2017

Robustness Evaluation for Phylogenetic Reconstruction Methods and Evolutionary Models Reconstruction of Tumor Progression

Jun Zhou
University of South Carolina

Robustness evaluation for phylogenetic reconstruction methods and evolutionary models reconstruction of tumor progression

by

Jun Zhou

Bachelor of Science
Nanjing University 2008

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Computer Science

College of Engineering and Computing

University of South Carolina

2017

Accepted by:

Jijun Tang, Major Professor

Homayoun Valafar, Committee Chair

John Rose, Committee Member

Jianjun Hu, Committee Member

Laszlo Szekely, Committee Member

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

Acknowledgments

I would like to gratefully and sincerely thank Dr. Jijun Tang for his guidance, understanding, patience, and most importantly, his friendship during my graduate studies at University of South Carolina. His mentorship was paramount in providing a well rounded experience consistent my long-term career goals. For everything you have done for me, Dr. Tang, I thank you. I would also like to thank all of the members of the our research group, giving me important suggestions and great environment to enjoy my work.

I would also like to thank Dr. Yu Lin for his assistance and guidance in research directions and endless help when I encounter difficulties. I am very grateful for the friendship and cooperation with him.

I would like to thank all my committee members: Dr Valafar, Dr Hu, Dr Rose and Dr Szekely. Thanks for their kind suggestions in my proposal and help in my research. I also want to thank the colleagues at Department of Computer Science and Engineering in University of South Carolina for the kind helps.

Finally, and most importantly, I would like to thank my parents for their faith in me and allowing me to be as ambitious as I wanted. It was under their watchful eye that I gained so much drive and an ability to tackle challenges head on.

Abstract

During evolutionary history, genomes evolve by DNA mutation, genome rearrangement, duplication and gene loss events. There has been endless effort to the phylogenetic and ancestral genome inference study. Due to the great development of various technology, the information about genomes is exponentially increasing, which make it possible figure the problem out. The problem has been shown so interesting that a great number of algorithms have been developed rigorously over the past decades in attempts to tackle these problems following different kind of principles. However, difficulties and limits in performance and capacity, and also low consistency largely prevent us from confidently statement that the problem is solved. To know the detailed evolutionary history, we need to infer the phylogeny of the evolutionary history (Big Phylogeny Problem) and also infer the internal nodes information (Small Phylogeny Problem). The work presented in this thesis focuses on assessing methods designed for attacking Small Phylogeny Problem and algorithms and models design for genome evolution history inference from FISH data for cancer data. During the recent decades, a number of evolutionary models and related algorithms have been designed to infer ancestral genome sequences or gene orders. Due to the difficulty of knowing the true scenario of the ancestral genomes, there must be some tools used to test the robustness of the adjacencies found by various methods. When it comes to methods for Big Phylogeny Problem, to test the confidence rate of the inferred branches, previous work has tested bootstrapping, jackknifing, and isolating and found them good resampling tools to corresponding phylogenetic inference methods. However, till now there is still no system work done to try and tackle this problem for small phylogeny. We tested the earlier resampling schemes and a new method inversion on different ancestral genome reconstruction methods and showed different resampling methods are appropriate for their corresponding methods.

Cancer is famous for its heterogeneity, which is developed by an evolutionary process driven by mutations in tumor cells. Rapid, simultaneous linear and branching evolution has been observed and analyzed by earlier research. Such process can be modeled by a phylogenetic tree using different methods. Previous phylogenetic research used various kinds of dataset, such as FISH data, genome sequence, and gene order. FISH data is quite clean for the reason that it comes form single cells and shown to be enough to infer evolutionary process for cancer development. RSMT was shown to be a good model for phylogenetic analysis by using FISH cell count pattern data, but it need efficient heuristics because it is a NP-hard problem. To attack this problem, we proposed an iterative approach to approximate solutions to the steiner tree in the small phylogeny tree. It is shown to give better results comparing to earlier method on both real and simulation data.

In this thesis, we continued the investigation on designing new method to better approximate evolutionary process of tumor and applying our method to other kinds of data such as information using high-throughput technology. Our thesis work can be divided into two parts. First, we designed new algorithms which can give the same parsimony tree as exact method in most situation and modified it to be a general phylogeny building tool. Second, we applied our methods to different kinds data such as copy number variation information inferred form next generation sequencing technology and predict key changes during evolution.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Ancestral Genome Inference and methods assessment

During evolutionary history, genomes get changed not only by events like DNA mutation, but also by other level of events such as genome rearrangement, duplication and gene loss. Since rearrangement events are rare and far less common than simple nucleotide mutations, they can be used to reconstruct evolutionary history extends far back to the evolutionary history.

The success of phylogenetic reconstruction demonstrates the power of revealing the evolutionary relation of a group of organisms by computational means. As phylogeny often takes the form of rooted binary tree, each internal node of the tree can be naturally regarded as the common ancestor of the living organisms descended from it. Fig 1.1 shows an simple phylogenetic tree inferred from 19 species of True flies (Diptera) family [159]. The predication of ancestral orders of these ancestors has been investigated in-depth and several methods have been developed for the task.

The small phylogeny problem (**SPP**) defines when the phylogenetic tree is given and the goal is only to reconstruct the ancestral genomes, while the **BPP** searches the most appropriate tree along with a set of ancestral genomes. In my thesis, we are interested in the small phylogeny problem. Majority of current methods solving SPP adopt either adjacency-based approach in which rearrangements are only implicitly considered or rearrangement-based approach that involves computing numerous instances of median problems. In particular, adjacency-based methods mainly focus

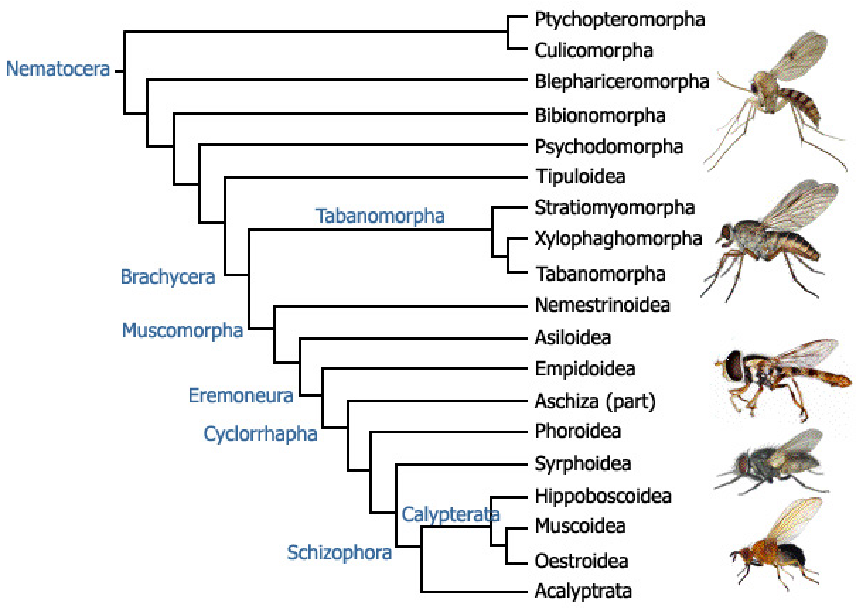


Figure 1.1 Phylogenetic inferred for True Flies

.

on the analysis of independent gene adjacencies extracted from the gene orderings, trying to calculate or estimate the score for each gene adjacency to present in an ancestral genome. The final step is to construct a proper gene/edge graph and rejoin discrete gene adjacencies back into contiguous ancestral regions (CAR) by optimizing the total score of the path. From another point of view, some methods employ a non-parametric way (parsimony) and suggest to use least number of changes to explain observed data; while the rests estimate the parameters in a parametric way and use posterior probabilities or likelihood to score the gene adjacencies. Table 1.1 generalizes on the difference between a number of methods for solving SPP with gene order data.

In the context of rearrangement-based parsimonious methods, the median problem can be formalized as follows: give a set of m genomes with permutations $\{x_i\}_{1\leq i\leq m}$ and a distance measurement d, find another permutation x_t such that the median score defined as $\sum_{i=1}^{m} d(x_i, x_t)$ is minimized. **GRAPPA** and **MGR** are two similar methods that implemented a selection of median solvers for phylogeny and ancestral gene-order inference. However solving even the simplest case of median problem when m equals to three is NP-hard for most distance measurements [15, 19, 139]. In detail, given a tree topology, these methods iteratively assigns median genomes to ancestral nodes in the tree until convergence. Then the set of gene-order assignments that minimizes the tree score are reported as the resulting ancestral genomes. Since the scoring procedure of GRAPPA involves solving numerous instances of median problems, a fast median solver is crucial. Exact solutions to the problem of finding a median of three genomes can be obtained for the inversion, breakpoint and DCJ distances [20, 122, 154]. Among all the median solvers, the best one is the DCJ median solver proposed by Xu and Sankoff (**ASMedian** [154]) based on the concept of adequate subgraph. Adequate subgraphs allow decompositions of a multiple breakpoint graph into smaller and easier graphs. Though the ASMedian solver could remarkably scale down the computational expenses of median searching, it yet runs very slow when the genomes are distant. On the other hand, **GASTS** and **SCJ** are two heuristic methods that scale up their capacities to handle high-resolution vertebrate genomes. GASTS is based on a fast and accurate heuristic for the inversion median [108] in which only a few of the simplest decompositions of adequate graphs are be solved; it provides a fast and robust scoring method for a fixed tree and demonstrated very high accuracy in the simulation experiments compared to MGR. Single-cut-or-join (SCJ) defines a breakpoint-like operation proposed under which the median problem and SPP can be solved in polynomial time. It utilized the Fitch's algorithm to solve the SPP in which each adjacency is viewed as a binary character of state either presence or absence and ultimately all adjacencies are determined in ancestral genomes. This is the only known evolutionary distance for which the SPP has a polynomial time solution.

Adjacency-base parsimonious method was formally introduced in **InferCARs** by Ma in 2006. It identifies a most-parsimonious scenario for the history of each individual adjacency, introduces weights to the graph edges to model the reliability of each adjacency and propose a greedy heuristic approach to look for vertex-disjoint paths in the graph, which represent the contiguous ancestral regions. Later Ma introduced an extended work **InferCARsPro** in the probabilistic framework for reconstructing ancestral order. The essential part of this method is to predict the posterior probability of an adjacency occurring in the ancestor based on an extended Jukes-Cantor model for breakpoints. However, neither of Ma's methods is able to hand the scenario where gene contents between input genomes are difference and Ma's methods can hardly guarantee complete assembly and hence often return a large number of CARs. Also Ma's methods require users to input a tree with branch lengths attached instead of estimating the branching lengths itself. To alleviate the problems, GapAdj is proposed to handle unequal gene contents and use TSP solver to replace the greedy heuristic to assemble gene adjacencies into completely assembled genomes at the sacrifices of accuracy. The core of GapAdj is the relaxation of direct adjacencies and allows pair of genes separated by up to a give number of genes. In this way, based on the assumption that small and local evolutionary events are more frequent than large and far-reaching operations, GapAdj expects to reconnect CARs by considering gapped adjacencies and gives good assembly of gene adjacencies. GapAdj can also treat datasets of unequal gene contents by first inferring the ancestral gene content through a natural process proposed in [56]. **PMAG** [69, 70] is an adjacency-based probabilistic framework of inferring the conditional probability of observing a gene adjacency in the target genome using the Bayes' theorem. Moreover they enhanced the general framework with a transition model and a re-root procedure. PMAG is shown to be significantly faster in running time and have a lower error rate than earlier method.

Since it is almost impossible to get the true evolutionary history when it comes to real data, to test the robustness of existing ancestral reconstruction method by

Table 1.1 Classification of current methods for solving gene order small phylogeny problem (SPP)

	Parsimonious	Probabilistic
Adjacency-based	InferCARs [83], GapAdj [47]	InferCarsPro [82], PMAG+ [70]
Rearrangement-based	GRAPPA [91], MGR [1, 14], GASTS [153], SCJ [9]	N/A

comparing to true situation is almost impossible. Methods must be developed to assess the quality of inferred ancestral genomes through resampling strategies. For phylogeny reconstruction, it is now a common practice to provide robustness evaluation of constructed tree edges. If the dataset contains DNA (or protein) sequences, the standard method is bootstrapping [39] which relies on resampling of the input columns. For gene order data, as the genome is viewed as one character, such approach cannot be readily applied and jackknifing was the first approach used to assess gene order phylogeny, which resamples the genomes by removing 40% of the genes from the dataset [120]. As some rearrangement methods use gene adjacencies as characters, Lin *et al.* [80] showed that bootstrapping on adjacencies can also be applied to rearrangement data. Shi *et al.* [119] tried another resampling strategy (called isolating), which randomly picks some genes and places them into new chromosomes. This approach simulates a major data error in genome assembly where some genes are misplaced. As no gene is excluded from the computation, using isolating achieves better performance than using jackknifing.

However, to our knowledge, there is no method to evaluate the quality of inferred ancestral genomes, although many tools are capable now to handle real world genomes.

1.2 Cancer and cancer phylogeny analysis

Cancer is recognized to be an evolutionary process driven by mutations in tumor cells [149]. These evolutionary processes include single-nucleotide variations, insertions and deletions, copy-number aberrations, structural variations and gene fusions [46]. Many experiments reveal considerable intra–tumor and inter–tumor heterogeneity [132], attributed to these evolutionary processes. Clinical implications of this heterogeneity, for example in drug resistance and disease diagnosis, has been well studied [132, 57].

Rapid, simultaneous linear and branching evolution in multiple sub-clones of cancer cells can be modeled by a phylogenetic tree [158]. Inferring such phylogenies facilitates the study of cancer initiation, progression, treatment, and resistance [3]. They can help pinpoint important changes that lead to the recurrence of some genome aberrations [6]. Phylogenies also aid in identifying genes crucial for the evolution and hence may contribute to developing better cancer treatment [104, 86, 94, 25].

Mutation patterns in cancer are characterized by frequent and widespread gains and losses of genomic material which is markedly different from what is observed in species or population level evolution [3]. In particular, gene copy number changes affect a larger fraction of the genome in cancers than do any other type of somatic genetic alteration [121, 160]. During tumor development, the gene copy number can increase or decrease, due to failures in DNA repair mechanisms (e.g., translesion synthesis and non-homologous end joining) [34, 115, 131, 17, 29, 113, 23, 152]. Another characteristic feature of tumor evolution is the high genetic heterogeneity found. Previous phylogenetic models for cancer, such as [86, 102, 156, 144, 59, 50], either do not account for these unique characteristics of cancer evolution or are not scalable and hence not of practical use. Thus there is need for development of new phylogenetic models with scalable algorithms that can adequately model cancer evolution. A step towards a scalable model for inferring tumor phylogeny by copy number variation was taken by Chowdhury *et al.* [27, 26] using FISH data.

Cancer is recognized to be an evolutionary process driven by mutations in tumor cells [149]. These evolutionary processes include single-nucleotide variations, insertions and deletions, copy-number aberrations, structural variations and gene fusions [46]. Many experiments reveal considerable intra–tumor and inter–tumor heterogeneity [132], attributed to these evolutionary processes. Clinical implications of this heterogeneity, for example in drug resistance and disease diagnosis, has been well studied [132, 57]. FISH data can be got from FISH technology. Each cell has a non-negative integer for each gene probe, the number of digits is the gene probes detected. For a general sample, FISH data of a system is a 2D-array, row number means how many patterns are found in it while column number is how many features are used.

Rapid, simultaneous linear and branching evolution in multiple sub-clones of cancer cells can be modeled by a phylogenetic tree [158]. Inferring such phylogenies facilitates the study of cancer initiation, progression, treatment, and resistance [3]. They can help pinpoint important changes that lead to the recurrence of some genome aberrations [6]. Phylogenies also aid in identifying genes crucial for the evolution and hence may contribute to developing better cancer treatment [104, 86, 94, 25]. Mutation patterns in cancer are characterized by frequent and widespread gains and losses of genomic material which is markedly different from what is observed in species or population level evolution [3].

In particular, gene copy number changes affect a larger fraction of the genome in cancers than do any other type of somatic genetic alteration [121, 160]. During tumor development, the gene copy number can increase or decrease, due to failures in DNA repair mechanisms (e.g., translesion synthesis and non-homologous end joining) [34, 115, 131, 17, 29, 113, 23, 152]. Another characteristic feature of tumor evolution is the high genetic heterogeneity found, the evolutionary process of current situation can be explained by phylogenetic analysis. A phylogeny is a term that represents the reconstructed evolutionary relationship of a set of organisms or cells in the form of a binary tree in which the given set of organisms are descendants placed at the leaves and internal nodes stand for extinct ancestors connected by the edges. Previous phylogenetic models for cancer, such as [86, 102, 156, 144, 59, 50], either do not account for these unique characteristics of cancer evolution or are not scalable and hence not of practical use. Thus there is need for development of new phylogenetic models with scalable algorithms that can adequately model cancer evolution.

Rectilinear steiner minimum tree

Given a weighted connected or undirected graph, a spanning tree of it is a subgraph that connects all the vertices together. A single graph can have many different spanning trees. The tree weight is the sum of the weights of the edges in that spanning tree. A minimum spanning tree (MST) or minimum weight spanning tree is then a spanning tree with weight less than or equal to the weight of every other spanning tree. There can be multiple minimum spanning trees in a graph. Fig.1.2 gave a good example of minimum spanning tree. The RSMT problem is a graph similar to minimum spanning tree and can be defined as follows. The RSMT problem is NP-complete [48] and is defined as follows.

Definition: RSMT(n, d)

Input: a weighted connected or undirected graph, nodes are values of different features.

Output: A minimum weight tree (or L_1 distance) including all the observed n nodes and, as needed, unobserved steiner nodes along with new values for each feature.

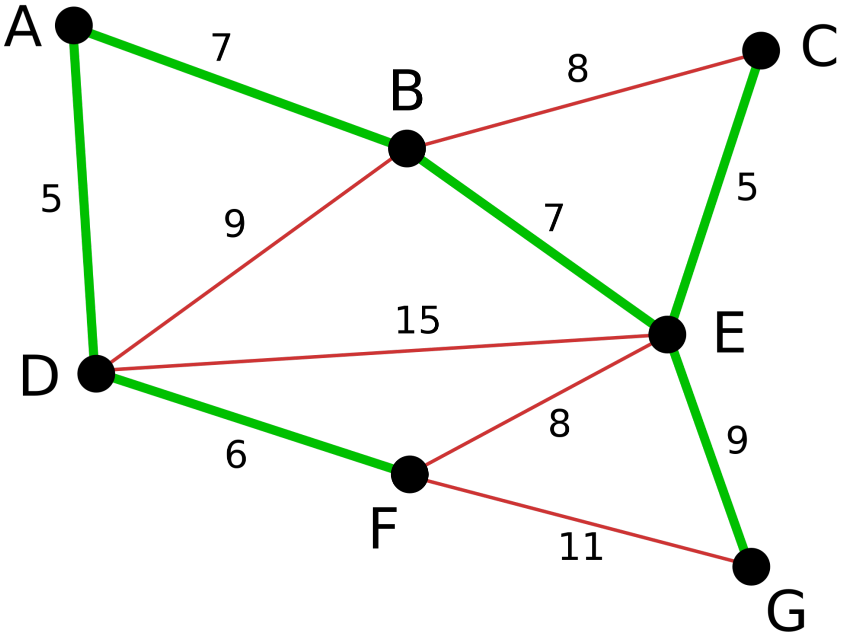


Figure 1.2 A simple example of minimum spanning tree for weighted graph(green color).

Chapter 2

Assessing Ancestral Genome Reconstruction Methods by Resampling

2.1 Background

During evolutionary history, genomes get changed not only by events like DNA mutation, but also by other level of events such as genome rearrangement, duplication and gene loss. Since rearrangement events are rare and far less common than simple nucleotide mutations, they can be used to reconstruct evolutionary history extends far back to the evolutionary history.

Handling rearrangement events is algorithmically very difficult: it took almost a decade to find the first polynomial algorithm to compute the inversion distance [63], while distances for sequence data are simple to define and have been thoroughly studied. Current algorithm development is focused on biological events with better methods for distance and median computations, using a unifying framework of double-cut-and-join (DCJ) [157]. On the other hand, several method using the idea of encoding the genome structure into binary sequences [81] and maximum-likelihood approaches have been developed and shown that they have reached performance that is comparable to the best sequence-based methods.

Most of existing genome analysis methods can not only be used to reconstruct phylogenies, but also be used to obtain an inference of ancestral genomes, which has a wide range of applications, such as to decide the evolution of genome structures [31], to infer the genome rearrangement rate and mechanism [143], to infer the orthology assignment [133], and to estimate gene orders of incompletely sampled genomes [71].

Since it is almost impossible to get the true evolutionary history, we must develop methods to assess the quality of inferred phylogenies and ancestral genomes, through resampling strategies. For phylogeny reconstruction, it is now a common practice to provide robustness evaluation of constructed tree edges. If the dataset contains DNA (or protein) sequences, the standard method is **bootstrapping** [39] which relies on resampling of the input columns. For gene order data, as the genome is viewed as one character, such approach cannot be readily applied and **jackknifing** was the first approach used to assess gene order phylogeny, which resamples the genomes by removing 40% of the genes from the dataset [120]. As some rearrangement methods use gene adjacencies as characters, Lin *et al.* [80] showed that bootstrapping on adjacencies can also be applied to rearrangement data. Shi *et al.*[119] tried another resampling strategy (called **isolating**), which randomly picks some genes and places them into new chromosomes. This approach simulates a major data error in genome assembly where some genes are misplaced. As no gene is excluded from the computation, using isolating achieves better performance than using jackknifing.

However, to our knowledge, there is no method to evaluate the quality of inferred ancestral genomes, although many tools are capable now to handle real genomes. In our work, we use simulations to test the performance of existing methods under various resampling strategies. From our experiments, we find that resampling is a valid strategy to assess the quality of ancestral genomes and worst adjacencies can be removed.

Genome Rearrangements

Given a set of n genes $\{1, 2, \cdots, n\}$, a genome can be represented by an *ordering* of these genes. To indicate the strandedness of genes, each gene is assigned with an orientation that is either positive, written i, or negative, written $-i$. Two genes

i and j are said to be *adjacent* in genome G if i is immediately followed by j, or, equivalently, $-j$ is immediately followed by $-i$.

Denote the head of a gene i by i^h and its tail by i^t. We refer $+i$ as an indication of direction from head to tail ($i^h \rightarrow i^t$) and otherwise $-i$ as ($i^t \rightarrow i^h$). There are a total of four scenarios for two consecutive genes a and b in forming an *adjacency*: $\{a^t, b^t\}$, $\{a^h, b^t\}$, $\{a^t, b^h\}$, and $\{a^h, b^h\}$. If gene c is at the first or last place of a linear chromosome, then we have a corresponding singleton set, $\{c^t\}$ or $\{c^h\}$, called a *telomere*.

Let G be a genome with signed ordering $\{g_1, g_2, \cdots, g_n\}$, an *inversion* (also called *reversal*) between indexes i and j ($i \leq j$) of produces a new genome with linear ordering

$$g_1, g_2, \cdots, g_{i-1}, -g_j, -g_{j-1}, \cdots, -g_i, g_{j+1}, \cdots, g_n$$

. There are additional operations for multi-chromosomal genomes, such as *translocation* (one end segment in one chromosome is exchanged with one end segment in the other chromosome), *fission* (one chromosome splits and becomes two), and *fusion* (two chromosomes combine to become one).

Yancopoulos *et al.* [157] proposed a universal double-cut-and-join (DCJ) operation that can be used to represent inversions, translocations, fissions, and fusions. Researchers everywhere have adopted the DCJ model in their work because of its mathematical simplicity and because of its observed robustness in practice.

A genome can also be expressed as a multiset of adjacencies and telomeres. By using 1 (0) to indicate presence (absence) of an adjacency, we can encode and transfer genomes into binary sequences (Table 2.1 shows an example)

Table 2.1 Example of encoding two genomes $G_1 : (1, 2, -3)$ and $G_2 : (3, -2, 1)$ into binary sequences.

	$\{1^h\}$	$\{1^t, 2^h\}$	$\{2^t, 3^t\}$	$\{3^h\}$	$\{2^h, 1^h\}$	$\{1^t\}$
G_1	1	1	1	1	0	0
G_2	0	0	1	1	1	1

Ancestral genome reconstruction

Scientists have been working on reconstructing extinct ancestral genomes for quite a long time. For example, some such work has been used to predict protein functional shift and positive selections[92]. Ancestral genome reconstruction normally assumes a given phylogeny tree and detailed information for leaf genomes, with the quest to infer genomes on the internal nodes (ancestors). There are now two different groups of methods: event-based and adjacency-based methods.

Event-based methods

Event-based methods are based on maximum parsimony and find the best tree (and its associated ancestral genomes) that minimizes the total number of events. The simplest tree has only three leaves and one internal node, thus form the median problem which is defined as follows: given three genomes, find a single genome that minimizes the sum of the pairwise distances between itself and each of the three given genomes. However solving even this simplest case is NP-hard for most distance measurements [42], thus many current event-based methods use an iterative improvement approach, based on the computation of medians defined on internal nodes.

Exact solutions to the median problem are available for inversion, breakpoint and DCJ distances [109, 155]. Among all existing median solvers, the best is the DCJ median solver proposed by Xu and Sankoff (`ASMedian` [155]) based on the concept of adequate subgraph. The best method for hanlding more than three genomes is `GASTS` [153], which is a heuristic based on `ASMedian` and can quickly score a fixed phylogenetic tree, as demonstrated on a set of vertebrate genome with over 2,000 genes.

Adjacency-based methods

The other type of model that handles gene adjacencies relies (Table 2.1) on two separate steps. First, the weight or probability that a gene adjacency is present in a genome is computed independently. Then those gene adjacencies are assembled into a valid ancestral genome. `InferCAR` [83] and its probabilistic version `InferCARsPro` are the pioneering methods based on this model. In this model, all combinations of gene adjacencies are considered, and their probabilities are computed by a variant of the Fitch parsimony algorithm. Finally a greedy heuristic is used for to assemble the genes into a valid genome.

Later by relaxing the constraint of gene adjacency to gapped adjacency, `GapAdj` is proposed using a rigorous score for each potential ancestral adjacency (a, b), reflecting the maximum number of times a and b can be adjacent for any setting of ancestral genomes, as well as an algorithm to generate more reliable amount of chromosomes.

PMAG is till now the best probabilistic method introduced by Hu *et al.* [68, 70] which has no need of branching length input and is parameter free. It first transfers rearrangement data into adjacency data as leaf genome adjacencies; then these leaf adjacencies are used to infer the probability of each adjacency in the ancestral genomes. Based on the probabilities of adjacencies, we can construct a Traveling Salesperson Problem to find a path with the minimum TSP score, which can then be transferred back into the ancestral gene order.

Simulation is widely used to assess the quality of various ancestral genome reconstruction methods, since in simulations true ancestral gene orders are known. However, when it comes to real biological data, the situation would be quite different due to the difficulty of knowing the true ancestors. Until now, only Ma *et al.*[83] has tried to measure the percentage of true adjacencies out of all the adjacencies in the constructed genomes produced by his own method. The question whether there is a universal resampling scheme good for all ancestral genome reconstruction methods, or different resampling schemes should be used for different methods is still not answered.

Several resampling approaches have been proposed to assess the tree edges generated by phylogenetic inference methods. If the method is sequence based, then classical bootstrapping is suggested. A new resampling scheme which combines jackknifing and bootstrapping is used by Lin[80] and it has a similar support value to classical bootstrapping on sequence data based methods, but can also be applied for rearrangement data based phylogeny inference methods. Shi *et al.* also tried isolating [119] and jackknifing[120] on rearrangement data and both are shown to work well on gene order based methods.

2.2 Methods and result

We borrow similar approaches to assess the quality of inferred ancestral gene orders. The following four resampling strategies are tried in our experiments:

Jackknifing is introduced by Quenouille *et al.*[107] and further improved by Tukey *et al.*[142]. For rearrangement data, a jackknifing event takes one gene out of the original gene order without putting it back. For this approach, a resampling rate of $k\%$ means that $r\%$ genes will be removed and thus only $(100 - r)\%$ of the original genes will participate in the computation. For example, for resampling rate of 40%, {1,2,3,4,5} might become {1,3,4} after jackknifing.

We can also try jackknifing on adjacency data, which means that only a portion of existing adjacencies will be left after jackknifing. For example, when $r = 33.3$ $r = 33.3$ jackknifing is applied on the adjacencies shown in Table 2.1 (i.e. 1/3 adjacencies will be removed), the new adjacency list may become the one shown in Table 2.2. Compared to Table reft1, the second and fifth columns are removed.

The bootstrapping operation was first tried by Soltis *et al.*[125], and later Holmes *et al.*[65] introduced the concept of bootstrapping. Given a genome with n genes,

Table 2.2 Example of jackknifing on adjacencies on Table 2.1, where the second and the fifth columns are removed.

	$\{1^h\}$	$\{2^t, 3^t\}$	$\{3^h\}$	$\{1^t\}$
G_1	1	1	1	0
G_2	0	1	1	1

$X = \{x_1, x_2, ..., x_n\}$, the data input after bootstrapping would become a fictional data set $Y = \{y_1, y_2, ..., y_n\}$ constructed by sampling with replacement from X, so each item in Y can be any item in X. It is possible for one gene in X to show up several times or not at all in Y. Bootstrapping can be applied to both rearrangement and adjacency data. However, only bootstrapping on the adjacency data is meaningful because bootstrapping on gene order will treat genes as isolated letters with no order information while the order, or adjacency information, should be kept. Using Table 2.1 as an example again, after bootstrapping, the table may become that shown in Table 2.3. In this example, the first column is duplicated, while the fifth column is removed, but the total number of column remains 6.

Isolating [119] was first used by Shi *et al.* for robustness testing on phylogeny inferring methods from gene order data. A gene is isolated by applying a double-cut event to the original genome which removes one gene, and make it a new chromosome. Compared to jackknifing, isolating retains the chosen genes thus introduces smaller disturbance. For this method, we define the resampling rate r as the percentage of genes being isolated.

We can also apply random evolutionary events on the input genomes to introduce disturbance of data, thus providing another resampling technique. As inversion is viewed as the dominant event, we can randomly choose a part of the genome of any possible length within a maximum length limit and reverses it. This approach keeps all of the gene information, so it will still have the same gene order content after resampling. However, there are some decisions to make: should we apply the same amount of events on every input genome? should we limit the length of an inversion?

To simplify the procedure, in our work, we use only one parameter, which is the resampling rate r defined as the number of inversion events applied divide by the number of genes. For every genome, we will apply the same amount of events, and each of these events can start and end at any genes. We call this method *Inversion* in later sections.

Experimental design

Our model tree simulation follows Lin's[79] birth-death model. The model trees are generated with the following parameters: the number of genomes (m) is 100, the number of genes (n) is 1000. We use two different tree diameters (d) of 500 and 2000. For each combination of parameters, 10 phylogenies are generated and for each phylogeny, the same root node is evolved by random double-cut-and-join (DCJ) events with respect to the tree diameter d.

Given a reconstruction method, if the true ancestors are known, to assess its quality, we can compare those adjacencies correctly recovered (true positive) from those missed (negative) and wrongly inferred (false positive). However, since we do not have any ancestor information in real data analysis, we have to utilize the above resampling methods using the following procedures:

1. For each dataset, we will generate new datasets by applying one of the above resampling methods;

2. We will then apply the same reconstruction method to obtain ancestors on the disturbed input;

3. The above two steps will be repeated k times and for a given ancestor.

After k results are obtained, we will examine each ancestor (i.e. each internal node of a given tree) in turn, and assess the quality of an adjacency by computing the times f that this adjacency appears in the k inferred gene orders; if $\frac{f}{k}$ is larger than a cut-off threshold, this adjacency is viewed as potentially good.

In the experiments, for each resampling method, we apply corresponding resampling events to each data set and produce k replicas where k = 100, from these replicas we use three different ancestral sequence reconstruction methods, PMAG, GapAdj and GASTS to get the gene orders of the ancestral genomes.

We use Receiver-Operator-Characteristic(ROC) curves to assess the quality of each resampling approaches. The frequency f of a gene adjacency in an ancestral genome G is regarded as its confidence value. For each confidence threshold t, if $\frac{f}{k}$ is greater than t, we retain the gene adjacency in G; otherwise, we discard that gene adjacency. The value of t ranges from 0.10 to 1.00 with an interval of 0.02.

Suppose the set of adjacencies left after checking with the threshold is A_t and the set of adjacencies in the true ancestral nodes is A_T. Then $specificity$ is treated as the proportion of the adjacencies in A_T and also in A_t ,as the expression is $|A_T \bigcap A_t|/|A_t|$, while *sensitivity* is the proportion of adjacencies in A_t and also in A_T, expression as $|A_T \bigcap A_t|/|T|$. The ROC map is drawn as the relationship of $1 - specificity$ and *sensitivity*.

We apply bootstrapping the same way as classical bootstrapping, which is only valid for the adjacency-based methods, such as PMAG and GapAdj. Jackknifing is also applied on the input data for PMAG both on rearrangement and adjacency data level and only on rearrangement data level for GASTS and GapAdj. Isolating and inversion are used on all the input data for the three methods. For the resampling rates, given n genes, the rates of jackknifing vary from 0.05n to 0.40n with an interval of 0.05n, the rates of isolating and inversion are both within the range 0.01n to 0.13n with an interval of 0.01n. Bootstrapping has only one requirement that the number of columns should be kept the same (Table 2.3).

Table 2.3 Example of bootstrapping on adjacencies on Table 2.1, where the first column is duplicated and the fifth columns is removed. The total number of columns remains to be 6.

	$\{1^h\}$	$\{1^h\}$	$\{1^t, 2^h\}$	$\{2^t, 3^t\}$	$\{3^h\}$	$\{1^t\}$
G_1	1	1	1	1	1	0
G_2	0	0	0	1	1	1

Experimental Results

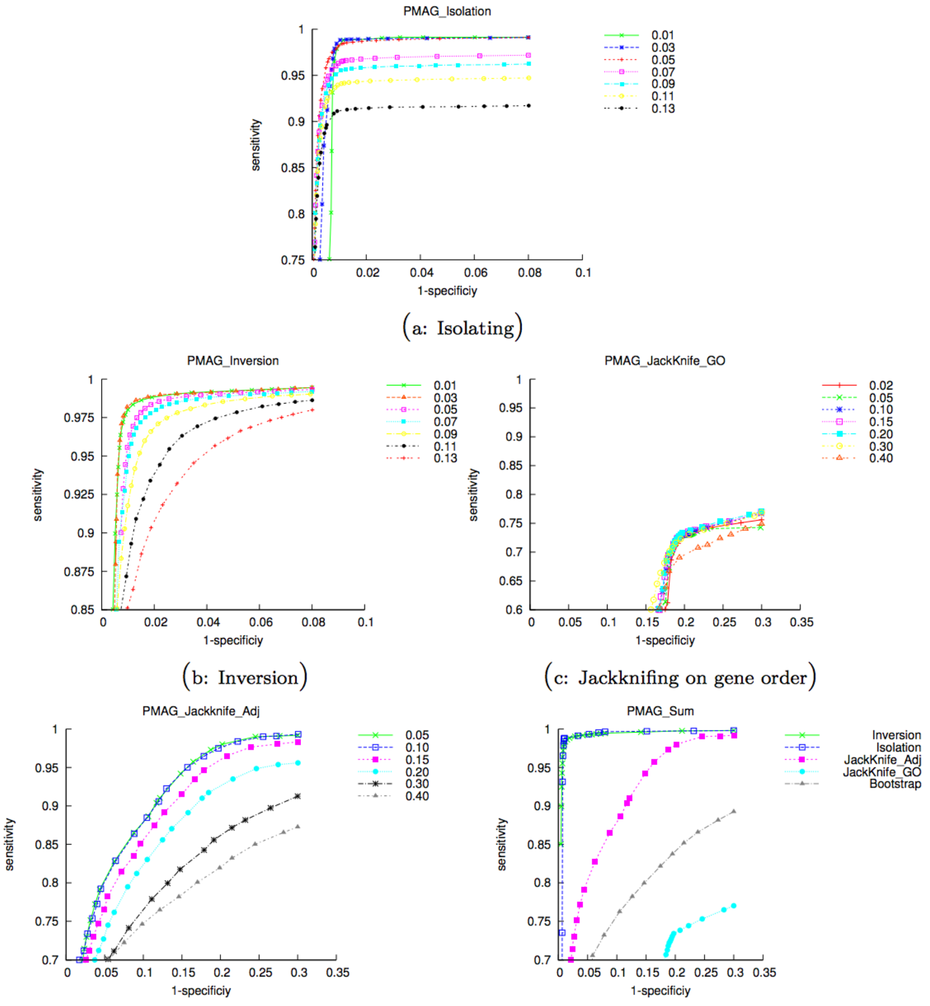
Figures 2.1 and 2.3 to 2.6 and **??** show the ROC curves for each reconstruction method, using various resampling methods. We first compare the performance of different resampling rates under the same resampling method, we then choose the rate with the largest area under its corresponding ROC curve, and summarize them to determine whether a resampling approach is valid for a given reconstruction method.

Figures 2.1 and 2.2 show the ROC curves for PMAG, with tree diameters (d) equal to 2000 and 500 events respectively. These two figures have quite similar results thus we only analyze the results for $d = 2000$ (Figure 2.1).

Fig.2.1(a) and 2.1(b) show the ROC curves of resampling using isolating and applying random inversion events. In these figures, with increasing resampling rates, the area under ROC curves increase first and then decrease; with resampling rates of 0.03, both methods have the largest areas under the ROC curves.

Fig.2.1(c) shows the ROC curves of jackknifing methods on gene orders. The result shows the ROC areas have no big difference when the jackknifing rates are between 0.02 0.30. Fig.2.1(d) shows the ROC curves of using jackknifing on adjacencies, where the ROC areas decreases with increasing jackknifing rates. These two figures suggest that both jackknifing methods have poor performance for PMAG.

We summarize the ROC curve results for PMAG in Fig.2.1(e), using the best rate for each resampling method. As shown in the figure, for PMAG, the resampling techniques of isolating and inversion are nearly the same and dominate the other two resampling schemes.



(a: Isolating)

(b: Inversion)

(c: Jackknifing on gene order)

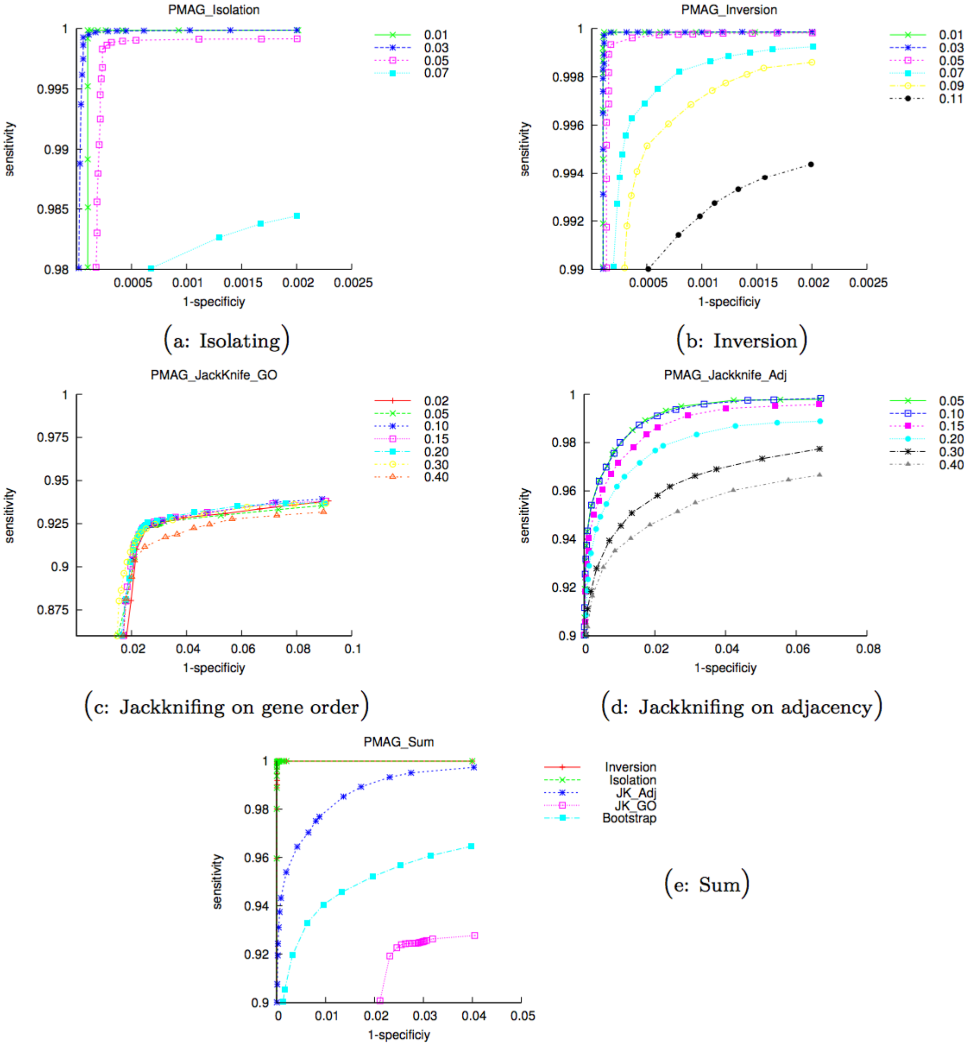
(d: Jackknifing on adjacency)

(e: Sum)

Figure 2.1 Result for PMAG (Diameter=2000). This figure shows the ROC curves using various resampling techniques for PMAG. (a) Isolating, (b) Applying random inversion events, (d) Jackknifing on gene orders, (d) Jackknifing on adjacencies. (e) Summary, the resampling rates with the largest ROC area for each resampling method are shown here. From the results, isolating and inversion is better than other resampling methods for PMAG.

Results for GapAdj

As the results for both tree diameters are similar, we only show those from $d = 2000$ in Figure 2.3. Compared to those shown in Fig.1, the performance of isolating and inversion approaches are quite similar, with the best resampling rates to be 0.03 again. However, jackknifing on input gene orders seems to have better performance



(a: Isolating)

(b: Inversion)

(c: Jackknifing on gene order)

(d: Jackknifing on adjacency)

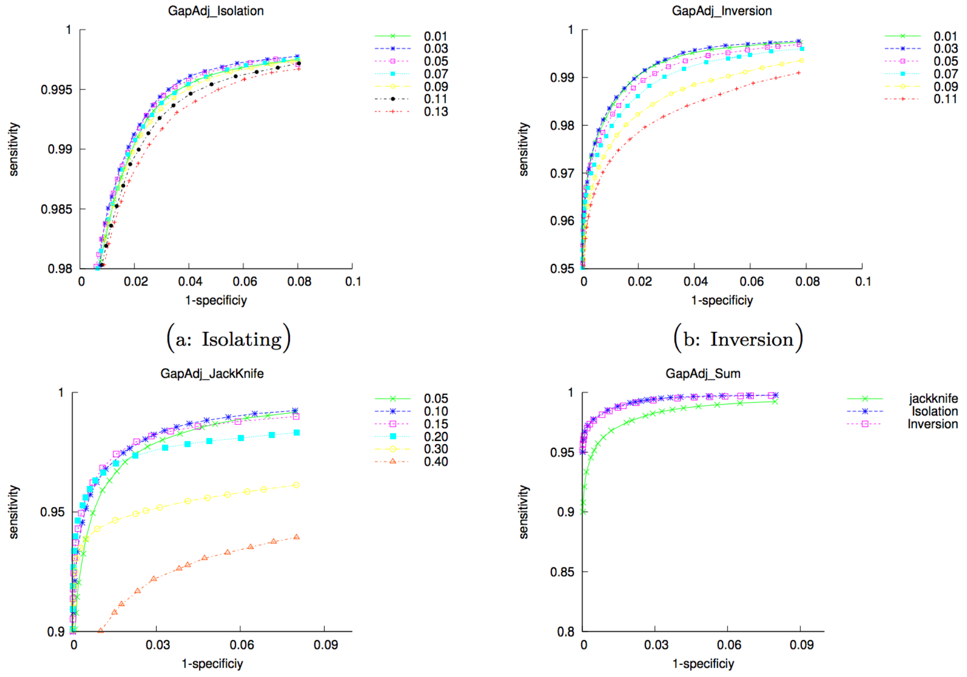
(e: Sum)

Figure 2.2 Result for PMAG (Diameter=500). This figure shows the ROC curves using various resampling techniques for PMAG. (a) Isolating, (b) Applying random inversion events, (d) Jackknifing on gene orders, (d) Jackknifing on adjacencies. (e) Summary, the resampling rates with the largest ROC area for each resampling method are shown here. From the results, isolating and inversion is better than other resampling methods for PMAG.

for GapAdj, and the best resampling rate is around 0.03, i.e. 3% genes can be removed with reasonable quality retained.

Fig. Figure 2.3(d) shows all the result of all resampling methods for GapAdj. The detailed index are: 3% isolating rate, 3% inversion rate, 10% jackknifing rate on gene order, The other two methods are not shown here because they are not applicable to

GapAdj. As shown in the figure, isolating and inversion are again nearly the same and obviously dominate jackknifing.



(a: Isolating) (b: Inversion)

(c: Jackknifing on gene order) (d: Sum)

Figure 2.3 Result for GapAdj (Diameter=2000). This figure shows the ROC curves using various resampling techniques for GapAdj. (a) Isolating, (b) Applying random inversion events, (d) Jackknifing on gene orders. (d) The sum of the results, the resampling rates with largest ROC area for each resampling method are shown here. From the result, isolating and inversion is better than jackknifing for GapAdj ancestral genome reconstruction method.

Results for GASTS

As the results for both tree diameters are similar, we only show those from $d = 2000$ in Fig. Figure 2.5. Compared to those shown in Figure 2.5, the performance of isolating and inversion is again quite similar, with the best resampling rates to be 0.03 for both methods. The performance of jackknifing on gene orders is greatly improved and as shown in figure Figure 2.5(d), isolating, inversion and jackknifing have very similar quality.

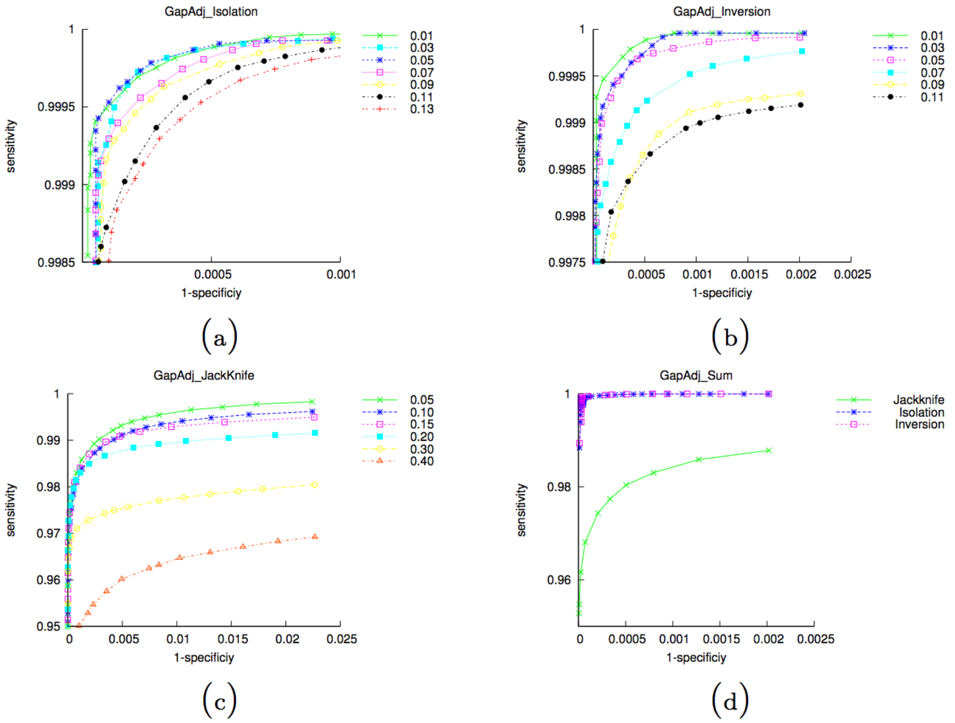
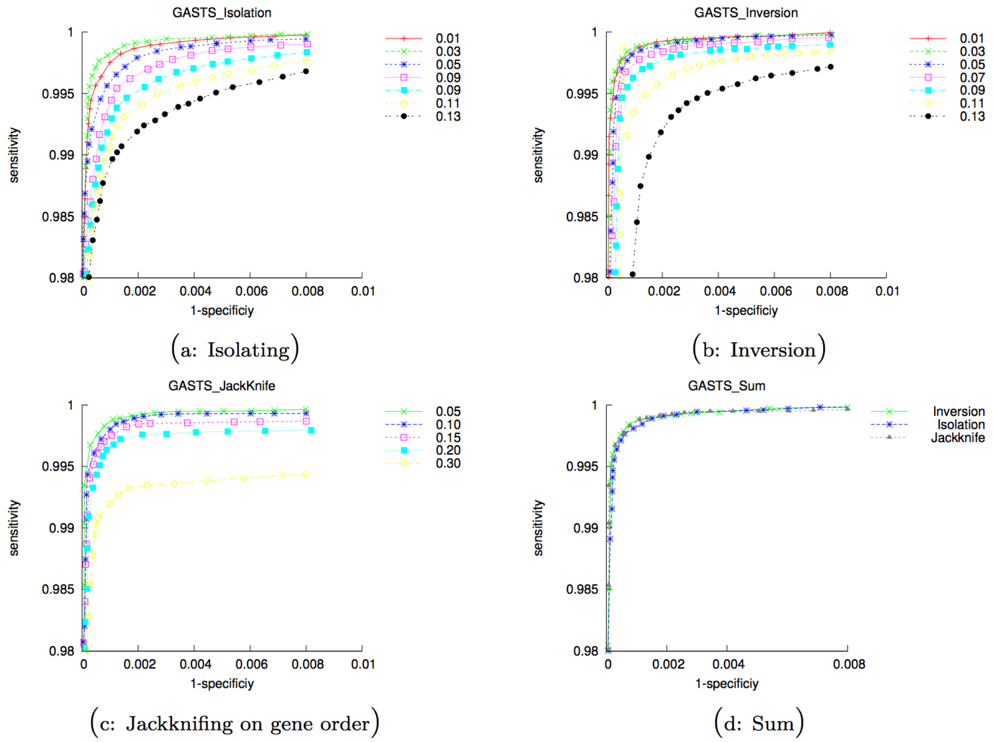


Figure 2.4 Result for GapAdj (Diameter=500). This figure shows the ROC curves of various resampling techniques for GapAdj. (a) Isolating, (b) Applying random inversion events, (c) Jackknifing on gene orders, (d) Summary, the resampling rates with the largest ROC area for each resampling method are shown here. From the results, isolating and inversion is better than other resampling methods for PMAG.

FP and FN results

The ROC figures do not specify which cut-off threshold should be used, i.e. above which threshold an adjacency can be determined to be “true”. We pick the best resampling rates for different reconstruction and resampling methods from Figures 2.1 to 2.3 and 2.5, and show the FP and FN rates in Table 2.4. Suppose D is a set of gene adjacencies encoded from an ancestral genome and D’ represents the corresponding genome inferred from the leaf data. A gene adjacency in D is missing in D’ if D’ does not contain such an adjacency defining the same connection; such a gene adjacency is called a false negative (FN). The false negative rate (FNR) measures the proportion of false negative gene adjacencies with respect to the total number of gene adjacencies



(a: Isolating)

(b: Inversion)

(c: Jackknifing on gene order)

(d: Sum)

Figure 2.5 Results for GASTS (Diameter=2000). This figure shows the ROC curves using various resampling techniques for GapAdj. (a) Isolating, (b) Applying random inversion events, (c) Jackknifing on gene orders. (d) The sum of the results, the resampling rates with largest ROC area for each resampling method are shown here. From the result, isolating, inversion and Jackknifing have similar effect on GASTS ancestral genome reconstruction method.

in D. The false positive (FP) and false positive rate (FPR) are defined similarly, by swapping D and D'. The FP and FN values are used to identify the best threshold for different methods. In the table, the original FP and FN without resampling; the FP and FN value under different threshold of 70, 75, 80, 85 are shown.

In phylogenetic reconstructing from DNA or protein sequence data, the threshold value of 80% is widely used. From Table 2.4, the threshold value of 75% has the best balance of FP and FN, thus can be used to filter bad adjacencies.

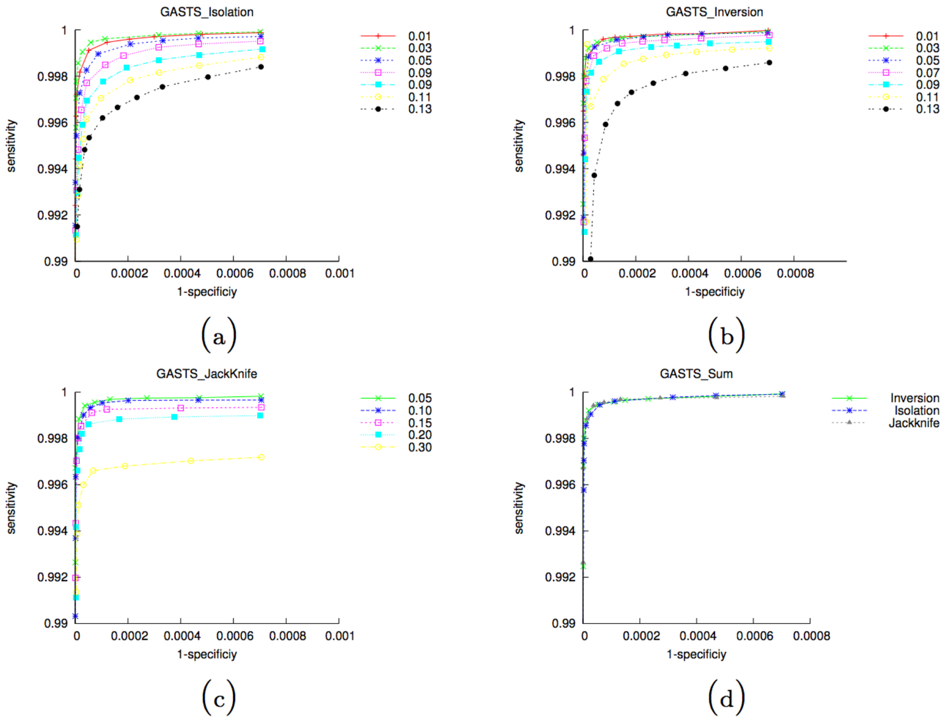


Figure 2.6 Result for GASTS (Diameter=500). This figure shows the ROC curves of various resampling techniques for GASTS. (a) Isolating, (b) Applying random inversion events, (d) Jackknifing on gene orders, (e) Summary, the resampling rates with the largest ROC area for each resampling method are shown here. From the result, isolating, inversion and Jackknifing have similar effect on GASTS ancestral genome reconstruction method.

2.3 Conclusions

In our work, we conduct extensive experiments to evaluate various resampling techniques in assessing the quality of inferred ancestral genomes. From the experimental results, isolating and applying random inversions are shown to produce a better ROC curve for PMAG and GapAdj compared to other resampling schemes. Thus both isolating and inversion are suggested for adjacency-based ancestral reconstruction method. For event-based parsimony methods (GASTS), we find that jackknifing, isolating, and inversion produce similar ROC curve if they have similar resampling rates. The reason for such results is because both PMAG and GapAdj only keep part of leaf adjacency information after resampling to construct the ancestral genome.

Table 2.4 FP/FN rates of PMAG, GASTS and GapAdj, under various threshold values and resampling methods.

Thresholds	0%		70%		75%		80%		85%	
Diameter-2000	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN
PMAG-Inversion	1.0	1.2	1.0	2.1	1.0	2.1	0.9	2.2	0.9	2.2
PMAG-Isolating	1.0	1.2	1.0	2.1	1.0	2.1	1.0	2.2	0.9	2.2
GapAdj-Inversion	10.7	10.7	0.6	14.2	0.5	17.9	0.5	23.4	0.2	31.2
GapAdj-Isolating	10.7	10.7	0.8	12.4	0.7	15.3	0.7	19.4	0.4	25.2
GASTS-Inversion	0.19	0.19	0.06	0.22	0.05	0.24	0.05	0.26	0.05	0.28
GASTS-Isolating	0.19	0.19	0.05	0.24	0.05	0.27	0.04	0.31	0.03	0.36
GASTS-Jackknifing	0.19	0.19	0.03	0.33	0.02	0.39	0.01	0.53	0.01	3.78

Both jackknifing and bootstrapping will remove all the information about certain adjacencies, while isolating and inversion still keep part of the adjacency information of the leaf genomes, which is the reason isolating and inversion can outperform other strategies for PMAG and GapAdj. Jackknifing on gene orders, compared to jackknifing on adjacencies, has more effect on adjacencies because it not only removes part of original adjacencies, but also generates a lot of incorrect adjacencies, thus it has the worst performance and should not be used. For GapAdj, it can recover some of this impact by allowing certain distance for adjacencies which could be the reason that the difference is smaller compared to PMAG. The FP and FN data show that the best threshold for different resampling schemes is around 75%.

Chapter 3

An iterative approach for phylogenetic analysis of tumor progression using cancer data

3.1 Introduction

Traditionally, cancer research focused mainly on the identification of oncogenes and tumor suppressor genes. However, in the last decades it become more and more well-known that disruption of normal differentiation is an essential contributor of tumorigenesis. Not total cellular maturation is now treated as a symbol of human cancers [62], and the level of differentiation of a tumor plays a key role for diagnosis, prognosis, and treatment. The identification of phenotypic markers and gene expression profiles, for instances, correlated with maturation has enabled researchers to link the expansion of malignant cells to certain stages of hematopoietic differentiation [**bennett1976thesiss**].

Nevertheless, success to define cancer stages have proven difficult due to an incomplete understanding of differentiation pathways from normal cells into mesenchymal and epithelial tissues. Another difficulty to solve this problem is to collect the data for different tumor stages to identify the maturation stages they correspond [87]. In a true tumor development, only a part of cells undergoes differentiation and current methods do not allow isolation of those cells during the differentiation process from the bulk of unchanged cells.

Fluorescence in situ hybridization (FISH) give us a chance to probe copy numbers of small numbers of gene markers in a group of single cells per study. Previous studies have shown that a single tumor can have hundreds of genetically distinct cell types [124, 137, 64]. Development in single-cell sequencing research which offers a more complete picture of the genome than FISH but for fewer cells supported this conclusion of great inter-cellular heterogeneity [147]. The extensive heterogeneity of tumor suggests that tumor phylogeny approaches and models need to scale to hundreds or thousands of taxa per tumor to produce reliable tools for evolutionary analysis of single tumors. Before Chowdhury *et al.*'s studies, phylogenetic model inference on single cell data has been achieved only for specialized datasets involving just limited probes per cell. [102, 86]

Recently, Chowdhury *et al.* successfully modeled the progression of tumor progression using FISH copy number to the Rectilinear Steiner Minimum Tree (RSMT) problem, and proposed both exact and heuristic algorithms to reconstruct phylogenetic trees modeling the development of cancer cell patterns [27, 26]. The RSMT problem for phylogenetic analysis is explained more in the following definition.

Problem 3.1.1. The RSMT problem for gene copy number:

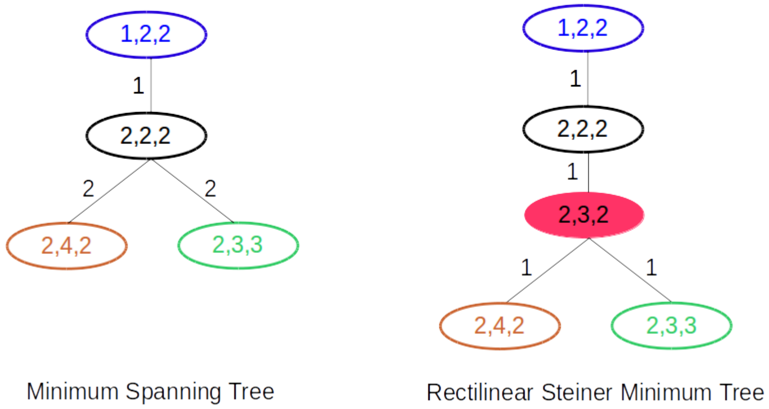
Definition: RSMT(n, d)

Input: FISH data of n cell count patterns on d gene probes for a given patient

Output: A minimum weight tree with the rectilinear metric (or L_1 distance) including all the observed n cell count patterns and, as needed, unobserved Steiner nodes along with their cell count patterns for d probes, Steiner nodes here is used to represent missing node during gene copy number change process.

The single ancestor could be, for example, cell count pattern with a copy number count of 2 for each gene probe (a healthy diploid cell) [27, 26]. The RSMT problem is NP-complete [48]. Note that if all possible cell count patterns in cancer cells are present as the input, then the RSMT is simply the minimum spanning tree, since no additional Steiner nodes are needed. Since both the minimum spanning tree and the minimum spanning network (as the union of all minimum spanning trees) can be constructed efficiently, previous heuristics have approximated RSMT by adding additional Steiner nodes to the minimum spanning network [27, 26]. For example, Fig.3.1 shows an instance of 4 cell count patterns on 3 genes, and the RSMT can be obtained by adding a Steiner node to the minimum spanning tree. However, both algorithms do not scale well with the number of gene probes, making them impractical to handle dozens of gene probes — a typical number of genes in one complicated signal pathway. their above heuristic is likely to be trapped in a local optimum if there are multiple possible Steiner nodes that can be introduced, since the order in which the Steiner nodes are added may affect the resulting tree weight. A similar model based on the Steiner Minimum Tree has also been introduced to study the “small phylogeny" problem at both the sequence level [114] and the gene order level [11]. A special case of the “small phylogeny" problem is called the median problem — given three sequences (or permutations), find the configuration of a median genome

	Gene A Copy #	Gene B Copy #	Gene C Copy #
Count Pattern 1	2	2	2
Count Pattern 2	1	2	2
Count Pattern 3	2	4	2
Count Pattern 4	2	3	3



Minimum Spanning Tree

Rectilinear Steiner Minimum Tree

Figure 3.1 (Top) The input data of 4 cell count patterns on 3 genes. (Bottom left) The minimum spanning tree has weight 5. (Bottom right) the RSMT has weight 4. The Steiner node in RSMT is colored in red.

to minimize the sum of the pairwise distances between the median and three input ones [41]. Sankoff *et al.* proposed iterative approaches to approximate solutions to the Steiner tree, which iteratively solve the median problem for one internal vertex at a time, and to make improvement until a local optimum is found [114, 11]. We propose similar heuristics to approximate solutions to the RSMT problem through iteratively optimizing the median version of RSMT problem. Moreover, the new iterative approach incorporates two key observations, the generation of median instances and the order of iterative optimizations of median instances, which generalizes the previous approaches of using median solvers to approximate solutions to the Steiner tree [114, 11] and takes into consideration specific characterization and challenging in the RSMT problem.

We make some contribution to address the need for algorithms for evolutionary model inference for tumor phylogenetics capable of handling large single-cell datasets, with specific application to FISH copy number data. We show here a new heuristic to attack the RSMT problem, which is inspired by iterative approaches to approximate solutions to the Steiner tree in the “small phylogeny" problem [114, 11]. Experimental results from both simulated and real tumor data show that our approach outperforms the previous heuristic algorithm in approximating better solutions for the RSMT problem.

In Chapter 2, 3, 4 and 5, we used the datasets for testing. We used both the real cervical cancer[148] and breast cancer[64] data samples and simulation samples generated through the same process described in the the supplemental material of the previous study by Chowdhury *et al.* [27]. The cervical cancer data contain four gene probes LAMP3[75], PROX1[150], PRKAA1[72] and CCND1[45], and the breast cancer data contain eight gene probes COX-2[67], MYC[151], CCND1[45], HER-2[138], ZNF217[98] ,DBC2[60], CDH1[10] and p53[145]. All those genes are chosen because they are considered as important factors for cancer growth inhibition or promotion. The cervical cancer data is from 16 lymph positive patients (both primary and metastasis tumors) and 15 lymph negative patients, making 47 samples in total. The breast cancer data is from 12 patients with both IDC and DCIS and 1 patient with only DCIS, making 25 samples in total. More details of this FISH data set can be found in Chowdhury *et al.* [27].

Chowdhury *et al.* proposed an inefficient exact algorithm and a efficient heuristic algorithm to reconstruct phylogenetic trees modeling the development of cancer cell patterns [27]. Since the inefficient exact algorithm can not finish most of the test samples with a reasonable amount of time, we compare our iterative approach to the efficient heuristic algorithm [27]. In the following text, we refer to the efficient heuristic algorithm as **FISHtrees**, and refer to our iterative approach as **iFISHtrees**.

3.2 Methods and result

Iterative approach to attack the RSMT problem

Below we described our approach for building a phylogenetic tree by using copy number change information from FISH data. As input data, each cell or data input has some non-negative integer count of each gene probe. Given two cell count patterns $(x_1, x_2, \ldots, x_d)$ and $(y_1, y_2, \ldots, y_d)$, the pairwise distance under the rectilinear metric (or L_1 distance) is defined as $|x_1 - y_1| + |x_2 - y_2| + \ldots + |x_d - y_d|$, where $x_i, y_i \in \mathbb{N}$. The weight of a tree with nodes labeled by cell count patterns is defined as the sum of all branch lengths under the rectilinear metric. Since the distance between two cell count patterns under the rectilinear metric represents the number of single gene duplication and loss events between them, a minimum weight tree, including Steiner nodes if needed, explains the n observed cell count patterns of d probes with minimum total number of single gene duplication and loss events, from a single ancestor.

Two instances of RSMT(3,d) are shown in Fig.3.2. Given three count patterns

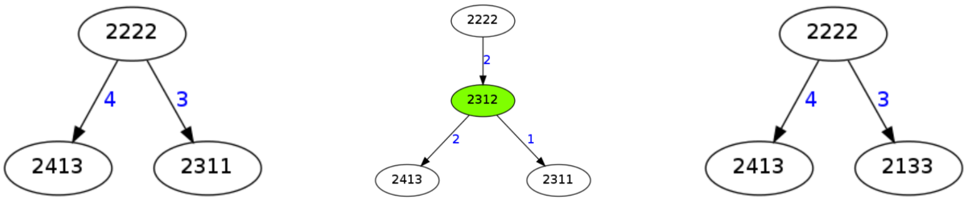


Figure 3.2 Instances of RSMT(3,d) and the introduction of the steiner node as the median.

in Fig.3.2(a), a Steiner node is introduced in Fig.3.2(b) with reduced the weight of the tree (i.e., the number of single gene duplication and loss events) from 7 to 5. Fig.3.2(c) shows an instance that no Steiner node is introduced. The median version of RSMT problem can be solved in linear time.

Theorem 3.1. *RSMT(3,d) can be solved in time O(d).*

Proof. Given three cell patterns $(x_1^1, x_2^2, \ldots, x_d^3)$, $(x_1^2, x_2^2, \ldots, x_d^2)$ and $(x_1^3, x_2^3, \ldots, x_d^3)$, RSMT$(3, d)$ returns a cell count pattern $(m_1, m_2, \ldots, m_d)$ such that $\sum_{i=1}^{3}\sum_{j=1}^{d} |x_j^i - m_j|$ is minimized, where $x_j^i, m_j \in \mathbb{N}$. Two instances of RSMT(3,d) are shown in Fig.3.2. Given three cell count patterns in Fig.3.2(a), a Steiner node is introduced in Fig.3.2(b) with reduced the weight of the tree (i.e., the number of single gene duplication and loss events) from 7 to 5. Fig.3.2(c) shows an instance that no Steiner node is introduced. Since the count for each gene probe is independent, we can optimize m_j independently which minimizes $\sum_{i=1}^{3} |x_j^i - m_j|$, respectively, and m_j simply equals to the median of x_j^1, x_j^2 and x_j^3. Thus $(m_1, m_2, \ldots, m_d)$ can be constructed in time $O(d)$ and if it differs from all three input cell count patterns then a Steiner node with cell count pattern $(m_1, m_2, \ldots, m_d)$ has to be introduced. On the other hand, $\sum_{j=1}^{d} min_{y \in \mathbb{N}} \sum_{i=1}^{3} |x_j^i - y|$ is a lower bound for the minimum weight of any steiner tree on three input cell count patterns, and $\arg\min_{y \in \mathbb{N}} \sum_{i=1}^{3} |x_j^i - y| = m_j$, thus the above construction is optimal under the rectilinear metric. □

Sankoff *et al.* studied iterative approaches to approximate solutions to the Steiner tree, which solve the median problem for one internal vertex at a time, and iteratively make improvement until a local optimum is found [114, 11]. For each internal node in the current (binary) tree, the input for a median instance consists its three immediate neighbors [11].

We first observed that considering only triplets that are immediate neighbors of the same internal node may prevent the escape from a local optimal. For example, the tree in Figure 3.3 is a local optimal with respect to all median optimizations on its internal nodes, but the tree can be further improved by introducing potential steiner nodes for all the possible triplets, not necessarily connected as a star shape, as shown in Figure 3.3(b). Since RSMT(3,d) can be solved efficiently, our approach iteratively checks all potential triplets in the tree, instead of only the triplets introduced by immediate neighbors of internal nodes.

We further observed that the order how the steiner nodes are added to the tree may also affect the minimizing the weight of the resulting tree. Figure 3.3(a) shows the original tree before iterative optimization, and Figure 3.3 (b) and (c) show the introduction of steiner nodes through two different orders. Compared to Figure 3.3 (c), Figure 3.3 (b) first introduced a steiner node 21422282 which prevents adding new potential steiner nodes in the later stage.

We define an inference score for each potential steiner node to model the inference between potential steiner nodes. The *steiner count* of any node in the current tree is define as the number of triplets which contains this node and requires the introduction of a steiner node to optimize the tree weight. The *inference score* for each potential steiner node with respect to a triplet is thus defined as the sum of *steiner counts* of three nodes in that triplet. At each iterative, the potential steiner node with minimum *inference score* is added to minimize the inference upon other potential steiner nodes with respect to the current tree. An example is shown in Figure 3.4.

Our iterative algorithm starts from a Minimum Spanning tree built from the set

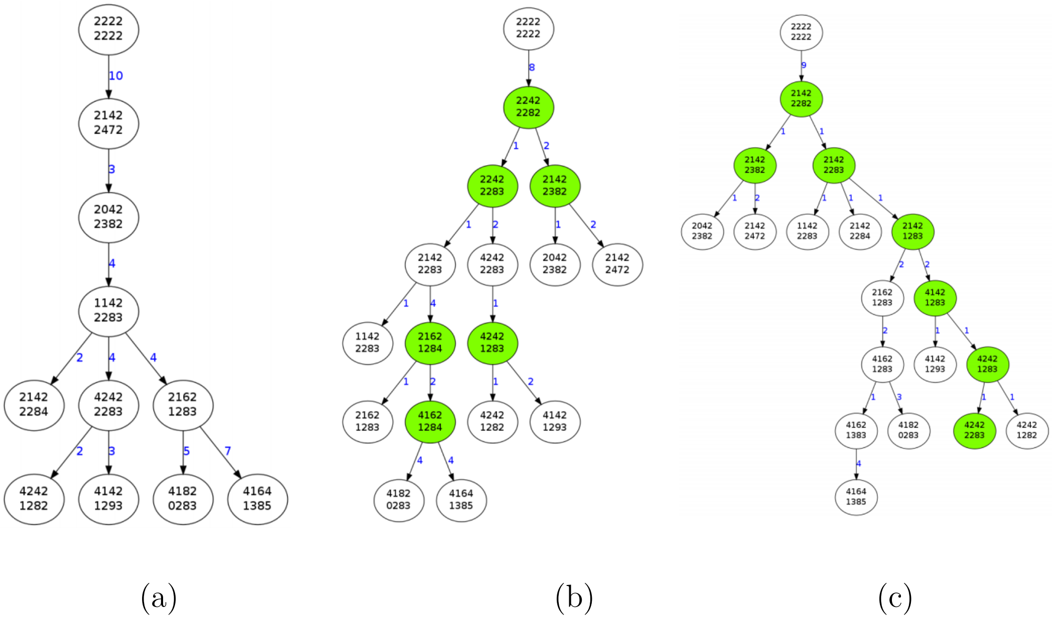


Figure 3.3 Different orders of adding steiner nodes result in different weights of the resulting trees.

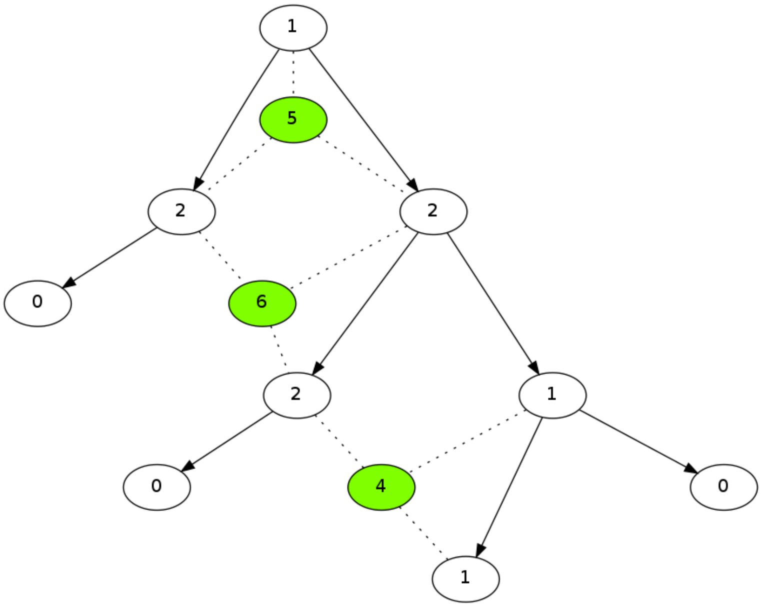


Figure 3.4 The definition of *steiner count* of the node in the current tree and the *inference score* of potential steiner nodes to be added.

of input cell count patterns, select a median instance at a time, and iteratively make improvement until a local optimum is found. The detailed description is shown in Algorithm 13.

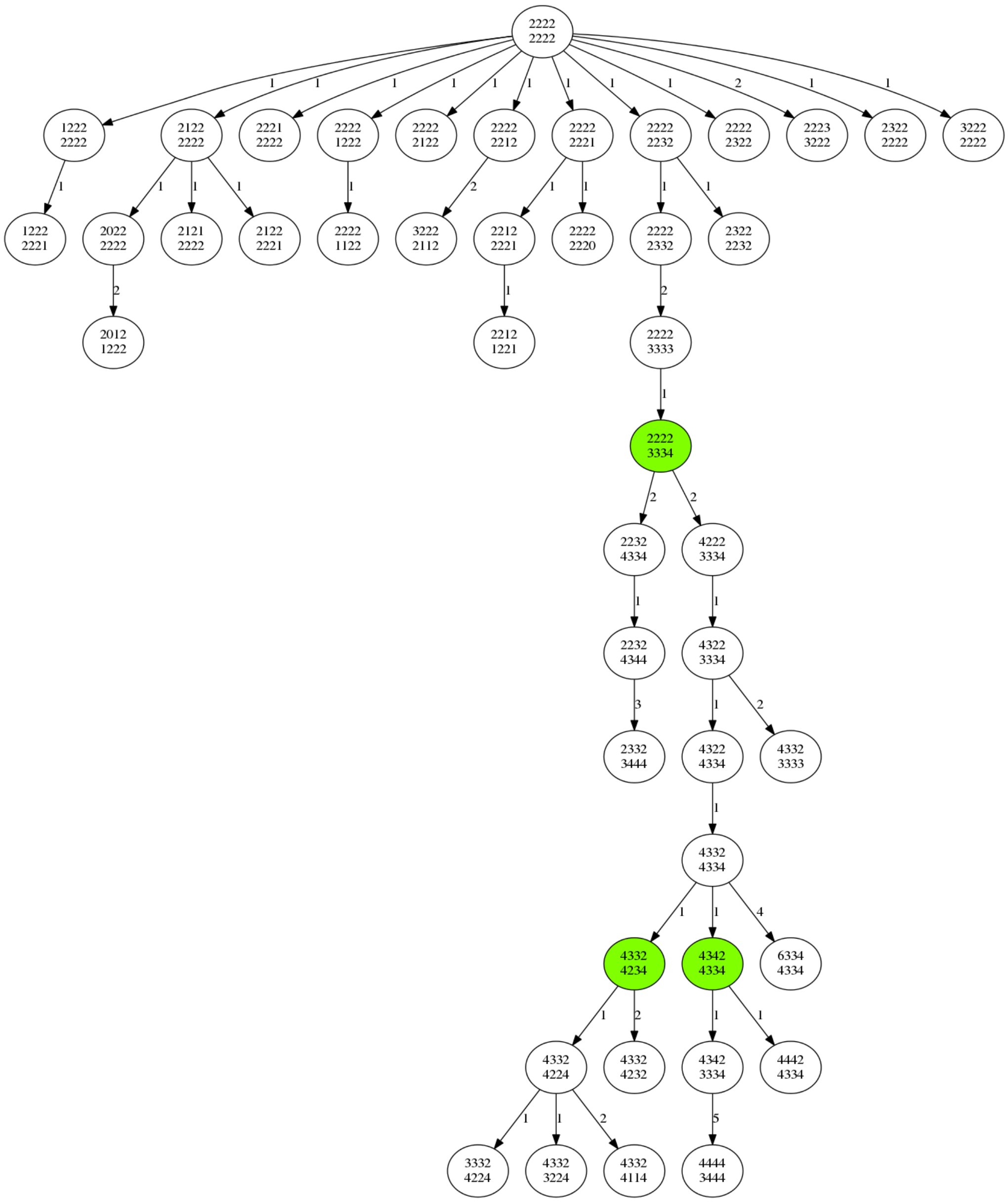
```
Input: a set of $k$ cell count patterns on $d$ gene probes
Output: a tree with additional steiner nodes if needed and $k$ nodes that
correspond to $k$ input cell count patterns respectively
Initialization: the initial tree $T_0$ = a Minimum Spanning tree on $k$ cell count
patterns under the rectilinear metric
Iteration: from tree $T_i(V_i)$ on node set $V_i$ to $T_{i+1}(V_{i+1})$ on node set $V_{i+1}$
  Identify the set $S$ of potential steiner nodes from all possible triplets in $T_i$
  While $S$ is not empty
      Select the potential steiner node $p$ with minimum inference score in $S$
      Build a Minimum Spanning tree on $\{V_i \cup p\}$ as $T(V_i \cup p)$
      If the weight of $T(V_i \cup p)$ is lower than the weight of $T_i(V_i)$
          $T_{i+1}(V_{i+1}) = T(V_i \cup p)$
      Else
          $S = S \setminus \{p\}$
Exit condition: $S$ is empty
```

Algorithm 1: An iterative algorithm to approximate solutions for RSMT

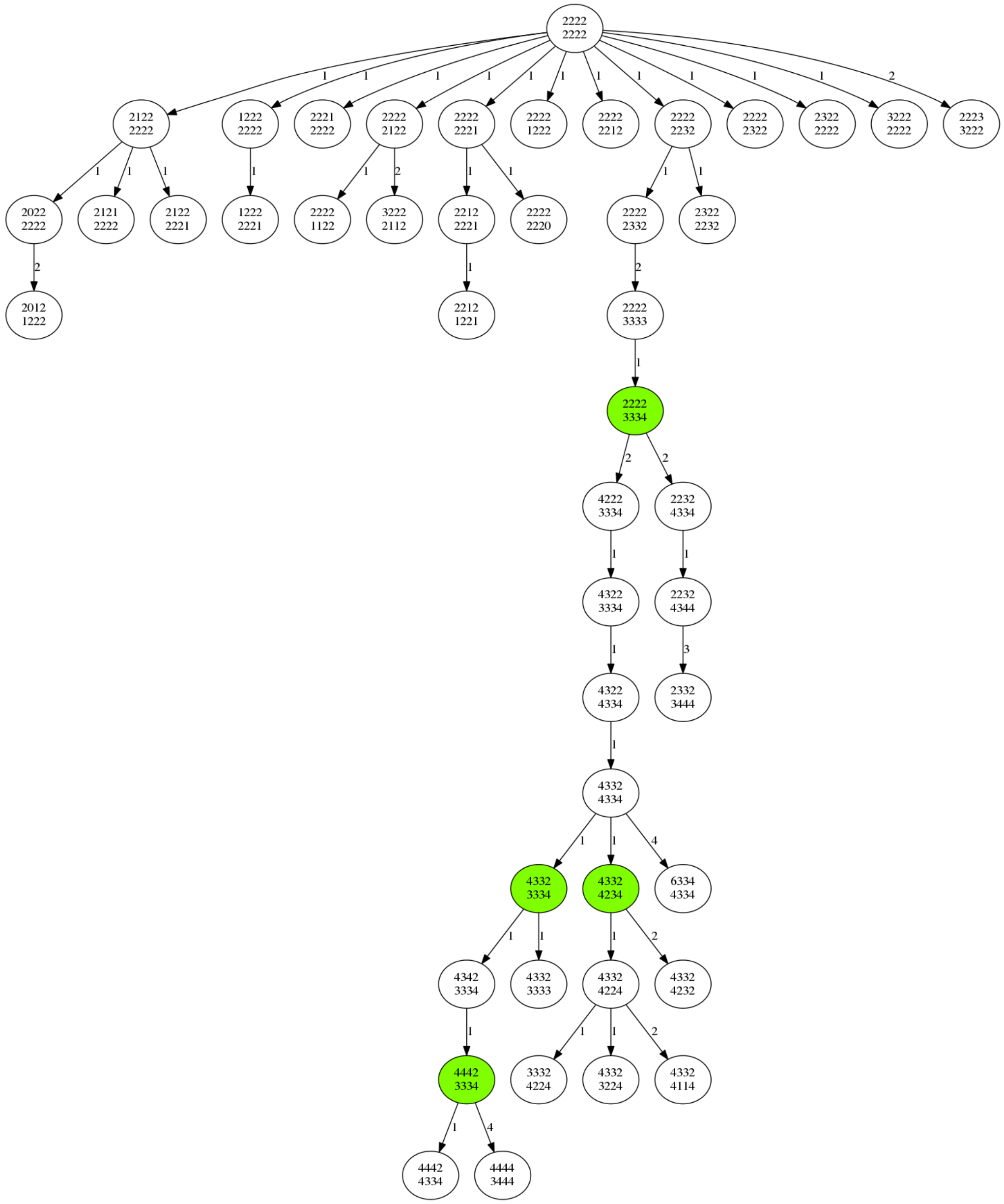
Experimental Results

Real cancer data

There are 25 data samples the breast cancer dataset, and our iterative approach **iFISHtrees** performs better than **FISHtrees** in 14 sample, ties in 10 samples, and performs worse in 1 sample. Table 3.1 summarizes the comparison between **FISHtrees** and **iFISHtrees** (ties are not included due to the space limit and the better tree weight is shown in bold). Fig. 3.5 show two trees constructed by **FISHtrees** and **iFISHtrees** reconstructed from the DCIS cancer sample from patient 13, respectively. For example, the steiner node 44423334 is introduced by iteratively checking all potential triplets in **iFISHtrees** rather than checking only the triplets introduced by immediate neighbors of internal nodes, which allows **iFISHtrees** to escape from the local optimal that has tapped **FISHtrees**.


2222 2222
1
1
1
1
1
1
1
1
1
2
1
1
1222 2222
2122 2222
2221 2222
2222 1222
2222 2122
2222 2212
2222 2221
2222 2232
2222 2322
2223 3222
2322 2222
3222 2222
1
1
1
1
1
2
1
1
1
1
1222 2221
2022 2222
2121 2222
2122 2221
2222 1122
3222 2112
2212 2221
2222 2220
2222 2332
2322 2232
2
1
2
2012 1222
2212 1221
2222 3333
1
2222 3334
2
2
2232 4334
4222 3334
1
1
2232 4344
4322 3334
3
1
2
2332 3444
4322 4334
4332 3333
1
4332 4334
1
1
4
4332 4234
4342 4334
6334 4334
1
2
1
1
4332 4224
4332 4232
4342 3334
4442 4334
1
1
2
5
3332 4224
4332 3224
4332 4114
4444 3444

(a)



(b)

Figure 3.5 Phylogenetic trees constructed by **FISHtrees**(a) and **iFISHtrees**(b) from the DCIS breast cancer sample of patient 13, respectively. Each node in the tree is labeled by a cell count pattern of eight gene probes COX-2, DBC2, MYC, CCND1, CDH1, p53, HER-2 and ZNF217. Nodes colored in green represent inferred Steiner nodes while other nodes represent input cell count patterns.

Table 3.1 Comparison on the real dataset for breast cancer samples.

Case #	Initial		FISHtrees		iFISHtrees	
	Node #	Tree weight	Node #	Tree weight	Node #	Tree weight
B1_IDC	119	230	135	213	132	**212**
B1_DCIS	143	259	158	**241**	159	242
B2_IDC	104	238	124	217	123	**216**
B3_DCIS	106	72	80	100	80	**98**
B4_IDC	110	232	129	214	129	**213**
B6_IDC	85	116	90	112	90	**111**
B7_IDC	59	128	73	116	71	**113**
B7_DCIS	76	202	84	186	83	**184**
B9_IDC	94	251	121	222	119	**217**
B9_DCIS	76	177	89	164	89	**162**
B10_DCIS	95	154	89	146	89	**145**
B11_DCIS	80	144	87	136	84	**135**
B12_IDC	112	212	124	201	123	**200**
B13_IDC	84	140	92	133	92	**131**
B13_DCIS	43	66	47	63	47	**62**

Similarly, our iterative approach **iFISHtrees** performs better than **FISHtrees** in sample 29, ties in 16 samples, and performs worse in 2 samples, out of 41 samples in the cervical cancers datasets. Table 3.2 summarizes the comparison between **FISHtrees** and **iFISHtrees** (ties are not included due to the space limit and the better tree weight is shown in bold).

Simulation data

We also test on simulated datasets generated for different number of gene probes (4, 6 and 8) and for different tree growth factors (0.4 and 0.5) [27]. For each pair of parameters, we simulate 200 samples with cell count patterns varying from 75 to 150. Table 3.3 summarizes the comparison of between **FISHtrees** and **iFISHtrees** from these simulation datasets, and in average **iFISHtrees** outperforms **FISHtrees** on all of them. Moreover, we also generate simulated datasets for relatively larger number of gene probes (e.g., 12 and above), **FISHtrees** started taking too much

Table 3.2 Comparison on the real dataset for real cervical cancer samples.

Case #	Initial		FISHtrees		iFISHtrees	
	Node #	Tree weight	Node #	Tree weight	Node #	Tree weight
C5	140	208	153	**195**	151	196
C9	130	144	131	143	132	**142**
C10	72	87	72	87	73	**86**
C12	63	72	63	72	64	**71**
C15	66	75	67	74	68	**73**
C21	63	77	67	**73**	65	74
C27	49	60	50	59	52	**57**
C29	76	85	78	83	78	**82**
C32	160	216	167	209	169	**207**
C34	67	88	72	83	73	**82**
C37	71	74	72	73	73	**72**
C42	157	207	164	199	166	**198**
C45	126	183	136	172	140	**169**
C46	87	116	92	110	93	**109**
C49	128	166	132	162	133	**161**
C51	76	83	76	83	83	**76**
C53	64	82	67	82	66	**79**
C54	123	152	129	146	130	**145**

Table 3.3 Comparison on simulated datasets.

Probe #	Growth factor	**FISHtrees =iFISHtrees**	**FISHtrees >iFISHtrees**	**FISHtrees <iFISHtrees**
4	0.4	176	23	1
6	0.4	161	30	9
8	0.4	162	31	7
4	0.5	182	18	0
6	0.5	160	31	9
8	0.5	152	32	6

time to produce solutions while our iterative approach **iFISHtrees** still scales well with dozens of gene probes, and thus we did not include the comparison results for larger number of gene probes.

3.3 Conclusions and discussion

Chowdhury *et al.* successfully modeled the progression of tumor progression using FISH copy number to the Rectilinear Steiner Minimum Tree (RSMT) problem, and proposed both exact and heuristic algorithms to reconstruct phylogenetic trees modeling the development of cancer cell patterns [27]. We show that the RSMT problem can be solved in linear time when there are only three input cell count patterns. Inspired by the iterative approaches to approximate solutions to the Steiner tree in the “small phylogeny" problem [114, 11], we propose a new iterative algorithm to approximate solutions of the RSMT problem. Moreover, our new iterative approach extends the generation of median instances, and also takes into account the order of iterative optimizations of median problem. Experimental results from both simulated and real tumor data show that our approach outperforms the previous heuristic algorithm in approximating better solutions for the RSMT problem and may provide insights into more likely tumor progression pathways.

Chapter 4

Maximum parsimony analysis of gene copy number changes in tumor phylogenetics

4.1 Background

Chowdhury *et al.* successfully modeled tumor progression using FISH copy number to the Rectilinear Steiner Minimum Tree (RSMT) problem. Their heuristic is also likely to be trapped in a local optimum if there are multiple possible Steiner nodes that can be introduced, since the order in which the Steiner nodes are added may affect the resulting tree weight. Although we try to solve this issue by adding potential Steiner nodes by better order as shown in Chapter 3, it is still like to be trapped by local optimal.

Phylogenies provide great help in the analysis of many fine-scale genetic data. The use of phylogenetics has become more and more frequent, essential in a varies of research fields such as medical research, drug discovery, epidemiology, and population dynamics [99]. For example, phylogenetics gave considerable assistance in predicting evolution of human influenza A [16], understanding the genetic evolution of HIV [111], identifying new viruses such as SARS [85], reconstructing and identifying ancestral proteins [22]. Later, phylogenies were used to study the evolutionary history in popular human diseases [90, 102, 126, 110, 27, 26]. Among these studies in particular, Chowdhury *et al.* used single-cell sampled data from affected individuals as FISH data. The survey of Beerenwinkel *et al.* summarized the uses of molecular phylogenetics [21].

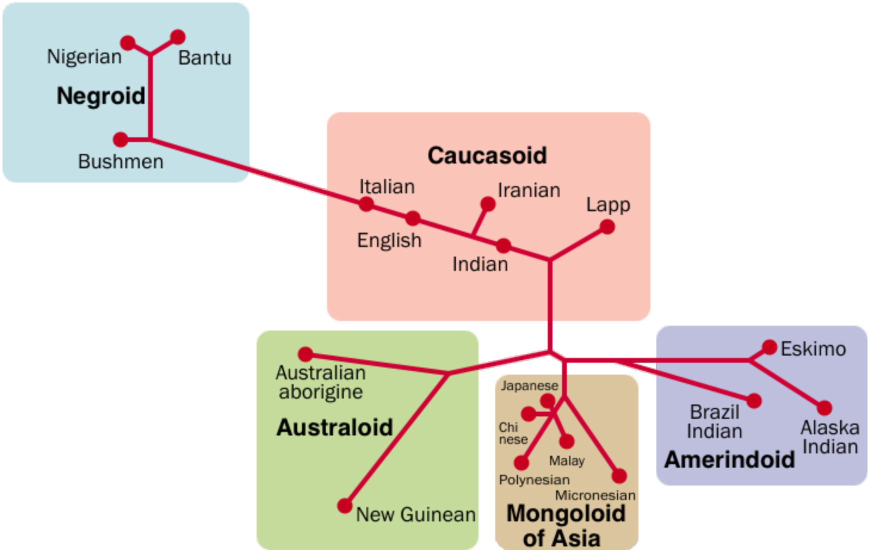


Figure 4.1 This phylogenetic tree made as an estimate of 18 world human groups by a neighbor-joining method based on 23 kinds of genetic information. It was made by Saitou Naruya professor at the National Institute for Genetics (2002).

Phylogenetics research studies the hierarchical evolutionary relationships among species, or taxa, by means of data such as DNA, RNA, amino acids, or FISH copy number. The research results are usually shown as a weighted tree, called a phylogeny, whose leaves represent the observed taxa, internal vertices represent the ancestors, edges stand for the estimated evolutionary relationships, and edge weights represent evolutionary distance between the two connected taxa [40]. Fig.4.1 is an example of phylogeny tree get from one distance-based phylogenetic inference method Neighbor-join. [112]

Methods for phylogenetic reconstruction can be roughly classified into three groups according to the criterion they follow.

- Distance-based methods:
  - Neighbor-join [112], FastME [22] and TIBA [48]
- Parsimony-based methods:
  - BPanalysis [6], [91], MGR [14],SCJ [9], PAUP [135], TNT [53].

- Probability-based methods:
  - RAxML [127]

TNT(Tree analysis using New Technology) [55] is a program available for Windows, MacOS or Linux. It has very efficient tree-searching algorithms for large data sets of great number of taxa. Parsimony is its only available optimality criterion. It implements many new heuristic search methods, such as the ratchet and sectorial searches. It can also be used for tree manipulation and diagnosis. Most real data matrices have too many taxa (i.e. more than about 100 taxa) to be analyzed by exact methods therefore a search for the most parsimonious trees must be conducted, TNT could be a good choice.

4.2 Methods and result

We developed our new RSMT solver by transferring it to be the MPT problem. We studied our MPT problem based on TNT and extensively tested on our new tool, the result shows our new tool provides more accurate phylogenetic trees comparing to earlier methods.

Problem 4.2.1. The small phylogeny problem (Maximum parsimony tree): Given a set Γ of n taxa and values $d_{ij} \geq 0$ for all pairs of taxa $i, j \in \Gamma(i \neq j)$, find a phylogeny $t* \in T$ that solves the problem:

$$\min_{t \in T} w(T) = \sum_{e \in E(T)} w_e$$

where $w_e \geq 0$ for all $e \in E(t*)$, $E(t)$ denotes the set of edges of a phylogeny $t \in T$, w_e is the weight of edge $e \in E(t)$.

The MPT problem is also NP complete [32] but heuristics like TNT [53], have largely overcome computational limitations and allow reconstructions of large trees and the use of continuous characters [54]. The copy number of each gene can be treated as continuous characters and TNT can be used to find the minimum weight phylogenetic tree. Note that, given n observed cell count patterns as the input (leaf nodes), MPT introduces (n-2) unobserved internal nodes, while the minimum spanning tree does not introduce any unobserved nodes.

In general, there may be multiple optimal solutions for the MPT problem, e.g., the internal nodes labeled by different cell count patterns. In any MPT with all nodes labeled by cell count patterns, a branch is called *trivial* if its length is 0 under the rectilinear metric. For any MPT, an unobserved internal node is a Steiner node if and only if it is labeled by a distinct cell count pattern other than any input cell count patterns. If we contract all trivial branches in MPT, the remaining unobserved internal nodes will be the Steiner nodes in RSMT. See Figure 4.2 for an example.

	Gene A Copy #	Gene B Copy #	Gene C Copy #
Count Pattern 1	2	2	2
Count Pattern 2	2	1	1
Count Pattern 3	4	2	3
Count Pattern 4	4	1	3

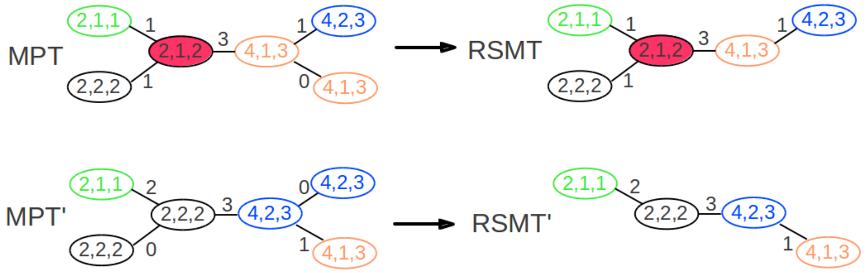


Figure 4.2 (Top) the input data of 4 cell count patterns on 3 genes. (Bottom) two maximum parsimony trees MPT and MPT', both of weight 6, are shown on the left. Nodes with identical cell count patterns are shown in the same color in both MPT and MPT'. The corresponding RSMT and RSMT', both of weight 6, are shown on the right, and the Steiner node in RSMT is colored in red.

The MPT, as obtained above, may contain up to (n-2) Steiner nodes. Following the philosophy of parsimony, we will also seek to minimize these artificially introduced nodes, although this step does not reduce the final tree weight and is not required by the formal definition of RSMT (which does not place any explicit constraints on the number of Steiner nodes). In fact, all the previous heuristics [27, 26, 162] also implicitly do not add unnecessary Steiner nodes and are biased towards a parsimonious solution due to their incremental way of adding Steiner nodes to an initial tree with no Steiner nodes.

Given any MPT, if the internal nodes are labeled by cell count patterns, the RSMT can be derived by contracting all its trivial edges; but the MPT obtained does not have labels assigned to the internal nodes. Hence the problem reduces to finding the best possible labels for internal nodes that does not increase the weight. The dynamic programming (DP) method of [134] can be adapted to find the internal labels, but modifications are needed to account for the rectilinear metric and its implications on the total tree weight. Our algorithm proceeds by finding whether a leaf label can be reused in its parent (or "lifted") for each leaf in the tree. The node with the lifted pattern is chosen to be the root node and the leaf is removed. In the bottom–up phase of the DP, labels from other leaves are propagated up the tree by using ranges of cell count patterns that can maintain the leaf cell counts without increasing the tree weight. In the top–down phase, cell count values are assigned to the internal nodes and a candidate tree is generated by contracting trivial edges. Several such candidate trees are generated by selecting different root nodes from lifted leaves. We choose a candidate tree with minimum number of Steiner nodes, without increasing tree weight.

The complete algorithm is presented in Algorithm 2 and a detailed example is shown in Fig.4.3.

The data we used to test the performance of our new tool contains real cancer data and simulation data. The real data and simulation data generation is the same as explained in Chapter 3. Figure 4.4 shows three approximate RSMT trees for the cervical cancer sample of patient 29, constructed by **FISHtree** (Figure 4.4(a),

```
Input: MPT with optimal weight $W_{opt}$
Output: RSMT with optimal weight $W_{opt}$
  For each $Leaf$ in MPT
      $Parent(Leaf)$: the parent node of $Leaf$ in MPT
      $MPT \setminus Leaf$: the tree obtained by removing $Leaf$, rooted at
$Parent(Leaf)$
            (Figure 4.3(a))
      Compute the ranges of possible values in internal nodes in $MPT \setminus Leaf$
            (DP bottom-up phase; Figure 4.3(b))
      Assign the cell count pattern of $Leaf$ to $Parent(Leaf)$
      Determine all the values for all other internal nodes in MPT
            (DP top-down phase; Figure 4.3(c))
      Contract all trivial branches in $MPT \setminus Leaf$ and derive $RSMT^*$
            (Figure 4.3(d))
      If the weight of $RSMT^*$ is equal to $W_{opt}$
          Store $RSMT^*$ as a candidate RSMT
  Return a candidate RSMT with the minimum number of Steiner nodes
```

Algorithm 2: Algorithm to derive RSMT from MPT

tree weight = 83), **iFISHtree** (Figure 4.4(b), tree weight = 82) and **mpFISHtree** (Figure 4.4(c), tree weight = 81), respectively. In this figure, we refer to previous heuristics as **FISHtree** [27, 26][1] and **iFISHtree** [162], and we refer to our Maximum-Parsimony based approach as **mpFISHtree**. We also refer to the exact method [27] as **Exact**.

Real Cancer Datasets

We use both the real cervical cancer and breast cancer which the same as the data used by Chowdhury *et al.* Table 4.1 and Table 4.2 summarize the comparison of **FISHtree**, **iFISHtree** and **mpFISHtree** for breast cancer samples and cervical cancer samples, respectively (and the best tree weights are shown in bold). Note that **mpFISHtree** of the three heuristic methods has the best performance in all the samples. Figure 4.4 shows three approximate RSMT trees for the cervical cancer

[1] We use the best result derived from the heuristic option in [27] and the option PLOIDY_LESS_HEURISTIC in [26] that also approximate RSMT under the case of gene copy number changes of single probes.

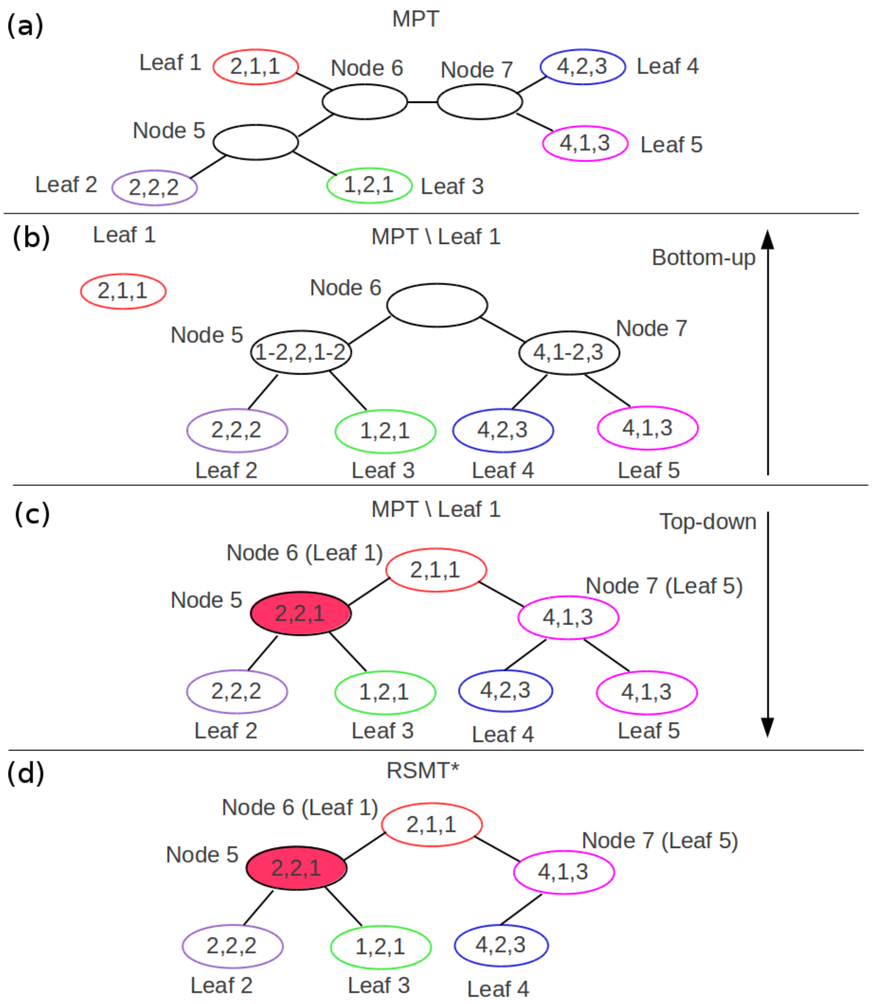


Figure 4.3 An example to test whether $Leaf_1$ can be optimally "lifted" to its parent node $Node_6$ in MPT. (a) a MPT on 5 leaves and 3 internal nodes. (b) $Leaf_1$ and compute the ranges of possible values to internal nodes, except $Node_6$, in $MPT \setminus Leaf_1$ in a bottom-up phase. (c) Assign the cell count pattern of $Leaf_1$ to the root of $MPT \setminus Leaf_1$, and determine the values for other internal nodes in $MPT \setminus Leaf_1$ in a top-down phase. (d) Contract all trivial branches in $MPT \setminus Leaf_i$ and derive $RSMT^*$. Nodes with identical cell count patterns are shown in the same color and the Steiner node in RSMT* is colored in red.

sample of patient 29, constructed by **FISHtree** (Figure 4.4(a), tree weight = 83), **iFISHtree** (Figure 4.4(b), tree weight = 82) and **mpFISHtree** (Figure 4.4(c), tree weight = 81), respectively.

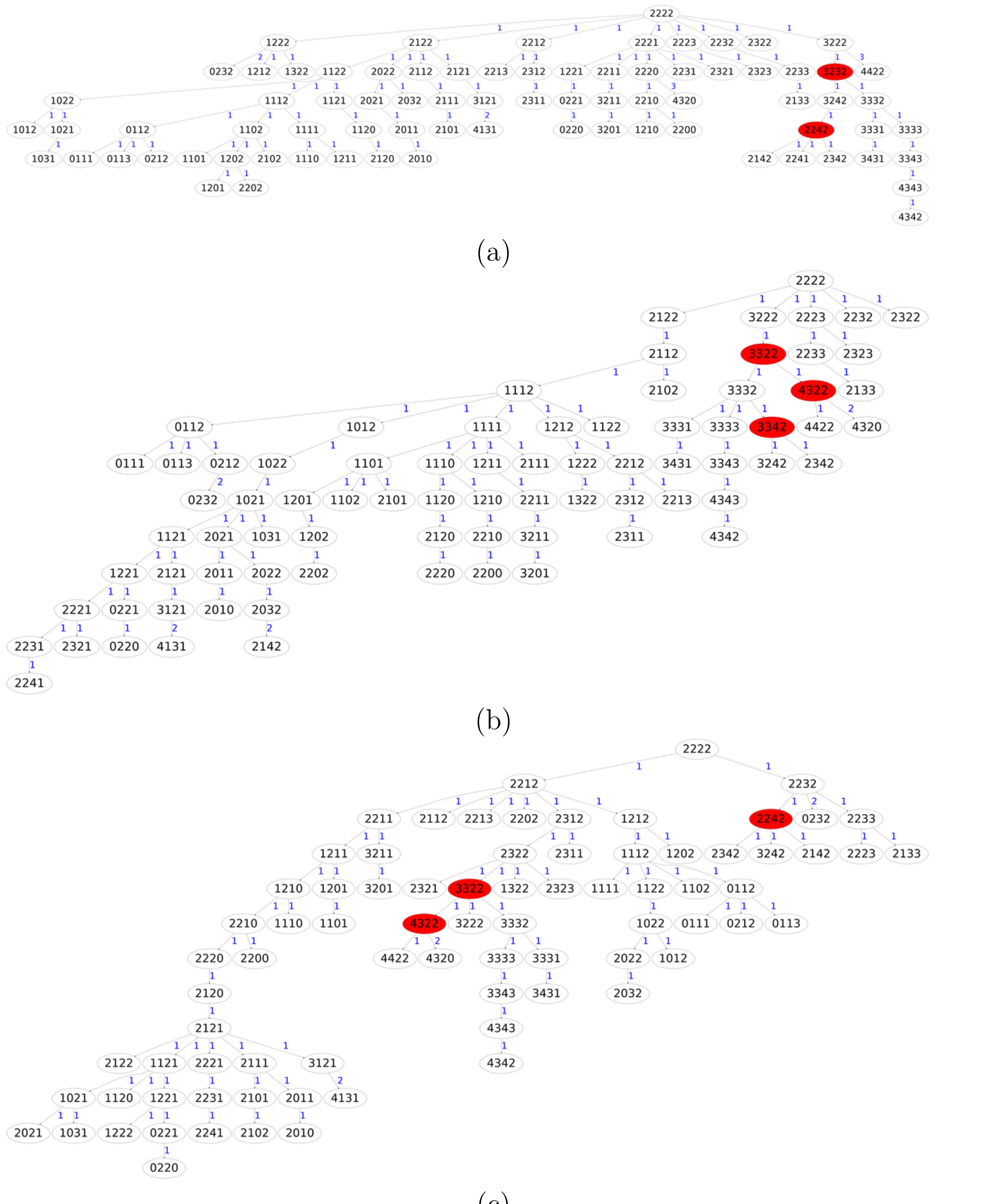


Figure 4.4 Given the metastatic cervical cancer sample of patient 12, (a) approximate RSMT constructed by **FISHtree** with weight 83, (b) approximate RSMT constructed by **iFISHtree** with weight 82 and (c) approximate RSMT constructed by **mpFISHtree** with weight 81. Each node in the tree is labeled by a cell count pattern of four gene probes LAMP3, PROX1, PRKAA1 and CCND1. Each white node represents an input cell count pattern, and each red node represents an inferred Steiner node. Branch lengths are shown in blue.

Table 4.1 Comparison on the real datasets for breast cancer samples (**Exact** results are not available due to the time limitation). The best tree weights are shown in bold for each sample. The number of Steiner nodes is shown in parenthesis. 7 breast cancer samples have ties in tree weights and thus are not included due to the space limit.

Case #	Tree weight (# Steiner nodes)		
	FISHtree	iFISHtree	mpFISHtree
B1_IDC	213 (15)	212 (13)	**211** (19)
B1_DCIS	241 (14)	242 (15)	**239** (22)
B2_IDC	217 (15)	216 (20)	**211** (22)
B2_DCIS	56 (2)	56 (2)	**55** (3)
B3_DCIS	100 (7)	**98** (7)	**98** (10)
B4_IDC	214 (16)	**213** (17)	**213** (17)
B6_IDC	112 (4)	**111** (4)	**111** (6)
B7_IDC	116 (8)	**113** (12)	**113** (12)
B7_DCIS	186 (13)	184 (14)	**182** (22)
B9_IDC	222 (22)	217 (25)	**213** (30)
B9_DCIS	164 (12)	163 (13)	**161** (15)
B10_IDC	128 (4)	128 (4)	**127** (4)
B10_DCIS	146 (6)	**145** (8)	**145** (9)
B11_DCIS	136 (6)	135 (7)	**134** (7)
B12_IDC	201 (9)	200 (10)	**198** (15)
B12_DCIS	161 (9)	161 (10)	**158** (13)
B13_IDC	132 (7)	**131** (8)	**131** (8)
B13_DCIS	63 (3)	**62** (4)	**62** (4)

Simulated cancer data

4.3 Discussion

The Rectilinear Steiner Minimum Tree (RSMT) has been shown to be a good model for progression of cancer cells using FISH cell count pattern data [27, 26]. Efficient heuristics are necessary to obtain approximations to RSMT since finding the optimal solution is NP–hard. We present a new algorithm to approximate the RSMT based on Maximum Parsimony (MP) phylogeny reconstruction. Our experiments on synthetic and real datasets demonstrate the superiority of our algorithm over previous methods in obtaining better parsimonious models of cancer evolution. RSMT instances found

Table 4.2 Comparison on the real datasets for cervical cancer samples. The best tree weights are shown in bold for each sample. The number of Steiner nodes is shown in parenthesis. 24 cervical cancer samples have ties in tree weights and thus are not included due to the space limit.

Case #	Tree weight (# Steiner nodes)			
	FISHtree	iFISHtree	mpFISHtree	Exact
C5	195 (13)	196 (12)	**194** (13)	**194** (13)
C6	82 (2)	82 (2)	**81** (5)	**81** (4)
C8	103 (6)	103 (6)	**100** (9)	**100** (8)
C9	143 (1)	**142** (2)	**142** (5)	**142** (2)
C10	87 (0)	**86** (1)	**86** (1)	**86** (1)
C12	72 (1)	**71** (2)	**71** (2)	**71** (2)
C13	150 (5)	150 (5)	**149** (7)	**149** (7)
C15	74 (1)	**73** (2)	**73** (2)	**73** (2)
C18	127 (4)	127 (4)	**126** (6)	**126** (6)
C21	**73** (4)	74 (3)	**73** (5)	**73** (4)
C27	59 (1)	**57** (3)	**57** (2)	**57** (3)
C29	83 (2)	82 (3)	**81** (3)	**81** (3)
C30	118 (9)	118 (9)	**116** (9)	**116** (10)
C32	209 (7)	207 (9)	**205** (14)	**205** (13)
C34	83 (5)	**82** (6)	**82** (6)	**82** (6)
C35	67 (1)	67 (1)	**66** (2)	**66** (3)
C42	199 (7)	198 (9)	**197** (12)	**197** (11)
C45	172 (10)	**169** (13)	**169** (14)	**169** (15)
C46	110 (5)	109 (6)	**108** (8)	**108** (7)
C49	162 (4)	**161** (5)	**161** (7)	**161** (7)
C53	80 (3)	**79** (4)	**79** (4)	**79** (4)
C54	146 (6)	145 (7)	**144** (10)	**144** (9)

Table 4.3 Comparison on simulated datasets: number of times and percentage that the best scoring tree (including ties) is obtained by the four methods. **Exact** results for datasets with over four gene probes are not available due to the time limitation.

Probe #	Growth factor	Best score count (Best score percentage)			
		FISHtree	iFISHtree	mpFISHtree	Exact
4	0.4	92 (46%)	137 (68.5%)	196 (98%)	200
6	0.4	70 (35%)	98 (49%)	194 (97%)	N/A
8	0.4	41 (20.5%)	69 (34.5%)	196 (98%)	N/A
4	0.5	93 (46.5%)	130 (65%)	194 (97%)	200
6	0.5	68 (34%)	99 (49.5%)	196 (98%)	N/A
8	0.5	40 (20%)	64 (32%)	195 (97.5%)	N/A

by our method (as well as previous heuristics) have multiple solutions with the same tree weight and additional constraints are needed to choose one from them. We choose the parsimonious solution of minimizing the Steiner nodes introduced by MP reconstruction. Proving that our method produces the solution with the minimum number of Steiner nodes and exploring other strategies to choose from multiple RSMT solutions remain open problems. The RSMT instances usually have multiple optimal solutions in terms of the overall tree weights, and other sources of information may be used to distinguish them, e.g., minimizing the number of the Steiner nodes (although we are not sure whether the algorithm in Section 4.2 guarantees the minimum number of Steiner nodes introduced). We may further other optimization strategies to explore multiple optimal solutions.

Chapter 5

Tumor phylogenetic study considering of large scale change

5.1 Background

It is quite clear that evolution is fundamental to cancer treatment problem such as drug resistance[44]. There has been extensive work on tumor phylogenetics, however, the study of algorithms for reconstructing tumor evolution for large numbers of single cells has limitation comparing to great advances in data generation. The most method used for single-cell tumor phylogenetics remains the use of simple generic phylogeny algorithms such as neighbor-joining [112] that are not set to model the patterns of copy number changes based on the theory that chromosome abnormalities are the key changes. We developed algorithms to find copy-number phylogenies from a group of cells with several probes. The earlier work was limited to a simple model that tumor cells only evolve gaining or losing a single copy number on a single probe each step. While large scale gene changes(including duplication of the entire chromosome or genome) are commonly observed in cancer development, whole genome duplication can be observed in around 37% cancer sample by The Cancer Genome Atlas Pan-Cancer study [160]. In real tumors, gene copy numbers changes can be summarized as the following three events. The first is single gene duplication/loss event as the kind of event only considered in previous two chapters. The second is chromosome duplication/loss event, in which a gene changes on the chromosome level while the chromosome changed might contain multiple FISH markers. The last one is whole genome duplication event in which all the gene markers doubles after one operation. The theoretical changes is summarized in Fig.5.1. The work in this chapter tries to design the new scalable phylogenetic algorithms to fit more realistic models of tumor-like evolution of using hundreds of single cells per tumor. Chowdhury *et al.* recently extended the evolutionary model of tumor progression by gene copy number changes of single probes to all probes, jointly, on a gene, a chromosome and the whole genome [26]. In the following, we show how to extend the heuristics for RSMT to derive approximate solutions for DSMT.

5.2 Method and result

In this chapter, we develop a new method to advance the theory of phylogenetic inference for cell populations in solid tumor. The data we used here is the same as the last two chapter which is assessed by multicolor fluorescence in situ hybridization(FISH). We still try to identify a most parsimonious tree of copy number changes on single cell copy-number heterogeneity with considering large scale change. The main result needed is the method being able to infer minimum distances between two cell patterns by considering copy number changes on the three previous mentioned levels. In the work of Chowdhury*et al.* [26], the large scale changes, chromosome and whole genome level changes are identified. We follow the idea from Chowdhury to first identify possible large scale duplications. Specifically, given a tree reconstructed by [26] for DSMT, we first locate all branches containing large scale duplications (including both chromosomal and whole genome duplications). We then remove such branches, and thus split the tree into disjoint subtrees. For each subtree, we use only the leaf genomes as the input and reconstruct a new RSMT tree by using the above two heuristics (described in Chapter 2 and 3). Finally, we re-insert the removed branches and thus assemble the reconstructed RSMT subtrees into a new tree which is our approximate solution for DSMT. Again the most time consuming part is phylogenetic

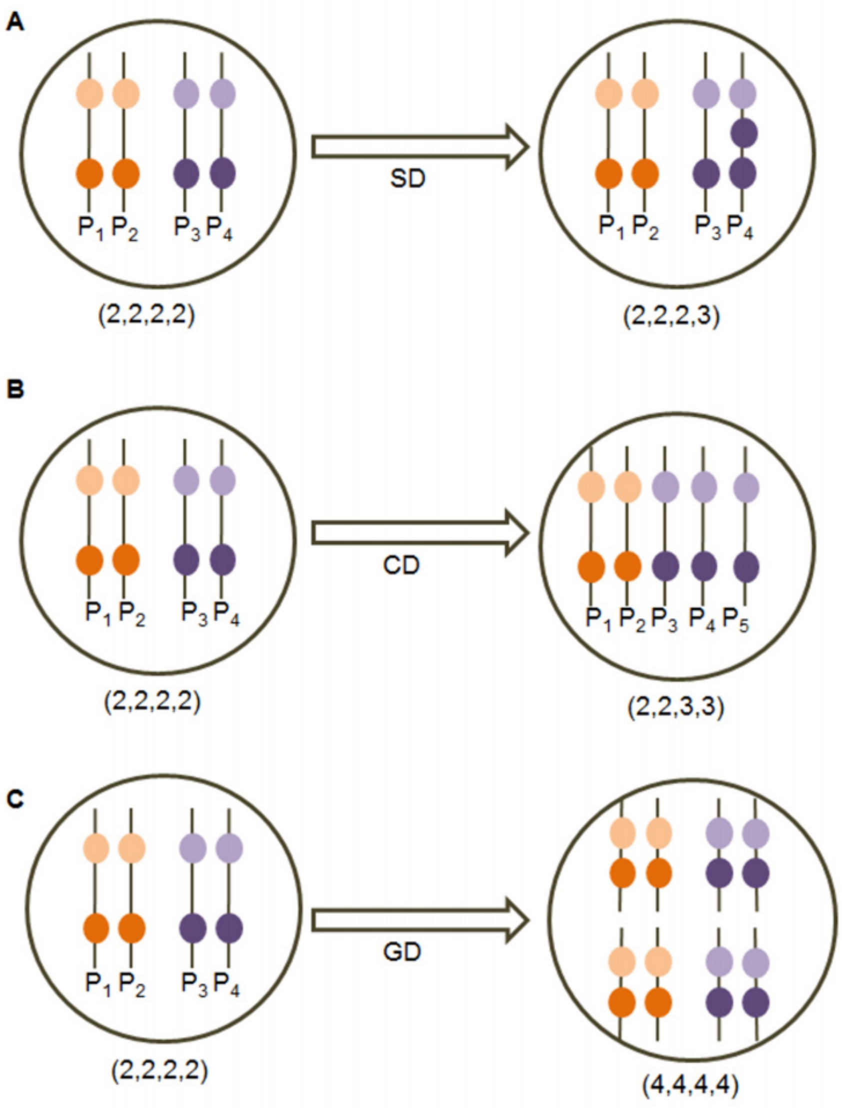
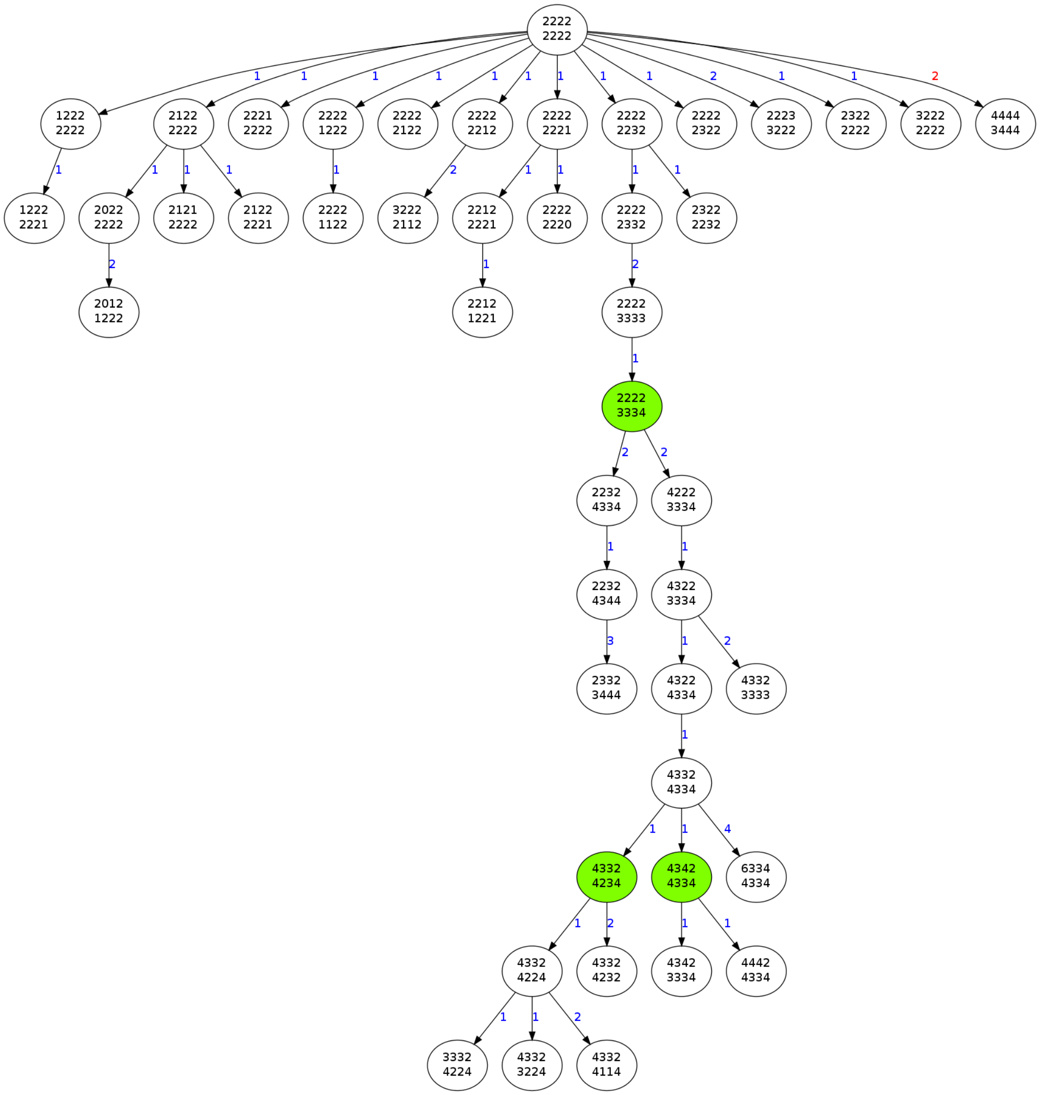
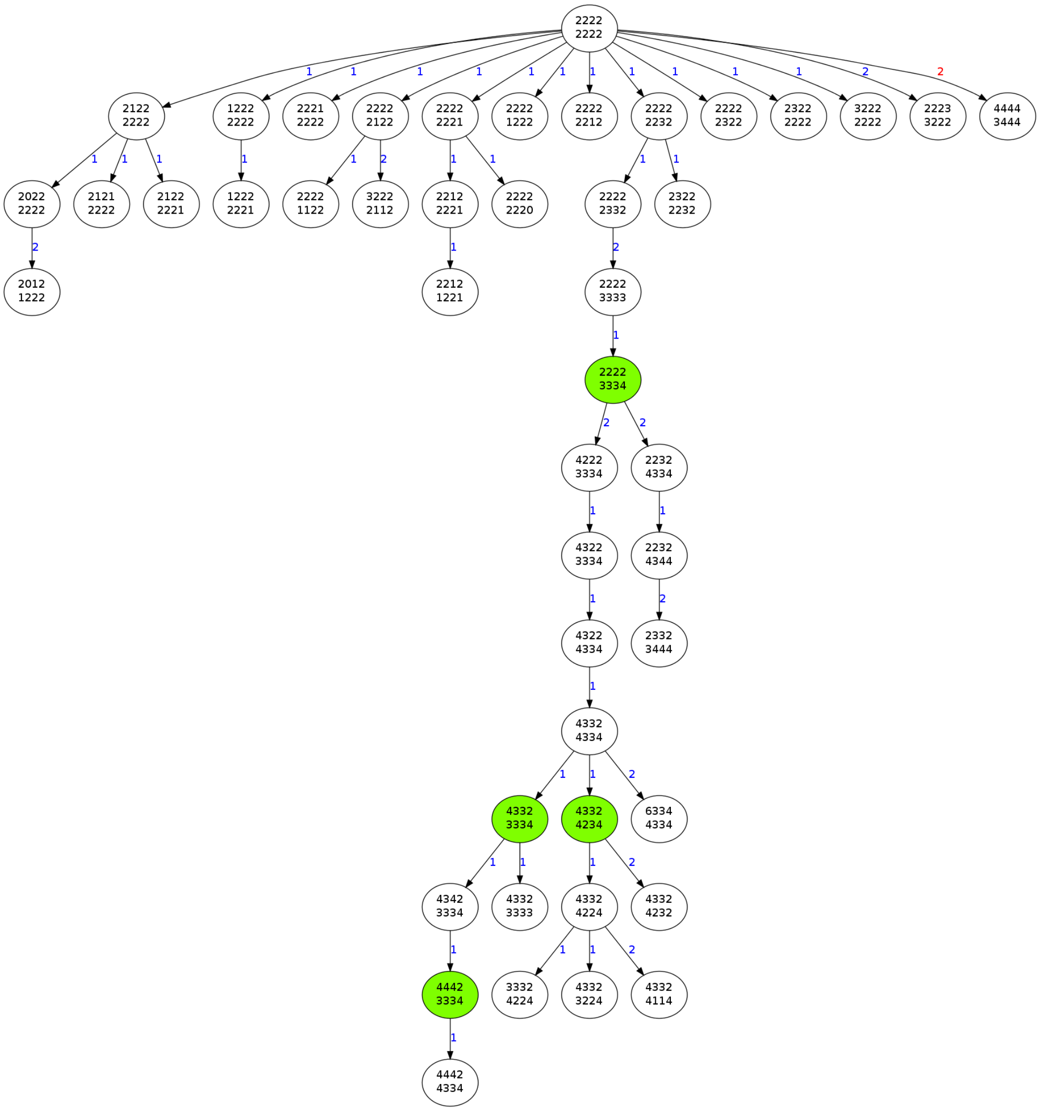


Figure 5.1 Graph shows the three mechanisms of hypothetical copy number changes. The original gene copy numbers are all 2 named as P1, P2 and P3, P4 respectively. After the single gene duplication event, the copy number of a gene located on P4 gets increased by 1(A). The chromosome P4 gets duplicated and the cell has one extra copy of that chromosome as chromosome P5 after the single chromosome duplication event(B). All the chromosomes are duplicated and the total number of chromosomes in The new cell is twice the number of chromosomes in the original cell after the whole genome duplication event(C).

inference using TNT software. Since **TNT** is quite efficient, so average time spent on the whole tree inference is much shorter comparing to the earlier methods.


2222 2222
1 1 1 1 1 1 1 1 1 2 1 1 2
1222 2222
2122 2222
2221 2222
2222 1222
2222 2122
2222 2212
2222 2221
2222 2232
2222 2322
2223 3222
2322 2222
3222 2222
4444 3444
1 1 1 1 1 2 1 1 1 1
1222 2221
2022 2222
2121 2222
2122 2221
2222 1122
3222 2112
2212 2221
2222 2220
2222 2332
2322 2232
2 1 2
2012 1222
2212 1221
2222 3333
1
2222 3334
2 2
2232 4334
4222 3334
1 1
2232 4344
4322 3334
3 1 2
2332 3444
4322 4334
4332 3333
1
4332 4334
1 1 4
4332 4234
4342 4334
6334 4334
1 2 1 1
4332 4224
4332 4232
4342 3334
4442 4334
1 1 2
3332 4224
4332 3224
4332 4114

(a)



(b)

Figure 5.2 Given the metastatic breast cancer sample of patient 13(DCIS sample), (a) approximate RSMT constructed by **FISHtree** with weight 61 and (b) approximate RSMT constructed by **mpFISHtree** with weight 63. Each node in the tree is labeled by a cell count pattern of eight gene probes CDH1, COX2, DBC2,Her2, MYC, ZNF217, p53 and CCND1. Each white node represents an input cell count pattern, and each red node represents an inferred Steiner node. Branch lengths are shown in blue while the edge with WGD is labeled in red.

As in Chapter 2 and 3, we compare the trees inferred by our new tools and earlier methods by using real data and simulation. Here, similarly, we compare our tool to FISHtree from Chowdhury *et al.* by applying them to cervical and breast cancer data. We also did comparison based on simulation data. For the DSMT problem, we compare **FISHtree** [26] and **MPTtree**, since **MPTtree** outperforms **MSTtree** for RSMT. We first gave one example on metastatic breast cancer data sample of patient 13 in Fig. 5.2. We summarize the results on breast cancer samples and cervical cancer samples in Table 5.1 and Table 5.2 (better tree weights are shown in bold). We summarize the results on simulation data samples in Table 5.3 (better tree weights are shown in bold). Similarly, **MPTtree** outperforms **FISHtree** in both real cancer and simulation data. Note that DSMT problem is NP-hard and so obtaining optimal solutions can be very difficult. Although the improvements in terms of tree weights appear small, coming closer to the optimal tree even by a few units is challenging. The improvements are more clearly seen on simulated data in the following section.

Table 5.1 Comparison on the real datasets for DSMT on breast cancer samples: number of times and percentage that the best scoring tree (including ties) is obtained by **FISHtree** and **MPTtree**.

Cell Line	DSMT Best score	
	FISHtree	MPTtree
B1_IDC	217	**206**
B1_DCIS	150	**140**
B2_IDC	203	**189**
B3_DCIS	99	**97**
B4_IDC	203	**193**
B5_IDC	64	**63**
B6_IDC	108	**106**
B6_DCIS	**42**	43
B7_IDC	116	**115**
B10_IDC	125	**123**
B11_DCIS	122	**121**
B12_IDC	125	**123**
B12_DCIS	162	**149**
B13_IDC	132	**129**
B13_DCIS	63	**61**

Table 5.2 Comparison on the real datasets for DSMT on cervical cancer samples: number of times and percentage that the best scoring tree (including ties) is obtained by **FISHtree** and **MPTtree**.

Cell Line	DSMT Best score	
	FISHtree	MPTtree
C6	82	**81**
C8	95	**93**
C18	126	**122**
C24	**201**	204
C29	80	**76**
C34	**81**	82
C53	75	**71**

Table 5.3 Comparison on simulated datasets for DMST: number of times and percentage that the best scoring tree (including ties) is obtained by **FISHtree** and **MPTtree**.

Probe #	Growth factor	DMST Best score count (Best score percentage)	
		FISHtree	MPTtree
4	0.4	175 (87.5%)	191 (95.5%)
6	0.4	145 (35%)	194 (97%)
8	0.4	101 (50.5%)	199 (99.5%)
4	0.5	178 (89%)	189 (94.5%)
6	0.5	147 (73.5%)	193 (96.5%)
8	0.5	93 (46.5%)	200 (100%)

5.3 Discussion

This chapter presents novel theory and algorithms for reconstructing evolutionary process from gene copy numbers in solid tumors. The new tool is built based on a model which incorporates changes at the scale of single gene probes, full chromosomes, or whole genome level. The novel approach utilizes the work of Chowdhury *et al* which can infer the edges contain whole genome duplication and our earlier tool, MPtree, which can infer maximum parsimony tree. Experimental results on simulated data confirm the ability of the new methods to improve phylogenetic inference accuracy relative(lower tree weight) to earlier models by applying more universal events and maximum parsimony tree. Application to real human tumor data, cervical and breast cancer data, shows that these extended evolutionary models are able to yield more parsimonious tree reconstructions. In the next chapter and future work, we will try to extend the theory developed here to handle even more realistic models and more challenging data types such as high throughput data. One way is to improve more upon the heuristic approximations to better approach the goal of finding the large scale change globally. The evolutionary models, likewise, might be further extended to go beyond the three events mentioned above such as insertion, inversion. The data sets contains only gene copy number information while we can get more information about the cancer data such as direction of each alleles. Further more, single-cell sequencing has the potential to eventually become practical for tumor phylogenies, it should worth to extend the tool developed here to single-cell sequencing data. Since the current status of the technologies will introduce error in the data, we should develop tools to tolerate the error-prone data. Finally, we hope to make more use of these new tools to further explore the biological insights by gaining from more accurate tumor phylogenies from the tools.

Chapter 6

Application of Phylogenetic Analysis to tumor prognosis

6.1 Background

Cancer is a multi-step process which is driven by genetic mutations and epigenetic alterations of DNA which can be easily shown in Fig. 6.1. By acquiring the start mutations, a subset of cells get a faster growth rate and ultimately we have multiple subpopulations of genomically distinct cell types forming the tumor itself. Afterwards, the tumor may become invasive, with cells migrating and colonizing to other parts of the body and potentially becoming life threatening because some other key mutations happened during the process. Recent research suggests a branch-type evolutionary mechanism [2] instead of a linear model of evolution [38, 97]. Another important character about tumor is that it is a heterogeneous system [132]. There are experimental evidences of both intratumor and intertumor heterogeneity. Earlier studies have revealed distinct mutations in closely related tumors and large heterogeneity within single tumors. Navin *et al.* [95, 94] demonstrated that a single breast cancer biopsy may contain multiple intermixed tumor populations that differ by major structural chromosomal gene amplifications. Gerlinger *et al.* [49] showed that a single biopsy may underestimate the somatic mutational landscape of a tumor. Campbell *et al.* [18] identified some amplification of cancer genes occurs predominantly in early cancer development and all tumors studied harbored a dominant clone that was distinct from other subclones. One explanation in their work is that some particular genotypes might drive metastasis. The intertumor and intratumor heterogeneity can be utilized on both diagnosis and treatment of cancer [57, 132].

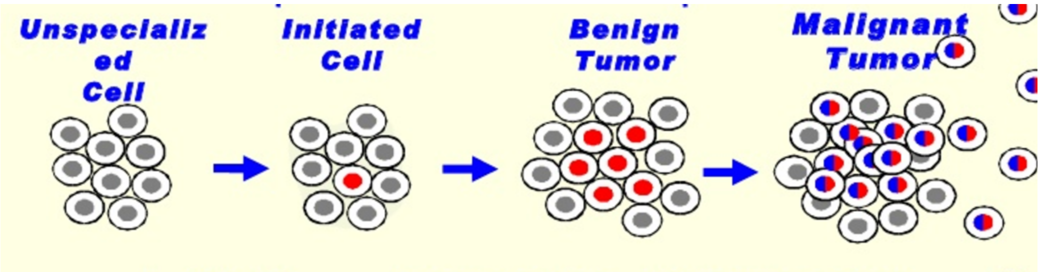


Figure 6.1 Emergence and progression of cancer.

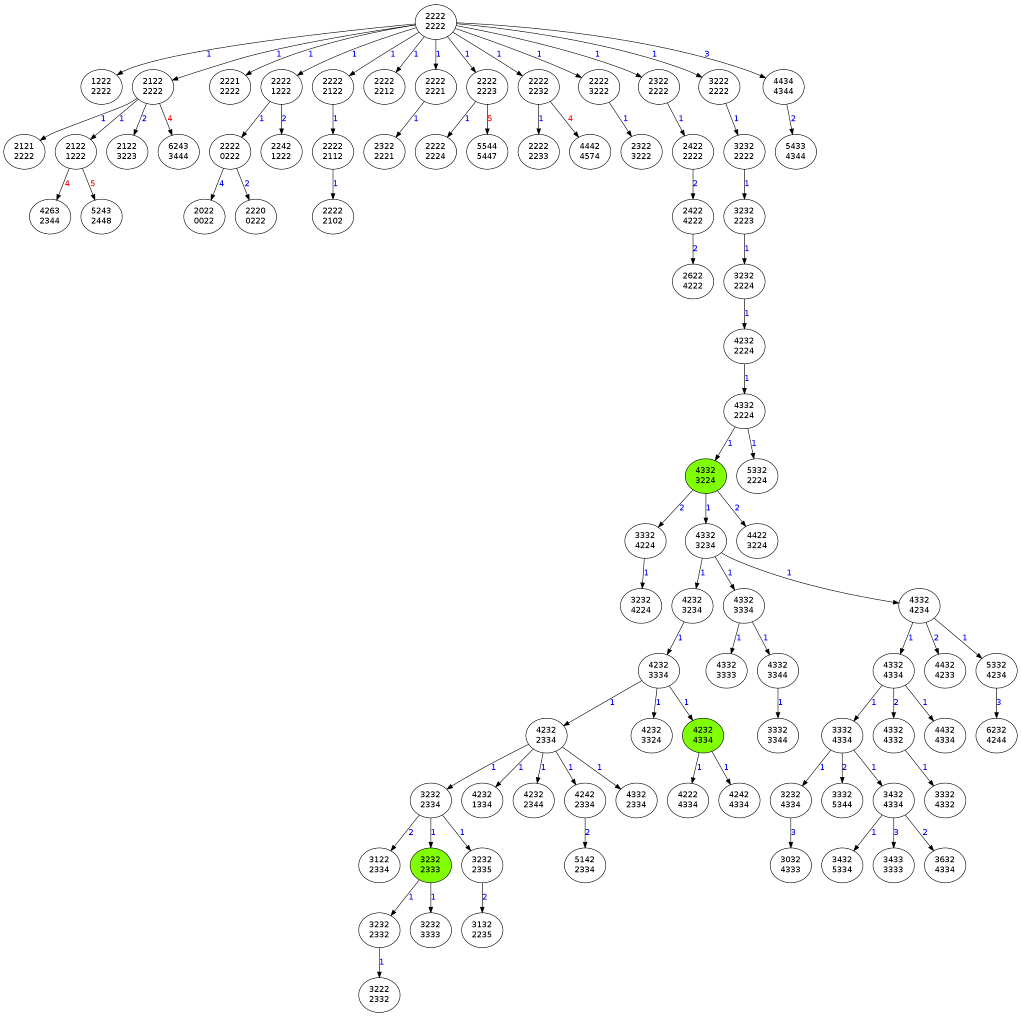
As cancer is a continuously evolving system, the success in treatment of patients depends not only the current state, but also what state it will be over the course of the treatment. From the "evolutionary" perspective, the prospection state or the changes during treatment can be predicted, thus the treatment can be expected to change. However, to identify the cancer progression process has various problems in the diagnosis, prognosis and treatment of the disease, such as sampling bias and biomarkers discovery [132]. At present, normally targeted treatment with chemotherapy is done based on the primary lesion, which may have been collected a long time ago before the treatment. Such a treatment strategy may cease to work well for the reason of dramastic changes during treatment.

The biomarker discovery approaches that combine prediction of gene function with the help of genetic or transcriptomic analyses of tumor tissue often depend on tumor biopsies collected from the primary or metastatic site of the tumor to prioritize the identification of candidate biomarkers for validation which can easily lead to sampling bias [132]. Early detection of cancer is much more useful and crucial than treatment of the disease at late stages [123]. This observation from earlier research shows that many earlier detected cancers never posed a threat to the patient [12]. On the other side, overtreatment of cancers can become a more dangerous threat to patient heath [13]. The assumption that initiation and progression of cancer is result of changes in a number of cellular functions and pathways caused by several "driver" genes has been commonly accepted [62, 61]. So identification of the d̈riverg̈enes gives possibility to design drug or therapeutics targeting these genes for each patient specifically [33, 100]. To date, more than 100 targeted therapeutics have been identified for the d̈riverg̈enes, thanks to the great advance in large-scale sequencing. Although targeted therapeutics [24] has succeeded a lot, however, for most patients it is only short-term recovery because the emergence of drug resistance [44].

The third character of tumor is the rate of mutations which is called hypermutability [64]. The best known example of hypermutability is mutation in the TP53 gene observed in a majority of human solid tumors [58], which leads to extensive gain, loss and rearrangement of whole or large fractions of chromosomes during cell division. Changes in cancer genomes affect both a single nucleotide and large segments of genes, resulting in Single Nucleotide Variation (SNV) and Copy Number Aberrations (CNA). SNVs are variations within genomes at single base levels. CNAs result in larger rearrangements in chromosome segments and are an effective force behind the larger genomic and phynotypic heterogeneity observed in cancer. In cells, copy numbers of genes may change as a result of Single Gene Duplication, Chromosome Duplication or Whole Genome Duplication events. There are current two widely accepted sources for rearrangement, chromothripsis and chromoplexy. Chromothripsis refers to the phenomenon of massive genomic rearrangements as a result of a single catastrophic effect during the cell lifetime [128]. Chromoplexy is another source of complex genomic rearrangements where several strands of DNA are broken and then ligated together to form a new configuration. There has been evidence of chromoplexy both at clonal and subclonal level in prostate tumors [5]. Tumor cells accumulate mutations and undergo large genomic rearrangements during the cell cycle to form distinct subpopulations of cells. Due to this evolutionary nature of the progression of cancer, phylogenetic inference tools started widely being used to study the evolutionary relationships for cancer research [7]. Desper *et al.* firstly proposed "oncogenetic tree" which uses a wild-type node as the tree root and other nodes representing a cell type which arises after genomic changes [36].

Then a maximum weight branching-based tree construction algorithm was proposed [37]. Desper *et al.* proposed a distance based tree construction method to estimate dependencies among cancer driving events [35, 74]. Von Heydebreck *et al.* proposed a maximum likelihood based model for probabilistic Bayesian formulation of the oncogenetic tree model [144, 74, 136, 8, 140]. Next generation sequencing makes SNVs data available in a population of cancer cells and the size of SNVs data makes it computationally difficult to infer a complete model of tumor progression. There are already a lot research being done to infer phylogenies from SNV frequencies [97, 129, 161].Subclone identification is required for inferring phylogenies based on SNV frequencies data. A Variational Bayes mixture and a Hidden Markov Model are proposed to identify the subclones [88, 43]. For inference of tumor phylogeny from gene expression data, Schwartz *et al.* [116, 141] proposed an unmixing of tumor samples and then used the inferred mixtures of individual tumor states to identify possible evolutionary relationships among tumor cells. Subramanian *et al.* [130] developed a pipeline and a novel Hidden Markov Model for inferring tumor phylogenies from the deconvoluted heterogeneous tumor samples. There are several other methods used to infer phylogenetic history of single tumors profiled at regional level [93, 78, 59].

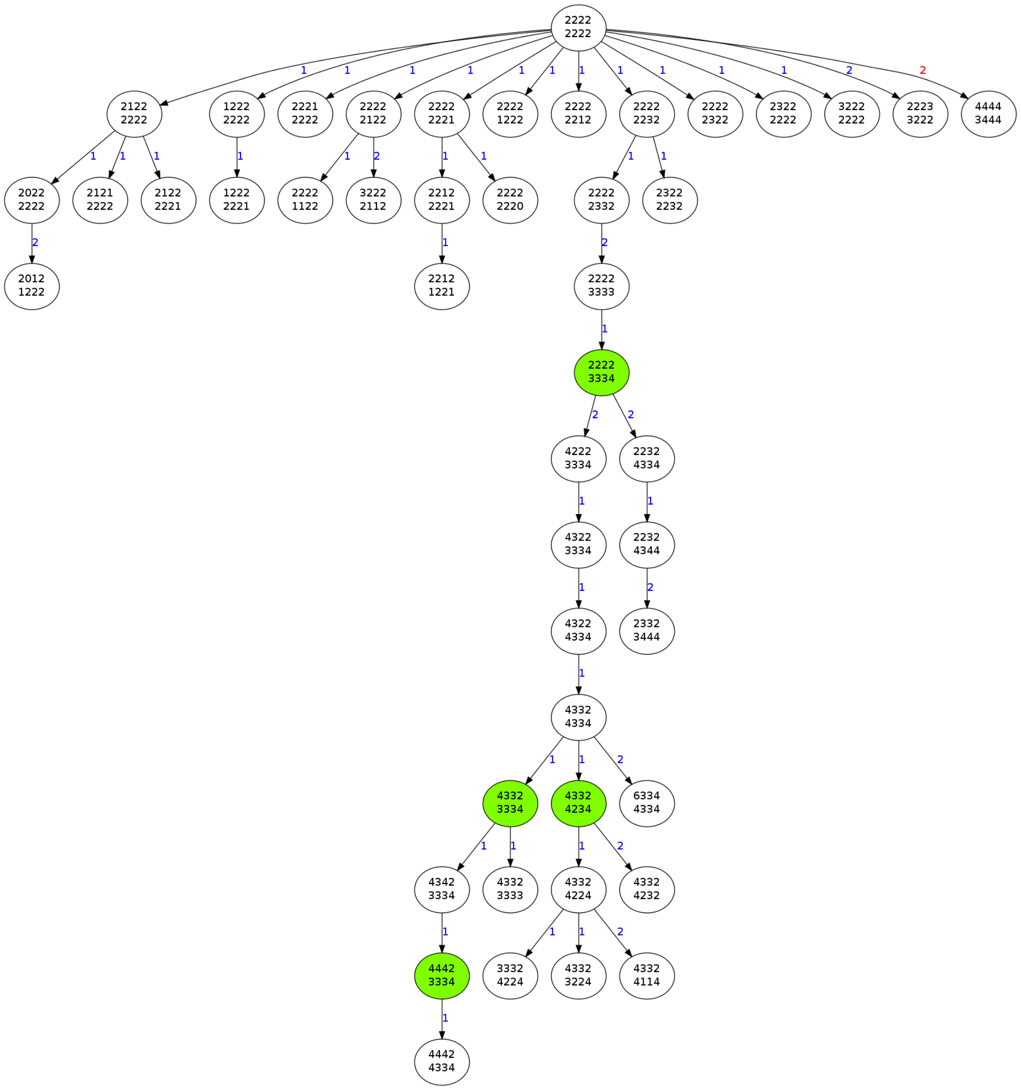
In this chapter, we explored the prognostic value of tumor phylogenies inferred by our newest mpFISHtree which can take care of large scale changes. For the experiments, we did tumor phylogenetic inference for the cervical and breast cancer data using DMST solver–mpFISHtree. The example tree can be seen in Fig. 6.2 which compares the primary and metastatic tumor sample from breast cancer patient 13. We can see that the tree from primary tumor is more balanced and has more nodes. The different topologies of the trees may prove that the cells in different state of



(a)

tumor development have different selective pressures.

Phylogenetic models is valuable for distinguishing driver genes from passengers in tumor genomic data. We tried to apply our mpFISHtree upon cervical and breast cancer to tentatively identify the driver or key changes which lead to metastasis. We tried to analysis of metastasis based on gene gain and loss pattern difference between primary and metastatic tumor. The major molecular factors during past research lead to metastasis are HOX7, EFGR, PDGFR, LAMP3 and several others [4, 75]. The other several factors except LAMP3 are universal helpful for all kind of tumor metastasis. LAMP3 overexpression is shown to be associated with an enhanced metastatic



(b)

Figure 6.2 Comparison between trees inferred from primary and metastatic tumor of patient 13 by using our DSMT solver–MPTtree.

potential and may be a prognostic factor for cervical cancer [75]. PROX1 is a key regulator of lymphatic endothelial cell commitment during embryonic development [28] while the function of CCND1 and PRKAA1 have population difference. For breast cancer with lung metastasis, overexpression of several gene markers (EREG, COX2, ID1, CXCL1, COX2, EREG and MMP1) are shown to have some correlation with metastasis [89]. MYC indirectly increase metastasis potential through activating microRNA miR-9 [84].

6.2 Methods and results

To identify the key changes leading to metastasis, We applied our algorithm on each of the samples in each of the datasets separately with considering of each SD, GD and WGD events. We show the boxplots for the inferred parameter values in the cervical cancer dataset in Fig. 6.3 for each marker. The boxplot for each marker is gain/loss ratio inferred from pairs of 16 primary and 16 metastatic samples comparison. We calculated the total number of times each gene showed a higher gain/loss ratio between the primary and metastatic samples out of the 16 pairs. We summarized the result in table 6.1, upper cervical cancer part. There are total 12 pairs of primary and 12 metastatic samples. We similarly show the boxplots for the inferred parameter values in the breast cancer datasets in Fig. 6.4 for the eight markers. We summarized the result in table 6.1, bottom breast cancer part.

From the cervical cancer result shown in Fig. 6.3 and Tab. 6.1, we can see that LAMP3 has the most significant difference between primary and metastasis dataset, 11 metastasis samples have higher gain/loss ratio while only 2 primary has higher ratio. Fig. 6.3 shows the ranges of the difference between metastasis cancer comparing to primary cancer. We examined whether the statistically significant results are consistently due to higher proportion of gain in primary or in metastatic samples. Out of 11 sample comparisons of LAMP3, seven were due to higher proportion of gain in metastasis and two were due to higher proportion of gain in primary samples. For PROX1, thirteen different samples were observed, 8 of them are due to higher proportion of gain in metastatic samples and five was due to higher proportion of gain in primary samples. This result suggests that "higher gain of LAMP3" is associated with the metastatic stage of cervical cancer, while "higher gain of PROX1" is associated with the primary tumor.

From the breast cancer result shown in Fig. 6.4 and Tab. 5.1, we can see that MYC has the most significant difference between primary and metastasis dataset, nine metastasis samples have higher gain/loss ratio while only one primary has higher ratio. Fig. 6.3 shows the ranges of the difference between metastasis cancer comparing to primary cancer. We examined whether the statistically significant results are consistently due to higher proportion of gain in primary or in metastatic samples. Out of ten sample comparisons of MYC, nine were due to higher proportion of gain in metastasis and one were due to higher proportion of gain in primary samples. For CDH1, CCND1, thirteen different samples were observed, eight of them are due to higher proportion of gain in metastatic samples and five was due to higher proportion of gain in primary samples. This result suggests that "higher gain of MYC" is associated with the metastatic stage of cervical cancer, while "higher gain of PROX1" and "higher gain of CCND1" are associated with the primary tumor.

6.3 Conclusion and discussion

We showed our developed algorithms for the problem of inferring tumor phylogeny and in this chapter we apply the mpFISHtree to single-tumor phylogenetic tree inference at the cellular level, with specific application to inferring multiscale copy number evolution from single-cell FISH data. We can see that by apply the new model to cervical and breast cancer, the potential driver or key changes can be found. So we can get the conclusion that the resulting models provide insight into tumor-specific

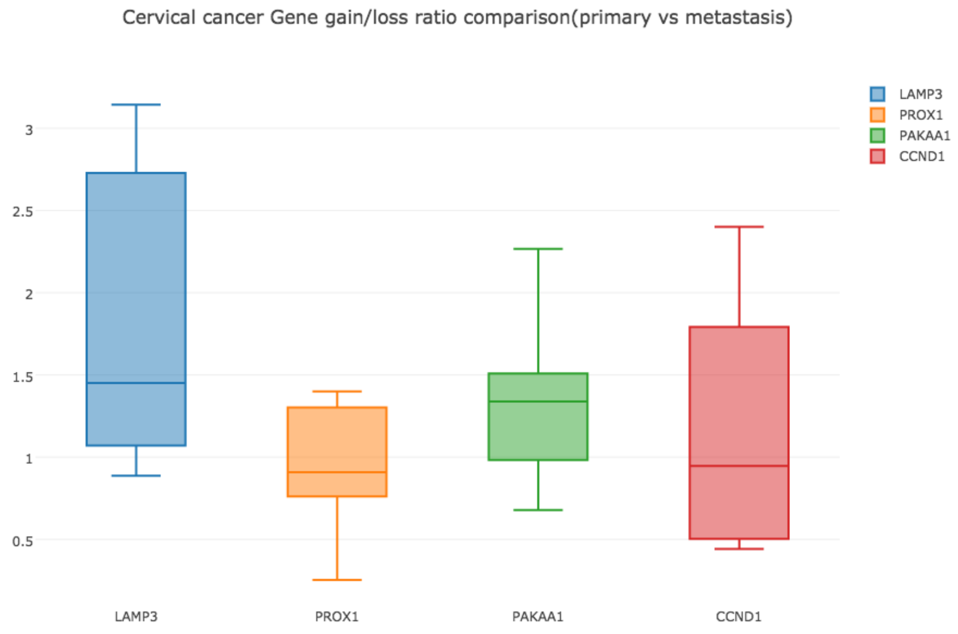


Figure 6.3 Comparison of 16 pairs of primary and metastasis cervical cancer data. Each boxplot represents the gain/loss ratio comparison between primary and metastasis tumor sample. The range are for the up to 16 data input from 16 samples.

Table 6.1 The count of samples with higher gain/loss ratio for certain gene marker(primary vs metastatic). The table contains four gene markers for cervical cancer and eight for breast cancer.

Cancer type	Gene marker	Primary	Metastatic
Cervical	LAMP3	2	**11**
	PROX1	8	5
	PRKAA1	7	5
	CCND1	6	6
Breast	COX-2	5	7
	CCND1	8	4
	HER-2	4	6
	ZNF217	2	7
	DBC2	3	8
	CDH1	7	3
	p53	5	6
	MYC	1	**9**

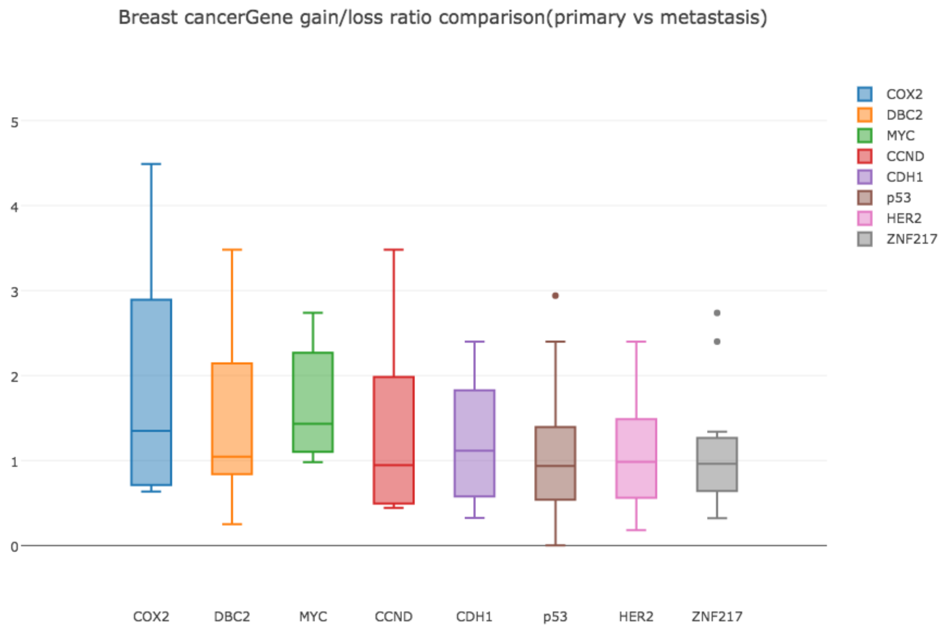


Figure 6.4 Comparison of 12 pairs of primary and metastasis breast cancer data. Each boxplot represents the gain/loss ratio comparison between primary and metastasis tumor sample. The range are for the up to 12 data input from 12 samples.

variation and lead to improved prediction of future tumor progression in multiple tumor types.

Chapter 7

Application of Phylogenetic Analysis on CNV data

7.1 Background

There are currently two data sources for clean single cell data source: fluorescence in situ hybridization (FISH) and single cell sequencing. FISH data is collected based on FISH (Fluorescent In Situ Hybridization) technology which shows presence or absence of specific DNA sequence in a cell. It has the advantage that it yields single-cell resolution profiles instead of average information of a group of cells. Pennington *et al.* proposed couple of methods to infer phylogenetic inference of single tumors by using FISH data in single tumors [101, 103, 102]. Martins *et al.* developed computational methods to predict the temporal order of somatic events of the genes BRCA1, PTEN, and p53 using FISH data from 55 BRCA1 breast cancers. For single cell sequencing data, there are also a lot of research has been done. Navin *et al.* [94] used the neighbor joining algorithm to infer evolutionary histories of cancer lineages on low-coverage single nucleus sequencing. Xu *et al.* [156] found common mutations among cells and also significant difference between mutation pattern between cancer cells in the tumor by using single cell exome sequencing. Hou *et al.* [66] found that this neoplasm of a JAK2-negative myeloproliferative neoplasm patient represented a monoclonal evolution by whole-exome single cell sequencing. FISH data has the shortcoming that it is not high-through output data which make it hard to be generally used, while it is hard get enough cell sample for single sequencing data for technology limitations. So we need to modify our method to other more general available dataset such as data got by next generation sequencing technology. Fast growing new technology such as high throughput genomic, epigenetic and proteomic sequencing gave researchers a lot options. [105, 77, 51]. Gene duplications and copy number variation is important in understanding gene function, such as along the Y-chromosome of *Rattus* where they caused elevation of blood pressure and affected kidney function [106]. In addition, whole genome duplication is a major evolutionary event which occurred in several eukaryotic lineages, including plants, fungi and animals [73]. In particular, the monocot and core eudicot ancestor shows this kind of major event [30]. Likewise for animals, recent convincing evidence has shown that the whole genome duplication took place in the common ancestor of all extant teleosts [52].

Rearrangement events in cancer lead to frequent and widespread gains and losses of genomic material which is markedly different from that observed at the species or population level [3]. Cancers, such as ovarian cancer, typically display a large number of genomic rearrangement events which may lead to copy number variation (CNV) in genes and large genomic segments, and is a source for analysis of evolutionary relationships of cancer genomes in their "ecological" context of space and time [121, 160]. During tumor development, copy number can also increase or decrease, mainly from failures of DNA repair mechanisms (e.g., translesion synthesis and non-homologous end joining) [34, 115, 131, 17, 29, 113, 23, 152]. We can represent CNV data as a two-dimensional matrix where each row represents a cancer genome and each column represents a marker; a value of the matrix corresponds to the number of copies of a marker in a genome.

Epithelial ovarian cancer is considered a heterogeneous cancer as it comprises at least five distinct histological subtypes, the most common and well-studied ovarian cancer is high-grade serous ovarian cancer(**HGSOC**) [146]. Data from TCGA Research Network greatly helped in understanding the genetic composition of HGSOC where genomic, epigenomic, and transcriptomic expression data from 489 patients with HGSOC were analyzed for mRNA, microRNA, DNA copy number, and sites of DNA methylation. Ovarian cancer is shown to have a low number of non-synonymous mutations and high level of somatic copy number alterations by the studies using paired-end next generation sequencing and high-density SNP arrays comparing to other tumor types [76, 96].

Recently, Schwarz *et al.* proposed a method called MEDICC to jointly solve the problems of phasing (infer major and minor copy-number) and tree reconstruction based on the minimum evolution criterion [118]. The MEDICC method is developed upon copy number variation data, which is the average count for each gene and are widely used for biological research. We decided to do some tentative research using CNV data. Schwarz *et al.* [117] tried to correlate level of tumor heterogeneity with clinical outcome. Copy number information from 17 patient with HGSOC (135 metastatic sites) at the time of diagnosis and following chemotherapy. They designed a tool named minimum event distance for intratumor copy number comparisons (**MEDICC**) to build an evolutionary tree for each patient with at least three tumor samples. 7.1 briefly shows the process of their experiment and the resulting tree by using neighbor-joining. Their research endorsed significant intratumor heterogeneity and most tumors shows a branched evolutionary pattern. But we think the tool MEDICC is not ideal to build phylogenetic tree for the patients compared to our tool and thus can lead to poor prediction for key point of cancer emergence and right treatment strategy. 7.2 shows two examples of spatial and temporal heterogeneity by using MEDICC which show strong overall conservation. We extend our tools to deal with CNV data to show that our tool is able to deal with highthroughput data.

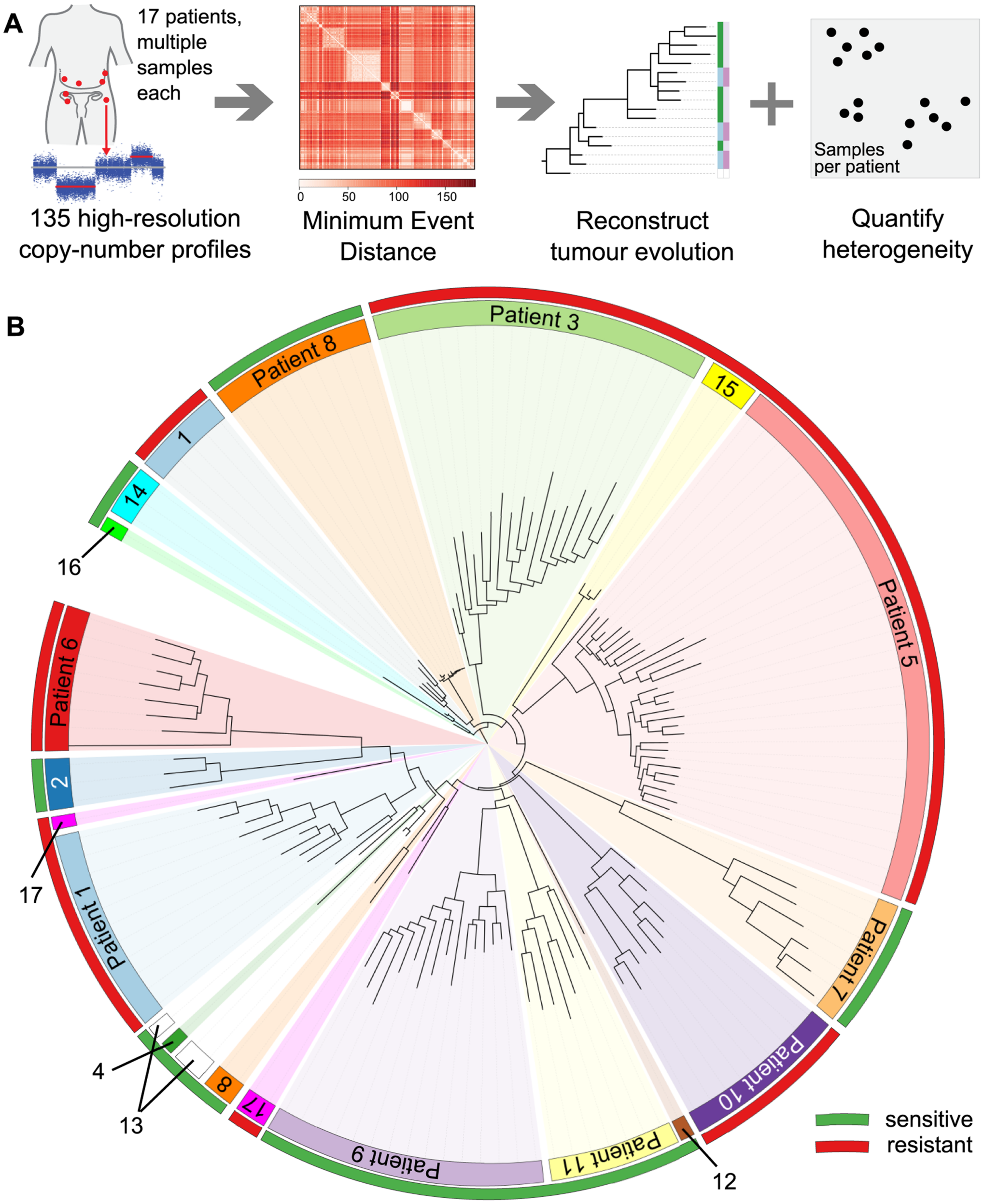


Figure 7.1 Overview of Schwarz *et al.*'s analysis on the clinical data. (A) Profiles from 135 metastatic site from 17 patients with HGSOC, the process of MEDICC algorithm, tumor evolution phylogeny and quantify heterogeneity. (B) Significant patient-specific intra-tumor genetic heterogeneity based on neighbor-joining method. The outer circular bar indicates resistant versus sensitive to treatment based on survival: red, resistant; green, sensitive.

7.2 Methods

We tried to apply our tool upon CNV data get from Schwarz *et al.*'s work. Before we apply our tool to the dataset, we first did some data treatment. Since the input data

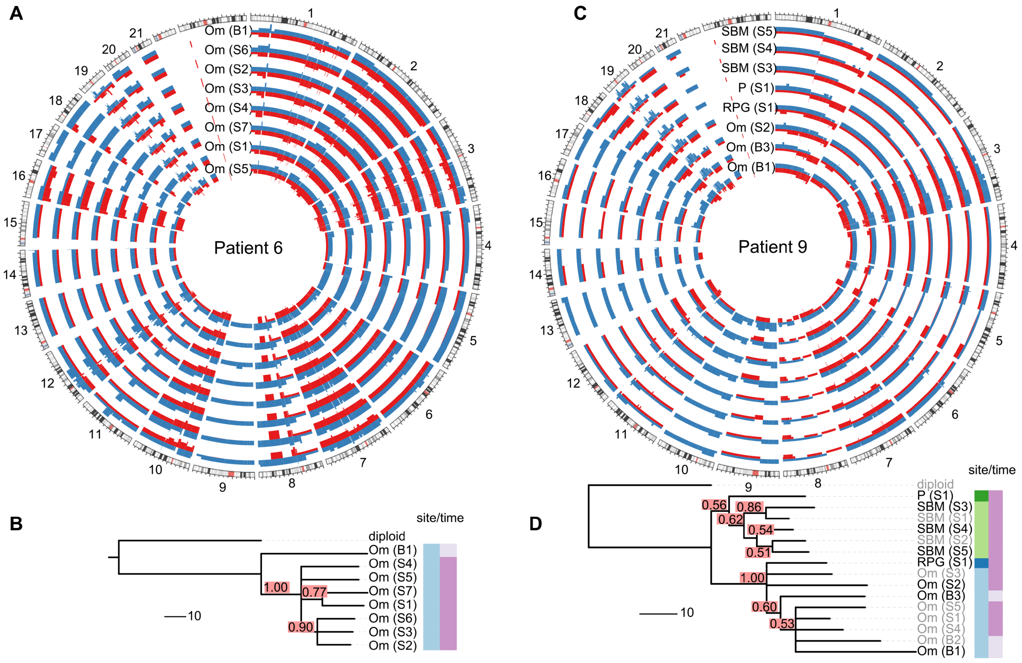
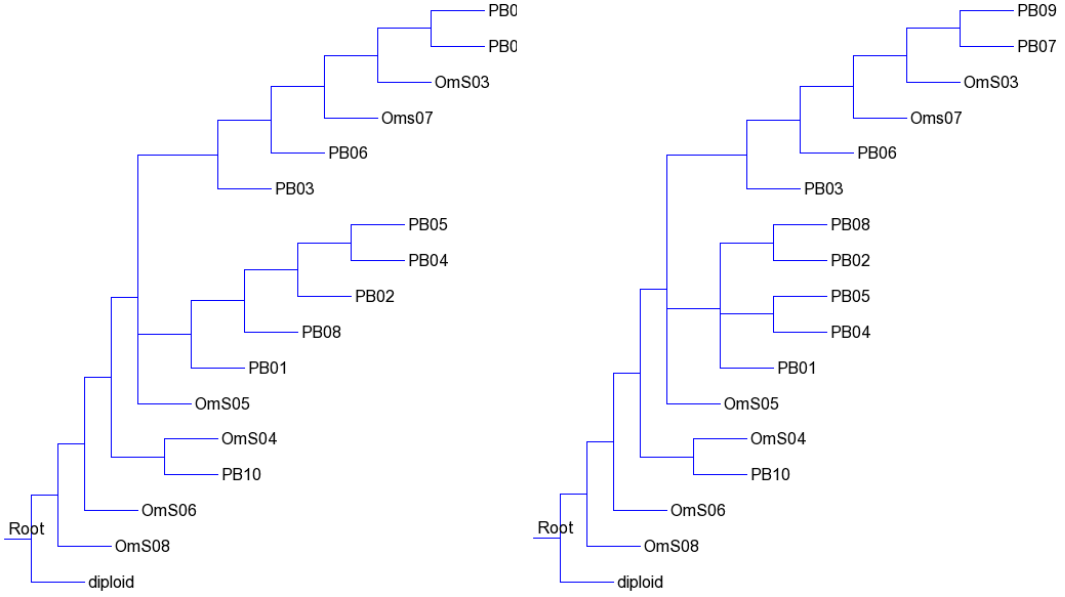


Figure 7.2 Overview of Schwarz *et al.*'s analysis on the clinical data. (A) Profiles from 135 metastatic site from 17 patients with HGSOC, the process of MEDICC algorithm, tumor evolution phylogeny and quantify heterogeneity. (B) Significant patient-specific intra-tumor genetic heterogeneity based on neighbor-joining method. The outer circular bar indicates resistant versus sensitive to treatment based on survival: red, resistant; green, sensitive.

has hundreds of gene markers, we tried to reduce some redundancy information as our tool work better with fewer gene markers. First we tried to combines the genes which always change together as one gene. Secondly, we tried to limit the maximum gene copy number allowed. We limit the maximum copy number to be 30 as the most gene copy number is within this limit. Then we did tree inference upon the cleaned dataset and compared the trees inferred by our tool with the trees from MEDICC. There are totally 17 patient data samples in their dataset.

7.3 Result and Conclusion

We showed one example result tree using MEDICC and mpFISHtree in Fig. 7.3 and its RF distance is just 1 which should be considered quite similar to each other. We



(a)

(b)

Figure 7.3 Comparison between trees inferred by our mpFISHtree and MEDICC of patient 1. The distance between the two trees are just 1 which are almost the same.

got the same tree in fourteen sample and the RF distance for the other three data samples are all 1. So we can see that our tool can be easily extend to be applied on CNV data and the trees inferred by our tool is worth trust. Thanks to the high speed of tnt, the speed of our tool is much faster than MEDICC.

Chapter 8

Conclusions and Future Directions

8.1 Thesis conclusions

There have been various kinds of phylogenetic tools for gene order data and they can be divided into event-based and probability-based methods. But there is no detailed work being done to assess the quality of inferred ancestral genomes. Since isolating and inversions are shown to produce a better ROC curve for PMAG and GapAdj compared to other resampling schemes, we think that both isolating and inversion are better resampling methods for adjacency-based ancestral reconstruction method. We think that jackknifing, isolating, and inversion are all suitable resampling methods for event-based parsimony methods such as GASTS. We also show that the best threshold for different resampling schemes is around 75% from the FP and FN analysis. An evolutionary understanding of tumor has long been recognized as important for studying heterogeneity and problems in cancer treatment, such as, drug resistance after treatment. There are already a lot work done for tumor phylogenetic study. however, it is such a hard question that there is still strong eager for reliable tools to be developed. In this thesis, we aim to implement novel algorithms and tools to infer the underlying tumor evolutionary history. We first tried to solve the DSMT problem which only involved single gene copy number change. To find the parsimony phylogeny with missing nodes during evolutionary, steiner node, we start from two direction, minimum spanning tree and maximum parsimony tree. In Chapter 2, we solved the probe from minimum spanning tree side. We first built minimum spanning tree out of the input nodes which stands for the cell types. Then identify all the potential steiner nodes which can reduce the minimum spanning tree weight by it themselves. Finally, we tried to add the nodes into the minimum spanning tree in an optimized order when they can reduce the tree weight. In Chapter 3, we tried to solve the problem from maximum parsimony tree side. We first treat all the input data as leaf nodes to build the phylogenetic tree. Then we further retract the zero weight edges which also remove all the duplicate or redundant nodes. To further fit for the parsimony purpose, for each retraction, we chose the one leads to fewer steiner nodes. To accomplish the purpose of also considering large scale changes, we employed the large scale change edges which can be found by Chowdhury's work. After we found edges with large scale changes, we divide the cell patterns into different groups and apply maximum parsimony phylogenetic inference method to each group and further link each subgroup to form the result tree by linking the edges with large scale changes. All the tools developed by us are shown to work well on real data, including breast and cervical cancer, and simulation data comparing to earlier methods. Applying the new methods on the real tumor data resulted in inference of more parsimonious models of tumor progression comparing to earlier methods. We used the our new tool for cancer stage differentiate and found key remarks or driver gene which can be potentially used to improve tumor diagnosis and prognosis. We also tried to extend our work on highthroughput data. Our mpFISHtree gave similar tree compared to MEDICC, the new software designed for phylogenetic inference on cancer CNV data while the speed of our tool is much faster than MEDICC.

8.2 Future directions

The work in this thesis makes some contribution towards scalable algorithms for phylogenetic inference for single tumors. As the same as the work of Chowdhury *et al.*, the algorithms developed in this thesis currently consider only single gene, chromosome level and whole genome level changes. However, gene copy number change events can happen with other mechanisms too, such as all kinds of rearrangement events, chromothripsis, chromoplexy etc. The evolutionary model should be further extended to consider these and other mutational mechanisms by which copy number profiles of tumor cells by combining with some other data types, such as pair end data. Fig. 8.1 shows an example of possible change history of cancer evolutionary history. The work presented in this thesis focused on FISH data. Although FISH is currently the only technology to reliably profile gene copy numbers across hundreds to thousands of cells per patient in sizable patient populations, it has its own limitations: limitation on number of genetic markers and limitation on data types. Other types of variations, such as single nucleotide variations, which are also important in the evolution of tumors, cannot be captured using FISH. Although single cell sequencing technology is still far from becoming practical for the number of cells needed for the questions we examine. However, that it will eventually become the dominant technology for single cell cancer studies. There would thus be value in extending the theory developed here to single cell sequencing data. Since our tools can handle larger number and variety of markers, the main problem would be the noise arising from single-cell amplification. There are also some other data types such as whole genome sequencing data which our tools can be extend to apply on. Although we tried it on CNV data, it is far away from enough since it is just tentative trial.

In the thesis, we have taken some first steps towards using tumor phylogenetics as a source of features to predict driver gene changes. While these results show the promise of these directions, they also suggest many avenues for improvement. In the future we will try to apply on other data types for more important biological findings. Successful cancer treatment depends on future evolutionary trajectory of a cancer especially after drug treatment. The phylogeny inferred by our tool makes advances in predicting future evolutionary trajectories of individual tumors as some to

	Marker A	Marker B	Marker C
Copy Number Profile 1	2	2	2
Copy Number Profile 2	2	1	3
Copy Number Profile 3	2	1	2

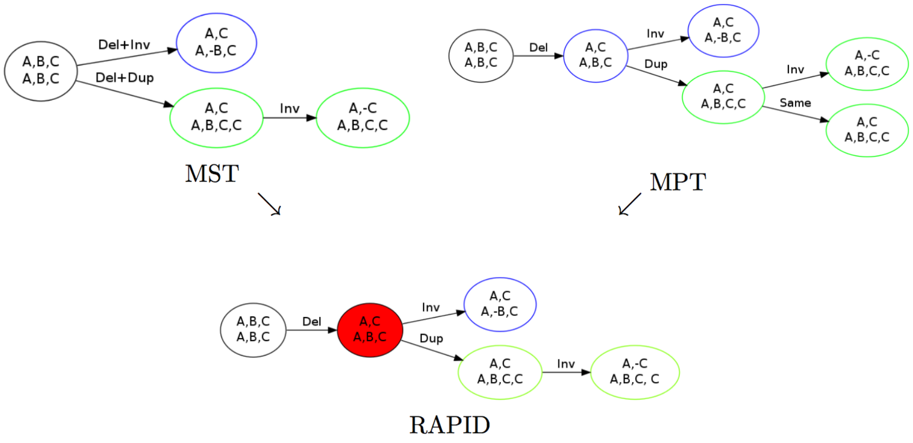


Figure 8.1 One possible phylogenetic history considering rearrangements.

data cell pattern can be treated as future cell pattern. The phylogenies inferred also suggests important feature of the process how tumor is evolving such as driver gene. In summary, our work represents an important development in tumor phylogenetic study as a new source of clinical guidance for individualized patients.

Bibliography

[1] Max A Alekseyev and Pavel A Pevzner. “Breakpoint graphs and ancestral genome reconstructions”. In: *Genome research* 19.5 (2009), pp. 943–957.

[2] Kristina Anderson et al. “Genetic variegation of clonal architecture and propagating cells in leukaemia”. In: *Nature* 469.7330 (2011), pp. 356–361.

[3] Camille Stephan-Otto Attolini and Franziska Michor. “Evolutionary theory of cancer”. In: *Annals of the New York Academy of Sciences* 1168.1 (2009), pp. 23–51.

[4] SW Aziz and MH Aziz. “Cervical Cancer Metastasis”. In: *Introduction to Cancer Metastasis* 3.100,000 (2016), p. 77.

[5] Sylvan C Baca et al. “Punctuated evolution of prostate cancer genomes”. In: *Cell* 153.3 (2013), pp. 666–677.

[6] Michael Baudis. “Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data”. In: *BMC cancer* 7.1 (2007), p. 226.

[7] Niko Beerenwinkel et al. “Cancer evolution: mathematical models and computational inference”. In: *Systematic biology* 64.1 (2015), e1–e25.

[8] Niko Beerenwinkel et al. “Learning multiple evolutionary pathways from cross-sectional data”. In: *Journal of computational biology* 12.6 (2005), pp. 584–598.

[9] Priscila Biller, Pedro Feijão, and João Meidanis. “Rearrangement-based phylogeny using the Single-Cut-or-Join operation”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 10.1 (2013), pp. 122–134.

[10] Walter Birchmeier and Jürgen Behrens. “Cadherin expression in carcinomas: role in the formation of cell junctions and the prevention of invasiveness”. In: *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1198.1 (1994), pp. 11–26.

[11] Mathieu Blanchette, Guillaume Bourque, and David Sankoff. “Breakpoint phylogenies”. In: *Genome Informatics* 8 (1997), pp. 25–34.

[12] Archie Bleyer and H Gilbert Welch. “Effect of three decades of screening mammography on breast-cancer incidence”. In: *New England Journal of Medicine* 367.21 (2012), pp. 1998–2005.

[13] Christine Bouchardy et al. “Undertreatment strongly decreases prognosis of breast cancer in elderly women”. In: *Journal of clinical oncology* 21.19 (2003), pp. 3580–3587.

[14] Guillaume Bourque and Pavel A Pevzner. “Genome-scale evolution: reconstructing gene orders in the ancestral species”. In: *Genome research* 12.1 (2002), pp. 26–36.

[15] David Bryant. “The complexity of the breakpoint median problem”. In: *Centre de recherches mathematiques* (1998).

[16] Robin M Bush et al. “Predicting the evolution of human influenza A”. In: *Science* 286.5446 (1999), pp. 1921–1925.

[17] Keith W Caldecott. “Single-strand break repair and genetic disease”. In: *Nature Reviews Genetics* 9.8 (2008), pp. 619–631.

[18] Peter J Campbell et al. “The patterns and dynamics of genomic instability in metastatic pancreatic cancer”. In: *Nature* 467.7319 (2010), pp. 1109–1113.

[19] Alberto Caprara. “Formulations and hardness of multiple sorting by reversals”. In: *Proceedings of the third annual international conference on Computational molecular biology*. ACM. 1999, pp. 84–93.

[20] Alberto Caprara. “On the practical solution of the reversal median problem”. In: *Algorithms in Bioinformatics*. Springer, 2001, pp. 238–251.

[21] Daniele Catanzaro et al. “Classifying the Progression of Ductal Carcinoma from Single-Cell Sampled Data via Integer Linear Programming: A Case Study”. In: (2015).

[22] Belinda SW Chang and Michael J Donoghue. “Recreating ancestral proteins”. In: *Trends in ecology & evolution* 15.3 (2000), pp. 109–114.

[23] J Ross Chapman, Martin RG Taylor, and Simon J Boulton. “Playing the end game: DNA double-strand break repair pathway choice”. In: *Molecular cell* 47.4 (2012), pp. 497–510.

[24] Paul B Chapman et al. “Improved survival with vemurafenib in melanoma with BRAF V600E mutation”. In: *n Engl J Med* 2011.364 (2011), pp. 2507–2516.

[25] Yu-Kang Cheng et al. “A mathematical methodology for determining the temporal order of pathway alterations arising during gliomagenesis”. In: *PLoS computational biology* 8.1 (2012), e1002337.

[26] Salim Akhter Chowdhury et al. “Algorithms to model single gene, single chromosome, and whole genome copy number changes jointly in tumor phylogenetics”. In: *PLoS Comput Biol* 10.7 (2014), e1003740.

[27] Salim Akhter Chowdhury et al. “Phylogenetic analysis of multiprobe fluorescence in situ hybridization data from tumor cell populations”. In: *Bioinformatics* 29.13 (2013), pp. i189–i198.

[28] Anca Maria Cimpean et al. “Prox 1, VEGF-C and VEGFR3 expression during cervical neoplasia progression as evidence of an early lymphangiogenic switch.” In: *Histology and histopathology* 27.12 (2012), pp. 1543–1550.

[29] James E Cleaver, Ernest T Lam, and Ingrid Revet. “Disorders of nucleotide excision repair: the genetic and molecular basis of heterogeneity”. In: *Nature Reviews Genetics* 10.11 (2009), pp. 756–768.

[30] Avril Coghlan et al. “Chromosome evolution in eukaryotes: a multi-kingdom perspective”. In: *TRENDS in Genetics* 21.12 (2005), pp. 673–682.

[31] L. Cui et al. “Adaptive evolution of chloroplast genome structure inferred using a parametric bootstrap approach”. In: *BMC Evol. Biol* 6.13 (2006).

[32] William HE Day. “Computational complexity of inferring phylogenies from dissimilarity matrices”. In: *Bulletin of mathematical biology* 49.4 (1987), pp. 461–467.

[33] JS De Bono and Alan Ashworth. “Translating cancer research into targeted therapeutics”. In: *Nature* 467.7315 (2010), pp. 543–549.

[34] Rinne De Bont and Nik Van Larebeke. “Endogenous DNA damage in humans: a review of quantitative data”. In: *Mutagenesis* 19.3 (2004), pp. 169–185.

[35] Richard Desper et al. “Distance-based reconstruction of tree models for oncogenesis”. In: *Journal of Computational Biology* 7.6 (2000), pp. 789–803.

[36] Richard Desper et al. “Inferring tree models for oncogenesis from comparative genome hybridization data”. In: *Journal of computational biology* 6.1 (1999), pp. 37–51.

[37] Jack Edmonds. “Optimum branchings”. In: *Journal of Research of the National Bureau of Standards B* 71.4 (1967), pp. 233–240.

[38] Eric R Fearon, Bert Vogelstein, et al. “A genetic model for colorectal tumorigenesis”. In: *Cell* 61.5 (1990), pp. 759–767.

[39] Joseph Felsenstein et al. “Confidence limits on phylogenies: an approach using the bootstrap”. In: *Evolution* 39.4 (1985), pp. 783–791.

[40] Joseph Felsenstein and Joseph Felenstein. “Inferring phylogenies”. In: (2004).

[41] Guillaume Fertin. *Combinatorics of genome rearrangements*. MIT press, 2009.

[42] G. Fertin et al. *Combinatorics of Genome Rearrangements*. MIT Press, 2009.

[43] Andrej Fischer et al. “High-definition reconstruction of clonal composition in cancer”. In: *Cell reports* 7.5 (2014), pp. 1740–1752.

[44] R Fisher, L Pusztai, and C Swanton. “Cancer heterogeneity: implications for targeted therapeutics”. In: *British journal of cancer* 108.3 (2013), pp. 479–485.

[45] Maofu Fu et al. “Minireview: Cyclin D1: normal and abnormal functions”. In: *Endocrinology* 145.12 (2004), pp. 5439–5447.

[46] P Andrew Futreal et al. “A census of human cancer genes”. In: *Nature Reviews Cancer* 4.3 (2004), pp. 177–183.

[47] Yves Gagnon, Mathieu Blanchette, and Nadia El-Mabrouk. “A flexible ancestral genome reconstruction method based on gapped adjacencies”. In: *BMC bioinformatics* 13.Suppl 19 (2012), S4.

[48] Michael R Garey and David S. Johnson. “The rectilinear Steiner tree problem is NP-complete”. In: *SIAM Journal on Applied Mathematics* 32.4 (1977), pp. 826–834.

[49] Marco Gerlinger et al. “Intratumor heterogeneity and branched evolution revealed by multiregion sequencing”. In: *N Engl j Med* 2012.366 (2012), pp. 883–892.

[50] Moritz Gerstung et al. “Quantifying cancer progression with conjunctive Bayesian networks”. In: *Bioinformatics* 25.21 (2009), pp. 2809–2815.

[51] Robert J Gillies, Daniel Verduzco, and Robert A Gatenby. “Evolutionary dynamics of carcinogenesis and why targeted therapy does not work”. In: *Nature Reviews Cancer* 12.7 (2012), pp. 487–493.

[52] Stella MK Glasauer and Stephan CF Neuhauss. “Whole-genome duplication in teleost fishes and its evolutionary consequences”. In: *Molecular Genetics and Genomics* 289.6 (2014), pp. 1045–1060.

[53] Pablo A Goloboff, James S Farris, and Kevin C Nixon. “TNT, a free program for phylogenetic analysis”. In: *Cladistics* 24.5 (2008), pp. 774–786.

[54] Pablo A Goloboff, Camilo I Mattoni, and Andres Sebastian Quinteros. “Continuous characters analyzed as such”. In: *Cladistics* 22.6 (2006), pp. 589–601.

[55] Pablo Goloboff, Steve Farris, and Kevin Nixon. “TNT (Tree analysis using New Technology).” In: (2000).

[56] Jonathan L Gordon, Kevin P Byrne, and Kenneth H Wolfe. “Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern Saccharomyces cerevisiae genome”. In: *PLoS Genet* 5.5 (2009), e1000485–e1000485.

[57] Mel Greaves and Carlo C Maley. “Clonal evolution in cancer”. In: *Nature* 481.7381 (2012), pp. 306–313.

[58] MS Greenblatt et al. “Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis”. In: *Cancer research* 54.18 (1994), pp. 4855–4878.

[59] Chris D Greenman et al. “Estimation of rearrangement phylogeny for cancer genomes”. In: *Genome research* 22.2 (2012), pp. 346–361.

[60] Masaaki Hamaguchi et al. “DBC2, a candidate for a tumor suppressor gene involved in breast cancer”. In: *Proceedings of the National Academy of Sciences* 99.21 (2002), pp. 13647–13652.

[61] Douglas Hanahan and Robert A Weinberg. “Hallmarks of cancer: the next generation”. In: *cell* 144.5 (2011), pp. 646–674.

[62] Douglas Hanahan and Robert A Weinberg. “The hallmarks of cancer”. In: *cell* 100.1 (2000), pp. 57–70.

[63] S. Hannenhalli and P.A. Pevzner. “Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals)”. In: 1995, pp. 178–189.

[64] Kerstin Heselmeyer-Haddad et al. “Single-cell genetic analysis of ductal carcinoma in situ and invasive breast cancer reveals enormous tumor heterogeneity yet conserved genomic imbalances and gain of MYC during progression”. In: *The American journal of pathology* 181.5 (2012), pp. 1807–1822.

[65] Susan Holmes. “Bootstrapping phylogenetic trees: theory and methods”. In: *Statistical Science* (2003), pp. 241–255.

[66] Yong Hou et al. “Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm”. In: *Cell* 148.5 (2012), pp. 873–885.

[67] LR Howe et al. “Cyclooxygenase-2: a target for the prevention and treatment of breast cancer.” In: *Endocrine-related cancer* 8.2 (2001), pp. 97–114.

[68] Fei Hu. “Probabilistic Reconstruction of Phylogeny and Ancestral Genomes from Gene Order Data”. In: (2013).

[69] Fei Hu, Lingxi Zhou, and Jijun Tang. “Reconstructing ancestral genomic orders using binary encoding and probabilistic models”. In: *Bioinformatics Research and Applications*. Springer, 2013, pp. 17–27.

[70] Fei Hu et al. “Probabilistic reconstruction of ancestral gene orders with insertions and deletions”. In: *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* 11.4 (2014), pp. 667–672.

[71] J. Huan et al. “Reconstruction of ancestral gene order following large-scale genome duplication and gene loss.” In: *Proc. 2nd IEEE Comput. Systems Bioinfo. Conf. (CSB’03)*. IEEE Press.IEEE Press, 2003, pp. 484–485.

[72] Fung Yu Huang et al. “Semi-quantitative fluorescent PCR analysis identifies PRKAA1 on chromosome 5 as a potential candidate cancer gene of cervical cancer”. In: *Gynecologic oncology* 103.1 (2006), pp. 219–225.

[73] Olivier Jaillon, Jean-Marc Aury, and Patrick Wincker. “"Changing by doubling", the impact of Whole Genome Duplications in the evolution of eukaryotes”. In: *Comptes rendus biologies* 332.2 (2009), pp. 241–253.

[74] Feng Jiang et al. “Construction of evolutionary tree models for renal cell carcinoma from comparative genomic hybridization data”. In: *Cancer research* 60.22 (2000), pp. 6503–6509.

[75] Hiroyuki Kanao et al. “Overexpression of LAMP3/TSC403/DC-LAMP promotes metastasis in uterine cervical cancer”. In: *Cancer research* 65.19 (2005), pp. 8640–8645.

[76] Cyriac Kandoth et al. “Mutational landscape and significance across 12 major cancer types”. In: *Nature* 502.7471 (2013), pp. 333–339.

[77] Antonija Kreso and John E Dick. “Evolution of the cancer stem cell model”. In: *Cell stem cell* 14.3 (2014), pp. 275–291.

[78] Eric Letouzé et al. “Analysis of the copy number profiles of several tumor samples from the same patient reveals the successive steps in tumorigenesis”. In: *Genome biology* 11.7 (2010), R76.

[79] Yu Lin, Vaibhav Rajan, and Bernard ME Moret. “Fast and accurate phylogenetic reconstruction from high-resolution whole-genome data and a novel robustness estimator”. In: *Journal of Computational Biology* 18.9 (2011), pp. 1131–1139.

[80] Yu Lin, Vaibhav Rajan, Bernard ME Moret, et al. “Bootstrapping phylogenies inferred from rearrangement data.” In: *Algorithms for Molecular Biology* 7.1 (2012).

[81] Y. Lin et al. “Maximum likelihood phylogenetic reconstruction from high-resolution whole-genome data and a tree of 68 eukaryotes”. In: World Scientific Pub., 2013, pp. 285–296.

[82] Jian Ma. “A probabilistic framework for inferring ancestral genomic orders”. In: *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on*. IEEE. 2010, pp. 179–184.

[83] Jian Ma et al. “Reconstructing contiguous regions of an ancestral genome”. In: *Genome Research* 16.12 (2006), pp. 1557–1565.

[84] Li Ma et al. “miR-9, a MYC/MYCN-activated microRNA, regulates E-cadherin and cancer metastasis”. In: *Nature cell biology* 12.3 (2010), pp. 247–256.

[85] Marco A Marra et al. “The genome sequence of the SARS-associated coronavirus”. In: *Science* 300.5624 (2003), pp. 1399–1404.

[86] Filipe C Martins et al. “Evolutionary pathways in BRCA1-associated breast tumors”. In: *Cancer discovery* 2.6 (2012), pp. 503–511.

[87] Igor Matushansky et al. “A developmental model of sarcomagenesis defines a differentiation-based classification for liposarcomas”. In: *The American journal of pathology* 172.4 (2008), pp. 1069–1080.

[88] Christopher A Miller et al. "SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution". In: *PLoS Comput Biol* 10.8 (2014), e1003665.

[89] Andy J Minn et al. "Genes that mediate breast cancer metastasis to lung". In: *Nature* 436.7050 (2005), pp. 518–524.

[90] Navodit Misra et al. "Generalized buneman pruning for inferring the most parsimonious multi-state phylogeny". In: *Research in Computational Molecular Biology*. Springer. 2010, pp. 369–383.

[91] Bernard ME Moret et al. "New approaches for reconstructing phylogenies from gene order data". In: *Bioinformatics* 17.suppl 1 (2001), S165–S173.

[92] K Müller et al. "Evolution of carnivory in Lentibulariaceae and the Lamiales". In: *Plant Biology* 6.4 (2004), pp. 477–490.

[93] Nicholas E Navin and James Hicks. "Tracing the tumor lineage". In: *Molecular oncology* 4.3 (2010), pp. 267–283.

[94] Nicholas Navin et al. "Inferring tumor progression from genomic heterogeneity". In: *Genome research* 20.1 (2010), pp. 68–80.

[95] Nicholas Navin et al. "Tumour evolution inferred by single-cell sequencing". In: *Nature* 472.7341 (2011), pp. 90–94.

[96] Cancer Genome Atlas Research Network et al. "Integrated genomic analyses of ovarian carcinoma". In: *Nature* 474.7353 (2011), pp. 609–615.

[97] Serena Nik-Zainal et al. "Mutational processes molding the genomes of 21 breast cancers". In: *Cell* 149.5 (2012), pp. 979–993.

[98] Genevieve H Nonet et al. "The ZNF217 gene amplified in breast cancers promotes immortalization of human mammary epithelial cells". In: *Cancer research* 61.4 (2001), pp. 1250–1254.

[99] Lior Pachter and Bernd Sturmfels. "The mathematics of phylogenomics". In: *SIAM review* 49.1 (2007), pp. 3–31.

[100] Mark D Pegram, Gottfried Konecny, and Dennis J Slamon. "The molecular and cellular biology of HER2/neu gene amplification/overexpression and the clinical development of herceptin (trastuzumab) therapy for breast cancer". In: *Advances in Breast Cancer Management*. Springer, 2000, pp. 57–75.

[101] Gregory Pennington, Stanley Shackney, and Russell Schwartz. “Cancer phylogenetics from single-cell assays”. In: *Dept. Comput. Sci., Pittsburgh, PA, USA, Carnegie Mellon Univ., Tech. Rep. CMU-CS-06-103* (2006).

[102] Gregory Pennington et al. “Reconstructing tumor phylogenies from heterogeneous single-cell data”. In: *Journal of bioinformatics and computational biology* 5.02a (2007), pp. 407–427.

[103] G Pennington et al. “Expectation-maximization method for reconstructing tumor phylogenies from single-cell data”. In: *Computational Systems Bioinformatics Conference (CSB).* 2006, pp. 371–380.

[104] Erin D Pleasance et al. “A comprehensive catalogue of somatic mutations from a human cancer genome”. In: *Nature* 463.7278 (2009), pp. 191–196.

[105] Kornelia Polyak and Andriy Marusyk. “Cancer: clonal cooperation”. In: *Nature* 508.7494 (2014), pp. 52–53.

[106] Jeremy Prokop et al. “Gene Duplication and Sequence Variants of the Rattus norvegicus Y-Chromosome can Alter Kidney Function”. In: *The FASEB Journal* 29.1 Supplement (2015), pp. 665–11.

[107] Maurice H Quenouille. “Approximate tests of correlation in time-series”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 11.1 (1949), pp. 68–84.

[108] Vaibhav Rajan et al. “Heuristics for the inversion median problem”. In: *BMC bioinformatics* 11.Suppl 1 (2010), S30.

[109] A. Caprara. “The Reversal Median Problem”. In: *INFORMS J. Computing* 15.1 (2003), pp. 93–113.

[110] Markus Riester et al. “A differentiation-based phylogeny of cancer subtypes”. In: (2010).

[111] Howard A Ross and Allen G Rodrigo. “Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration”. In: *Journal of virology* 76.22 (2002), pp. 11715–11720.

[112] Naruya Saitou and Masatoshi Nei. “The neighbor-joining method: a new method for reconstructing phylogenetic trees.” In: *Molecular biology and evolution* 4.4 (1987), pp. 406–425.

[113] Julian E Sale, Alan R Lehmann, and Roger Woodgate. “Y-family DNA polymerases and their role in tolerance of cellular DNA damage”. In: *Nature Reviews Molecular Cell Biology* 13.3 (2012), pp. 141–152.

[114] David Sankoff, Robert J Cedergren, and Guy Lapalme. “Frequency of insertion-deletion, transversion, and transition in the evolution of 5S ribosomal RNA”. In: *Journal of Molecular Evolution* 7.2 (1976), pp. 133–149.

[115] Orlando D Schärer. “DNA Interstrand Crosslinks: Natural and Drug-Induced DNA Adducts that Induce Unique Cellular Responses”. In: *ChemBioChem* 6.1 (2005), pp. 27–32.

[116] Russell Schwartz and Stanley E Shackney. “Applying unmixing to gene expression data for tumor phylogeny inference”. In: *BMC bioinformatics* 11.1 (2010), p. 42.

[117] RF Schwarz et al. “Phylogenetic quantification of intra-tumor heterogeneity predicts time to relapse in high-grade serous ovarian cancer”. In: *PLoS Medicine (in revision)* (2013).

[118] Roland F Schwarz et al. “Phylogenetic quantification of intra-tumour heterogeneity”. In: *PLoS Comput Biol* 10.4 (2014), e1003535.

[119] Jian Shi et al. “Isolating-a new resampling method for gene order data”. In: *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2011 IEEE Symposium on*. IEEE. 2011, pp. 1–6.

[120] Jian Shi et al. “Using jackknife to assess the quality of gene order phylogenies”. In: *BMC bioinformatics* 11.1 (2010), p. 168.

[121] Adam Shlien and David Malkin. “Copy number variations and cancer”. In: *Genome Med* 1.6 (2009), p. 62.

[122] Adam C Siepel and Bernard ME Moret. “Finding an optimal inversion median: experimental results”. In: *Algorithms in Bioinformatics*. Springer, 2001, pp. 189–203.

[123] Robert A Smith et al. “American Cancer Society guidelines for the early detection of cancer”. In: *CA: a cancer journal for clinicians* 52.1 (2002), pp. 8–22.

[124] Matija Snuderl et al. “Mosaic amplification of multiple receptor tyrosine kinase genes in glioblastoma”. In: *Cancer cell* 20.6 (2011), pp. 810–817.

[125] Pamela S Soltis, Douglas E Soltis, et al. “Applying the bootstrap in phylogeny reconstruction”. In: *Statistical Science* 18.2 (2003), pp. 256–267.

[126] Srinath Sridhar et al. “Algorithms for efficient near-perfect phylogenetic tree reconstruction in theory and practice”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 4.4 (2007), pp. 561–571.

[127] Alexandros Stamatakis. “RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models”. In: *Bioinformatics* 22.21 (2006), pp. 2688–2690.

[128] Philip J Stephens et al. “Massive genomic rearrangement acquired in a single catastrophic event during cancer development”. In: *cell* 144.1 (2011), pp. 27–40.

[129] Francesco Strino et al. “TrAp: a tree approach for fingerprinting subclonal tumor composition”. In: *Nucleic acids research* 41.17 (2013), e165–e165.

[130] Ayshwarya Subramanian, Stanley Shackney, and Russell Schwartz. “Inference of tumor phylogenies from genomic assays on heterogeneous samples”. In: *BioMed Research International* 2012 (2012).

[131] Jung-Suk Sung and Bruce Demple. “Roles of base excision repair subpathways in correcting oxidized abasic sites in DNA”. In: *Febs Journal* 273.8 (2006), pp. 1620–1629.

[132] Charles Swanton. “Intratumor heterogeneity: evolution through space and time”. In: *Cancer research* 72.19 (2012), pp. 4875–4882.

[133] K. M. Swenson, N. D. Pattengale, and B. M. E. Moret. “A framework for orthology assignment from gene rearrangement data.” In: *Proc. 3rd RECOMB Workshop on Comparative Genomics (RECOMBCG’05).* Volume 3678 of *Lecture Notes in Bioinformatics*: Volume 3678 of *Lecture Notes in Bioinformatics*, 2005, pp. 153–166.

[134] David L Swofford and Wayne P Maddison. “Reconstructing ancestral character states under Wagner parsimony”. In: *Mathematical Biosciences* 87.2 (1987), pp. 199–229.

[135] DL Swofford. “PAUP*: phylogenetic analysis using parsimony, version 4.0 b10”. In: (2003).

[136] Aniko Szabo and Kenneth Boucher. “Estimating an oncogenetic tree when false negatives and positives are present”. In: *Mathematical biosciences* 176.2 (2002), pp. 219–236.

[137] Nicholas J Szerlip et al. “Intratumoral heterogeneity of receptor tyrosine kinases EGFR and PDGFRA amplification in glioblastoma defines subpopulations with distinct growth factor response”. In: *Proceedings of the National Academy of Sciences* 109.8 (2012), pp. 3041–3046.

[138] Ming Tan and Dihua Yu. “Molecular mechanisms of erbB2-mediated breast cancer chemoresistance”. In: *Breast Cancer Chemosensitivity.* Springer, 2007, pp. 119–129.

[139] Eric Tannier, Chunfang Zheng, and David Sankoff. “Multichromosomal genome median and halving problems”. In: *Algorithms in Bioinformatics.* Springer, 2008, pp. 1–13.

[140] Ali Tofigh et al. “A global structural EM algorithm for a model of cancer progression”. In: *Advances in neural information processing systems.* 2011, pp. 163–171.

[141] David Tolliver et al. “Robust unmixing of tumor states in array comparative genomic hybridization data”. In: *Bioinformatics* 26.12 (2010), pp. i106–i114.

[142] John W Tukey. “Bias and confidence in not-quite large samples”. In: *Annals of Mathematical Statistics.* Vol. 29. 2. INST MATHEMATICAL STATISTICS IMS BUSINESS OFFICE-SUITE 7, 3401 INVESTMENT BLVD, HAYWARD, CA 94545. 1958, pp. 614–614.

[143] T. Vision. “Gene order evolution in plants: a slow but sure shuffle”. In: *New Phytologist* 168 (2005), pp. 51–60.

[144] Anja Von Heydebreck, Bastian Gunawan, and László Füzesi. “Maximum likelihood estimation of oncogenetic tree models”. In: *Biostatistics* 5.4 (2004), pp. 545–556.

[145] Karen H Vousden and David P Lane. “p53 in health and disease”. In: *Nature reviews Molecular cell biology* 8.4 (2007), pp. 275–283.

[146] Vivian Wang et al. “Ovarian cancer is a heterogeneous disease”. In: *Cancer genetics and cytogenetics* 161.2 (2005), pp. 170–173.

[147] Yong Wang et al. “Clonal evolution in breast cancer revealed by single nucleus genome sequencing”. In: *Nature* 512.7513 (2014), pp. 155–160.

[148] Darawalee Wangsa et al. “Fluorescence in situ hybridization markers for prediction of cervical lymph node metastases”. In: *The American journal of pathology* 175.6 (2009), pp. 2637–2645.

[149] Robert Weinberg. *The biology of cancer*. Garland science, 2013.

[150] Jeffrey T Wigle and Guillermo Oliver. “Prox1 function is required for the development of the murine lymphatic system”. In: *Cell* 98.6 (1999), pp. 769–778.

[151] Anita Wolfer and Sridhar Ramaswamy. “MYC and metastasis”. In: *Cancer research* 71.6 (2011), pp. 2034–2037.

[152] Stefanie Wolters et al. “Loss of Caenorhabditis elegans BRCA1 Promotes Genome Stability During Replication in smc-5 Mutants”. In: *Genetics* 196.4 (2014), pp. 985–999.

[153] Andrew Wei Xu and Bernard ME Moret. “GASTS: parsimony scoring under rearrangements”. In: *Algorithms in Bioinformatics*. Springer, 2011, pp. 351–363.

[154] Andrew Wei Xu and David Sankoff. “Decompositions of multiple breakpoint graphs and rapid exact solutions to the median problem”. In: *Algorithms in Bioinformatics*. Springer, 2008, pp. 25–37.

[155] A.W. Xu. “A fast and exact algorithm for the median of three problem—a graph decomposition approach”. In: 16.10 (2009), pp. 1369–1381.

[156] Xun Xu et al. “Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor”. In: *Cell* 148.5 (2012), pp. 886–895.

[157] Sophia Yancopoulos, Oliver Attie, and Richard Friedberg. “Efficient sorting of genomic permutations by translocation, inversion and block interchange”. In: *Bioinformatics* 21.16 (2005), pp. 3340–3346.

[158] Lucy R Yates and Peter J Campbell. “Evolution of the cancer genome”. In: *Nature Reviews Genetics* 13.11 (2012), pp. 795–806.

[159] David K Yeates and Brian M Wiegmann. *The evolutionary biology of flies*. Columbia University Press, 2005.

[160] Travis I Zack et al. “Pan-cancer patterns of somatic copy number alteration”. In: *Nature genetics* 45.10 (2013), pp. 1134–1140.

[161] Habil Zare et al. “Inferring clonal composition from multiple sections of a breast cancer”. In: *PLoS Comput Biol* 10.7 (2014), e1003703.

[162] Jun Zhou et al. “An iterative approach for phylogenetic analysis of tumor progression using fish copy number”. In: *International Symposium on Bioinformatics Research and Applications*. Springer. 2015, pp. 402–412.