel. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *SCIENCE*, 258(5083):818–821, OCT 30 1992.

[9] D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B. M. Ljung, J. W. Gray, and D. G. Albertson. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, 20: 207 – 211, 1998.

[10] B. Ylstra, P. van den Ijssel, B. Carvalho, R. H. Brakenhoff, and G. A. Meijer. BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). *Nucleic acids research*, 34(2):445–450, 2006.

[11] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14(13):1675–1680, 1996.

[12] Y. Wang, J. Hayakawa, F. Long, Q. Yu, A. H. Cho, G. Rondeau, J. Welsh, S. Mittal, I. De Belle, E. Adamson, M. McClelland, and D. Mercola. "promoter array" studies identify cohorts of genes directly regulated by methy-

lation, copy number change, or transcription factor binding in human cancer cells. *Annals of the New York Academy of Sciences*, 1058:162–185, 2005.

[13] S. C. Schuster. Next-generation sequencing transforms today's biology. *Nature Methods*, 5(1):16–18, December 2007.

[14] E. R. Mardis. Next-generation DNA sequencing methods. *Annual review of genomics and human genetics*, 9(1):387–402, June 2008.

[15] R. A. Stein. Next-generation sequencing update. *Genetic Engineering and Biotechnology News*, 28(15), 2008.

[16] J. C. Venter et. al. The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507):1304–1351, February 2001.

[17] E. S. Lander et.al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

[18] T. M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

[19] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Macmillan, New York, 1994.

[20] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[21] C. Jong. *Machine learning for human cancer research*. PhD thesis, VU University, Amsterdam, 2006.

[22] S. Myllykangas, J. Tikka, T. Böhling, S. Knuutila, and J. Hollmén. Classification of human cancers based on DNA copy number amplification modeling. *BMC Medical Genomics*, 1:15, 2008.

[23] I. R. Kirsch. *The Causes and Consequences of Chromosomal Aberrations*. CRC Press, first edition, December 1992.

[24] J. A. Lee, C. M. B. Carvalho, and J. R. Lupski. A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. *Cell*, 131(7):1235 – 1247, 2007.

[25] F. Zhang, M. Khajavi, A. M. Connolly, C. F. Towne, S. D. Batish, and J. R. Lupski. The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nature Genetics*, 41(7):849–853, 2009.

[26] J. Tikka, J. Hollmén, and S. Myllykangas. Mixture modeling of DNA copy number amplification patterns in cancer. In Francisco Sandoval, Alberto Prieto, Joan Cabestany, and Manuel Graña, editors, *Computational and Ambient Intelligence*, volume 4507 of *Lecture Notes in Computer Science*, pages 972–979. Springer Berlin / Heidelberg, 2007.

[27] A. Isaksson, M. Wallman, H. Göransson, and M. G. Gustafsson. Cross-validation and bootstrapping are unreliable in small sample classification. *Pattern Recognition Letters*, 29(14):1960–1965, 2008.

[28] U. M. Braga-Neto and E. R. Dougherty. Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20(3):374–380, February 2004.

[29] J. H. Wolfe. Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, 5:329–350, 1970.

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

[31] B. S. Everitt and D. J. Hand. *Finite mixture distributions*. Chapman and Hall, London; New York:, 1981.

[32] P. R. Adhikari and J. Hollmén. Patterns from multiresolution 0-1 data. In *Proceedings of the ACM SIGKDD Workshop on Useful Patterns*, UP '10, pages 8–16, New York, NY, USA, 2010. ACM.

[33] P. R. Adhikari and J. Hollmén. Preservation of statistically significant patterns in multiresolution 0-1 data. In Tjeerd Dijkstra, Evgeni Tsivtsivadze, Elena Marchiori, and Tom Heskes, editors, *Pattern Recognition in Bioinformatics*, volume 6282 of *Lecture Notes in Computer Science*, pages 86–97. Springer Berlin / Heidelberg, 2010.

[34] K. Puolamäki, M. Fortelius, and H. Mannila. Seriation in Paleontological Data Using Markov Chain Monte Carlo Methods. *PLoS Computational Biology*, 2(2):e6, 2006.

[35] A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas. Assessing data mining results via swap randomization. *ACM Transactions on Knowledge Discovery from Data*, 1(3):14, 2007.

[36] T. Brijs, G. Swinnen, K. Vanhoof, and G. Wets. Using association rules for product assortment decisions: a case study. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 254–260, New York, NY, USA, 1999. ACM.

[37] G. J. McLachlan and D. Peel. *Finite mixture models*, volume 299 of *Probability and Statistics – Applied Probability and Statistics Section*. Wiley, New York, 2000.

[38] S. D. Poisson. *Recherches sur la probabilité des jugements en matiŕe criminelle et en matiŕe civile: Prćdés des rgles gńŕales du calcul des probabilitś.* Elibron Classics, 1837.

[39] P. Deb. Finite mixture models. Summer north american stata users' group meetings 2008, Stata Users Group, 2008.

[40] R.L. Scheaffer and L.J. Young. *Introduction to Probability and its Applications*. Duxbury Pr, 2009.

[41] P. Boehner. Collected articles on Ockham. *OFM, St Bonaventure*, 1958.

[42] M. Gyllenberg, T. Koski, E. Reilink, and M. Verlann. Non-Uniqueness in Probabilistic Numerical Identification of Bacteria. *Journal of Applied Probability*, 31(2):542–548, 1994.

[43] M. Carreira-Perpiñán and S. Renals. Practical Identifiability of Finite Mixtures of Multivariate Bernoulli Distributions. *Neural Computation*, 12(1):141–152, 2000.

[44] J. Macqueen. Some Methods for Classification and Analysis of Multivariate Observations. *In Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley*, pages 281–297, 1967.

[45] S. Lloyd. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[46] C. Andrieu, N. Freitas, A. Doucet, and M. I. Jordan. An Introduction to MCMC for Machine Learning. *Machine Learning*, 50(1-2):5–43, 2003.

[47] R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. *SIAM J. Comput.*, 38:1141–1156, July 2008.

[48] S. Vempala and G. Wang. A Spectral Algorithm for Learning Mixtures of Distributions. In *Journal of Computer and System Sciences*, pages 113–123, 2002.

[49] G. H. Golub and C. F. Van Loan. *Matrix Computations (Johns Hopkins Studies in Mathematical Sciences)(3rd Edition)*. The Johns Hopkins University Press, 3rd edition, October 1996.

[50] G. Golub and W. Kahan. Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis*, 2(2):205–224, 1965.

[51] D. Ormoneit and V. Tresp. Averaging, maximum penalized likelihood and Bayesian estimation for improving Gaussian mixture probability density estimates. *IEEE Transactions on Neural Networks*, 9(4):639–650, 1998.

[52] S. Geisser. A Predictive Approach to the Random Effect Model. *Biometrika*, 61(1):101–107, 1974.

[53] F. Monsteller and J. Tukey. Data Analysis including statistics. *In Lindzey G. and Aronson E., editors, Handbook of Social Psychology, Vol-2, Addison-Wesley*, 1968.

[54] R. Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *IJCAI*, pages 1137–1145, 1995.

[55] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, second edition, June 2005.

[56] S. D. Gay. *Datamining in proteomics: extracting knowledge from peptide mass fingerprinting spectra*. PhD thesis, University of Geneva, Geneva, 2002.

[57] B. G. Lindsay and M. L. Lesperance. A review of semiparametric mixture models. *Journal of Statistical Planning and Inference*, 47(1-2):29 – 39, 1995.

[58] A. Juan and E. Vidal. Bernoulli Mixture Models for Binary Images. *International Conference on Pattern Recognition*, 3:367–370, 2004.

[59] A. Juan and E. Vidal. On the use of Bernoulli mixture models for text classification. *Pattern Recognition*, 35(12):2705 – 2710, 2002.

[60] M. Saeed and H. Babri. Classifiers based on Bernoulli mixture models for text mining and handwriting recognition tasks. In *IJCNN*, pages 2169–2175, 2008.

[61] M. Aitkin, D. Anderson, and J. Hinde. Statistical Modelling of Data on Teaching Styles. *Journal of The Royal Statistical Society. Series A (General)*, 144(4), 1981.

[62] J. R. Pollack, C. M. Perou, A. A. Alizadeh, M. B. Eisen, A. Pergamenschikov, C. F. Williams, S. S. Jeffrey, D. Botstein, and P. O. Brown. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics*, 23(1):41–46, 1999.

[63] S. Knuutila, Y. Aalto, K. Autio, A. Björkqvist, W. El-Rifai, S. Hemmer, T. Huhta, E. Kettunen, S. Kiuru-Kuhlefelt, M.L. Larramendy, T Lushnikova, O. Monni, H. Pere, J. Tapper, M. Tarkkanen, A. Varis, V. Wasenius, M. Wolf, and Y. Zhu. DNA Copy Number Losses in Human Neoplasms. *Gynecologic Oncology*, 155(2):683–694, 1999.

[64] S. Myllykangas, J. Himberg, T. Böhling, B. Nagy, J. Hollmén, and S. Knuutila. DNA copy number amplification profiling of human neoplasms. *Oncogene*, 25(55):7324–7332, 2006.

[65] J. Hollmén and J. Tikka. Compact and understandable descriptions of mixtures of bernoulli distributions. *Lecture Notes in Computer Science including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*, 4723 LNCS:1–12, 2007.

[66] P. M. V. Rancoita, M. Hutter, F. Bertoni, and I. Kwee. Bayesian DNA copy number analysis. *BMC Bioinformatics*, 10(1):10, 2009.

[67] B. D'haene, J. Vandesompele, and J. Hellemans. Accurate and objective copy number profiling using real-time quantitative PCR. *Methods*, 50(4):262–270, 2010.

[68] E. Despierre, D. Lambrechts, P. Neven, F. Amant, S. Lambrechts, and I. Vergote. The molecular genetic basis of ovarian cancer and its roadmap towards a better treatment. *Gynecologic Oncology*, 117(2):358–365, 2010.

[69] C. W. P. Louis. Flexible, Robust, And Efficient Human Speech Recognition. In *In Proceedings of The XIVth International Congress of Phonetic Sciences*, pages 9–16, 1997.

[70] L. Wall. Perl: Practical Extraction and Report Language. Website, 1987. `http://www.perl.org/` : Last Accessed: 15 Nov 2010.

[71] National Center for Biotechnology Information. Human Genome Project. Website, February 2010. `http://www.ncbi.nlm.nih.gov/projects/mapview/` : Last Accessed: 15 Nov 2010.

[72] F. Leisch. FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. *Journal of Statistical Software*, 11(i08), 2003.

[73] T. Benaglia, D. Chauveau, D. R. Hunter, and D. Young. mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29, 2009.

[74] Mathworks. Matlab: the language of technical computing. Website, 1994. `http://www.mathworks.com/products/matlab` : Last Accessed: 15 Nov 2010.

[75] E. Quigley and S. Hawkins. *The Complete Linux Shell Programming Training Course (CD-ROM Boxed-Set)*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2000.

[76] R. Project. The R project for statistical computing. Website, 1997. `http://www.r-project.org` : Last Accessed: 15 Nov 2010.

[77] J. Hollmén. *BernoulliMix: Program package for finite mixture models of multivariate Bernoulli distributions*, May 2009. `http://www.cis.hut.fi/jHollmen/BernoulliMix` : Last Accessed: 15 Nov 2010.

[78] M. Baudis. Online database and bioinformatics toolbox to support data mining in cancer cytogenetics. *Biotechniques*, 40(3), March 2006.

[79] G. W. Stewart. *Matrix Algorithms: Volume 1, Basic Decompositions*. Society for Industrial Mathematics, 1998.

[80] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *SIGMOD '93: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, New York, NY, USA, 1993. ACM.

[81] H. Mannila, H. Toivonen, and A. I. Verkamo. Efficient algorithms for discovering association rules. In Usama M. Fayyad and Ramasamy Uthurusamy, editors, *AAAI Workshop on Knowledge Discovery in Databases (KDD-94)*, pages 181–192, Seattle, Washington, 1994. AAAI Press.

[82] A. Gallo, P. Miettinen, and H. Mannila. Finding Subgroups having Several Descriptions: Algorithms for Redescription Mining. In *SDM*, pages 334–345, 2008.

[83] D. Burdick, M. Calimlim, and J. Gehrke. MAFIA: A Maximal Frequent Itemset Algorithm for Transactional Databases. In *In ICDE*, pages 443–452, 2001.

[84] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[85] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, September 1998.

[86] F. Rosenblatt. *Principles of neurodynamics; perceptrons and the theory of brain mechanisms*. Spartan Books Washington, 1962.

[87] G. Huang, Q. Zhu, and C. Siew. Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1-3):489–501, December 2006.

[88] K. Gurney. *An Introduction to Neural Networks*. Taylor & Francis Inc., Bristol, PA, USA, 1997.

[89] Z. Zivkovic and F. van der Heijden. Recursive unsupervised learning of finite mixture models. *PAMI*, 26(5):651–656, May 2004.

[90] G. J. Mclachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley-Interscience, first edition, November 1996.

[91] W. Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Wiley, 2007.

Appendix A

Chromosome Nomenclature

> *If people never did silly things, nothing intelligent would ever get done.*
>
> — Ludwig Wittgenstein
>
> *Austrian philosopher (1889 - 1951)*

There is a standardized naming scheme or nomenclature to address the different areas in the genome defined by the International System for Human Cytogenetic Nomenclature (ISCN) [1]. This naming scheme is used by the domain experts and found in the literature when addressing the parts of the genome. The history of chromosome nomenclature dates back to 1971 when a meeting in Paris decided the basic nomenclature for the bands in the chromosome. Hence, the nomenclature is often referred to as Paris nomenclature and some names have been adopted from French.

A chromosome is divided into two arms by the centromere: the p arm which stands for petit (meaning small in French) is the longer arm and the arm q which stands for queue. The regions are named q1, q2, q3 or p1, p2, p3 starting from the centromere and moving towards the edges. Regions are often separated by specific and consistent landmarks which possess distinct morphological characters such as the ends of the chromosome arms, the centromere and certain bands. The regions are further divided into bands such as q11 (pronounced as 'q-one-one' not 'q-eleven'). The bands are further divided into sub-bands such as q11.1 or even sub-sub bands such as q11.11. This naming scheme is hierarchical and irregular.

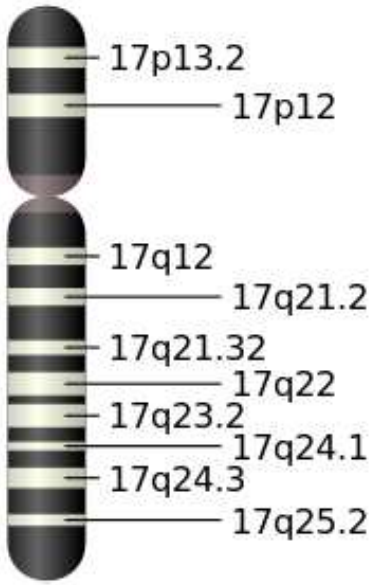

Figure A.1: Nomenclature of Chromosome bands of Chromosome 17 in resolution 400.

An example of chromosome nomenclature is shown in Figure A.1 which is an example case in chromosome 17. For example, area 17q21.32 means chromosome 17, arm 'Q', region 21, band 3 and sub-band 2. It is important to note that length of each region, band and sub-band varies.

ISCN has also defined Ideograms for G-banding patterns for normal human chromosomes at five different resolutions [1]. Five different resolutions of chromosomes mean that a chromosome is divided into different parts in different resolutions. In resolution 400, for example, the chromosome is divided into 393 different parts and in resolution 850, chromosome is divided in 862 different parts. With respect to the chromosome bands region q21 in resolution 400, for example, is divided into q21.1, q21.2, q21.31, q21.32 and q21.33 in resolution 850. However, region q22 in resolution 400 remains undivided in resolution 850 as well. The division of the chromosome bands is determined by the resolution of naming scheme which depends on the properties of the genome.

Appendix B

Results on Each Chromosome

> *There is no such thing as failure. There are only results.*
>
> — Tony Robbins
>
> *American self-help author(1960-)*

Synopsis

This chapter presents a summary of results of experiments on all 22 chromosomes. Number of components required to fit the data (J) along with their respective likelihood ($\mathcal{L}$) in all three types of data: upsampled, downsampled and combined.

Chromosome 1						
Resolution	**Upsampled**		**Downsampled**		**Combined**	
	J	$\mathcal{L}$	J	$\mathcal{L}$	J	$\mathcal{L}$
400	6	-5.9194	7	-7.0697	7	-7.0344
550	7	-12.9252	8	-9.9172	8	-9.9790
700	7	-16.4747	9	-12.399	8	-13.1283
850	7	-14.6505	5	-13.065	7	-13.7521

Chromosome 2						
Resolution	**Upsampled**		**Downsampled**		**Combined**	
	J	$\mathcal{L}$	J	$\mathcal{L}$	J	$\mathcal{L}$
400	4	-6.0295	7	-8.0169	6	-8.1897
550	5	-11.8613	7	-11.7851	7	-11.6974
700	7	-16.5517	7	-13.4636	7	-13.4568
850	7	-20.3705	7	-15.0203	7	-16.5597

Chromosome 3						
Resolution	**Upsampled**		**Downsampled**		**Combined**	
	J	$\mathcal{L}$	J	$\mathcal{L}$	J	$\mathcal{L}$
400	4	-6.2442	6	-7.0561	7	-6.7363
550	7	-7.6133	6	-8.7447	7	-8.4167
700	4	-10.4383	7	-10.5652	6	-11.2252
850	7	-12.4494	7	-11.9871	7	-12.9220

Chromosome 4						
Resolution	**Upsampled**		**Downsampled**		**Combined**	
	J	$\mathcal{L}$	J	$\mathcal{L}$	J	$\mathcal{L}$
400	6	-5.5425	7	-6.7628	8	-6.5087
550	4	-7.3582	6	-6.9300	8	-7.1889
700	3	-14.3333	7	-11.7569	7	-12.3751
850	6	-14.3317	6	-11.3934	7	-11.6412

Chromosome 5						
Resolution	**Upsampled**		**Downsampled**		**Combined**	
	J	$\mathcal{L}$	J	$\mathcal{L}$	J	$\mathcal{L}$
400	5	-3.8016	7	-5.2970	7	-5.2339
550	4	-12.158	8	-9.3618	6	-9.5914
700	8	-21.0161	7	-12.6904	7	-12.4653
850	6	-20.7898	6	-12.8497	7	-13.0176

Chromosome 6						
Resolution	**Upsampled**		**Downsampled**		**Combined**	
	J	$\mathcal{L}$	J	$\mathcal{L}$	J	$\mathcal{L}$
400	3	-6.8321	7	-6.0988	6	-6.4244
550	5	-9.3474	7	-8.2348	6	-8.8388
700	4	-14.9753	6	-10.317	6	-11.9083
850	6	-17.0915	6	-12.560	5	-13.1997

Chromosome 7						
Resolution	**Upsampled**		**Downsampled**		**Combined**	
	J	$\mathcal{L}$	J	$\mathcal{L}$	J	$\mathcal{L}$
400	7	-4.7596	7	-4.4072	6	-4.8041
550	7	-6.4318	7	-5.6954	7	-5.9767
700	5	-12.160	7	-7.5511	5	-8.8348
850	4	-18.9840	4	-8.6133	7	-9.7109

Chromosome 8						
Resolution	**Upsampled**		**Downsampled**		**Combined**	
	J	$\mathcal{L}$	J	$\mathcal{L}$	J	$\mathcal{L}$
400	4	-4.9276	6	-4.1155	6	-4.3724
550	4	-6.2317	7	-4.9172	6	-5.3038
700	4	-10.5181	7	-7.1678	7	-7.4469
850	8	-8.1046	8	-7.1619	7	-7.4235

Chromosome 9						
Resolution	**Upsampled**		**Downsampled**		**Combined**	
	J	$\mathcal{L}$	J	$\mathcal{L}$	J	$\mathcal{L}$
400	6	-3.8209	7	-4.3096	5	-4.7419
550	5	-5.2903	7	-4.9811	6	-5.5177
700	8	-10.620	6	-8.6071	5	-9.5910
850	7	-10.0603	7	-9.5175	6	-9.5272

Chromosome 10						
Resolution	**Upsampled**		**Downsampled**		**Combined**	
	J	$\mathcal{L}$	J	$\mathcal{L}$	J	$\mathcal{L}$
400	8	-3.2834	6	-4.3964	6	-4.3978
550	8	-2.7821	7	-6.3721	7	-6.2640
700	8	-8.8074	8	-8.2315	6	-9.0516
850	8	-13.3051	8	-11.1141	6	-11.0203

Chromosome 11						
Resolution	**Upsampled**		**Downsampled**		**Combined**	
	J	$\mathcal{L}$	J	$\mathcal{L}$	J	$\mathcal{L}$
400	4	-3.6813	6	-4.7850	6	-4.7648
550	8	-3.9603	8	-7.5050	6	-8.1162
700	4	-15.6520	7	-11.1570	6	-11.4460
850	4	-13.3610	4	-11.2810	9	-11.4400

Chromosome 12						
Resolution	**Upsampled**		**Downsampled**		**Combined**	
	J	$\mathcal{L}$	J	$\mathcal{L}$	J	$\mathcal{L}$
400	6	-4.1444	7	-4.6736	8	-4.6166
550	8	-5.0504	8	-6.2282	8	-6.3146
700	6	-13.9440	7	-10.961	9	-10.5580
850	5	-16.9440	5	-10.738	10	-10.7330

Chromosome 13						
Resolution	**Upsampled**		**Downsampled**		**Combined**	
	J	$\mathcal{L}$	J	$\mathcal{L}$	J	$\mathcal{L}$
400	5	-3.6969	6	-3.4472	7	-3.3926
550	5	-4.3378	8	-3.8797	9	-3.7812
700	9	-3.9688	7	-4.7221	8	-4.6245
850	6	-8.7558	6	-7.3815	6	-7.9231

Chromosome 14						
Resolution	**Upsampled**		**Downsampled**		**Combined**	
	J	$\mathcal{L}$	J	$\mathcal{L}$	J	$\mathcal{L}$
400	5	-3.7348	6	-3.4965	6	-3.4845
550	5	-4.3009	6	-3.7693	8	-3.6269
700	7	-4.6961	7	-4.3075	7	-4.3181
850	4	-7.6186	4	-6.0092	7	-5.6381

Chromosome 15						
Resolution	**Upsampled**		**Downsampled**		**Combined**	
	J	$\mathcal{L}$	J	$\mathcal{L}$	J	$\mathcal{L}$
400	6	-3.1646	5	-3.8836	8	-3.4694
550	4	-4.9117	9	-4.1355	7	-4.4979
700	5	-4.8904	9	-4.4926	7	-4.8368
850	8	-4.3434	8	-6.3936	8	-5.5853

Chromosome 16						
Resolution	**Upsampled**		**Downsampled**		**Combined**	
	J	$\mathcal{L}$	J	$\mathcal{L}$	J	$\mathcal{L}$
400	5	-3.5699	4	-3.7158	6	-3.3564
550	7	-3.1739	6	-3.3044	6	-3.3510
700	5	-4.7573	6	-4.0875	6	-4.1563
850	11	-3.3864	11	-4.5890	6	-4.6062

Chromosome 17						
Resolution	**Upsampled**		**Downsampled**		**Combined**	
	J	$\mathcal{L}$	J	$\mathcal{L}$	J	$\mathcal{L}$
400	6	-3.3910	7	-3.2701	6	-3.4884
550	8	-3.2570	7	-3.4897	7	-3.5734
700	6	-4.4526	6	-4.7788	8	-4.4173
850	8	-4.3136	8	-4.5374	6	-5.2015

Chromosome 18						
Resolution	**Upsampled**		**Downsampled**		**Combined**	
	J	$\mathcal{L}$	J	$\mathcal{L}$	J	$\mathcal{L}$
400	4	-3.4651	7	-2.5073	6	-2.7273
550	5	-3.5862	7	-3.0942	6	-3.2695
700	6	-3.5142	6	-3.2159	7	-3.2043
850	9	-3.6741	9	-3.8413	7	-3.7792

Chromosome 19						
Resolution	**Upsampled**		**Downsampled**		**Combined**	
	J	$\mathcal{L}$	J	$\mathcal{L}$	J	$\mathcal{L}$
400	3	-3.7135	6	-2.4045	6	-2.5358
550	5	-2.8951	6	-2.4154	7	-2.4565
700	4	-5.1835	6	-2.9831	7	-3.0044
850	4	-4.0183	4	-3.0227	5	-3.3670

Chromosome 20						
Resolution	**Upsampled**		**Downsampled**		**Combined**	
	J	$\mathcal{L}$	J	$\mathcal{L}$	J	$\mathcal{L}$
400	5	-2.9799	6	-2.4055	6	-2.5126
550	6	-2.9034	6	-2.2939	6	-2.5126
700	5	-3.6122	7	-2.6078	6	-2.9454
850	5	-4.0085	5	-3.1918	7	-3.1907

Chromosome 21						
Resolution	**Upsampled**		**Downsampled**		**Combined**	
	J	$\mathcal{L}$	J	$\mathcal{L}$	J	$\mathcal{L}$
400	4	-1.6774	5	-1.5461	5	-1.6076
550	6	-1.6447	4	-1.8109	5	-1.7402
700	5	-2.1391	5	-1.9207	7	-1.7489
850	5	-2.1391	5	-1.8478	6	-2.0233

Chromosome 22						
Resolution	**Upsampled**		**Downsampled**		**Combined**	
	J	$\mathcal{L}$	J	$\mathcal{L}$	J	$\mathcal{L}$
400	7	-2.1946	4	-1.7142	3	-1.9641
550	7	-2.2315	5	-2.1204	4	-2.3481
700	7	-2.7416	6	-2.4106	5	-2.5080
850	8	-2.4068	8	-2.4156	7	-2.3680

Appendix C

Datasets

> *Although we often hear that data speak for themselves, their voices can be soft and sly.*
>
> — F. Mosteller, S. Fienberg, R. Rourke
>
> *from Beginning Statistics with Data Analysis*

Synopsis

This chapter presents the visualization of the dataset used in this thesis. There were two different chromosomal aberrations dataset in two different resolutions: 400 and 850. Here, the chromosomal aberrations in resolution 400 is depicted for the whole genome while the chromosomal aberrations for dataset in resolution 850 is omitted because of large dimension of dataset. The chapter also tabulates variations in the number of chromosome bands (regions) across different resolutions.

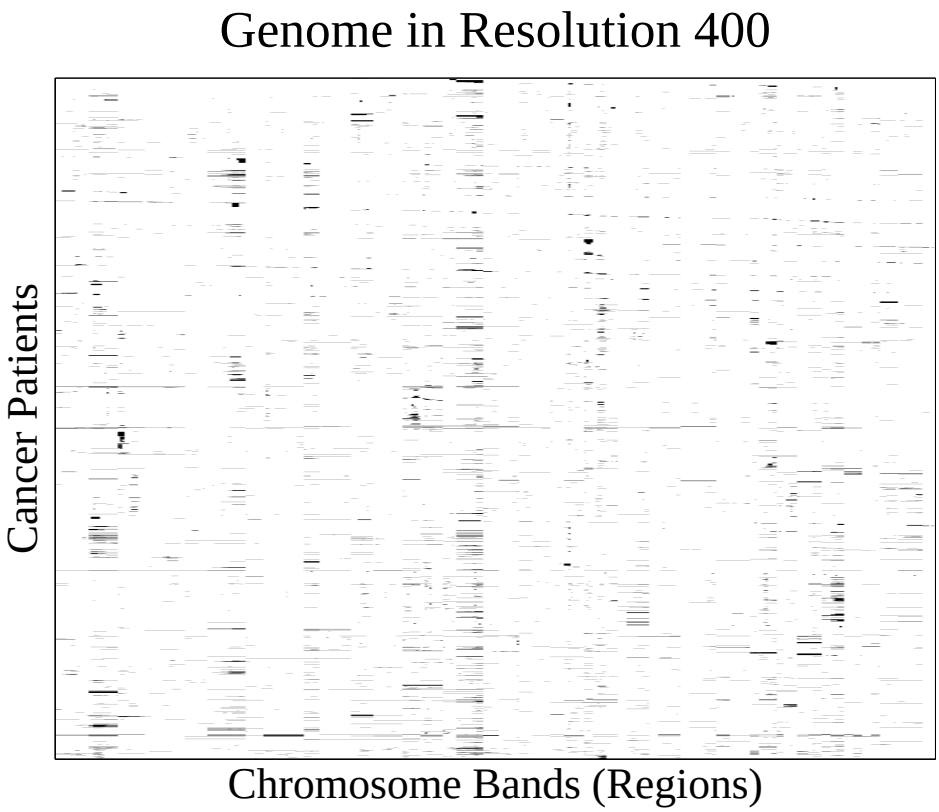

Figure C.1: Genome in Resolution 400. X-axis are spatial coordinates of the chromosome regions. In resolution 400, there are 393 different regions. Y-axis are the cancer patients numbering 4590. Each row represents one sample of the aberrations pattern for a cancer patient and each column represents one of the chromosome bands (regions). $\overline{X} = (X_{ij})$, $X_{ij} \in \{0, 1\}$. In figure dark color denotes the presence of aberrations and the white color denotes the absence of chromosomal aberrations. The data is very sparse and skewed. For example, Elementwise AND operation over all the samples in the data results in a zero vector.

Chromosome	Resolution			
	400	**550**	**700**	**850**
1	28	42	61	63
2	30	40	50	62
3	27	36	50	62
4	26	30	45	47
5	21	33	43	45
6	23	33	44	48
7	18	26	34	44
8	18	26	40	40
9	16	22	39	43
10	14	28	34	42
11	15	30	34	36
12	15	26	39	41
13	14	20	24	36
14	14	18	24	32
15	16	22	24	32
16	15	15	21	25
17	12	14	22	24
18	9	14	16	20
19	11	11	19	19
20	10	10	18	20
21	8	10	12	14
22	8	12	16	16
X	19	28	38	40
Y	6	10	11	11

Table C.1: Variation of number of chromosome bands in each chromosome in four different resolutions. Table captures the differences in the number of chromosome bands across resolutions. Table also shows that some of the chromosomes in two different resolutions have the same number of chromosome bands. For example, chromosome 19 has 11 bands in resolution 400 & 550 and 19 bands in resolution 700 & 850.