

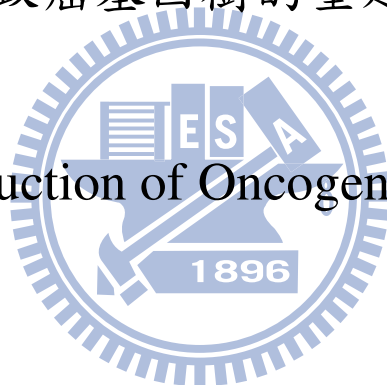
國立交通大學

應用數學系

碩士論文

致癌基因樹的重建

Reconstruction of Oncogenetic Trees



研究生：周彥伶

指導教授：傅恆霖 教授

中華民國一百年六月

致癌基因樹的重建


Reconstruction of Oncogenetic Trees

研究生：周彥伶 Student: Yen-Lin Chou
指導教授：傅恆霖 教授 Advisor: Hung-Lin Fu

國立交通大學

應用數學系

碩士論文



A Thesis
Submitted to Department of Applied Mathematics
College of Science
National Chiao Tung University
in Partial Fulfillment of the Requirements
for the Degree of
Master
in
Applied Mathematics
June 2011
Hsinchu, Taiwan, Republic of China

中華民國一〇一〇年六月

致癌基因樹的重建

研究生：周彥伶

指導教授：傅恆霖 教授

國立交通大學

應用數學系

摘要

比較型基因組雜交分析技術(CGH)，是一種分子生物學上測量腫瘤細胞中，染色體複製倍數(增生或減少)的方法。在生物領域中，通常樹模型是用於研究生物的演化、物種間的進化關係。近幾十年來，生物學家認為癌症發生的過程和DNA的增生或減少息息有關，而且不單單只是某一個基因改變，而是一連串不同基因改變事件發生。因此，也將樹模型當作分析CGH資料的工具，用來研究、探索癌症的發展過程(carcinogenesis)。在數學的基礎下，我們可以從CGH的資料去推導出數學模型。Desper 及其合作者提出了分支樹和距離樹兩種模型，比Vogelstein 等學者提出的直腸癌的單路徑模型更加精確。在這篇論文中，我們介紹了分支樹和距離樹兩個不同的樹模型，比較兩者的相異處。另外，也利用最大概似估計法，來建立引起腫瘤、癌症的樹模型(oncogenetic tree model)。

中華民國一百年六月

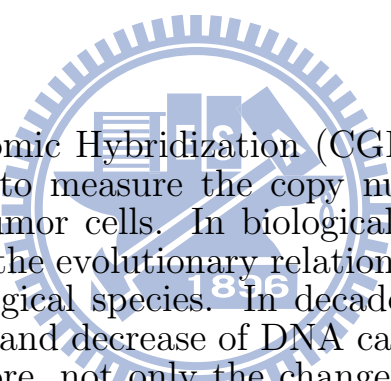
Reconstruction of Oncogenetic Trees

Student: Yen-Lin Chou

Advisor: Hung-Lin Fu

Department of Applied Mathematics
National Chiao Tung University

June, 2011



Comparative Genomic Hybridization (CGH) is a molecular cytogenetic method to measure the copy number aberrations of chromosomal in tumor cells. In biological domain, tree model is usually used on the evolutionary relationship between organic and different biological species. In decades, the biologists believe that increase and decrease of DNA cause the occurrence of cancer. Furthermore, not only the change of one genetic event results in the cancer occurrence, but also serial alteration of different genes. Therefore, tree model is used as the tool to analyze CGH data and explore carcinogenesis. The mathematical model from CGH data can be inferred under mathematical ground. Desper *et al.* stated two models: maximum-weight branching model and distance-based tree model. These two models are more accurate than path model of colorectal-tumor which was brought up by Vogelstein *et al.*. In this paper, we introduce two different tree models and compare the differences between them. In addition, the oncogenetic tree model is reconstructed by using the idea of maximum likelihood.

謝誌

感謝主，漫長的兩年半，終於完成碩士論文了！

兩年半前，我來到陌生的新竹，一開始傅老師就讓我有機會認識學長姊，也給予我工作的機會。在 seminar 的過程中，傅老師通常會給我一些特別的意見，激發我的大腦，除了學業上的知識，老師提的人生道理、保持健康的方法，總是會適時地提醒我。感謝您，我的指導教授——傅恆霖老師。

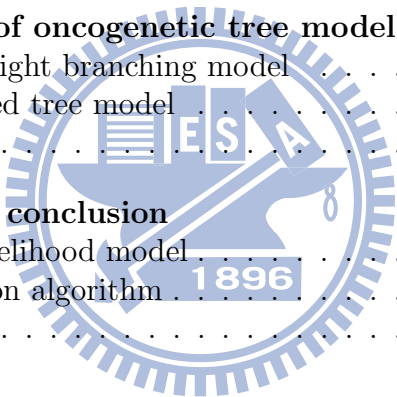
感謝老師讓我成為「傅家子弟」，同時也感謝同為傅家子弟的學長姊們，特別是賓賓學長、惠蘭學姊，在作研究的過程中，即使很忙碌，還是抽空指導我不足的地方，提供非常多的幫助；另外，謝謝敏筠讓我在研究生生活多了籃球的色彩，讓我有機會在球場上奔跑；也謝謝明輝、惠娟、貓頭、Robin、智懷、軒軒等學長姊，在學習過程中提供許多幫助，也給予我論文口試方面的建議。還有在研究生生活一路陪伴我的瑩晏、小吵、小育、小太、俞欣、還有許多應數所同學們、學弟妹們，謝謝大家在這段時間的照顧與幫助，讓我的研究生活更加充實。

感謝上帝帶領我來到新竹，接受到許多不同的挑戰，同時也給予我許多資源與幫助。感謝梅竹團契的輔導們、畢契們、學契們的陪伴與代禱，以及中正教會會友們的關心與幫助，讓我在陌生的新竹感受到極大的溫暖。在寫論文的過程中，生物醫學的部分，以樂幫忙惡補我不足的地方；此外，宜均、晟安、以及我的好室友米琪也幫忙校正我的英文，謝謝你們陪我一起完成論文，梅竹團契真的是人才輩出啊！

最後，感謝我的家人，一路上支持我完成我的學業，雖然無法常常相聚在一起，但是透過網路、電話，讓我知道你們的關心與支持，謝謝爸爸、媽媽、姊姊、哥哥、妹妹，I love you!

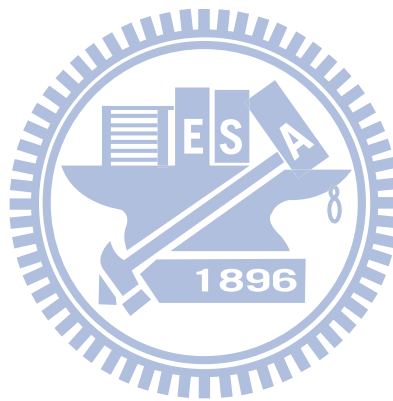
Contents

Abstract (in Chinese)	i
Abstract (in English)	ii
Acknowledgement	iii
Contents	iv
List of Figures	v
1 Introduction	1
1.1 Motivation and background	1
1.2 CGH method	2
1.3 Tree model	4
1.4 Definition	5
2 Reconstruction of oncogenetic tree model	9
2.1 Maximum-weight branching model	9
2.2 Distance-based tree model	11
2.3 Comparison	12
3 Main result and conclusion	16
3.1 Maximum likelihood model	16
3.2 Reconstruction algorithm	19
3.3 Conclusion	21



List of Figures

1	oncogenetic path for colorectal cancer	4
2	Maximum-weight branching model for renal cancer cell [5]	13
3	Distance-based tree model for renal cancer cell [5]	14
4	(a) an oncogenetic tree T ; (b) a subtree T_j of T	18
5	an example of oncogenetic tree in a step of algorithm	19



1 Introduction

1.1 Motivation and background

Cancer has become the main cause for human death around the world. According to the indication of statistical data, cancer continues to hold the top of human dead causes in Taiwan for 28 years, which bring patients great sufferings. There are a lot of factors that are likely to induce cancers, for instance, heredity, diet habit, endocrine disorder, stimulus of chemical substances, etc. Generally, the earlier detect occurrence of cancer, the cure result would be better. Hence, biologists devote on oncogenic mechanism research.

In general, a normal cell has a cell cycle, and check points among cell cycle can determine the growth and replication of cells. Because the mechanisms fail in tumor cells, the cells proliferate abnormally. When growth control mechanisms became abnormal, the cell would be out of control and reproduce a great number of cells, which is called a *tumor*.

Tumors are classified into two types: *benign tumors* and *malignant tumors*. Both of them can proliferate without control. Furthermore, malignant tumors can also penetrate into nearby cells, tissues, even body's organ and may be lethal. Consequently, malignant tumors, which are called *cancer*, are more dangerous than benign tumors.

Deoxyribonucleic acid (DNA) is a self-replicating linear molecule with a large molecular weight, which is contained in living cells. It is a carrier of genetic information. The chain of DNA contains the bases adenine (A), thymine (T), cytosine (C), guanine (G). The sequence of the bases along the

chain forms a genetic code directing the synthesis of RNAs and proteins.

According to biological central dogma, a gene is a specific fragment of DNAs, and it will be transcribed into a RNA fragment, which will then be translated into proteins. Proteins are not only the critical consistent of the organisms' structure but also can maintain vital mechanisms in organisms, including cell cycle check points.

Despite there are many factors may result in some mutations of DNA, DNA repair system largely reduces the abnormal DNA amount. When DNA mutations are too serious that cannot be repaired, the protein expression might be affected. The cell cycle could then be abnormal, causing cancer at this time. This is why we confer about the genetic alterations for cancer studies.

1.2 CGH method

Comparative genomic hybridization (CGH) is an important experimental method in molecular cytogenetic. It is derived from florescence *in situ* hybridization (FISH). They mark DNAs in a normal cell and a tumor cell individually by probes with different colors. Then mix the dyed DNAs, and put them into the chromosomes of the target cell which is in metaphase. After hybridization, the corresponding DNA pairs would bind together.

The mutative DNAs in the tumor cell proliferate massively. The amplified DNAs would be more, but the deleted DNAs are less. In the hybridization, the locations of amplified DNAs may bind with DNAs of the tumor cell, and the locations of deleted DNAs bind with DNAs of the normal cell.

Chromosome is one of the small, rod-shaped, deeply staining bodies

in a cell nucleus. Each chromosome consists of a single long molecule of DNA associated with proteins. DNA are scattered throughout the nucleus ordinarily, but they would be centralized and become chromosomes among cell division.

Consequently, by the order numbers of chromosomes in goal cell, the chromosomes which show color of the tumor cell are amplified chromosomes, and the chromosomes which show color of the normal cell are deleted chromosomes.

During the duplicating process, the chromosomes are in pairs and so do the DNAs in a normal cell, because of this, it is the feature used in CGH. In different situation such as amplifications, deletions, or normal conditions, the process of hybridization are different and the fluorescence on DNA display differently. These chromosome variations observed in a tumor are called *copy number aberrations (CNAs)*.

Therefore, the gains or losses of DNAs in tumor cells and the DNA locations of abnormal copies can be detect by these fluorescent markers. A series data of CNAs would be obtained after these experiment.

Based on the information of DNA CNAs, we specify the relationships between genetic alterations and diseases or cancers. CNAs play a critical role in medical science. Furthermore, CGH can efficiently detect the amplifications and deletions of all DNA fragments in a tumor at a time. Therefore, CGH is used widely in research.

However, after accumulating a large amount of laboratory data, how to analyze these data in order to explain the relationship between genetic alterations and tumorigenesis becomes another bigger topic.

1.3 Tree model

There are 23 pairs of chromosomes in a human cell, number 1 to number 22, X , and Y . Because of chromosome Y not contained in females, we do not consider the mutations of chromosome Y . The longer arm of a chromosome is denoted by q , and the shorter one is denoted by p . Each chromosome has the longer arm q , but the shorter arm is not in chromosome 13, 14, 15, 21, 22. As a result, there are 41 regions that we can consider, and each one will probably amplify or delete. Then the chromosome variations are at most 82 possible conditions.

The models early proposed are based on the path model stated by Vogelstein *et al.* [3]. From a large CGH database, we can infer that colorectal cancer is mainly lead by four genetic events $+2q$, $-7p$, $-13q$, $-6q$. The model of colorectal cancer is built with pathway by these four genetic events. And these four events from the data show a causal link. When $+2q$ occurs, $-7p$ has greater chances of occurring. When $-7p$ occurs, then $-13q$ has greater chances of occurring. When $-13q$ occurs, $-6q$ is more likely to occur. Therefore, the four genetic events are important indicators of cancer detection.

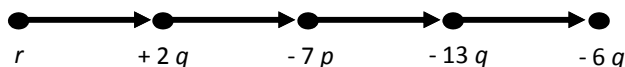


Figure 1: oncogenetic path for colorectal cancer

However, while widely used, the path model still has many shortcomings. According to statistics from biological research, the genetic events which occur during the formation of tumor cells do not so much resem-

ble a chain, rather the accumulation of multiple alterations. Moreover, the genetic events are possible form many different branches. Therefore, tree topology would be more suitable than path topology as a model for the analysis of the carcinogenic process.

1.4 Definition

A *graph* is an ordered pair $G = (V, E)$, where V is a finite set of elements called *vertices* and E is a set of unordered pairs of vertices called *edges*. In general, the vertex set is denoted by $V(G)$ or V , and the edge set is denoted by $E(G)$ or E .

A graph P_n with vertex set $V(P_n) = \{v_0, v_1, \dots, v_{n-1}\}$ and edge set $E(P_n) = \{(v_i, v_{i+1}) \mid i = 0, 1, \dots, n-2\}$ is call an $v_0 - v_{n-1}$ *path*. The *length* of P_n is the number of $E(P_n) = n - 1$. A graph C_n with vertex set $V(C_n) = \{v_0, v_1, \dots, v_{n-1}\}$ and edge set $E(C_n) = \{(v_i, v_{i+1}) \mid i = 0, 1, \dots, n-2\} \cup \{(v_0, v_{n-1})\}$ is call a *cycle*.

A graph G is *connected* if there exists an $u - v$ path in G for any $u, v \in V(G)$. A *subgraph* of a graph G is a graph H such that $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$. A *spanning subgraph* of a graph G is a subgraph with vertex set $V(G)$. The *components* of a graph G are its maximal connected subgraphs.

A *tree* or *branching* is a connected graph which has no cycle. A *spanning tree* of graph G is a spanning subgraph of G without cycles.

The *degree* of vertex v in a graph G is the number of edges incident to v . The vertex whose degree is 1 in a tree is called a *leaf*. The set which is a collection of leaves of a tree T is called the *leaf set* of T , denoted by $L(T)$.

A *directed graph* or *digraph* is an ordered pair $D = (V, E)$ with vertex set V and edge set E which is a set of ordered pairs of vertices. We call the ordered pair as *directed edge* or *arc*. Given a directed edge $e = (u, v)$, the first vertex u of e is called the *tail* of e and the second vertex v is called the *head* of e . Furthermore, u is called a *parent* of v , denoted by $par(v)$, and v is called a *child* of u , denoted by $ch(u)$.

In a directed tree T , if the vertex $v \in V(T)$ has no parent, then we called v is the *root* of tree T .

A digraph P_n with vertex set $V(P_n) = \{v_0, v_1, \dots, v_{n-1}\}$ and directed edge set $E(P_n) = \{(v_i, v_{i+1}) \mid i = 0, 1, \dots, n-2\}$ is called a *directed path* from v_0 to v_{n-1} ($v_0 - v_{n-1}$ directed path).

In a directed tree T , let r be the root of T , then there exists a directed path from r to v for every vertex $v \in V(T)$. Every vertex on the $r - v$ directed path unless v is called an *ancestor* of v . For any two distinct vertex u, v , if w is an ancestor both of u and v , then w is called the *common ancestor* of u, v . If the vertex w is an common ancestor of u, v with maximum length path from root to w , then we call w is the *least common ancestor* of u, v , denoted by $l.c.a.(u, v)$.

In a directed graph D , if there exists a directed path P_4 with vertex set $V(P_4) = \{v_0, v_1, v_2, v_3\}$ and directed edge set $E(P_4) = \{e_i = (v_i, v_{i+1}) \mid i = 0, 1, 2\}$, then e_0 is a *predecessor edge* of e_1 , denoted by $pre(e_1)$, and e_2 is a *successor edge* of e_1 , denoted by $su(e_1)$. A *leaf edge* e is a directed edge if $su(e) = \emptyset$.

In this paper, all graphs we mentioned are directed. Therefore, we use "graph" instead of "directed graph", "edge" instead of "directed edge",

”tree” instead of ”directed tree”, and so on.

Before discussing with the tree models, we define the probability distribution on a tree.

Definition 1.4.1. *Given a vertex set V , a function P which assign non-negative real number to all subset of V such that $\sum_{S \subseteq V} P(S) = 1$ is called a probability distribution P on 2^V .*

Definition 1.4.2. *An oncogenetic tree $T = (V, E, r, p, L)$ is a rooted directed tree with $0 < p(e) \leq 1, \forall e \in E$, where V is the vertex set of tree, E is a set of pairs of vertices, r is the root of tree (no genetic alterations), L is the important genetic event set, and $p(e)$ is the probability that edge e is present independently.*

Each oncogenetic tree T is useful as the generator of distribution on 2^V , and we denote the distribution as P_T .

In the process of causing tumorigenesis, there are some causal link between different genetic events, or different alterations probably occur in different stages. So we will divide into two mathematical models, untimed and timed.

Definition 1.4.3. *A pure untimed oncogenetic tree is a tree T with a probability $p(e)$ attached to each edge e . This tree generates observations on mutation presence/absence the following way : each edge e is independently retained with probability $p(e)$; the set of vertices that are still reachable from root r gives the set of the observed genetic alterations.*

Definition 1.4.4. *A pure timed oncogenetic tree is a tree T with a rate $\lambda(e)$ attached to each edge e and an observation-time distribution ϕ on $+$. This*

tree generates observations on mutation presence/absence the following way : first the time of observation t is drawn from ϕ and the transition time along each edge e is drawn independently from an exponential distribution with rate $\lambda(e)$. The set of vertices that are reachable from root r along a path for which the sum of transition times is less than t gives the set of the observed genetic alterations.

Next, there are some notations for probabilities.

Notation : An oncogenetic tree $T = (V, E, r, p, L)$ with distribution P , define

- $p_i = P(v_i \text{ occurs}), \text{ for } i = 1, \dots, n; p_0 = 1$
- $p_{ij} = \begin{cases} P(v_i \text{ and } v_j \text{ occur}) & , \text{ for } i \neq j, i, j = 1, \dots, n \\ p_i & , \text{ o.w.} \end{cases}$
- $p_{i|j} = P(v_i \text{ occurs} \mid v_j \text{ occurs}), \text{ for } i \neq j, i, j = 1, \dots, n$
- $p_{i \vee j} = P(v_i \text{ or } v_j \text{ occur}), \text{ for } i \neq j, i, j = 1, \dots, n$

, where $v_0 = r$.

2 Reconstruction of oncogenetic tree model

Since each timed oncogenetic tree with path topology has an equivalent untimed oncogenetic tree [11], by the biological statistics, we can get the correlative genetic events of a specific tumor type, but we cannot have the occurrence order of all genetic events immediately in an individual tumor. Supposed that each alteration occurs at most once in every tumor at a time. Besides, a genetic event may occur in different stages, but we don't consider the number of occurrences of each alteration. We just care about if the alteration occurs in an individual tumor or not. Thus, we will talk about untimed cases in this thesis.

How to use the reconstruction of tree model to find the optimal oncogenetic tree for given CGH data is an crucial problem. In this section, we shall talk about the reconstruction problem as the following:

Input :

A set L of genetic events, and k samples from a distribution p over 2^L .

Output :

An oncogenetic tree $T = (V, E, r, p, L)$ with $L \subset V$, such that P_T is an approximation of P .

2.1 Maximum-weight branching model

In maximum-weight branching model, each vertex should be an important genetic event. Suppose there are n genetic events (including root r), we construct a directed complete graph K_n on V . Then there are $n(n-1)$ possible direct edges, but the oncogenetic tree obtained from maximum-weight branching model just has $n-1$ edges finally. So the reconstruc-

tion of maximum-weight branching model is to find the maximum spanning branching such that the sum of all edges weights in the branching is maximum.

Now, we introduce the definition of edge weight. Let $T = (V, E, r, p, L)$ be an oncogenetic tree. The weight function is a mapping from p to real number for the pairs in V^2 . In the maximum-weight branching model, there is a vertex as the root, and other vertices are important events, i.e, $V = L \cup r$. The directed edge e_{ij} means the cause-effect, in other terms, if event i occurs, then the occurrence of event j is more possible. So, the weight w_{ij} should reflect the possibility from event i to event j .

1. The weight should reflect the likelihood ratio for i and j occurring together. The likelihood of assumption over the likelihood which could occur is $\frac{p_{ij}}{p_i p_j}$.
2. The weight should reflect that it is more possible to occur first for event i . If $p_i > p_j$, then the high possibility of an edge from i to j may exist with $\frac{p_i}{p_i + p_j}$.

Then, we combine the above two to obtain

$$\frac{p_i}{(p_i + p_j)} \times \frac{p_{ij}}{p_i p_j} = \frac{p_{ij}}{p_j (p_i + p_j)}. \quad (*)$$

The logarithm (*) of is chosen instead of the above-mentioned weight for proving reconstruction algorithm works. Thus, we define the weight function on the directed edge from event i to event j :

$$w_{ij} = \log \frac{p_{ij}}{p_j (p_i + p_j)} = \log p_{ij} - \log(p_i + p_j) - \log p_j.$$

The spanning tree with the sum of all directed edges maximum is the optimal branching tree in this model.

Definition 2.1.1. *Let $T = (V, E, r, p, L)$ be an oncogenetic tree. Then T is not skewed if for all distinct vertices $v_i, v_j, v_k \in V$ where $v_k = \text{l.c.a.}(v_i, v_j)$, then $p_{i|j} < p_{i \vee j|k}$.*

Theorem 2.1.1. [11] *Let $T = (V, E, r, p, L)$ be a non-skewed oncogenetic tree. Then the maximum branching over V with respect to the weight defined by $w_{ij} = \log \frac{p_{ij}}{p_j(p_i + p_j)}$ is precisely T .*

2.2 Distance-based tree model

The idea of distance-based tree model comes from the combination of Cavender-Farris trees [6, 8] and path metrics [7]. And, we use distance-based algorithms from the phylogenetic literature. In this model, the important genetic event set L is exactly the leaf set of oncogenetic tree, and there are some unknown (or hidden) genetic events as internal nodes of tree.

So far, there is no algorithm which can ensure that the output of reconstruction algorithm is exactly the real oncogenetic tree. Actually, the reconstruction which was brought up by Desper *et al.* [12] is an approximation.

Now, we consider the distance by path metrics. The logarithm of edge probability is negative since it is small than 1. For all edge e , we define the distance $d(e) = -\log p(e)$. If x and y are two leaves of tree, then we define the distance

$$d(x, y) = \log \frac{p_x p_y}{p_{xy}^2} = -2 \log p_{xy} + \log p_x + \log p_y.$$

Given a sample S , we obtained the probability \hat{p}_x, \hat{p}_{xy} from sample S for all genetic events x, y , where \hat{p}_x is the observed probability of x , and \hat{p}_{xy} is the observed joint probability of x and y . Next, we calculate the distance $\hat{d}(x, y) = -2\log \hat{p}_{xy} + \log \hat{p}_x + \log \hat{p}_y$ for each events x and y . Finally, we use a tree-fitting algorithm to find optimal oncogenetic tree T' such that the metric $d_{T'}$ is close to the metric \hat{d} .

In [12], Desper *et al.* use the L_1 distance [9] between two trees to measure the approximation. Let p_{min} be the minimum p_x for all genetic event x .

Theorem 2.2.1. [12] *Suppose that the input data are k samples from the distribution p_T of an oncogenetic tree T . Our oncogenetic tree reconstruction algorithm converges to a tree T_* and distribution p_{T_*} such that the expected L_1 distance between p_T and p_{T_*} is $O(\frac{|L|^2}{\sqrt{kp_{min}}})$.*

Thus, we infer that if the number of samples is considerable enough, then the output tree by the reconstruction algorithm is approaching to real tree.

2.3 Comparison

A tree can be regarded as a collection of path. Then maximum-weight branching models are developments of path models. However, on an oncogenetic tree derived from maximum-weight branching model, each vertex may not have a single direction, but the vertex likely occur on many paths. Thus, we can observe the heterogeneity of a tumor in this model.

As a result of reconstruction of maximum-weight branching model, we can get the information shown as below:

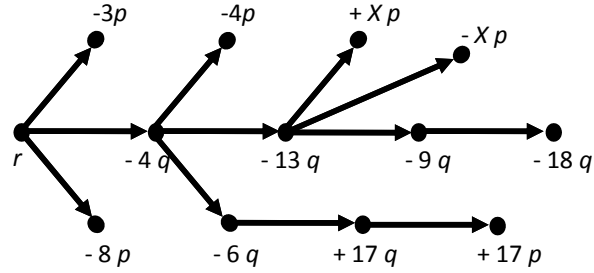


Figure 2: Maximum-weight branching model for renal cancer cell [5]

1. Just like path models, there exists a causal relationship between the vertices on each path of an oncogenetic tree. That is, if the predecessor occurs, then the possibility of successor occurrence increase.
2. Events on a branch would be marked as a subclass of the tumor.

Distance-based tree model is discussed from another direction. The distances between various vertices are calculated, then we find the approximation of an optimal tree by the distance matrix.

As a result of reconstruction of distance-based tree model, we can get the information as below:

1. Each edge e has length $-\log p(e)$ on the oncogenetic tree, and the horizontal distance between endpoints of e is proportional to the length of e .
2. The vertex which is more closed to root would be the event which occurs earlier.
3. Just like maximum-weight branching model, events on a branch can be marked as a subclass of the tumor.

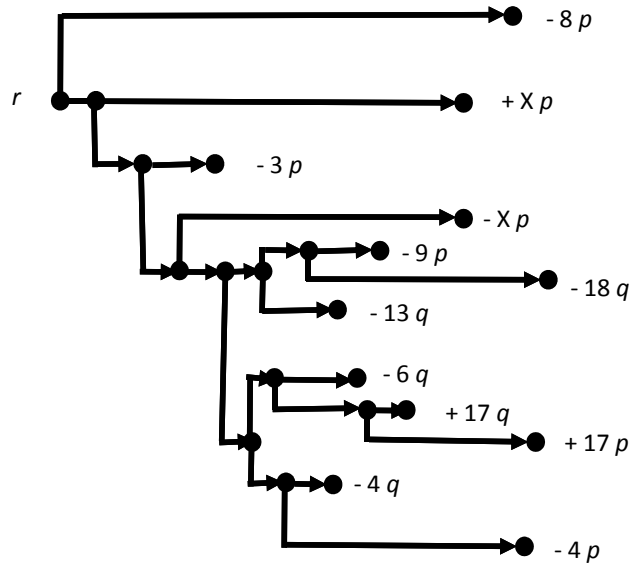


Figure 3: Distance-based tree model for renal cancer cell [5]

Although maximum-weight branching model and distance-based model construct the oncogenetic tree according to different mathematical ways, they have partial common features and their own advantages. Especially, maximum-weight branching model shows the cause and effect between events. On the other hand, distance-based model focuses on the correlation between each two events. Therefore, both two models are commonly used on most of cancer researchs.

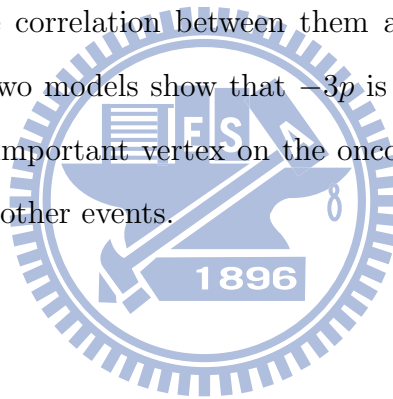
Due to the advantages of two models, the reliability of oncogenetic trees increases after matching the pairwise inferred relationships. Then the reliability of oncogenetic trees reconstructed by both two models would increase. We can deduce from the influence of genetic alterations on the cancer by the oncogenetic trees derived from tree models.

We know that the early events are closed to the root in the two models.

However, the distance-based tree model is more accurate because of the horizontal distance from root to the vertex. For example, by the oncogenetic trees in Figure 2 and 3, we know that the events $+3p$ and $-4q$ closed to the root are early genetic events.

Furthermore, the events with strong correlations are assembled in a branch in both models. Figure 2 and Figure 3 show that there are subtrees with $-13q$, $-9p$, and $-18q$ in both two models. The subtrees with $-6q$, $+17q$, and $+17p$ also exist in both two models, and so on. We also observe that $-4q$ is very closed to $-4p$, and so do $+17q$ to $+17p$.

On the contrary, $-8p$ and $-3p$ are independent with other vertices. This implies that the correlation between them and all other events are very small. Both of two models show that $-3p$ is very closed to the root. However, $-4q$ is the important vertex on the oncogenetic trees because it is highly pertinent to other events.



3 Main result and conclusion

In section 2.2, we infer tree models by distance-based methods from phylogenetics. The maximum likelihood model adopt the suggest of distance-based model.

3.1 Maximum likelihood model

We start with building a binary matrix M from given CGH data information. The row index corresponds to each genetic event and the column index corresponds to each tumor. The entries of M stand for the occurrence relation, i.e., $m_{ij} = 1$ if and only if the genetic event i occurs in the tumor j . According to this matrix we can organize the relationship between genetic events and tumor. The column vector m_j shows the variations of tumor j , called the profile of j .

In this section, we consider the variation of an individual tumor with conditional probability.

Definition 3.1.1. *Let X_v be a random variable with value in $\{0, 1\}$ for all vertex v of tree, such that $X_v = 1$ if $v = r$; otherwise, $P(X_v = 1 | X_{par(v)} = 1) = p_e$, and $P(X_v = 1 | X_{par(v)} = 0) = 0$. Notice that $X_v = 0$ implies that v isn't observed at the tumor in the experiment. Conversely, $X_v = 1$ implies that v is observed in the tumor.*

The subtree of T rooted at v is denoted by $T^{(v)}$. The leaf set of $T^{(v)}$ is denoted by $L(T^{(v)})$. Then we define q_e , the conditional probability, as below:

$$\begin{aligned}
q_e &:= P(X_l = 0, \forall l \in L(T^{(v)}) \mid X_{par(v)}=1) \\
&= \begin{cases} (1 - p_e) & , \text{ for all leaf edge } e \\ (1 - p_e) + p_e \prod_{k \in su(e)} q_k & , \text{ o.w.} \end{cases}
\end{aligned}$$

, where e is an edge from $par(v)$ to v .

The probability q_e is separated into two conditions: (i) $par(v)$ occurs, but v does not. (ii) Both $par(v)$ and v occurs, but all leaves in $L(T^{(v)})$ do not. And we use the recursion to compute part (ii). Trivially, if e is a leaf edge, i.e., $su(e) = \emptyset$, then $q_e = 1 - p_e$. For example, in Figure 4 (a), $q_{e_3} = (1 - p_{e_3}) + p_{e_3} q_{e_7} q_{e_8}$.

Next, we discuss the condition of alterations in a individual tumor. Given an oncogenetic tree $T = (V, E, r, L, p)$ and an observed data matrix M from the experiments, we have some notations:

1. A set L_j is collecting the alterations observed in tumor j , i.e., $L_j := \{v \in L \mid m_{vj} = 1\}$.
2. The subtree of T rooted at r and spanned by $r \cup L_j$ is denoted by T_j .
3. $E'(T_j) = \{e \in E(T) \setminus E(T_j) \mid pre(e) \in E(T_j)\}$.
4. $m_e = |\{j \mid e \in E(T_j)\}|$.
5. $n_e = |\{j \mid e \in E'(T_j)\}|$.

The profile $m_{.j}$ respects to all occurrences of alterations in tumor j . Then the probability of profile $m_{.j}$ is

$$P(m_{.j}) = \prod_{e \in E(T_j)} p_e \prod_{e \in E'(T_j)} q_e. \quad (1)$$

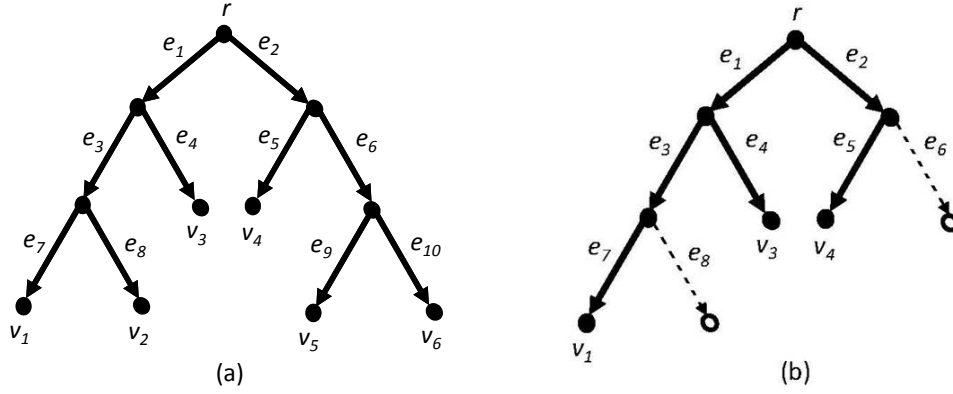


Figure 4: (a) an oncogenetic tree T ; (b) a subtree T_j of T

For example, an oncogenetic tree T is the part (a) of Figure 4. Suppose alterations v_1, v_3, v_4 are observed in tumor j , i.e., $\{v_1, v_3, v_4\}$ is the profile of tumor j . And the part (b) is the subtree T_j of T . $E(T_j) = \{e_1, e_2, e_3, e_4, e_5, e_7\}$, and $E'(T_j) = \{e_6, e_8\}$. So, we calculate the probability of profile of tumor j which is $P(m_j) = p_{e_1}p_{e_2}p_{e_3}p_{e_4}p_{e_5}p_{e_7}q_{e_6}q_{e_8}$.

We can obtain a series of genetic data from CGH experimental result. Each tumor has its own data, then tumors are taken as independent between each other. There is an equation formed by (1) for each tumor's profile. Thus, we define the likelihood of the oncogenetic tree as the product of tumors' probabilities:

$$\mathcal{L}(T; M) := \prod_j \left(\prod_{e \in E(T_j)} p_e \prod_{e \in E'(T_j)} q_e \right) = \prod_{e \in E(T)} (p_e^{m_e} q_e^{n_e}).$$

In next section, we'd like to construct a reconstruction algorithm for a given fixed matrix M . Then the likelihood of an oncogenetic tree T will be denoted by $\mathcal{L}(T)$ instead of $\mathcal{L}(T; M)$.

3.2 Reconstruction algorithm

We start with introducing some definitions and notations. Given a tree T ,

1. $AE_v = \{e \in E(T) \mid e \text{ is in the path from the root of tree to } v\}$ is called the *ancestor edge set* of v .
2. $CE_v = \{e \in E(T) \mid e = vv' \text{ for some child vertex } v' \text{ of } v\}$ is called the *child edge set* of v .
3. A *free vertex* is an internal node which has only one child. In the following algorithm, each tree contains at most one free vertex, then we denote the free vertex set by $f(T)$. For instance, the vertex v_1 in Figure 5. is the free vertex of T .
4. Let $T = (V, E)$ be a tree with vertex set V and edge set E in the following algorithm.

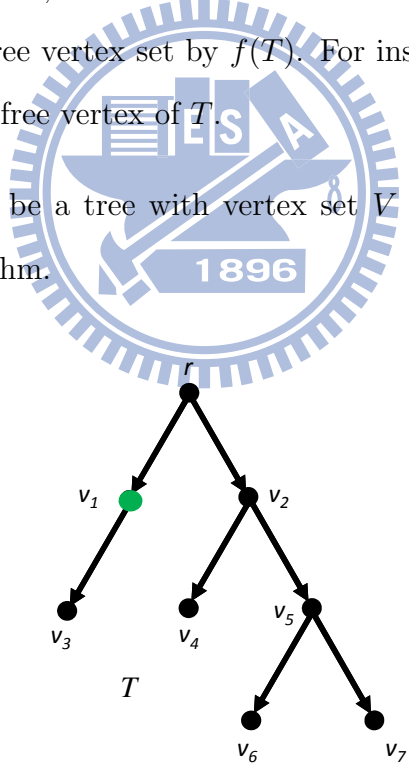


Figure 5: an example of oncogenetic tree in a step of algorithm

Input:

r : no genetic alteration.

L : the set of important events.

I : the set of hidden events (as the internal node in the tree).

Output:

A binary tree T with maximal likelihood.

Algorithm 1 RECONSTRUCTION-BY-MAXIMUM-LIKELIHOOD

```

1:  $V \leftarrow \{r\}$ .
2: Choose an event in  $L$ , denoted by  $v$ .
3:  $V \leftarrow V \cup \{v\}$ ,  $E \leftarrow \{rv\}$ .
4: while ( $|L \setminus V| \neq 1$ ) do
5:   (i) Choose an event in  $L \setminus V$ , denoted by  $v$ .
6:   (ii)  $V \leftarrow V \cup \{v, v'\}$ ,  $E \leftarrow E \cup f(T)v', v'v$ , where  $v'$  is the
       parent of  $v$ . %  $v'$  is as an "hidden" event.
7:   (iii) Count the likelihood  $\mathcal{L}(T)$ .
8:   (iv)  $a \leftarrow f(T)$ , and  $DE \leftarrow AE_a \cup CE_a$ 
9:     Step1 (create a new tree  $T' = (V', E')$  from  $T$ )
10:    (a)  $E' \leftarrow E$ ,  $V' \leftarrow V$ , choose an edge  $e \notin DE$ , which
        separate  $T$  into two components  $T_1$  and  $T_2$  where
         $r$  is contained in  $T_1$ ,  $E' \leftarrow (E' \setminus \{e\})$ 
11:    (b)  $DE \leftarrow DE \cup \{e\}$ .
12:    (c) Connect the root  $u$  of  $T_2$  and vertex  $a$  by a new edge,
         $E' \leftarrow E' \cup \{au\}$ .
13:    (d) Count the likelihood  $\mathcal{L}(T')$ .
14:    Step2 (Compare the likelihood of  $T$  and  $T'$ )
15:    if  $\mathcal{L}(T') < \mathcal{L}(T)$  and  $DE \neq E(T)$  then
16:      go to Step1.
17:    else if  $\mathcal{L}(T') < \mathcal{L}(T)$  and  $DE = E(T)$  then
18:      Return  $T$ .
19:    else
20:       $T \leftarrow T'$ ,  $DE \leftarrow \{au\}$ ,  $a \leftarrow f(T)$ ,  $DE \leftarrow DE \cup AE_a \cup CE_a$ ,
      and go to Step1.
21:    end if
22:  end while
23: Add the last event  $v$ ,  $V \leftarrow V \cup \{v\}$ ,  $E \leftarrow E \cup \{f(T)v\}$ .
24: Return  $T$ .

```

The main idea of this algorithm:

In step 1 (rearrangement), we cut an edge e of $T \setminus DE$, such that the original tree separate two components T_1 and T_2 , where root of T is contained in T_1 . And then, we link these two components with a new edge $e' \neq e$. For keeping the order of vertices in each component, we connect the two components T_1, T_2 with $f(T)$ and root of T_2 . Next, we'd like to find the tree with larger likelihood. So, we compare the original tree T and the new tree T' in step 2.

(I) If $\mathcal{L}(T') < \mathcal{L}(T)$ and $DE \neq E(T)$, it implies that T is better than T' in this size, and there are more choice of edges for cutting. Then we go back to rearranging step again. At the same time, e is taken in DE for avoiding a tree occurring more than once.

(II) If $\mathcal{L}(T') < \mathcal{L}(T)$ and $DE = E(T)$, it implies that T is better than T' . But there is no edge which could be cut. So, the tree T is the optimal tree with maximal likelihood in this size. We will add a pair of vertices to T in order to proceed rearranging step with next size.

(III) If $\mathcal{L}(T') > \mathcal{L}(T)$, it implies that the T' is better than T . So, we use T' in place of T , and then go back to rearranging step.

3.3 Conclusion

Every time rearranging is to figure out the tree with maximum likelihood in this algorithm. It keeps sorting until there is no tree with larger likelihood. By following the algorithm steps, the likelihood of the tree does not increase in the end. Although it may be the local maximum, it is the optimum condition for the algorithm.

Maximum likelihood model has the common point with the two models in previous section. The genetic events assembled in a branch are more correlative between each other. That is, the genetic events which triggered together are gather in the same branch.

The previous two models can analyze the effect on the tumor by genetic alterations. However, maximum likelihood model analyze the data about a specific type of tumors since the observed matrix will collect a specific type of cancer of many individual tumors. Therefore, the maximum likelihood model can integrate correlation between genetic alterations in a specific type of cancer.

Through these models, we know which genetic alterations would cause cancer lesions easily. Recently, many researchers use tree model to explore and analyze the mechanism of tumor.

We can combine the relevant biomedical information. Not only CGH data but also the results of other experiments, e.g., array-CGH, fluorescence *in situ* hybridization (FISH), detection methods of single nucleotide polymorphisms, can reconstruct the tree model. It would deepen the knowledge of tumor developing mechanism, and are good for preventing. Furthermore, it helps earlier detecting, diagnosis, and treatment.

References

- [1] A. Kallioniemi, O.P. Kallioniemi, D. Sudar, D. Rutovitz, J.W. Gray, F. Waldman and D. Pinkel, Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, Volume 258, No.5083, pp.818-821, 1992.
- [2] A. von Heydebreck, B. Gunawan, L. Füzesi, Maximum likelihood estimation of oncogenetic tree models, *Biostatistics*, Volume 5, No.4, pp.545-556, 2004.
- [3] B. Vogelstein, E.R. Fearon, S.R. Hamilton, S.E. Kern, A.C. Preisinger, M. Leppert, Y. Nakamura, R. White, A.M. Smits, J.L. Bos, Genetic alterations during colorectal-tumor development. *N. Engl. J. Med*, Volume 319, No.9, pp.525-532, 1988.
- [4] D.L. Swofford, G.J. Olsen, Phylogeny reconstruction, in: D.M. Hillis and C. Moritz, Editors, *Molecular Systematics*, Sinauer Associates, Sunderland, pp.411-501, 1990.
- [5] F. Jiang, R. Desper, C. H. Papadimitriou, A. A. Schäffer, O.-P. Kallioniemi, J. Richter, P. Schraml, G. Sauter, M. J. Mihatsch, H. Moch, Construction of evolutionary tree models for renal cell carcinoma from comparative genomic hybridization data, *Cancer Research*, Volume 60, No. 22, pp.6503-6509, 2000.
- [6] J.A. Cavender, Taxonomy with confidence, *Mathematical Biosciences*, Volume 40, Issues 3-4, pp.271-280, August 1978.

- [7] J.-P. Barthélemy and A. Guénoche, *Trees and Proximity Representations*, Wiley, New York, 1991.
- [8] J.S. Farris, A probability model for inferring evolutionary trees, *Syst. Zool.*, Volume 22, pp.250-256, 1973.
- [9] M. Farach, S. Kannan, Efficient algorithms for inverting evolution, *Proc. 28th ACM Symp. Theory Comput.*, pp230-236, 1996.
- [10] M.S. Waterman, T.F. Smith, M. Singh, and W.A. Beyer, Additive evolutionary trees, *Journal of Theoretical Biology*, Volume 64, Issue 2, pp.199-213, 21 January 1977.
- [11] R. Desper, F. Jiang, O.P. Kallioniemi, H. Moch, C.H. Papadimitriou, A.A. Schäffer, Inferring tree models for oncogenesis from comparative genome hybridization data, *J. Computational Biology*, Volume 6, No. 1, pp.37-51, 1999.
- [12] R. Desper, F. Jiang, O.P. Kallioniemi, H. Moch, C.H. Papadimitriou, A.A. Schäffer, Distance-based reconstruction of tree models for oncogenesis. *J. Computational Biology*, Volume 7, No. 6, pp.789-803, 2000.
- [13] R.E. Tarjan, Finding optimum branchings, *Networks*, Volume 7, No. 1, pp.25-35, Spring 1977.