Genomic Signal Processing Methods for Detection of Copy Number Variation from Array CGH Data and Phylogenetic Classification using Protein Sequences

Thesis submitted to
Cochin University of Science and Technology
in partial fulfillment of the requirements for the award of the degree of

Doctor of Philosophy

in the Faculty of Technology

By
Anu Sabarish R

Under the guidance of
Dr. Tessamma Thomas

DEPARTMENT OF ELECTRONICS
COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY
KOCHI- KERALA, INDIA 682022
September 2017

Genomic Signal Processing Methods for Detection of Copy Number Variation from Array CGH Data and Phylogenetic Classification using Protein Sequences

Ph.D Thesis under the Faculty of Technology

Author

Anu Sabarish R.
Research scholar
Department of Electronics
Cochin University of Science and Technology
Kochi-682022
Kerala, India
email: anusabarish@cusat.ac.in

Supervising guide

Dr. Tessamma Thomas
Emeritus scientist
Department of electronics
Cochin University of Science and Technology
Kochi-682022
Kerala, India
email: tess@cusat.ac.in

Department of Electronics
Cochin University of Science and Technology
Kochi- 682022

September 2017

DEPARTMENT OF ELECTRONICS
COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY
KOCHI-22

Certificate

This is to certify that this thesis entitled **Genomic Signal Processing Methods for Detection of Copy Number Variation from Array CGH Data and Phylogenetic Classification using Protein Sequences** is a bonafide record of the research work carried out by **Mr. Anu Sabarish R** under my supervision in the Department of Electronics**,** Cochin University of Science and Technology. The results presented in this thesis or part of it has not been presented for the award of any other degree. It is also certified that all the relevant corrections and modifications suggested by the audience during the pre-synopsis seminar and recommended by the Doctoral Committee of the candidate has been incorporated in this thesis.

Kochi-22
18-09-2017

Dr.Tessamma Thomas
(Supervising Guide)
Department of Electronics
CUSAT
Kochi- 682022

DECLARATION

I hereby declare that this thesis entitled **Genomic Signal Processing Methods for Detection of Copy Number Variation from Array CGH Data and Phylogenetic Classification using Protein Sequences** is based on the original research work carried out by me under the supervision of **Dr.Tessamma Thomas** in the Department of Electronics**,** Cochin University of Science and Technology. The results presented in this thesis or part of it has not been presented for the award of any other degree.

Kochi-22
18-09-2017

Anu Sabarish R
Research Scholar
Department of Electronics
CUSAT
Kochi- 682022

Acknowledgement

It is my pleasure and privilege to thank the many individuals who made this thesis possible. I thank God Almighty for his immense blessings all throughout my journey. I express my deepest sense of gratitude to my research guide, Dr.Tessamma Thomas, who has guided me through this work with valuable advice, immense patience and support throughout.

I thank from the bottom of my heart Dr. Supriya M. H, Head of the Department of Electronics, Cochin University of Science and Technology for the support and goodwill extended to me and for extending the facilities in the department for my research work.

I would like to place on record my sincere gratitude to Professors, Dr. K. Vasudevan, Dr. P. R. S Pillai, Dr. P. Mohanan, Dr. C. K. Aanandan, Dr. James Kurian, for their valuable advice and continuous encouragement. I'm also grateful to the teaching faculty of Department of Electronics, Dr. Bijoy Antony Jose, Arun A. Balakrishnan, Mithun Haridas T. P for their support. I thank all other teaching and non-teaching staff and technical staff of the Department of Electronics and the staff of the Administrative Section of CUSAT for their support and cooperation.

I also acknowledge the help rendered by my co researchers Dr.Deepa Sankar, Dr. Reji A. P, Dr. Praveen. N, Dr. Anatharesmi, Dr. Sethunadh, Dr. Deepa J, Dr. Nobert Thomas, Tina P.G, Suja S, Deepthi, Sangeetha R and Binthiya of Audio and Image Research Lab, Department of Electronics, CUSAT. I take this opportunity to thank Paulbert Thomas, Lindo A.O, Dr. Ullas G.K. Anju P. Mathews, Dr. Sarin V.P, Dr. Dinesh R, Nijas M, Sreenath S, and Deepak at the Department of Electronics for all their support.

It is beyond words to express my gratitude to my family for their unflagging love and support throughout my life. I cannot ask for more from my parents, Ravindran G and Lalitha Kumari V. G, as they have been an epitome of unconditional love and encouragement towards my academic pursuits. I am indebted to my wife, Vandana B who provided me with indispensable support and

motivation. I'm thankful to my son, Harisankar S who has been a great source of calmness and happiness during tough times. I am also grateful to all my dear family members and friends for their patience and encouragement during the period of research.

I also thank all my colleagues in BSNL who have extended their cooperation and support for my work. Finally, I would like to thank everybody who was important to the successful realization of the thesis, as well as expressing my apology that I could not mention personally one by one.

Anu Sabarish R

Abstract

In recent years, the rapid development in the field of DNA sequencing techniques has necessitated the evolution of new computational methods for processing the genomic data. In this context, signal processing techniques have a significant role in developing new methods for genomic study. In this work, the variations in genomic sequences are studied using signal processing methods and novel methods are developed for the detection and localization of copy number variations using Array CGH data and phylogenetic classification of organisms using protein sequence.

A novel method, named Edge Enhancement and Segmentation (EES), is developed for the detection and localization of Copy Number Variation (CNV) in the genome, using Array CGH fluorescence intensity data. CNVs are instances where a cell has an abnormal number of copies of certain sections of the genome due to loss or gain of DNA. The medical significance of CNV is widely recognized to be associated with a number of disorders like Alzheimer's, Autism and Schizophrenia. The EES method performs an edge enhancement filtering prior to the segmentation of log ratio data into regions of discrete copy number levels. These regions are then categorized as normal or aberrant regions using thresholding method. Performance of the EES method is studied using simulated data and real Array CGH data where it clearly illustrated its superior performance in detecting the regions involving CNVs, compared to other established methods.

An alignment free method for phylogenetic classification using a frequency domain approach is also developed. Phylogenetics is the study of evolutionary relationship among organisms. Sequence alignment methods have been the most important part of the phylogenetic analysis methods. As the size and amount of sequences increases, the computational time and complexity for those methods become a challenge. The newly developed method, named Single Protein Power Spectral Density (SPPSD) method, infers the phylogenetic relationship among organisms using the distance between power spectral densities of the numerical representation of the amino acid sequence of a protein obtained from different

organisms. The study is then extended to develop a method, named Consensus Phylogeny using Principal Component Analysis (CPPCA), where phylogenetic relationship between organisms is inferred using sequences of multiple proteins. The capability of the new methods in capturing the underlying pattern of relationship between organisms is demonstrated using different sets of organisms, the results of which showed better conformance to the taxonomical classification of organisms.

Contents

Abbreviations

Array CGH	: Array Comparative Genomic Hybridization
AUC	: Area Under the Curve
BAC	: Bacterial Artificial Chromosome
BCCL	: Breast Cancer Cell Line
cDNA	: complementary DNA
CGH	: Comparative Genomic Hybridization
CNV	: Copy Number Variation
CPPCA	: Consensus Phylogeny using PCA
DGGE	: Denaturing Gradient Gel Electrophoresis
DHPLC	: Denaturing High Performance Liquid Chromatography
DNA	: Deoxyribonucleic acid
DFT	: Discrete Fourier Transform
DSP	: Digital Signal Processing
DWT	: Discrete Wavelet Transform
EES	: Edge Enhancement and Segmentation
EIIP	: Electron Ion Interaction Potential
FFP	: Feature Frequency Profile
FFT	: Fast Fourier Transform
FISH	: Fluorescent In-Situ Hybridization
FoSTes	: Fork Stalling and Template switching
FPR	: False Positive Rate
GBM	: Glioblastoma Multiforme
GSP	: Genomic Signal Processing
HBA	: Alpha hemoglobin
HBB	: Beta hemoglobin
HIV	: Human Immunodeficiency Virus
HMM	: Hidden Markov Model
IC	: Information Correlation
LCR	: Low Copy Repeats
LTI	: Linear Time Invariant

MCMC : Markov Chain Monte Carlo
mRNA : messenger RNA
miRNA : micro RNA
MMEJ : Micro homology Mediated End Joining
MRe : Majority Rule extended
MSA : Multiple Sequence Alignment
MSC : Multi-Species Coalescent
MVF : Minimum Variance Filter
NAHR : Non Allelic Homologous Recombination
NCBI : National Center for Biotechnology Information
NGS : Next Generation Sequencing
NHEJ : Non Homologous End Joining
NJ : Neighbor Joining
OUT : Operational Taxonomic Unit
PAC : P1 bacteriophage based Artificial Chromosome
PC : Principal Component
PCA : Principal Component Analysis
PCR : Polymerase Chain Reaction
PIC : Partial Information Correlation
PMD : Pelizaeus-Merzbacher Disease
PSD : Power Spectral Density
RMSE : Root Mean Square Error
RNA : Ribonucleic acid
ROC : Receiver Operating Characteristics
rRNA : ribosomal RNA
SNP : Single Nucleotide Polymorphisms
SNR : Signal to Noise Ratio
snRNA : small nuclear RNA
snoRNA : small nucleolar RNA
SPPSD : Single Protein Power Spectral Density
SSCA : Single Strand Conformation Analysis
SSE : Sum of Squared Error
STR : Short Tandem Repeats

SVD	: Singular Value Decomposition
SWT	: Stationary Wavelet Transform
TPR	: True Positive Rate
tRNA	: transfer RNA
UPGMA	: Unweighted Pair Group Method with Arithmetic average
VNTR	: Variable Number Tandem Repeats
YAC	: Yeast Artificial Chromosome

List of Figures

List of Tables

Chapter 1

Introduction

This chapter provides a brief overview of Genomic Signal Processing discipline. Some of the basic concepts in molecular biology related to the present work like DNA, RNA, chromosome, gene, protein and genomic sequence variation are introduced. An insight into the objectives of this study is also provided. The chapter concludes with a summary of contributions of the thesis and its organization.

1.1 Genomic Signal Processing

Genomic Signal Processing (GSP) is a highly interdisciplinary area of study related to the extraction, processing and analysis of genomic data. Genomic data can be images such as Fluorescent In-Situ Hybridization (FISH) images, microarray images and sequences such as expression data, DNA (deoxyribonucleic acid) sequence, protein sequence etc. The genome analysis aims to unveil the complete DNA sequence of an organism and to annotate and analyze the important features in them. The human genome project and its successful completion have spurred the field of DNA sequencing technology resulting in its rapid growth. The latest techniques like Next Generation Sequencing (NGS) methods have been able to perform the DNA sequencing very quickly. The vast amount of genomic data generated using these methods have resulted in the emergence of a new area called computational genomics or computer assisted genomic information processing. Digital Signal Processing (DSP) techniques such as clustering, classification, pattern recognition etc. have already been used with significant success in the field of genomic signal processing. And this success has opened up a unique and challenging opportunity for the digital signal processing researchers to play a significant role in genomic analysis. A brief introduction to some of the basic concepts in molecular biology related to this study is discussed here.

1.2 Basic concepts of molecular biology

Genomes can be understood as the biological information needed to create and sustain a living organism. This information is stored in a long sequence of nucleic acid which determines the hereditary nature of the organism. In most organisms, including the human beings, the genome or

the genetic material is made up of DNA. In certain viruses, RNA (ribonucleic acid) constitutes the genome. With the help of complex interactions these sequences have the capacity to build new structures and to perform necessary functions at appropriate place and time to maintain the life of a living organism. The following part explains the three basic molecules involved in genomes and its expression.

1.2.1 Deoxyribonucleic acid (DNA)

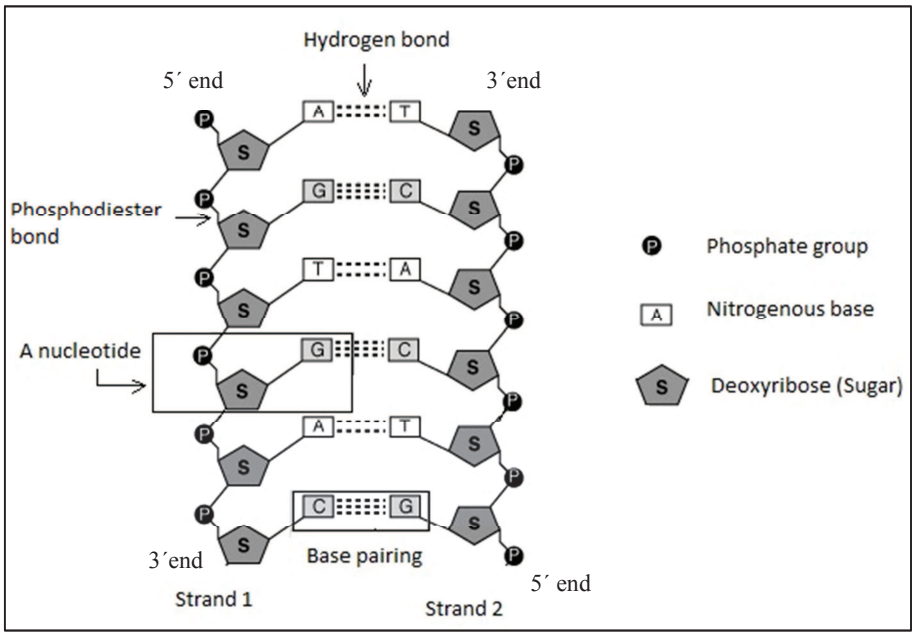


Figure 1.1 Base pairing between the nucleotides on complementary strands of DNA polynucleotide.

The DNA molecule is a double helix structure constituted by two polynucleotide chains. The two chains or strands run antiparallel. Each of this polynucleotide chain is a linear polymer made up of many linked monomer units called nucleotides. Each nucleotide comprises of three

components: deoxyribose (sugar), a nitrogenous base and a phosphate group. There are four types of nucleotides based on the type of nitrogenous base attached to the sugar molecule. The four bases are adenine (A), guanine (G), cytosine (C), and thymine (T). The nucleotides are linked together by phosphodiester bonds between the phosphate group and sugar molecule of nearby nucleotides. The two strands of the helix are kept together by the base pairing between the bases in opposite strands by hydrogen bonds (Fig 1.1).

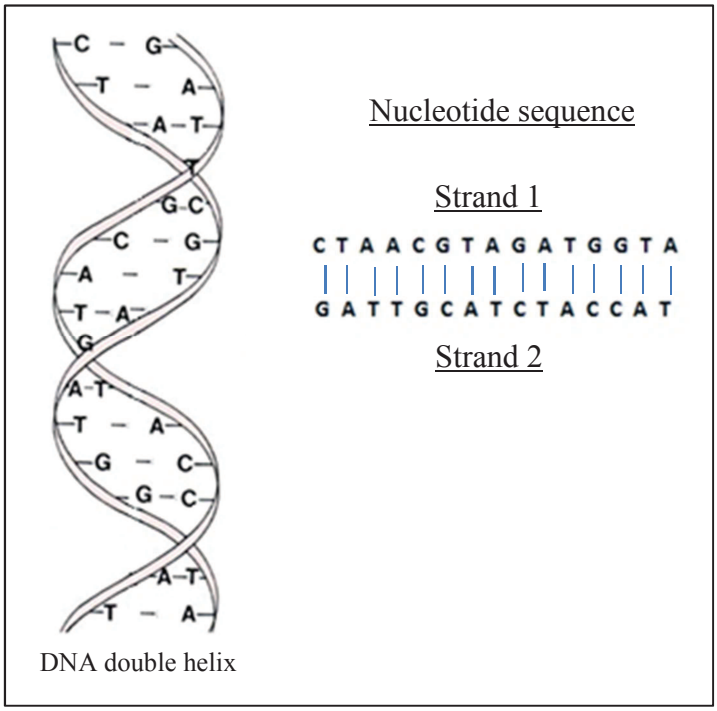


Figure 1.2 DNA double helix and complementarity of sequences on the two strands

DNA bases pair up in a unique way, with adenine forming a pair only with thymine (A-T) and cytosine pairing only with guanine (C-G). The base sequence of the two strands are complementary to each other as the

nucleotide of one strand has to link to the nucleotide of the other as shown in Fig 1.2. This property also allows each strand of DNA to act as a template to create exact copies of parent DNA polynucleotide in the daughter cells by the process of DNA replication.

1.2.2 Ribonucleic acid (RNA)

RNA is also a polynucleotide made up of nucleotide subunits. The RNA differs from the DNA by the fact that the sugar molecule in the RNA is ribose and instead of thymine, RNA contains uracil (U). The base pairing rule also changes accordingly where adenine (A) pairs with uracil (U) instead of thymine (T). RNA molecule is generally single stranded and is also shorter in length. RNAs can be classified into coding RNA and noncoding RNA. The coding RNA includes one type of molecule called messenger RNA or mRNA. They are obtained from protein coding genes in DNA by a process called transcription and have the capability of getting converted into protein sequence by the process called translation. Noncoding RNA is also known as functional RNA as they perform many important roles in the functioning of a cell. Noncoding RNA includes different types of RNA molecules such as, ribosomal RNA (rRNA), transfer RNA (tRNA), micro RNA (miRNA), small nuclear RNA (snRNA) and small nucleolar RNA (snoRNA) (Brown, 2007).

1.2.3 Chromosome

The DNA polynucleotide chain of an organism that contains several billions of nucleotides are found within each cells in short, coiled and condensed form. The entire DNA content is split into a set of molecules known as a chromosome. Each organism has its own unique number of chromosomes.

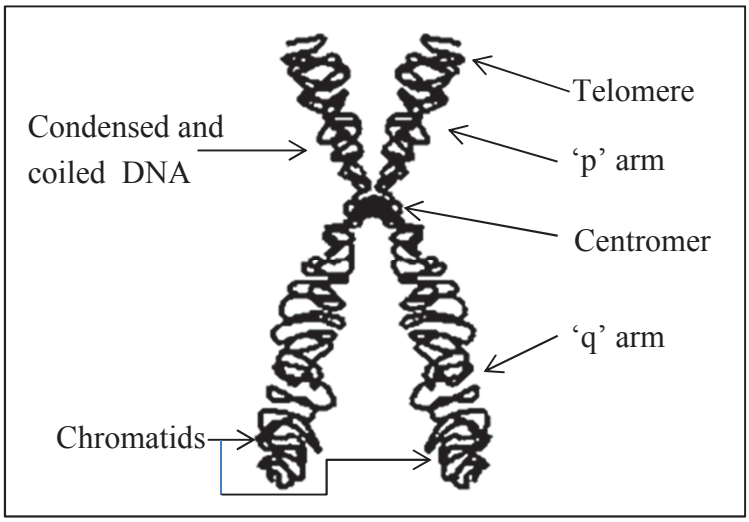


Figure 1.3 Chromosome structure.

A eukaryotic chromosome consists of two chromatids, each having a short 'p' arm and a long 'q' arm, joined together at the centromere. The terminal region of each chromatid is termed telomere. The structure of a eukaryotic chromosome is shown in Fig 1.3.

1.2.4 Gene

The DNA sequence can be divided into two types of regions, gene region and intergenic region. Each gene contains a set of information for the generation of a specific protein with an associated function. Though the entire set of genes are present in all the cells, only a selected number of genes are expressed in each cell based on the cell function and the remaining ones are turned off. The gene region can again be divided into two sub regions called introns and exons**.** During the process of protein synthesis, the intron regions are removed and the exons are joined together to form an mRNA sequence which in turn gets translated into appropriate protein.

1.2.5 Protein

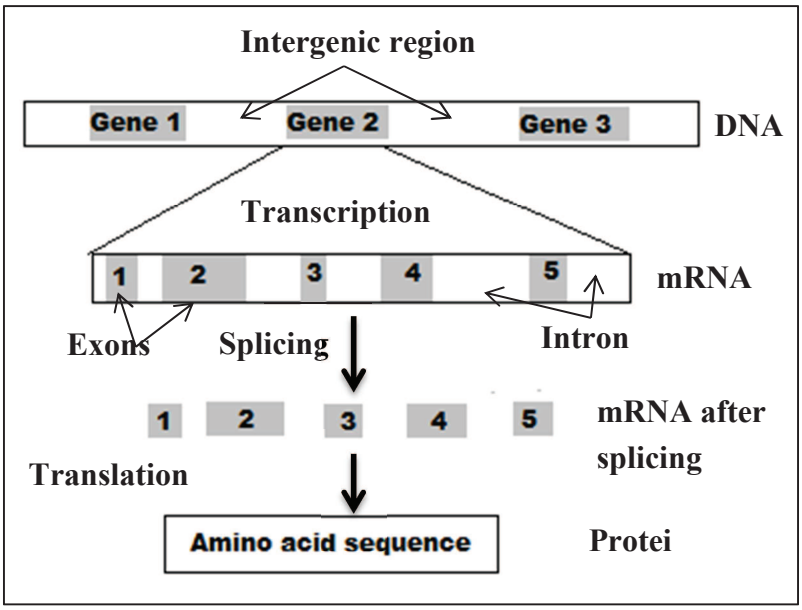


Figure 1.4 Central dogma of molecular biology

A protein is also a polymer made up of monomeric subunits called amino acids. The mRNA after splicing contains a sequence of nucleotides which again can be divided into groups of three adjacent bases. This grouping of 3 adjacent bases, also known as a triplet, comprises a codon. Various different combinations of bases can result in a total of 64 codons which comprises the genetic code. Each codon encodes one of the 20 amino acid in a many to one mapping. In a cell, each codon from an expressed gene instructs the cell mechanism to synthesize an amino acid. Every protein sequence starts with a unique start codon, ATG (methionine), which indicates the beginning of the protein coding part of a gene. The protein synthesis continues until a stop codon is reached which indicates the end of protein coding region. There are three codons namely, TAG, TAA and TGA

that can act as a stop codon. These amino acids join together and represent the protein corresponding to the original gene. This flow of genetic information within a cell from DNA to mRNA and from mRNA to protein is known as the central dogma of molecular biology (Fig 1.4). The information stored in DNA in the form of nucleotide sequence determines the sequence of mRNA molecules, which in turn specifies the protein that gets synthesized for various cellular functions. The proteins perform a wide range of functions determined by the sequence of amino acid constituting it.

1.3 Genomic sequence variation

A complete genome sequence of a species is a representative sequence based on the DNA collected from a few samples of a single species. Most of the sequence is same from one individual to the next. Genetic sequence variations are differences in the DNA sequences from one individual to another. The differences observed are relatively small and are resulted from the interaction of various different forces involved in evolutionary process. This involves processes such as mutation, selection, recombination and also various demographic factors. So the genome can be considered as a dynamic entity that changes over time as a result of the cumulative effects of various processes (Brown, 2007). Variations can be found throughout the genome, on every one of the chromosomes in an individual. Certain regions of the genome exhibit large number of variations and are termed as hot spots. The parts of the genome that exhibits less variation between individuals are termed as stable. The majority of variations is found outside the gene regions and does not affect an individual's characteristics. They are generally harmless and have a chance to accumulate without causing any problems. Gene regions, meanwhile,

tend to be stable and alterations in those regions are often harmful. As a result, the chances for the alterations to be passed on to the next generations are lesser and thereby to accumulate the effect.

Genome variations can be classified into different types according to the size and type of variation. The classification based on size is small scale, large scale and whole chromosome variations. Small scale variations include Variable Number Tandem Repeats (VNTR) and Single Nucleotide Polymorphisms (SNP). VNTR are regions of genome with a segment of DNA sequence, represented as a block, which repeats itself in tandem. Individual alleles of the genome will have varying number of repeats and can be uniquely identified by its number of repetition. Based on the length of the segment or block it is classified into two groups, microsatellites and minisatellites. Microsatellites are repeats of sequence with length less than 5 base pairs. It is also known as Short Tandem Repeats (STR) and Simple Sequence Repeats (SSR). Minisatellites involve longer blocks, usually with a sequence length of 10-20 base pairs. An SNP is a sequence variation caused by change in a single nucleotide position observed with a frequency greater than 1%. Copy Number Variations (CNV) are large scale variations involving regions of sequences with several kilo base pairs in length. Duplication and deletions of large lengths of sequences results in abnormal number of copies of one or more genes in an individual which is termed as a copy number variation. Whole chromosome variations result in conditions called aneuploidy and polyploidy. Aneuploidy is a condition where a cell exhibits an increase or decrease in the total number of chromosomes. Aneuploidy can be categorized as monosomy, trisomy, tetrasomy etc., depending on the number of copies of chromosomes. Polyploidy involves a change of one or more complete sets of chromosomes. In the case of a

diploid organism the normal number of chromosome can be denoted as 2X. Monosomy, trisomy and tetrasomy refers to the presence of 2X-1, 2X+1 and 2X+2 number of chromosomes respectively. And polyploidy can be denoted by a chromosome number of 2X multiplied by n, where n denotes the number of sets of chromosomes.

DNA sequence variations are also described as mutations and as polymorphisms. Polymorphism or SNP is defined as DNA variants detectable in more than 1 % of the population. A mutation can be defined as a change in the DNA sequence whose frequency is lesser than 1 % (Karki et al., 2015). Polymorphic sequence variants usually do not cause diseases. They are usually found outside of genes and are neutral in effect. Those found within gene coding region may influence characteristics such as height and colour which possess lesser medical importance. Meanwhile polymorphism may have impact on disease susceptibility and drug response.

The effects of these sequence variations are important and vary widely (Darwin, 1859). Most of the variations are harmless, but certain DNA variations affect the phenotypic traits and can be harmful. The most important phenotypic trait is fitness, which indicates, the ability to survive and reproduce. The effects of the variants can be classified into 4 classes. They can be lethal mutations, deleterious mutations, neutral mutations and advantageous mutations. Lethal mutations directly results in the loss of capability to survive. Deleterious reduces the overall fitness level compared to normal phenotype. Neutral ones do not affect the fitness and can be considered as normal phenotype. And finally the advantageous ones increase the fitness and thereby enhance the chances of natural selection. The sequence variant that is neutral or beneficial in effect will be preserved while the harmful ones are discarded. DNA sequence variation can cause

various diseases which are broadly categorized as mendelian disease and complex disease. Mendelian diseases are caused by variations in a single gene and are also termed as monogenic disease. Complex diseases on the other hand are caused from interactions involving multiple genes and hence are termed polygenic.

Studies on genome variation have proved helpful in a wide range of applications like screening for genetic diseases, DNA fingerprinting, evolutionary studies and genome mapping. Methods for identifying sequence variation can be broadly classified as screening methods and diagnostic methods. The methods include DNA sequencing, Denaturing High Performance Liquid Chromatography (DHPLC), Single Strand Conformation Analysis (SSCA), Denaturing Gradient Gel electrophoresis (DGGE), allele specific hybridization, allele specific PCR, restriction endonuclease digestion, southern blot, fluorescent in-situ hybridization etc.

1.4 Application of signal and image processing methods in genomics

Signal processing approach deals with techniques for analysis, recovery, representation, transformation, extraction, modeling and understanding various types of signals. It deals with a wide range of signals, from audio to image, video, biological, seismic and many others. The application of signal/image processing techniques for the processing and analysis of genomic data is discussed in this section. The two key areas that have received dominant attention are genomic sequence analysis and micro array analysis. The sequence analysis involves the processing of DNA, RNA or protein sequence information for the detection and classification of structures or relevant regions. Microarray analysis meanwhile enabled the

genome wide analysis, where tens of thousands of gene can be monitored simultaneously, to extract and compare gene expression information using image data.

Application of the signal processing approaches to analyze genomic information is greatly helped by the capability of representing the genomic data in digital form. The genomic sequences like DNA, RNA and protein are represented as character strings where each character is one out of a finite set of alphabets. In the case of a DNA sequence, the finite set is {A,T,C,G} and in the case of RNA it is {A,U,C,G}. In the case of protein, the alphabet is of size 20 which is the set of all 20 amino acids. By assigning appropriate numerical values to these letters, the alphabetical sequence can be represented as a numerical sequence facilitating the application of signal processing techniques for analysis. An overview of such applications for genomic analysis is described here. DNA sequence comparison is one of the basic area of study to understand the structure and function of sequences. In (Veljkovic et al., 1985) DFT analysis of protein sequences were employed to identify characteristic peak frequencies for functionally related sequences. A similar attempt was made in (Trad et al., 2002) to identify the characteristic frequency band of protein using wavelet transforms. Ramachandran and Antoniou (2008) went onto to locate the hot spots which are critical in determining the functionality of a protein using digital filters. Fourier analysis has been used successfully for identifying specific structures in the DNA sequence. Protein coding DNA regions were located using fourier transform analysis by Anastassiou (2001). Vaidyanathan and Yoon later extended the exon detection method using digital filters (2002). Various other computational techniques such as artificial neural networks (Ma et al., 2001; Hatzigeorgiou et al., 1996) and wavelet transforms (Trad et

al., 2002; Zhao et al., 2001) have already been used to explore the genomic sequences.

The increased use of microarrays and their unprecedented advantages made the role of image processing techniques critical for consistent and reliable extraction of information. The main image processing task involved in the microarray image processing can be summarized as, identification of the spots, their boundaries and to measure the fluorescence intensity value of the spots. Various image processing techniques have been applied for the gridding, segmentation and quantification processes to perform the three tasks described above. Fully automatic gridding methods have been described in (Jain et al., 2002; Wang et al., 2003; Wang et al., 2005) for the identification of spots. Once the gridding is completed the next task is to estimate the boundary of the spot to distinguish the spot from its background. The tool ScanAnalyze (Eisen, 1999) performs this segmentation using a simple concept of placing a mask of fixed radius over the spot location. Later various other algorithms were introduced to segment spots with varying radius and irregular shapes. Seed region growing method was introduced by Yang et al. (2002) for microarray segmentation to good effect. Various other approaches for segmentation has been attempted in (Chen et al., 1997; Demirkaya et al., 2005; Rahnenführer and Bozinov, 2004), each possessing its own merits. A number of signal and image processing techniques have been successfully applied to genomic data analysis. Yet, much advancements in the field of signal processing is needed, by introducing novel techniques to perform accurate, quick and fully automatic analysis of sequences to fulfill the expectations of this interdisciplinary area of research.

1.5 Objectives

The main objective of this work is to develop novel signal processing algorithms to study genomic sequence variations. The work is divided into two parts as given below.

i) Development of a genomic signal processing method for the detection and localization of copy number variations using Array CGH data.
ii) Development of a genomic signal processing method for phylogenetic classification of organisms using protein sequence.

1.6 Summary of contributions

The primary aim of CNV detection methods is to process the Array CGH data to track the variations in the data identify significant deviations and locate those critical regions of alterations. The task of converting the raw log ratio of fluorescence intensities of spots on Array CGH slide into discrete copy number values involves different stages like denoising, segmentation and copy number value assignment. . Most of the existing methods give emphasis to either the denoising part or the segmentation part. The segmentation based approaches suffer in the presence of high noise levels and the smoothing based method fails to provide accurate classification of regions into normal and aberrant ones. The focus of this work is to develop a novel method combining the smoothing step and segmentation step followed by a status assignment step for the detection and localization of copy number aberration. To take care of this problem a novel EES (Edge Enhancement and Segmentation) method is developed.

Determination of evolutionary divergence between species is the main step to understand the pattern of evolution of organisms. The molecular phylogenetic analysis methods determine the evolutionary divergence between the various species from a measure of genomic sequence similarity. In the second part of this study, a frequency domain approach for protein sequence similarity analysis is developed. Based on the analysis, a Single Protein Power Spectral Density method (SPPSD) is developed to infer the phylogenetic relationship between organisms using amino acid sequence of a single protein. With the aim of improving the consistency of inferred phylogenetic relationship, a new method called, Consensus Phylogeny using Principal Component Analysis (CPPCA), is developed to generate a phylogenetic tree by combining phylogeny information obtained from multiple proteins.

1.6.1 Edge Enhancement and Segmentation (EES) method

The EES method employs a step by step approach for CNV detection. The algorithm includes, local edge enhancement filtering, K-means clustering based segmentation and a threshold based decision making. The EES method performs the edge enhancement filtering operation using a local filtering algorithm named, Minimum Variance Filtering (MVF). It is based on the search for local homogeneity in small neighborhoods. The filter also implements an iterative filtering option with an automatic stopping criterion. In the second stage, EES method uses a K-means clustering approach to achieve the segmentation of the log intensity ratio data into discrete, non-overlapping segments corresponding to different copy number levels. The segmented levels are then categorized as normal or aberrant regions using a threshold based decision function.

1.6.2 Single Protein Power Spectral Density (SPPSD) method

A new signal processing based frequency domain approach for protein sequence similarity analysis is developed. The method, called SPPSD, uses an alignment free method to measure protein sequence similarity and to infer phylogenetic relationship between different species using a single protein. It employs a numerical mapping method using the Electron Ion Interaction Potential values of amino acids for converting a protein sequence into numerical form. The distance between Power Spectral Densities (PSD) of these numerical sequences are used as a measure of genetic distance for the construction of phylogenetic tree. This method infers the evolutionary relationship between organisms using sequences of a single protein.

1.6.3 Consensus Phylogeny using Principal Component Analysis (CPPCA) method

A new method, called CPPCA, is developed to infer phylogenetic relationship from sequences of multiple proteins. Amino acid sequences corresponding to different proteins are obtained from the organisms under study. Genetic distances are calculated using the SPPSD method for each of the protein. The CPPCA method then combines the genetic distances obtained for individual proteins using Principal Component Analysis (PCA) to create a consensus distance metric. Finally a phylogenetic classification of the organisms is achieved using the consensus distance.

1.7 Outline of the thesis

The thesis deals with the study of genomic sequence variation, with two different perspectives, cytogenetic and phylogenetic. The chapters are divided into two sections as described below,

The first section constituted by chapters from 2 to 5, explains the development of a molecular cytogenetic analysis technique to detect and locate regions affected by large scale structural variations that are observed in genomic sequence of organisms called Copy Number Variations (CNV). The section also illustrates the application of the method in identifying regions involved in tumor genesis using Array CGH data from real cancer cell lines.

The second section includes chapter 6 to chapter 8 and discusses the development of a frequency domain technique for phylogenetic classification of organisms. An alignment free method for inferring evolutionary relationship between organisms using amino acid sequence of a single protein is presented. This section also explains a new method for generating consensus phylogeny using multiple proteins.

Chapter 2: An overview of cytogenetic analysis techniques.

This chapter aims to provide a brief overview on various cytogenetic analysis techniques. The basic concepts of Chromosome banding, Fluorescent In-Situ Hybridization (FISH), Comparative Genomic Hybridization (CGH) and Array Comparative Genomic Hybridization (Array CGH) are explained. It is followed by an introduction to Copy Number Variation, mechanisms contributing to CNV, their significance and methods of detection.

Chapter 3: Array Comparative Genomic Hybridization.

A detailed description of Microarray and Array comparative genomic hybridization technologies is given in the chapter. The different types of microarray and the various steps involved in an Array CGH experiment are also explained. The data analysis step in a microarray experiment where the data generated through a series of steps is processed by statistical or computational means to draw meaningful results is illustrated. Major computational approaches used for the detection of CNV and a review of literature describing the applications of these methods for the CNV detection is also presented.

Chapter 4: Development of a new method for the detection of CNV from Array CGH data.

The development of a new method called EES (Edge Enhancement and Segmentation), for the detection and localization of copy number variations in the genome using log ratio data obtained from Array CGH experiment is explained here. The EES method performs an edge enhancement filtering prior to the segmentation of log ratio data into regions of discrete copy number levels. Performance of the EES method is verified with the help of simulated log ratio data and log ratio data obtained from real Array CGH experiment. A comparison of the results obtained using the EES method with other established methods are performed and the improvement in the performance of the EES method is also highlighted.

Chapter 5: Application of EES method on Array CGH data of human cell lines.

This chapter demonstrates the application of the EES algorithm for the analysis of real Array CGH data for the detection and localization of

copy number variations. EES method is applied on three real data sets to identify the alterations. It provides for an extensive and reliable evaluation of the method's performance in medical and clinical applications.

Chapter 6: Phylogenetic analysis.

A brief overview of evolutionary mechanisms and phylogenetic analysis is provided. It also describes tree representation, basic tree terminologies involved and tree construction methods. A review of literature describing various alignment free method of phylogenetic analysis is also discussed.

Chapter 7: Development of an alignment free method for phylogenetic classification using a single protein.

A frequency domain approach for protein sequence similarity analysis is described. Based on which, a Single Protein Power Spectral Density (SPPSD) method is developed to infer phylogenetic relationship between species using the amino acid sequence of a single protein. The application of SPPSD method on different set of samples for inferring phylogenetic relationship is also presented.

Chapter 8: Development of a consensus method for constructing phylogenetic tree using multiple protein sequences.

A review of literature describing various approaches of phylogenetic analysis using multiple proteins is provided. The development of a new method called Consensus Phylogeny using Principal Component Analysis (CPPCA) for combining information from multiple proteins for the generation of a consensus phylogenetic tree is discussed. Finally the CPPCA

method is applied on sample datasets for the generation of a consensus phylogenetic tree.

Chapter 9: Conclusion and future scope

In this concluding chapter, a summary of the research work along with important observations and results is presented. Directions for further study are also discussed.

Chapter 2

An overview of cytogenetic analysis techniques

An overview of various human cytogenetic analysis techniques is presented. The basic concepts of Chromosome banding, Fluorescent In-Situ Hybridization (FISH), Comparative Genomic Hybridization (CGH) and Array Comparative Genomic Hybridization (Array CGH) are discussed in this chapter. An introduction to Copy Number Variation (CNV), mechanisms contributing to CNV, their significance and methods for detection are also provided.

2.1 Introduction

Cytogenetics is a branch of study that developed from the union of two separate streams – cytology and genetics. It aims at understanding the hereditary mechanisms within a cell through various methods used in cytology and genetics. Hereditary determinants are found both within and outside the nucleus of a cell. Apart from chromosomes which are present within nucleus, certain extra nuclear constituents like mitochondria, plastids, plasmids present in the cytoplasm have demonstrated their inheritance capabilities (Schulz-Schaeffer, 1980). Since chromosomes contain most of the genetic material, the study of cytogenetics can be directly associated with the analysis of structure, function, number, and other variations of the chromosomes. It also includes the study of chromosomal variations associated with diseases and various techniques employed for the analysis of chromosomes and detection of abnormalities in them. The most important and challenging task faced at the initial stages of cytogenetic study was to determine the exact number of chromosomes in an individual and how to distinguish the individual chromosomes. The initiation of human cytogenetic study can be traced back to the discovery that normal human cells contain 46 chromosomes by Tjio and Levan (1956). From those initial stages, over the last few decades the field has witnessed introduction of several path breaking technologies that led to the current state of understanding. Chromosome banding, fluorescence in-situ hybridization, comparative genomic hybridization and single nucleotide polymorphism arrays are some of the major techniques that contributed to the growth of this field of research.

2.2 Historical perspective of human cytogenetic techniques

2.2.1 Chromosome banding

Chromosome banding techniques produce a series of banding patterns along the length of chromosomes. Chromosome banding is performed at the metaphase stage of the cell cycle when the chromatins are highly condensed and are best viewed. At this stage the 2 chromatids are joined at the centromere and ended by telomeres. The individual chromosomes can be identified by their uniqueness in banding pattern, size and location of centromere (Heim and Mitelman, 2015). Chromosome banding also allows identification of specific segments of chromosomes. Deletions, inversions, translocations and duplications within an individual chromosome can also be identified using the banding techniques. There are different types of banding methods which differ in the staining procedures employed and the regions of chromosomes highlighted by them.

Figure 2.1 The 46 chromosomes extracted from human embryonic lung fibroblast tissues by Tjio and Levan

Staining

Table 2.1 Types of staining

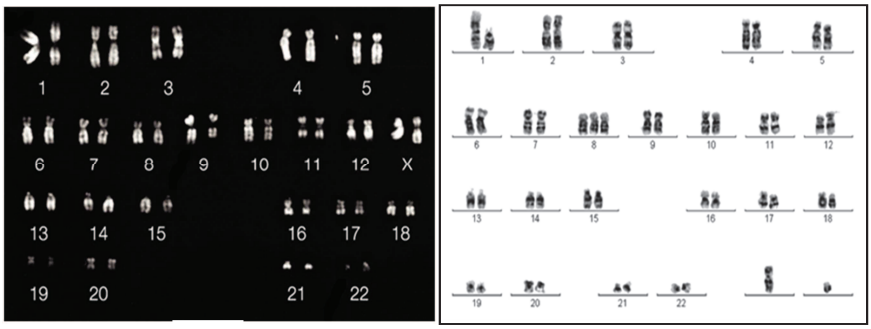
Stain	**Imaging technique**
Visible light dye	Light microscopy
Fluorescent dye (Fluorochrome)	Fluorescence microscopy
Heavy metals	Electron microscopy , Coherent X-ray diffraction

One of the most important step involved in banding schemes is the process of staining. Staining techniques are used to increase the contrast of metaphase chromosomes and create bands due to differential staining along the chromosomes. The stains bind to the chromosome due to covalent and non-covalent interactions with the DNA. Metal stains usually involve covalent bonding. Non covalent interactions include electrostatic interactions and hydrogen bonding. The different types of staining used are described in the (Table 2.1). The choice of staining is based on the imaging technique employed in the method.

Q-banding

Q-banding or Quinacrine banding (Caspersson, 1971) developed by Torbjörn Caspersson is the first among the various banding techniques. Light and dark horizontal bands are produced along the length of chromosomes by using a fluorochrome stain called quinacrine mustard which was later replaced by another dye quinacrine di-hydrochloride. The dark and bright bands (Q-bands) are produced as a result of differential fluorescence. The AT rich regions of the chromosomes appeared brighter

and the GC rich regions appeared as darker regions. The unique banding of chromosome permits the identification of all 22 pairs of autosomes and (X, Y) sex chromosomes. The technique also allowed identification of structural abnormalities and specific identification of the extra chromosomes. The trisomy of chromosome number 21 associated with Down's syndrome was identified using the technique.



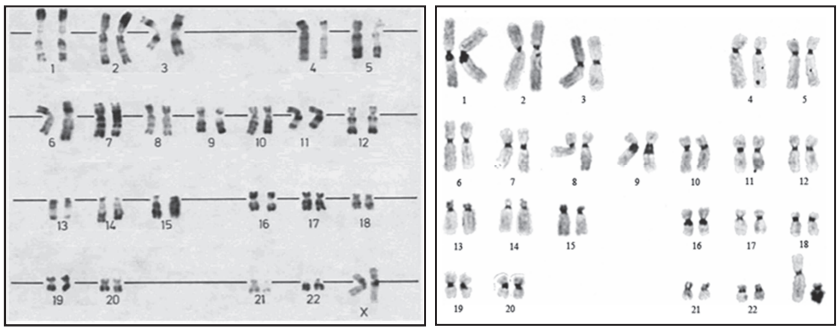
(a) (b)

Figure 2.2 a) Q-banding. b) G-banding

G-banding

G-banding was developed by Seabright (1971) and is the most widely used banding technique. G-banding uses a non-fluorescent, visible light dye called Giemsa, for staining. In this method before the staining, the chromosomes are subjected to a pretreatment process. This pretreatment process performs a differential digestion along the length of the chromosome which facilitates the formation of banding pattern at the Giemsa staining stage. The heterochromatin regions which are AT-rich regions produce dark bands. The darker regions which are called positive G-bands represent regions with higher hydrophobicity in the chromosome. The

lighter regions are called the negative G-bands and are usually GC-rich regions.



(a) (b)

Figure 2.3 a) R-banding. b) C-banding

R-banding

R-banding is also known as reverse banding. Giemsa stain is used to reveal the GC-rich regions of the chromosome. The AT-rich regions are denatured by heating in the pretreatment stage leaving the GC-rich regions unchanged. Thus darker bands are produced in GC-rich regions and generate a banding pattern opposite to that produced by G-banding. GC specific fluorochromes can also generate a similar banding pattern. They are mainly useful in analyzing telomere regions and can identify deletions and translocations in the telomere regions of chromosomes.

C-banding

C-banding is performed to stain the centromere part of chromosomes and other parts containing heterochromatin. Giesma stain is used for generating banding pattern after pretreatment of DNA. Regions containing heterochromatin resist degradation during pretreatment due to their relatively dense packing. Dark bands are generated at the AT-rich

centromere part of chromosome which is densely packed. It can be used to detect increases, decreases, inversions or rearrangements of heterochromatic regions.

2.2.2 Fluorescent In-Situ Hybridization (FISH)

The introduction of in-situ hybridization signaled the start of molecular level analysis in the field of cytogenetics. The procedure allowed researchers to identify the positions of specific DNA sequences on chromosomes. In-situ hybridization is a method that uses labeled DNA or RNA sequences to identify the presence or location of specific nucleic acid sequence on chromosomes. Fluorescent In-Situ Hybridization (FISH) was introduced by Langer-Safer et al. (1982) using fluorescent probes to detect specific DNA sequence in chromosomes. FISH technique is based on the ability of DNA strands to bind selectively to its complementary sequence and not to other sequences. Based on the intended target, a target specific probe is generated. In the next step both the target and probe are denatured, where the strands are separated. The probe sequence is labeled with fluorescent dyes and is mixed with the denatured target sequence. When the two complementary sequences in probe and target find each other, they bind together and hybridization occurs. The probe-target hybrid is washed and visualized with the help of a fluorescent microscope. The identified target can be observed as fluorescent spot or cluster. Detection of structural and numerical variations of chromosome like duplication and deletion are efficiently done with FISH. Deletion of sequence in the region complementary to the probe prevents hybridization which will be indicated by the absence of fluorescent emission. Meanwhile a duplication event increases the hybridization and is indicated by an increased fluorescence of the target sequence. Unlike many earlier methods, FISH do not need the

extraction of DNA as it works with intact chromosomes. The introduction of FISH greatly improved the resolution of aberration analysis to the sequence level. FISH technique can be used to detect chromosomal abnormalities involving mega base pairs of DNA in size. But it can only screen for a limited number of chromosomal abnormalities at one time. FISH has a wide range of application from gene mapping to detection of various chromosomal aberrations like gene deletion and duplication. It has emerged as the most commonly used and proven method in routine diagnostics for identifying genetic changes.

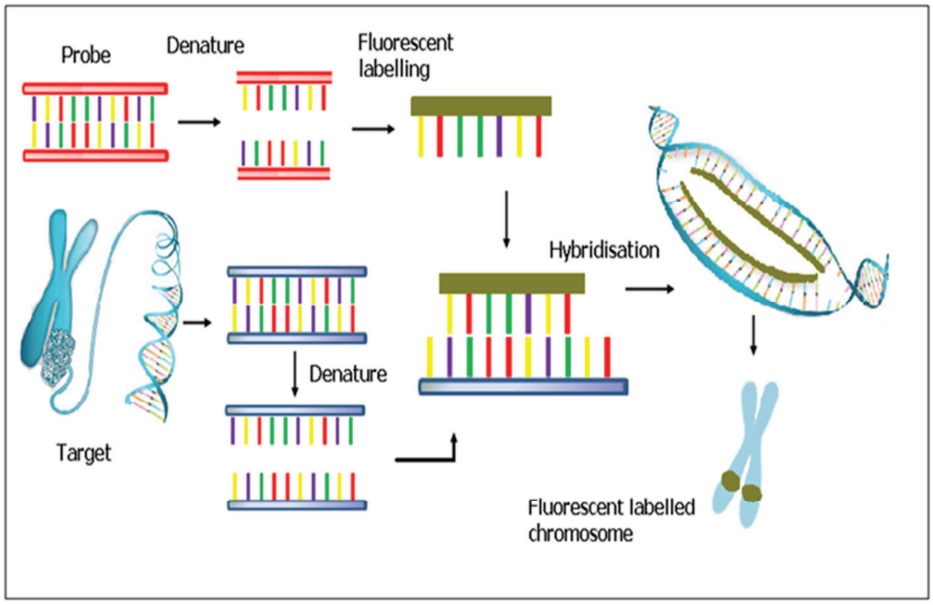


Figure 2.4 Schematic representation of FISH technique

2.2.3 Comparative Genomic Hybridization (CGH)

Comparative genomic hybridization is a cytogenetic technique introduced by Kallioniemi et al. (1992) for the analysis of chromosomes. CGH method made it possible to test for a wide variety of abnormalities

throughout the chromosome with a single experiment. CGH performs a comparative analysis of the test DNA with a normal, reference DNA. It can also be termed as competitive in-situ hybridization involving two samples from different sources fluorescently labeled with two different colours. The test DNA and the reference DNA, which is obtained from a normal individual, are isolated and independently labeled with fluorochromes. Commonly the test DNA is labeled with red fluorochrome and the reference DNA is labeled with green fluorochrome. Both the sequences are subjected to denaturation to convert double stranded DNA to single DNA strands. The two samples are mixed in 1:1 ratio and are applied to normal human metaphase chromosome slide preparation, where competitive hybridization takes place. The DNA on the metaphase slide is also obtained from a normal individual. The relative amount of test DNA and reference DNA that got hybridized depends on their relative amounts in the mixture. Using fluorescence microscopy, the colour variation of fluorescence are visualized and recorded. The red to green fluorescence ratio along the length of chromosome on the slide represents the amount of corresponding DNA sections in the test DNA sample. In the case of the test DNA being a normal one, the amount of both samples will be equal resulting in a ratio=1 and a yellow colour at the given location. Ratio value <1, represents a loss of genetic material in the test DNA. The resulting colour will be shifted towards green. Similarly a gain in test DNA will be indicated by a ratio >1 and the fluorescence shifting towards red. This enables a quick and easier diagnosis of chromosome copy number alterations for clinical diagnosis.

CGH offered significant improvement in terms of resolution obtained (2-4 Mbp) compared to the earlier methods like FISH and chromosome banding which were limited in their resolution by the

requirement of fluorescence microscope and expertise needed in interpreting the acquired image. This was also helped by improvements in other related techniques like polymerase chain reaction (PCR) and micro dissection which enabled the detection of smaller chromosomal abnormalities. However, to detect aberrations in the order of kilo base pairs required newer and better methods. And the solution was provided by Array CGH or Array Comparative Genomic Hybridization. It combines the microarray technology with the principle of competitive in-situ hybridization used in CGH (Solinas-Toldo et al., 1997).

2.2.4 Microarray based Comparative Genomic Hybridization or Array CGH

Microarray based Comparative Genomic Hybridization or Array CGH is an advanced technology that allows detection of submicroscopic chromosomal alterations that cannot be detected with the help of a microscope. These include micro-deletions and micro-duplications. This high-resolution method facilitates the exploration of the genome to map and measure relative changes in genome sequences. Array CGH technology also allows genome-wide scanning of differences in DNA copy numbers by simultaneous monitoring of thousands of genes throughout the genome. Thus it initiated a shift in approach from locus specific analysis to genome wide analysis. The principle behind the Array CGH technology is the same as that of CGH, which is competitive in-situ hybridization. In CGH microarrays, DNA from the test cell is directly compared with the DNA from the normal cell using several thousands of small DNA fragments, with known identity and genomic position. The main factor that differentiates Array CGH from conventional chromosome CGH method is the difference

in the design of probe or the support for hybridization. Instead of metaphase chromosomes which are used as probes in traditional CGH method, Array CGH uses a matrix of short DNA sequences such as cDNA, BAC clones and oligonucleotides arranged on a slide. Since the chromosomes have a highly condensed and coiled structure, the usage of metaphase chromosome limits the detection of smaller aberrations of size less than 20 Mb (Pinkel et al., 1998), which can be dealt with by smaller probe sequences. The chromosome CGH method also required high expertise to identify aberrations from images.

The first array based CGH implementation was demonstrated by Solinas-Toldo et al. (1997), where DNA fragments (large genomic clones) were immobilized into an array of spots on a glass slide and competitive hybridization was performed. The ability of the Array CGH method using genomic DNA probes in detection and mapping of copy number abnormalities associated with disease phenotype was demonstrated by Pinkel et al. (1998). Changes in copy number was identified and measured by relating it with the ratio of fluorescence intensity of test and reference dyes on the array. Microarray implementation of CGH using cDNA probes was demonstrated by Pollack et al. (1999). A genome wide analysis for the detection of copy number variations and its association with the breast cancer occurrences was performed using cDNA probes and the results are validated by comparison with expression data. Heiskanen et al. (2000) used the cDNA based technique to detect the presence of gene amplifications in several cancer cell lines. Later short oligonucleotide probes where introduced in the place of cDNA and genomic clone probes to study copy number aberrations by Carvalho et al. (2004).

2.3 Copy Number Variation

Structural variations in human chromosomes play a significant role in the human diversity, genetic diseases and disease susceptibility. These structural variations are caused by a number of reasons such as deletion, duplication, translocation, insertion, inversion etc. It can also affect sequences of a few kilo base pairs to mega base pairs in length.

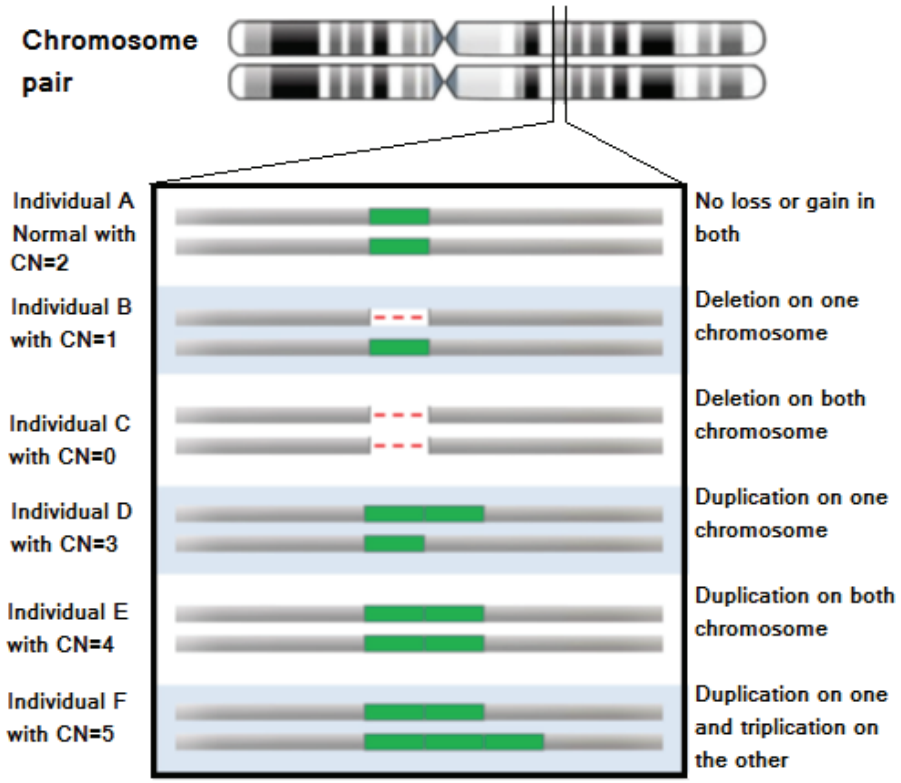


Figure 2.5 Copy Number Variation (CN = Copy Number)

Copy number variation (CNV) is a type of structural variation occurring due to changes in DNA of a genome, resulting in the cell having an abnormal number of copies of one or more sections of DNA. In a normal

human being the DNA sequence or genes are found in two copies in a genome. CNVs are instances where this can be 1, 3 or more copies in number or even absent in a genome. The term CNV is defined to include instances of deletion or duplication event involving sequences of length >1 kilo base pairs (Redon et al., 2006; Feuk et al., 2006; Freeman et al., 2006). CNVs can be of two types- copy number loss due to deletion of large segments of the genome resulting in fewer than the normal number of copies of DNA and copy number gain due to duplication resulting in more than the normal number of copies of DNA. Most of the CNVs are benign variants that are harmless. But certain deletions and duplications can result in abnormalities, when they involve regions containing developmental genes.

2.3.1 Mechanisms contributing to CNV

Many different mechanisms are found to be related to the formation of copy number variants, which can be categorized into homologous and non-homologous mechanisms.

Homologous mechanisms

Non allelic homologous recombination (NAHR) on chromosomes is a source for the occurrence of CNV (Redon et al., 2006). It is a form of homologous recombination that occurs between non allelic sequences having high sequence similarity due to the presence of low copy repeats (LCRs). Deletions and duplications of DNA sequence can occur during the genetic rearrangement, due to misalignment. The occurrences of large scale variations in regions associated with segmental duplications caused due to tandem repeats of segments and duplication during transposition are also reported in Iafrate et al. (2004).

Non homologous mechanisms

This can again be classified into replicative and non-replicative mechanisms. CNV can occur during non-homologous recombination events like non homologous end joining (NHEJ) and micro homology mediated end joining (MMEJ) (Hastings et al., 2009). These are non-replicative mechanisms that can result in genetic rearrangement. Replicative mechanisms are also linked with occurrences of CNV. Replication slippage and FoSTes (Fork stalling and template switching) are mechanisms that come under this category. Replication slippage happens during DNA replication when sequences between homologous segments are either deleted or duplicated. This occurs within a replication fork and hence involves deletion or duplication of shorter sequences. Another replication based mechanism responsible for rearrangement is fork stalling and template switching (FoSTes) (Lee et al., 2007). The proposed mechanism happens during the replication process and occurs between replication forks. During replication the replication fork can stall under stress. The lagging strand then switches the template to another nearby replication fork with micro homology. The role of this mechanism in Pelizaeus-Merzbacher disease (PMD) has been analyzed and elucidated.

The earliest identified CNV associated with a phenotype variation was reported in *drosophila melanogaster* where the duplication of the Bar gene caused the Bar eye phenotype (Bridges, 1936). Studies on large scale structural variations in human genomes by Iafrate et al. (2004) reported imbalances in 255 loci. Among these identified loci, 56% overlapped with regions coding for genes. Some of the CNVs reported included regions associated with disorders like Cri du chat syndrome, Spinal muscular atrophy and Prader-Willi and Angelman syndrome. Meanwhile Sebat et al.

(2004) identified 221 CNVs in human genome, of which several variation are found to be associated with various disorders. Redon et al. (2006) constructed a first generation CNV map of the human genome. DNA from 270 individuals is analyzed and 1447 regions having CNVs are identified. Relevance of CNVs in diseases such as DiGeorge, Williams–Beuren, Prader–Willi, Smith–Magenis and Angelman syndromes are also identified. CNVs are also found to be associated with susceptibility to HIV, with certain traits like Alzheimer's, Autism and phenotypes like Schizophrenia and Psoriasis (Zhang et al., 2009). These studies performed on human population helped to unravel the complexity of genetic variations and provided a framework for further analysis to understand the mechanisms resulting in CNVs and its significance. These variations are also found to have significant impact on the process of evolution where beneficial variations are inherited due to natural selection (Locke et al., 2003) and in maintaining genetic diversity within populations (Sebat et al., 2004). The existence of gene families has been attributed to the positive selection received by duplications (Redon et al., 2006).

2.3.2 Detection of CNV

Cytogenetic techniques employed to detect the CNV can be broadly classified as methods for detecting microscopic structural variations and methods for submicroscopic structural variations. G- banding and C-banding are examples of commonly used microscopic structural variation detection method. They are widely employed for identifying variations that are microscopically visible like aneuploidy. The resolution of these methods is determined and limited by the capacity of optical microscopes employed for visualizing the chromosome. In order to overcome these limitations FISH technique was introduced. It uses region specific probes to detect

smaller aberrations. Advance in molecular technologies like CGH and microarray technology made it possible to detect sub microscopic variations with better accuracy. One main advantage of CGH based method was that no prior knowledge of type and location of structural variation were needed. Genome wide survey of CNVs is made possible by the introduction of Array CGH technique. Bacterial artificial clones (BAC) were initially used in these types of arrays. Various new techniques like oligonucleotide array, representational oligonucleotide microarray, SNP array are also used for detecting structural variations.

Array CGH allows a genome wide analysis of the variations which can meet the ever growing demand of processing speed. The method uses an array of probes attached on a substrate to quantify the target genomes which hybridizes to the probes. The probes on the hybridized array when scanned produce an image with the fluorescence intensity of each probe location proportional to the amount of hybridization. A comparison of these image with an image obtained from a reference diploid genome allows to identify the regions of copy number gain or loss based on the comparative increase or decrease of fluorescence intensity of each probe location on the array. The accuracy and efficiency of the copy number variation localization depends a lot on the Array CGH image processing and data analysis techniques, both of which are covered in detail in the following chapters. A block representation of different steps involved in the detection of CNV using Array CGH is given in Fig 2.6.

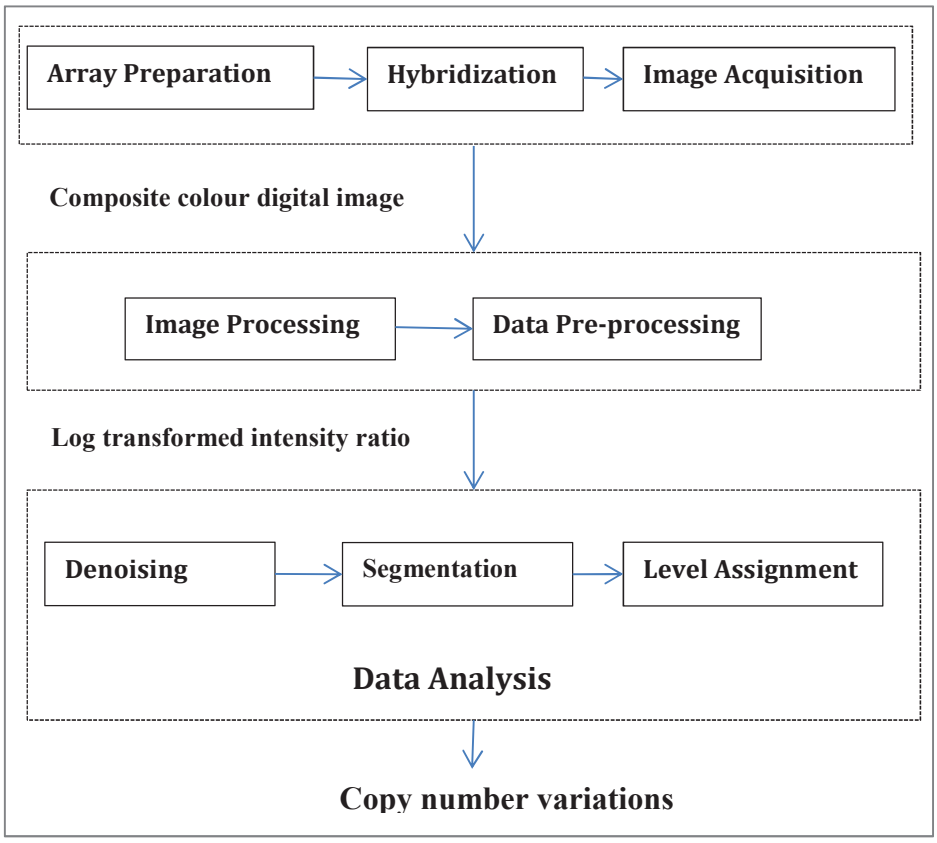


Figure 2.6 A block representation of different steps in the Array CGH based CNV analysis

2.4 Significance of the study

Recent advances in genomics have greatly enhanced the understanding of roles played by various genetic factors like genomic imbalances or alterations in the occurrence and progression of cancers. Array CGH technique has made it possible to perform genome wide screening for all possible sites with significant variations in copy number in a single experiment. This has resulted in a greater demand for improved techniques for processing high density microarray images to estimate

regions of copy number gains and losses. Copy number variation (CNV) studies have helped the field of medical genetics by providing a way to identify disease causing genomic alterations in a large number of diseases. Now, CNV analysis has become a routine procedure for clinical diagnostics and has led to a significant increase in the detection of chromosomal abnormalities (Coughlin et al., 2012). The results of CNV analysis are used for prenatal decision making and genetic counseling. These variations that involve regions containing dosage sensitive genes, oncogenes etc. can cause genetic disorders and complex diseases like autism, immune deficiency and tumors. However, a large number of CNVs within the human genome are found to have no association with any adverse phenotypic outcomes for the individuals. Improved resolution of CNV detection methods can reveal the existence of smaller CNVs in the genome. Microarray based CNV detection has become universally accepted owing to its comprehensive nature of genome wide analysis capability. In spite of these advantages, there are still several challenges like, to limit the number of ambiguous and false findings and maximize the significant or true findings. Another requirement is to increase the consistency and repeatability of the CNV detection methods. Algorithms that can automatically identify the aberration regions and estimate the copy number value of the region is better suited to tackle the need of consistency and repeatability. The noise present in the array image plays a crucial role in determining the CNV detection accuracy of different methods. Over the last few years, different CNV detection methods have been developed with most of them having relative merits and demerits. Automatic CNV detection and localization technique capable of labeling genomic regions as normal, gain and loss is essential and must have repeatability and consistency in the presence of varying noise levels.

2.5 Summary

This part of the work aims to develop a new method for detection and localization of copy number variation using the $\log_2$ratio data from Array CGH. A brief introduction to the branch of cytogenetics and different cytogenetic techniques involved is provided. The chapter also introduces the term copy number variation, various mechanisms resulting in CNV and the significance of studying the CNV in genome. A detailed discussion on Array CGH methodology is provided in the next chapter.

Chapter 3

Array Comparative Genomic Hybridization

A detailed description of Microarray and Array comparative genomic hybridization technologies is given in the chapter. Different types of microarray and the different steps involved in an Array CGH experiment are explained. The data analysis part of the experiment that deals with the interpretation of experimental results for gaining valuable information about the genetic anomalies is discussed. The major tasks involved in the data analysis stage for the detection of copy number variation are also illustrated. Finally, the different computational approaches used for the detection of CNV and a review of literature describing various methods for the detection of CNV are provided.

3.1 Introduction

Array CGH or Array Comparative Genomic Hybridization is a molecular cytogenetic technique capable of identifying submicroscopic chromosomal variations. It is an efficient, high throughput technique that allows genome wide analysis for chromosomal imbalances. Competitive hybridization of differentially labeled test and reference genomic DNA sample on probes arrayed on a glass substrate forms the basis of the Array CGH technique. Advancements in the field of microarray technology, in the areas like printing techniques, surface technology, labeling techniques and automatic visual analyzers, have contributed significantly towards the transformation of in-situ hybridization techniques into a high throughput technique capable of genome wide analysis.

3.2 Microarray Technology

The genome project aimed at sequencing the genomes of model organisms has initiated an unprecedented amount of activities in the field of genome exploration which resulted in the introduction of advanced sequencing techniques. As a result, there is an explosion in the volume of DNA sequence data generated in recent years. The interpretation of this huge amount of genomic data needs a mechanism for comprehensive analysis, where instead of analyzing one gene at a time, methods to process whole genome in a single experiment is needed. Microarray technology facilitated the monitoring of thousands of genes in parallel. The fundamental principle behind the technology is the inherent specificity of the DNA sequences to recognize and bind to its complementary sequence to form a stable structure. Edwin Mellor Southern was the first to employ this specificity for detecting specific sequences among a complex mixture of

DNA fragments (Southern, 1975). The first cDNA microarray was developed at Stanford University by Schena et al. (1995). Forty five complementary DNA (cDNA) clones from the model organism, *Arabidopsis thaliana* are obtained and are printed on glass slides using robots. These clones act as gene specific target probes. Fluorescently labeled mRNA samples are then applied onto the slides and allowed to hybridize. The expression measurement of each of the 45 genes is then obtained by scanning the corresponding cDNA clone location on the array. Further developments in the field of microarray have made it an automatic choice for genome exploration.

The terminology used here describes the DNA attached to the array substrate as a 'probe' and the labeled DNA that hybridizes to the array is termed as 'target'. A DNA microarray consists of a solid substrate, on which an array of probes complementary to the intended targets is deposited in a grid pattern. The cDNA sequence, which is the intended target to be probed, is separately obtained from reference (normal) and test samples. Both the samples are labeled with differently coloured fluorescent dyes. The fluorescently labeled targets are mixed and are applied to the array for hybridization to take place. A fluorescence scanner is used to excite the dyes with the help of two laser sources. On excitation, the dyes produce fluorescent emission from the hybridized spots. Two independent monochrome image of the array is obtained corresponding to each of the two colours (normally red and green). These monochrome images are then combined to create the red-green composite image of the microarray. The fluorescence intensities of each spot correspond to levels of hybridization of targets to probes on that location of the slide. A green spot in the composite microarray image indicates that the gene present in greater concentration in

the reference sample compared to test sample, while an abundance of gene in test sample will produce a red spot. An equal amount of gene in both samples will result in a yellow spot and a gene that is absent in both samples will appear as a black spot. The intensity information obtained is processed and transformed into required data format for further analysis.

3.2.1 Methods of microarray fabrication

The fabrication of microarray originated basically from two different manufacturing approaches. They are deposition based arrays (spotted arrays) and in-situ synthesized arrays. The spotted arrays are based on the Southern's technique described previously. The cDNA clones or oligonucleotide sequences are first synthesized and then using robots they are deposited on the array substrate in a grid form. In the case of in-situ synthesis, probes are synthesized directly on the substrate. Photolithographic synthesis was the first such technique that introduced a reliable way for in-situ synthesis of arrays (Fodor et al., 1991; Pease et al., 1994). It uses a light directed oligonucleotide synthesis where light and light-sensitive masks are used to build a sequence with one nucleotide at a time. The photo-protected surface is selectively de-protected by illumination through a mask. After the controlled removal, the addition of a single specific base is performed. This cycle of de-protection and nucleotide addition is repeated until the sequences of every probe become fully synthesized. In situ synthesis is also performed with the help of ink jet printing and electrochemical synthesis.

3.2.2 Single channel and dual channel arrays

Two different approaches are employed in the microarray experiments, single channel arrays and dual channel arrays. In dual channel

(two colours) arrays, two samples, reference (healthy) and test (mutant) are differentially labeled, mixed and allowed to hybridize on the same array. It involves a comparative hybridization and the scanned fluorescence value of the dots on array yields a ratio of gene expression between the samples (Russell et al., 2008). In the case of single channel (one colour) arrays the samples are labeled, and hybridized on separate slides where hybridization takes place. They are then independently scanned and the data is finally combined for comparison.

3.3 Microarray based CGH

Array CGH is a combination of the micro array technology with the comparative genomic hybridization technique. Microarray slide is prepared by depositing and immobilizing thousands of DNA fragments in an evenly spaced grid form on a substrate. The DNA probes can be genomic fragments of different types such as large genomic clones obtained from Bacterial Artificial Chromosome (BAC), P1 bacteriophage based Artificial Chromosome (PAC), plasmids, cosmids, cDNA or short oligonucleotides. The selection of appropriate probe is based on the required resolution and the size of the target sequence to be probed. In the next step, DNA from a reference sample and the test sample are obtained and dyed with two different fluorophores (Cy5 & Cy3). The differentially labeled samples are mixed and are allowed to hybridize competitively with the probes arranged on the array substrate. After hybridization, the image of array is obtained and the relative fluorescence intensities at each of the spots are quantified. The ratio of fluorescence intensities of Cy3 and Cy5 indicates the amount of a specific DNA fragment in test sample in comparison to the reference sample. An increased amount of test DNA indicates a duplication event and

a reduced amount indicates a deletion event. In other words, targets with normalized test intensities significantly greater than reference intensities indicate copy number gains in the test sample at those positions. Similarly, significantly lower intensities in the test sample are signs of copy number loss. In studies aimed at detecting copy number aberrations, the data to be analyzed is obtained in the form of a normalized $\log_2$ ratio of the intensity. For a diploid organism the copy number data will have a median value equal to zero representing a normal DNA without any copy number loss or gain. For a single copy number gain the $\log_2$ ratio would be 0.58 and for a single copy loss it would be -1. An insight into the magnitude and location of the copy number aberrations of the DNA can be obtained from the analysis of the log ratio data. The goal is to effectively identify locations of gains or losses of DNA copy number that will make it easier to characterize the genomics diseases as well as to identify the targets for treatment.

3.3.1 Types of CGH microarrays

The microarrays used in CGH applications can be classified into two, based on the aim of the experiment. They are whole genome arrays and targeted arrays. Whole genome arrays are designed to interrogate the entire genome of an organism in a single experiment. This is mainly useful when screening for copy number aberrations in the genome, where prior information about its presence and location is not available. Multiple samples affected by cancer (or any other abnormality) are analyzed and the recurrent regions of copy number variations are observed to identify the genes or regions associated with the cancer development. They can detect all types of structural variants as it covers the entire genome, including intragenic, intronic and exonic regions. Meanwhile the targeted arrays restrict the analysis to certain regions of the genome and will not capture the

full complexity of all variations, which often involves intronic and intragenic regions. Targeted arrays are mainly used when diagnosing for a disease phenotype. The targets are designed specifically for a specific gene or region of genome with known clinical significance and the arrays would not be having probes spanning the entire length of genome.

3.3.2 Technological approaches

There are different approaches used for the implementation of Array CGH technique based on the properties of DNA deposited on the substrate as probes and on the sequences to be interrogated by the targets. The source of DNA sequences that are immobilized on the glass substrate can be mainly grouped into three classes: genomic clones, cDNA clones and oligonucleotides.

Genomic clones

Genomic clones are replicates of DNA segments generated from the organism of interest using DNA cloning. DNA cloning is a method that allows the generation of pure sample of a gene or DNA segment, separate from all other genes or DNA segments. Many different types of cloning vectors are used for the efficient replication of the required DNA fragment. The size of the required clone is a critical factor in the selection of the vector to be used. The different types of vectors and the corresponding genome insert size that they can carry are given in Table 3.1. A specific DNA fragment is inserted into a vector DNA capable of carrying a foreign DNA fragment and replicating inside a host. The resulting recombinant DNA is then introduced into a host organism. The most commonly used hosts are bacterium, *Escherichia coli* or the yeast, *Saccharomyces cerevisiae*. All the vectors, other than Yeast Artificial Chromosomes (YAC)

mentioned in the table 3.1 use *Escherichia coli* for propagation while YACs use *Saccharomyces cerevisiae* as the host. The inserted DNA then undergoes replication within the host and passes copies of them to their daughter cells after cell division. This process of replication and cell division will result in a colony of host with the recombinant DNA. Since the hosts have a faster growth rate, large number of genetically identical host organisms are obtained, each carrying a copy of the recombinant DNA with the foreign DNA fragment insert. The genes or DNA fragments are then separated from the clone for use as probes in microarray slides. An example of the process is illustrated in Fig 3.1, where the vector used is a plasmid.

Table 3.1 Comparison of vector type and probe size

Cloning Vector	**Size of genome insert**
Plasmids	Upto 15 kb
Cosmids	Upto 45 kb
P1 bacteriophage	70 -100 kb
P1 Artificial chromosome (PAC)	130-150 kb
Bacterial artificial chromosome (BAC)	120-300 kb
Yeast artificial chromosome (YAC)	0.2 -2 Mb

cDNA clones

The cDNA or complementary DNA sequence can be described as DNA copies of mRNAs and hence contain only the coding regions of the gene. Usually this process of generating a complementary sequence corresponding to the mRNAs is performed with the help of reverse

transcriptase enzyme which is an RNA dependent DNA polymerase. The cDNA sequences are then cloned using vectors to generate large number of copies of the required probe sequence. The cDNA sequences offer a target sequence of size 0.5 to 2 kb which is considerably smaller than genomic clones.

Oligonucleotides

Oligonucleotides are short DNA sequence with a single strand having a few nucleotide residues. They are synthesized by chemical process where reactive residue on a growing nucleotide chain is prevented from reacting using a protection group. To add a specific base to the sequence, first a controlled de-protection is done followed by a base addition. This cycle of de-protection and base addition is performed until the required sequence is synthesized. This synthesized oligonucleotide can then be deposited on the slide. Oligonucleotide array can also be synthesized in-situ, as described in section 3.2.1.

3.3.3 Quality measures

There are certain criterion that needs to be take care of while designing an Array CGH experiment such as resolution, specificity, sensitivity and SNR. These can also be termed as quality measures.

Resolution

Resolution of an array can be defined as the smallest aberration it can detect. The resolution of detection of CNVs is mainly determined by the spacing between the DNA probes on the array and the size of the probes. Resolution can also be increased by using overlapping probes. As the size and spacing of DNA probe on the array decreases, the resolution increases.

Specificity

It is the ability to detect the targeted sequence to be interrogated without cross hybridization to non-target sequence with sequence similarity. It measures the amount of allowable sequence similarity without the possibility of cross hybridization event. Specificity is directly related to the length of the target sequence on the array substrate. Shorter sequences increases cross hybridization and hence reduce the specificity.

Sensitivity

Sensitivity is the ability to correctly classify a hybridization intensity value as an aberration and specifies the minimum intensity variation required to discriminate the aberrant ones from normal ones. Sensitivity increases with the hybridization efficiency of the probe and its target, the affinity between the molecules and their purity. It is also directly related to the length of target sequence and also on the abundance of the target sequence in the sample.

Signal to Noise Ratio

It is a measure of the ability to distinguish the spots from the background. High SNR is achieved by higher amount fluorescence intensity of hybridized spots on array and a lower intensity of the background. The hybridization signal intensity depends on the size of sequence on the spot and increases with the length of sequences. It also depends on the purity of both of the sequence.

3.4 Stages of Microarray based CGH study

The activities involved in a typical microarray experiment can be grouped into 7 steps: probe selection, sample preparation and labeling,

hybridization and washing, image scanning, image processing, data preprocessing and data analysis. The laboratory part of the experiment involving the first three steps is shown in Fig 3.1. The data extraction and analysis steps are shown in Fig 3.2.

3.4.1 Probe selection

One of the important steps is the selection of the most appropriate probe based on the aim of the study and the array technology. It is the probe that acts as the sensor to detect and measure the amount of DNA sequence in the sample to be tested. The selection of the probe determines the reliability and resolution of the experiment. Probes can be either cDNA clones or short oligonucleotides. cDNA probes are better suited for high throughput studies aimed at monitoring a large number of genes on a genome level. For targeted studies aimed at monitoring certain specific genes, shorter oligonucleotide probes can be a better option. Another criterion is the selection of the length of probe depending on the intended target (Tomiuk, 2001). Apart from these there are other factors like specificity, sensitivity, noise and bias that need to be considered while selecting a probe (Russell et al., 2008).

3.4.2 Sample preparation and Labeling

DNA microarray sample preparation includes, isolation and purification of DNA segments from the two samples, test and reference. Isolation step intends at obtaining a set of short DNA probes effectively covering the entire genome. The sample DNA is denatured to separate the strands to create fluorescent nucleotide incorporating single strands of DNA. In order to distinguish between the DNA sequences that bind to the array of probes, the samples are labeled differently. The most commonly

preferred labeling method is using fluorescent cyanine dyes Cy3 and Cy5 which are excited by green and red lasers respectively (Zhang, 2006). The widely used system of labeling uses Cy3 dyes for the reference sample from healthy tissue and Cy5 dye for the test sample.

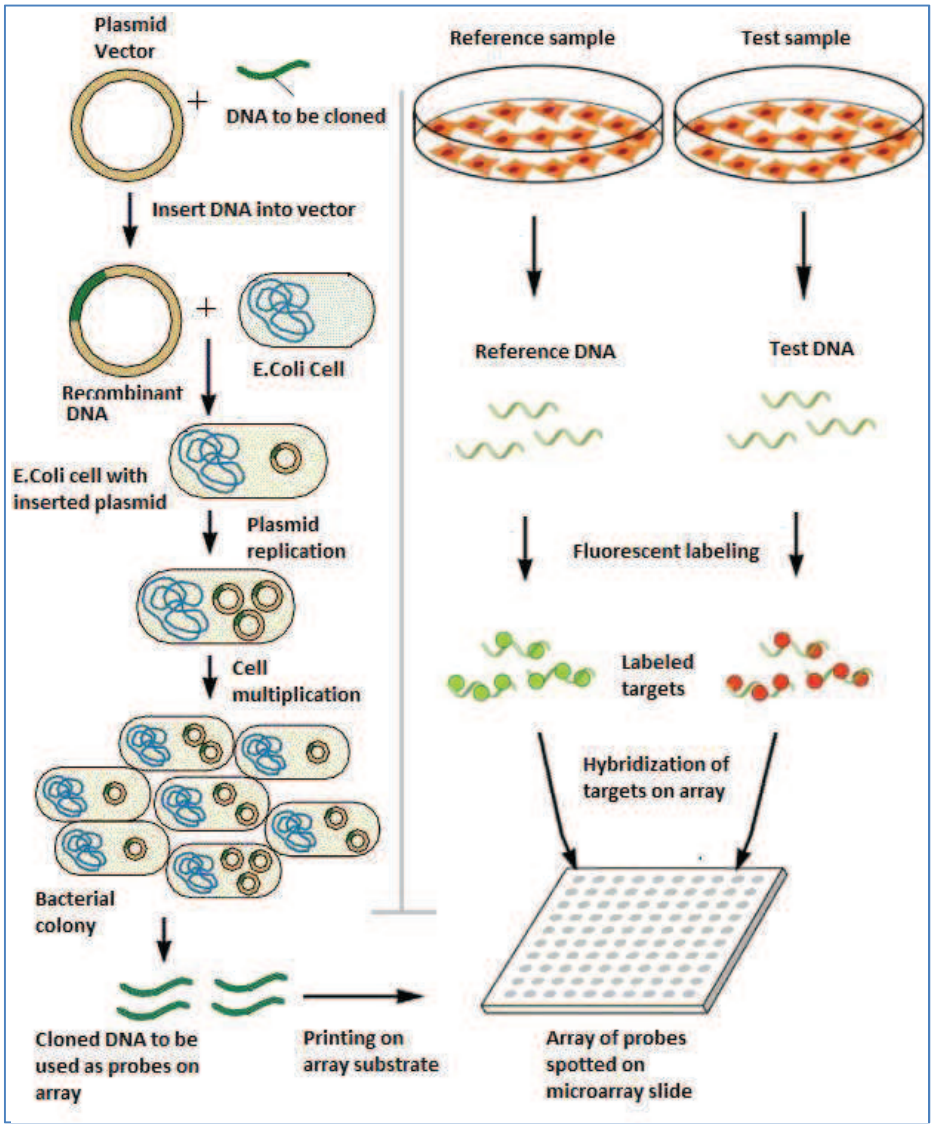


Figure 3.1 Probe, target preparation and hybridization on microarray slide

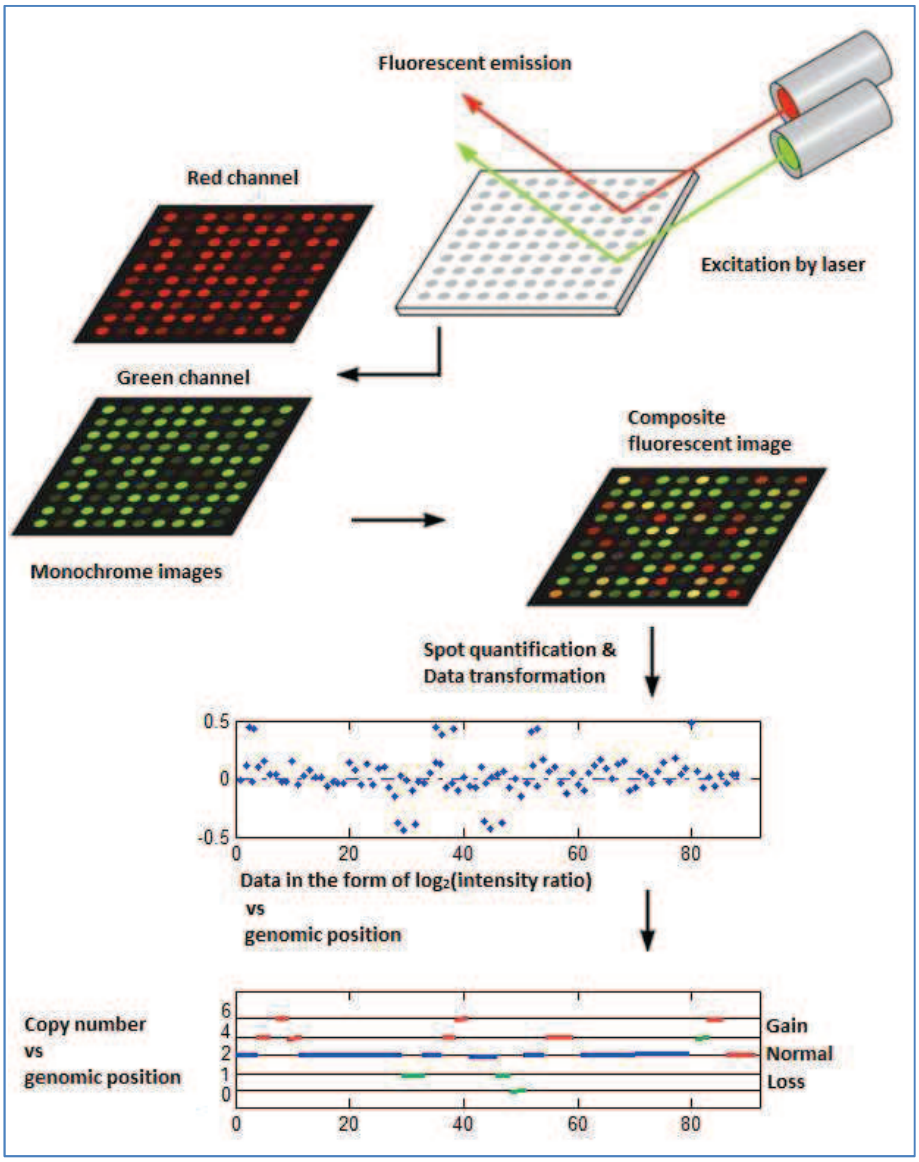


Figure 3.2 Microarray image scanning & processing, data analysis stages

3.4.3 Hybridization and Washing

Hybridization reaction involves the interaction of labeled samples with the probes on the array. The process is driven by the affinity of a DNA strand to bind to its complementary strand. The array is treated before hybridization to minimize background due to nonspecific binding of nucleic acids in sample with that on array surface. The labeled samples are then mixed together and applied to the array at a specific temperature and for a predetermined duration. After the hybridization phase, the array is washed to remove unbound or weakly bound samples. The hybridization, washing and subsequent drying have to be performed uniformly across the array surface.

3.4.4 Image scanning

Once the hybridization process is over, the array is scanned to determine the amount of labeled DNA bound to the immobilized probe on the array. The fluorescent dyes emit light when excited with lasers and the microarray scanner uses two separate lasers to excite the fluorophores on the sample. The excitation wavelength for Cy3 is 550nm and emission wavelength is 580nm and for Cy5 the excitation wavelength is 650 nm and emission wavelengths is 670nm. Separate monochrome image, corresponding to red channel (Cy5) and green channel (Cy3) are obtained. They capture the fluorescence intensity information proportional to the strength of hybridization of test DNA and reference DNA respectively. These images are then combined to create a composite colour digital image of the surface of the hybridized array for further quantification of the hybridization level.

3.4.5 Image processing

The image processing techniques primarily aim at removing the irrelevant details and thereby enhancing the features of interest for a given problem. The Array CGH data processing starts with image acquisition using laser scanners and ends with the data analysis to interpret it. The composite red-green image obtained from the raw data needs to be processed to extract the exact intensity information of the spots on array. In the ideal scenario, (composite image in Fig 3.3) the array must have sub arrays with equal size and spacing, uniform size, shape and spacing of spots within a sub array, uniform background intensity and no contamination. But in real microarray experiments, the images obtained are not ideal ones and needs of image processing steps to locate the spots and measure the fluorescence intensities.

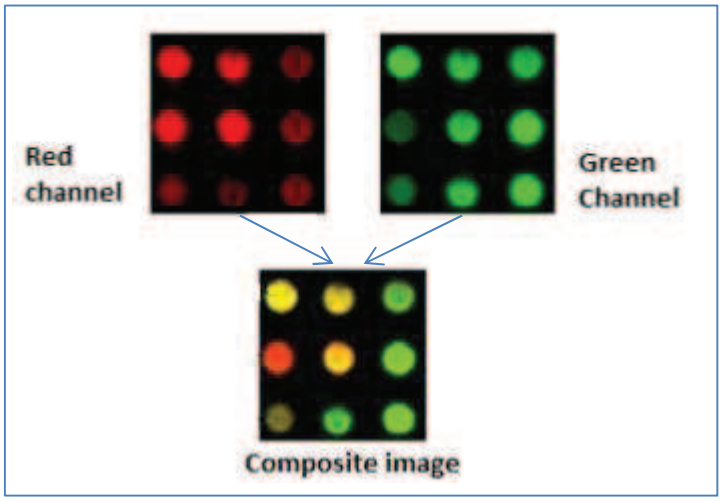
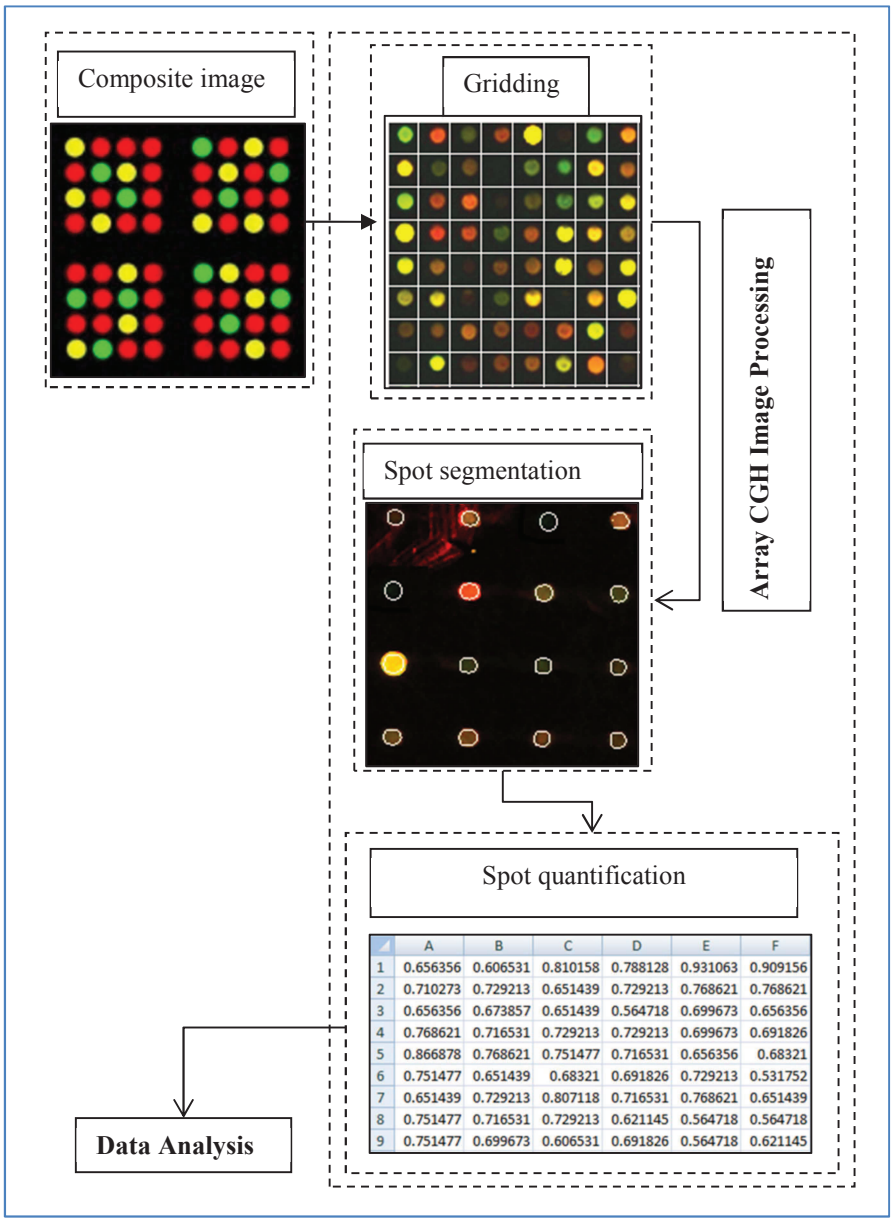


Figure 3.3 Composite array image obtained by combining red channel and green channel images



	A	B	C	D	E	F
1	0.656356	0.606531	0.810158	0.788128	0.931063	0.909156
2	0.710273	0.729213	0.651439	0.729213	0.768621	0.768621
3	0.656356	0.673857	0.651439	0.564718	0.699673	0.656356
4	0.768621	0.716531	0.729213	0.729213	0.699673	0.691826
5	0.866878	0.768621	0.751477	0.716531	0.656356	0.68321
6	0.751477	0.651439	0.68321	0.691826	0.729213	0.531752
7	0.651439	0.729213	0.807118	0.716531	0.768621	0.651439
8	0.751477	0.716531	0.729213	0.621145	0.564718	0.564718
9	0.751477	0.699673	0.606531	0.691826	0.564718	0.621145

Figure 3.4 Array CGH image processing

The processing consists of 3 basic steps: Grid placement, Spot segmentation and Spot quantification. Automatic gridding methods, locate

the sub arrays and the spots within them successfully even in the presence of commonly found noise and contaminations. The segmentation method then automatically fixes the seed and the thresholds based on the characteristics of the individual spots and separate the foreground intensities from the background. Quantification of the fluorescence intensity of the spots is then performed to obtain a numerical value representing the intensity information.

Grid Placement

The location of individual spots within the image is identified along with their size in this initial step. Most arrays are arranged as a series of sub arrays each containing rows and columns of spots. The grid placement step aims at finding the position of array within the image, location of the sub arrays within the array and the distance between the rows and columns within a sub array.

Spot Segmentation

After positioning the grids and identifying the location of spots in the image, spot segmentation involves distinguishing and separating the pixels representing the spot from the background. The segmentation process aims at classifying the pixels into foreground pixels and background pixels based on certain criteria. As the density of the microarray increases this estimation of foreground and background pixels becomes increasingly difficult. Fixed circle segmentation, adaptive circle segmentation, adaptive shape segmentation, K-means segmentation and histogram segmentation are different types of techniques used for spot segmentation in microarray images.

Spot Quantification

The pixels belonging to the spot (foreground) and the background are distinguished in the segmentation step. This final stage of the image analysis aims at obtaining a quantitative measure of the foreground and background fluorescence intensity value for both channels of all spots. The pixel values from both channels are combined to calculate the composite fluorescence intensity value of a given spot. One simple approach of calculating the intensity value of a spot is to use the mean value of pixels within the segmented spot. Similarly the background value can be calculated as the median of pixels around the segmented spot.

3.4.6 Data preprocessing

Data preprocessing is an essential step before any downstream statistical analysis or clustering approaches. There are three major steps involved in this processing: Correction, Transformation and Normalization.

Data Correction

It involves mechanisms to remove errors introduced in the raw intensity data due to various experimental procedures. Presence of background noise due to nonspecific binding, debris left after washing of array and noise from scanner is common errors found in raw intensity data. Background noise correction is required to prevent biased results due to over estimation of the foreground intensity value. Missing intensity data of spots is another such error that can arise due to various reasons like, contamination of slide, higher background intensity compared to foreground intensity and zero intensity spots. Missing data estimation is another essential operation to ensure the reliability of the downstream analysis. Various methods are used for this estimation which includes replacing

missing values with zero, replacing with row average, using K-nearest neighbor method and using imputation (Zhang, 2006).

Data Transformation

Data transformation involves the conversion of raw intensity information of spots to another scale of measurement to prevent misleading results due to skewed distribution of intensity values. Most commonly used transformation methods are transformations to ratio of intensity and logarithmic ratio of intensity. Intensity ratio metric is the ratio intensity value for Cy5 and Cy3 or in other words, the ratio of red fluorescence intensity to green fluorescence intensity. It is given by the equation 3.1.

$$\text{Intensity ratio} = R/G \qquad (3.1)$$

Logarithmic ratio is a log transformation given by the logarithm to base 2 of the intensity ratio. One of the features of this transformation is a nearly gaussian distribution obtained for the transformed data. It can also represent a wide range of intensity of values from zero to infinity. Log intensity ratios are most commonly used form of transformation especially in the case of expression measurements and copy number variation analysis.

$$\text{Log ratio} = \log_2(\text{intensity ratio}) = \log_2(R/G) \qquad (3.2)$$

Data Normalization

The main aim of normalization is to remove any systematic errors introduced in intensity measurements. The normalization methods can be classified as global schemes and intensity dependent schemes. One of the frequently used normalization method is LOWESS or locally weighted linear regression, which is an intensity dependent normalization method.

3.4.7 Data analysis

Data analysis is the final step in a microarray experiment where the data generated through a series of steps is processed by statistical or by other computational means to draw meaningful results. The objective of the experiment can be identifying the differentially expressed genes, detecting DNA copy number variation, mapping of binding sites, detecting polymorphisms etc. The selection of computational or processing approach largely depends upon on the specific objective of the study. In the case of copy number variation analysis, the aim is to identify the regions of aberration in the genome such as a loss or gain of DNA segments. The data analysis then becomes a task of keeping track of the log ratio values and identifying the significant deviation of its value from normal level. The problem of converting the raw log ratio of fluorescence intensities of spots on Array CGH slide into discrete copy number values is challenging. Computational methods are required for drawing meaningful conclusions from Array CGH data by partitioning the genome sequence into segments of discrete and separate copy number levels. The task of inferring the DNA copy number of genomic regions from the noisy Array CGH data involves the following steps- denoising, segmentation and copy number level assignment. A block representation of the different steps involved is shown in Fig 3.5.

Denoising

The term noise can be described as disturbances in the signal that are of no interest. These noises are unavoidable in the case of real world signals which are introduced due to a multitude of reasons. In order to process these data for inferring meaningful information, the noise must be removed or

reduced. Denoising step performs the recovery of the original signal that has been contaminated by noise. The denoising thus aims to remove the random fluctuations in the log intensity ratio which makes the detection of aberration boundaries difficult. Most of these are local trends originating due to various different reasons including errors in array printing, labeling, hybridization efficiency and scanning of arrays. Therefore denoising of the data as a pre-processing step is essential to precisely infer the patterns of aberrations in the samples.

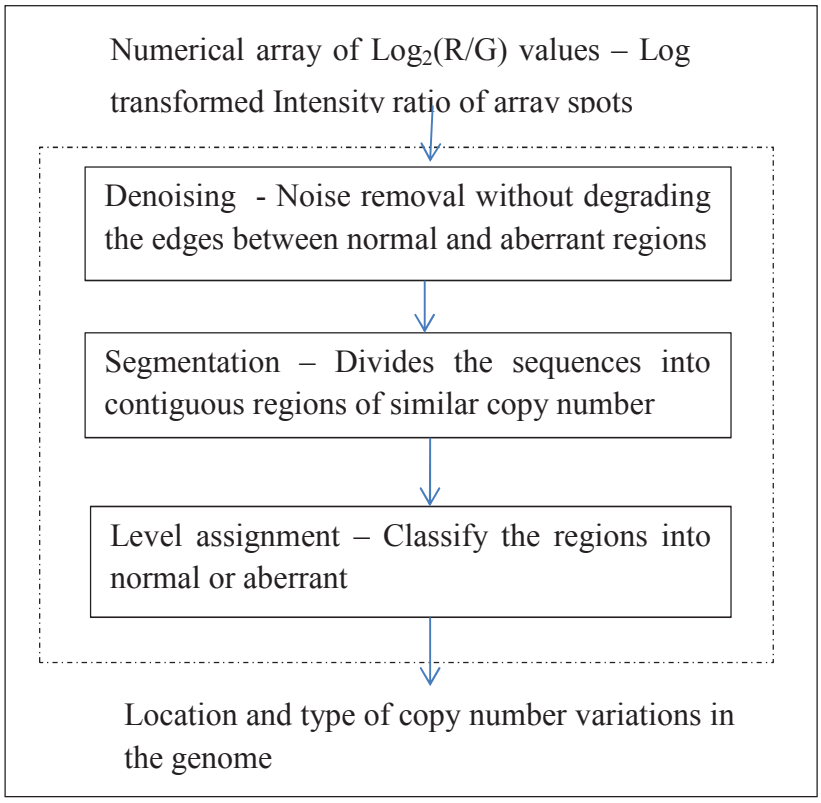


Figure 3.5 Data analysis steps for CNV detection

Segmentation

The segmentation process is the most critical task involved and has received most attention. The segmentation involves splitting genomes into discrete regions or segments. A segment can be defined as a region with probes exhibiting different signal intensity patterns compared to that of the adjacent regions. In other words, the process splits the DNA sequence into continuous regions where copy number is a constant separated by discontinuities or copy number change points. This converts the copy number signal into a piecewise constant signal, where the magnitude of the constant region represents the number of copies of DNA in that region and the discontinuities between the constant levels indicating the locations of copy number variations.

Level Assignment

The segmentation step identifies the copy number change point locations, but do not classify the associated genomic alterations as gains or losses. Since the primary objective of Array CGH analysis is to identify regions of copy number gain and loss, copy number level assignment or copy number status assignment methods have been proposed following the segmentation step. This step labels the individual segments with their inherent copy number and assigns a status such as normal, gain and loss. For that, an estimation of the copy number of the segments already identified is performed. Then using methods such as clustering and thresholding, the level assignment is done. Thus actual aberration locations in the genome can be determined and the information can be used for further downstream analysis of affected genes etc.

3.5 A review of various computational approaches for CNV detection using Array CGH

A detailed discussion of the various approaches that laid the framework for rapid improvements in the computational methods for the detection of copy number variation is needed for better understanding of the problem. Various methods have been proposed for the implementation of copy number detection. They can be mainly classified as thresholding based methods, smoothing methods, segmentation methods and clustering methods. There are a large number of algorithms dealing with the topic with corresponding strength and weaknesses. A direct comparison of the performances of these methods is made cumbersome by the complexity of the algorithms and varying perspective with which they are designed.

Threshold based approaches were initially proposed to divide the genome into individual segments and for classifying them. The individual segments can be characterized by certain parameters like mean, median or variance. Abrupt changes or steps of size higher than a threshold value can be used to mark the breakpoints or boundaries of the segment. Finally the segments are assigned a level based on the value of mean or median of the segment.

The goal of the clustering methods is to group the individual points into a finite number of separate distinct clusters. In the case of the Array CGH log ratio values, the individual clusters represent group of points with normal copy number, copy number gain and copy number loss. Hidden Markov Model (HMM) based approach falls under this category where the transitions of copy number along the clone locations are modeled as a HMM model with underlying hidden states representing the copy number status of

the clone. Other clustering approaches like K-means clustering and hierarchical clustering are also used for grouping the data points.

Segmentation based methods aim to divide the genome into segments of different copy number value. The process then identifies the edge co-ordinates of segments which are locations corresponding to copy number variations. The major tasks involved in the segmentation based approaches are determining the parameter affected, localization and segmentation of the sequence based on some optimization criteria. Since segmentation approach is the most widely used and discussed method for the breakpoint detection problem, there are numerous algorithms that have been proposed. The algorithms differ in the selection of optimization criteria and the methods used for optimization. Methods like genetic algorithm, dynamic programming and binary segmentation have been used for reaching the optimum criteria with varying computational complexity. Segmentation methods can be classified into online and offline approaches. The online approach is based on local optimization while offline methods provide a global optimal segmentation.

The intensity ratio data obtained from the Array CGH experiment usually has stretches of constant value with abrupt jumps merged in noise. This sort of a problem doesn't allow the use of conventional linear filtering approaches that fails to preserve the edges and hence the need of local edge preserving smoothing methods. The smoothing algorithms model the data as a series of linear segments with unknown abrupt boundaries which needs to be identified. The smoothing methods aim at reducing the local variations or noise by comparing each data point with a defined neighborhood. This local smoothing helps in remove unwanted outliers that results in false breakpoints. Once the outliers and noise is removed, the data can be

visualized as flat segments with intermediate steps which allow easier identification of change points.

One of the initial efforts in the analysis of copy number variation using microarray technology can be attributed to Hodgson et al. (2001) where a model based maximum likelihood approach was employed. The method involved the classification of data points into three possible states or conditions namely- normal, increased and decreased copy number. Pollack et al. (2002) proposed a threshold based method for the identification of gain and loss regions in breast cancer cell line. The algorithm performed a local smoothing followed by threshold based detection. Weiss et al. (2003) also applied a threshold based implementation to identify gains and losses in copy number. The threshold was fixed based on the criteria of false discovery rate.

Hidden Markov Models (HMM) based methods to cluster the clones into homogenous groups with same underlying copy number was then introduced (Snijders et al., 2003; Fridlyand et al., 2004; Sebat et al., 2004). The data points are modeled as an HMM where the log intensity ratio represents the observed state corresponding to hidden state which is the underlying copy number. The HMM based approach was extended by Guha et al. (2008), where a Bayesian HMM method for the analysis of Array CGH data to make inferences about gains and losses in copy number was proposed. HMM based methods make use of the spacial coherence of the signal. It aims at grouping the genome coordinates into finite number of hidden states corresponding to normal and different aberrant states and thereby allowing to identify the breakpoints.

The clustering technique for genome wide screening of genetic alteration proposed by Wang et al. (2005), employed a variation of standard agglomerative clustering algorithm. The algorithm named cluster along chromosomes (CLAC) builds hierarchical clustering trees along each chromosome arms such that the gain and loss regions are separated into different branches. The method also proposes the use of a moving window mean filtering on the raw data prior to clustering where the gain and loss regions are selected based on certain criteria to keep the false discovery rate at desired level.

Jong et al. (2003) introduced a statistical segmentation method using genetic algorithm for chromosomal breakpoint detection. The method used a maximum likelihood criterion for the optimization of the model fitting problem. It proposed a genetic local search algorithm for the selection of the most probable partition of the data for a given number of break points. Circular binary segmentation (CBS) proposed by Olshen et al. (2004) can be termed as one of the pioneering contribution in this area of copy number variation detection. It is also the first major statistical segmentation approach for the task of breakpoint detection. It is a modified form of binary segmentation and performs a non-parametric estimation of change point locations. The method aims at identifying all the change points that partitions the chromosome into segments where copy numbers are constant. Then the copy numbers of the segments are estimated. Picard et al. (2005) introduced a new statistical method for the analysis of the of Array CGH data using a segmentation approach. The method models the copy number ratio as a gaussian distribution affected by abrupt changes at unknown locations. A new procedure for the estimation of the number of segments

using a penalty term which is adaptive to the data is employed, followed by a dynamic programming algorithm to locate the change points.

The smoothing algorithms used in the CNV detection can be categorized into statistical smoothing techniques and signal processing techniques. In the initial stages statistical methods were generally applied but recently more and more signal and image smoothing methods have proved to be efficient in CNV detection. Hupe et al. (2004) introduced a Gaussian model-based approach called GLAD, which uses an adaptive weights smoothing for breakpoint estimation followed by a region assignment method. A quantile smoothing method was proposed by Eilers and De Menezes (2004). The method is based on the minimization of error in L1 norms (sum of absolute errors) rather than in L2 norm (sum of squared errors). This change in penalty term has resulted in a data with flat segments and sudden jumps compared to the more rounded off edges in previous smoothing process. Wang et al. (2009) describes a non-parametric smoothing technique using a mean shift based algorithm. The algorithm performs a locally adaptive noise removal by preserving the abrupt change points. Hsu et al. (2005) introduced another promising approach based on wavelet transform for the smoothing of Array CGH data. Wavelet analysis uses a representation of data in the form of a linear combination of dilated and translated wavelet functions. Here maximal overlap discrete wavelet transforms (MODWT) using Haar wavelets are used for the denoising. The analysis involves 3 steps: the decomposition of data into various frequency sub bands, thresholding of wavelet coefficients and wavelet reconstruction to produce the denoised data. The thresholding step is the crucial step that determines the performance of various wavelet denoising methods and SURE thresholding method was employed in the method for better edge

preservation. Wang, Y. and Wang, S. (2007) extended the wavelet approach to use stationary wavelet transform (SWT) followed by a modified form of universal thresholding for copy number detection. A dual-tree complex wavelet transform method with the bivariate shrinkage estimator (DTCWTi-bi) was later proposed by Nguyen et al. (2007) followed by further extension of their work by using stationary wavelet packet transform with dependent laplacian bivariate shrinkage estimator for Array CGH data smoothing (Nguyen et al., 2010).

A computational tool named, CGH-Plotter, which employs a three stage procedure for detecting constant levels in Array CGH data was put forward by Autio et al. (2003). The three stages include a moving mean/median filter, k-means clustering and finally a dynamic programming algorithm to determine the actual constant levels representing the baseline, amplicon and deletion regions. Myers et al. (2004) also employed a similar three stage approach for the identification of changes. The method named, Chromosomal Aberration Region Miner (ChARM), involves an edge detection filter that identifies the potential regions, an EM (Expectation Maximization) algorithm that finds the maximum likelihood breakpoints in these potential regions and a statistical analysis to determine the significant ones.

A comparison of three segmentation approaches in the analysis of Array CGH data is performed by Willenbrock and Fridlyand (2005). Performance of an HMM based approach, CBS and GLAD are compared on simulated and real data. According to the results CBS method is reported to be better than others in the detection of copy number alterations. It has also an advantage of having better sensitivity and low false detection rate. HMM showed good performance for smaller aberrations and GLAD was more

suited for wider ones. A detailed and independent study of the performances of a large number of the previously described algorithms in detecting copy number variations is attempted by Lai et al. (2005). 11 different algorithms are used for the analysis of Array CGH data. The methods include CGHseg (Picard et al., 2005), Quantreg (Eilers and De Menezes, 2005), CLAC (Wang et al., 2005), GLAD (Hupe et al., 2004), ACE (Lingjaerde et al., 2005), GA (Jong et al., 2003), ChARM (Myers et al., 2004), HMM (Fridlyand et al., 2004), Wavelet (Hsu et al., 2005), Lowess (Beheshti et al., 2003) and CBS (Olshen et al., 2004). Among these, CGHseg, CBS, HMM, GA are segmentation based methods and Lowess, Wavelet, Quantreg belongs to smoothing based methods. Meanwhile, CLAC, GLAD, ChARM and ACE employ a combination of smoothing and segmentation for CNV detection. The methods are tested on simulated data as well as on real data and the results are compared based on various performance criteria. Better computational speed was observed for smoothing algorithms based on wavelets and lowess. CGHseg and CBS performed consistently well when the noise level was low while the smoothing methods were superior in the presence of higher noise level. But the interpretation of results was difficult in the case of smoothing methods compared to the segmentation methods. Based on this, the study recommended an optimal combination of the smoothing step and segmentation step for improved CNV detection.

3.6 Summary

In this chapter, a detailed description of microarray technique and microarray based comparative genomic hybridization is presented. The various stages of an Array CGH experiment are explained. The first three stages can be classified as the laboratory part of the experiment and involves

the preparation of probe and target samples followed by hybridization of probe with the target and washing of the array. The next steps of image scanning, image processing, data preprocessing performs the data extraction operation. The final data analysis part involves gaining meaningful information from the array data obtained. In this work the data analysis part deals with the detection of CNV locations in the genome under study. The important steps involved in the detection of CNV are highlighted. A review of various computational methods used for the analysis of Array CGH data for the detection of copy number variation is also presented. In the next chapter a new method for detection of CNV from the Array CGH log ratio data is described.

Chapter 4

Development of a new method for the detection of CNV from Array CGH data

The development of a new method called Edge Enhancement and Segmentation (EES), for the detection and localization of copy number variations in the genome using log ratio data obtained from Array CGH experiment is explained. The EES method performs an edge enhancement filtering prior to the segmentation of log ratio data into regions of discrete copy number levels. Performance of the EES method is verified with the help of simulated log ratio data and log ratio data obtained from real Array CGH experiment. A comparison of the results obtained using the EES method with other established methods are performed and the improvement in the performance of the EES method is highlighted.

4.1 Introduction

A wide variety of computational methods have been developed for the analysis of Array CGH data for copy number variation detection. The primary aim of these methods is to track the variations in the log ratio data, identify the significant deviations and locate those critical regions corresponding to copy number variation. The task of converting the raw log ratio of fluorescence intensities of spots on Array CGH slide into discrete copy number values is very important and challenging. Computational methods are required for drawing meaningful conclusions from Array CGH data, by partitioning the raw data into segments of discrete and separate copy number levels efficiently. The computational task of inferring the copy number variations involves different stages like denoising, segmentation and copy number level assignment. A clear distinction of these individual tasks may not be visible in various computational approaches. The differences arise from the individual goals of each algorithm, as some of the algorithm gives emphasis to the denoising part while some others emphasize on identifying the points of variations that mark the locations of aberrations. The focus of this work is to develop a novel method which combines the merits of edge preserving denoising and segmentation tasks, followed by a level assignment step for the detection of copy number aberration. The problem can be summarized as a process of segmentation of noisy intensity ratio data into discrete levels separated with abrupt jumps and finally determining the value of these discrete levels to distinguish normal and aberrant genomic regions.

4.2 Denoising

Array CGH data can be modeled as a mixture of noise free copy number variation signal with an additive noise. Smoothing techniques are computationally efficient methods for processing the copy number data and considers the data as a non-stationary signal with sharp transitions and singularities at end points. The smoothing techniques apply various local smoothening methods which aim to remove the random fluctuations in the log intensity ratio that are not due to copy number changes in the sample. The presence of noise makes it difficult to detect the aberration boundaries and estimate the copy number value which necessitates a noise removal step prior to further analysis.

Gaussian noise

Gaussian noise is a statistical noise model that follows a normal probability distribution function. Normal distribution is a probability distribution that is often used to represent real valued random variables whose distributions are not known and whose values are clustered around a mean. In statistics the prominence of normal distribution is due to multiple reasons. The first one is the central limit theorem, which states that the sum of a number of independent and identically distributed random variables with finite variances will tend to a normal distribution as the number of variables grows. The other reason is the tractability, where a linear combination of two normally distributed independent random variables also results in a normally distributed variable. The probability density function of a gaussian random variable, x, is given by,

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad (4.1)$$

where μ is the mean of distribution and σ its standard deviation.

Physical quantities that are obtained as a result of the sum of many independent processes such as measurement errors often have distributions that can be approximated to a normal distribution. The noise found in copy number data is caused due to various independent sources and are often modeled as gaussian noise.

Smoothing and sharpening filters

The filtering operation can be defined in two domains: frequency domain and time domain or spatial domain. The frequency domain filtering concept has its root in the use of FFT (Fast Fourier Transform) for filter implementation in signal processing applications. The spatial filtering operation is directly performed on the pixels of an image or on the values of a data sequence. The spatial filters can be either a smoothing filter or a sharpening filter. The smoothing filters are used for noise removal and are referred to as low pass filters. Mean filter and median filter are examples of smoothing filters. The sharpening filters are also known as high pass filters, and are used to enhance fine details such as edges. Sobel filter is an example of sharpening filter. Many practical applications require filters that can achieve both of these complementary enhancement techniques. The CNV detection operation involves identification of sharp changes in copy number level from the log ratio of fluorescent intensity that is degraded by the presence of unwanted noise. The problem then becomes a task of smoothening out the noise in the data without blurring the sharp edges corresponding to copy number changes, which requires a combination of smoothening and sharpening filters. There are two major categories of filters - linear and nonlinear filters that are used to meet the required objectives of

filtering. The selection of the right type of filter depends on the nature of the data and the objective of filtering.

Linear & nonlinear filters

A linear filter is a system that processes the time varying input data to produce output data, where the output is a linear combination of input. Linear systems follow certain fundamental principles like principle of superposition, shift invariance, causality and stability. An LTI system (Linear Time Invariant) can be uniquely represented by its impulse response, h (n), or it frequency response H (ω). The output of any LTI filter can be expressed as the convolution of the input signal with the impulse response. The convolution of two sequences x[n] and h[n] is defined by the equation:

$$y[n] = \sum_{k=-\infty}^{\infty} x[k]h[n-k] \qquad (4.2)$$

The corresponding frequency response is given by:

$$Y(\omega) = X(\omega)H(\omega) \qquad (4.3)$$

Mean filter or averaging filter is a commonly used example of linear filtering method. The mean filtering replaces each value in the noisy data with the mean or average value of its neighbors to produce the filtered data. Nonlinear filters are systems that do not follow the linearity principles, where the output is not a linear function of inputs. The simplest nonlinear filter is a median filter, where the output depends on the ordering of the input values from smallest to largest or vice versa. The median filter has the ability to remove outliers while preserving the edges. The problem of filtering Array CGH data deserves attention due to the limitation of

conventional linear filtering methods in dealing with the abrupt discontinuities. Conventional fourier transform based methods often perform the denoising by low pass filters which are suitable for signals without abrupt jumps. But the Array CGH data are characterized by abrupt jumps which contain very valuable information regarding the CNV locations. This necessitates non-linear filtering approaches for denoising without degrading the abrupt changes in copy number levels of the log ratio data.

4.3 Cluster analysis

Cluster analysis divides the data into groups that share common characteristics. It provides an abstraction from large amount of data values to the groups or clusters in which those data values belong. It also allows obtaining a suitable representative value to characterize each cluster. This value which is the representative of other data objects in the cluster can act as a basis for further detailed analysis. Thus instead of applying data analysis steps to the entire dataset, it can be applied to a reduced data set containing the representative values of each cluster. This feature of cluster analysis is particularly useful in Array CGH log ratio analysis for the detection of changes in copy number values. Array CGH log ratio data includes log ratio of fluorescence intensity values of entire spots on a micro array slide. This data is considerably large, when an entire genome is being probed. So cluster analysis allows grouping of log ratio values with similar intensity values together with a representative value that characterizes the group. This representative value of each cluster allows classifying the spots within each cluster as normal or aberrant genome loci.

4.3.1 Hierarchical clustering and partitional clustering

Clustering can be classified into types based on whether the clusters are nested or unnested. Partitional clustering groups the data into non overlapping clusters. Hierarchical clustering allows clusters to have sub clusters and the entire structure takes a tree topology. A node in the tree represents a cluster. Each non leaf node is the union of its sub clusters and the root of the tree is the set containing entire data.

4.3.2 Hard Clustering and soft clustering

In hard clustering, each data point is assigned to a single cluster. It is also known as exclusive clustering. Soft clustering can be either overlapping or fuzzy clustering. In overlapping clustering, a data point can simultaneously belong to more than one cluster. In fuzzy clustering each data point belongs to every cluster with an associated probability value with a constraint that the sum of all probabilities is equal to one.

4.3.3 Types of clustering algorithms

Depending on the methodology or the set of rules for defining the similarity among data points, clustering algorithms can be classified into different types. Among the numerous types, the most popular algorithms are: Centroid based, Density based and Graph based algorithms.

Centroid based algorithms

These are algorithms in which, the similarity is determined by the closeness of a data point to the centroid of the clusters. In these models, the number of clusters is a pre-requisite and model iteratively performs the cluster assignment of data points until it satisfies an optimization criteria. K-

means clustering belongs to this category of algorithm. The centroid value of a cluster in K-means approach is the mean value of the cluster members.

Density based algorithms

These algorithms try to identify the regions where there is a high density of data points surrounded by low density areas. The algorithm isolates the different high density regions from one another and assigns the data points within these regions in the same cluster. DBSCAN is one of the most popular density based approach used for clustering.

Graph based algorithms

In a graph based model, the nodes represent the data points and the links shows the connection between them. A cluster is then represented as a set of points that are connected to one another. Contiguity based clustering is an example where the points are connected only if they are within a specified distance.

4.3.4 K-means Clustering

K-means clustering is a partitional clustering technique that groups objects or values based on attributes into K number of distinct groups, represented by their centroid values. The first step in the algorithm is the determination of the number of clusters, 'K'. Then each of these clusters is assigned with an initial centroid value. The centroid value chosen can be a random selection or based on any available information about the distribution of the data points to be classified. For every data point in the sequence, distances to all centroids are calculated and the point is assigned to the closest centroid. This cluster assignment process is performed for all the data points. The collection of points assigned to the centroid defines the

new cluster. Based on the assignment, new centroids are calculated for each cluster. Using the new centroid, the cluster assignment of data points is repeated. This process of cluster assignment and centroid updation is repeated until there is no change in the cluster membership. Most often used optimization criteria is the minimization of the Sum of Squared Error (SSE) which aims to minimize the sum of squared distance between data points and cluster centroids to which the data point belongs.

Selection of proper initial centroids is a key step in the K-means clustering algorithm and plays an important role in the final clustering. Random selection of initial centroids often results in sub-optimal clustering. Another approach is to perform multiple runs using different initial centroids and select the clustering with the minimum SSE. In order to assign a point to the closest centroid, a distance measure, to quantify the closeness is required. Euclidean distance between the point and centroids is often used as the measure of closeness. Manhattan distance, squared euclidean, jaccard measure and cosine distance are some of the alternative distance measures employed for the cluster assignment. After the cluster assignment, the cluster centroid is updated for each cluster. This recalculation of the centroid depends a lot on the selection of optimization criteria and the distance measure in the algorithm. When euclidean distance is used as the measure of closeness, it has been shown that the centroid that minimizes the SSE criteria is the cluster mean value. So for each centroid updation step, new centroids are calculated using the mean value of the points within each cluster. After several iterations the K-means algorithm achieves a local minimum for SSE for a specific set of centroid and clusters. The K-means clustering thus groups the entire set of data points into K groups characterized by their cluster centroids.

4.4 Development of Edge Enhancement and Segmentation (EES) method for the detection of CNV

A novel method called Edge Enhancement and Segmentation (EES), for the detection and localization of copy number variations in the genome using log ratio data obtained from Array CGH experiment is developed. The EES method performs the edge enhancement by retaining the edges while removing the noise, using an edge preserving filter prior to the segmentation of log ratio data into regions of discrete copy number levels. The method involves the following stages:

1) Denoising using an edge preserving filter.
2) Segmentation using a K-means clustering based approach.
3) Copy number level assignment using thresholding method.

Array CGH technique uses competitive hybridization of the test DNA with a normal reference DNA to detect copy number variations. The ratio of fluorescence intensities emitted by the differentially labeled test and reference DNA indicates the amount of a specific DNA segment in the test sample compared to the reference sample. The intensity ratio is often transformed into a logarithmic scale of base 2. The log ratio data obtained from an Array CGH experiment is modeled as a piecewise constant signal immersed in gaussian noise. The piecewise constant approximation of the Array CGH data ensures the importance of the sharp or abrupt jumps in the signal corresponding to the changes in DNA copy number. In the first step of the algorithm, the noisy log ratio data is denoised without degrading the informative edge points. The denoising performs the smoothing action on constant regions and performs the sharpening action on the edges. This helps to recover the locations of copy number changes which were

otherwise difficult to locate due to the presence of noise. This step is followed by a segmentation step for transforming the data into discrete levels or segments separated by edge points. In this algorithm the segmentation operation performs the role of a signal quantizer where the input data is transformed to a quantized output data. And in the final step the individual segments, characterized by the segment mean values, are classified as a region with copy number loss, gain or normal. This classification and copy number level assignment is performed using thresholding technique.

4.4.1 Minimum Variance Filter (MVF)

The EES method uses a local edge enhancing filter called Minimum Variance Filter (MVF) for noise removal and edge sharpening. The MVF (Tomita and Tsuji,1977) performs the denoising using a search for homogeneity in small neighborhoods. The raw copy number data obtained from Array CGH can be considered as a non-stationary signal mixed with additive white gaussian noise. MVF uses a sliding window to create smaller segments where local stationarity can be assumed. Filtering operation aims to obtain an estimate of the original noise free value of each data point.

Let *x(n)* be the noisy log ratio data, *x'(n)* be the noise free data which is to be estimated by the filter and 'W' be the sliding window of size *L*. In order to filter the data point represented by *x(i),* a neighbourhood is defined around it, using the sliding window 'W'. The sliding window is placed so that *x(i)* is its center point. Within this neighbourhood, smaller sub-windows (W_s) are defined. The sub-windows are created in such a way that, all the sub-windows contain the data point *x(i)* as one of its member.

[Fig. 4.1 shows an example, where *x(5)* is the point being processed, size of sliding window, *L*=5, the three sub-windows defined within W is shown as W_{s1} , W_{s2} , and W_{s3} . The sub-windows have a size of 3 and all the sub windows contain *x(5)*].

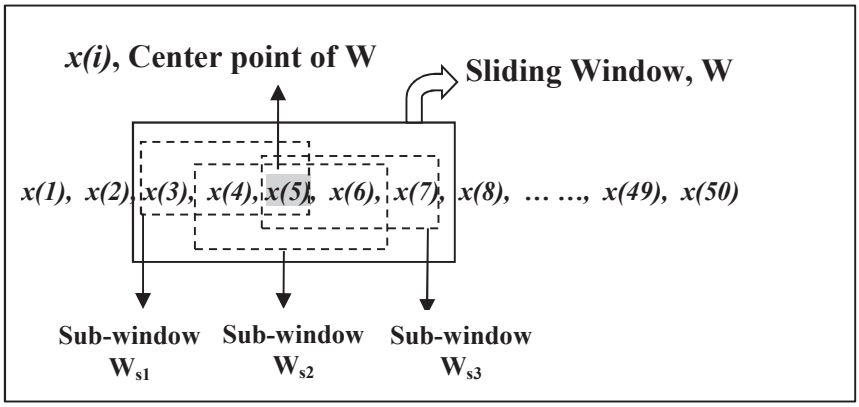


Figure 4.1 An example showing the positioning of sliding window and sub-windows for MVF

The sub-windows act as units for the search of homogeneity in the neighbourhood. Here the measure of homogeneity is the signal variance within the sub-window. The signal variance in each of the sub-window is obtained and the sub-window with minimum signal variance represents the most homogenous region in the neighbourhood. The mean value of the points in this most homogenous sub-window represents the noise free estimate of the center point in the filtered signal, denoted by *x'(i)*. Then the sliding window is shifted by one point to the right, to calculate the filtered value of *x(i+1)*. A new neighbourhood is defined by the window W, centered on the point *x(i+1)* and the process is repeated to obtain the filtered value. This process is continued until all the points in the sequence are filtered.

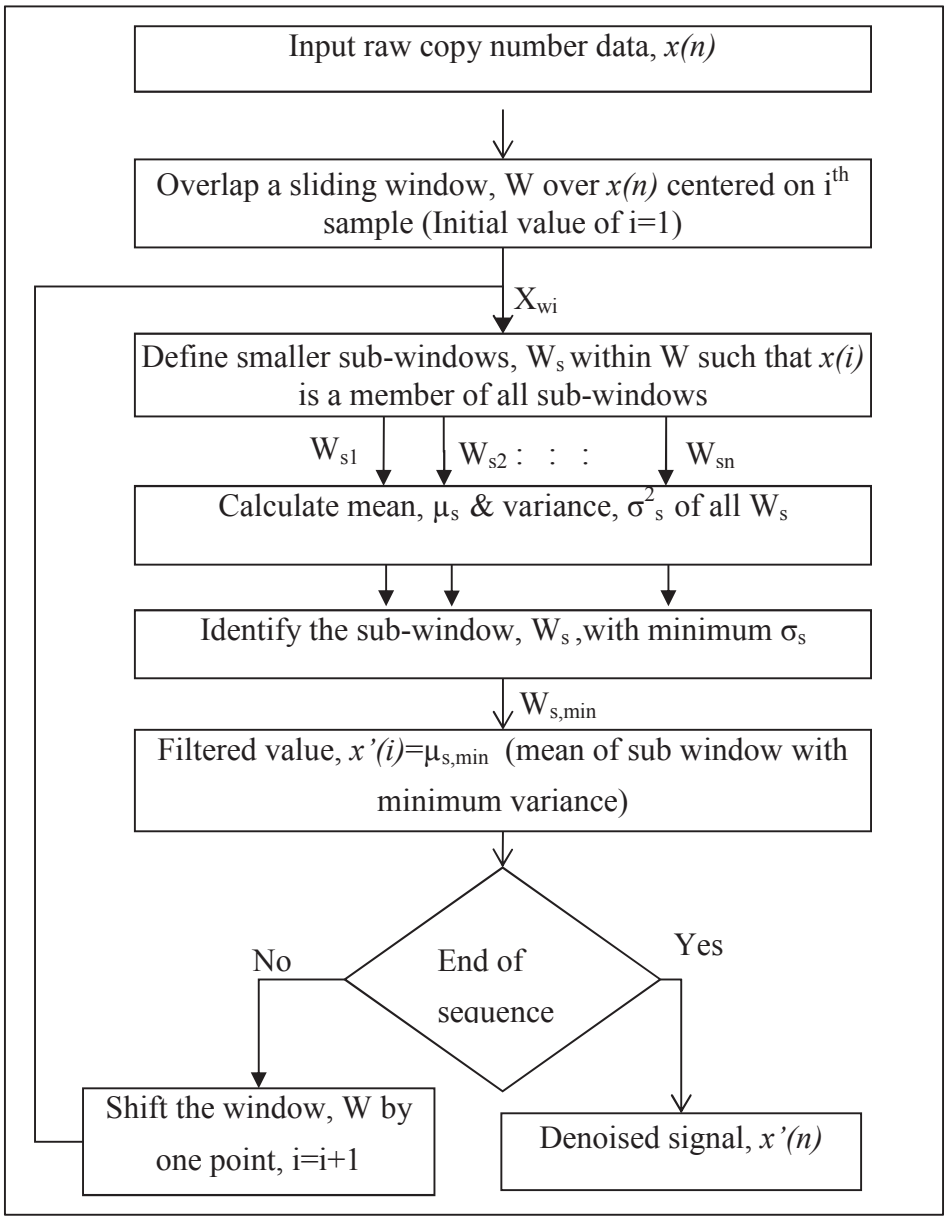


Figure 4.2 Minimum Variance Filtering algorithm

The MVF uses a parameter, β, which determines the smallest width of aberration that can be retained without smoothening out. The sub windows

defined within the neighbourhood has a size equal to β. The parameter, β, also decides the size of sliding window, *L*= 2β-1. The value of β is proportional to the smoothing effect and inversely proportional to the resolution of the method which is the smallest width of aberration that can be retained without degradation. A flow chart representation of Minimum Variance Filtering algorithm is shown in Fig 4.2.

The MVF algorithm also includes an option for iterative filtering with an automatic stopping criterion. The stopping criterion is implemented using a minimum value for mean squared difference (ε) between successive iteration result. After every iteration, the mean squared difference between the new noise free estimate and the previous estimate is calculated. When the difference becomes negligible ($\varepsilon \leq 0.001$, which is the stopping criterion), the iteration is stopped. The iterated MVF is particularly helpful in the case of highly noisy log ratio data. The effect of parameter, β is shown in Fig. 4.3 [a-d] and the effect of iterated filtering for a data with high noise content is shown in Fig. 4.4[a-c] in EES method implementation in section 4.5.1.

The mean squared error between successive iteration step is calculated as given below (Eq. 4.4), where, *x(n)* is the raw log ratio data, $x_j'(n)$ is the noise free estimate after j^{th} iteration and N be the length of the signal. Then the mean squared difference (ε) is given by,

$$\varepsilon = \frac{1}{N}\sum_{i=1}^{N}(x_j'(i) - x_{j-1}'(i))^2 \tag{4.4}$$

Stopping criterion : $\varepsilon \leq 0.001$.

The step by step description of the MVF algorithm is as follows:

1) Let *x(n)* be the noisy raw log ratio data
2) Select the parameter, β, which determines the minimum resolution of the filter.
3) Define a sliding window, W, of length $L = 2\beta-1$,
4) Extend the input data using border replication on both sides.
5) Overlap the sliding window, W, centered on the first data point.
6) The data point which is being processed denoted by *x(i)*.
7) Define smaller sub-windows (W_s) of length β within 'W' such that the point *x(i)* is contained in all sub-windows (Fig.4.1).
8) Calculate the signal mean and variance within each sub-window.
9) Identify the most homogenous sub-window ($W_{s,min}$) with minimum signal variance.
10) Obtain the mean signal value $\mu_{s,min}$ of the sub window, $W_{s,min}$ identified in step 8.
11) The new noise free estimate, *x'(i)* of the point *x(i)* is assigned the value $\mu_{s,min}$.
12) Shift the sliding window 'W' by one point and go to step 6 to obtain the noise free estimate, *x'(i+1),* for the next point.
13) Repeat until all points are processed.
14) For iterated filtering using MVF, Step 4 to 11 is repeated until the stopping criterion, $\varepsilon \leq 0.001$ is satisfied.

4.4.2 Segmentation using K-means clustering

Array CGH data are characterized by constant or flat regions with a finite number of instantaneous sharp jumps. The segmentation task in this scenario aims to find the location of the jumps which will allow the

discretization of the signal to different levels. This problem can also be considered as a step fitting problem. A running median filter was introduced by Beyer and Tukey (1981), in one of the earliest attempts to solve this change point detection problem. Several different approaches are reported in literature to solve the break point detection problem (Jong et al., 2003; Picard et al., 2005). In the presented method, the log ratio data is modeled as a gaussian random process with a parameter θ having M discontinuities. Here the parameter of interest (θ) is the mean (μ) value of a segment or a contiguous set of points in the log ratio data. The M edge points, partitions the data into M+1 segments, within which the parameter θ remains constant. The EES method uses a K-means clustering based approach to achieve this partitioning of denoised log ratio data.

K-means clustering is an algorithm to group objects or values based on attributes into K number of distinct group with the criteria of minimizing the sum of squares of the distance between data points and cluster centroids. For a given denoised data sequence, $X' = \{x'_1, x'_2, \dots x'_n\}$, the K-means clustering aims to group the n data values into K clusters, where (K≤ n) so that the following criteria is minimized.

Optimization criteria: $\text{argmin}_C(\sum_{i=1}^{K}\sum_{x' \in C_i}\|x' - m_i\|^2)$ (4.5)

where, $C = \{C_1, C_2, \dots C_k\}$, represents the clustering pattern and each cluster C_i is characterized by its centroid m_i and the set of points belonging to the cluster.

The criterion aims at selecting the pattern of clustering 'C' that minimizes the argument in the bracket which is the sum of squares of within cluster distance. The algorithm proceeds through the following steps.

Step 1: Determine the value of 'K', where K represents the number of clusters.

Step 2: Assign initial values for cluster centroids.

The centroids are represented as: m_i^t, where *'m'* stands for the mean of the cluster *'i'* in the 't'th iteration. The initial values of centroids are denoted by, $m_1^1, m_2^1, m_3^1, \ldots m_k^1$.

Step 3: Assign each data value to the cluster with closest centroid value.

Each data value 'x" is assigned to a cluster 'C_i' using the criteria

$$\text{argmin}_i(\|x' - m_i\|^2)\,,\ 1 \le \text{i} \le \text{K} \qquad (4.6)$$

The assignment task during the t[th] iteration can be represented as,

$$C_i^t = \{x' : \|x' - m_i^t\|^2 \le \left\|x' - m_j^t\right\|^2 \forall\, j\,, 1 \le j \le K\} \qquad (4.7)$$

Step 4: Recalculate the new centroids for each cluster.

The mean value of the points assigned to each cluster is calculated and selected as the updated value of centroid. The new centroid m_i^{t+1} is defined as,

$$m_i^{t+1} = \mu_i^t\ , 1 \le i \le K \qquad (4.8)$$

Step 5: Repeat the steps 3 & 4 until there is no change in the cluster assignment for any of the data points.

Step 6: Create two sets CC and CM from the final clustering of data points.

CC is the set cluster centroids defined as,

$$CC = \{m_1, m_2, m_3, \ldots m_K\} \qquad (4.9)$$

where, m_i , represents the mean value of cluster 'i'.

CM is a set that contains cluster membership information of all the data points and is defined as,

$$CM = \{c_1, c_2, c_3, \ldots c_n\} \quad (4.10)$$

where $'c_i'$ indicates the cluster to which the data element *x(i)* is assigned.

Step 7: Identify break points or location of significant changes in log ratio data.

The set of breakpoints B is defined as,

$$B = \{i: c_i \neq c_{i-1} \forall\, i\,, 2 \leq i \leq n,\ c_i \in CM\} \quad (4.11)$$

Step 8: Using the breakpoints in *B,* the denoised log ratio data, X', is divided into multiple segments.

Step 9: Every data point in a segment is then replaced with their mean to create the new segmented data, *Y*.

Discretization of the denoised log ratio data ($\log_2$(R/G)) is achieved in this step. The log ratio data is the logarithm to base 2 of the ratio of red fluorescence intensity (test) to green fluorescence intensity (reference) (Eq.3.1). As the data is log transformed, the value has a very small range and in most cases will be in the range -1 to 1. Hence the data points have close values. To quantize these points, clustering is done with an interval of 0.1. To implement this, the peak to peak variation of the data is obtained, and this amplitude range is divided into intervals of magnitude 0.1. The number of intervals with at least 2 data points is considered as a cluster to obtain the number of clusters, K. Thus the selection of the appropriate number of clusters for a given set of data points is purely a subjective task. Instead of selecting a single predetermined value for K, the value of K is obtained from the distribution of the log ratio values in the denoised data. The median value of the intervals becomes the initial centroids for the K-

means clustering algorithm. The K-means function treats each observation in the data as an object having a location in space. It then finds a grouping where, objects within a cluster are as close to each other as possible, and as far from objects in other clusters as possible. Each cluster in the partition is defined by the spatial position of the cluster members and its centroid. Thus the clustering process results in two output data. First one, CC, is a set of cluster means and the second one, CM, is a sequence representing the cluster membership information of every data point. The data is then divided into smaller segments with each segment having a magnitude equal to their mean value. To perform this segmentation, first the location of the breakpoints or segment boundaries must be determined. A continuous stretch of data points belonging to same cluster represent a segment. The positions in log ratio data where there is a change in cluster association or cluster membership represent a breakpoint or segment boundary. These breakpoints are obtained from the array CM, as described above in Eq.4.11. The set of breakpoints B contains the indices of elements in array CM where there is a change in value (change in cluster). These value changes indicate the boundary of a segment. Once the boundaries of segments are identified, data points in every segment are replaced by their mean value. This operation results in the conversion of data into discrete segments. The K-means clustering based segmentation thus partitions the input data into segments of discrete levels, corresponding to different copy numbers.

4.4.3 Copy number level assignment using thresholding

Copy number level assignment deals with the classification of the individual segments obtained by segmentation into a region with copy number loss, gain or normal value. Ideally the log copy number ratio will have a value of 0 for normal DNA regions with two copies of genome. The

aberrant regions may have values less than 0, for genome regions affected by loss of 1 or 2 copies. For regions of genome with duplication of 1 or more copies, the value will be greater than 0. The table 4.1 given below shows these changes in log ratio value for copy number changes due to deletion and duplication events. The copy number value for the discrete segments is calculated from the log ratio as given below,

$$\text{DNA copy number} = \mathbf{2^{1+Log_2(R/G)\ value}} \qquad (4.12)$$

where, (R/G) value is the ratio of red to green fluorescence intensity as described in section 3.4.6, Eq.(3.1). The status of a region of genome can be {loss, normal, gain} based on the log ratio value. The segmentation step achieves the segmentation of intensity ratio data into regions of constant values. A meaningful interpretation necessitates assigning copy number values and copy number status to these segments. The level assignment step aims to achieve this objective.

Table 4.1 Log ratio value and corresponding DNA copy number

Loss/Gain	**Log_2(R/G) value**	**DNA copy number**
No deletion/duplication	0	2
Single copy loss	-1	1
Single copy gain	0.58	3
Two copy gain	1	4
Three copy gain	1.32	5
Four copy gain	1.58	6

Thresholding

The thresholding technique offers a simple yet effective method by applying a reasonable threshold to the average ratio over a continuous segment of genome. The threshold acts as a decision function which determines whether the segment is normal, amplified or deleted. Nakao et al. (2004) have suggested a threshold value for chromosome amplifications and deletions to be defined as having a gain or loss. The threshold values for classification of regions is given below,

$|\text{Log}_2(R/G)| > |0.225|$ for aberration region

$|\text{Log}_2(R/G)| < |0.225|$ for normal region (4.13)

A median $|\text{Log}_2(R/G)| > 0.14$ for all clones on the arm represents a loss or gain involving the entire length of chromosome arm. The same thresholds are employed in the EES method as default, for classification of regions into normal and aberrant regions. Provision for allowing user defined thresholds for decision making is also incorporated in the EES method.

4.5 Implementation of the EES method for the detection of CNV

As mentioned in section 4.4, the input Array CGH log ratio data given by, $X = \{x_1, x_2, \ldots x_n\}$ is denoised using MVF in the first step of EES method to obtain noise free estimate, *X'*. The denoised data *X'* is then discretized to form the segmented data, *Y,* with a finite number of constant segments with abrupt breakpoints using a K-means clustering based approach. The individual segments of data, *Y,* are then assigned copy number values based on the segment mean values and are assigned a copy number status by applying threshold. Thus EES method performs the

detection and localization of copy number variation in the genome using the Array CGH log ratio data and returns the location of variations, their type (deletion or duplication) and the magnitude of variations.

An example of the raw log ratio data obtained from an Array CGH experiment is given in Table 4.2. The table shows the values corresponding to a small portion from chromosome 1 of the test genome. The first column indicates the chromosome number; the second and third column indicates the starting and ending position of the genomic loci corresponding to the spots in the array. The fourth column shows the $\log_2(R/G)$ value of the spot.

Table 4.2 An example of raw log ratio data.

Chromosome	**POS.start**	**POS.end**	**Log_2ratio**
1	786434	786839	0.63241
1	794658	795080	0.237744
1	816673	817242	-0.00393
1	818403	818847	-0.2785
1	829319	829630	-0.08125
1	973923	974304	-0.02255
1	973923	974304	0.448109
1	973947	974326	-0.14733
1	1011293	1011776	-0.24212
1	1026861	1027384	0.490703
:	:	:	:
:	:	:	:

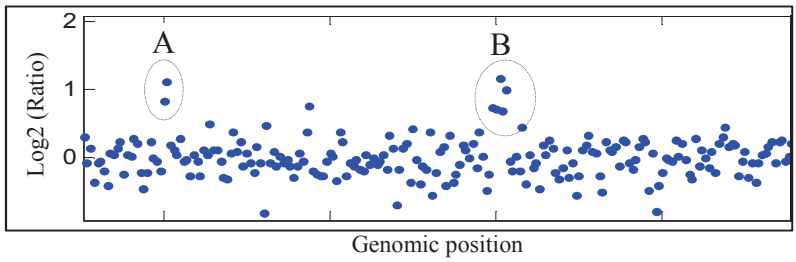


Figure 4.3 (a) An example of raw log ratio plot with CNV

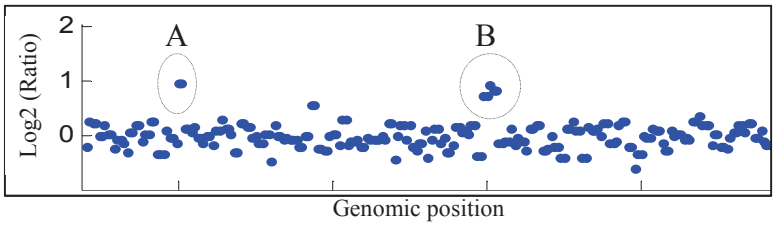


Figure 4.3 (b) Denoised data with β=3 & No. of iteration =1

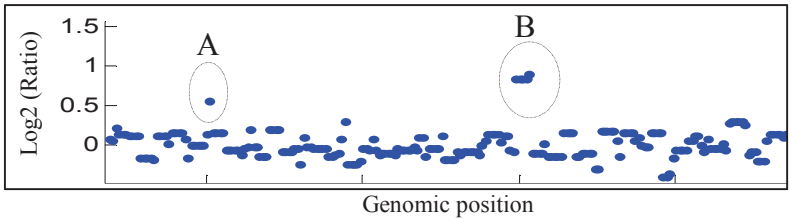


Figure 4.3 (c) Denoised data with β=7 & No. of iteration =1

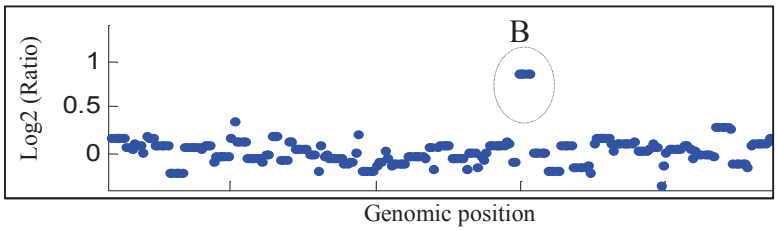


Figure 4.3 (d) Denoised data with β=9 & No. of iteration =1

A sample plot of another input data with CNV is shown in Fig 4.3a. There are 2 CNV in the sequence marked A, B with aberration width 2 and 5 respectively. The data in Fig 4.3a is then processed using the MVF to remove the noise in the signal and to enhance the edge points. The parameter, β, determines the degree of smoothing of signal. The denoised data corresponding the Fig 4.3a using MVF with β=3 is shown in Fig 4.3b. Fig 4.3c,d shows the result with β=7 and with β=9 respectively. It can be observed that the magnitude of the first CNV with width 2 is retained well with β=3. For β=7, the magnitude of the first CNV is reduced but it is still recognizable as the original aberration has high magnitude, which prevents

it from being totally smoothened out. For β=9, it can be seen that the CNV of width 2 is totally smoothened out. This highlights the effect of β, whereby the aberrations of width greater than β is guaranteed to be retained. In the case of aberrations with width less than β, the chances of retention depends on the magnitude of aberration and the values of neighborhood points. So the parameter, β, can be defined as the smallest width of aberration that will be retained by MVF without degradation.

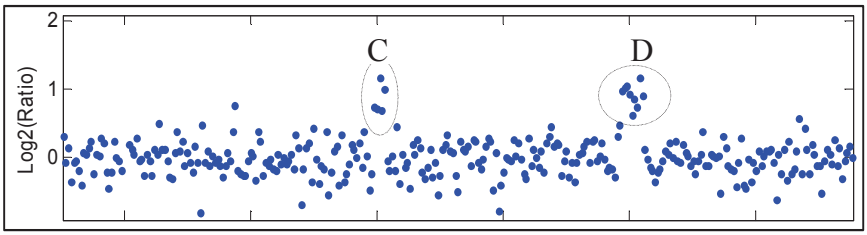


Figure 4.4 (a) Raw log ratio plot with 2 CNV of width 5 and 10.

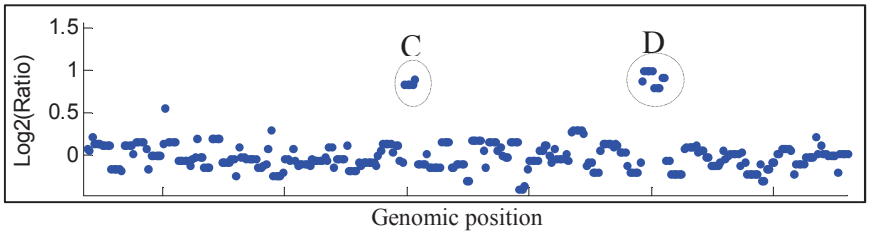


Figure 4.4 (b) Denoised data with β=9 & No. of iteration =1

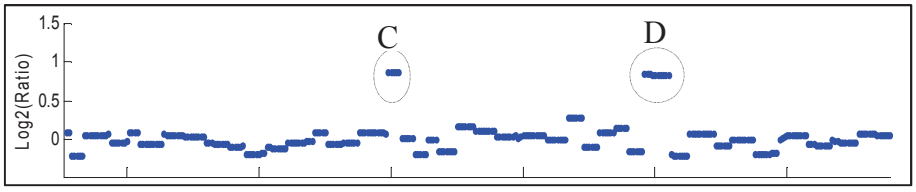


Figure 4.4 (c) Denoised with β=9 & iterated filtering with automatic stopping (No. of iteration =4)

The MVF also has an option for iterated filtering. The iterated filtering employs an automatic stopping criterion ($\varepsilon \leq 0.001$). The Fig 4.4a shows an example of a raw log ratio data with 2 CNVs with width 5 and 10.

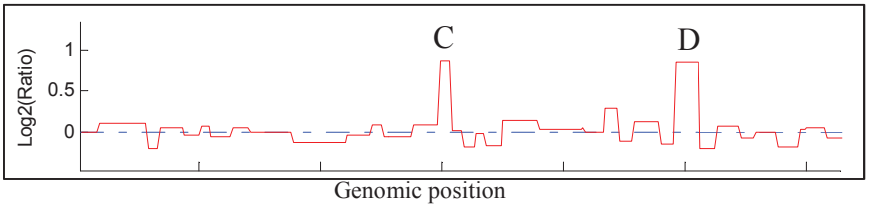


Figure 4.4 (d) Log ratio data after segmentation

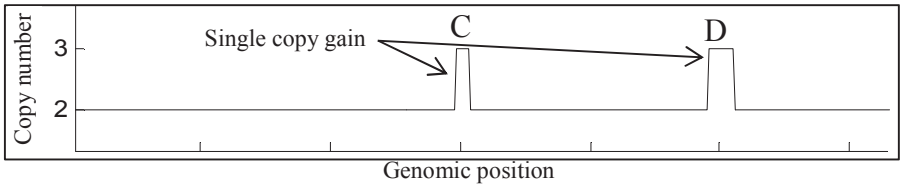


Figure 4.4 (e) Copy number status estimated by the EES method

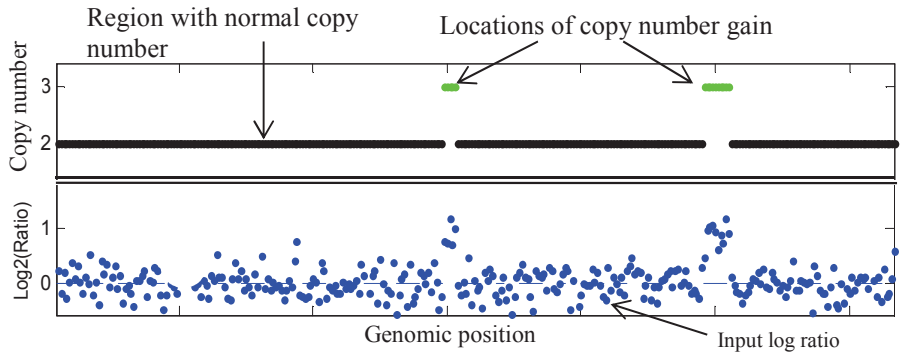


Figure 4.4 (f) Input log ratio data and copy number status estimated by the EES method

A comparison of the result of MVF with single iteration and iterated filtering with automatic stopping criteria is shown in Fig 4.4b and Fig 4.4c respectively. For this case, the stopping criterion is reached with four iterations. It can be seen that the iterated filtering results in a more

smoothened signal compared to single step filtering. A comparison of Fig 4.4a and Fig 4.4c clearly highlights the better detectability of the edges in the filtered signal, by making the transition sharper in the filtered data compared to the input data. The denoised data is then segmented into discrete levels using the K-means clustering based method described in 4.4.2. Fig 4.4d shows the result of segmentation step on the denoised data shown in Fig 4.4c. The segmentation process can be described as a process where the denoised data is quantized into discrete levels. In Fig 4.4d, the data is divided into segments with discrete magnitude. Each segment represents a continuous stretch of points belonging to a single cluster and has a magnitude equal to their mean value. This operation results in the conversion of data into a set of segments represented by their mean value. Finally, using Eq.4.12 each of the segments is assigned a copy number value and using a thresholding function they are assigned a status based on the log ratio value as given in Eq.4.13. The status can be one from the following set {'Loss', 'Normal', 'Gain'}. The Fig 4.4e shows the result of thresholding and shows the locations and magnitude of copy number variations. Fig 4.4f shows the input log ratio data and the corresponding copy number status estimated by the EES method. The blue dots show the distribution of input log ratio data. The region shown in green indicates the genomic loci with copy number gain and region shown in black indicates the genomic loci with normal copy number.

4.6 Results and Discussion

The EES algorithm is evaluated using standard simulation analysis and using real data. Three different performance analysis schemes are employed for simulation study as given below,

(i) Root Mean Square Error (RMSE) analysis.

(ii) Analysis of the resolution of detection.

(iii) Analysis of Receiver Operating Characteristics (ROC) and Area Under the Curve (AUC).

4.6.1 Generation of simulated data

The simulated data sets are created for the study. The dataset 1 is used to RMSE analysis and dataset 2 is used to study the effect of aberration width on the performance of EES method. Analysis of ROC and AUC is performed with the help of dataset 3.

Dataset 1

A new simulated data model that closely resembles the real Array CGH log ratio data is designed for RMSE analysis. Real Array CGH contains regions with normal copy number, with single copy loss, single copy gain and gain of multiple copies, represented by log ratio levels with amplitude, 0, -1, 0.58 and ≥1 respectively. Dataset 1 includes all these 4 levels and an additional level with amplitude -0.5 representing smaller loss. The sample data sequence is designed with a length of 100 and with 5 levels of amplitude representing the different levels discussed above. The data also has aberrations at both ends to evaluate the ability of the algorithm to detect aberrations at boundaries. The template for noise free simulated data thus designed is shown in Fig 4.5a. The data has a length of 100, contains multiple aberrations of widths 3 and 5. The data has 5 amplitude levels: -1,-0.5, 0.58 and 1. This forms a noise free data template, 'x' to which random noise of different levels is added to create noisy simulated data sequences. Gaussian noise of 3 different levels, ($\sigma = 0.1, 0.15$ & 0.2) are added to the template to generate noisy samples (Fig 4.5b-d) as given below.

The gaussian noise with noise standard deviation, σ, is represented by, 'n_σ'.

Eg: Noise with σ =0.1 is represented by $n_{0.1}$.

The noisy data sample, $y_{(\sigma)}$, with gaussian noise 'σ' is obtained by,

$$y_{(\sigma)} = x + n_\sigma \qquad (4.14)$$

For each noise level, a sample set of 20 noisy data sequences is generated as shown below.

$$Y_{(\sigma)} = \{y_{(\sigma)}^1, y_{(\sigma)}^2, y_{(\sigma)}^3 \ldots, y_{(\sigma)}^{20}\} \qquad (4.15)$$

where $Y_{(\sigma)}$ represents a set of 20 samples of noisy data with noise standard deviation 'σ'.

The three sample sets, $Y_{(0.1)}$, $Y_{(0.15)}$, $Y_{(0.2)}$ together constitutes the simulated dataset 1 for RMSE analysis.

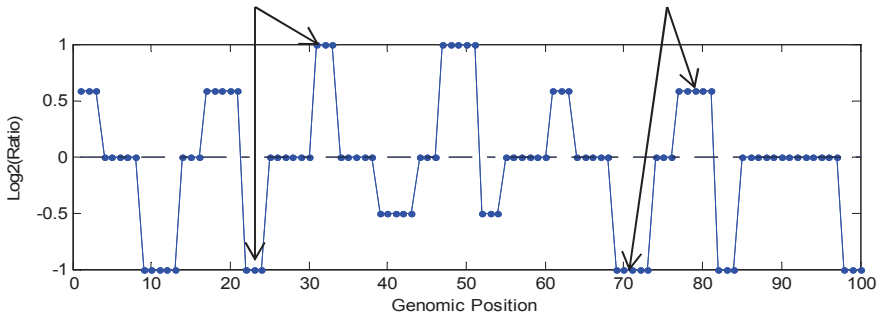


Figure 4.5a Noise free simulated data for RMSE analysis

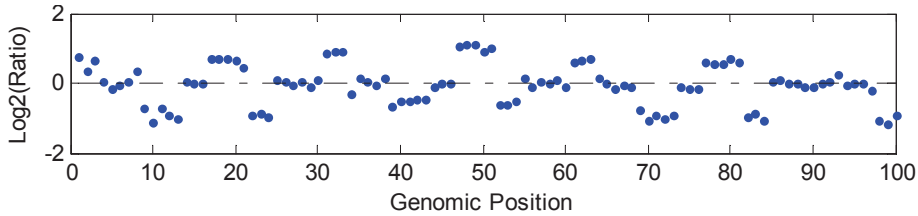


Figure 4.5b Noisy data with σ =0.1

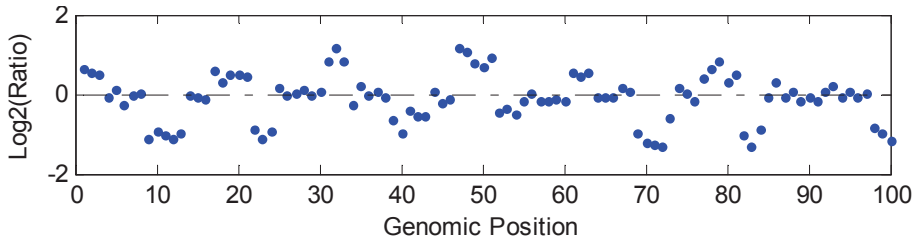


Figure 4.5c Noisy data with σ =0.15

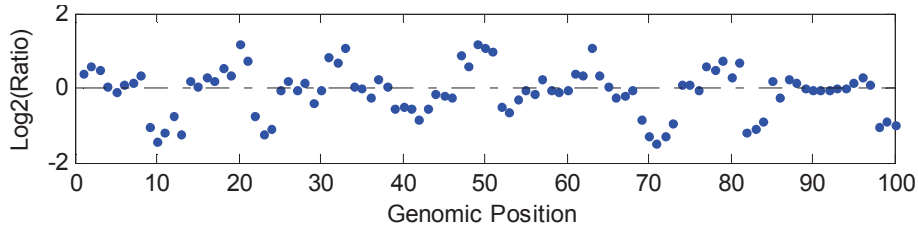


Figure 4.5d Noisy data with σ =0.2

Dataset 2

This simulated data involves a sequence of length 500 with aberrations of increasing width. Five data templates with each having a length of 100 and one aberration at the center, is created. Aberrations of increasing widths of 2, 5, 10, 20 and 40 are inserted at the center of the different templates. The five templates are joined together to obtain the simulated noise free data of length 500, containing a total of 77 aberrant data points or probes of unit amplitude in 5 aberration locations. The noise free data thus created is shown in Fig 4.6. Different levels of gaussian noise with noise standard deviation, $\sigma = 0.1, 0.15, 0.2, 0.225$ & 0.25 are then superimposed on the data to obtain noisy simulated data, $Y_{(0.1)}$, $Y_{(0.15)}$, $Y_{(0.2)}$, $Y_{(0.225)}$ & $Y_{(0.25)}$. The dataset 2 is used to study the effect of aberration width on the performance of the EES algorithm and to evaluate its resolution of detection.

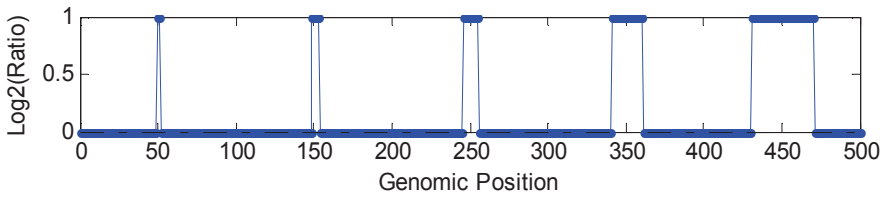


Figure 4.6 Noise free simulated data for studying the effect of aberration width

Dataset 3

The noise free simulated data generated for ROC analysis is similar to the noise free template of dataset 1 with a length of 100, containing 5 signal levels. For ROC analysis, gaussian noise of 3 different levels, ($\sigma = 0.25, 0.5$ & 1) are used. Noisy simulated data, $Y_{(0.25)}, Y_{(0.5)}$ & $Y_{(1)}$ are created by adding these noise to the noise free template (Fig.4.5e-g). Compared to the dataset 1, higher levels of noise are used here to study the worst case scenarios.

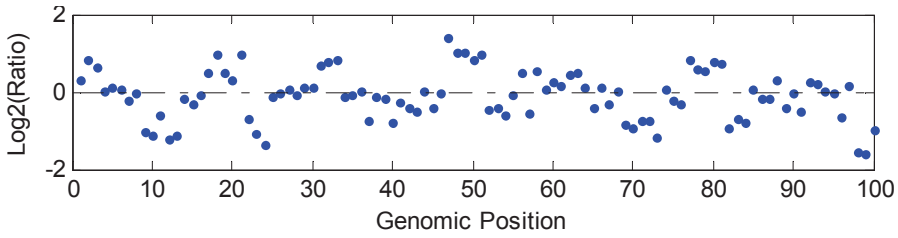


Figure 4.5e Noisy data with σ =0.25

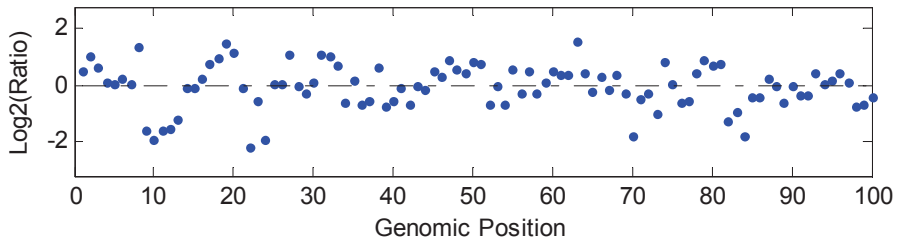


Figure 4.5f Noisy data with σ =0.5

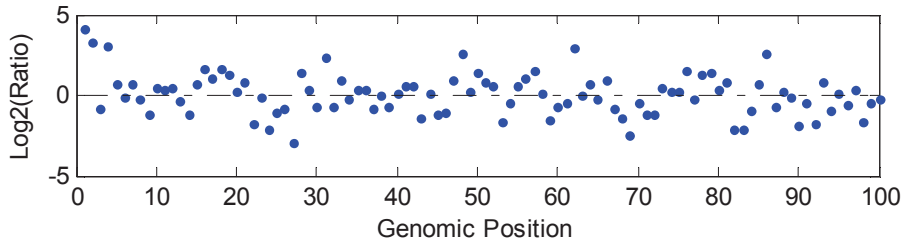


Figure 4.5g Noisy data with σ =1

4.6.2 Real Array CGH dataset

Dataset 4

Array CGH log ratio data from several real samples are obtained and analyzed. An example of real Array CGH data (Neve et al., 2006) from chromosome 17 of breast cancer cell line sample, BT474 is used here for illustrating the performance of EES method. It includes the genomic loci from 2484 Mb to 2566 Mb of BT474 genome. This constitutes the dataset 4. A detailed analysis of real Array CGH data using the EES method is presented in the next chapter.

4.6.3 Root Mean Square Error analysis

RMSE can be described as a measure of the difference between the actual observed values and the values estimated by a model. It quantifies the estimation accuracy of the method. The noisy Array CGH data (*y(t)*) can be considered as a mixture of the original noise free data (*x(t)*) with additive gaussian noise (*n(t)*).

$$y(t) = x(t) + n(t) \qquad (4.16)$$

The noise filter can be considered as a model which produces an estimate $\hat{x}(t)$ of the original noise free data from the noisy input data. The

RMSE is a measure of the difference between the original data, *x(t)* and the noise free estimate, $\hat{x}(t)$ of the model and is calculated as shown below.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{x}(i) - x(i))^2} \tag{4.17}$$

where $x(t)$ is the original noise free data, N is the length of the data $x(t)$ and $\hat{x}(t)$ is the filter output which is an estimate of the noise free data.

RMSE represents the estimation error of the MVF method used in the EES algorithm. RMSE value for the MVF is calculated using the dataset 1, consisting of sample set with three different noise levels, described in section 4.6.1. A noisy simulated data sample, $y_{(\sigma)}$ is obtained and denoised using the MVF filter to obtain an estimate $\hat{x}$ of original noise free data, *x*. Using this estimate, the RMSE value is calculated. Similarly, RMSE values for all the 20 sequences corresponding to a noise level is obtained and the mean value for these 20 sequences in a sample set is taken as the RMSE corresponding to a noise level as shown below.

$$RMSE_{Y_{(\sigma)}} = \frac{1}{20}\sum_{i=1}^{20} RMSE_{y^i_{(\sigma)}} \tag{4.18}$$

Lowess (Beheshti et al., 2003), quantreg (Eilers and De Menezes, 2005) and wavelet (Hsu et al., 2005) are three widely used denoising methods. The sequences in dataset1 are denoised using these three methods with the help of CGHweb tool (Lai et al., 2008). RMSE for all the 3 noise levels are calculated using the above approach. Fig 4.7 [a-d] shows the noisy data $y_{(0.1)}$ and the estimated noise free data for different methods. The

input noisy data is shown as blue dots and the filtered value is shown by black lines. The figure shows that the MVF method tracks the data more closely than other methods and has the least estimation error. Meanwhile the lowess method smoothens all the edges and hence produces the largest error in estimation. When the noise content increases, the RMSE of all the filters increases accordingly.

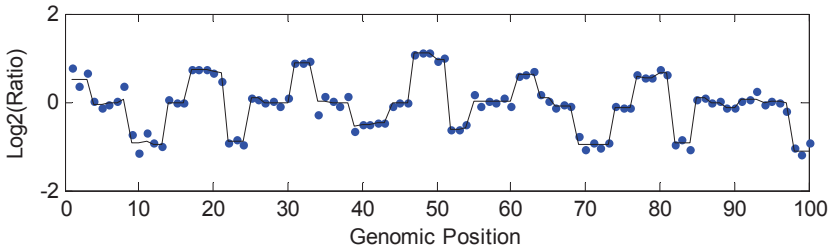


Figure 4.7a Result of MVF filter for simulated dataset 1 with noise σ = 0.1

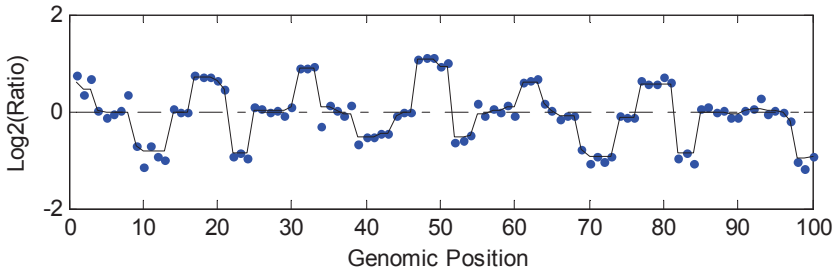


Figure 4.7b Result of quantreg method for simulated dataset 1 with noise σ = 0.1

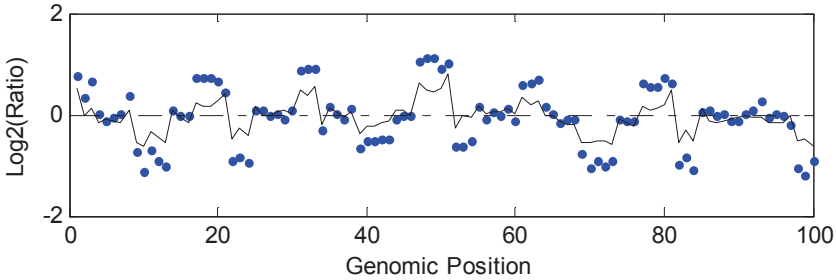


Figure 4.7c Result of wavelet method for simulated dataset 1 with noise σ = 0.1

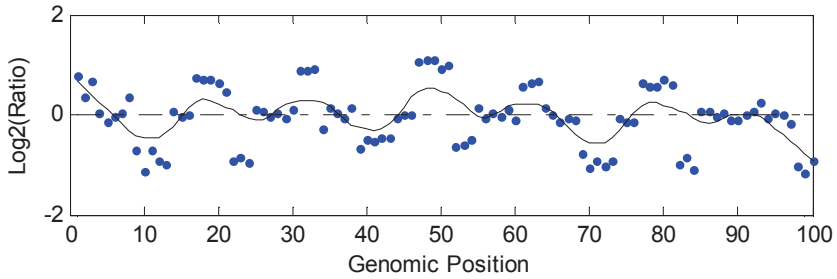


Figure 4.7d Result of lowess method for simulated dataset 1 with noise σ = 0.1

Table 4.3 Quantitative comparison of RMSE values of different filtering methods

Noise std dev(σ) →	**σ = 0.1**	**σ = 0.15**	**σ = 0.2**
MVF	0.0786	0.1149	0.1671
Quantreg	0.0933	0.1257	0.1759
Wavelet	0.3279	0.3567	0.3693
Lowess	0.4164	0.4171	0.4221

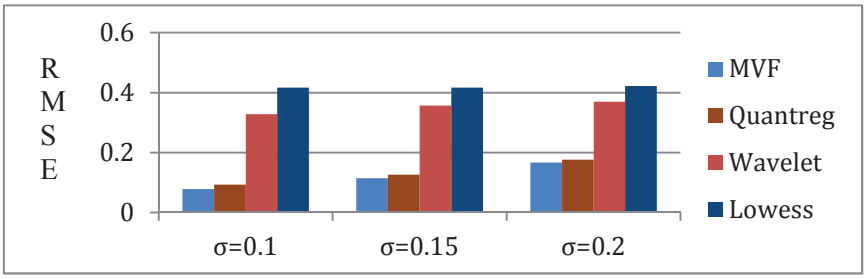


Figure 4.8 Comparison of RMSE values of MVF with other methods.

The table 4.3 shows a comparison of RMSE obtained for different denoising methods. It shows that the MVF filter has the least RMSE for all the three noise levels. The MVF offers 5% to 15.75% improvement in RMSE compared to quantreg method for a noise standard deviation 0.2 to 0.1. Similarly MVF offers a performance improvement of 54% to 76% over

wavelet method and 60% to 81% over lowess method (Table 4.4). A comparison of RMSE of the different filtering methods is shown using bar graph, in Fig 4.8.

Table 4.4 Percentage improvement in RMSE value when using MVF compared to other methods.

Noise std dev(σ) ⟶	**σ = 0.1**	**σ = 0.15**	**σ = 0.2**
Quantreg	15.75%	8.6%	5%
Wavelet	76%	67.78%	54.75%
Lowess	81.1%	72.45%	60.4%

4.6.4 Analysis of the resolution of detection

The resolution of detection of the EES method is defined as the smallest width of aberration that can be retained without degradation and is assured to be detected. The performance of the EES algorithm in detecting aberrations of different widths is observed here. The simulated dataset 2 described in section 4.6.1 is used. The sample data is of length 500 and has aberration of 5 different widths, superimposed with 5 levels of gaussian noise. The simulated data is processed using the EES method for the detection of aberrations. The results showed that the EES method detects all the five aberrations irrespective of the superimposed noise. In the simulated data there were 77 data points/probes within the aberration region. The EES algorithm identified all the 77 probes correctly for the data with σ = 0.1 & 0.15 and detected 76 probes for the data with σ = 0.2, 0.225 & 0.25. The

algorithm did not produce any false positives for all the noise level except σ = 0.25, where 1 non aberrant probe is classified as a FP.

Table 4.5 No. of TP & FP probes identified by the EES algorithm

Noise level →	**σ = 0.1**	**σ = 0.15**	**σ = 0.2**	**σ = 0.225**	**σ = 0.25**
True Positives	77	77	76	76	76
False Positives	0	0	0	0	1
Actual Positives	77	77	77	77	77

Table 4.6 No. of TP & FP probes obtained by different methods for σ = 0.25.

Noise level (σ=0.25)	**EES**	**CGHseg**	**CBS**	**Quantreg**	**Wavelet**	**Lowess**
True Positives	76	76	74	73	74	73
False Positives	1	1	1	1	1	3
Actual Positives	77	77	77	77	77	77

The table 4.5 lists the number of True Positive (TP) and False Positive (FP) probes identified by the EES algorithm. A TP result is one that detects an aberration when it is actually present and a FP is one that detects an aberration when it is actually normal. The simulated dataset 2 is also processed with five other methods for comparison. The methods are lowess, wavelet, quantreg, CBS (Olshen et al., 2004) and CGHseg (Picard et al., 2005). A comparison of the results obtained for various methods for a simulated data with noise σ = 0.25 is shown in Fig 4.9 [a-f]. Fig.4.9a shows

the result of EES method, where all the five aberrations are successfully identified. The EES method could detect the smallest aberration with a width of two data points even from the noisy data with σ = 0.25. The table 4.6 shows a comparison of the number of True Positive (TP) and False Positive (FP) probes identified by various methods for a simulated data with noise σ = 0.25.

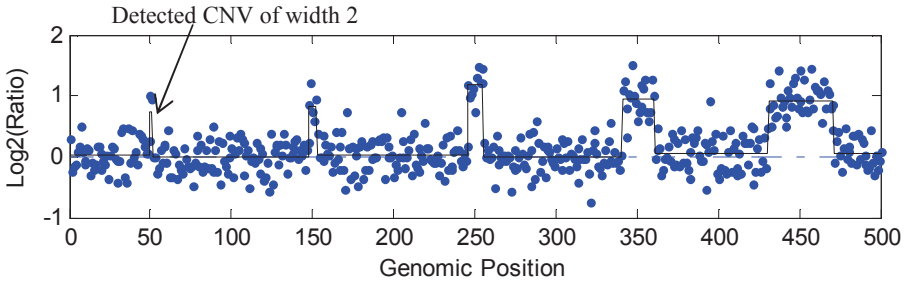


Figure 4.9a Result of EES algorithm for simulated dataset 2 with σ = 0.25

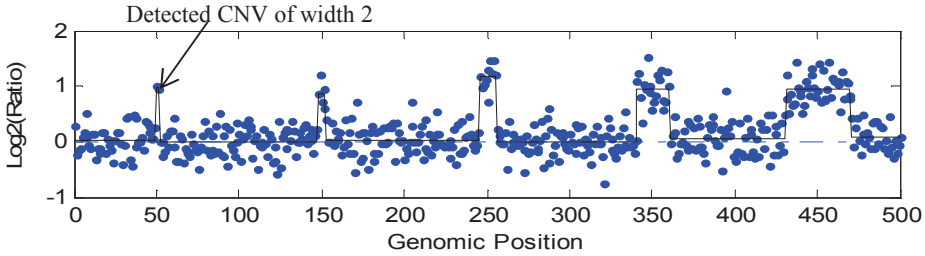


Figure 4.9b Result of CGHseg algorithm for simulated dataset 2 with σ = 0.25

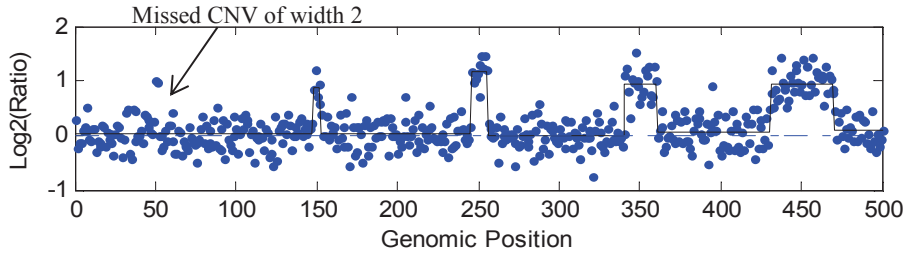


Figure 4.9c Result of CBS algorithm for simulated dataset 2 with σ = 0.25

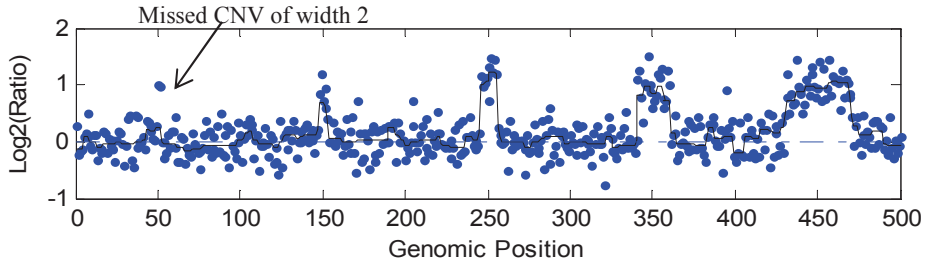


Figure 4.9d Result of quantreg algorithm for simulated dataset 2 with σ = 0.25

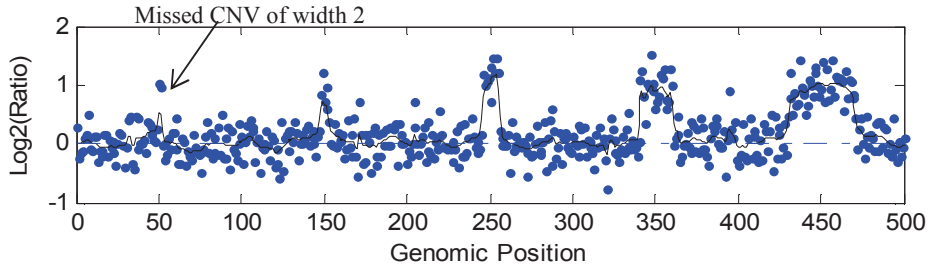


Figure 4.9e Result of wavelet algorithm for simulated dataset 2 with σ = 0.25

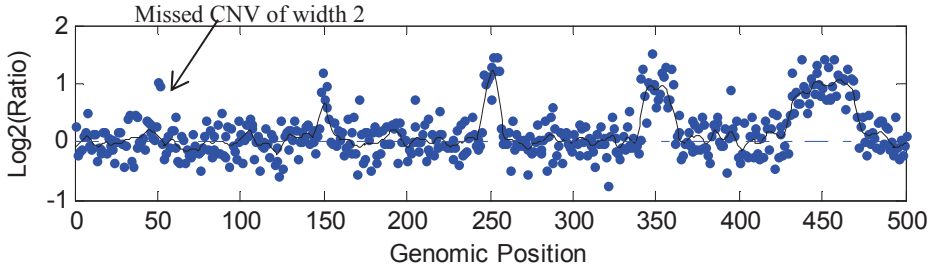


Figure 4.9f Result of lowess algorithm for simulated dataset 2 with σ = 0.25

All the methods detected aberrations of larger widths but, apart from the EES algorithm, only CGHseg could detect the smallest one. The smoothing algorithms like lowess, quantreg and wavelet failed to detect aberration of width 2 as it get smoothened out. This clearly highlights the effectiveness of MVF employed in the EES algorithm. In this study, the MVF was implemented with β=3 and without iteration. The MVF performed the denoising of data without degrading the aberration boundaries and thus enhancing the detectability of even smaller copy

number changes. The analysis of the effect of aberration width on the performance of the EES method illustrated the ability to detect CNVs with a resolution of two data points.

4.6.5 Analysis of Receiver Operating Characteristics (ROC) and Area Under the Curve (AUC)

ROC analysis is the most effective method for comparing the performance of different algorithms in separating aberrant and non-aberrant regions in an Array CGH profile. The ROC curve is also known as a relative operating characteristic curve as it facilitates a comparison of two operating characteristics, the True Positive Rate (TPR) and the False Positive Rate (FPR). The ROC curve is drawn by plotting the TPR against the FPR at various thresholds. The TPR is also known as sensitivity and the FPR can be calculated as (1 - specificity). Thresholds are fixed to classify every probe as either aberrant or normal. A log ratio value greater than the threshold classify the corresponding probe as aberrant and vice versa. TPR is defined as the ratio of, number of probes/data points inside the aberration whose estimated values are above the threshold level, to the actual number of probes in the aberration. FPR is defined as the ratio of, number of probes outside the aberration whose estimated values are above the threshold level, to the total number of probes outside the aberration. The equation for calculating TPR and FPR are given below,

$$TPR = \frac{No.of\ probes\ inside\ aberration\ with\ estimated\ value > threshold}{Total\ no.of\ probes\ inside\ aberration}$$

$$FPR = \frac{No.of\ probes\ outside\ aberration\ with\ estimated\ value > threshold}{Total\ no.of\ probes\ outside\ aberration} \quad (4.19)$$

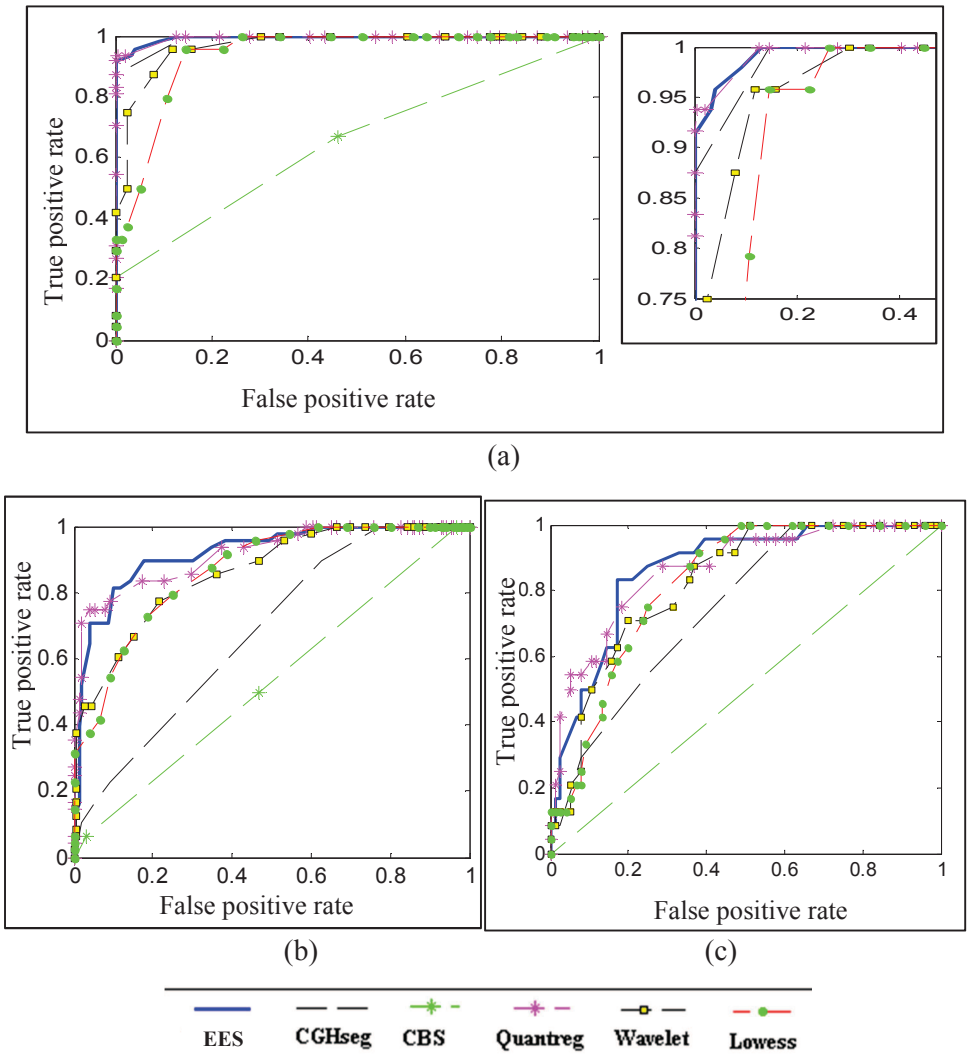


Figure 4.10 ROC plot for simulated dataset 3 with (a) σ = 0.25 (b) σ = 0.5 (c) σ = 1

To plot the ROC curve, the threshold value for aberration is varied from the minimum log-ratio value to the maximum. For each threshold value, the algorithm classifies some of the points as aberration and the remaining as normal. Based on this classification, a TPR and FPR are calculated for the particular threshold value. This pair of value is

represented by a point in the 2-D space. The set of TPRs and FPRs obtained for every threshold are then plotted to obtain the ROC profile of the algorithm for a particular aberration width and SNR. The analysis allows to study the tradeoff between specificity and sensitivity.

The simulated dataset 3 generated as described in 4.6.1 is used for ROC analysis. The noise free data template consists of probe sequence of length 100 with 5 levels of amplitude. Three levels of gaussian noise with, $\sigma = 0.25, 0.5 \,\&\, 1$ are added to the template to generate noisy samples, $Y_{(0.25)}, Y_{(0.5)}, Y_{(1)}$. The data Y is then processed by the EES method and as described above the threshold for classifying the region into an aberration is gradually varied from the -2 to 2. The TPR, FPR value for all thresholds are calculated and plotted on the ROC curve.

The same simulated data is applied to methods CGHSEG, CBS, quantreg, wavelet, and lowess and results are observed. The ROC curves for these methods are also plotted. The ROC curves for all noise levels are generated and are shown in Fig 4.10 [a-c]. Fig 4.10a shows the result for data with σ = 0.25. It can be noted that the EES method, quantreg and CGHseg offers similar performance at this noise level. But as the noise level increases, (Fig 4.10b,c) , the performance of CGHseg get degraded compared to the other two methods. This can be attributed to the fact that CGHSeg does not involve denoising step which is affecting its performance at higher noise levels. The same can be said about the CBS method which is the poorest among all the five methods compared. For all the three noise levels, EES algorithm shows the best performance followed by quantreg. For lower noise content or for signals with high SNRs, all the tested algorithms except CBS exhibited similar level of performance. So for the

ROC analysis, simulated data is formed with higher noise content ($\sigma > 0.25$). The results showed that the EES method results in more accurate classification of regions into normal and aberrant regions.

The Area Under the Curve (AUC) is a metric that can be used for comparing the ROC curves with AUC=1 being ideal and AUC=0.5 being the poorest. The AUC for all the six methods is calculated and is shown in Table 4.7.The AUC demonstrates the ability of the test to correctly classify the aberrant and non-aberrant regions. The AUC value for EES method is higher than all other method, with quantreg method showing closely similar capability (Table 4.7). The results obtained from the ROC curves and AUC metric clearly highlight the effectiveness of the EES algorithm in classifying probes in a noisy Array CGH data into aberrant and normal regions compared to other methods.

Table 4.7 Comparison of AUC for different algorithms.

	$\sigma = 0.25$	**$\sigma = 0.5$**	**$\sigma = 1$**
EES method	0.995	0.922	0.867
CGHSEG	0.99	0.7	0.74
CBS	0.651	0.53	0.5
Quantreg	0.995	0.917	0.864
Wavelet	0.968	0.86	0.82
Lowess	0.94	0.86	0.82

4.6.6 Analysis using real Array CGH data

Chromosome 17 is one of the smallest and most densely gene loaded human chromosome and is frequently rearranged in human breast tumors. The q-arm or the long arm of chromosome 17 is known to harbor complex combinations of gains and losses.

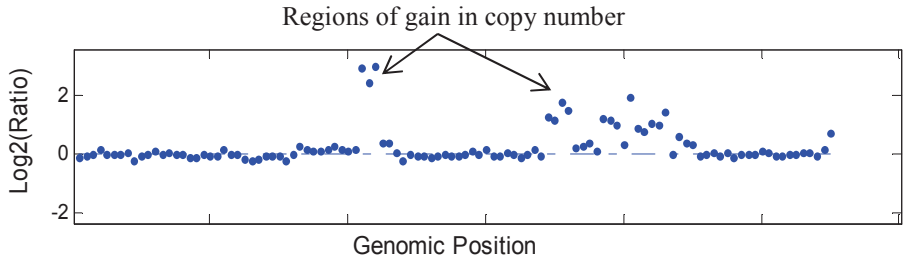


Figure 4.11a Log ratio plot of BT474/chromosome 17

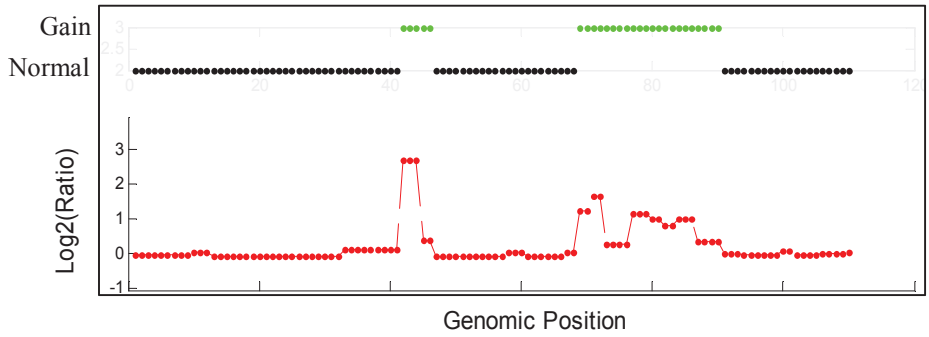


Figure 4.11b CNV observed using EES method in BT474/chromosome 17

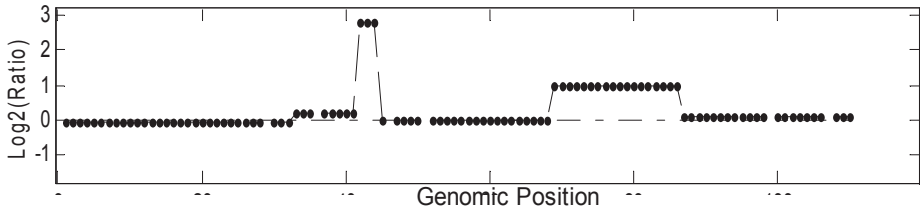


Figure 4.11c Result of CBS for BT474/chromosome 17

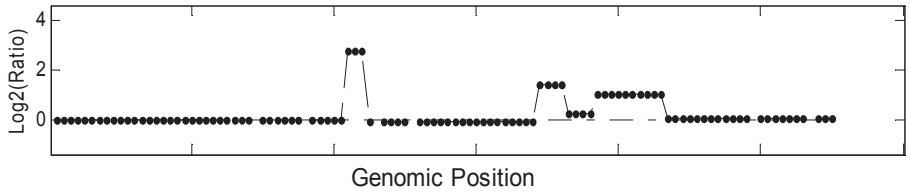


Figure 4.11d Result of CGHseg for BT474/chromosome 17

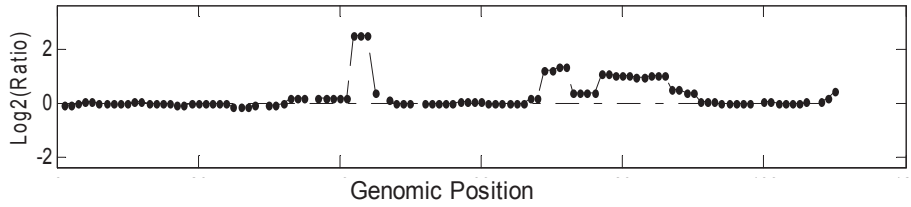


Figure 4.11e Result of quantreg for BT474/chromosome 17

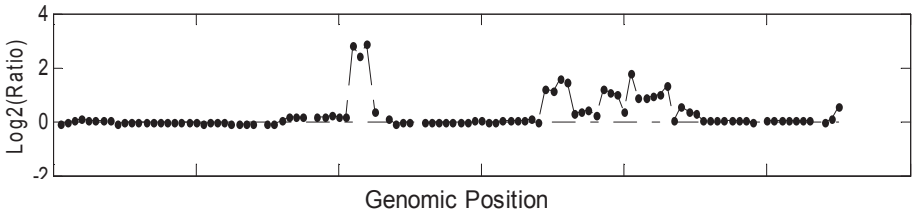


Fig 4.11f Result of wavelet for BT474/chromosome 17

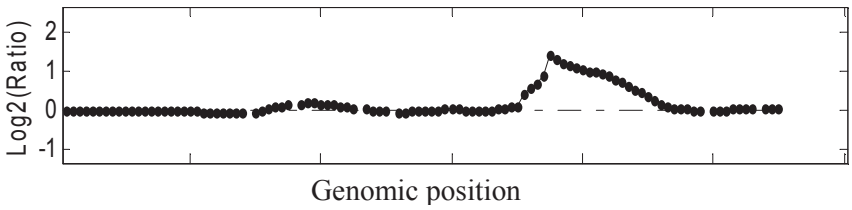


Fig 4.11g Result of lowess for BT474/chromosome 17

The log ratio obtained for the chromosome 17 of BT474 is shown in Fig 4.11a. The data is processed using the EES method and the result obtained is shown in Fig 4.11b.The result obtained using the other algorithms are given in Fig 4.11[c-g].

It can be observed from the figures that all the methods detected the two regions of copy number gains in the chromosome, but the resolution of the detection is different. CBS method shows the second amplification as a single aberration of same amplitude. CGHseg and Quantreg show it as three amplifications of different amplitudes. The wavelet method and the EES method predict a finer breakdown of the region. But the result of wavelet method is difficult to interpret compared to the other method. The EES method detects the entire region as amplified, but clearly divides the region into smaller segments with different levels of amplification as reported by Orsetti et al. (2004).

4.7 Summary

A new method for the detection and localization of CNV from Array CGH data is presented. The new method called EES method, achieves the objective of CNV detection using an edge enhancing filtering, K-means clustering based segmentation approach and threshold based copy number status assignment. The algorithm performs the denoising using a Minimum Variance Filter which is based on the search for local homogeneity in small neighborhoods. A K-means clustering based approach is used to segment the log intensity ratio data into discrete segments corresponding to different copy number values. The segmented levels are then categorized as normal or aberrant regions using thresholds. A comparison of RMSE of the MVF with other smoothing methods is performed and is found superior. The MVF offered 5% to 15.75% improvement in RMSE compared to Quantreg method, 54% to 76% over Wavelet method and 60% to 81% over Lowess method. The method also proved better in detecting smaller aberrations of width as small as 2 data points which is made possible by the better edge

retention capability. The ROC analysis is performed and AUC values for different ROC curves are also calculated. The results obtained from the ROC curves and AUC metric clearly highlighted the effectiveness of the EES algorithm in classifying a probe into aberrant or normal region while using simulated data. To study the performance of the new EES method on real data, the analysis was extended to real Array CGH data, and found to be the best compared to other five methods, to detect CNV. A detailed analysis of copy number variations in real Array CGH data using the EES method is described in the next chapter.

Chapter 5

Application of EES method on Array CGH data of human cell lines

This chapter demonstrates the application of the EES method for the analysis of real Array CGH data for the detection and localization of copy number variations. The three cell lines tested are: Coriell cell line, Breast cancer cell line and Glioblastoma multiforme cell lines. The analysis demonstrates the capability of EES method in clinical applications to detect copy number alterations in genome.

5.1 Introduction

The EES algorithm presented in the previous chapter is applied on real Array CGH datasets for the analysis and detection of copy number aberrations to gain clinically beneficial information. The algorithm is applied to three real data sets. The data sets used here are diploid cell lines obtained from human. The data sets are: (i) Coriell cell line data obtained from fibroblast cell lines (Snijders et al., 2001), (ii) Breast cancer cell line data (Neve et al., 2006) and (iii) Glioblastoma Multiforme data (Bredel et al., 2005).

5.2 Real Array CGH datasets from human

5.2.1 Coriell cell line data

The first real data set selected for analysis is the Coriell cell line BAC Array CGH data, described in Snijders et al. (2001). It is widely used as a standard data set for evaluating the performance of Array CGH algorithms as the exact locations of aberrations are already known. It consists of data from 15 fibroblast cell strains containing single copy aberrations whose locations are cytogenetically mapped. The normalized log ratio data for the cell line is obtained with CGH array using BAC clones. The data set includes only single copy aberrations called monosomies and trisomies. Monosomy is a condition where a portion or entire chromosome has only one copy and trisomy is a condition where it is present in three copies. Since the variations are already mapped ones, the true copy number changes are known for these cell lines for validating the results of this analysis. The data set also have one more advantage that the noise levels are low and the variations are strong and involve only single copy changes.

There are totally 23 aberrations identified by spectral karyotyping in the data set (Snijders et al., 2001).

5.2.2 Breast cancer cell line data

The second dataset selected for analysis is a Breast Cancer Cell Line (BCCL) database. Neve et al. (2006) has developed a model system to study the breast cancer cell lines, which exhibit the same heterogeneity in copy number and expression abnormalities as the primary tumors. The system includes CGH array data consisting of 51 breast cancer cell lines representing the recurrent features of 145 primary breast cancer cases. The CGH array was generated using BAC clones to obtain a data set with resolution of 1Mb. This dataset do not have an accepted or proven set of known aberration locations. Analysis of gene expression and genome copy number in these cell lines has identified 66 candidate therapeutic target genes. The dataset shows that 55 out of the 66 genes are amplified and over expressed in at least one cell line in the dataset. Here an attempt to identify these 55 amplified genes using EES algorithm is made to validate the capability of the algorithm.

5.2.3 Glioblastoma multiforme data

The Glioblastoma Multiforme (GBM) is the most commonly occurring form of brain tumor. The Array CGH data for GBM samples employed in the analysis is obtained from Bredel et al. (2005). It involves a high resolution genome-wide mapping of alterations using Array CGH methodology to profile copy number variations across 42000 mapped human cDNA clones in 54 gliomas of varying tumor grade. The analysis also identifies recurrent pattern of aberrations and various mutual relationships in different types of gliomas and performs the characterization

of aberrations involved in glioma-genesis. The GBM data set is very noisy, making the aberration detection complicated. It also involves a large variety of copy number variations like low amplitude gains, losses and regions with multiple gains. Recurrent patterns of distinct chromosomal aberrations and interrelationships of several alterations are also observed in the analysis. The observations and results of Bredel et al. (2005) forms a basis for validating the results obtained using the EES algorithm.

5.3 Results & Discussion

5.3.1 Application on Coriell cell lines

The data consists of 15 cell strains containing partial as well as whole chromosome aneuploidies. The array consists of 2460 BAC and P1 clones covering the 22 autosomal chromosomes and 2 sex chromosomes. Cytogenetic mapping has identified 23 locations of aberrations in the dataset (Table 5.1a). The microarray study performed by Snijders et al. (2001) on the dataset have identified and listed 22 out of the 23 aberrations and reported no evidence for an aberration in chromosome 15 of GM07081 in the Array CGH data. Using the EES algorithm all the chromosomes in each of the 15 cell lines is tested and the copy number variations are observed. The complete list of copy number variations identified by the algorithm is shown in Table 5.1a. The table also lists the locations of the cytogenetically mapped aberrations on the 15 cell lines and the type of variation (monosomy or trisomy). The CNVs detected by the EES method are shown in Fig 5.1 a-u.

[E.g. Fig 5.1a shows the profile for chromosome 3 of the sample-GM03563, represented as GM03563/3. The points shown on the profile as blue dots are the input $\log_2$ratio values for the clones or spots on the array

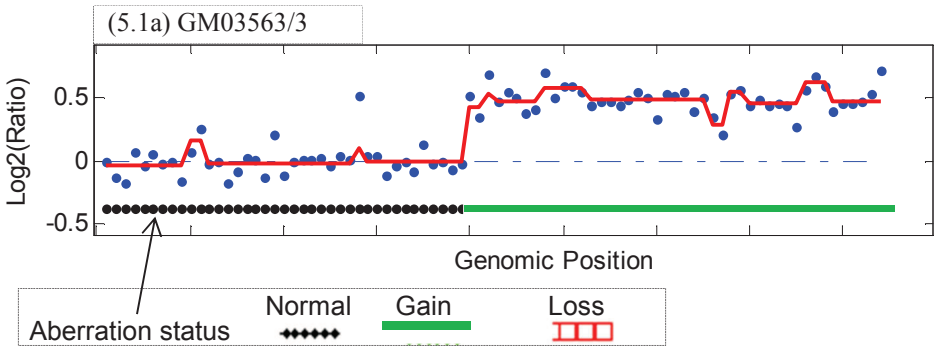
corresponding to the chromosome 3. The red line indicates the copy number value estimated by the EES method in the same logarithmic scale. The CNV status across the genomic positions is shown at the bottom of the plot where the status can be a loss, gain or normal indicated by red, green and black colours respectively. The EES method detected the region from 3q12 to 3qter (qter means end of q arm) as having a single copy gain or trisomy.]

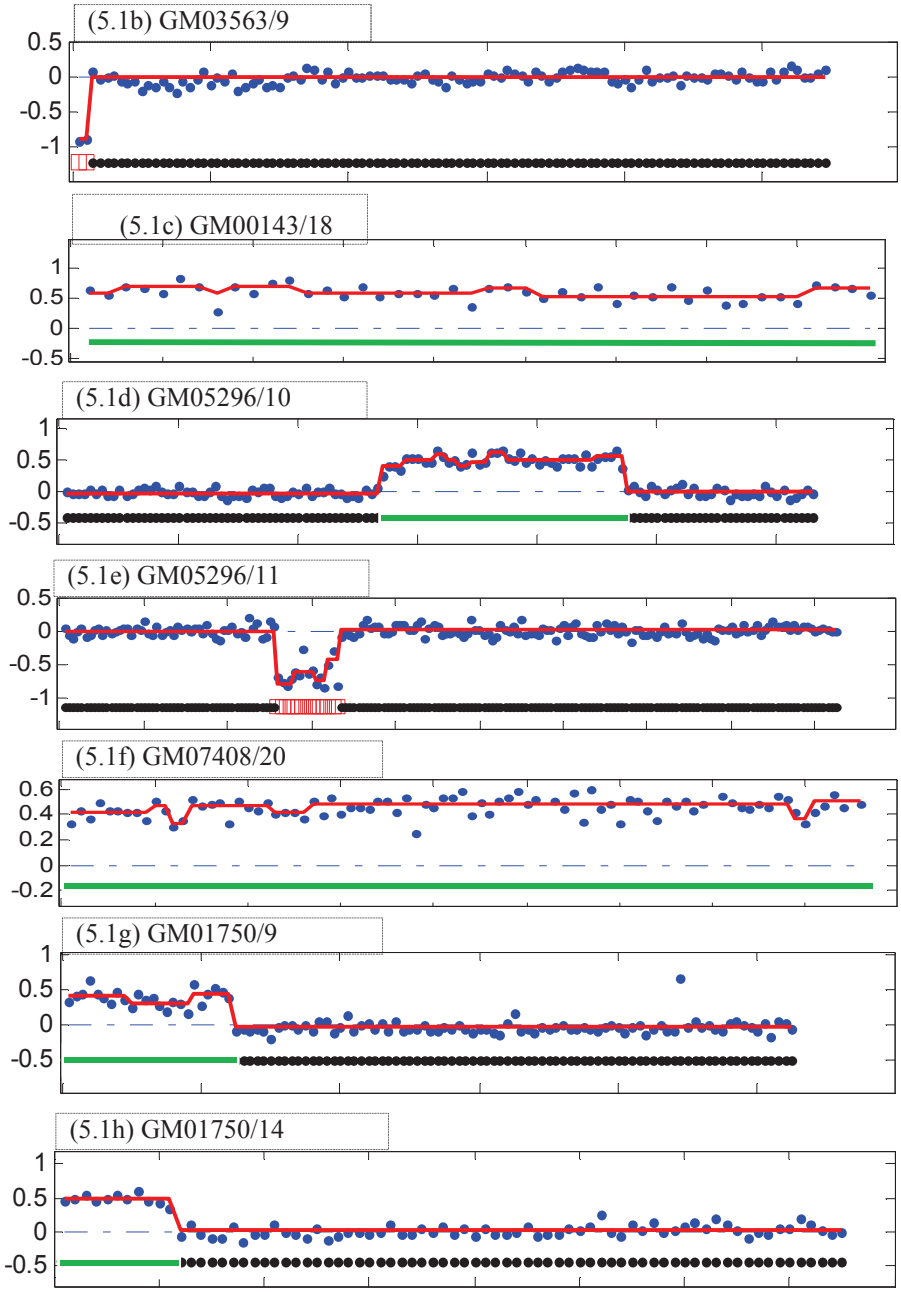
Table 5.1a List of cytogenetically mapped aberrations in the 15 coriell cell lines with their locations on the genome and the aberrations detected by the EES method.

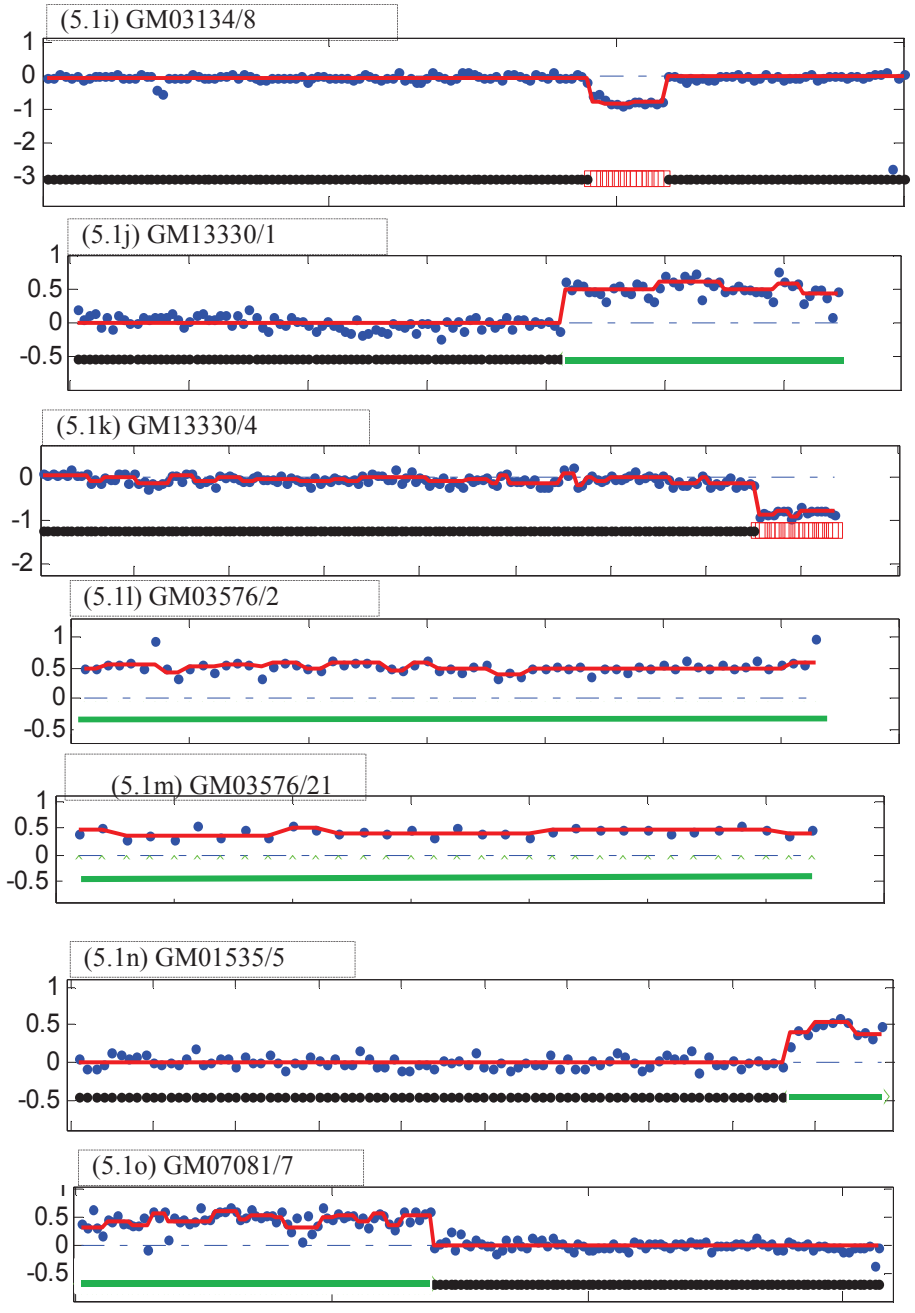
Cell Line	**Chromosome**	**Mapped Aberrations**	**Detected regions by EES**
GM03563	3	Trisomy 3q12-3qter	Trisomy 3q12-3qter
	9	Monosomy 9pter-9p24	Monosomy 9pter-9p24
GM00143	18	Trisomy on whole 18	Trisomy on whole 18
GM05296	10	Trisomy 10q21-10q24	Trisomy 10q21.3-10qter
	11	Monosomy 11p12-11p13	Monosomy 11p12-11p13
GM07408	20	Trisomy on whole 20	Trisomy on whole 20
GM01750	9	Trisomy 9pter-9p24	Trisomy 9pter-9p21
	14	Trisomy 14pter-14q21	Trisomy 14pter-14q12
GM03134	8	Monosomy 8q13-8q22	Monosomy -8q13-8q21.1
GM13330	1	Trisomy 1q25-1qter	Trisomy 1q22-23-1qter
	4	Monosomy 4q35-4qter	Monosomy 4q34-4qter
GM03576	2	Trisomy on whole 2	Trisomy on whole 2
	21	Trisomy on whole 21	Trisomy on whole 21

GM01535	5	Trisomy 5q33-5qter	Trisomy 5q34- 5qter
	12	Monosomy 12q24-12qter	Missed
GM07081	7	Trisomy 7pter-7q11.2	Trisomy 7pter-7q11.2
	15	Monosomy 15pter-15q11.2	Missed (not detected by any method)
GM02948	13	Trisomy on whole 13	Trisomy on whole 13
GM04435	16	Trisomy on whole 16	Trisomy on whole 16
	21	Trisomy on whole 21	Trisomy on whole 21
GM10315	22	Trisomy on whole 22	Trisomy on whole 22
GM13031	17	Monosomy 17q21.3-17q23	Monosomy 17q21.3-17q24
GM01524	6	Trisomy 6q15-6q25	Trisomy 6q12-6q22.3

It can be observed that the EES method successfully identified the loss in chromosome 9 of GM03563 which includes only two points (Fig 5.1b) which illustrates that a minimum resolution of aberration of 2 points can be identified by the EES method as illustrated with simulated data in the previous chapter.




(5.1b) GM03563/9
0.5
0
-0.5
-1
(5.1c) GM00143/18
1
0.5
0
-0.5
(5.1d) GM05296/10
1
0.5
0
-0.5
(5.1e) GM05296/11
0.5
0
-0.5
-1
(5.1f) GM07408/20
0.6
0.4
0.2
0
-0.2
(5.1g) GM01750/9
1
0.5
0
-0.5
(5.1h) GM01750/14
1
0.5
0
-0.5


(5.1i) GM03134/8
1
0
-1
-2
-3
(5.1j) GM13330/1
1
0.5
0
-0.5
(5.1k) GM13330/4
0
-1
-2
(5.1l) GM03576/2
1
0.5
0
-0.5
(5.1m) GM03576/21
1
0.5
0
-0.5
(5.1n) GM01535/5
1
0.5
0
-0.5
(5.1o) GM07081/7
0.5
0
-0.5

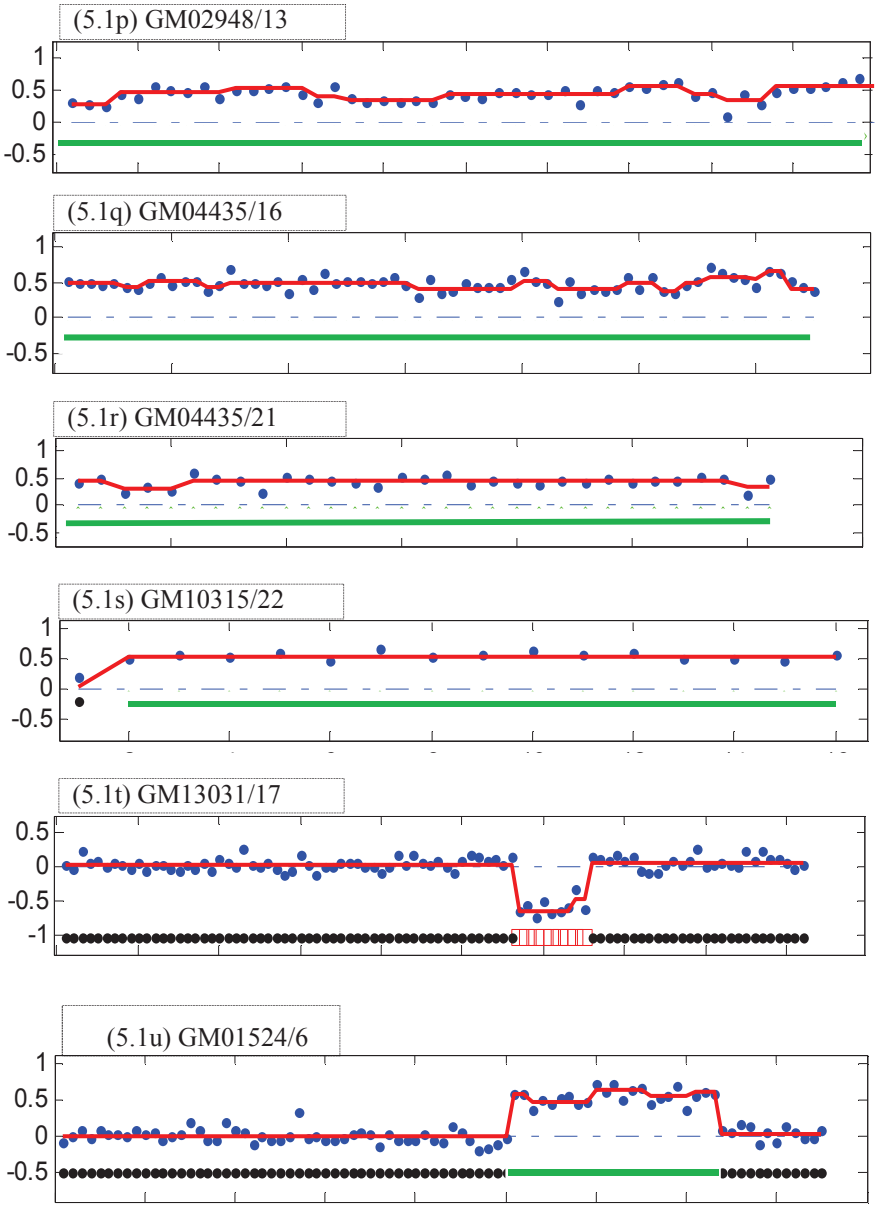


Figure 5.1[a-u] Results obtained for Coriell cell line data using the EES method.

The EES method detected all the aberrations reported by Snijders et al. (2001) except one single point aberration on chromosome 12 of GM01535. The single point aberration is not detected, as the algorithm is designed to exclude such cases as it is difficult to categorize a single point variation as a valid variation or not. It is demonstrated by Fig 5.2 a-b.

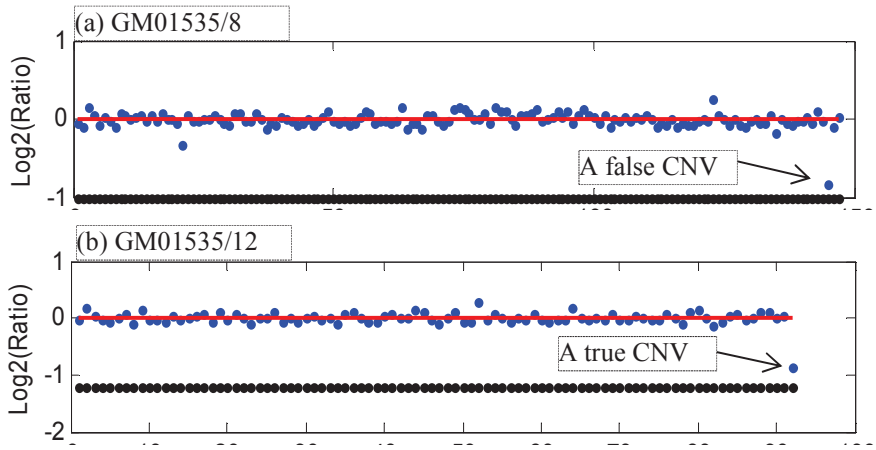


Figure 5.2 a) GM01535/chromosome 8 profile. b) GM01535/chromosome 12 profile.

The Fig 5.2b shows the profile of chromosome 12 of sample GM01535. The plot shows the single point loss at the end point which is the variation undetected by the EES method. Fig 5.2a shows the log ratio plot of chromosome 8 of sample GM01535 and similar to the chromosome 12, chromosome 8 also exhibits a single point loss at the end point. In the case of chromosome 12 the loss is a valid CNV, while the loss in chromosome 8 is not a valid CNV as per the cytogenetic mapping result. The EES method is designed so as to avoid classifying such single point variations as valid CNV to reduce false positive identification. The EES method thus identified 21 out of the 22 aberrations, reported by Snijders et al. (2001), successfully including all the 8 entire chromosome aneuploidy. The dataset is also

analyzed using Wavelet, CBS, Quantreg and CGHseg methods (Annexure A3). The Quantreg and CBS methods detected 21 out of the 23 known aberration with 19 and 7 false positives respectively. The Wavelet and CGHSeg methods detected 22 out of the 23 known aberrations with 82 and 9 false positive detections respectively. A comparison of the true positives and false positives detected by the different methods is given in Table 5.1b. It is noteworthy from the Table 5.1b that the EES method is the only one that achieved the result without any false positive detection.

Table 5.1b Comparison of True & False CNVs detected by different methods.

	Wavelet	**Quantreg**	**CBS**	**CGHSeg**	**EES**	**Actual**
True Positives	22	21	21	22	21	23
False Positives	82	19	7	9	0	0

5.3.2 Application on Breast Cancer Cell Line (BCCL) data

In this analysis, the dataset involves all chromosomes from the 51 breast cancer cell line samples. Analysis of the entire dataset is performed using the EES algorithm. All the 55 candidate genes reported to be amplified and over expressed in the dataset are located in 4 chromosomes- 8, 11, 17 and 20. Hence only results obtained for these 4 chromosomes are discussed here.

Table 5.2 List of the over expressed genes in BCCL database, detected by the EES algorithm as amplified

SPFH2	LOC441347	STARD3	AP3M2	***GOLGA7***
PROSC	CCND1	PNMT	VDAC3	***SLD5***
BRF2	FADD	PERLD1	FLJ20291	***1KBKB***
RAB11FIP1	PPFIA1	ERBB2	PPARBP	***POLB***
ASH2L	CTTN	GRB7	RAB22A	***FNTA***
LSM1	NADSYN1	GSDML	VAPB	***ACACA***
BAG4	N-PAC	PSMD3	STX16	
DDHD2	DDX52	ZNF217	NPEPL1	**CSTF1**
WHSC1L1	TBC1D3	BCAS1	GNAS	**THAP1**
FGFR1	PCGF2	RAE1	TH1L	
ADAM9	PSMB3	RNPC1	C20orf45	
MYST3	PIP5K2B	TMEPAI		

The EES method successfully detected all the 55 genes as amplified in at least one cell line. The list of over expressed genes detected as amplified by EES method is shown in Table 5.2. The results of Neve et al. (2006) shows that 47 amplifications or gene copy number gain are detected on the same cell line, where an over expression of the gene is observed. Meanwhile, 8 amplifications are detected for the same gene but in other cell

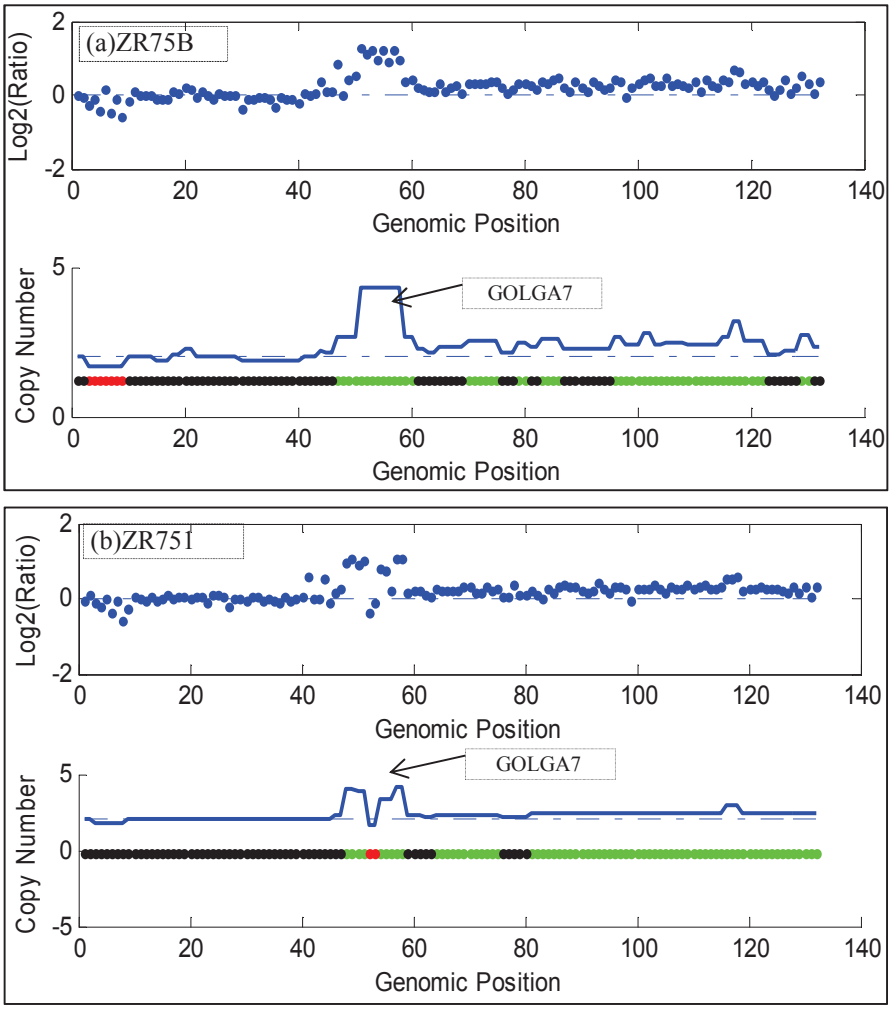


Figure 5.3 Amplification of GOLGA7 containing clone in chromosome 8 of (a)ZR75B and (b)ZR751 samples.

lines where high gene expression was not observed, which is not expected and needs further biological validation. EES algorithm detected 53 of these genes to be amplified in the same cell line in which an over expression was observed and only 2 are detected on other cell lines. For example, the gene GOLGA7 is over expressed in 7 cell lines but Neve et al. (2006) did not

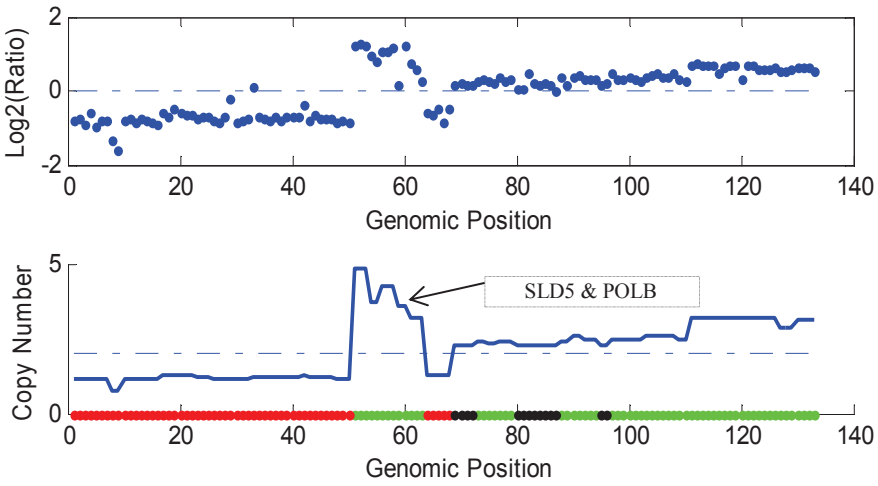


Figure 5.4 Amplification of SLD5 & POLB in chromosome 8 of CAMA1 sample

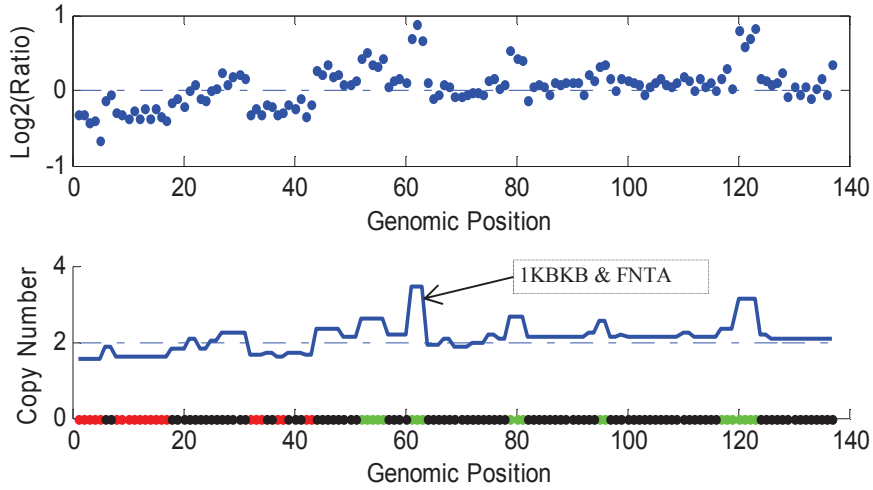


Figure 5.5 Amplification of 1KBKB & FNTA in chromosome 8 of SUM185PE sample.

detect amplification in any of these 7 cell lines. Instead the gene is found amplified on another 4 cell lines where over expression is not reported. The EES algorithm was able to detect the amplification of GOLGA7 in the cell line ZR75B itself, which also showed over expression of the gene. Fig 5.3

shows amplification of GOLGA7 containing clone in chromosome 8 of ZR75B and ZR751 samples.

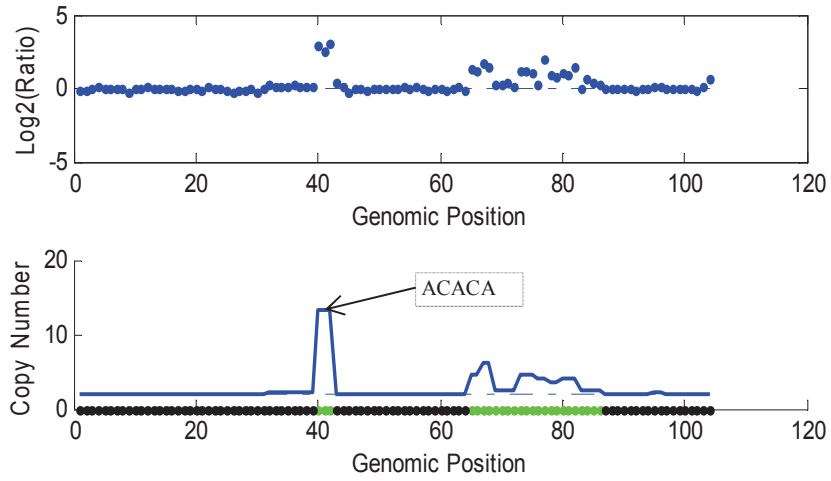


Figure 5.6 Amplification of ACACA in chromosome 17 of BT474 sample.

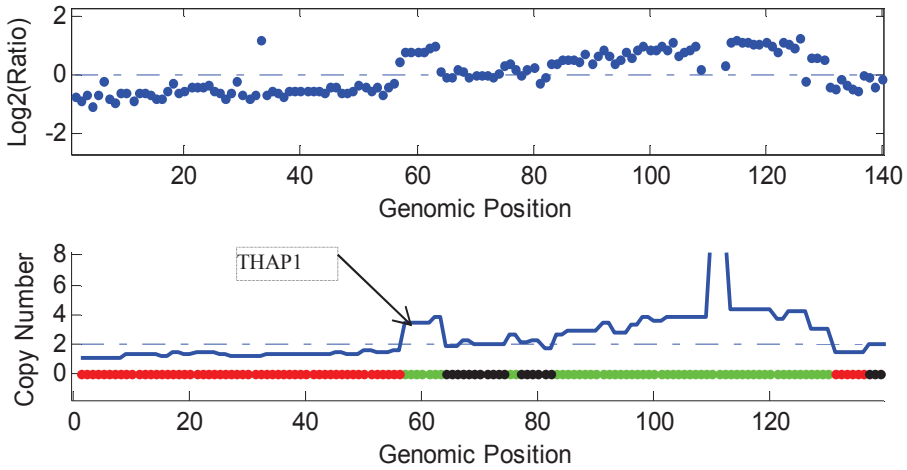


Figure 5.7 Amplification of THAP1 in chromosome 8 of HCC1954 sample

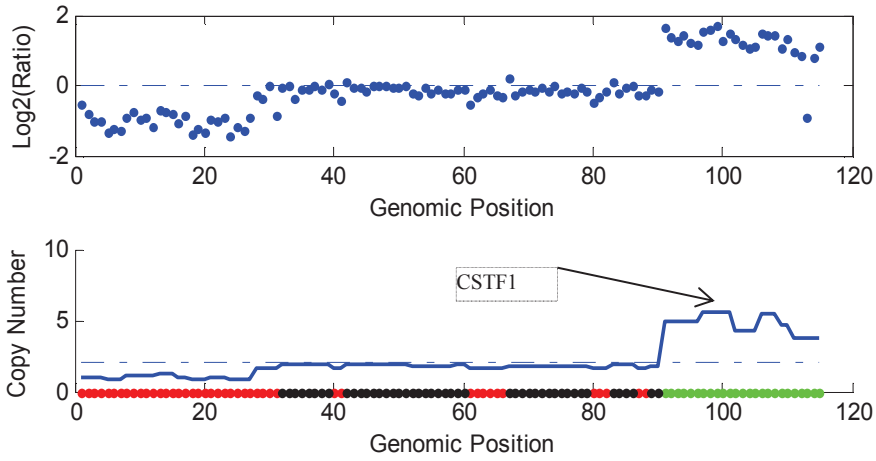


Figure 5.8 Amplification of CSTF1 in chromosome 20 of HCC1428 sample

[In the above figures, the plot shown on the top with blue dots, shows the input $\log_2$ratio values for the clones or spots on the array as blue dots. The plot below shows the copy number value estimated by the EES method in blue line at the same logarithmic scale. The status bar at the bottom of the plot shows the copy number status across genomic locations. The black, red and green colours indicate the normal regions, region with loss and gain respectively. The X-axis represents the location of clones along the chromosome.]

Similarly 5 more genes, SLD5, 1KBKB, POLB, FNTA and ACACA are also detected by EES algorithm as amplified in cell lines where a high gene expression is observed. The 6 additional genes which are detected as amplified by the EES method on the same cell line is shown in the column 5 of Table 5.2, in bold italics. Fig 5.4 shows the gain in copy number of SLD5 gene detected in chromosome 8 of CAMA1. It also shows the gain of gene POLB detected in chromosome 8 of CAMA1. Fig 5.5 shows amplification of 1KBKB and FNTA in chromosome 8 of SUM185PE sample. Fig 5.6

shows amplification of ACACA in chromosome 17 of BT474 sample. The two genes, THAP1 and CSTF1 in column 5 shown in bold letters are the ones which showed amplification in cell lines where an overexpression was not observed. Fig 5.7 shows the amplification of THAP1 in chromosome 8 of HCC1954 sample and Fig 5.8 shows amplification of CSTF1 in chromosome 20 of HCC1428 sample. The analysis clearly highlights the improved performance of the EES algorithm in the copy number detection by identifying several true aberrations which were missed by earlier analysis.

5.3.3 Application on Glioblastoma Multiforme (GBM) data

Array CGH profiles of 26 glioblastoma occurrences of varying grade are studied in this analysis. All the chromosomes in the 26 samples are analyzed using the EES method. The analysis of GBM data using the EES algorithm detected some characteristic genomic alteration regions associated with glioblastoma occurrences. These are:

i) Losses on chromosome 10.
ii) Losses of large portions of chromosome 13 and 22.
iii) Gains on whole chromosome 19 & 20.
iv) Loss of terminal region of p-arm of chromosome 9 (9pter region).

The Fig 5.9a shows the log ratio profile of chromosome 10 of GBM28 sample. The high signal variance of the data can be noticed from the plot. Fig 5.9b shows the copy number level estimated by the EES algorithm. The status bar at the bottom of the plot shows the copy number status across genomic locations. The black, red and green colours indicate the normal regions, region with loss and gain respectively. The X-axis represents the location of clones along the chromosome. The copy number estimated by

the algorithm shows a large number of losses throughout the chromosome 10 indicated by the red colour in the status bar at bottom.

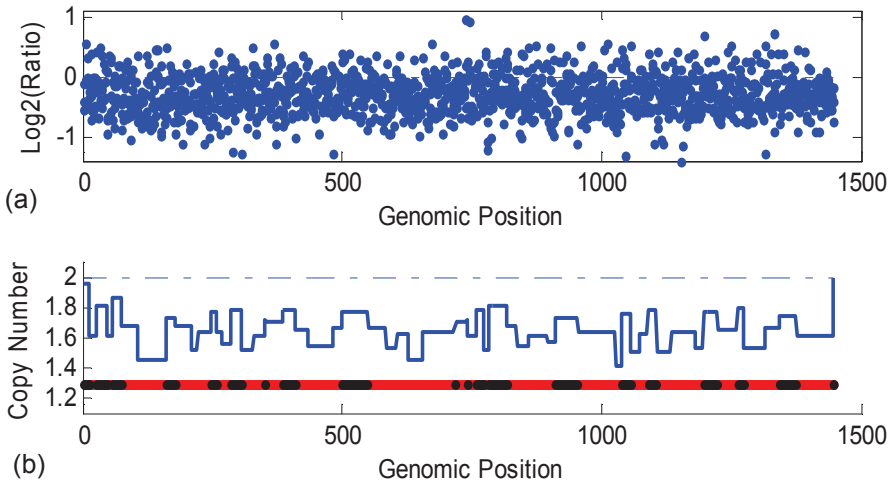


Figure 5.9 a) Log ratio data of chromosome10 in GBM28 b) Copy number estimated using EES.

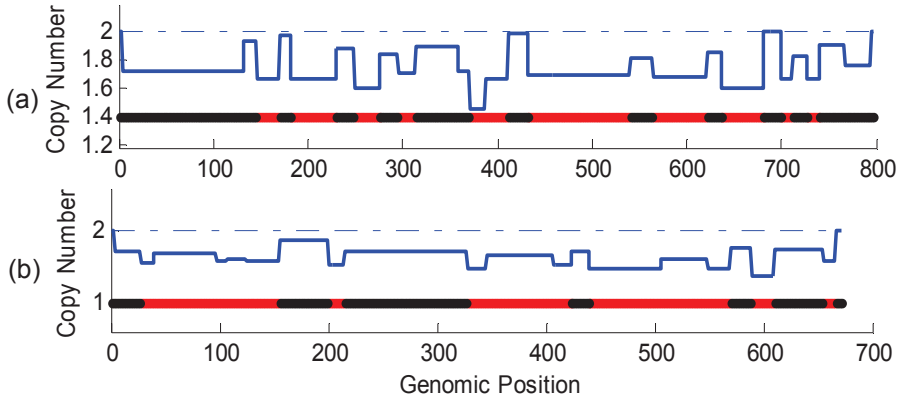


Figure 5.10 The copy number observed and its locations in GBM28 a) Chromosome 13 b) Chromosome 22.

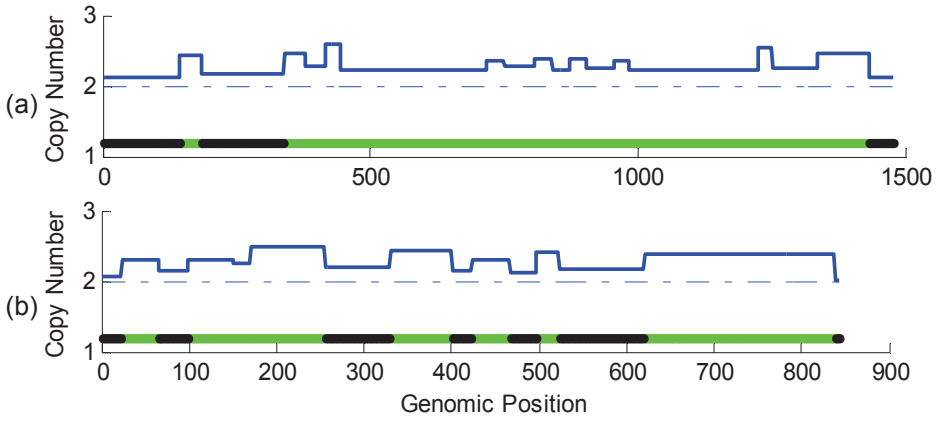


Figure 5.11 Whole chromosome copy number gain observed in GBM27
a) Chromosome 19 b) Chromosome 20

The Fig 5.10a shows the copy number losses observed and its locations in chromosome 13 of GBM28 and 5.10b shows the copy number losses estimated on chromosome 22 of GBM28. It can be observed that both of these chromosomes show losses similar to that observed in chromosome 10. Analysis of the GBM13 sample also showed similar losses in chromosome 10, 13 and 22.

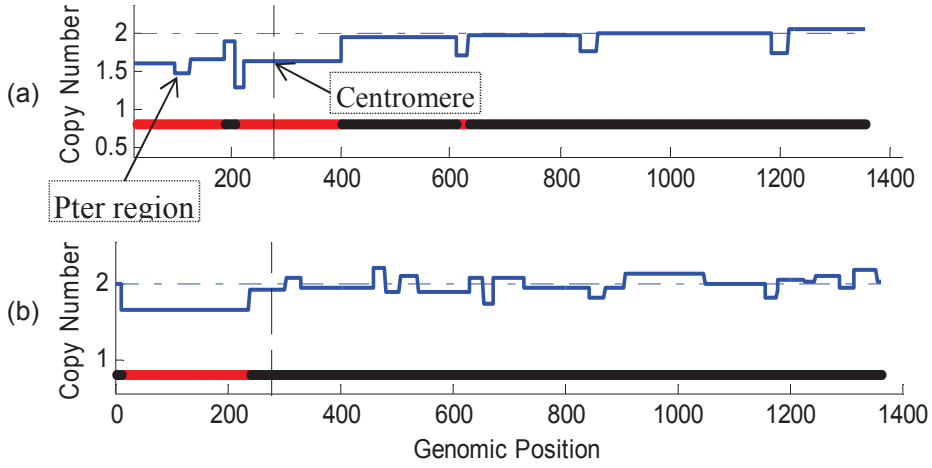


Figure 5.12 Copy number loss observed in pter region of chromosome9 of
(a) GBM21 (b) GBM31.

The analysis of GBM27 sample has shown whole chromosome gain in chromosome 19 and chromosome 20. The same alteration is also observed in six other samples of GBM. The Fig 5.11 shows the results obtained for GBM27. It can be seen that the entire length of chromosomes shows copy number value greater than two.

The CNV analysis of the GBM21 sample showed the loss of p-terminal region of chromosome 9 (9pter) as shown in Fig 5.12a. The vertical dashed line indicates the position of centromere of chromosome and the part to the left of centromere represents the p-arm and the part to the right represents the q-arm of the chromosome. Analysis of GBM31 also showed similar loss of 9pter region as shown in Fig 5.12b.

Several interesting correlation between alterations are also noted in the analysis. Two of the most evident correlated behavior of alterations observed in our analysis are given below.

i) Gains on chromosome 7 and losses on chromosome 10
ii) Gains on chromosome 20 and losses on chromosome 10

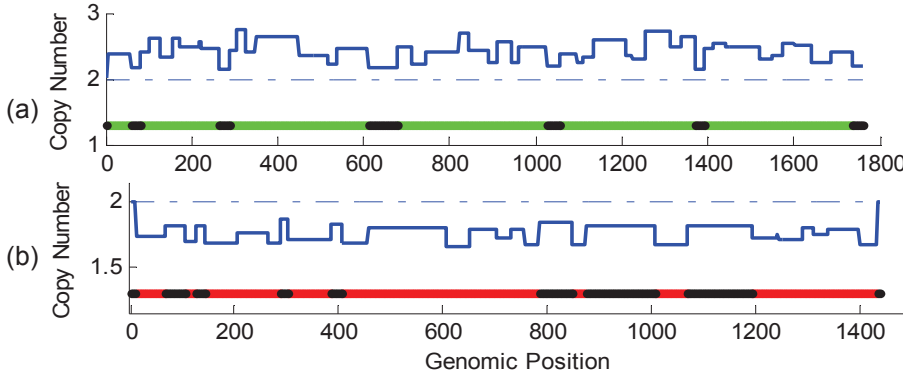


Figure 5.13 Correlated occurrence of entire chromosome gain of chromosome 7 and loss of chromosome 10 in GBM9 (a) chromosome 7 (b) chromosome 10

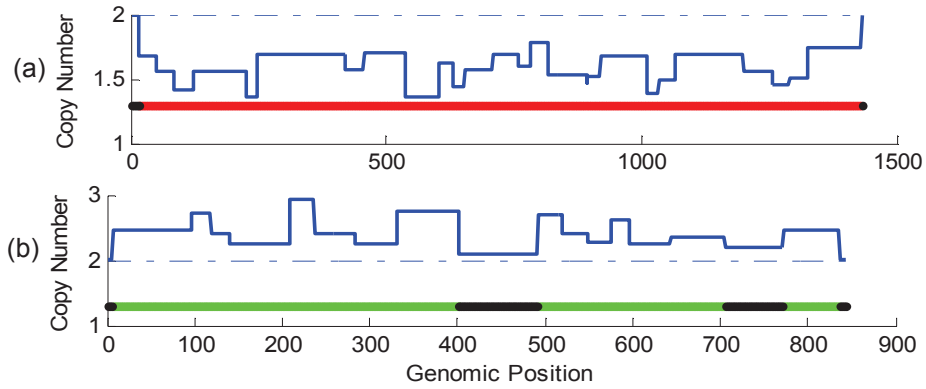


Figure 5.14 (a) Loss on chromosome 10 of GBM29 (b) Gain on chromosome 20 of GBM29.

Analysis of the relationship between alterations in glioblastoma samples showed correlated alterations in chromosome 7 and 10. In several samples, gains on chromosome 7 are found to be associated with a loss on chromosome 10 of the same sample. This is the most common correlated alteration, found in 11 GBM samples. Fig 5.13 shows an example of this co-occurrence in GBM9 sample.

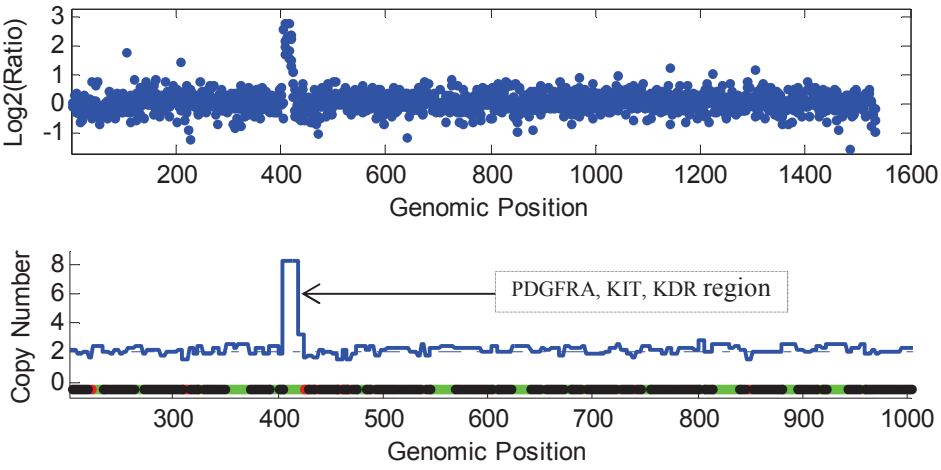


Figure 5.15 Co-amplification of PDGFRA, KIT, KDR in chromosome 4 of GBM6

The second co-occurrence of alteration observed is the correlated occurrences of gain on chromosome 20 with a loss on chromosome 10 of the same sample. Seven GBM samples showed this co-occurrence and an example of this co-occurrence observed in GBM29 is shown in Fig 5.14.

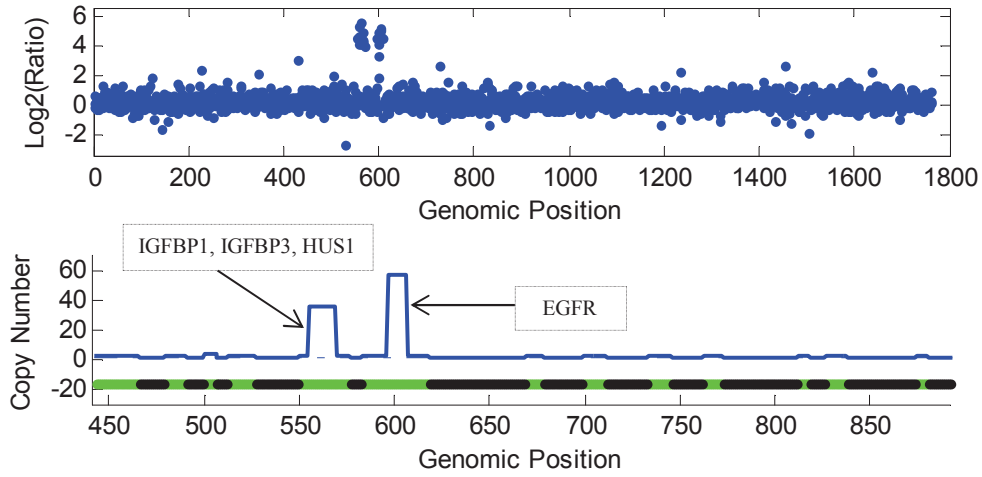


Figure 5.16 Co-amplification of EGFR, IGFBP1, IGFBP3 & HUS in chromosome 7 of GBM29

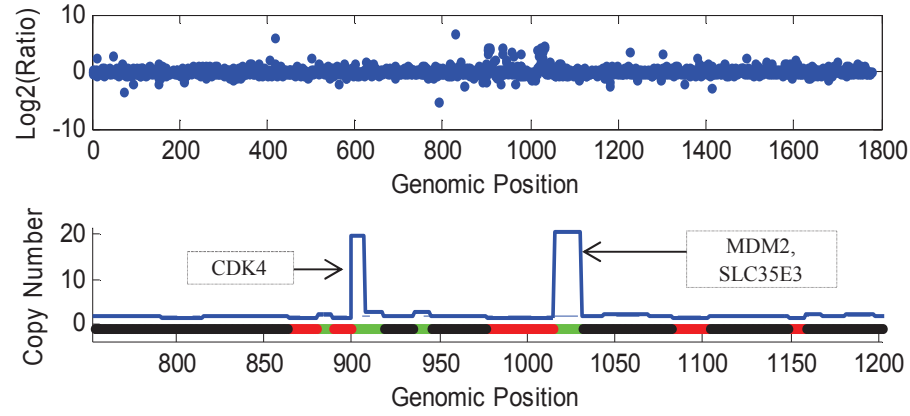


Figure 5.17 Co-amplification of CDK4 and MDM2, SLC35E3 in chromosome 12 of GBM22.

The high resolution Array CGH based analysis of copy number alterations has allowed the detection of genes and co-alteration of genes which may contribute to tumor genesis. These analyses have been able to identify a few genes which are closely related to glioma genesis. The GBM data contained several amplicons, of which the amplifications found around the region containing the genes such as PDGFRA, EGFR and CDK4 are well studied. The PDGFRA gene located on chromosome 4 was observed to be co-amplified with the oncogene KIT and a growth factor receptor gene KDR in GBM6 sample as shown in Fig 5.15. Meanwhile the amplification of the EGFR gene is observed in 11 GBM samples. A couple of glioblastoma samples showed co-amplification of the IGFBP1 and IGFBP3 genes with the EGFR gene. Fig 5.16 shows the co-amplification in chromosome 7 of GBM29. Seven GBM samples showed amplification of CDK4 gene while a couple of them showed co-amplification with MDM2 and a putative oncogene SLC35E3. Fig 5.17 shows this co-amplification in chromosome 12 of GBM22. All the results discussed above are supported by the results and observations of Bredel et al. (2005). More detailed analysis of the samples can be done with the presented algorithm, but is not attempted here as it is beyond the scope of this work. From the above discussions it can be clearly seen that the presented EES method can successfully detect and localize the copy number alterations from noisy array data that otherwise is a cumbersome task.

5.4 Summary

The EES algorithm is applied on three real Array CGH datasets for the analysis and detection of copy number aberrations. The first real data set selected for analysis is the Coriell cell line BAC Array CGH data. It consists

of 15 cell strains containing single copy aneuploidies involving partial as well as whole chromosome. The EES algorithm successfully identified 21 out of the 22 aberrations, reported by Snijders et al. (2001), including all the 8 entire chromosome aneuploidy without any false positives. The second dataset is a breast cancer cell line database described in Neve et al. (2006). It consists of 51 breast cancer cell lines. The dataset describes 55 genes as candidate therapeutic target genes which are amplified and over expressed in at least one cell line. EES algorithm is used to identify these genes in the dataset. Neve et al. (2006) detected 47 amplifications on the same cell line where an over expression of the gene is observed and 8 amplifications are detected for the same gene but in other cell lines where high gene expression was not observed. EES algorithm detected 53 genes as amplified in the same cell line in which an over expression was observed and only 2 are detected on other cell lines. The third data set is the glioblastoma multiforme (GBM) data. The Array CGH profiles of 26 glioblastoma occurrences (Bredel et al., 2005) are studied. The analysis of GBM data using the EES algorithm detected many characteristic genomic alterations associated with glioblastoma occurrences like - losses on chromosome 10, 13 and 22; gains on whole chromosome 19, 20; loss of 9pter region. It also showed correlated alterations like - gains on chromosome 7 and losses on chromosome 10; gains on chromosome 20 and losses on chromosome 10. The analysis also showed several other amplifications and deletions which are closely related to glioblastoma occurrences. All these results clearly illustrates that the EES method combining the merits of both smoothing and segmentation tasks, can successfully denoise, detect and localize the copy number variations from the log ratio data obtained from Array CGH experiments.

Chapter 6

Phylogenetic Analysis

A brief overview of evolutionary mechanisms, comparative genomics and phylogenetic analysis is provided. It also describes tree representation, basic terminologies involved and tree construction methods. Finally a review of literature describing various alignment free method of phylogenetic analysis is also discussed.

6.1 Introduction

The history of evolution of life on earth traces back to the very appearance of life in its most primitive form on this planet about 500 million years ago. According to the theory of evolution of life, all the complex, diverse living organisms that exist now and have got extinct share a common ancestor. The evolution of the genomes from the very origin to the highly complex form of present day has a history of millions of years of gradual transformation through various events. The first form of genome can be thought of as RNAs capable of self-replication and directing biochemical reactions (Brown, 2007). Later DNA genomes and genes with specific functions emerged leading to the current day complexity of genomes with tens of thousands of genes and several chromosomes. Darwin (1959) attributes the existence of a large number of species to the accumulation of hereditary modifications that are retained by nature through the process of natural selection. The variations incurred in a species that are favorable for the existence under a circumstance are retained and are passed on to subsequent generations. And this natural selection leads to the extinction of the organisms that failed to acquire favorable variations that enable them to adapt to the environment. The existence of the present day genomes with high complexity and wide variety can thus be attributed to the process of genomic variation and natural selection according to Darwin. With the advancement in whole genome sequencing methods, the biologists currently have unprecedented amount of information about the genome of many organisms for a comparative study. This comparative study, also called comparative genomics, has taken a central role to understand the evolutionary changes. The theory of evolution thus forms a basis for comparative genomics. Phylogenetics is the study of evolutionary

relationships between organisms for which comparative genomics can play a significant role. The following section describes an attempt to uncover the evolutionary relationship between organisms using comparative analysis of sequence variation.

6.2 Comparative genomics

Comparative genomics aims at comparing the genomic sequences from different species, to study their similarities and dissimilarities. The basic principle behind the comparative genomics analysis is that, the part of genome sequence that is critical for a biological function is highly conserved across different species. The comparison of sequences can be performed at different levels, of coding regions, noncoding regions and of the overall genome structure. The comparison of genome size, number of chromosomes and genes represent the first level. Comparative analysis at genome levels includes comparison of overall nucleotide statistics like G+C content, codon and amino acid biasing. Analysis of coding region aims at identifying the genes present in organisms and comparing the count of total genes, common genes and unique genes. The analysis of non-coding region studies the presence of regulatory elements. Comparative genomics is a powerful tool for studying evolutionary changes among organisms. It also helps to identify genes that are conserved and the genes that are unique to each organism. The multiple sequences from different organisms are aligned and the conserved regions are identified. The degrees of similarity or conservation among the different species give an indication regarding the recency of a common ancestor. The analysis has shown a high degree of sequence similarity among closely related organisms. The biggest challenge in the process is the alignment of sequences. The sequences due to its large

size, insertions, deletions, duplications and rearrangements make the manual alignment highly impractical and cumbersome. Hence the development of computational and visualization tools for genome scale sequence alignment and comparison is of highest priority and has been a topic of hot pursuit. The information gained by comparative genomics can help scientists to understand the structure and function of genes, and develop new methods for clinical applications and for understanding the evolutionary process.

6.3 Molecular evolution

The phenotype of organisms is determined by the genetic information they possess and its interaction with the environment. A study of changes in the genetic information allows to understand the mechanisms that results in the phenotypic variations of organisms. The genetic information is carried in most of the organisms by deoxyribonucleic acid (DNA). DNA bases sometimes undergo modification as a result of chemical changes initiated by environmental factors and due to action of DNA polymerase during duplication events. Most often these changes are identified and removed by cellular repair mechanisms. There are occasions when such changes escapes these mechanisms and results in the genetic information getting altered. This is known as mutation. The genetic code consists of 64 triplets of nucleotides called codons. Among 64 codons, 61 codons encodes for one of the 20 amino acids. This results in a redundancy in the code, where more than one codon encodes for most of the amino acids. All point mutations do not result in amino acid changes due to this redundancy of the genetic code. Transitions are mutations that occur as a result of a purine base (A, G) getting replaced by a purine or a pyrimidine (T, C) with a pyrimidine due to chemical reasons. Transversions are changes

where a purine is replaced by a pyrimidine and vice versa. Changes in DNA can also occur as a result of deletion or insertion of one or more nucleotides during the duplication event. Genetic recombination is another source of variation, where DNA segments are broken and recombined to produce new combinations. The recombination during meiotic division is a major source of variation in diploid organisms (Salemi et al., 2009). Apart from these there are various other sources of variation like gene duplication and lateral gene transfer that contributes to the changes in DNA or molecular evolution. Polymorphism is the occurrence of more than one variants of a gene with in the population of a species. This occurs when a gene mutation is passed on to the offspring and the new gene variant and the original gene coexist within a population. These different variants are known as alleles. Within a population some of the alleles are lost over the course of time, while some other may become prominent as a result of natural selection and subsequently leads to the evolution of a new species. The rate of genetic divergence is thus dependent on the rate of mutation or rate of genetic changes.

6.4 Scientific classification

Scientific classification, or taxonomy, is a hierarchical classification of living and extinct organisms based on various characteristics. There are eight levels of hierarchical in the system as shown in Fig 6.1a. Domain is the broadest division and the lowest and the basic unit of classification is species. The hierarchical groupings in between include kingdom, phylum, class, family, order and genus. Fig 6.1b shows an example of a full classification for human.

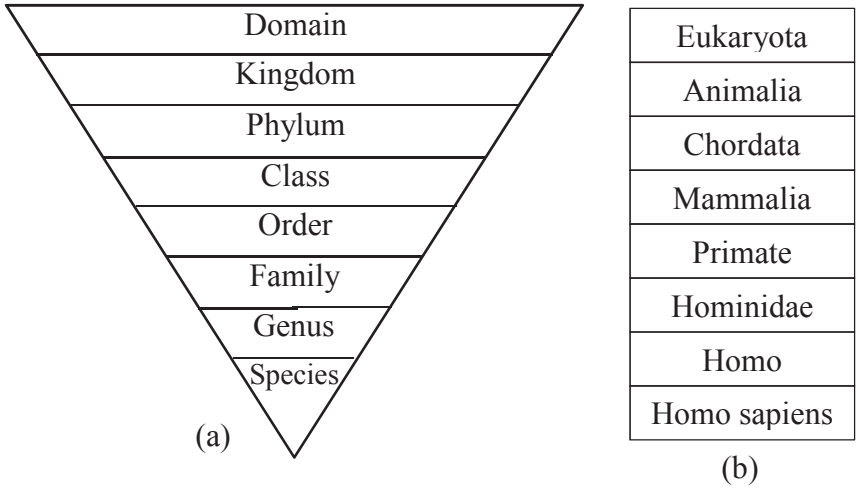


Figure 6.1 (a) Scientific classification of organisms. (b) An example: different hierarchical levels of human.

6.5 Phylogenetic analysis

Phylogenetics is the study of evolutionary relationship among organisms. There are different approaches to phylogenetic analysis, but the objective of all is to infer the evolutionary relationship between organisms and to infer the time of divergence from a common ancestor by studying the variable characters in the organisms involved. There exists relationship between all organisms and hence it is a relative concept where the degree of relationship may vary. An organism 'A' can be described to be more closely related to 'B' than another organism 'C', though all the three are mutually related. In the initial stages morphological features were used for comparison. The advancement in laboratory and analytical technologies made the task of sequencing the entire genome cost effective and quicker. As a result genome sequences of more and more organisms could be obtained allowing large scale comparisons and studies for better

interpretation of evolutionary pathway. Genome sequencing projects provides the exact sequences of nucleotides in DNA or RNA segments which ensure detailed and unambiguous data for molecular phylogenetics. Molecular data contains easily identifiable and independent character states and hence provides for a better quantitative comparison compared to the ones using morphological and embryological features that are far more subjective.

6.5.1 Molecular phylogenetics

Molecular phylogenetics infers the evolutionary relationship between organisms using molecular information. As genomes evolve through gradual accumulation of variations, the amount of difference between a pair of genome should indicate the evolutionary distance between them. Molecular phylogenetics aims to gain information regarding the evolutionary relationships by comparing the genome sequences. Zuckerkandl and Pauling, (1962) suggested that substitution rates were essentially constant within homologous protein over a large period of time. A molecular clock hypothesis was proposed which states that DNA and protein sequences evolve at a rate that is relatively constant over time. The accumulation of these changes can be compared to the steady ticking of a clock. Thus the number of differences between two homologous proteins can be well correlated to the amount of time since speciation caused them to diverge independently. This will facilitate to decipher the phylogenetic relationship between species and also the time of their divergence. In molecular phylogenetics the most common approach is the comparison of homologous gene sequences, using sequence alignment technique to establish the trend of evolution of the particular gene (Hill, et al., 1963; Nadler, 1995). For a comprehensive analysis, the organisms selected for

comparison must represent the various stages in the evolution from the most primitive form to the recent advanced form (Hill, et al., 1963). The molecular sequences in the form of character sequences can also be easily represented in numerical form which allows application of mathematical, statistical and signal processing methods for analysis. To calculate the percentage similarity between two sequences, the number of identical nucleotides or amino acids is counted, relative to the length of the sequence. The genes of closely related species usually have higher sequence similarity compared to the distantly related species.

6.6 Phylogenetic analysis using protein sequences

DNA sequences or nucleotide sequences are mostly compared to estimate the evolutionary relationship between organisms. Protein sequences also undergo evolutionary changes similar to that of DNA sequences. All genetic mutations in the genes will not lead to changes in protein sequences. Only those mutations that result in a change of amino acid are reflected in the protein sequence. This is due to redundancy of genetic code, whereby more than one codon can code for each amino acid. As a result, many of the changes in the DNA sequences will not result in protein sequence and thus functionality making such changes less significant. There are phylogenetic analysis approaches, where the changes in every third nucleotide in the protein coding sequence are neglected, owing to this fact (Salemi et al., 2009). It is also pointed out by Salemi et al. (2009), that use of protein sequence instead of DNA sequence significantly improves the signal to noise ratio of input data. Adachi and Hasegawa (1992) introduced a maximum likelihood method for inferring Protein Phylogeny, called MOLPHY PROTML, from amino acid

sequences. Matsuda (1995) also used a maximum likelihood method for tree construction from amino acid sequences. It uses genetic algorithms with scores derived from the log-likelihood of trees computed by the maximum likelihood method.

6.7 Phylogenetic Tree

Phylogenetic tree is the simplest and most useful form of representing the degree of evolutionary relationship between organisms and their time of divergence, in a visual form.

6.7.1 Tree terminologies

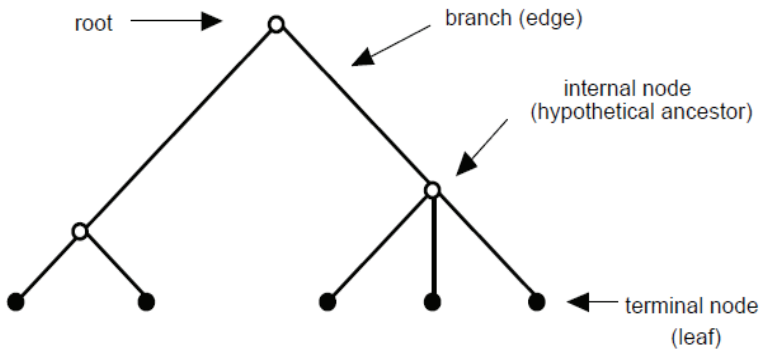


Figure 6.2 A tree representation and its components

Phylogenetic entities in the tree are commonly known as taxa. A tree consists of nodes connected by branches or edges. Terminal nodes or leaves or OTUs [Operational Taxonomic Units] represent sequences from organisms or organisms themselves. Internal nodes represent hypothetical ancestors from which the descendant taxa evolved. They typically represent extinct species that existed in the past. Root of the tree represents the ancestor of all the sequences. The length of the branches in a tree is termed branch length. It represents the amount of time or the amount of change

between two nodes. Topology of a tree is the branching structure of the tree without the branch lengths.

Cladogram

Cladogram is the most basic tree, which only shows the relative recency of common ancestry. For three sequences- X, Y, Z, the cladogram in Fig. 6.3a shows that sequences X and Y share a common ancestor more recently than either have with Z.

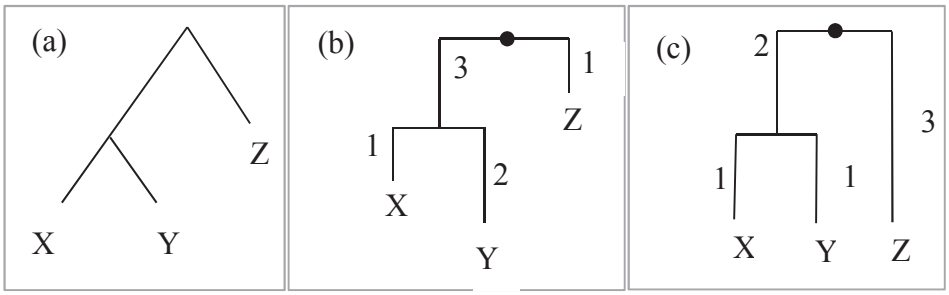


Figure 6.3 Form of trees (a) Cladogram (b) Phylogram (c) Dendrogram

Phylogram

Phylograms contains additional information about the branch lengths associated with each branch. Branch length reflects the amount of evolutionary changes between the sequences involved (Fig. 6.3b).

Ultrametric tree or Dendrogram

Ultrametric trees are phylograms where the leaf nodes are all equidistant from the root of the tree (Fig. 6.3c). They give an idea of evolutionary time, expressed either in years or as amount of sequence divergence using a molecular clock.

Rooted & Unrooted trees

Cladogram and phylogram can either be rooted or unrooted while an ultrametric tree is always rooted. A rooted tree has a unique node corresponding to the most recent common ancestor of all nodes in the tree. A rooted tree is a directed tree in the sense that all the other nodes descend from the root and this direction corresponds to evolutionary time. The closer a node is to the root; the older it is in time.

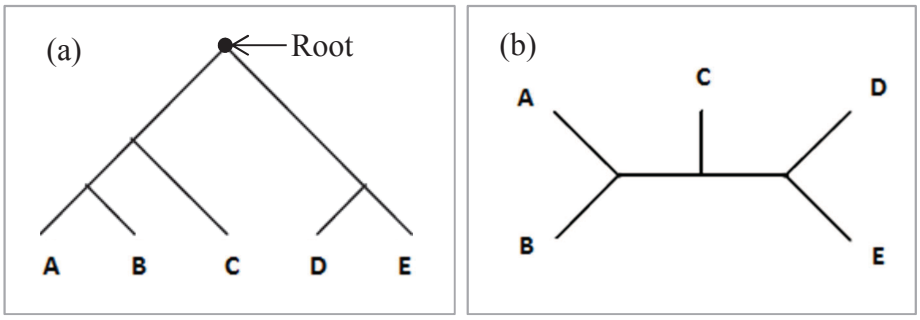


Figure 6.4 (a) Rooted tree (b) Unrooted tree

Unrooted trees do not have a node designated as root. They cannot depict the ancestor-descendant relationship as the rooted trees. The adjacent nodes on an unrooted tree need not be evolutionarily closely related and hence is not quite suited to represent phylogenetic relationship.

Tree representation using nested parenthesis

An easy text notation using nested parentheses for the representation of trees can be used. An internal node is represented by a pair or parentheses that enclose all descendants of that node. The Fig. 6.3a shows a tree with five nodes. Using the text notation the tree can be represented as (((A,B),C),D),E).

6.7.2 Phylogenetic tree construction

A species tree describes how the species evolved from a common ancestor and the speed of evolution of different lineages. The pattern of relationship can be easily represented using a tree which can be generated using either phenetic or cladistic method. Phenetics uses distance based methods and cladistics uses character based methods for phylogenetic tree construction. Maximum parsimony and maximum likelihood approaches are the commonly used character based methods while Neighbour Joining, Fitch-Margoliash and UPGMA are distance based approaches. Distance based method is a two-step process. The first step in the analysis is the computation of a distance matrix representing the genetic distances or evolutionary distances between all pairs of sequences. The second step is the construction of a tree from the distance matrix. The evolutionary distance or genetic distance between all pairs of sequences, obtained from the nucleotide or amino acid sequence dissimilarity is represented as the distance matrix. Under the influence of evolutionary pressures, sequences undergo changes in the course of time. Sequences derived from a common ancestor evolve independently of each other and eventually diverge. Evolutionary distance or genetic distance is the measure of this divergence and reflects a measure of the similarity between sequences. In analysis involving molecular clock, this genetic distance also indicates the time duration of divergence. The most common way of estimation of the sequence dissimilarity is using Multiple Sequence Alignment (MSA). Once all the pairwise distances between the sequences are computed, a tree topology can be inferred using different methods, like UPGMA, Fitch-Margoliash and Neighbor joining.

Multiple sequence alignment

Multiple sequence alignment has been the most important part of the phylogenetic analysis methods. It is the most common and standard way of comparing sequences and to measure the amount of similarities or dissimilarities between them. For sequence alignment, the DNA sequences or protein sequences are represented as rows within a matrix. Gaps are inserted between the residues to align the sequences so that similar characters occupy a column or homologous locations as much as possible. The alignment scores are generated to give a measure of alignment which represents the genetic distance between sequences. Except very short sequences, alignment of most sequences require computational approaches as it is very difficult, time consuming and complex.

Distance based methods for tree construction

Unweighted Pair Group Method with Arithmetic average (UPGMA) method starts by grouping the two nodes with smallest distance separating them into a single node. Then a new distance matrix is computed with the new node and the above process is repeated until all are grouped together. UPGMA generates rooted ultrametric trees (with molecular clock) assuming constant rate of change along the branches. Fitch-Margoliash with evolutionary clock uses a weighted least squares method for clustering. It generates different trees and a distance matrix is calculated for each of them. Then the optimal least squares tree is selected as the final tree structure. Neighbor Joining (NJ) algorithm, first selects the two most closely related nodes and group them under a single new node. Then calculate the distance of this new node to the remaining nodes and replace the two old nodes in the

matrix with new one. The process is repeated until all the nodes are included in the tree.

6.8 Evaluation of trees

Mainly two techniques are used to evaluate the reliability of the phylogenetic tree inferred using distance based methods. They are bootstrap analysis and jackknifing.

6.8.1 Bootstrap analysis

The bootstrap analysis is a technique used for assessing the accuracy of any statistical estimate. It uses a resampling technique to approximate the distribution of nodes in the tree. It was first employed by Felsenstein (1985) for the estimation of confidence intervals for phylogenies. The confidence for each subtree of a tree is based on the proportion of bootstrap trees showing the same subtree. In bootstrapping, new data sequences are obtained from the original by random sampling with replacements. Each original data point may be represented more than once or not at all in the new sequence called bootstrap replicate. The replicate sequence has the same length as that of original sequence. For each bootstrap replicate data set, a tree is constructed. The proportion of each branching pattern among all the bootstrap replicates is computed. This proportion is taken as the statistical confidence value of each subtree. Bootstrap analysis is a simple and effective technique to test the relative stability of groups within a phylogenetic tree. Under normal circumstances, considerable confidence can be given to branches or groups supported by more than 70%.

6.8.2 Jackknifing

An alternative resampling technique often used to evaluate the reliability of specific clades in the tree is jackknifing. Jackknife randomly purges half of the sites from the original sequences so that the new sequences will be half as long as the original. This resampling procedure typically will be repeated many times to generate numerous new samples. Each new sample set will be subjected to regular phylogenetic reconstruction. The frequencies of subtrees are counted from reconstructed trees. If a subtree appears in all reconstructed trees, then the jackknifing value is 100%; that is, the strongest possible support for the subtree. As for bootstrapping, branches supported by a jackknifing value less than 70% should be treated with caution.

6.9 A review of alignment free methods for sequence analysis

Multiple Sequence Alignment(MSA) has been the method of choice for several decades to compare genomic sequences. They are extensively used in phylogenetic analysis, genome annotation, gene prediction and protein structure analysis. Sequence alignment based methods are facing big challenges as the size and amount of sequences grow at a rapid rate. Computational complexity and time required for such methods are making it difficult to employ MSA methods for large volume of sequences. Sequence alignment assumes continuity between homologous segments which is adversely affected by events such as recombination and genetic shuffling. Alignment free methods are better suited to deal with large volume of data as they are computationally efficient and require less time. They are also able to handle sequence without continuity of homologous regions. They

are used for similarity searches, clustering and phylogenetics. Alignment free methods can be classified into four types, methods based on:

- Word frequency
- Substrings
- Information theory
- Graphical representation.

Several alignment free methods for calculation of genetic distance have been proposed. Chu et al. (2004) describes an alignment free phylogenetic analysis of complete genomes using correlation analysis of compositional vectors calculated from the frequency of amino acid strings. Genetic distance values for all species are obtained and phylogenetic relationships are inferred from the distance matrix. Another approach (Qi et al., 2004) uses frequency of amino acid K-strings in complete proteome for inferring evolutionary relatedness. A feature frequency profile (FFP) based method where the frequencies of l-mer features of whole genomes is illustrated in (Sims and Kim, 2011). Qi et al. (2012) represented DNA sequences by a dinucleotide frequency matrix/vector and comparisons between sequences are carried out by calculating the distances between these mathematical descriptors. Qi et al. (2011) proposed mathematical descriptors based on graph theory. The adjacency matrix of the directed graph is used to represent each DNA sequence. Similarity measures is obtained by taking both ordering and frequency of nucleotides into consideration. Ulitsky et al. (2006) introduced a new method for calculating the distance between sequences by computing the average lengths of maximum common substrings. In (Haubold et al., 2009), the number of substitutions per site between two DNA sequences using the shortest absent substring is used for mutation distance calculation. Information correlation

(IC) and partial information correlation (PIC) using the base correlation property of DNA sequence is described in Gao and Luo, (2012). An alignment free method for computation of genetic distance using DNA sequence to signal mapping function is employed by Borrayo et al. (2014).

6.10 Summary

An overview of the mechanism of evolution and phylogenetic analysis is provided in this chapter. Basic terminologies related to phylogenetic trees, methods of tree construction and tree evaluation are also described. Finally a brief review of alignment free methods of phylogenetic classification is also provided. A new signal processing based approach for protein sequence similarity analysis using frequency domain techniques is described in the next two chapters. An alignment free method to measure protein sequence similarity and to infer phylogenetic relationship between species using a single protein is developed. The study is then extended to develop a principal component analysis based method for combining phylogenetic information from multiple proteins for generating consensus phylogeny.

Chapter 7

Development of an alignment free method for phylogenetic classification using a single protein

A new approach for protein sequence similarity analysis using a frequency domain method is described. Based on the frequency domain approach, a Single Protein Power Spectral Density (SPPSD) method is developed. The method infers the phylogenetic relationship between species using the amino acid sequence of a single protein. The chapter also describes the application of the SPPSD method on different sample sets for inferring phylogenetic relationship.

7.1 Introduction

Determination of evolutionary divergence between species is a major task to understand the pattern of evolution of organisms. The molecular phylogenetic analysis methods focus on obtaining genetic distance between the various species from a measure of genomic sequence similarity. In this part, a new signal processing based method for phylogenetic classification using a single protein sequence, obtained from a collection of organisms, is described. The method uses an approach that calculates the protein sequence similarity without performing sequence alignment. The SPPSD method (Single Protein Power Spectral Density method) described here uses the distance between Power Spectral Densities (PSD) of protein sequences as a measure of genetic distance for the construction of phylogenetic tree.

7.2 Discrete Fourier Transform

Discrete Fourier Transform or DFT is a method of representing a discrete finite duration sequence x[n] as a linear combination of discrete complex exponentials restricted to a finite interval. The DFT X[k] of a sequence x[n] is defined by the analysis equation,

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi nk/N} \tag{7.1}$$

where $0 \leq k \leq N-1$ in the interval $0 \leq n \leq N-1$.

The X[k] is a complex sequence where the absolute value, |X[k]| represents the amplitude spectrum and arg (X[k]) represents the phase spectrum of the sequence x[n]. x[n] is a time domain or space domain representation of the sequence and X[k] is a frequency domain

representation of x[n] which shows the different frequency components present in the sequence.

7.3 Discrete Wavelet Transform

In Discrete Wavelet Transform (DWT) analysis, a signal is represented as the sum of wavelets with different locations and scales, with the coefficients indicating the strength of the contribution of the wavelet at the corresponding locations and scales. In DWT, any discrete time sequence f(n) of finite energy can be expressed in terms of the discrete time basis functions $\omega_{j,k}$(n) as,

$$f(n) = \sum_{j,k} d_j(k)\omega(2^j n - k), \qquad (7.2)$$

where, $d_j(k)$ represent the coefficient corresponding to scale j and location k.

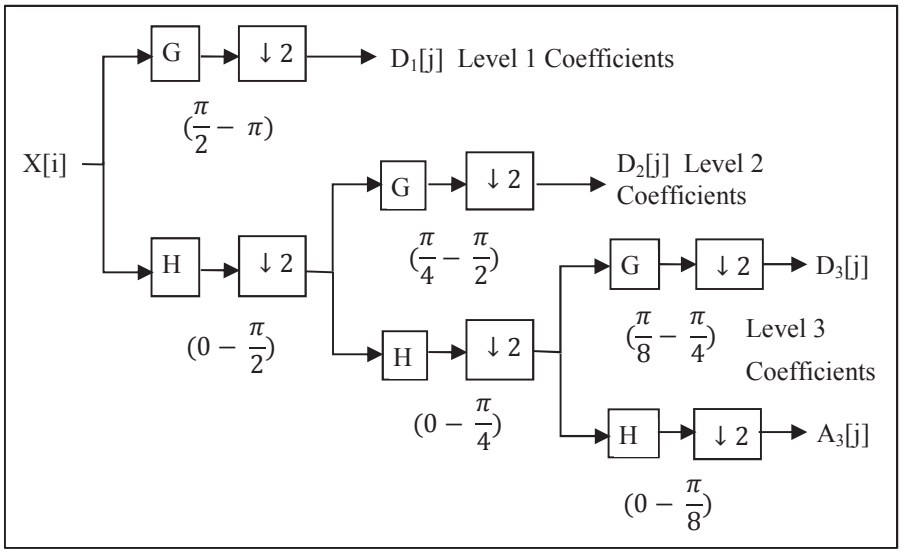


Figure 7.1a Schematic of a 3-level DWT decomposition tree

Discrete Wavelet Transform is implemented using Mallat algorithm where filter banks are used. A signal is decomposed into different frequency bands by successive high pass and low pass filtering of the time domain signal followed by subsampling. A schematic representation of a three-level DWT decomposition is shown in Fig.7.1a and the corresponding frequency characteristics is shown in Fig.7.1b. The discrete sequence X[i] is passed through a half band, high pass filter G and a low pass filter H, at each level (L) and is then down sampled by 2 to produce the detail information D_L and the coarse information A_L . A_L which is the low frequency components is again passed through the filters G and H to produce D_{L+1} and $A_{L+1.}$ The decimation by 2 is done since the filtered sequence at each level will have a frequency span one-half that of the original sequence and can be represented by half the number of samples. As a result DWT provides good time resolution at high frequencies and good frequency resolution at low frequencies.

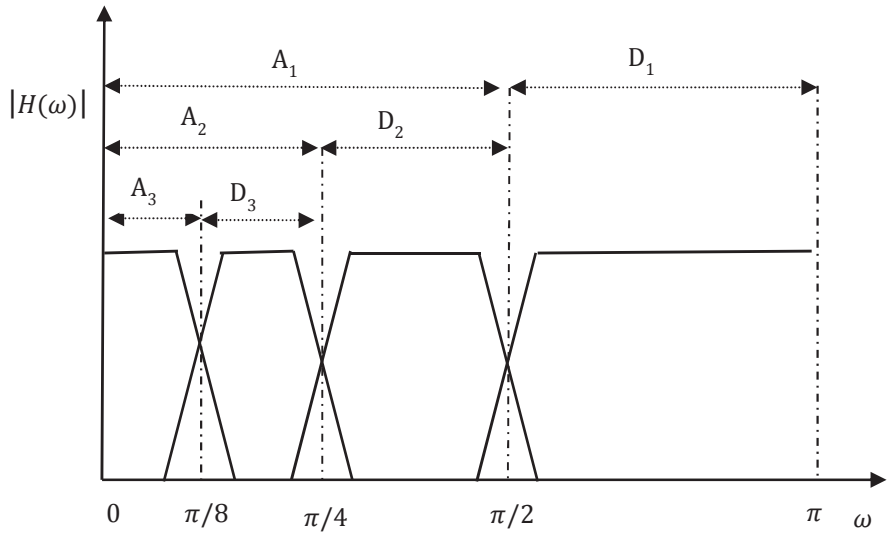


Figure 7.1b Frequency characteristic of 3-level DWT decomposition.

The filtering and decimation operation done at each level can be mathematically represented as:

$$D[j] = \sum_i X[i] \times G[2j - i]$$
$$A[j] = \sum_i X[i] \times H[2j - i] \quad (7.3)$$

where, D is the high pass detail filter output and A is the low pass coarse filter output.

7.4 Correlation analysis

The correlation function is a measure of similarity of one signal with respect to another as a function of time. The function is normalized such that its magnitude is always less than or equal to 1. The correlation coefficient R(i) is calculated as follows,

$$R(i) = \frac{\sum_0^{N-1} Y(n)X(n-i)}{\sqrt{\sum_0^{N-1} X^2(n) \times \sum_0^{N-1} Y^2(n)}} \quad (7.4)$$

where, X and Y are the two sequences and i represents the shift. The maximum value of R(i) is taken as the measure of similarity between the two sequences X and Y, denoted by $C_{xy.}$

7.5 Numerical transformation of amino acid sequences

The protein sequences are represented in the form of a sequence of characters, each representing a distinct amino acid (Fig 7.2). There are 20 different amino acids with which the proteins are formed. This alphabetic form of sequence data significantly curbs the application of mathematical

and signal processing analysis on sequence data. Various methods to circumvent this sequence metric problem have been employed by transforming the genomic sequences such as nucleotide sequence and amino acid sequence into some numerical form.

M A L W I R S L P L L A L L V F S G P G T S Y

Figure 7.2 Sequence of amino acid in a protein

In (Vaidyanathan, 2004), a method using indicator sequence for mapping DNA sequence is proposed. A numerical representation based on observation frequency is described in (Chu et al., 2004; Zhou et al., 2007). In (Glazier et al., 1995) a four dimensional pseudo random walk is used for numerical transformation. Another numerical mapping for nucleotide and amino acid sequence using complex numbers is described in Anastassiou, (2001). A numerical representation of DNA sequences using categorical periodograms is discussed in (Nair and Mahalakshmi, 2006). Ionization constant of amino acids are used for protein sequence analysis in (Cosic and Pirogova, 1998) and dielectric relaxation properties of amino acids are used in (Pirogova et al., 2003) for numerical mapping. Electron Ion Interaction Potential (EIIP) value of the amino acid is used for numerical mapping in (Lazovic, 1996; Ramachandran and Antoniou, 2008; Veljkovic et al., 1985). For a reliable representation, the numerical values assigned to each amino acid should represent the physical characteristic of the particular amino acid and should be relevant for the biological activity of these molecules (Veljkovic et al., 1985).

A multivariate statistical analysis of a large number of amino acid attributes to derive five major attribute factors and the corresponding factor

scores for amino acids is described in (Atchley et al., 2005). 494 amino acid parameters representing different physiochemical and biological properties were selected and using exploratory factor analysis a smaller number of factors that describe the structure of highly correlated variables is generated. 5 factors scores are obtained and can be used for numerical transformation of amino acid sequence into meaningful numerical equivalent. A numerical mapping scheme using Atchley factor score is used in (Vo et al., 2010; Marsella et al., 2009). A comparison of the informational capacity of various physicochemical, thermodynamic, structural and statistical parameters of amino acids are performed in (Lazovic, 1996) and it is shown that Electron Ion Interaction Potential (EIIP) is the most suitable known amino acid property that can be used in structure-function analysis of proteins.

The EIIP values for amino acids and nucleotides are calculated using the general model of pseudo potential described in (Veljkovic and Slavic, 1972):

$$\langle \vec{k} + \vec{q} | w | \vec{k} \rangle = \frac{\alpha(Z - Z_0)\sin(2\pi\beta_z\mu)}{2\pi\mu} \qquad (7.5)$$

where, q is the change in momentum of delocalized electron in the interaction with the potential w, Z is the atomic number, Z_0 is the atomic number of the inert element that begins the period, which includes the actual Z in the standard periodic table,

$$\mu = \frac{q}{2K_F} \qquad (7.6)$$

where, q is a wave number and K_F the corresponding Fermi momentum.

$$\beta_z = \frac{2(E_F)_z}{3\alpha(Z - Z_0)} \qquad (7.7)$$

where $(E_F)_Z$ is the corresponding Fermi energy.

The EIIP values of the 20 amino acids that form the linear polypeptide chain of each protein sequence (Cosic, 1994) are given in Table 7.1. Numerical sequence corresponding to the alphabetic sequence of protein is obtained by substituting the alphabets with the EIIP values.

Table 7.1 EIIP values of amino acids.

Amino acids	**Alphabet**	**EIIP**
Alanine (Ala)	A	0.0373
Cysteine (Cys)	C	0.0829
Aspartic acid (Asp)	D	0.1263
Glutamic acid (Glu)	E	0.0058
Phenylalanine (Phe)	F	0.0946
Glycine (Gly)	G	0.0050
Histidine (His)	H	0.0242
Isoleucine (Ile)	I	0
Lysine (Lys)	K	0.0371
Leucine (leu)	L	0
Methionine (Met)	M	0.0823
Asparagine (Asn)	N	0.0036
Proline (Pro)	P	0.0198

Glutamine (Gln)	Q	0.0761
Arginine (Arg)	R	0.0959
Serine (Ser)	S	0.0829
Threonine (Thr)	T	0.0941
Valine (Val)	V	0.0057
Tryptophan (Trp)	W	0.0548
Tyrosine (Tyr)	Y	0.0516

7.6 Protein sequence database

Most of the identified protein sequence data is available freely over the web at various online databases, one of which is the Entrez search and retrieval system of the National Center for Biotechnology Information (NCBI, 2017). ENTREZ is a user friendly text based cross database search and retrieval system. NCBI is a huge repository of nucleotide sequences, protein sequences, protein structures, complete genome sequences, expressed sequence tags etc. The amino acid sequences of proteins used in this study are obtained from the NCBI website. The sequences are obtained in FASTA format. FASTA format is a text based format for representing nucleotide sequences or amino acid sequences. FASTA representation of a sequence begins with a single line description starting with a '>' symbol. The description contains sequence names, accession number and any other information about the sequences. This is followed by lines of sequence data. The sequence data contains a sequence of alphabets, where each single letter corresponds to either a nucleotide or an amino acid. An example of protein sequence obtained in FASTA format is given in Fig 7.3.

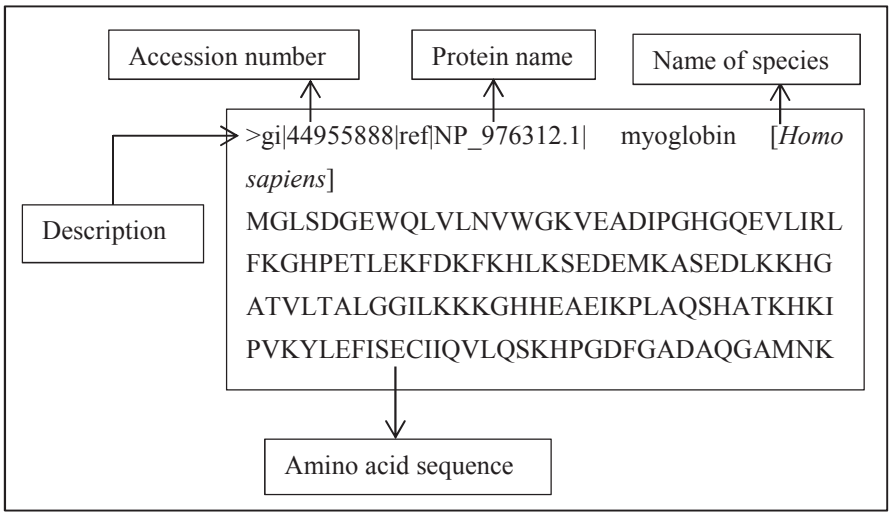


Figure 7.3 Myoglobin sequence of human being in FASTA format.

7.7 Frequency domain analysis of protein sequence similarity

Protein sequence comparison is performed to identify the similarities and dissimilarities between different protein sequences. This similarity search allows identification of amino acid residues that are critical for the biological function, structure and to infer phylogenetic relationship. In this section, a signal processing based method for protein sequence comparison and similarity search is described. The method is based on frequency domain analysis of numerically transformed protein sequences. A multi resolution analysis of protein sequence using Discrete Wavelet Transform is performed. It allows to compare the frequency spectrum of sequences at different resolutions. An L-Level DWT decomposition will provide L+1 sequences, each of which corresponds to the protein sequence component belonging to a particular band of frequency. A cross correlation between DWT coefficients of sequences at each level is performed. The correlation

coefficient thus obtained represents the measure of sequence similarity at each level.

7.7.1 Preparation of protein dataset

To study the characteristics of protein sequences in the frequency domain, multiple protein sequences from different species are obtained and analyzed. Here, two sets of sample are selected for illustrating the frequency domain characteristics of protein sequences. Each dataset is characterized by a single protein obtained from a set of species. The amino acid sequences of these proteins are obtained from the Entrez search and retrieval system of the National Center for Biotechnology Information (NCBI, 2017). The sequences are obtained in FASTA format.

Dataset 1

The first set of sample includes amino acid sequence of the protein prolactin obtained from 13 vertebrate species. The list of the species with their scientific name and common names are: *Papio anubis* (Baboon), *Bos taurus* (Bovine), *Felis catus* (Cat), *Pan troglodytes* (Chimpanzee), *Cervus elaphus* (Deer), *Gorilla gorilla* (Gorilla), *Homo sapiens* (Human), *Mus musculus* (Mouse), *Neovison vison* (Mink), *Struthio camelus* (Ostrich), *Ailuropoda melanoleuca* (Giant panda), *Rattus norvegicus* (Rat) and *Macaca mulatta* (Rhesus monkey).

Dataset 2

The second set of sample includes amino acid sequence of the protein somatotropin from 16 different vertebrate species. The list of the species with their scientific name and common names are: *Papio anubis* (Baboon), *Bos taurus* (Bovine), *Pangasianodon gigas* (Catfish), *Gallus*

gallus (Chicken), *Pan troglodytes* (Chimpanzee), *Cervus elaphus* (Deer), *Giraffa camelopardalis* (Giraffe), *Capra hircus* (Goat), *Gorilla gorilla* (Gorilla), *Homo sapiens* (Human), *Coturnix coturnix* (Quail), *Macaca mulatta* (Rhesus monkey), *Ovis aries* (Sheep), *Saimiri boliviensis boliviensis* (Squirrel monkey), *Meleagris gallopavo* (Turkey) and *Pongo abelii* (Urangutan).

7.7.2 Protein sequence similarity analysis using DWT

The sequence in the form character string is converted into numerical form using the EIIP mapping method described in section 7.5. The numerical sequence thus obtained is then normalized to zero mean and is subjected to DWT decomposition using Bior3.3 bi-orthogonal wavelets (Fig 7.4). Bior3.3 decomposition wavelet function and scaling function are very rugged and have many abrupt changes. They have been found suitable for the analysis of protein sequence which is also very rugged in nature (Trad et al., 2002). Using the method described in section 7.3, a 3 level DWT decomposition is performed to obtain coefficients corresponding to D_1, D_2, D_3 and A_3. These 4 coefficient sequences correspond to the different frequency component of the particular protein sequence with D_1, the highest frequency band and A_3, the lowest frequency band, as shown in Fig 7.1b. To measure the sequence similarity between a pair of homologous proteins, a cross correlation between DWT coefficients at each level is performed, as mentioned in section 7.4. Using Eq.(7.4), maximum value, (C_{xy}), of R(i) at each level is obtained, which gives a measure of the similarity between a pair of sequence X and Y.

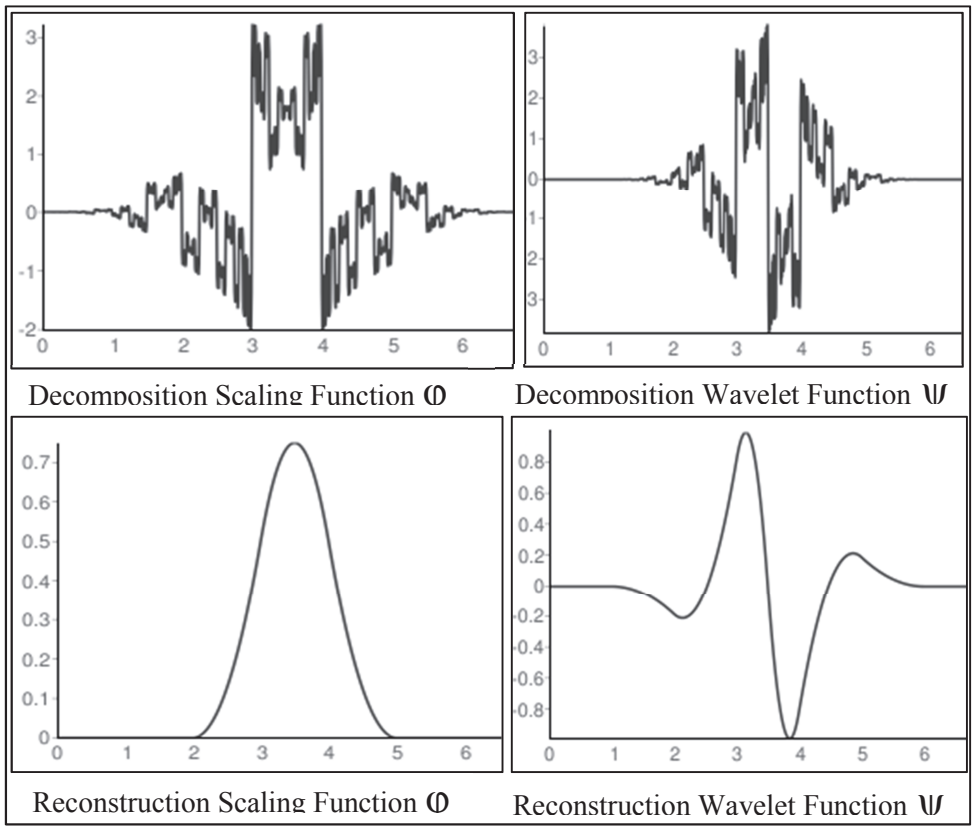


Figure 7.4 Bior3.3 Scaling and Wavelet functions

7.7.3 Application of the algorithm for similarity analysis

For the analysis of the frequency domain characteristics of the protein sequences, two datasets are employed. The dataset is designed so that similarity between closely related species as well as distantly related species can be studied. The dataset 1, given in section 7.7.1, includes prolactin sequence from 13 species. Among them, 12 species belongs to the class- mammalia and represents closely related organisms. One species, namely ostrich, belonging to the class- aves represents a species that is distantly related to others in the dataset. Ostrich, selected from another class, act as an external species. The 12 species from mammals includes

organisms belonging to 4 different order. Three experiments are designed, where pairwise similarity of every sequence is compared to:

i. Cat which belongs to the order carnivore.
ii. Human which belongs to the order primate.
iii. Ostrich which is the external species.

The first two experiments expect to obtain high similarity to species belonging to same order which are closely related. The third experiment expects to obtain very low similarity values for all species as the reference for comparison is a distantly related species.

The dataset 2 includes somatotropin sequence from 16 species. It includes 3 closely related groups of organisms and a single distant species, catfish from the class- actinopterygii. First group contains 7 species belonging to the order- primate, second has 5 species belonging to the order- artiodactyla and the third group includes 3 species from the class- aves. Four experiments are designed where pairwise similarity of every sequence is compared to:

i. Human which belongs to the order primate.
ii. Bovine which belongs to the order artiodactyla.
iii. Chicken which belongs to the class aves.
iv. Catfish which is external to all the groups and belongs to the class actinopterygii.

The first three experiments expect to obtain high similarity to closely related species. The third experiment expects to obtain very low similarity values for all species as the reference for comparison is a distantly related species.

7.7.4 Results & Discussion

Three experiments are performed on dataset 1 for the analysis of protein sequence similarity. The results of the experiments are shown in Fig 7.5(a-c). The y-axis shows the value of sequence correlation, C_{xy} ($0<C_{xy}<1$) and the legends A3, D3, D2, D1 represents the 4 DWT coefficients. In the first experiment prolactin sequences from 12 species are compared with that of cat using the DWT method. It can be observed from Fig 7.5a that cat shows highest similarity (>95%) to the organisms mink and panda, at all the 4 frequency bands. Both these organisms belong to the order- carnivora, to which the cat belongs. The second experiment compared prolactin sequence from human with all the other species. It shows similar result with the human sequence showing more than 95% similarity to baboon, chimpanzee, gorilla and rhesus monkey at all resolutions (Fig 7.5b). All the 5 species belongs to the same order- primate. The third experiment compared the sequence of ostrich with other species and the result clearly supported the expectations where it showed no significant sequence similarity with any of the species at any DWT levels (Fig 7.5c).

The second dataset involves 4 experiments and the results are shown in Fig 7.6(a-d). In the first experiment human somatotropin sequence is compared with the 15 other species' sequence. It can be seen from Fig. 7.6a, that all the 6 primates shows high sequence similarity values compared to other groups. The second experiment compares bovine sequence with the rest. The result shows, (Fig 7.6b) almost 100% similarity with all the 4 species belonging to the same order- artiodactyla. In the third experiment somatotropin sequence of chicken is compared with other sequences. The result obtained is shown in Fig 7.6c, where it clearly shows strong

correlation with the other two aves in the set, quail and turkey at all DWT coefficient levels.

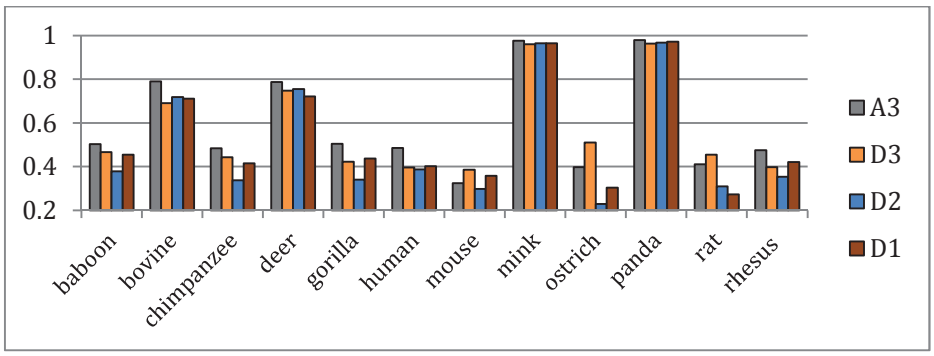


Figure 7.5 (a) Similarity of prolactin from cat with other species

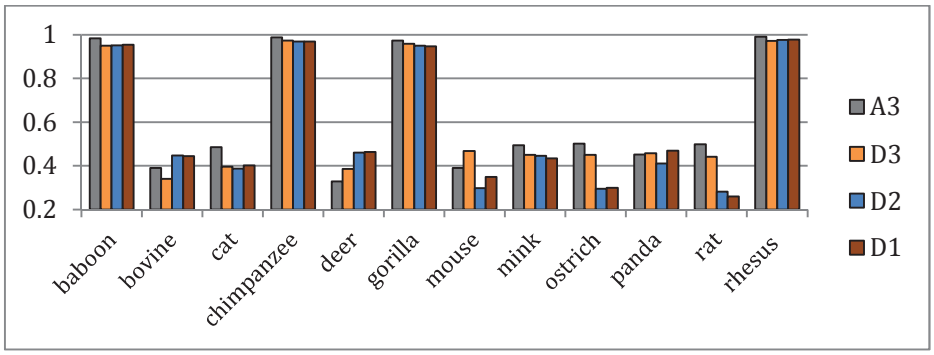


Figure 7.5 (b) Similarity of prolactin from human with other species

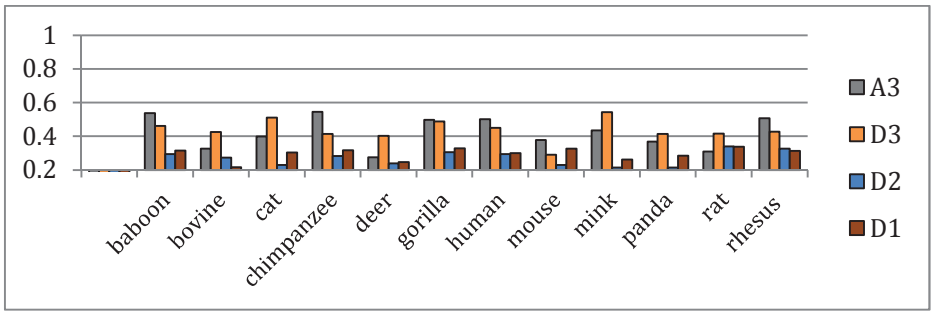


Figure 7.5 (c) Similarity of prolactin from ostrich with other species

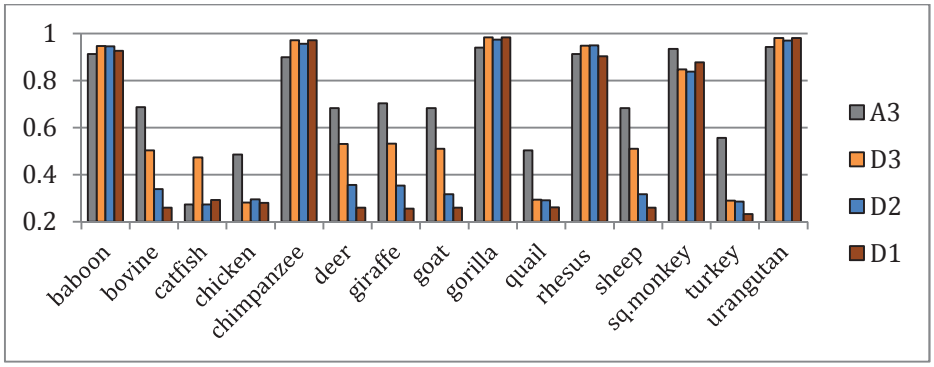


Figure 7.6 (a) Similarity of somatotropin from human with other species

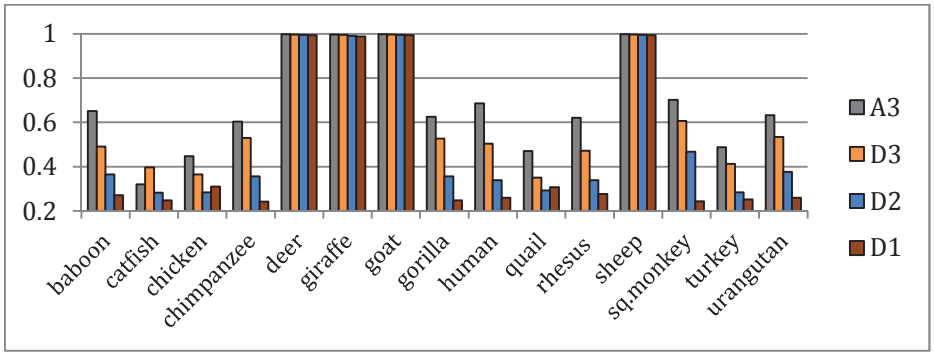


Figure 7.6 (b) Similarity of somatotropin from bovine with other species

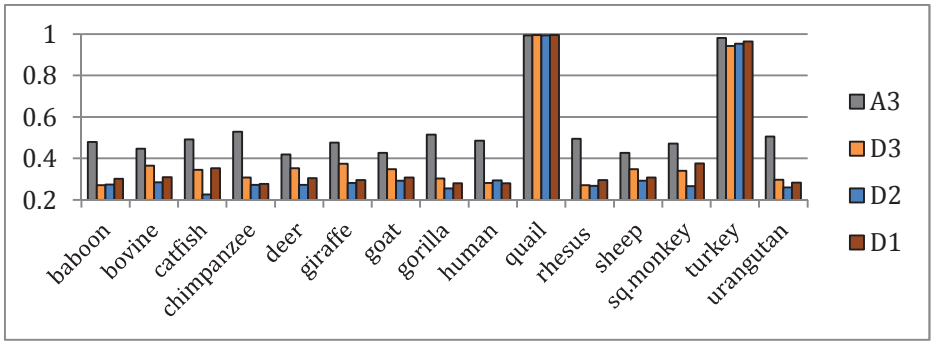


Figure 7.6 (c) Similarity of somatotropin from chicken with other species

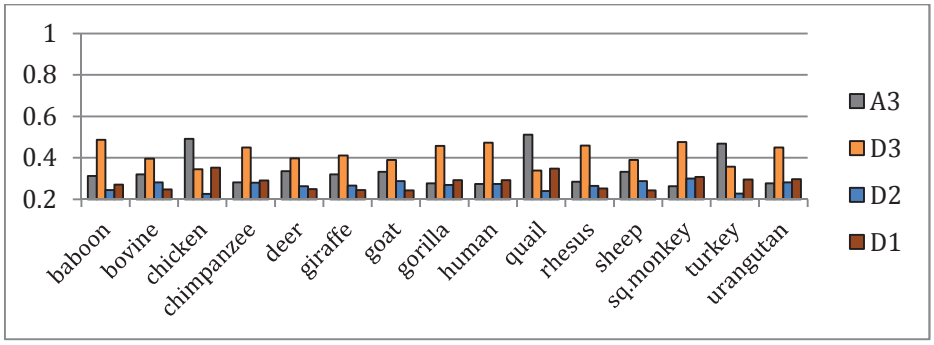


Figure 7.6 (d) Similarity of somatotropin from catfish with other species

The final experiment compares the sequence from a species, catfish with the others. The catfish is a species that is external to all the groups and represents a sample that is distantly related to all other species. As expected, the catfish somatotropin sequence showed low sequence similarity with all other sequences at all levels. The study clearly indicates the presence of strong spectral similarity in closely related species and also showed that the organisms that are distant in terms of evolutionary relationship have less sequence similarity. Similar result is also obtained using an analysis of sequence similarity using DFT of EIIP mapped amino acid sequence as described in appendix A.

The protein sequence similarity study in frequency domain showed that closely related organisms have very strong sequence similarity and the similarity reduces, as the diversity among the species under study increases. The study poins to the fact that, a quantitative measure of the frequency domain similarity between amino acid sequences of protein, that are numerically mapped using EIIP method, can act as a reliable genetic distance measure. The result of this frequency domain analysis of protein sequence similarity forms a basis for study of the evolutionary relationship between various organisms. In the next section a new method to calculate

genetic distance using the frequency domain similarity is introduced and is used for inferring phylogenetic relationship between organisms.

7.8 Single Protein Power Spectral Density (SPPSD) method for phylogenetic classification

The SPPSD method constructs a phylogenetic tree representing the evolutionary relationship among a group of species, using a frequency domain approach. The algorithm involves three steps. In the first step, amino acid sequence of a protein is obtained from a set of species and is transformed into a numerical sequence using the EIIP method described in section 7.5. In the second step, the pair wise distance between sequences is estimated from this numerical signal using the power spectral density functions of protein sequences. Finally, phylogenetic tree representing the evolutionary relationship is then constructed by UPGMA method using the genetic distance between the sequences. A confidence measure for the branching pattern obtained is calculated using bootstrap analysis for assessing the performance of the method.

7.8.1 Calculation of genetic distance from numerical sequence

Fourier transforms allows a signal which is a function of time or space to be represented by a sum of sinusoids having different frequencies and amplitudes. Such a representation of signal in frequency domain showing the frequencies that constitute the signal is called the spectrum of a signal. The Power Spectral Density (PSD) of a signal indicates the distribution of power of a signal as a function of frequency. PSD of a signal can also be defined as the fourier transform of the autocorrelation function. DWT analysis of protein sequences detailed in the previous section forms

the basis for genetic distance calculation using frequency domain method described here.

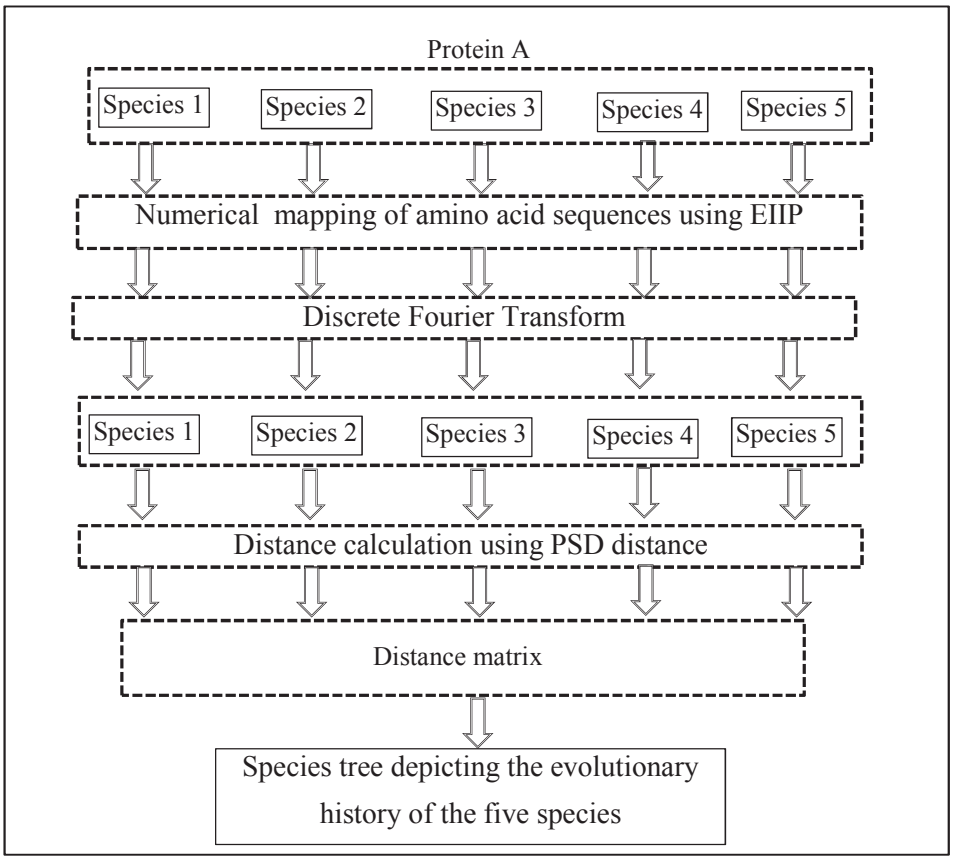


Figure 7.7a Schematic representation of the SPPSD method

Distance between Power Spectral Density

The genetic distance between two sequences is obtained from the distance between the PSD of the two sequences. The sequences X and Y, are transformed to frequency domain using DFT. X[k], Y[k] represents the frequency spectrum and P_X, P_Y represents the PSD of the two sequence. An estimate of PSD of the signal is obtained by taking the squared magnitude of fourier transform. The amount of difference between the PSD of sequences

is calculated using manhattan distance. The manhattan distance, D_{XY}, between the PSD of two sequences P_X and P_Y is defined as:

$$D_{XY} = \sum_{k} |P_X[k] - P_Y[k]| \quad (7.8)$$

The value, D_{XY}, represents power spectral distance between the pair of sequences. The distances between all the sequences are obtained and a distance matrix of all pair wise genetic distances is computed using this approach.

7.8.2 Phylogenetic tree construction using UPGMA method

UPGMA or Unweighted Pair Group Method with Arithmetic average (Sokal and Michener, 1958) is a clustering method that generates rooted ultrametric trees. It computes pair wise distance between all sequences and then selects the closest pair of sequences as siblings. The two nodes are then combined into a single node and the distance values in the matrix are updated with the new node. The process is repeated until all nodes are grouped and a tree is constructed with all the sequences. UPGMA method includes the following steps:

1) Identify the pair of nodes (A,B) with the smallest distance.
2) Group the nodes A and B into a single node AB.
3) Calculate the distance between the new node, AB from every other node, X denoted by $d(AB, X)$ as given below:

$$d(AB, X) = \frac{d(A, X) + d(B, X)}{2} \quad (7.9)$$

4) Remove the nodes A and B and update the distance matrix with new node.
5) Replace the distances with new distances.
6) Repeat from step 1 onwards until all nodes are grouped.

7.8.3 Bootstrap analysis

Bootstrap analysis is performed separately for each dataset. To perform the bootstrap analysis on a dataset, each protein sequence in the set is replicated N times. Here the value of N is chosen as 1000. Bootstrap replicate data sequences are obtained by random sampling of the original sequence with replacements. The replicate data sequences also have the same dimension as that of the original. Then using the SPPSD method, a phylogenetic tree is constructed for each replicate dataset. The N replicate dataset thus results in N trees. Using the N trees, a consensus tree is generated and the percentage of the trees in the set that has the same pattern for each branching is calculated. This percentage value denotes the bootstrap support for a particular branching pattern. A bootstrap support of 100% for a branching indicates that the particular branching is present in all the 1000 trees. Similarly a value of 50% means that the branching can only be found in 500 out of the 1000 replicate dataset.

7.9 Implementation of the SPPSD method to infer phylogeny

The amino acid sequence of the protein under study is obtained in alphabetical form. It is then transformed into a numerical sequence using EIIP values of the corresponding amino acids. The numerical sequence thus obtained is normalized to have a zero mean. The sequence is then transformed to frequency domain using FFT or Fast Fourier Transform. Using the fourier spectrum, the PSD of the sequence is obtained. The manhattan distance between a pair of PSD function is taken as a measure of genetic distance between the sequences. All the pair wise distances between sequences are obtained and are normalized such that $0 \leq D_{xy} \leq 1$. Using these pair wise distances, the distance matrix is generated and is used for

phylogenetic tree construction using UPGMA method. The different steps in the algorithm are given below:

1) Obtain amino acid sequences from different species.
2) Convert the alphabetical sequence into numerical equivalent using EIIP values.
3) Transform the normalized sequence into frequency domain using DFT.
4) Estimate the PSD of each sequence.
5) Calculate the Manhattan distance between a pair of PSD function to obtain the genetic distance.
6) Calculate the genetic distances between all pair of sequences using the power spectral distance.
7) A distance matrix with all pair wise distances is generated.
8) Using the distance matrix, Phylogenetic tree is constructed by UPGMA method.
9) Perform bootstrap analysis to evaluate the reliability of the tree topology.

7.10 Results & Discussion

Different set of samples are used to validate the capability of the new SPPSD method to infer the phylogenetic relationship using amino acid sequences corresponding to a single protein. The results obtained using SPPSD method is compared with three other established sequence alignment based phylogenetic analysis tools namely, CLUSTALW, COBALT and MEGA. Sequence alignment based methods are selected for comparison, as they are widely considered to be more accurate and reliable that alignment free approaches. Moreover, these selected methods are widely accepted and are popular tools for phylogenetic analysis.

CLUSTALW

CLUSTALW (Thompson et al., 1994) is one of the most widely used multiple sequence alignment program for DNA or protein sequences. It employs a progressive alignment technique, where most similar sequence pairs are first aligned, followed by alignment of the most similar pairs of collection of sequences. This is continued until all the sequences are included.

COBALT

It is a constraint based alignment tool for multiple alignment of protein sequences (Papadopoulos and Agarwala., 2007) available on NCBI website (NCBI, 2017). COBALT finds a collection of pairwise constraints derived from conserved domain database, protein motif database, and sequence similarity. It then combines these pairwise constraints, and incorporates them into a progressive multiple alignment.

MEGA

Molecular Evolutionary Genetics Analysis (MEGA) software performs comparative analysis of DNA and protein sequences for inferring evolutionary pathway and for creating phylogenetic trees (Kumar et al., 2016). MEGA software also includes the facility for performing bootstrap analysis to verify the strength of support for the branching pattern obtained.

7.10.1 Dataset 1

This set of data includes amino acid sequence of the protein myoglobin. The protein sequence is obtained from 6 different species belonging to the order- primates, which belong to the class- mammalia. The

accession number of the sequences, scientific name and common names of the species are given below.

A. P68084.2 - *Papio anubis* (Olive Baboon)
B. P02145.2 - *Pan troglodytes* (Chimpanzee)
C. P62734.2 - *Hylobates agilis* (Gibbon)
D. P02147.2 - *Gorilla beringei* (Eastern Gorilla)
E. NP_976312.1 - *Homo sapiens* (Human)
F. P02148.2 - *Pongo pygmaeus* (Bornean Urangutan)

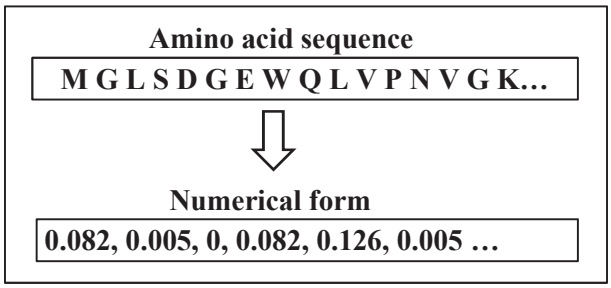


Figure 7.7b Numerical transformation of amino acid sequence

Table 7.2 Pair wise genetic distance between the 6 species in dataset 1

	A	B	C	D	E	F
A	0	0.965472	0.728923	1	0.904063	0.728923
B	0.965472	0	0.629851	0.496673	0.370552	0.629851
C	0.728923	0.629851	0	0.673157	0.567605	0
D	1	0.496673	0.673157	0	0.437633	0.673157
E	0.904063	0.370552	0.567605	0.437633	0	0.567605
F	0.728923	0.629851	0	0.673157	0.567605	0

All amino acid sequences are obtained in the FASTA format which is a sequence of alphabets. It is converted into numerical form using EIIP values as shown in Fig.7.7b. Using the methods described in section 7.8.1,

the genetic distance between each pair of protein sequence is obtained. A distance matrix containing all pair wise distance is thus formed for the 6 species and is shown in Table 7.2. The distance matrix is used to construct a phylogenetic tree of the 6 species using UPGMA method. The constructed phylogenetic tree depicting the evolutionary relationship between the species is given in Fig. 7.8a.

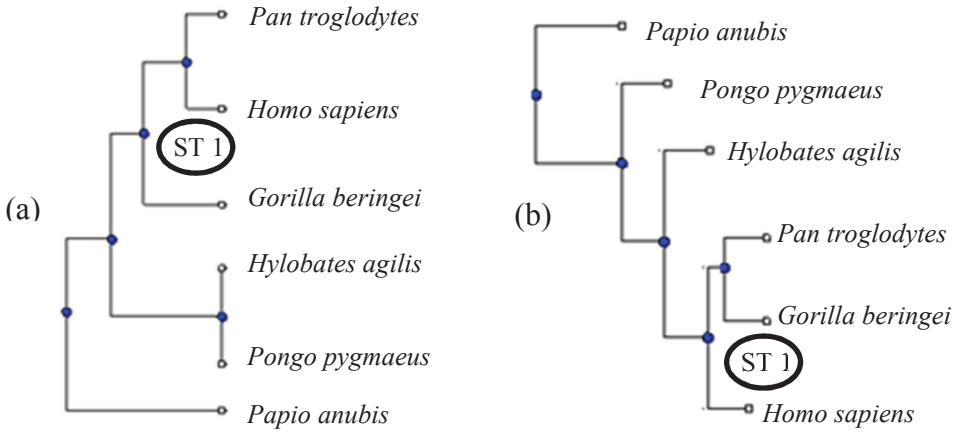


Figure 7.8 Phylogenetic tree inferred using dataset 1 by (a) SPPSD (b) COBALT

The UPGMA method generates ultrametric trees where the branch length gives an indication about the time of divergence. In the trees shown here, the original branch lengths obtained is not maintained for ease of representation as it is difficult to follow the same scale in all the figures due to space constraints. Hence the branch lengths do not represent the exact evolutionary distance between the species in these figures.

In order to evaluate the topology of the phylogenetic tree obtained using the SPPSD method, it is compared with tree obtained using standard and established sequence alignment methods. The methods used for comparison are COBALT, CLUSTALW and MEGA. The trees constructed using the three methods are given in Fig. 7.8 [b-d]. It can be observed that

the tree topology of the COBALT tree (Fig 7.8b) is different from the rest. The remaining three trees have almost similar topology. A variation is observed in the branching and grouping of 3 species included in the sub tree labeled ST1. It involves three species, *Pan troglodytes*, *Homo sapiens* and *Gorilla beringei*. The phylogenetic tree constructed by SPPSD method and MEGA (Fig 7.8d) groups *Pan troglodytes* and *Homo sapiens* together as most closely related species while, CLUSTALW (Fig 7.8c) groups *Gorilla beringei* and *Homo sapiens* together. The tree topology obtained by our method is exactly same as that of MEGA. It is also similar to the one by CLUSTALW in topology except for the subtree ST1.

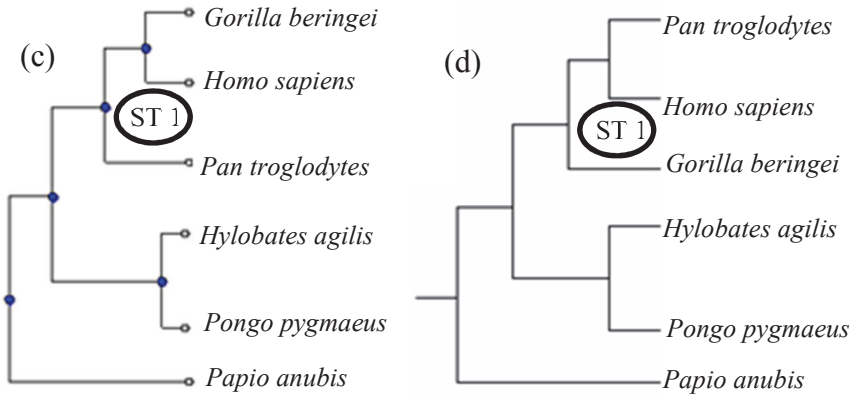


Figure 7.8 Tree for dataset 1 using (c) CLUSTALW (d) MEGA

The reliability of the tree structure is measured using a bootstrap analysis. The bootstrapped tree and the confidence values obtained for the branching pattern, using 1000 replicates for SPPSD method is shown in Fig 7.9a. A similar bootstrap analysis using MEGA is performed for comparison and is shown in Fig 7.9b. A confidence value of 50% is the minimum support for a reliable branching. It means that if we perform tree construction for 1000 replicate sequences, we expect to obtain a particular grouping of nodes 500 times. The subtree marked 'ST1' include 3 species,

Gorilla, Human and chimpanzee. It can be noticed that subtree ST1 is supported by a value of 33% in the tree generated by SPPSD method while it has only a support of 29% in MEGA bootstrap tree. It indicates that the lack of reliability of the subtree ST1. This explains the difference of topology of subtree ST1 among different trees. The bootstrap analysis also shows the improvement in reliability of SPPSD method in terms of the confidence values. All the remaining branches have more than 50% support in tree created using SPPSD, while the tree constructed by MEGA fails to match this value for all the branches.

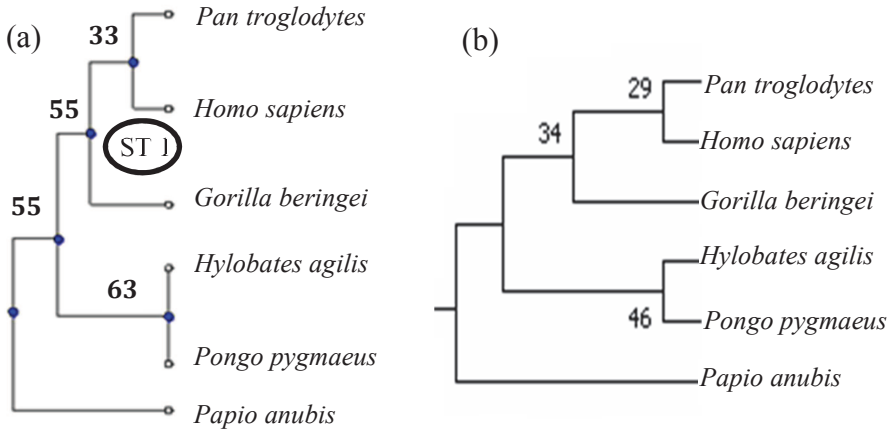


Figure 7.9 (a) Bootstrap tree for dataset 1 using SPPSD method (b) MEGA bootstrap tree and confidence values

7.10.2 Dataset 2

This set of data includes amino acid sequence of the protein myoglobin. The protein sequence is obtained from 10 different species which are more diverse than the first dataset. They belong to 2 different orders in the class-mammalia. The accession number of the sequences, scientific name and common names of the species are given below.

A. NP_976312.1 - *Homo sapiens* (Human)
B. P02151.2 - *Aotus trivirgatus* (Night Monkey)
C. P02152.2 - *Callithrix jacchus* (Common Marmoset)
D. P02145.2 - *Pan troglodytes* (Chimpanzee)
E. P68084.2 - *Papio anubis* (Olive Baboon)
F. NP_001072126.1 - *Ovis aries* (Sheep)
G. NP_776306.1 - *Bos taurus* (Cattle)
H. P68279.2 - *Tursiops truncates* (Bottlenose Dolphin)
I. P02173.2 - *Orcinus orca* (Killer Whale)
J. P02185.2 - *Physeter catodon* (Sperm Whale)

As described above, the phylogenetic tree for the 10 species is obtained using SPPSD method and the three other methods. The resulting phylogenetic trees are shown in the Fig 7.10 [a-d].

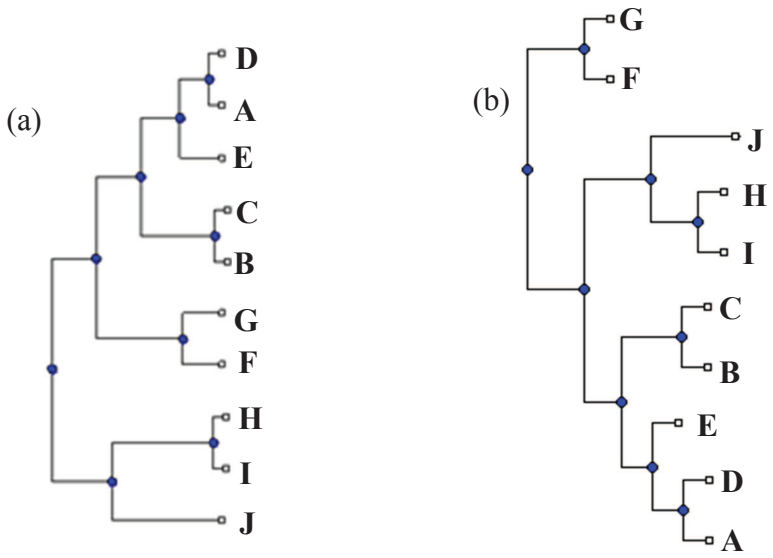


Figure 7.10 Tree for dataset 2 using (a) SPPSD (b) COBALT

A comparison of the phylogenetic trees obtained using the different methods shows that the tree obtained using SPPSD method is exactly same

as that obtained using MEGA. The topology of these tree, differ with the ones obtained using CLUSTALW and COBALT.

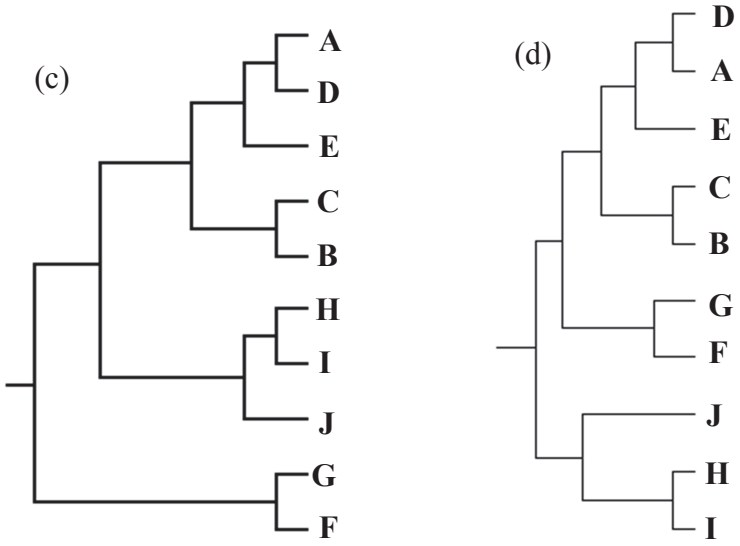


Figure 7.10 Tree for dataset 2 using (c) CLUSTALW (d) MEGA

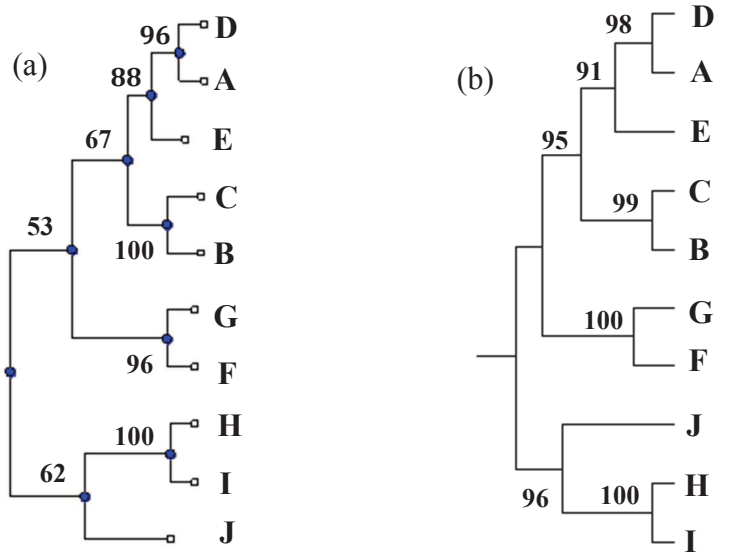


Figure 7.11 Bootstrap tree for dataset 2 using (a) SPPSD (b) MEGA

In SPPSD method the species belonging to aquatic mammals including, sperm whale, killer whale and bottle nose dolphin are identified as the group that is distant from others. But in the other two methods, the group containing cattle and sheep is identified as thedistant one. The former grouping has better conformance with taxonomical classification of these organisms. Bootstrap analysis using SPPSD method and MEGA also supports the corresponding tree topologies with good confidence values (Fig. 7.11 [a-b]).

7.10.3 Dataset 3

This dataset includes amino acid sequence of the protein Caveolin-1. The protein sequence is obtained from 7 different species. They belong to 3 different orders in the class- mammalia. The accession number of the sequences, scientific name and common names of the species are given below.

A. AAV83691.1- *Bos taurus* (Cattle)
B. AAR16246.1- *Pan troglodytes* (Chimpanzee)
C. AAH09685.1- *Homo sapiens* (Human)
D. Q2QLG6.1- *Callithrix jacchus* (Common marmoset)
E. AAR16290.1- *Mus musculus* (House mouse)
F. AAR16308.1- *Rattus norvegicus* (Norway rat)
G. ABI75290.1- *Ovis aries* (Sheep)

The phylogenetic tree obtained for the 7 species using the four different methods is shown in Fig 7.12[a-d]. The structure of the all other 3 trees is same and is different from that of SPPSD. The relationship shown by SPPSD, groups the species belonging to order glires and primates as most closely related. This grouping is more accurate as both of them belong to a

common superorder, euarchontoglires. The result of SPPSD has a better match with the taxonomic classification of the organisms compared to others.

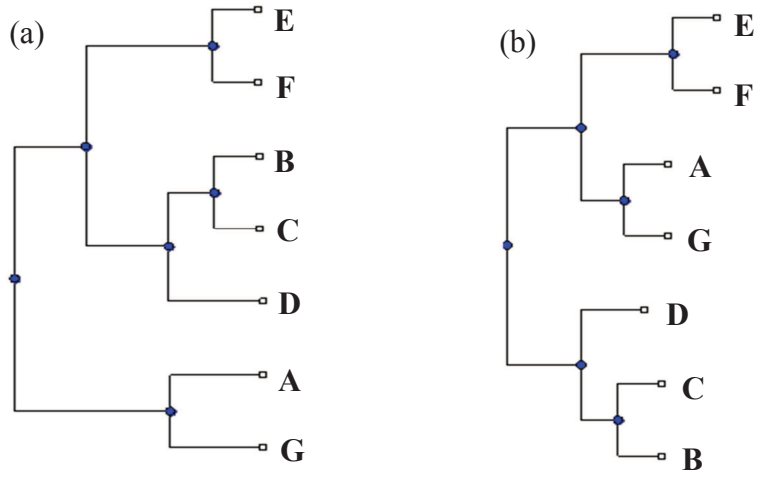


Figure 7.12 Tree for dataset 3 using (a) SPPSD (b) COBALT

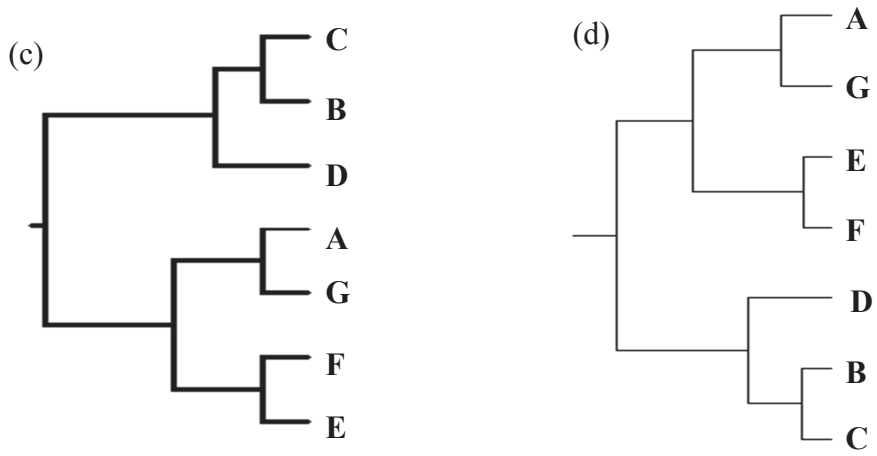


Figure 7.12 Tree for dataset 3 using (c) CLUSTALW (d) MEGA

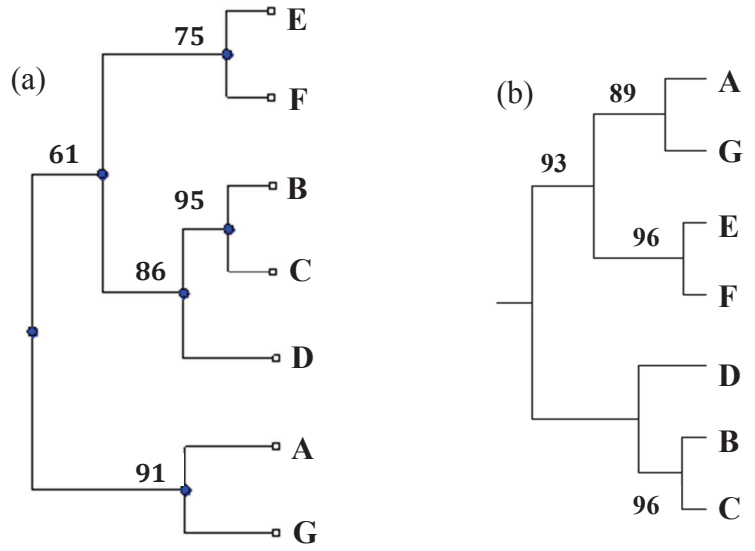


Figure 7.13 Bootstrap tree for dataset 3 using a) SPPSD b) MEGA

7.10.4 Dataset 4

This dataset includes amino acid sequence of the protein prolactin. The protein sequence is obtained from 6 different species. They belong to 3 different orders in the class- mammalia. The accession number of the sequences, scientific name and common names of the species are given below.

A. AAI48125.1- *Bos taurus* (Cattle)
B. P46403.1- *Felis catus* (Cat)
C. ADK11290.1- *Canis lupus familiaris* (Dog)
D. P06879.1- *Mus musculus* (Mouse)
E. P01237.1- *Rattus norvegicus* (Norway rat)
F. CAA53635.1- *Ovis aries* (Sheep)

The phylogenetic tree obtained for the 6 species using the all the four methods is having same topology and is shown below in Fig 7.14a. The

bootstrap tree obtained using SPPSD and MEGA are also exactly the same and is given in Fig. 7.14b

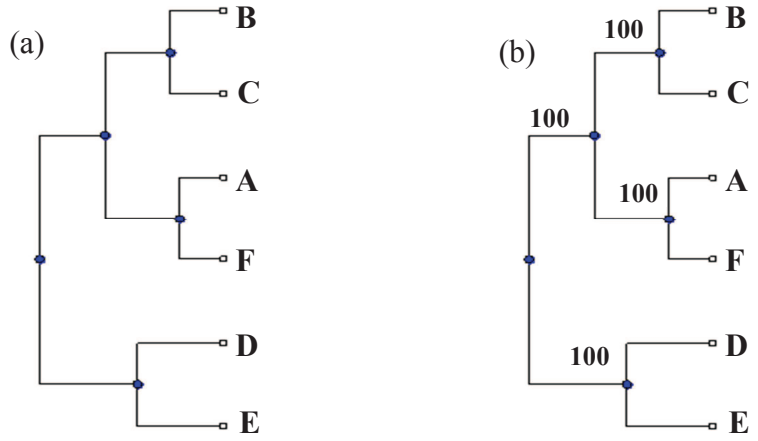


Figure 7.14 Dataset 4 (a) Tree topology for all methods. (b) Bootstrap tree.

The phylogenetic analysis of this particular set of protein sequences resulted in exactly the same tree topology for all the four methods employed. The bootstrap analysis also offered excellent results, where the entire bootstrap replicate resulted in the same tree topology giving a 100 % confidence in the structure.

7.10.5 Dataset 5

This dataset includes amino acid sequence of the protein somatotropin. The protein sequence is obtained from 11 different species. They belong to 4 different orders in the class- mammalia. The accession number of the sequences, scientific name and common names of the species are given below.

A. P01246.1- *Bos taurus* (Cattle)
B. P46404.1- *Felis catus* (Cat)

C. XP_004041225.1- *Gorilla gorilla gorilla* (Gorilla)
D. P01241.2- *Homo sapiens* (Human)
E. Q9GMB3.1- *Callithrix jacchus* (Common marmoset)
F. P06880.1- *Mus musculus* (House mouse)
G. P01244.1- *Rattus norvegicus* (Norway rat)
H. P33093.2- *Macaca mulatta* (Rhesus monkey)
I. NP_001009315.2- *Ovis aries* (Sheep)
J. P58343.1- *Saimiri boliviensis boliviensis* (Bolivian squirrel monkey)
K. XP_002827754.1- *Pongo abelii* (Sumatran orangutan)

The phylogenetic tree obtained for the 11 species using the four methods is shown in Fig 7.15[a-d]. A comparison of the 4 phylogenetic trees obtained shows that there is difference in the tree topology in two subtrees. The subtree ST1 consisting of C, K and D has two different topologies. SPPSD method groups the species C & K as closely related which is same as obtained by 2 other methods, MEGA and CLUSTALW. Only COBALT groups C & D as closely related. The second difference is with the subtree ST2 consisting of 5 species, A, I, B, F& G. Of the 4 methods, SPPSD method and CLUSALW generates same tree topology where A, I and B are included in a single clade with a common ancestor. The other two methods estimate a different topology where F, G and B forms a clade with a common ancestor. Bootstrap analysis using SPPSD method (Fig 7.16a) supports the grouping of C with K with a confidence value of 63 %. Meanwhile, the bootstrap analysis using MEGA (Fig 7.16b) shows a grouping of C with D with a confidence of 49%, which contradicts the phylogenetic tree, where C is paired with K. For the subtree ST2, bootstrap analysis using SPPSD method resulted in a 100 % confidence for the branching pattern while the bootstrap analysis using MEGA provided a

support of 66% for the alternate subtree structure estimated by MEGA and COBALT trees.

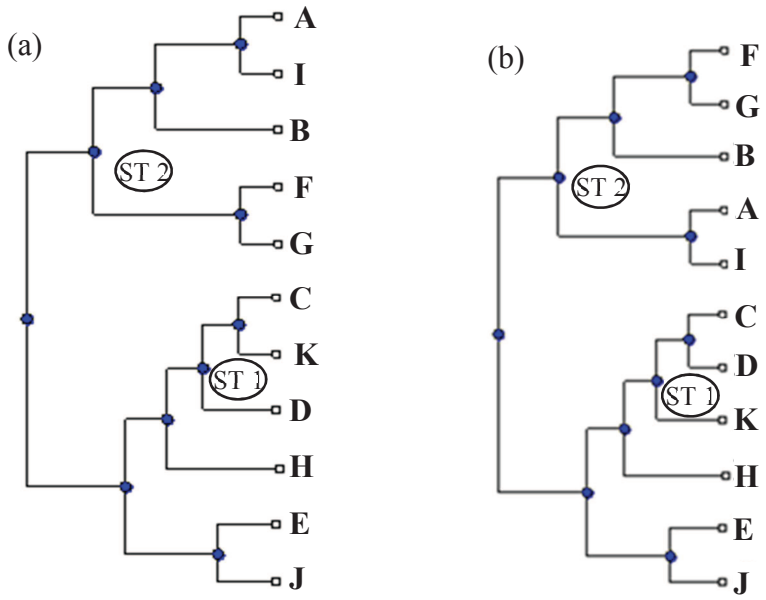


Figure 7.15 Tree for dataset 5 using (a) SPPSD. (b) COBALT

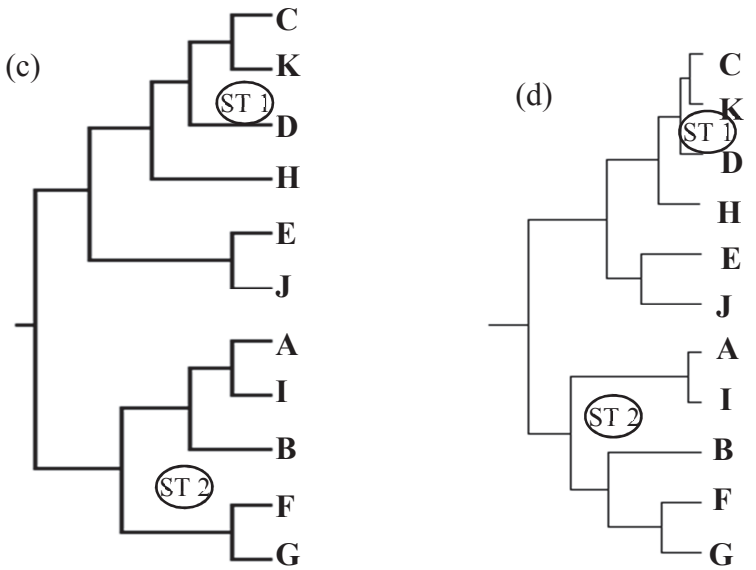


Figure 7.15 Tree for dataset 5 using (c) CLUSTALW (d) MEGA

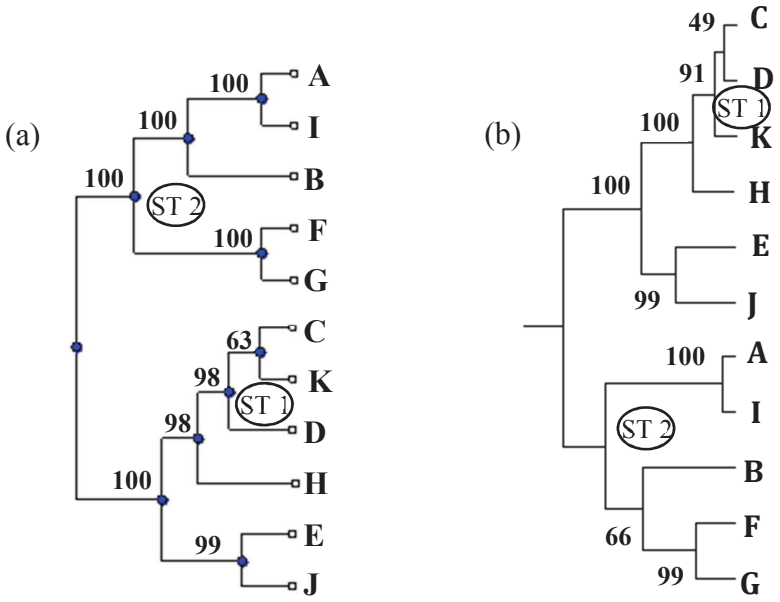


Figure 7.16 Bootstrap tree for dataset 5 using a) SPPSD b) MEGA

7.10.6 Dataset 6

This dataset includes amino acid sequence of the protein somatotropin. The protein sequence is obtained from 17 different species. They belong to 4 different orders under the class- mammalia and 2 orders under the class- Aves. The accession number of the sequences, scientific name and common names of the species are given below.

A. P01246.1- *Bos taurus* (Cattle)
B. P46404.1- *Felis catus* (Domestic cat)
C. P08998.2- *Gallus gallus* (Chicken)
D. P58756.1-*Pan troglodytes* (Chimpanzee)
E. P56437.1- *Cervus elaphus* (Red deer)
F. P33711.2-*Canis lupus familiaris* (Dog)

G. P67931.1- *Capra hircus* (Goat)
H. Q7YQB8.1- *Hippopotamus amphibius* (Hippopotamus)
I. P01241.2- *Homo sapiens* (Human)
J. P06880.1- *Mus musculus* (House mouse)
K. Q9PWG3.1- *Struthio camelus* (African ostrich)
L. Q8HYE5.1- *Ailuropoda melanoleuca* (Giant panda)
M. P01248.2- *Sus scrofa* (Pig)
N. P01244.1- *Rattus norvegicus* (Norway rat)
O. P33093.2- *Macaca mulatta* (Rhesus monkey)
P. NP_001009315.2- *Ovis aries* (Sheep)
Q. P22077.1- *Meleagris gallopavo* (Turkey)

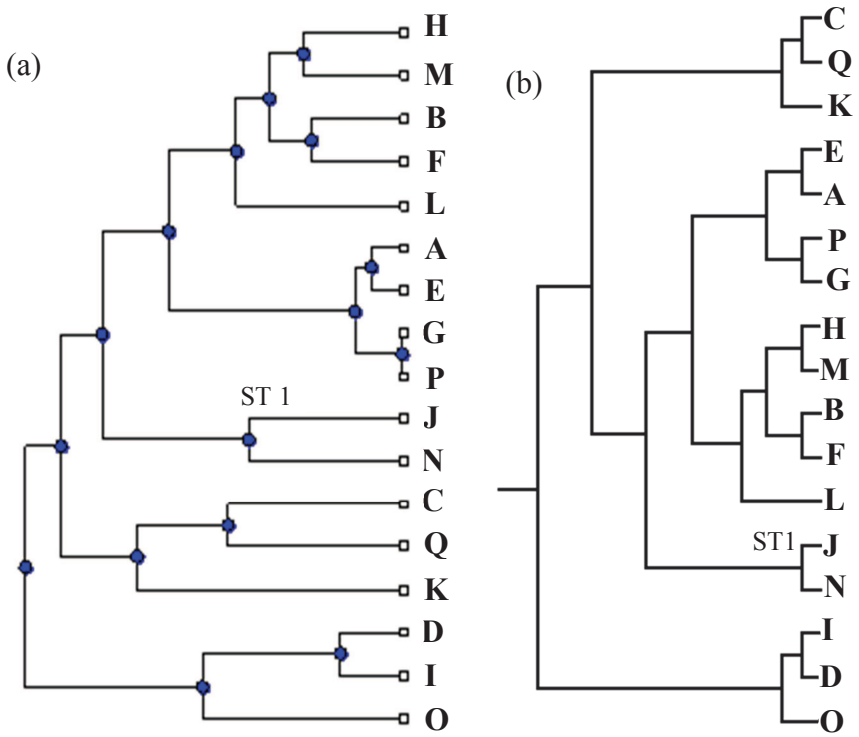


Figure 7.17 Tree for dataset 6 using (a) SPPSD (b) CLUSTALW

The phylogenetic tree obtained for the 17 species using the four methods is shown below in Fig. 7.17[a-d]. The phylogenetic tree obtained by SPPSD method for the dataset is exactly the same as that obtained using CLUSTALW. All the branching patterns follow similar pattern. The tree obtained using MEGA is also similar in topology except for the position of subtree ST1, consisting of two species J & N. The tree topology obtained using COBALT offers a very different branching pattern as that of all the other methods. The bootstrap analysis using SPPSD method (Fig 7.18a) and MEGA (Fig 7.18b) both supports the difference in the tree structures with low confidence values of 39% and 52% support which justifies the estimation of two different branching patterns for the subtree ST1.

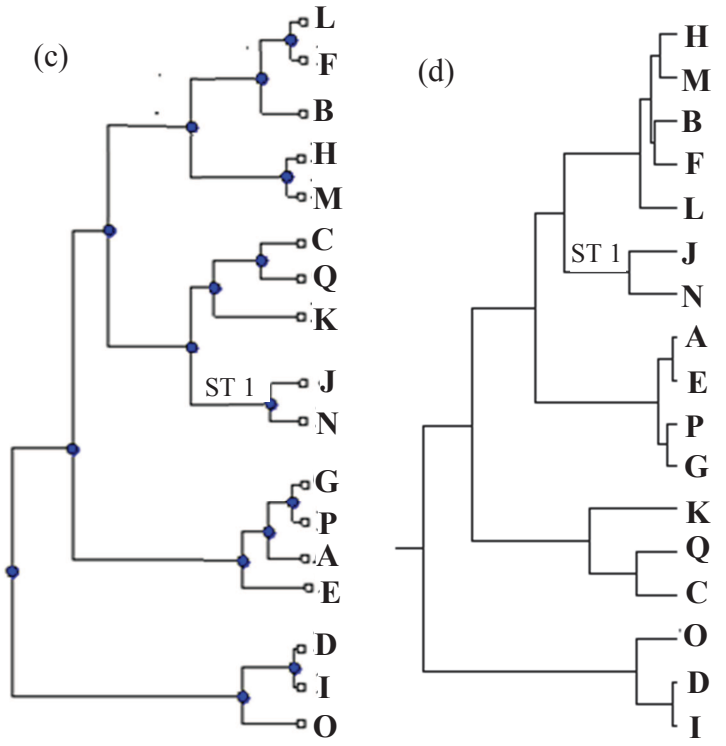


Figure 7.17 Tree for dataset 6 using (c) COBALT (d) MEGA

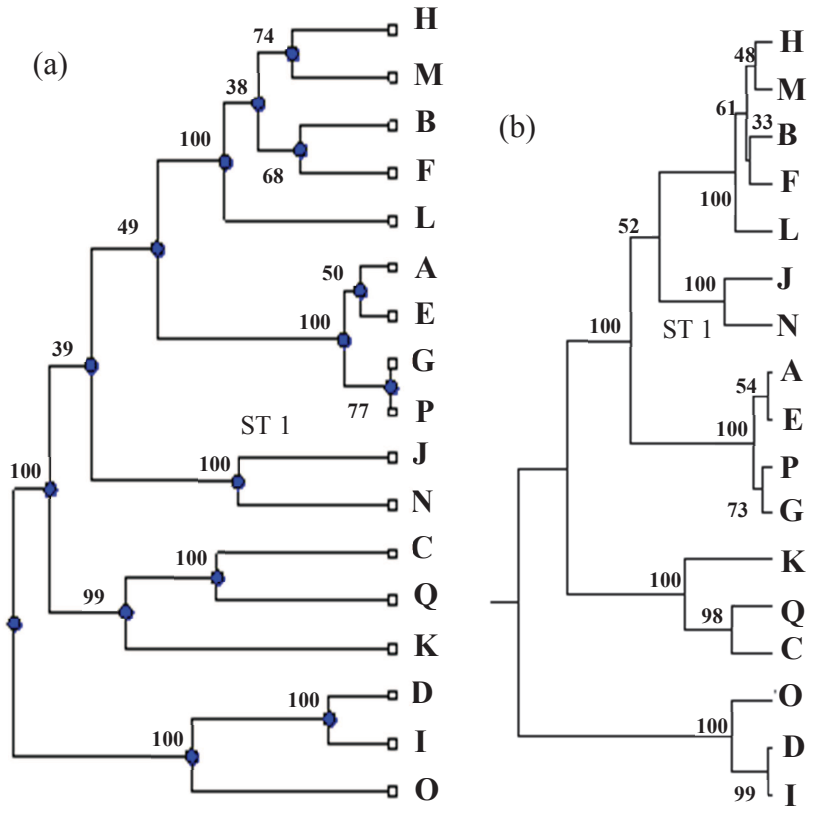


Figure 7.18 Bootstrap tree for dataset 6 using a) SPPSD b) MEGA

7.11 Summary

Understanding the evolutionary pattern of species as well as that of individual genes depends mainly on the analysis of similarities and variations in various characteristics. With the advancement of molecular technologies the sequence based analysis has gained wide attention. The molecular phylogenetic analysis methods focus on measurement of genetic distance between the sequences to uncover the amount of divergence between them. Amino acid sequences of proteins are used in this study. The

alphabetical sequence is converted to a numerical equivalent using EIIP method for applying signal processing algorithms. A frequency domain analysis of this numerical sequence is performed using DWT. The analysis of similarity between the sequences showed significant spectral similarity between closely related species and very poor spectral correlation for distantly related species. Based on the observation a new signal processing based approach for inferring evolutionary relationship using a single protein sequence obtained from a collection of organisms is developed. The method uses an alignment free approach to measure protein sequence similarity. The method named, SPPSD (Single Protein Power Spectral Density), uses the distance between power spectral densities (PSD) of numerically transformed protein sequences as a measure of genetic distance for the construction of phylogenetic tree. The method is applied to six datasets for validation. The tree generated using the SPPSD method for the sample datasets clearly illustrated the ability to capture the pattern of divergence of the individual proteins. The results are also compared with trees generated with methods such as, COBALT, CLUSTALW and MEGA. The reliability of the trees generated is again evaluated using a bootstrap analysis and the results are compared with the bootstrap analysis using MEGA. It can be observed from the results that, trees generated for the same set of species using different proteins may differ in topology as the pattern of evolution of a protein may differ from that of other proteins. So in the next chapter a new method for combining information from multiple proteins is developed.

Chapter 8

Development of a consensus method for constructing phylogenetic tree using multiple protein sequences

This chapter describes the development of a consensus method for phylogenetic analysis using multiple proteins. A review of literature describing various approaches of phylogenetic analysis using multiple proteins is provided. A new method called Consensus Phylogeny using Principal Component Analysis (CPPCA) for combining information from multiple proteins, for the generation of a consensus phylogenetic tree is developed. Finally the CPPCA method is applied on the sample datasets for the generation of consensus phylogenetic trees.

8.1 Introduction

A species tree represents the evolutionary history of a group of species and a gene tree represents the evolutionary history of a gene within a group of species. Speciation results in the emergence of new species and the gene lineages results in new genes. Gene trees explain how a gene evolves through various molecular events. Different genes do not necessarily have identical evolutionary histories and hence they may have difference in the tree topology and this difference is called gene tree incongruence. The rate of changes occurring in genes varies widely from one another. Some genes remain almost constant among the species under consideration while some genes vary too much so that a proper alignment is difficult. The different regions within the genomes can evolve differently (Maddison, 1997; Mallo and Posada, 2016; Posada, 2016), due to various biological phenomena like horizontal gene transfer, hybrid speciation and Multi-Species Coalescent (MSC) model. An internal node in a species tree represents a speciation event. Although species trees are not exactly same as the gene trees or protein trees, as the mutations and speciation events do not occur strictly at the same time, the gene trees are generally an accurate representation of species trees. But it is better to reconstruct gene trees from several genomic loci and combine these collections of gene trees using consensus methods to obtain a species tree. Phylogenetic tree based on multiple genes or proteins are consistent compared to the ones inferred from a single gene or protein as different rate of changes of sequences of different genes or proteins can lead to inconsistent topology of relationship.

This chapter describes a new approach to generate a phylogenetic tree by combining phylogeny information of multiple proteins. The method

uses an alignment free method for calculating pair wise genetic distances between protein sequences and generates a distance matrix for each protein. The individual distance matrices corresponding to different proteins are then combined using Principal Component Analysis (PCA) to obtain a consensus distance matrix. PCA is a dimensionality reduction technique which allows to combine the various trends in different matrices to produce a matrix which represents the underlying common pattern of distances in matrices. The phylogenetic tree representing the evolutionary relationship is then constructed using UPGMA method.

8.2 Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique which transforms high dimensional data to a lower dimensional subspace. PCA is an orthogonal transformation that transforms a data set with correlated variables into a smaller uncorrelated set of variables with less redundancy while retaining most of the useful information. PCA is equivalent to the rotation of original data space to a new coordinate space with new axes representing the principal components. The first Principal Component (PC) represents the direction of highest variance; the second PC represents the direction that maximises the remaining variance in the orthogonal subspace to the first component. This can be extended up to the adequate number of PC's that is required to represent the system in an optimal way. The amount of dimensionality reduction depends on this number of PC's selected and is very subjective.

The PCA can be considered as a linear transformation, P that transforms X into Y. Here, X represents the original data set and Y represents the new data set which is the projection of X on PC's. Both X

and Y are m x n matrices where m is the number of variables and n is the number of observations or samples. Now the task reduces to the estimation of the transformation function P. In order to quantify the redundancy in the variable, a covariance function C_x is defined as,

$$C_X = \frac{1}{n-1} XX^T \tag{8.1}$$

where C_x is an m x m matrix, with diagonal elements representing the variance of each variable and the off diagonal elements represents the covariance between the variables. The optimized data set without redundancy must have a covariance matrix with all off diagonal elements as zero. Hence the transformation matrix P should be selected such that C_Y is diagonalized. One method is to calculate the eigen vectors of C_x and forming a transformation matrix P with the eigen vectors as its rows. The eigen vectors also represent the PC's of the original data set X where, the significance of PC's is given by the corresponding eigen values. The first PC is the eigen vector with the largest eigen value and so on.

The principal components can also be obtained by using Singular Value Decomposition (SVD). The SVD of an n x m matrix X is represented by,

$$X^T \stackrel{\text{def}}{=} USV^T \tag{8.2}$$

Here, the columns of V are equivalent to the eigen vectors obtained from the covariance matrix, C_x, and are the PC's of X. The number of PC's extracted is same as the number of variables in original data. To achieve dimensionality reduction the first few meaningful components are retained. Eigen values provide a quantitative measure of the meaningfulness of components.

8.3 Review of phylogenetic tree construction methods using multiple genes.

The rapid growth in the availability of whole genome sequences of diverse organisms has made it possible to infer the species tree from multiple genes. Species tree generation method generates a single representative phylogenetic tree from phylogeny information of a collection of phylogenetic trees. The methods for combining multiple trees can be classified into concatenation or total evidence methods and summary or consensus methods. The concatenation or total evidence approach employs analysis of concatenated data sets. The consensus or summary methods aim to identify an estimate of the phylogeny by creating a consensus of the separate tree estimates.

In concatenated methods, multiple gene sequences are concatenated to create a combined sequence by joining together all the sequences of different genes from a species into one large single sequence. For each species such concatenated sequences are created. These concatenated sequences are then used to build a tree to reflect the evolutionary relationship of species. Salichos and Rokas (2013) extended the concatenation approach by selecting the most informative genes using a statistical method called bootstrap analysis. Sequences of most informative genes that carry the greatest amount of information with respect to evolutionary relationship are then concatenated to create one single sequence. Concatenated sequences are obtained for all the species and used for the construction of tree. Concatenation methods give accurate results when the individual gene trees have identical topologies. But when the

topology of the gene trees varies significantly, the approach is bound to fail as the approach assumes same evolutionary history for all the genes.

Phylogenetic analyses to construct species tree by summarizing a set of gene trees, estimated from individual gene sequence are called summary or consensus methods. This approach identifies the common substructures in the gene trees and constructs a single phylogenetic tree. Adams (1972) proposed the first consensus method. Consensus methods mainly belong to three types (Bryant, 2003), based on splits and clusters, cluster intersection methods and methods based on recoding. Strict consensus tree method (McMorris et al., 1983) and Majority rule tree (Barthelemy and McMorris, 1986) method are examples of methods based on splits and clusters. Adams tree and S-consensus trees (Stinebrickner, 1984) are based on cluster intersection methods. Consensus tree methods based on recoding converts the input trees into another form of data, like sequences and distances. The data in the new form are subsequently re-analysed using phylogenetic tree construction methods.

Distance based consensus methods estimates a species tree from the dissimilarity matrices of multiple gene trees. Distance matrix containing all the pair wise sequence dissimilarity is constructed for each gene. Then the individual distance matrices are combined to create a consensus matrix. The consensus matrix is then used to build the species tree. The average consensus tree of Lapointe and Cucumel (1997) and the Buneman tree (Buneman, 1971) are distance based methods. These methods try to compute a consensus matrix as the average of individual matrices. Liu and Yu (2011) have shown that the use of a distance matrix of average internode distances obtained from the collection of gene trees for constructing species tree. It is also proved that the method is statistically consistent one to

estimate the consensus species tree. Ane et al. (2007) employed a novel 2-stage Bayesian Markov chain Monte Carlo (MCMC) approach. In the first stage, a calculation of the posterior distribution of trees from single gene analyses is done. This posterior distribution along with a prior distribution on gene tree concordance is used in the second stage where a MCMC to estimate the joint probability distribution of the gene-to-tree map is employed. Ewing et al. (2008) describes a rooted triple approach to find the correct species tree. The method first generates the individual gene trees. Then extracts the rooted triple taxa trees from each gene tree and selects the most frequently occurring one as the species triplet tree. All the rooted species triples are then combined to produce a species tree for all taxa.

8.4 Consensus Phylogeny using Principal Component Analysis [CPPCA]

A new method named Consensus Phylogeny using Principal Component Analysis (CPPCA) to generate a species tree, by combining the phylogeny information of multiple proteins using Principal Component Analysis (PCA) is developed. In the first step the evolutionary relationship using each protein sequence is obtained using the SPPSD method described in chapter 7. In the next step this information from all the proteins is combined using PCA to obtain a consensus phylogeny. A schematic representation of the approach is shown in Fig. 8.1. The power spectral density based method for calculating pair wise genetic distances between protein sequences, as described in chapter 7, is used to create distance matrix for each protein in the first step. The individual distance matrices are then combined using PCA to obtain a consensus distance matrix. The PCA is a dimensionality reduction technique which combines the various trends

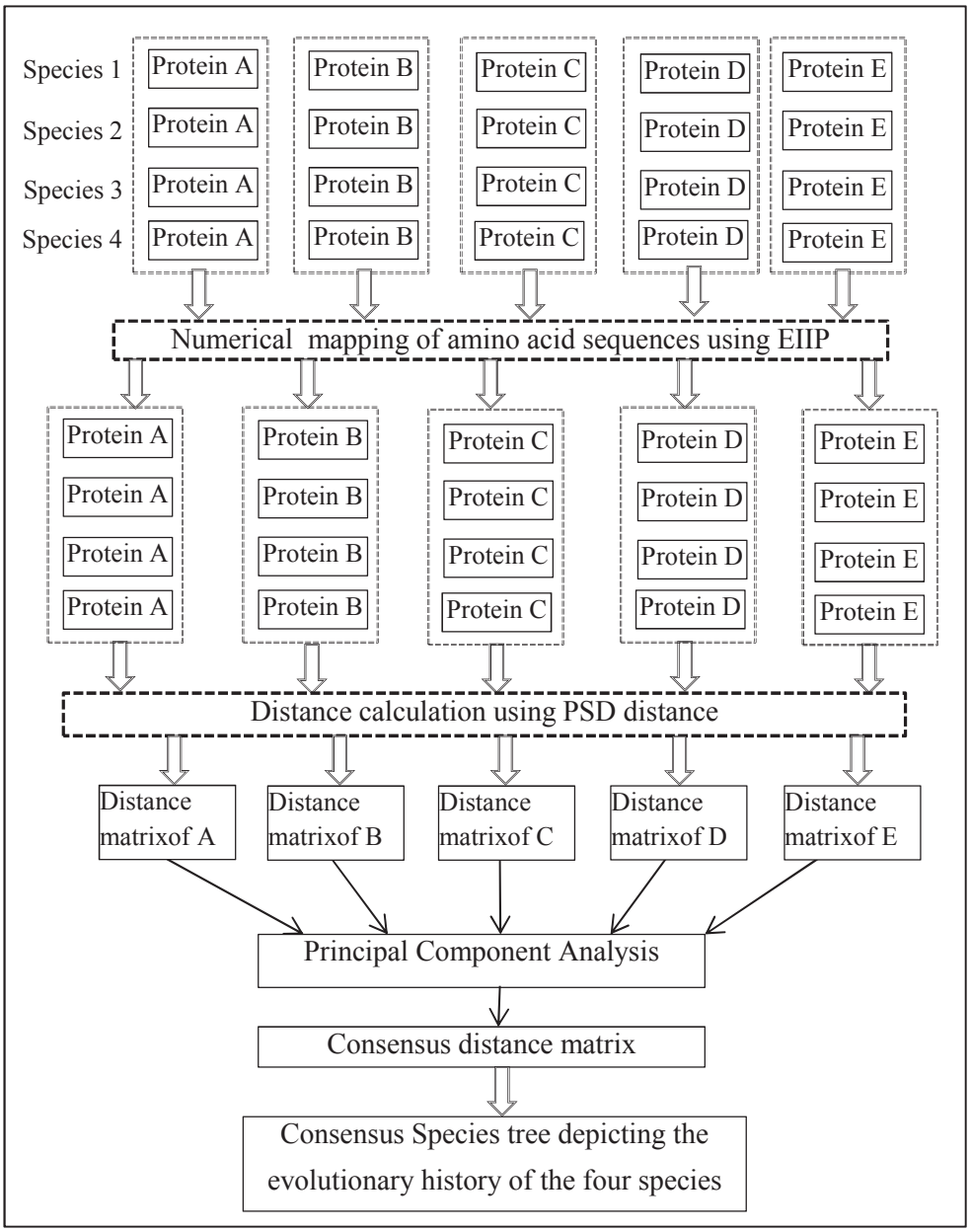


Figure 8.1 Schematic representation of the CPPCA method

in different matrices to produce a matrix which represents the underlying common pattern of distances in matrices. The consensus matrix thus

obtained is then used for the construction of phylogenetic tree representing the evolutionary relationship, by UPGMA method.

8.4.1 Creating a consensus distance matrix from multiple proteins using PCA

Using the SPPSD method separate distance matrices (D_i) are created corresponding to each protein. The distance matrices (D_i) that contains all pair wise distance between the different species can be denoted by,

$$D_i = \begin{bmatrix} D_{11} & \dots & D_{1k} \\ : & : & : \\ D_{k1} & \dots & D_{kk} \end{bmatrix} \qquad \text{i =1,2,3…N.} \qquad (8.3)$$

Here D_i is the distance matrix for the i^{th} protein, N is the number of proteins taken for analysis, k is the number of species under consideration, D_{xy} is the distance between species x and y.

The two dimensional distance matrix obtained is converted into a one dimensional distance vector (d_i) of length M as given below. (An example of this conversion illustrated in Fig 8.3 in section 8.4.2)

$$d_i = [d_{i1}, d_{i2}, \dots, d_{iM}] \qquad \text{i =1,2,3…N.} \qquad (8.4)$$

where, $M = k(k-1)/2$

Thus, N such one dimensional vectors are obtained, where N is the number of proteins selected for analysis. Then a joint data matrix (X) of size N x M is formed with each distance vector of size M occupying the rows of the new matrix as shown below.

$$X = \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1M} \\ d_{21} & d_{22} & \dots & d_{2M} \\ : & : & : & : \\ d_{N1} & d_{N2} & \dots & d_{NM} \end{bmatrix} \qquad (8.5)$$

A principal component analysis of the data set X is performed using singular value decomposition to obtain the eigen vector of X. The eigen vector with the highest eigen value is taken as the first principal component, PC_1. Then a projection of the data X onto PC_1 gives the distance vector Y_1 which can be taken as the consensus vector of all the proteins selected for the population under consideration.

$$Y_{N\times M} = P_{N\times N} X_{N\times M} \tag{8.6}$$

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1M} \\ y_{21} & y_{22} & \cdots & y_{2M} \\ \vdots & \vdots & \vdots & \vdots \\ y_{N1} & y_{N2} & \cdots & y_{NM} \end{bmatrix} = \begin{bmatrix} PC_1 \\ PC_2 \\ \vdots \\ PC_N \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1M} \\ x_{21} & x_{22} & \cdots & x_{2M} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NM} \end{bmatrix}$$

Y_1 is a one dimensional distance vector of the form, d_i, of length M which is converted back to distance matrix in the form D_i. Finally the phylogenetic tree based on N different proteins of the k populations is constructed using UPGMA method.

8.4.2 Protein sequence database and preparation of sample set

All the protein sequence data used for the study is obtained from the Entrez search and retrieval system of the National Center for Biotechnology Information (NCBI, 2017). The amino acid sequences of proteins are obtained in FASTA format.

Sample datasets

To study the performance of the PCA method for species tree construction, four different datasets are used. Each dataset is characterized by a set of proteins and a set of species. For each species, sequences corresponding to all proteins in the set are obtained. Thus, the protein

sample set includes sequences of a set of proteins obtained from different species.

Dataset 1

The first set of data for analysis includes amino acid sequences of 4 proteins obtained from 7 species. The proteins selected for this set includes, alpha hemoglobin (HBA), beta hemoglobin (HBB), cytochrome-b (Cyt-b) and myoglobin. All proteins are obtained from 7 species belonging to the class- mammalia and two different order- primate and rodentia. The scientific name and common name of the species selected are given below.

A. *Pan troglodytes* (Chimpanzee)
B. *Gorilla gorilla* (Gorilla)
C. *Homo sapiens* (Human)
D. *Callithrix jacchus* (Common Marmoset)
E. *Mus musculus* (Mouse)
F. *Rattus norvegicus* (Rat)
G. *Mus pahari* (Shrew Mouse)

Dataset 2

The second dataset includes amino acid sequences of 5 proteins obtained from 9 species. The proteins selected are: testin, myoglobin, lysozyme, caveolin-1 and cytochrome b. All the proteins are obtained from 9 species belonging to the class- mammalia and order- primate. The scientific name and common name of the species are given below.

A. *Papio anubis* (Baboon)
B. *Pan troglodytes* (Chimpanzee)
C. *Hylobates agilis* (Gibbon)
D. *Gorilla gorilla* (Gorilla)
E. *Homo sapiens* (Human)

F. *Callithrix jacchus* (Marmoset)
G. *Aotus trivirgatus* (Night monkey)
H. *Macaca mulatta* (Rhesus monkey)
I. *Saimiri sciureus* (Squirrel monkey)

Dataset 3

The third dataset includes amino acid sequences of 10 proteins obtained from 14 species. The proteins selected are: caveolin-1, caveolin-2, caveolin-3, cytochrome-b, flotillin, lysozyme, myoglobin, prolactin, somatotropin and testin. All proteins are obtained from 14 species belonging to the class- mammalia and five different orders- primate, rodentia, lagomorpha, carnivora and artiodactyla. The scientific name and common name of the species are given below.

A. *Papio anubis* (Baboon)
B. *Bos taurus* (Cattle)
C. *Felis catus* (Cat)
D. *Pan troglodytes* (Chimpanzee)
E. *Canis lupus familiaris* (Dog)
F. *Gorilla gorilla* (Gorilla)
G. *Homo sapiens* (Human)
H. *Callithrix jacchus* (Marmoset)
I. *Mus musculus* (Mouse)
J. *Oryctolagus cuniculus* (Rabbit)
K. *Rattus norvegicus* (Rat)
L. *Macaca mulatta* (Rhesus monkey)
M. *Ovis aries* (Sheep)
N. *Saimiri sciureus* (Squirrel monkey)

Dataset 4

The fourth dataset includes amino acid sequences of 3 proteins from a family of membrane protein called caveolin. They are: caveolin-1,

caveolin-2 and caveolin-3. They are obtained from 21 species belonging to three classes- mammalia, aves, actinopterygii and belongs to 10 different orders: primate, rodentia, carnivora, artiodactyla, lagomorpha, galliformes, passeriformes, cypriniformes, tetraodontiformes and perissodactyla. The scientific name and common name of the species are given below.

A. *Papio anubis* (Olive baboon)
B. *Bos taurus* (Bovine)
C. *Felis catus* (Cat)
D. *Gallus gallus* (Chicken)
E. *Pan troglodytes* (Chimpanzee)
F. *Danio rerio* (Zebra fish)
G. *Ailuropoda melanoleuca* (Giant panda)
H. *Equus caballus* (Horse)
I. *Homo sapiens* (Human)
J. *Callithrix jacchus* (Marmoset)
K. *Mus musculus* (Mouse)
L. *Aotus trivirgatus* (Night monkey)
M. *Takifugu rubripes* (Puffer fish)
N. *Coturnix japonica* (Japanese Quail)
O. *Oryctolagus cuniculus* (Rabbit)
P. *Rattus norvegicus* (Rat)
Q. *Macaca mulatta* (Rhesus monkey)
R. *Ceratotherium simum simum* (Rhinoceres)
S. *Ovis aries* (Sheep)
T. *Zonotrichia albicollis* (Sparrow)
U. *Leptonychotes weddellii* (Seal)

8.4.3 Implementation of the CPPCA method to infer consensus species tree

To evaluate the CPPCA method, four datasets are prepared. The first dataset includes 4 proteins obtained from 7 species forming a collection of

28 sequences. To start with, HBA sequence from all the 7 species is obtained and all pairwise distances between the sequences are calculated using the PSD method described in previous chapter. The distance matrix thus formed, which is a protein dissimilarity matrix constructed for the protein HBA, is given in Table 8.1.

Table 8.1 Distance matrix of 7 species corresponding to HBA

	A	B	C	D	E	F	G
A	0.000	0.549	0.000	0.566	0.646	0.864	0.694
B	0.549	0.000	0.549	0.206	0.798	0.934	0.880
C	0.000	0.549	0.000	0.566	0.646	0.864	0.694
D	0.566	0.206	0.566	0.000	0.829	0.960	0.923
E	0.646	0.798	0.646	0.829	0.000	0.929	0.321
F	0.864	0.934	0.864	0.960	0.929	0.000	1.000
G	0.694	0.880	0.694	0.923	0.321	1.000	0.000

The same distance matrix in Table 8.1 can also be represented as a triangular matrix as shown in Fig 8.2 [a-b].

A	0.000	0.549	0.000	0.566	0.646	0.864	0.694
B		0.000	0.549	0.206	0.798	0.934	0.880
C			0.000	0.566	0.646	0.864	0.694
D				0.000	0.829	0.960	0.923
E					0.000	0.929	0.321
F						0.000	1.000
G							0.000

A	0.000						
B	0.549	0.000					
C	0.000	0.549	0.000				
D	0.566	0.206	0.566	0.000			
E	0.646	0.798	0.646	0.829	0.000		
F	0.864	0.934	0.864	0.960	0.929	0.000	
G	0.694	0.880	0.694	0.923	0.321	1.000	0.000

Figure 8.2 Distance matrix of HBA represented as (a) Upper triangular matrix (b) Lower triangular matrix

The triangular matrix is then converted to a 1-D array by appending the rows one after the other to create a 1-D distance array of the protein HBA as shown in Fig. 8.3.

0.549	0.000	0.566	0.646	0.864	0.694	0.549	0.206	0.798	0.934	0.880	0.566	0.646	0.864	0.694
0.829	0.960	0.923	0.929	0.321	1.000									

Figure 8.3 Distance matrix of HBA as a 1-D vector.

Similarly, three distance matrices corresponding to the remaining three proteins are also generated and are converted to 1-D distance vector. Then a combined matrix is created with these 4 vectors as the 4 rows. This matrix forms the data on which a PCA is performed using SVD, to obtain the eigen vectors. The first PC, which is the eigen vector with the highest eigen value is obtained. And a projection of the combined data matrix in the direction of this first PC is performed and is taken as the consensus distance vector. The resulting consensus distance vector in the form of 1-D distance vector is then transformed back to the 2-D distance matrix form and is used to construct a consensus species tree as shown in Fig 8.4a.

Using similar steps, consensus tree is constructed for each of the remaining three datasets. The result of the phylogenetic classification obtained by the CPPCA method is compared with taxonomic classification for validation. Taxonomic classification for each of the dataset is obtained separately using the NCBI Taxonomy Common Tree tool. The tool creates a taxonomy tree which do not takes, the sequences selected, into consideration and thus represents an unbiased classification. The same datasets are used to generate species tree using concatenation method and distance matrix averaging method for comparison. In the concatenation method, all the selected protein sequences of a species are concatenated to create a single long sequence. Thus after concatenation there will be one sequence for each species, which is used to construct the phylogenetic tree. The distance matrix averaging method calculates individual distance

matrices for each protein and then an average distance matrix is created from the individual matrices, where distance between a pair of sequence is the mean of the distance between the corresponding species for all the proteins. In the last method, a consensus tree using extended majority rule is constructed. Individual trees for each protein are generated using CLUSTALW and a consensus tree is created using Majority Rule extended (MRe) with the help of PHYLIP program (Felsenstein, 2005). The CLUSTALW is a sequence alignment based method for phylogenetic analysis. The extended majority rule is the default option of the PHYLIP consensus tree method, where any set of species that appears in more than 50% of the trees is included in the consensus tree. Then the other sets of species in order of the frequency with which they appear is considered, until the tree is fully resolved. The method generates an unrooted tree. All the other methods described above generate rooted trees. So for comparison, outgroup rooting option is used by introducing an outgroup species to root the tree. Thus the consensus tree generated using CPPCA method is compared with consensus trees created using concatenation method, averaging method, PHYLIP - Majority Rule extended method and also with the taxonomic classification obtained using NCBI Taxonomy Common Tree tool. A detailed discussion of the results obtained and comparative performance of the CPPCA method is provided in the next section.

8.5 Results & Discussion

8.5.1 Dataset 1

The individual trees constructed for the 4 proteins in the first dataset are shown in Fig 8.4 [a-c]. It can be observed that the topology of the tree obtained using cytochrome-b and myoglobin are the same. The topology of

the trees constructed using hemoglobin-α & β differ from each other and from that of the other two proteins. It is worth noting that, even for a dataset that includes only 7 species, there is incongruence in the branching pattern and hence in the evolutionary pathway of individual proteins. The consensus tree obtained using the CPPCA method for the 4 proteins is shown in Fig 8.5a. It can be observed that, the smaller substructures or sub trees in the consensus tree are the most common pattern present in the individual trees. For example, the sub tree represented by (A,C) is present in all the 4 trees and the pattern (((A,C),B),D) and ((E,G),F) is present 3 of the 4 trees.

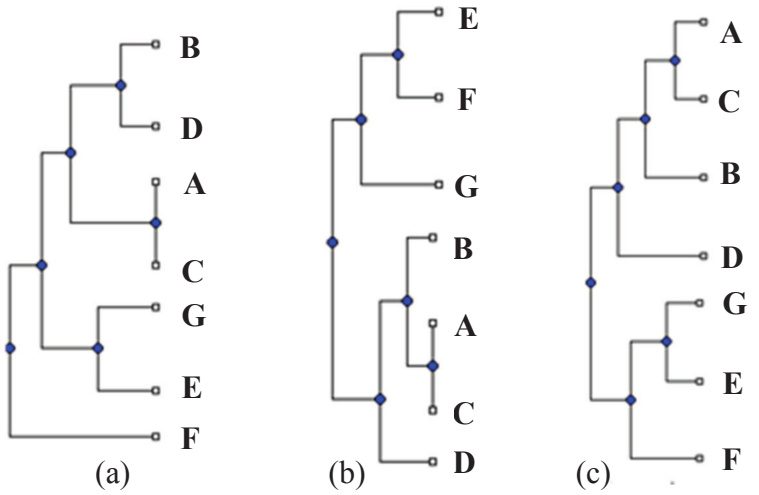


Figure 8.4 Phylogenetic tree generated using (a) hemoglobin-α. (b) hemoglobin-β. (c) cytochrome-b and myoglobin.

Fig 8.5b shows the taxonomic classification of the 7 species. From the figure, it can be seen that the 3 species, *Pan troglodytes, Gorilla* and *Homo sapiens* are shown to have a common ancestor node. It is due to the fact that all the 3 belongs to a sub family homininae represented by the ancestral internal node. From the figures, 8.5a and 8.5b it can be noted that the topology of the consensus tree generated by CPPCA method is the same

as that of the taxonomic classification of the species. The phylogenetic tree constructed using the three other methods, concatenation method, averaging method and PHYLIP - MRe also have the same topology as shown in Fig 8.5a.

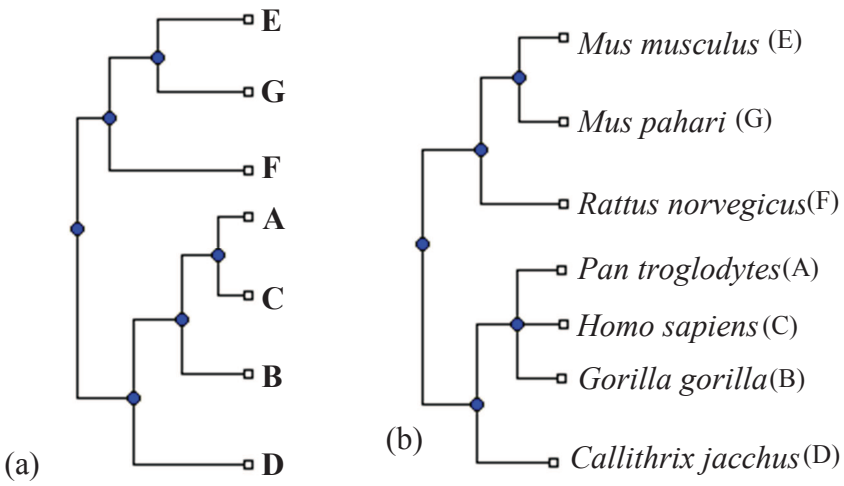


Figure 8.5 (a) Consensus phylogenetic tree created using the four proteins (Dataset1) by CPPCA, averaging, concatenation and PHYLIP – MRe methods (b) Taxonomic classification of the species using NCBI Taxonomy Common Tree.

8.5.2 Dataset 2

The second dataset involves 5 proteins from 9 primate species. The organisms selected in the set are closely related ones and will allow to verify the ability of the CPPCA method to resolve the relation between organisms when there is only small difference in the distance values between the different pairs of sequences. The Fig 8.6a shows the consensus tree constructed by the CPPCA method. The species F,G,I belong to 'new world monkeys', I,H belong to 'old world monkeys' and B,E,D,C belong to apes. This is a classification within primates based on several characteristics like size of organisms, capabilities etc. The tree constructed using concatenation

method and using PHYLIP - MRe is also having same structure as shown in Fig 8.6a. It can be noted that the CPPCA tree clearly categorizes each of the three groups as separate sub trees. The tree also hypothesizes a common ancestor for apes and old world monkeys which is supported by the taxonomic tree in Fig 8.6b and also by the result of concatenation tree and MRe tree. It can be observed that the substructure involving species F,G,I has a different branching pattern when compared to the taxonomic classification. But the pattern obtained in CPPCA is same as that of concatenated tree and MRe tree. It can also be noted that the tree generated by averaging method (Fig 8.6c) is having a totally different topology and is not considered for discussion as the result seems inaccurate.

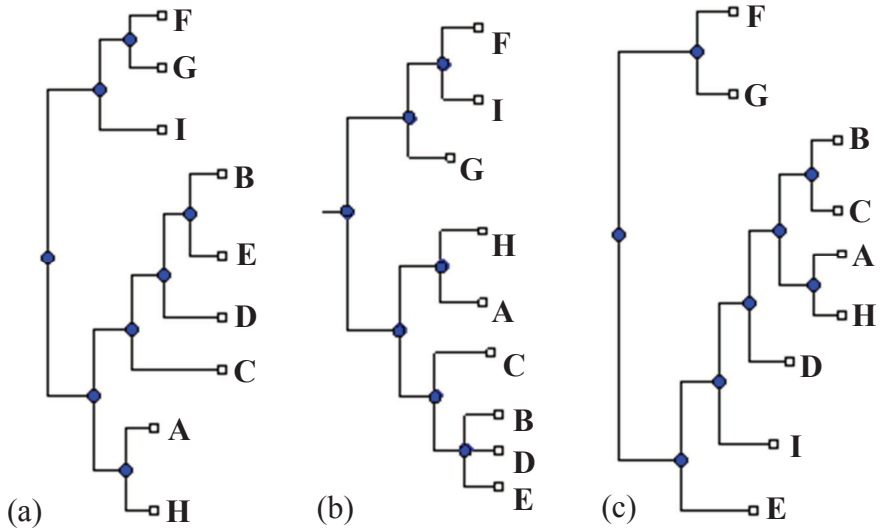


Figure 8.6 Classification of species in dataset 2 using (a) CPPCA, concatenation and PHYLIP – MRe methods. (b)taxonomy. (c) Averaging method

8.5.3 Dataset 3

The third dataset includes 10 proteins from 14 species forming a sample set of 140 sequences which is diverse in terms of proteins and species involved. The result of CPPCA obtained for the dataset is shown in Fig 8.7a. The taxonomic classification is shown in Fig 8.7b and MRe tree in Fig 8.7c. Fig 8.8a and 8.8b shows the tree obtained using concatenation method and averaging method respectively. There are species from five different orders in the set with {N,H,L,A,D,G,F} belonging to primate, {K,I} belonging to rodentia, {J} belonging to lagomorpha, {C,E} belonging to carnivora and {B,M} belonging to artiodactyla. All the five orders are clearly demarcated in the CPPCA tree as separate sub trees. The sub tree ((((G,D),F),(A,L)),(H,N)) and (I,K) are maintained in all the 4 consensus tree. The sub trees ((C,E),(B,M)) is supported by 3 out of the 4 trees.

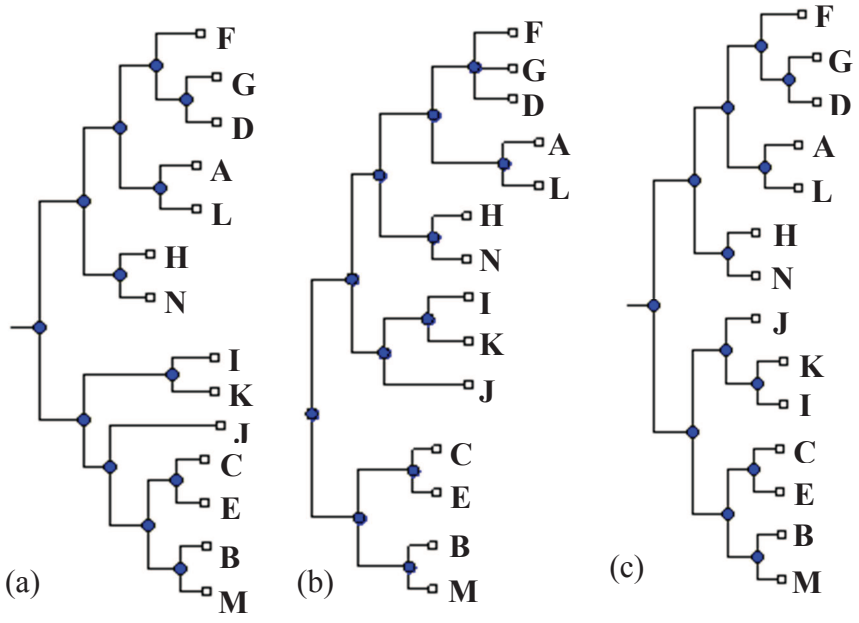


Figure 8.7 Classification of species in Dataset 3 based on (a) CPPCA method (b)Taxonomy (C) PHYLIP - MRe method

A comparison of the 4 consensus trees shows that the tree topology of the CPPCA and MRe tree are similar and is closer to the taxonomic classification. One interesting observation is that, all the 5 trees differ from each other. This highlights the evolutionary diversity and challenges in phylogenetic analysis. The group of species {C,E,B,M} and {J,K,I,N,H,L,A,D,G,F} belongs to two super orders, laurasiatheria and euarchontoglires respectively. In the taxonomic classification it is shown as separate branches. Such a grouping is not observed in any of the consensus tree which may be due to a difference in the evolutionary relationship from taxonomic relationship.

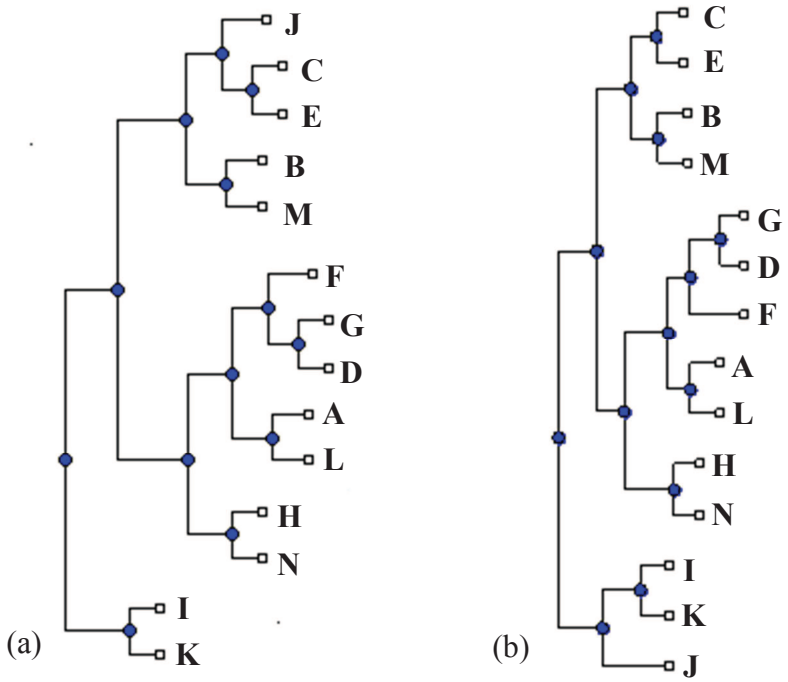


Figure 8.8 Classification of species in Dataset 3 based on (a) Concatenation method. (b) Averaging method

8.5.4 Dataset 4

The fourth dataset involves 3 proteins from a protein family called caveolin, obtained from 21 species. The caveolin family is selected for this analysis, as all the three members of the protein family are present in most chordates and as a result the complete sequences of these proteins from many species are easily available for analysis.

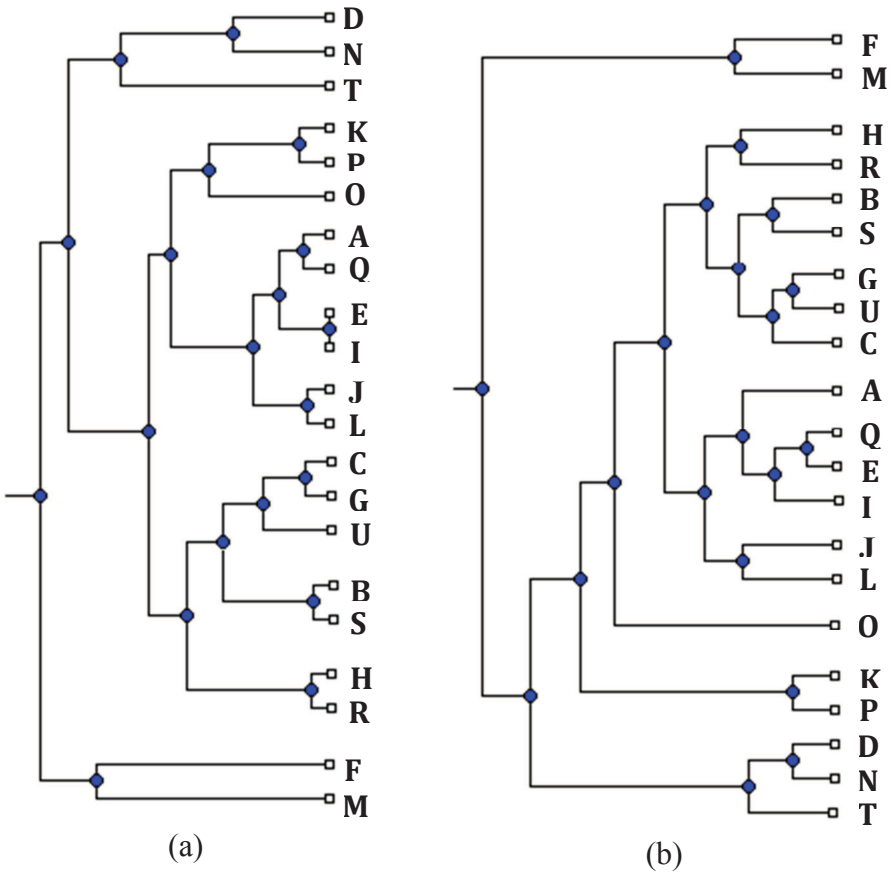


Figure 8.9 Classification of species in the dataset 4 based on (a) CPPCA method & averaging method (b) PHYLIP - MRe method

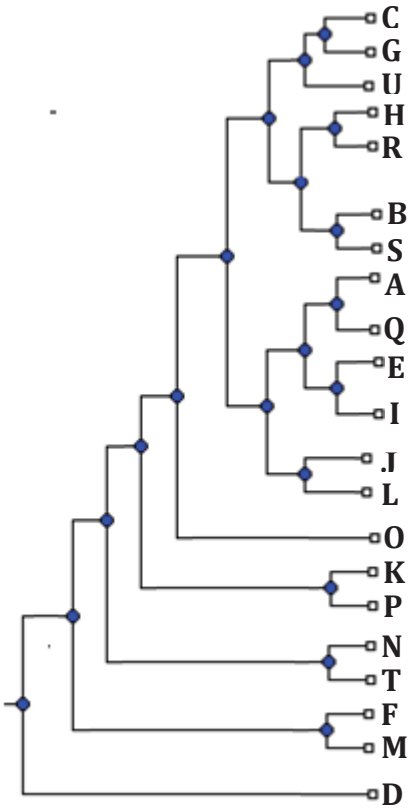


Figure 8.10 Consensus tree using concatenation method

The consensus phylogenetic tree constructed using CPPCA method for the sample set is shown in Fig 8.9a. The dataset contains organisms from 3 different classes, aves, actinopterygii and mammalia. The CPPCA method classified them into three groups, which is shown in the figure as 3 separate sub trees. Aves is represented by sub tree with terminal nodes D,N,T. Actinopterygii is shown as sub tree with terminal nodes F,M and mammalia by sub tree with terminal nodes K, P, O, A, Q, E, I, J, L, C, G, U, B, S, H, R.

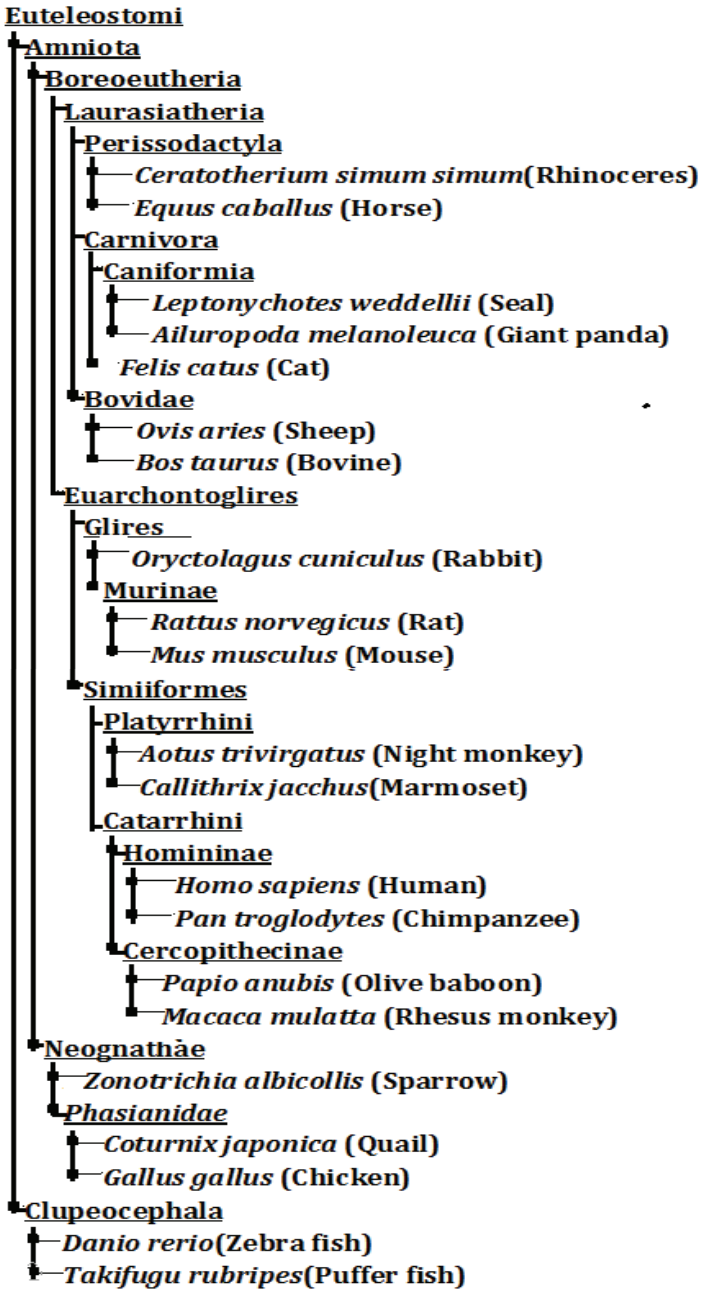


Figure 8.11 Taxonomic classification of the 21 species in dataset 4

The different orders in the set are also clearly distinguished as separate groups. ((L,J),((A,Q),(E,I))) represents the primate sub tree, (K,P) represents rodentia, ((C,G),U) represents carnivora, (B,S) represents artiodactyla, (R,H) represents perissodactyla and (N,D) represents galliformes. The remaining 4 orders have only one representative and hence not shown as a group. A comparison of the CPPCA with other consensus trees shows that the topology of the tree built using averaging method is same as that of CPPCA. The CPPCA shares some similarity with the MRe tree (Fig 8.9b) but differs for rodentia and primates. The MRe tree has a grouping (((Q,E),I),A) which is not supported by any of the trees and by the taxonomical classification (Fig 8.11). The topology of the concatenation tree shown in Fig 8.10 is much different compared to the other consensus trees and also not in conformance with the taxonomical classification. Meanwhile, the phylogenetic classification achieved using the CPPCA method is in accordance with the taxonomical classification except for the sub tree structure ((C,G),U) in the CPPCA, where the taxonomical classification supports a structure ((G,U),C). But it is notable that the sub tree structure ((C,G),U) is supported by 3 out of the 4 consensus trees.

The CPPCA method is applied to four datasets of varying composition for validation. The first dataset is a small sample set with less diversity. Even within such a small dataset, the individual proteins showed difference in the pattern of evolution. The second dataset included only species belonging to a single order and are much closely related with very small variation among the sequences. This was aimed at checking the ability of the method to resolve finer differences between sequences and deduce the correct relationship between them. The third dataset was a much diverse sample set and the diversity is illustrated by the consensus trees generated,

where all the four methods generated trees that differed from one another. The fourth dataset included sequences from more number of species but all the sequences belongs to a single family of membrane protein. The analysis clearly showed that the CPPCA method is effective in capturing the underlying pattern in multiple trees. The phylogenetic trees or the phylogenetic classification can be described as a hypothesis of the evolutionary past as there is still no established way to confirm the proposed relationships. The phylogenetic tree attempts to illustrate the details like, when different organisms evolved and the degree of relationships among them. Due to the absence of a perfectly known ground truth the comparison of the results is a tough task. The taxonomic classification categorizes the organisms based on similarities and dissimilarities of various characteristics. These similarities and dissimilarities are not random incidents but are results of inheritance and divergence from an ancestor species. Hence, taxonomic classification can be considered as a reflection of the evolutionary history, though it is not same as phylogenetic classification. In this study the consensus trees are compared with taxonomic classification for validation. Among the compared methods, the consensus tree generated using the CPPCA method clearly illustrated better conformance to the taxonomical classification for all the four datasets.

8.6 Summary

Phylogenetic tree represents the evolutionary pathway of organisms and the evolutionary pathway of species may be different from that of individual genes. So the construction of trees using sequences from multiple genomic loci using consensus methods improves the accuracy and reliability of the tress. A new signal processing based distance method for generating a

consensus tree from multiple protein sequences using principal component analysis is developed. The method named 'CPPCA' uses the PSD method described in the previous chapter for the generation of individual distance matrices for each protein which is then combined with PCA to generate a consensus distance matrix. The consensus distance matrix is used to construct the phylogenetic tree which is a summary of the evolutionary pathways of multiple proteins. The method is applied to four datasets of varying composition for validation. The consensus tree generated using the CPPCA method for the sample datasets clearly illustrated the ability to capture the most common underlying branching patterns in the individual trees.

Chapter 9

Conclusion and future scope

In this concluding chapter, a summary of the research study with important observations and results is highlighted. The directions for future study are also discussed.

9.1 Conclusion

The aim of this work was to develop signal processing algorithms to study genomic sequence variations. The research work is divided into two parts. The first part deals with the development of a novel method for the detection and localization of copy number variations in the genome, using Array CGH data. The second part involves the development of a new alignment free method for the phylogenetic classification of organisms using protein sequences.

A genome is the complete set of genetic information in an organism and the genome sequence of a species is a representative sequence based on multiple individuals of that species. Genomic variation refers to genetic differences among individuals of the same species. Some of these variations are beneficial to an organism, where it helps in adapting to the changes in environmental circumstances. These variations, thus helps in the survival of a population and plays a significant role in the evolutionary mechanism. There is another category of variations that are harmful to an organism, where it may cause disease, disease susceptibility or even affect an individual's capacity to survive. The huge amount of genomic data generated using latest sequencing techniques has necessitated the need of newer computational techniques. This research study aims to use digital signal processing methods to study the genomic sequence variation, to meet the research objectives.

The primary aim of copy number variation detection methods is to track the variations in the log fluorescence intensity data obtained from an Array CGH experiment, to identify significant deviations and to locate the critical regions corresponding to copy number variations. This involves

different stages like denoising of raw data, segmentation of data into discrete copy number level and assignment of copy number status to these regions. A wide variety of computational methods have been developed for the detection of copy number variation from log ratio data. Most of these methods give emphasis to either the denoising part or the segmentation part. As a result, the methods based on segmentation approaches fails to maintain the performance in the case of highly noisy log ratio data. Meanwhile, the smoothing based methods fail to provide accurate classification of regions into normal and aberrant regions. Most of the smoothing algorithms also cause smoothening out of smaller aberration and thus fail to detect CNVs of small widths. To overcome these disadvantages, a novel EES (Edge Enhancement and Segmentation) method is developed.

The EES method involves a step by step approach. It includes, a local edge enhancement filtering step, K-means clustering based segmentation step and a thresholding step. The EES method performs the edge enhancement filtering operation using a local filtering algorithm named, Minimum Variance Filtering (MVF). MVF is based on the search for local homogeneity of signal with the aim of preserving significant edges while performing the denoising operation. The denoising performance of the MVF method is compared with other smoothing based methods using RMSE analysis. MVF method showed an improvement of 5% to 15.75% in RMSE compared to Quantreg method, 54% to 76% over Wavelet method and 60% to 81% over Lowess method. The MVF method also proved the edge retention capability by detecting smaller aberrations of width as small as 2 data points. MVF method also includes an iterative filtering option with automatic stopping criterion for dealing with highly noisy data. In the second stage, EES method uses a K-means clustering approach to achieve

the segmentation of the log intensity ratio data into discrete, non-overlapping segments corresponding to different copy number levels. In the final step, the segmented levels are then categorized as normal or aberrant regions using a threshold based decision function. The default threshold value for classifying a region as aberrant is $|Log_2(R/G)| > |0.225|$. The ROC curves are plotted to evaluate the tradeoff between TPR and FPR of the EES method and compare it with other methods. The ROC curves and the area under the ROC curve metric clearly highlighted the superior performance of EES algorithm in classifying a probe into aberrant or normal region, at difference noise levels, while using simulated data. To study the performance of the new EES method on real data, the analysis was extended to real Array CGH data, and found to be the best compared to other five methods, to detect CNV. The ability of EES method in clinical application is illustrated by applying the method on three real Array CGH datasets for the analysis and detection of copy number variations.

The first data set selected for analysis is the coriell cell line BAC Array CGH data. The EES method successfully identified 21 out of the 22 aberrations without any false positive detection. The second dataset used is a breast cancer cell line database. The EES method detected all the 55 candidate therapeutic target genes which are amplified and overexpressed in at least one cell line. EES method could detect 53 of them as amplified in the same cell line in which an over expression was observed. The third data set is from glioblastoma multiforme database. The analysis of GBM data using the EES method detected many characteristic genomic alterations associated with glioblastoma occurrences like - losses on chromosome 10, 13 and 22; gains on whole chromosome 19, 20; loss of 9pter region. It also showed correlated alterations like - gains on chromosome 7 and losses on

chromosome 10; gains on chromosome 20 and losses on chromosome 10. The application of EES method on real Array CGH datasets clearly illustrated the effectiveness of the EES method in detecting copy number variations.

Understanding the evolutionary relationship between species is a major step in uncovering the pattern of evolution of organisms. The molecular phylogenetic analysis methods determine the evolutionary divergence between the various species from a measure of genomic sequence similarity. The most common approach for phylogenetic analysis involves the use of multiple sequence alignment techniques for obtaining a measure of similarity or dissimilarity between sequences. But, as the size and amount of sequences grow, the computational and time complexity of MSA techniques becomes unmanageable. In the second part of the thesis, an alignment free method to measure protein sequence similarity and to infer phylogenetic relationship between different species using a single protein is developed. Amino acid sequences of proteins are used in this study. The alphabetical sequence is converted to a numerical equivalent using EIIP method for applying signal processing algorithms. A frequency domain analysis of this numerical sequence is performed using DWT. The analysis of similarity between the sequences showed significant spectral similarity between closely related species and very poor spectral correlation for distantly related species. Based on the observation a new signal processing based approach for inferring evolutionary relationship using a single protein sequence obtained from a collection of organisms is developed. The method uses an alignment free approach to measure protein sequence similarity. The method named, SPPSD (Single Protein Power Spectral Density), uses the distance between Power Spectral Densities (PSD) of numerically

transformed protein sequences as a measure of genetic distance for the construction of phylogenetic tree. The SPPSD method is applied to different datasets and the results are compared with the results of other established multiple sequence alignment based methods such as, COBALT, CLUSTALW and MEGA. The tree generated using the SPPSD method for the sample datasets clearly illustrated its ability to capture the pattern of divergence of the individual proteins and in most cases, were in conformance with that of MEGA and CLUSTALW trees. Wherever it differed, the results of SPPSD method has showed better bootstrap confidence values and better matching with taxonomic classification.

The phylogenetic tree inferred for the same set of species using different proteins showed difference in topology. This difference has been attributed to the difference in the pattern of evolution of a protein from that of another. With the aim of improving the consistency of inferred phylogenetic relationship, a new method called, Consensus Phylogeny using Principal Component Analysis (CPPCA), is developed to generate a phylogenetic tree from multiple proteins. Amino acid sequences corresponding to different proteins are obtained from the organisms under study. Genetic distances are calculated using the SPPSD method for each of the protein. The CPPCA method then combines the genetic distances obtained for individual proteins using Principal Component Analysis, to create a consensus distance metric. Finally a phylogenetic classification of the organisms is achieved using the consensus distance. The CPPCA method is applied on different sample datasets for inferring phylogenetic relationship. The method was able to resolve finer differences between sequences belonging to closely related species with very small variation among them and successfully inferred relationship between them. The

method was also tested with dataset containing diverse species. For all the datasets, the CPPCA method clearly demonstrated its effectiveness in capturing the underlying pattern of relationship and showed better conformance to the taxonomical classification than other compared methods.

The research work successfully achieved the objectives of developing genomic signal processing based methods for the detection of copy number variation and phylogenetic classification of organisms. The task of copy number variation detection is solved using the EES method and the phylogenetic classification is achieved using SPPSD and CPPCA methods.

9.2 Scope for further study

With rapid developments in the field of array technology and next-generation sequencing technology, the future efforts need to focus on developing tools capable of analyzing data from both high density array platforms and next generation sequencing platforms. Array based methods offer a cost effective means for CNV analysis but lacks the capability to identify small CNVs. The next-generation sequencing based methods offer the ability to detect smaller CNVs but are expensive. So a computational framework incorporating analysis tools for both these platforms would provide for a better CNV detection capability.

Another area for further studies is the co-evolution of proteins or genes. Co-evolution is defined as interdependence of evolutionary pathway. The evolutionary pattern of many genes or proteins is dependent on others. At the molecular level, this can be due to specific co-adaptation between the

two co-evolving elements, where changes in one of them are compensated by changes in the other. The phylogenetic analysis methods that study the evolutionary pattern can be extended to focus on identifying co-evolving proteins. This will help to gain an insight into various physical interactions or functional relationships such as protein-protein interactions, protein functional sites etc.

Appendix

Appendix A1 illustrates estimation of the type of noise present in real Array CGH data. Appendix A2 describes an implementation of the EES method using Fuzzy C-Means clustering instead of K-Means clustering. A comparison of the performance of various CNV detection methods on real array CGH data is given in appendix A3 and protein sequence similarity analysis using discrete fourier transform is described in appendix A4.

A1. Estimation of noise type in Array CGH data

This analysis aims to study the type of noise present in Array CGH log ratio data. Histogram plot shows the underlying frequency distribution of a set of data. The noise type in log ratio data is studied by plotting the histogram of the noise and by observing the shape of the histogram. For gaussian noise, the histogram plot will have the shape of a gaussian probability distribution function. Here, an estimation of noise type is performed using noise signal extracted from a real array CGH database, Coriell cell line (Snijders et al., 2001) for which the actual copy number information is known. Let the noisy input log ratio data, Y be represented as,

$$Y=X+N \qquad \text{(A1.1)}$$

where X is the noise free actual data and N is the noise to be estimated.

For this database, X is known, so the noise N is obtained as,

$$N=Y-X \qquad \text{(A1.2)}$$

The noise component is extracted from the input data using the known copy number value information using the Eq. A1.2.

Eg: For GM03563 sample, chromosomes except 3 and 9 do not have copy number variations and can be considered to be having a constant zero level for the $\log_2$(R/G) ratio, with additive gaussian noise. The noise component is extracted and the histogram of the noise is plotted as shown in Fig A1.1a. The shape of the histogram obtained for GM03563 follows a gaussian probability distribution function. Similarly, noise histograms for other samples were also analyzed. Fig A1.1 b, c, d show the histogram plots for samples- GM00143, GM05296 and GM07408 respectively. The combined

histogram of noise extracted from all the samples in the database is shown in Fig.A1.2. The figures show a noise distribution that follows the shape of gaussian probability distribution function which supports the general assumption made in most literatures, where noise is modelled as gaussian. Hence the gaussian noise model is selected for denoising the Array CGH log ratio data for copy number variation analysis.

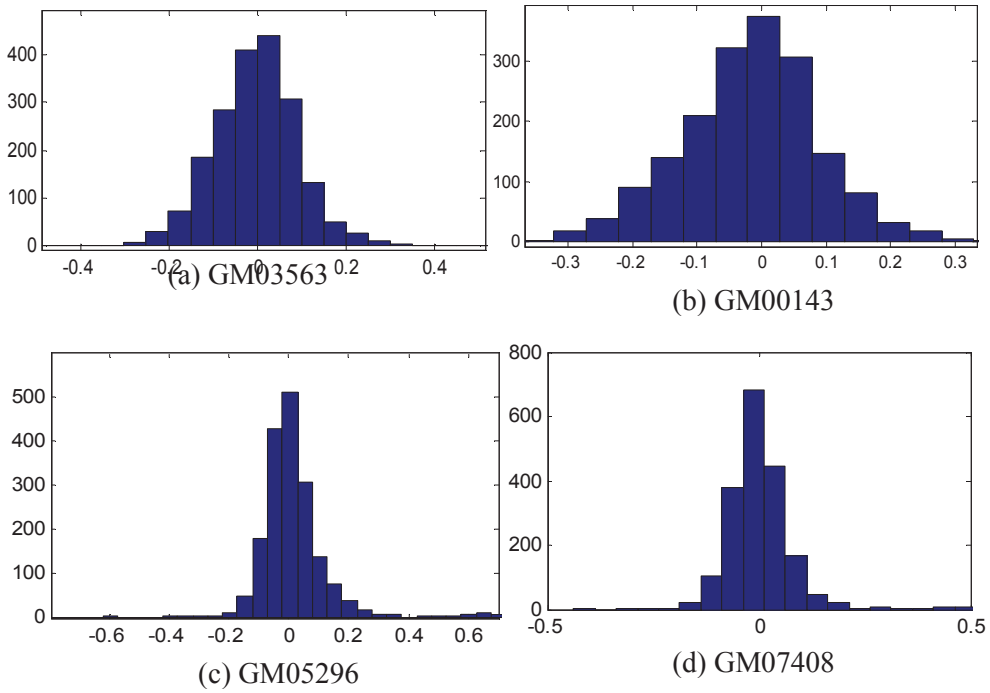


Figure A1.1 Histogram of the noise extracted from different Coriell cell line samples

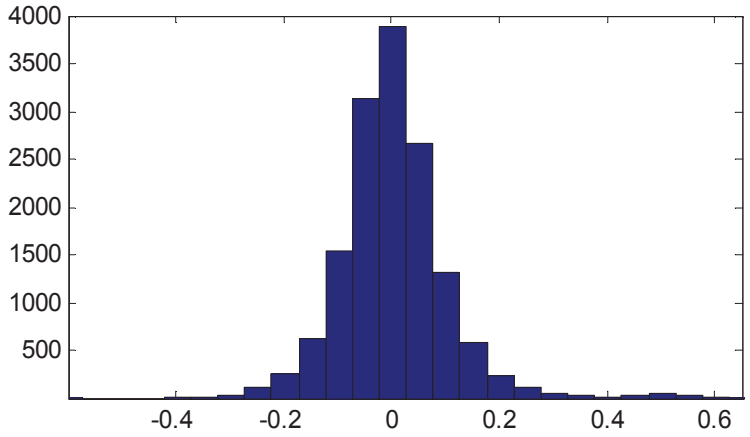


Figure A1.2 Histogram of the noise extracted from entire Coriell cell line samples

A2. Comparison of Fuzzy C-Means based implementation of EES method with K-Means based EES method

A2.1 Fuzzy C-Means Clustering

Fuzzy C-Means Clustering (FCM) is a soft clustering approach which allows a sample to belong to multiple clusters, with an associated membership value. This algorithm calculates the distance of each data point from every cluster centers. And based on this distance, the data point is assigned with a membership value for each cluster. This value will be higher if the data point is closer to a cluster center and vice versa. FCM does not assign exclusive membership of a data point to a given cluster; instead, it calculates the probability that a data point will belong to that cluster. The summation of membership value of each data point to all clusters should be equal to one. The FCM algorithm, aims to minimize the objective function, J given below,

$$J = \sum_{i=1}^{N} \sum_{j=1}^{C} \partial_{ij} \left\| x_i - c_j \right\|^2$$

where, C is the number of clusters, N is the number of data points, c_j, is the center vector for cluster j, and δ_{ij} is the degree of membership for the i^{th} data point x_i in cluster j, $\left\| x_i - c_j \right\|$ measures the closeness of the data point x_i to the center vector c_j of cluster j. The degree of membership for the i^{th} data point x_i in cluster j, given by δ_{ij} is calculated as,

$$\partial_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}}$$

$$1 < m < \infty, \ \sum_{j}^{C} \partial_{ij} = 1$$

Here, m is the fuzziness coefficient which determines the degree of overlap of clusters with one another and the default value of m is selected as 2. As the value of m increase, the overlap between clusters increases and vice versa.

A2.2 Analysis of the resolution of detection

The analysis of detection resolution described in section 4.6.4, is performed using FCM based approach. The dataset 2 described in section 4.6.1 is used to study the resolution of detection. The simulated data is processed using the EES method with FCM based segmentation instead of K-means segmentation approach. Fig.A2.1a and Fig.A2.1b shows the result of the method for dataset 2 with noise $\sigma = 0.1$ and 0.25 respectively. It can be observed that the FCM approach detects all the five aberrations as in the case of K-Means approach.

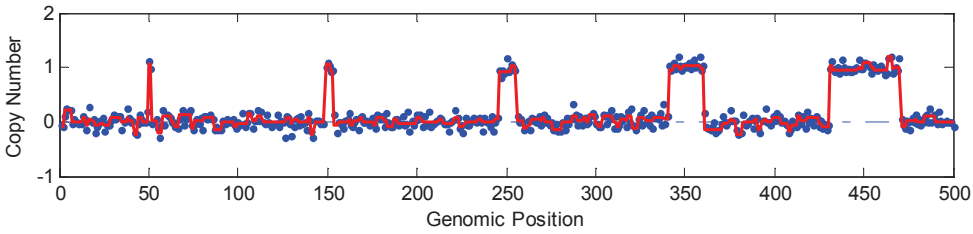


Figure A2.1a Result of EES algorithm using FCM for simulated dataset 2 with $\sigma = 0.1$

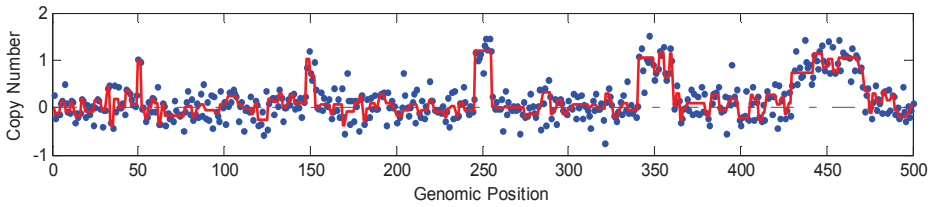


Figure A2.1b Result of EES algorithm using FCM for simulated dataset 2 with $\sigma = 0.25$

A comparison of the number of True Positives and False Positives identified by EES algorithm with K-Means approach and FCM approach is performed and is shown in Table A2.1. Both the methods produced similar results except for σ = 0.225, where, K-Means approach is slightly better in terms of FP detection.

Table A2.1 No. of TP & FP probes identified by the EES algorithm with K-Means and FCM approach

Noise level →	**σ = 0.1**	**σ = 0.15**	**σ = 0.2**	**σ = 0.225**	**σ = 0.25**
True Positives/ False Positives for K-Means	77/0	77/0	76/0	76/0	76/1
True Positives/ False Positives for FCM	77/0	77/0	76/0	76/1	76/1
Actual Positives	77	77	77	77	77

A2.2 Analysis of Receiver Operating Characteristics (ROC) and Area Under the Curve (AUC)

ROC analysis is also performed to compare the FCM based and K-Means based approaches. It is performed using the steps described in section 4.6.5 using the dataset 3 (section 4.6.1). The noise free data template consists of probe sequence of length 100 with 5 levels of amplitude. Two levels of gaussian noise with, σ= 0.25 & 0.5 are added to the template to generate noisy samples, $Y_{(0.25)}, Y_{(0.5)}$. The data Y is then processed by the EES method using both K-Means and FCM based approaches and the threshold for classifying the region into an aberration is gradually varied

from the -2 to 2. The TPR, FPR values for all thresholds are calculated and plotted on the ROC curve (Fig.A2.2a & Fig.A2.2b), the blue and red plots show the ROC curve corresponding to K-Means and FCM respectively. The Area Under the Curve (AUC) is calculated for both approaches and is shown in Table A2.2.

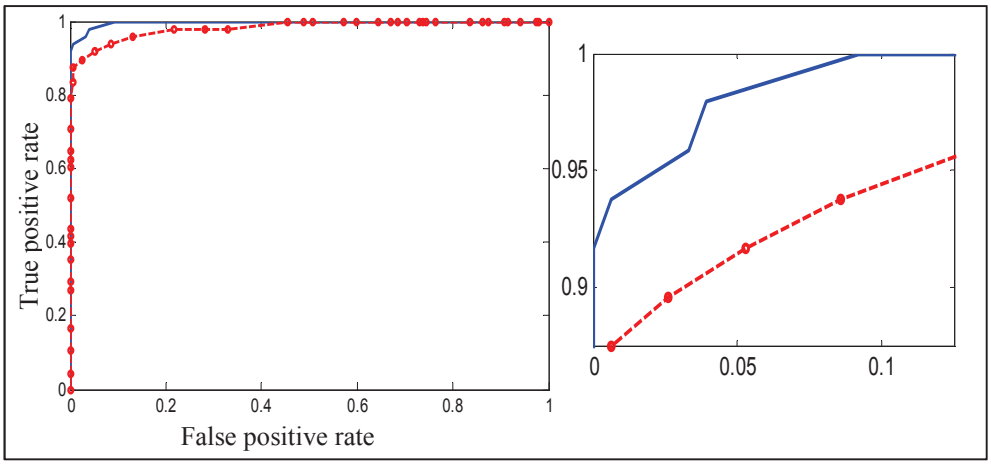


Figure A2.2a ROC plot for simulated dataset 3 with σ = 0.25.

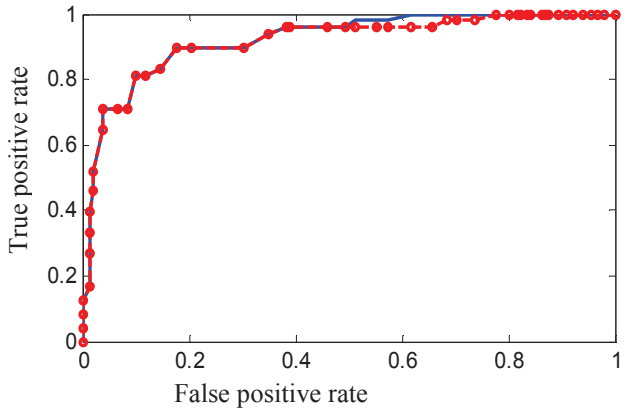


Figure A2.2b ROC plot for simulated dataset 3 with σ = 0.5

The ROC plot and the AUC values corresponding to the two approaches show that both, K-Means and FCM based methods produce similar results with K-Means slightly better.

Table A2.2 Comparison of AUC for K-Means and FCM approaches

	σ = 0.25	**σ = 0.5**
K-Means	0.995	0.922
FCM	0.983	0.915

A2.3 Analysis of real Array CGH data

Coriell cell line data

The coriell cell line dataset described in section 5.2.1 is used in this analysis using FCM based segmentation. Using the FCM based EES algorithm, all the chromosomes in each of the 15 cell lines are tested and the copy number variations are observed.

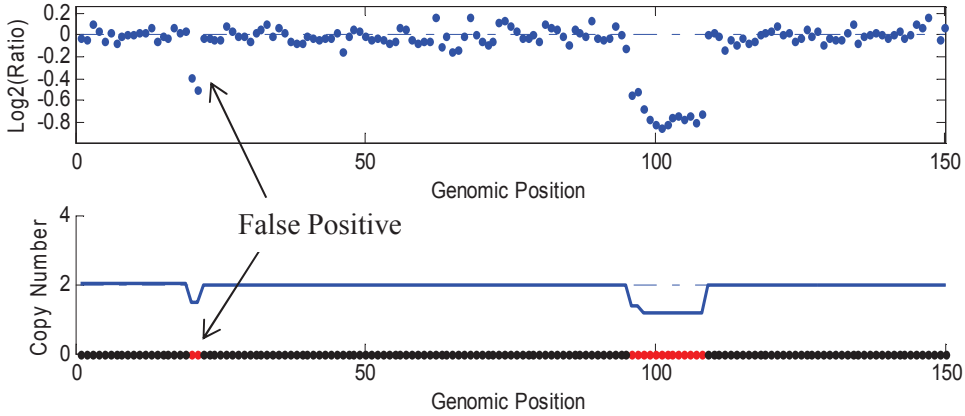


Figure A2.3 Results obtained for Coriell cell line sample GM03134, chromosome 8 using the FCM based EES method.

The method gave the same result as that of K-Means based approach and identified 21 out of the 22 aberrations as shown in Table 5.1. In addition, the FCM based method detected two monosomies in chromosome number 8 of the sample, GM03134, as shown in Fig. A2.3. According to cytogenetically mapped results, there is only one monosomy in this region and the first aberration shown in Fig A2.3 is a false positive detection. Meanwhile, K-Means based approach resulted in no FP detection (Fig 5.1i) compared to FCM based approach. The results of these analysis show that Fuzzy-C Means clustering can be incorporated in the EES algorithm without much difference in the performance of CNV detection.

A3. Analysis of CNV in Coriell cell line database using alternate methods.

Analysis of copy number variation in a real data set, Coriell cell line BAC Array CGH data, described in Snijders et al. (2001), using other established methods is performed. It is used for evaluation of performance of Array CGH algorithms as the exact locations of aberrations are already known. The data set is analyzed using Wavelet, Quantreg, CBS and CGHSeg methods. Table A3.1 given below shows the result of different methods in identifying the mapped aberrations. It can be noted that all the methods failed to detect the monosomy in chromosome 15 of GM07081. Table A3.2 shows the number of aberrations detected by the different methods and comparison of TP & FP counts. The Wavelet method detected 22 out of the 23 known aberrations and resulted in 82 false detections. The Quantreg and CBS detected 21 out of the 23 known aberration and resulted in 19, 7 false positives respectively. Among the methods CGHSeg performed better by detecting 22 out of 23 known aberrations along with 9 false positives. A comparison of this result with that obtained using the EES method described in section 5.3.1 shows that EES method failed to detect the single point aberration in GM01535, chromosome 12 (Fig5.2a) while CGHSeg and Wavelet detected it. This capability of CGHSeg and Wavelet method also resulted in wrong classification of 8 single point and 78 single point changes as valid aberrations.

Table A3.1 List of aberrations in the 15 coriell cell lines detected by different methods.

Cell Line / Chromosome	**Mapped Aberrations**	**Regions detected by**			
		Wavelet	**Quantreg**	**CBS**	**CGHSeg**
GM03563/ 3	Trisomy 3q12-3qter	Detected	Detected	Detected	Detected
GM03563/ 9	Monosomy 9pter-9p24	Detected	Detected	Detected	Detected
GM00143/18	Trisomy on whole 18	Detected	Detected	Detected	Detected
GM05296/10	Trisomy 10q21-10q24	Detected	Detected	Detected	Detected
GM05296/11	Monosomy 11p12-11p13	Detected	Detected	Detected	Detected
GM07408	Trisomy on whole 20	Detected	Detected	Detected	Detected
GM01750/9	Trisomy 9pter-9p24	Detected	Detected	Detected	Detected
GM01750/14	Trisomy 14pter-14q21	Detected	Detected	Detected	Detected
GM03134/8	Monosomy 8q13-8q22	Detected	Detected	Detected	Detected
GM13330/1	Trisomy 1q25-1qter	Detected	Detected	Detected	Detected
GM13330/4	Monosomy 4q35-4qter	Detected	Detected	Detected	Detected

GM03576/2	Trisomy on whole 2	Detected	Detected	Detected	Detected
GM03576/21	Trisomy on whole 21	Detected	Detected	Detected	Detected
GM01535/5	Trisomy 5q33-5qter	Detected	Detected	Detected	Detected
GM01535/12	Monosomy 12q24-12qter	Detected	Missed	Missed	Detected
GM07081/7	Trisomy 7pter-7q11.2	Detected	Detected	Detected	Detected
GM07081/15	Monosomy 15pter-15q11.2	Missed	Missed	Missed	Missed
GM02948/13	Trisomy on whole 13	Detected	Detected	Detected	Detected
GM04435/16	Trisomy on whole 16	Detected	Detected	Detected	Detected
GM04435/21	Trisomy on whole 21	Detected	Detected	Detected	Detected
GM10315/22	Trisomy on whole 22	Detected	Detected	Detected	Detected
GM13031/17	Monosomy 17q21.3-17q23	Detected	Detected	Detected	Detected
GM01524/6	Trisomy 6q15-6q25	Detected	Detected	Detected	Detected

Table A3.2 No.of aberrations detected by different methods and comparison of TP & FP counts.

Cell Line	**Copy number gains and losses detected by**					**Actual CNVs**
	Wavelet	**Quantreg**	**CBS**	**CGHSeg**	**EES**	
GM03563	8 gain 3 loss	2 gain 1 loss	2 gain 1 loss	1 gain 1 loss	1 gain 1 loss	1 gain 1 loss
GM00143	6 gain 3 loss	2 gain 4 loss	1 gain	1 gain	1 gain	1 gain
GM05296	5 gain 5 loss	1 gain 1 loss	1 gain 1 loss	1 gain 3 loss	1 gain 1 loss	1 gain 1 loss
GM07408	6 gain 2 loss	1 gain	1 gain	1 gain	1 gain	1 gain
GM01750	4 gain	3 gain	2 gain	3 gain	2 gain	2 gain
GM03134	4 loss	2 loss	1 gain 3 loss	3 loss	1 loss	1 loss
GM13330	4 gain 2 loss	2 gain 3 loss	1 gain 1 loss	1 gain 1 loss	1 gain 1 loss	1 gain 1 loss
GM03576	5 gain	2 gain	2 gain	2 gain	2 gain	2 gain
GM01535	4 gain 3 loss	1 gain	1 gain	2 gain 1 loss	1 gain	1 gain 1 loss
GM07081	11 gain 8 loss	1 gain	1 gain	1 gain	1 gain	1 gain 1 loss

GM02948	5 gain 1 loss	4 gain 2 loss	3 gain	1 gain	1 gain	1 gain
GM04435	8 gain 3 loss	2 gain 1 loss	2 gain 1 loss	2 gain 1 loss	2 gain	2 gain
GM10315	2 gain 1 loss	1 gain 1 loss	1 gain	1 gain	1 gain	1 gain
GM13031	1 gain 3 loss	2 loss	1 loss	2 loss	1 loss	1 loss
GM01524	2 gain 2 loss	1 gain	1 gain	1 gain	1 gain	1 gain
Total True Positives	**22**	**21**	**21**	**22**	**21**	**23**
Total False Positives	**82**	**19**	**7**	**9**	**0**	**0**

A4. Protein sequence similarity analysis using DFT

The aim of this work is to illustrate protein sequence similarity analysis using DFT of numerically mapped amino acid sequences. The algorithm involves three steps. In the first step, amino acid sequence of a protein is obtained from a set of species and is transformed into a numerical sequence using the EIIP method described in section 7.5. In the second step, Discrete Fourier Transform (DFT) of this numerical signal is obtained, which represents the signal in frequency domain showing the frequencies that constitute the signal. In the final step of analysis, the Manhattan distance between DFT of a pair of sequence is obtained, which gives a similarity measure between them. The similarity analysis is performed on two datasets described in section 7.7.3.

A4.1 Results & Discussion

Three experiments are performed on dataset 1 for the analysis of protein sequence similarity. The results of the experiments are shown in Fig A4.1(a-c). The y-axis shows the sequence similarity, C_{xy} ($0<C_{xy}<1$). In the first experiment prolactin sequences from 12 species are compared with that of cat using the DFT method. It can be observed from Fig A4.1a that cat shows highest similarity to the organisms, mink and panda. Both these organisms belong to the order- carnivora, to which the cat belongs. The second experiment compared prolactin sequence from human with all the other species. It shows similar result, with the human sequence showing highest similarity to baboon, chimpanzee, gorilla and rhesus monkey. (Fig A4.1b). All the 5 species belongs to the same order- primate. The third experiment compared the sequence of ostrich with other species and the

result clearly supported the expectations where it showed no significant sequence similarity with any of the species (Fig A4.1c).

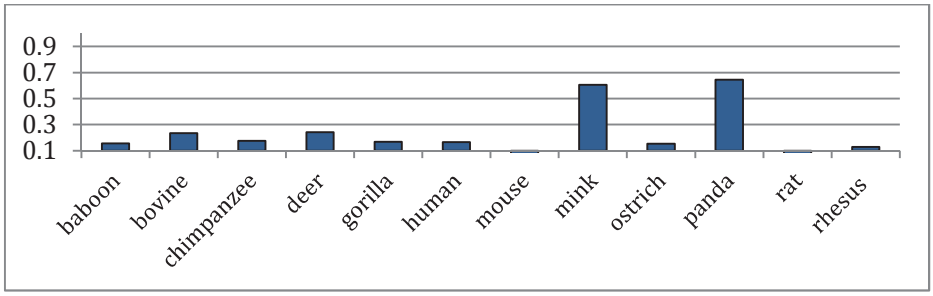


Figure A4.1 (a) Similarity of prolactin from cat with other species

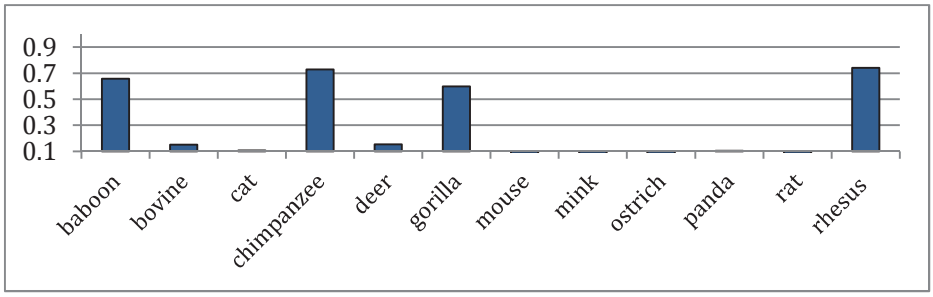


Figure A4.1 (b) Similarity of prolactin from human with other species

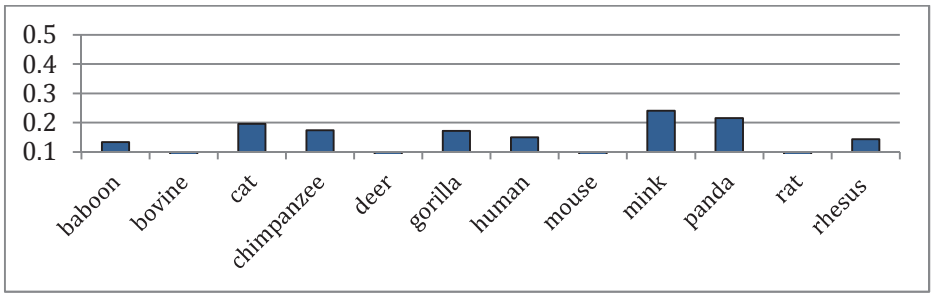


Figure A4.1 (c) Similarity of prolactin from ostrich with other species

The second dataset involves 4 experiments and the results are shown in Fig A4.2(a-d). In the first experiment human somatotropin sequence is compared with 15 other species' sequence. It can be seen from Fig. A4.2a,

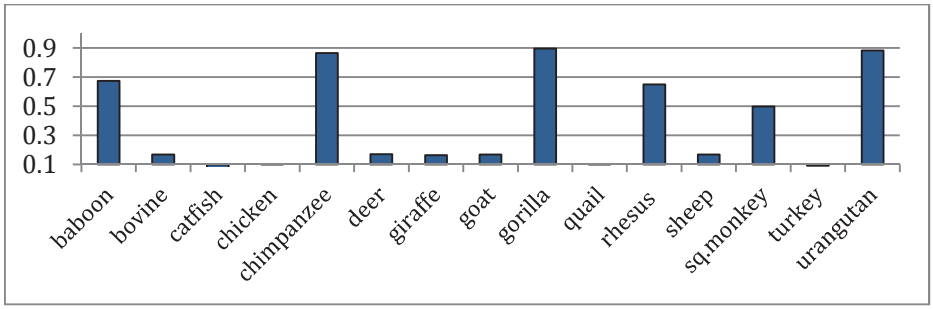


Figure A4.2 (a) Similarity of somatotropin from human with other species

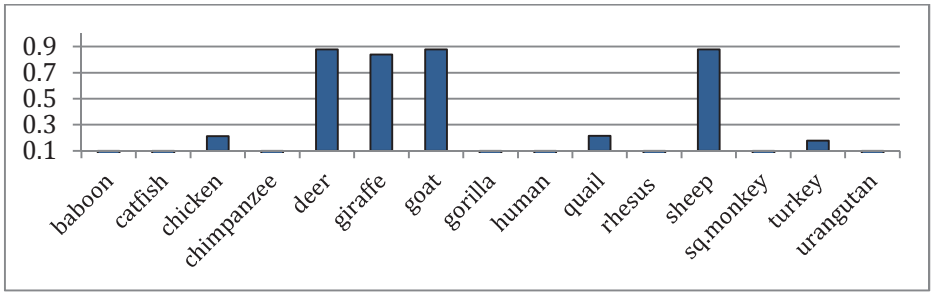


Figure A4.2 (b) Similarity of somatotropin from bovine with other species

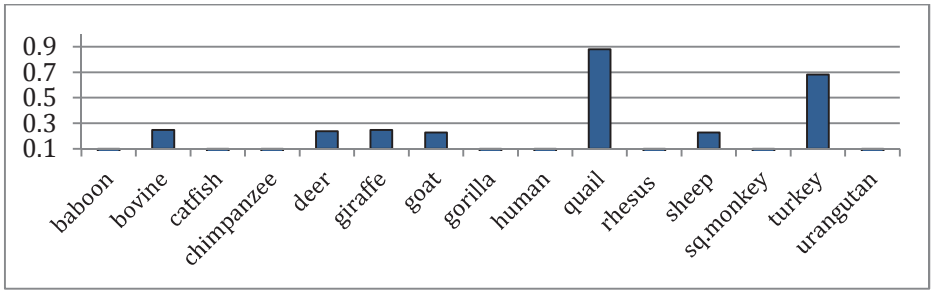


Figure A4.2 (c) Similarity of somatotropin from chicken with other species

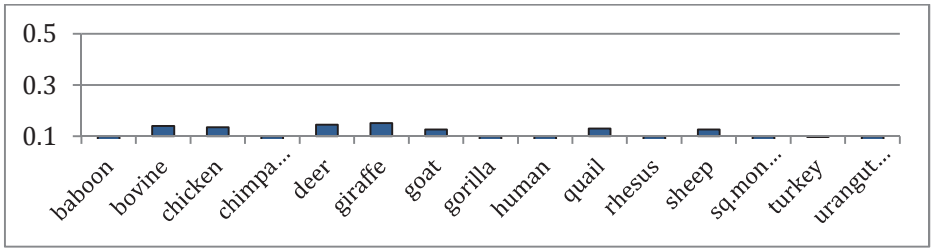


Figure A4.2 (d) Similarity of somatotropin from catfish with other species

that all the 6 primates show high sequence similarity values compared to other groups. The second experiment compares bovine sequence with the rest. The result shows, (Fig A4.2b) almost 90% similarity with all the 4 species belonging to the same order- artiodactyla. In the third experiment somatotropin sequence of chicken is compared with other sequences. The result obtained is shown in Fig A4.2c, where it clearly shows strong similarity with the other two aves in the set, quail and turkey. The final experiment compares the sequence from a species, catfish with the others. The catfish is a species that is external to all the groups and represents a sample that is distantly related to all other species. As expected, the catfish somatotropin sequence showed low sequence similarity with all other sequences. The study clearly indicates the presence of strong spectral similarity in closely related species and also showed that DFT based approach also offers same result obtained using DWT based analysis described in section 7.7.2.

References

Adachi, J. and Hasegawa, M., 1992. Protml: Maximum likelihood inference of protein phylogeny. *Tokyo: Computer Science Monographs of the Institute of Statistical Mathematics*.

Adams III, E.N., 1972. Consensus techniques and the comparison of taxonomic trees. *Systematic Biology*, *21*(4), pp.390-397.

Anastassiou, D., 2001. Genomic signal processing. *IEEE signal processing magazine*, 18(4), pp.8-20.

Ané, C., Larget, B., Baum, D.A., Smith, S.D. and Rokas, A., 2006. Bayesian estimation of concordance among gene trees. *Molecular biology and evolution*, *24*(2), pp.412-426.

Atchley, W.R., Zhao, J., Fernandes, A.D. and Drüke, T., 2005. Solving the protein sequence metric problem. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(18), pp.6395-6400.

Autio, R., Hautaniemi, S., Kauraniemi, P., Yli-Harja, O., Astola, J., Wolf, M. and Kallioniemi, A., 2003. CGH-Plotter: MATLAB toolbox for CGH-data analysis. *Bioinformatics*, *19*(13), pp.1714-1715.

Barthélemy, J.P. and McMorris, F.R., 1986. The median procedure for n-trees. *Journal of classification*, *3*(2), pp.329-334.

Beheshti, B., Braude, I., Marrano, P., Thorner, P., Zielenska, M. and Squire, J.A., 2003. Chromosomal localization of DNA amplifications in neuroblastoma tumors using cDNA microarray comparative genomic hybridization. *Neoplasia*, *5*(1), pp.53-62.

Beyer, H., 1981. Tukey, John W.: Exploratory Data Analysis. Addison-Wesley Publishing Company Reading, Mass.—Menlo Park, Cal., London, Amsterdam, Don Mills, Ontario, Sydney 1977, XVI, 688 S. *Biometrical Journal*, *23*(4), pp.413-414.

Borrayo, E., Mendizabal-Ruiz, E.G., Vélez-Pérez, H., Romo-Vázquez, R., Mendizabal, A.P. and Morales, J.A., 2014. Genomic signal processing methods for computation of alignment-free distances from DNA sequences. *PloS one*, *9*(11), p.e110954.

Bredel, M., Bredel, C., Juric, D., Harsh, G.R., Vogel, H., Recht, L.D. and Sikic, B.I., 2005. High-resolution genome-wide mapping of genetic alterations in human glial brain tumors. *Cancer research*, *65*(10), pp.4088-4096.

Bridges CB., 1936. The Bar 'gene': a duplication. *Science*. 83, pp.210–211.

Brown, T.A., 2007. *Genomes 3*. 3rd ed. Newyork & London: Garland science publishing.

Bryant, D., 2003. A classification of consensus methods for phylogenetics. *DIMACS series in discrete* mathematics *and theoretical computer science*, *61*, pp.163-184.

Buneman, P. 1971. The Recovery of Trees from Measures of Dissimilarity. In *Mathematics in the Archaeological and Historical Sciences*, F. R. Hodson, D. G. Kendall and P. Tautu (Eds), Edinburgh: Edinburgh University Press. pp. 387–395.

Carvalho, B., Ouwerkerk, E., Meijer, G.A. and Ylstra, B., 2004. High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides. *Journal of clinical pathology*, *57*(6), pp.644-646.

Caspersson, T., Lomakka, G.and Zech, L., 1971. The 24 fluorescence patterns of the human metaphase chromosomes—distinguishing characters and variability. *Hereditas*, 67(1), pp.89-102.

Chen, Y., Dougherty, E.R. and Bittner, M.L., 1997. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical optics*, 2(4), pp.364-374.

Chu, K.H., Qi, J., Yu, Z.G. and Anh, V.O., 2004. Origin and phylogeny of chloroplasts revealed by a simple correlation analysis of complete genomes. *Molecular biology and evolution*, *21*(1), pp.200-206.

Cosic, I., 1994. Macromolecular bioactivity: is it resonant interaction between macromolecules?-theory and applications. *IEEE Transactions on Biomedical Engineering*, *41*(12), pp.1101-1114.

Cosic, I. and Pirogova, E., 1998, October. Application of ionisation constant of amino acids for protein signal analysis within the resonant recognition model. In *Engineering in Medicine and Biology Society, 1998. Proceedings of the 20th Annual International Conference of the IEEE* (Vol. 2, pp. 1072-1075). IEEE.

Coughlin, C.R., Scharer, G.H. and Shaikh, T.H., 2012. Clinical impact of copy number variation analysis using high-resolution microarray technologies: advantages, limitations and concerns. *Genome medicine*, *4*(10), p.80.

Darwin, C. R., 1859. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: John Murray.

Demirkaya, O., Asyali, M.H. and Shoukri, M.M., 2005. Segmentation of cDNA microarray spots using markov random field modeling. *Bioinformatics*, 21(13), pp.2994-3000.

Eilers, P.H. and De Menezes, R.X., 2004. Quantile smoothing of array CGH data. *Bioinformatics*, *21*(7), pp.1146-1153.

Eisen, M., 1999. ScanAlyze user manual. Stanford University, USA.

Ewing, G.B., Ebersberger, I., Schmidt, H.A. and Von Haeseler, A., 2008. Rooted triple consensus and anomalous gene trees. *BMC evolutionary biology*, *8*(1), p.118.

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, *39*(4), pp.783-791.

Felsenstein, J., 2005. PHYLIP (phylogeny inference package) version 3.6. Distributed by Author. Department of Genome Sciences, University of Washington,Seattle.
http://evolution.genetics.washington.edu/phylip.html

Feuk, L., Carson, A.R. and Scherer, S.W., 2006. Structural variation in the human genome. *Nature reviews. Genetics*, *7*(2), pp.85.

Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T. and Solas, D., 1991. Light-directed, spatially addressable parallel chemical synthesis. *science*, pp.767-773.

Freeman, J.L., Perry, G.H., Feuk, L., Redon, R., McCarroll, S.A., Altshuler, D.M., Aburatani, H., Jones, K.W., Tyler-Smith, C., Hurles, M.E. and Carter, N.P., 2006. Copy number variation: new insights in genome diversity. *Genome research*, *16*(8), pp.949-961.

Fridlyand, J., Snijders, A.M., Pinkel, D., Albertson, D.G. and Jain, A.N., 2004. Hidden Markov models approach to the analysis of array CGH data. *Journal of multivariate analysis*, *90*(1), pp.132-153.

Gao, Y. and Luo, L., 2012. Genome-based phylogeny of dsDNA viruses by a novel alignment-free method. *Gene*, *492*(1), pp.309-314.

Glazier, J.A., Raghavachari, S., Berthelsen, C.L. and Skolnick, M.H., 1995. Reconstructing phylogeny from the multifractal spectrum of mitochondrial DNA. *Physical review E*, *51*(3), p.2665.

Guha, S., Li, Y. and Neuberg, D., 2008. Bayesian hidden Markov modeling of array CGH data. *Journal of the American Statistical Association, 103*(482), pp.485-497.

Hastings, P.J., Lupski, J.R., Rosenberg, S.M. and Ira, G., 2009. Mechanisms of change in gene copy number. *Nature reviews. Genetics*, *10*(8), pp.551.

Hatzigeorgiou, A., Mache, N. and Reczko, M., 1996, November. Functional site prediction on the DNA sequence by artificial neural networks. In Intelligence and Systems. *IEEE Int. Joint Symp. Intell. Syst.* pp.12-17.

Haubold, B., Pfaffelhuber, P., Domazet-Losˇo, M. and Wiehe, T., 2009. Estimating mutation distances from unaligned genomes. *Journal of Computational Biology*, *16*(10), pp.1487-1500.

Heim, S. and Mitelman. F., 2015. *Cancer cytogenetics: chromosomal and molecular genetic aberrations of tumor cells*. 4th ed. Wiley Blackwell.

Heiskanen, M.A., Bittner, M.L., Chen, Y., Khan, J., Adler, K.E., Trent, J.M. and Meltzer, P.S., 2000. Detection of gene amplification by genomic hybridization to cDNA microarrays. *Cancer Research*, *60*(4), pp.799-802.

Hill, R.L., Buettner-Janusch, J. and Buettner-Janusch, V., 1963. Evolution of hemoglobin in primates. *Proceedings of the National Academy of Sciences*, *50*(5), pp.885-893.

Hodgson, G., Hager, J.H., Volik, S., Hariono, S., Wernick, M., Moore, D., Albertson, D.G., Pinkel, D., Collins, C., Hanahan, D. and Gray, J.W., 2001. Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nature genetics*, *29*(4), pp.459.

Hsu, L.I., Self, S.G., Grove, D., Randolph, T., Wang, K., Delrow, J.J., Loo, L. and Porter, P., 2005. Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, *6*(2), pp.211-226.

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

Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W. and Lee, C., 2004. Detection of large-scale variation in the human genome. *Nature genetics*, *36*(9), p.949.

Jain, A.N., Tokuyasu, T.A., Snijders, A.M., Segraves, R., Albertson, D.G. and Pinkel, D., 2002. Fully automatic quantification of microarray image data. *Genome research*, 12(2), pp.325-332.

Jong, K., Marchiori, E., Van Der Vaart, A., Ylstra, B., Weiss, M. and Meijer, G., 2003. Chromosomal breakpoint detection in human cancer. *Lecture Notes in Computer Science*, pp.54-65.

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

Karki, R., Pandya, D., Elston, R.C. and Ferlini, C., 2015. Defining "mutation" and "polymorphism" in the era of personal genomics. *BMC medical genomics*, 8(1), pp.37.

Kumar, S., Stecher, G. and Tamura, K., 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Molecular biology and evolution, 33(7), pp.1870-1874.

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

Lai, W.R., Choudhary, V., and Park, P. J., 2008. CGHweb: a tool for comparing DNA copy number segmentations from multiple algorithms. *Bioinformatics,* 24(7), pp.1014–1015.

Langer-Safer, P. R., Levine, M., Ward, D. C., 1982. Immunological method for mapping genes on Drosophila polytene chromosomes. *Proceedings of the National Academy of Sciences,* 79(14), pp.4381.

Lapointe, F.J. and Cucumel, G., 1997. The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa. *Systematic Biology*, *46*(2), pp.306-312.

Lazović, J., 1996. Selection of amino acid parameters for Fourier transform-based analysis of proteins. *Bioinformatics*, 12(6), pp.553-562.

Lee, J.A., Carvalho, C.M. and Lupski, J.R., 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *cell*, 131(7), pp.1235-1247.

Lingjærde, O.C., Baumbusch, L.O., Liestøl, K., Glad, I.K. and Børresen-Dale, A.L., 2004. CGH-Explorer: a program for analysis of array-CGH data. *Bioinformatics*, *21*(6), pp.821-822.

Liu, L. and Yu, L., 2011. Estimating species trees from unrooted gene trees. *Systematic biology*, *60*(5), pp.661-667.

Locke, D.P., Segraves, R., Carbone, L., Archidiacono, N., Albertson, D.G., Pinkel, D. and Eichler, E.E., 2003. Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome research*, *13*(3), pp.347-357.

Ma, Q., Wang, J.T., Shasha, D. and Wu, C.H., 2001. DNA sequence classification via an expectation maximization algorithm and neural networks: a case study. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 31(4), pp.468-475.

Maddison, W.P., 1997. Gene trees in species trees. *Systematic biology*, *46*(3), pp.523-536.

Mallo, D. and Posada, D., 2016. Multilocus inference of species trees and DNA barcoding. *Phil. Trans. R. Soc. B*, *371*(1702), p.20150335.

Marsella, L., Sirocco, F., Trovato, A., Seno, F. and Tosatto, S.C.E., 2009. REPETITA: detection and discrimination of the periodicity of protein solenoid repeats by discrete Fourier transform, *Bioinformatics,* Vol.25, pp.285-295.

Matsuda, H., 1995. Construction of phylogenetic trees from amino acid sequences using a genetic algorithm. *Genome Informatics*, *6*, pp.19-28.

McMorris, F.R., Meronk, D.B. and Neumann, D.A., 1983. A view of some consensus methods for trees. In *Numerical Taxonomy* (pp. 122-126). Springer Berlin Heidelberg.

Nadler, S.A., 1995. Advantages and disadvantages of molecular phylogenetics: A case study of ascaridoid nematodes. *Journal of Nematology*, *27*(4), p.423.

Nair, A.S. and Mahalakshmi, T., 2006. Are categorical periodograms and indicator sequences of genomes spectrally equivalent?. *In silico biology*, *6*(3), pp.215-222.

Nakao, K., Mehta, K.R., Fridlyand, J., Moore, D.H., Jain, A.N., Lafuente, A., Wiencke, J.W., Terdiman, J.P. and Waldman, F.M., 2004. High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. *Carcinogenesis*, *25*(8), pp.1345-1357.

National Center for Biotechnology Information. *NCBI*.[online] Available at: http://www.ncbi.nlm.nih.gov/ [Accessed till Aug. 2017].

Neve, R.M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F.L., Fevr, T., Clark, L., Bayani, N., Coppe, J.P., Tong, F. and Speed, T., 2006. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer cell*, *10*(6), pp.515-527.

Nguyen, N., Huang, H., Oraintara, S. and Vo, A., 2007, October. A new smoothing model for analyzing array CGH data. In *Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on* (pp. 1027-1034). IEEE.

Nguyen, N., Huang, H., Oraintara, S. and Vo, A., 2010. Stationary wavelet packet transform and dependent Laplacian bivariate shrinkage estimator for array-CGH data smoothing. *Journal of Computational Biology*, *17*(2), pp.139-152.

Olshen, A.B., Venkatraman, E.S., Lucito, R. and Wigler, M., 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, *5*(4), pp.557-572.

Orsetti, B., Nugoli, M., Cervera, N., Lasorsa, L., Chuchana, P., Ursule, L., Nguyen, C., Redon, R., Du Manoir, S., Rodriguez, C. and Theillet, C., 2004. Genomic and expression profiling of chromosome 17 in breast cancer reveals complex patterns of alterations and novel candidate genes. *Cancer research*, *64*(18), pp.6453-6460.

Papadopoulos, J.S. and Agarwala, R., 2007. COBALT: constraint-based alignment tool for multiple protein sequences. Bioinformatics, 23(9), pp.1073-1079.

Pease, A.C., Solas, D., Sullivan, E.J., Cronin, M.T., Holmes, C.P. and Fodor, S.P., 1994. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proceedings of the National Academy of Sciences*, *91*(11), pp.5022-5026.

Picard, F., Robin, S., Lavielle, M., Vaisse, C. and Daudin, J.J., 2005. A statistical approach for array CGH data analysis. *BMC bioinformatics*, *6*(1), p.27.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.L., Chen, C., Zhai, Y. and Dairkee, S.H., 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature genetics*, *20*(2).

Pirogova, E., Simon, G.P. and Cosic, I., 2003. Investigation of the applicability of dielectric relaxation properties of amino acid solutions within the resonant recognition model. *IEEE transactions on nanobioscience*, *2*(2), pp.63-69.

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

Pollack, J.R., Sørlie, T., Perou, C.M., Rees, C.A., Jeffrey, S.S., Lonning, P.E., Tibshirani, R., Botstein, D., Børresen-Dale, A.L. and Brown, P.O., 2002. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences*, *99*(20), pp.12963-12968.

Posada, D., 2016. Phylogenomics for systematic biology. *Systematic biology*, *65*(3), pp.353-356.

Qi, J., Wang, B. and Hao, B.I., 2004. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *Journal of molecular evolution*, *58*(1), pp.1-11.

Qi, X., Wu, Q., Zhang, Y., Fuller, E. and Zhang, C.Q., 2011. A novel model for DNA sequence similarity analysis based on graph theory. *Evolutionary bioinformatics online*, *7*, p.149.

Qi, X., Fuller, E., Wu, Q. and Zhang, C.Q., 2012. Numerical characterization of DNA sequence based on dinucleotides. *The Scientific World Journal*, *2012*.

Rahnenführer, J. and Bozinov, D., 2004. Hybrid clustering for microarray image analysis combining intensity and shape features. *BMC bioinformatics*, 5(1), pp.47.

Ramachandran, P. and Antoniou, A., 2008. Identification of hot-spot locations in proteins using digital filters. *IEEE Journal of selected topics in signal processing*, Vol. 2, No.3, pp.378–389.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W. and Cho, E.K., 2006. Global variation in copy number in the human genome. N*ature*, *444*(7118), pp.444.

Russell, S., Meadows, L.A. and Russell, R.R., 2008. *Microarray technology in practice*. Academic Press.

Salemi, M., Lemey, P. and Vandamme, A.M. eds., 2009. The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. Cambridge University Press.

Salichos, L. and Rokas, A., 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, *497*(7449), p.327.

Schena, M., Shalon, D., Davis, R.W. and Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *SCIENCE-NEW YORK THEN WASHINGTON*, pp.467-467.

Schulz-Schaeffer, J., 1980. *Cytogenetics: plants, animals, humans*. New York: Springer-Verlag.

Seabright, M., 1971. Rapid banding technique for human chromosomes. *Lancet*,.2(7731). pp.971-972.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M. and Navin, N., 2004. Large-scale copy number polymorphism in the human genome. *Science*, *305*(5683), pp.525-528.

Sims, G.E. and Kim, S.H., 2011. Whole-genome phylogeny of Escherichia coli/Shigella group by feature frequency profiles (FFPs). *Proceedings of the National Academy of Sciences*, *108*(20), pp.8329-8334.

Snijders, A.M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A.K., Huey, B., Kimura, K. and Law, S., 2001. Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature genetics*, *29*(3), p.263.

Snijders, A.M., Fridlyand, J., Mans, D.A., Segraves, R., Jain, A.N., Pinkel, D. and Albertson, D.G., 2003. Shaping of tumor and drug-resistant genomes by instability and selection. *Oncogene*, *22*(28), p.4370.

Sokal, R. and Michener, C., 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*. Vol.38, pp.1409–1438.

Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Döhner, H., Cremer, T. and Lichter, P., 1997. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes, chromosomes and cancer*, *20*(4), pp.399-407.

Southern, E.M., 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of molecular biology*, *98*(3), pp.503-517.

Stinebrickner, R., 1984. s-Consensus trees and indices. *Bulletin of Mathematical Biology*, *46*(5-6), pp.923-935.

Thompson, J.D., Higgins, D.G. and Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic acids research, 22(22), pp.4673-4680.

Tjio, J.H. and Levan, A., 1956. The chromosome number of man. *Hereditas* Vol.42, Issue.1-2.

Tomita, F. and Tsuji, S., 1977. Extraction of Multiple Regions by Smoothing in Selected Neighborhoods, *IEEE Trans. on Systems, Man and Cybernetics*, Vol.7,Issue:2, pp.107-109.

Tomiuk, S. and Hofmann, K., 2001. Microarray probe selection strategies. *Briefings in bioinformatics*, *2*(4), pp.329-340.

Trad, C. H., Fang, Q. and Cosic, I., 2002. Protein sequence comparison based on the wavelet transform approach. *Protein Eng.*, Vol. 15, pp.193–203.

Ulitsky, I., Burstein, D., Tuller, T. and Chor, B., 2006. The average common substring approach to phylogenomic reconstruction. *Journal of Computational Biology*, *13*(2), pp.336-350.

Vaidyanathan, P.P. and Yoon, B.J., 2002, October. Gene and exon prediction using allpass-based filters. In *Proc. IEEE Workshop on Gen. Sig. Proc and Stat.*

Vaidyanathan, P.P., 2004. Genomics and proteomics: A signal processor's tour. *IEEE circuits and systems magazine*, *4*(4), pp.6-29.

Veljković, V. and Slavić, I., 1972. Simple general-model pseudopotential. *Physical Review Letters*, *29*(2), p.105.

Veljkovic, V., Cosic, I., Dimitrijevic, B. and Lalovic, D., 1985. Is it possible to analyze DNA and protein sequences by the methods of digital signal processing?. *IEEE Transactions on Biomedical Engineering*, Vol.32, No.5, pp.337-341.

Vo, A., Nguyen, N. and Huang, H., 2010. Solenoid and non-solenoid protein recognition using stationary wavelet packet transform. *Bioinformatics*, *26*(18), pp.i467-i473.

Wang, L.Y., Abyzov, A., Korbel, J.O., Snyder, M. and Gerstein, M., 2009. MSB: a mean-shift-based approach for the analysis of structural variation in the genome. *Genome research*, *19*(1), pp.106-117.

Wang, P., Kim, Y., Pollack, J., Narasimhan, B. and Tibshirani, R., 2005. A method for calling gains and losses in array CGH data. *Biostatistics*, *6*(1), pp.45-58.

Wang, X.H., Istepanian, R.S. and Song, Y.H., 2003. Application of wavelet modulus maxima in microarray spots recognition. *IEEE transactions on nanobioscience*, 2(4), pp.190-192.

Wang, Y., Shih, F. and Ma, M., 2005, July. Precise gridding of microarray images by detecting and correcting rotations in subarrays. In *Proceedings of the 8th Joint Conference on Information Science,* pp.1195-1198.

Wang, Y. and Wang, S., 2007. A novel stationary wavelet denoising algorithm for array-based DNA Copy Number data. *International Journal of Bioinformatics Research and Applications*, *3*(2), pp.206-222.

Weiss, M.M., Snijders, A.M., Kuipers, E.J., Ylstra, B., Pinkel, D., Meuwissen, S.G., van Diest, P.J., Albertson, D.G. and Meijer, G.A., 2003. Determination of amplicon boundaries at 20q13. 2 in tissue samples of human gastric adenocarcinomas by high-resolution microarray comparative genomic hybridization. *The Journal of pathology*, *200*(3), pp.320-326.

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

Yang, Y.H., Buckley, M.J., Dudoit, S. and Speed, T.P., 2002. Comparison of methods for image analysis on cDNA microarray data. *Journal of computational and graphical statistics*, 11(1), pp.108-136.

Zhang, A., 2006. *Advanced analysis of gene expression microarray data* (Vol. 1). World Scientific Publishing Co Inc.

Zhang, F., Gu, W., Hurles, M.E. and Lupski, J.R., 2009. Copy number variation in human health, disease, and evolution. *Annual review of genomics and human genetics*, *10*, pp.451-481.

Zhao, J., Yang, X.W., Li, J.P. and Tang, Y.Y., 2001. DNA sequences classification based on wavelet packet analysis. In *Wavelet Analysis and Its Applications*. Berlin, Heidelberg: Springer, pp. 424-429.

Zhou, L.Q., Yu, Z.G., Nie, P.R., Liao, F.F., Anh, V.V. and Chen, Y.J., 2007, August. Log-correlation distance and Fourier transform with Kullback-Leibler divergence distance for construction of vertebrate phylogeny using complete mitochondrial genomes. In *Natural Computation, 2007. ICNC 2007. Third International Conference on* (Vol. 2, pp. 304-308). IEEE.

Zuckerkandl, E. and Pauling, L., 1962. Molecular disease, evolution, and genic heterogeneity. In: Kasha M, Pullman B (eds) *Horizons in biochemistry*. New York: Academic Press, pp 189–225.

List of Publications

Anu Sabarish R and Tessamma Thomas, "A frequency domain approach to protein sequence similarity analysis and functional classification", *Signal & Image Processing: An International Journal (SIPIJ).*Vol.2, No.1, March 2011.

Anu Sabarish R and Tessamma Thomas, "Molecular phylogeny analysis using correlation distance and spectral distance", *Int. J. Data Mining and Bioinformatics*, Vol. 10, No. 4, 2014.

Anu Sabarish R and Tessamma Thomas, "Construction of phylogenetic tree from multiple gene trees using principal component analysis", *Int. Journal of Electronics and Communication Engineering & Technology*. Vol.5, Issue.12, Dec-2014.

Anu Sabarish R and Tessamma Thomas, "Dual Step Algorithm for the Detection of Copy Number Variation using Minimum Variance Filtering and K-means Clustering", Submitted to *Int. J. Data Mining and Bioinformatics.*

Resume

Anu Sabarish R
Research Scholar
Department of Electronics,
CUSAT, Cochin-682022
E-mail :anusabarish@cusat.ac.in

Area of Interest:

Signal processing, Image processing, Genomic signal processing, Data mining and Pattern recognition.

Academic Profile:

Period	Course	University/ Board	Grade	Institution
Currently pursuing	Ph.D	CUSAT		Dept. of Electronics CUSAT.
2005-2007	M.Tech (Digital Electronics)	CUSAT	CGPA 9/10	Dept. of Electronics CUSAT.
2001-2005	B.Tech (Electronics & Communication)	Calicut University	First Class	Govt. Engineering College, Thrissur.

Professional Experience:

1. Wipro Technologies, Cochin
   Employed as Project Trainee during July 2006 to May 2007
   Design For Testability in VLSI Domain
2. Wipro Technologies, Cochin
   Employed as Project Engineer during Jan 2008 to July 2009
   VLSI Design and Design For Testability
3. Bharat Sanchar Nigam Ltd, Kerala
   Employed as Junior Telecom Officer since Apr 2010

Publications:

International Journals: 3

Personal Information:

Date of birth : 30-10-1983
Sex : Male
Permanent Address: Sivapriya,
Kadayanickadu P.O,
Kottayam,
Kerala-686541.