Construction of Evolutionary Tree Models for Oncogenesis of Endometrial Adenocarcinoma

Lei Chen

Submitted by Lei Chen to the University of Skovde as a thesis towards the degree of M.Sc. by examination and thesis in the Department of Humanities and Informatics.

January 2005

I certify that all material in this thesis which is not my own work has been identified and that no material is included for which a degree has already been conferred upon me.

Lei Chen

Table of contents

Construction of Evolutionary Tree Models for Oncogenesis of Endometrial Adenocarcinoma

Lei Chen

Abstract

Endometrial adenocarcinoma (EAC) is the fourth leading cause of carcinoma in woman worldwide, but not much is known about genetic factors involved in this complex disease. During the EAC process, it is well known that losses and gains of chromosomal regions do not occur completely at random, but partly through some flow of causality. In this work, we used three different algorithms based on frequency of genomic alterations to construct 27 tree models of oncogenesis. So far, no study about applying pathway models to microsatellite marker data had been reported. Data from genome–wide scans with microsatellite markers were classified into 9 data sets, according to two biological approaches (solid tumor cell and corresponding tissue culture) and three different genetic backgrounds provided by intercrossing the susceptible rat BDII strain and two normal rat strains. Compared to previous study, similar conclusions were drawn from tree models that three main important regions (I, II and III) and two subordinate regions (IV and V) are likely to be involved in EAC development. Further information about these regions such as their likely order and relationships was produced by the tree models. A high consistency in tree models and the relationship among p19, Tp53 and Tp53 inducible protein genes provided supportive evidence for the reliability of results.

Keywords: algorithms, cancerous progress, endometrial adenocarcinoma, microsatellite marker, tree models

1 Introduction

Endometrial adenocarcinoma (EAC) is the fourth most common type of malignant tumor in woman (Cavanagh et al., 1999). EAC is a complex disease as environmental and other non-genetic factors play critical roles in many stages in the neoplasitc process (Olah, 1999; Noumoff & Faraqi, 1993). It is well known that EAC is genetically predisposed, and consequently specific genetic abnormalities and activation of oncogenes determine the etiology of EAC (Esteller et al., 1999; Suzuki et al., 1997). However, so far not many specific genetic changes have been identified and characterized during tumorigenesis and not much is known about the links among genetic changes, such as the order and causalities of the genetic changes.

Carcinogenesis is a multistep process which develops through the accumulation of genetic mutations (Balmain, 2002; Vogelstein & Kinzler, 1993). In a normal cell, the first mutation results in limited expansion of the progeny of a single cell. One of these cells is likely to acquire a second mutation, resulting in further loss of growth control. Each successive mutation provides a further growth advantage. Eventually, the cell will become malignant, enabling it to invade through surrounding normal tissues and metastasize to other organs (Pathak, 1999; Hanahan & Weinberg, 2000). Once a set of critical genetic alterations develops, the cancer cell goes “out of control” and starts to accumulate seemingly random alterations (Desper et al., 1999). The whole number of mutations that are present in a tumor cell population is about 10,000 (Tomlinson et al., 2002), however, as mentioned above, the total number of pivotal mutations that make transformation of a single normal cell into a malignant derivative is 3 to 7 (Figure 1A). Obviously, it is difficult to identify the primary causative mutations among 10,000, the majority of which are produced by genomic instability.

It is desirable to identify genetic changes causally associated with the particular tumor type and the relationships among them, as well as to find in which order the significant genetic alterations happen. More complex mathematic expression based on the simplest probability of genetic events can reflect the relationship among the biology events (such as the probability of event *a* and *b* occurring together or the probability of event *a* causing event *b*) (Desper et al., 1999; 2000).

Application of mathematical pathway models to genetic data from tumor sets provides new information on the interrelationships among genetic changes during tumor progression. The models can be used to make predictions of:

- which chromosomal regions are most likely to harbor important genes for tumor initiation (genetic events that tend to occur early);
- which events may be important for progression (genetic events that tend to occur later);
- which events tend to occur together;
- the likely order of events (Moch & Mihatsch, 2002; Desper et al., 1999).

Models for tumor progression pathways would be of obvious value to the early diagnosis and treatment of cancer.

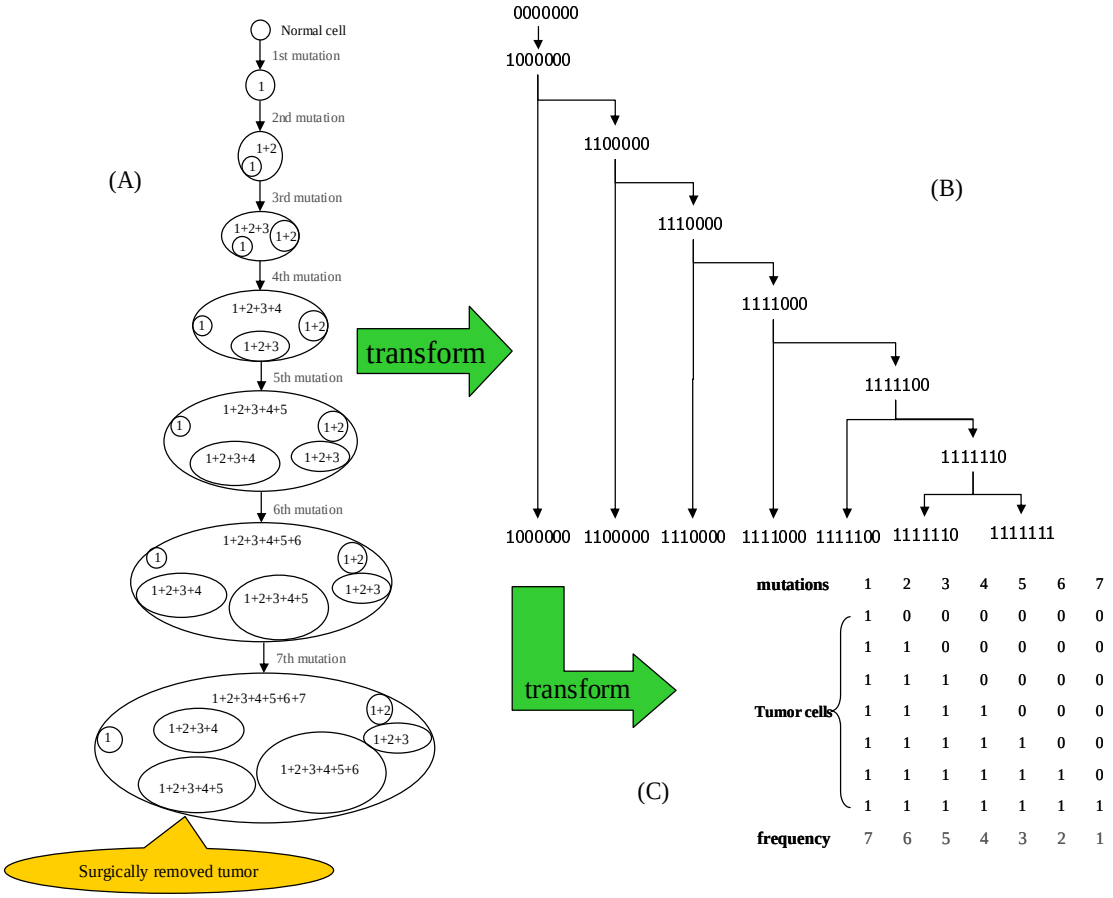

Figure 1. Frequencies of mutations represent the order of genetic changes. (A) Clonal evolution of tumor cells. (B) Binary vector representation of genetic events. (C) Frequency of the genetic events. Figure 1 (A) shows a schematic representation of the clonal evolution of tumor cells and the multi-mutations (3~7 mutations) procedure. Neoplasms originate from a single normal cell. The first mutation ('mutation 1') is thought to start the tumorigenic process. This change will be present in all descendants of the cell that acquired this mutation. The second mutation provides a further growth advantage. Cells containing mutations 1+2 will thus gradually overgrow or replace cells containing only mutation 1. Cells containing mutations 1+2+3 will overgrow cells containing mutations 1+2 and mutation 1. The same thing will happen with mutations 4, 5, 6, 7, etc. Accumulation of subsequent genetic changes is proposed to drive the tumor cell population towards advanced malignancy. In figure 1 (B), the whole biological process was represented in mathematical language (binary: 1 represents that the mutation occurs, 0 means that the mutation does not occur). Figure 1 C shows the calculation of frequency of genetic events from figure 1 B. For example, the frequency 7 of mutation 1 means mutation 1 occurs in all tumor cells; while mutation 7, the frequency of which is 1, occurs only at the last stage in tumor cells. Obviously, mutations with higher frequency are the earlier events during tumorigenesis. Consequently, the frequencies of mutations represent the order of genetic changes. Here, only crucial genetic changes are shown in the figure 1, while many irrelevant genetic changes, which occurred during tumorigenesis because of genetic instability, were omitted. In mathematical terms, these genetic changes are random events and most of these can be filtered out because of their very low frequency (Brodeur et al., 1982).

There are many different topologies for pathway models, such as linear array, ring, binary-tree, H-tree, star, 2D mesh, n-cube, cube-connected cycles, and complete graph (directed acyclic graph models, fully connected graph, etc) (Radmacher et al., 2001).

Simple path models for various cancers have been proposed in recent years. The most successful prototype of such models is the Vogelstein et al. (1988) model, which inferred that the progression of colorectal cancer can be described by a chain of four genetic events which appears to be an indicator of colorectal cancer. Once the first of these

events occurs, it increases the probability of the second event occurring, and when the second event occurs, the probability of the third increases, and so on. Unfortunately, attempts to find similar path models for other types of cancer have not been successful (Kuukasjarvi et al., 1997). Previous studies (Radmacher, 2001) suggest that this is because straight-line chain models do not suffice to capture oncogenesis in other types of cancers than colorectal cancer.

The failure of the simple straight-line path model inspired researchers to attempt more general models, like distance-based tree models, branching tree models, contingency tree models and directed acyclic graph models, etc (Desper et al., 1999; 2000; Radmacher, et al., 2001). Desper et al. (1999) proposed a rooted branching tree model of genetic aberrations in cancer, in which events appear as internal vertices or as leaves. The branching tree model, which was applied to a renal cancer data set, predicts which events occur early (those near the root) and which events may characterize tumor subgroups (those that cluster together in a subtree). Desper et al. (2000) proposed a distance-based phylogenetic tree model. In this model, all events are all leaves of the tree. Leaf nodes represent the detected genetic events; while internal nodes represent hypothetic events which are either the degree of correlation between a pair of events or events hidden from the current detection methods. The alterations predicted to be early are near the root, and sets of events that may characterize subgroups should cluster together in subtrees. The studies on tree models provided promising results.

However, to the best of my knowledge, no paper has so far reported any study about applying pathway models to microsatellite marker data collected during tumorigenesis. In order to gain more knowledge about the specific genetic changes during EAC tumorigenesis, in this paper tree models were applied to data derived by microsatellite analysis, which uses simple sequence repeat (SSR) polymorphisms as markers for detecting allelic imbalance or loss of heterozygosity. Microsatellite analysis has higher resolution compared to comparative genomic hybridization (CGH), which detects segmental DNA copy number changes, while microsatellite analysis allows the identification of specific gene alterations.

The performance of chosen tree models was evaluated by degree of consistency between different tree models and comparing the results to the known knowledge about the EAC progress. The information given by the tree models, i.e. which regions are the most important and the relationships between the important regions, was analyzed.

2 Materials and Methods

All biological experiments (described in chapters 2.1-2.6) had been implemented prior to this project at CMB-genetics at Goteborg University and School of Life Science at the University of Skovde, and all biological experimental data was provided by Professor Karin Klinga Levan.

2.1 Animal model

An inbred animal model is a useful tool for genetic research as it can reduce the complexity of genetic analysis (Balmain & Nagase, 1998; Nagase, et al., 1999). Intercrosses of susceptible rat strains with normal strains are used for establishing associations between genetic markers and quantitative traits distinguishing the EAC phenotype. The inbred BDII/Han rat strain is genetically predisposed to EAC, with an incidence of more than 90% in virgin females as early as 24 months of age (Deerberg and Kaspareit, 1987; Kaspareit-Rittinghausen et al., 1987). BN/Han and SPRD-Cu3/Han (non-EAC-susceptible inbred rat strains) rarely develop endometrial carcinomas with an incidence lower than 10% (Behboudi et al., 2001). BDII females were crossed with BN and SPRD animals. Some of the F1-animals were intercrossed to produce F2 populations. Some males of the F1 offspring were backcrossed to BDII females to produce the backcross progeny. In each cross progeny, EAC developed spontaneously in a number of animals.

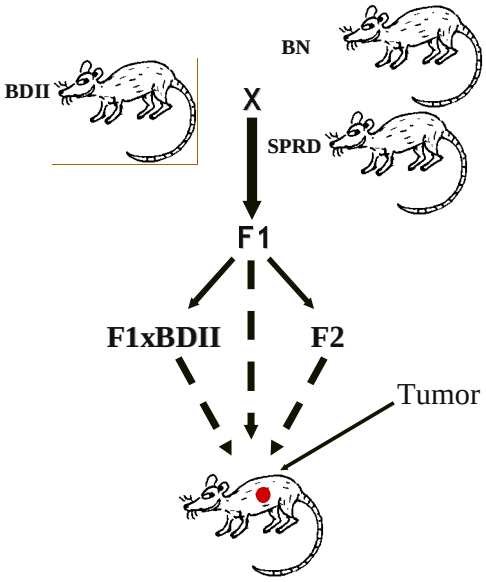

Figure 2. Schematic map of intercrosses of susceptible rat strains (BDII) with normal strains (SPRD and BN).

2.2 Rat chromosome 10 (RNO 10) and EAC development

In cancer research, RNO10 is of particular interest because several important cancer-related genes have been mapped to this chromosome (Chen et al., 1996; Koelsch, 1998; Szpirer et al., 1991). Helou et al. (2000) revealed patterns of nonrandom cytogenetic changes in the tumor materials. One recurrent change was deletion in the proximal part of rat chromosome 10 (RNO 10), usually accompanied by copy number increase in the distal part. Thus, there is the possibility that some of the cancer genes located on RNO10 are involved in EAC development.

2.3 Tissue samples and DNA extraction

The methods have been described by Behboudi (2001) in a previous study. Briefly, uterine tumor specimens and matched normal liver and/or spleen tissue samples were collected from animals for DNA extraction and cell culture establishment (Helou et al.,

2001). All tumors were characterized pathologically. Small pieces of fresh tumor tissue were used to set up primary cell cultures. DNA was extracted from the cell and tissue samples using a standard phenol-based method in the Genepure™ 341 Nucleic Acid Purification System (PE Applied Biosystems, Foster City, CA).

2.4 AI/LOH study

In the previous study (Behboudi, 2001), a total of 112 microsatellite markers were selected from various databases according to published linkage maps (RATMAP, RGD, and The Wellcome Trust Centre for Human Genetics-Rat Mapping Resources), and the corresponding primer pairs were synthesized by a commercial supplier (SIGMA-Genosystem, Cambridge, UK).

During the primary screening of the markers, based on the estimated map position and reproducibility of the PCR reactions, a subset of 47 informative DNA microsatellite markers covering the entire RNO10 was selected for LOH study in the tumor cell cultures. For the analysis of solid tumor DNA, some of the microsatellite markers were discarded and others added in order to generate a panel of 45 polymorphic microsatellite markers distributed along the proximal two-thirds of RNO10. Total only 51 microsatellite markers were selected, because some of them could be used for both TC and ST.

To analyze LOH in solid tumor DNAs, Behboudi (2001) adopted the method of Cawkwell et al. (1993), in which the degree of LOH is determined based on the ratio between the allelic peaks in the tumor sample and those in the corresponding normal tissue sample. Allelic imbalance ratio (AIR) was calculated by equation (1) for each marker in all of the informative paired samples.

$$AIR = \frac{T_1 - T_2}{N_1 - N_2} \quad (1)$$

T_1 and T_2 are the peak area values of two alleles for the tumor sample, while N_1 and N_2 are the peak area values for the tumor and normal sample. Whenever a matched tumor/normal allelic ratio was above 1.0, the inverted ratio was used in order to obtain AIR values in the range of 0–1. A cut-off AIR value of 0.6 was used.

2.5 Genetic linkage map and physical map

A layout of the order of genes (loci) as well as the distance between them is called a genetic map. The distances in a genetic map are determined according to the recombination fraction between two loci. The unit of measure is Morgans (or Centi-Morgans-cM), representing the recombination frequency between the two locations. One cM is one recombination event per 100 meiosis (Broman & Weber, 2000). Thus if two regions of the genome are 10cMs apart, there should be 10 recombination event between these two regions in 100 meiosis. The genetic map distance between two genes therefore determines the frequency at which those genes are expected to recombine. A genetic map can be contrasted to a physical map of a chromosome, where the distance between two

genes is measured in base pairs (or kilo-base pairs: kb). In humans, 1 cM on a genetic map corresponds to about 1-2 Mb of DNA (1 to 2 million base pairs).

2.6 Data categories and preprocessing

The total 96 data samples were made up of 67 tumor solid samples (ST) and 29 tissue culture samples (TC). According to different genetic backgrounds, the data can be divided into the following categories:

Table 1. Data source and categories.

Genetic Background	ST	TC	TOTAL
SPRDCu3xBDII F2 cross	4	4	8
SPRDCu3xBDII Backcross	35	11	46
BNxBDII Backcross	28	14	42
TOTAL	67	29	96

The 51 polymorphic microsatellite markers selected in the previous research (Behboudi, 2001) represent 51 genetic events during the EAC process. As mentioned previously, most of genetic events may appear randomly in tumor progression because of genomic instability. In this work, a method proposed by Brodeur et al. (1982), which estimates the significance of an event by comparing its observed frequency with the frequency obtained in random simulations, was used to filter out random events in tumor progression.

2.7 Tree models and the corresponding algorithms

Since branching tree model and distance-based tree model were proposed by Desper et al., (1999; 2000), they were applied to CGH data for many kinds of cancer and provided promising results. In this thesis, these tree models were adopted for microsatellite marker data of ECA. In tree models, three corresponding algorithm are the maximum weight branching algorithm (MWB), Neighbor-Joining (NJ) and Fitch-Margoliash (FM).

Tree models are based on a probability matrix. Probability is one of most common and useful statistical concept used to analyze biological data. Biological events can be described by probability, and probability can reflect the intrinsic features of many types of biological events. In tree models, for any events *a, b,* let p_a be the probability of event *a*, and let p_{ab} be the probability that event *a* and *b* occur together in the same tumor. For the root, let p_{root} =1, and $p_{a,root} = p_a$.

In the branching tree model, the probability matrix is a weight matrix drawn by the weight function. For each pair (*a*, *b*) of genetic events, the weight function was defined as follows by Desper (1999):

$$w(a,b) = \log(p_{ab}) - \log(p_a + p_b) - \log(p_b) \quad (2)$$

The weight function reflects the desirability of having *b* as a direct descendant of *a* in the tree. Thus, it includes the likelihood ratio for *a* and *b* occurring together. The asymmetry property of the weight matrix shows that when event *a* occurs more often than event *b*, it is more likely to have an edge from *a* to *b* than vice-versa (Desper, 1999).For each pair of genetic events, the closer the relationship, the higher the weight score.

Based on the weight matrix, the maximum weight tree can be pursued. Consequently, we adopted the MWB (Edmonds, 1967; Karp, 1971) that finds the rooted tree where the total weight (the sum of the weights of all edges in the tree) is maximized. The oncotrees software[1] was used to construct branching tree models.

In the distance-based tree model, the probability matrix is a distance matrix drawn by the distance function. For each pair (*a*, *b*) of genetic events, the distance function was defined as follows:

$$d(a,b) = -2\log p_{ab} + \log p_a + \log p_b \quad (3)$$

The distance function reflects that the distance of each pair (*a*, *b*) is a symmetric and the distance of the same events *(a, a)* is 0. For each pair of genetic events, the closer the relationship, the lower is the distance score.

Based on the distance matrix, two common algorithms were used to construct phylogenetic trees. NJ (Saitou and Nei, 1987) constructs a phylogenetic tree by using the following equation:

$$S_{mn} = \frac{\sum (d_{im} + d_{in})}{2(N-2)} + \frac{d_{mn}}{2} + \frac{\sum d_{ij}}{N-2} \quad (4)$$

S_{mn} represents the new distance between each vertex pair (*m*, *n*); *i*, *j* represent all vertices except *m* and *n*. $\sum (d_{im} + d_{in})$ is the sum of distances from *m* to *i* and from *n* to *i*. *N* is the number of vertices.

The FM algorithm (Fitch and Margoliash, 1967), which is used to compute actual branch lengths, estimates a phylogenetic tree assuming additivity of branch lengths.
In a phylogenetic tree (distance-based tree), all detected genetic events which occur are leaves of the tree, while internal nodes represent hypothetic events. The events that are predicted to be early are near the root, and sets of events that may characterize subgroups should cluster together in subtrees (Desper et al., 2000).

The programs FITCH and NEIGHBOR from the PHYLIP package v3.6 (Felsenstein et al., 1989) were used to construct distance-based tree models.

2.8 Evaluating tree models

In order to validate the reliability, derived tree models were evaluated from the following three aspects

[1] http://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/cgh. html

2.8.1 Using algorithms to evaluate tree models

There are three algorithms for deriving tree models. Each algorithm makes its own assumptions concerning tumor evolution. If several algorithms give essentially the same tree, the topology is not very dependant on the evolutionary assumptions. Consequently, this gives more confidence in the tree model.

2.8.2 Statistical evaluation of tree models

Since trees derived by different algorithms were compared and evaluated in this work, it becomes important to assess the confidence of the different tree models. If different trees disagree to some extent, we need to know which can be trusted more. Therefore, a bootstrapping procedure was used to analyze tree models statistically and evaluate their degree of confidence.

SEQBOOT is a general bootstrapping and data set translation tool included in the PHYLIP 3.6 package (Efron and Tibshirani, 1993). In this thesis, it was used to generate 100 data sets that are resampled versions of the input data set. The reliability of a monophyletic clustering can be judged by looking at the percentage of bootstrap trees that support each node. In the context of phylogenetics, monophyletic groups supported by bootstrap values of around 50% or lower are usually not entirely stable (Efron et al., 1996). This bootstrapping procedure only corrects for sampling errors, not for systematical errors, such as errors in the evolutionary assumptions of the method.

2.8.3 Evaluating tree models by comparing with previous knowledge

There are many public databases concerning the rat genome. In this thesis, the information was derived from the following databases. Physical locations of each linkage position were derived from http://www.ratmap.org. Corresponding genes were identified by using http://www.ensembl.org or http://www.ncbi.nlm.nih.gov/mapview/. The primary function of genes was found through http://rgd.mcw.edu/genes/ and NCBI databases.

The lengths of the markers used in this work vary from 98 bp to 218 bp. Generally, the length of a microsatellite markers is about 100~200 bp (Subramanian et al., 2002). But for many genes, the length can reach several kilo base pairs (kb), and in some cases even several hundred kb (Casey, 1992). For instance, the length of the Ebf gene is 393 kb. Obviously, a marker is a sign of a gene. The farther the distance between the gene and the marker, the lower the reliability of the relationship is. It is tricky problem that how long distance can be reflected by a marker. So far, there are no certain criteria for this problem.

The range covered by markers can be described by a genetic linkage map or a physical map. The relative genes can be found according to the linkage map position on which used markers located. For the physical map, the length of the whole chromosome 10 is about 110 Mb. In order to ensure high resolution, only a 50 kb range was extended around each marker. Compared to previous study (Nordlander et al., in press) in which 10 Mb range around each region midpoint was used, 50kb was quite smaller. Consequently, the relationship between genes in the extended 50kb ranges and corresponding markers is highly reliable.

Results and Analysis

The method of Brodeur et al. (1982) selected the non-random genetic changes for the different data sets. No non-random events were found in data sets based on the SPRDCu3xBDII F2 genetic background. Consequently, no tree-model was built for the SPRDCu3xBDII F2 data set.

The frequencies of non-random events in the different data sets are shown in Figure 3 in which linkage positions of detected markers used to represent genetic events were derived from the Ensembl database [1]. Position 48 occurs 9 times in the 9 data sets. According to descending order of frequency, with ties shown in parentheses, the events were 48, (46, 40), (49, 61, 62, 63), (66, 56, 28), 25, (26, 27, 30), (36, 37, 42, 60), (68, 79, 100) and (32, 73). Obviously, the genetic events with high frequency are likely to be more important. An evidence to support this is that no cancer-related genes were found on the positions with frequency lower than 40%. It is easy to see that there are three regions that include the genetic events with high frequency (see Figure 3).

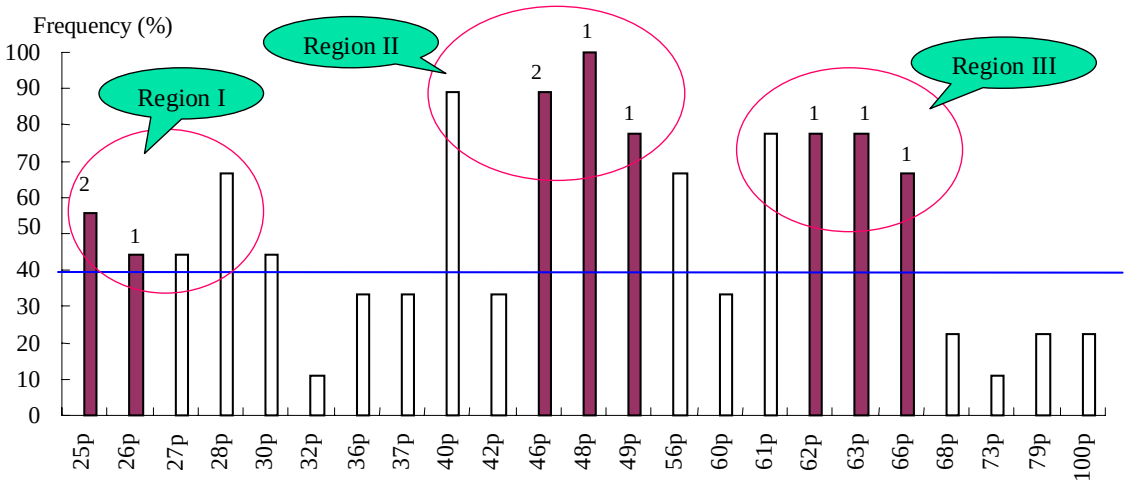

Figure 3. Genetic events in linkage map position. The frequency of non-random events in all data sets. Numbers above filled bars show how many cancer-related genes were found on positions. Unfilled bars correspond to positions where no cancer-related genes were found. The linkage map regions harboring the cancer-related genetic events indicated by circles.

In total, 27 tree models were derived (9 branching tree models and 9 distance-based tree models based on NJ and FM, respectively). Common features could be seen in the branching tree based on the total data (Figure 4). Positions that were nearby on the chromosome also appear together in the trees, regardless of which method was used. There were three major regions: region I (25, 26, 27, 28 and 30), region II (40, 46, 48 and 49), and region III (61, 62, 63 and 66). In addition, there were two subordinate regions: region IV (36 and 37) and region V (73 and 79). Region I and III branch from region I, in other words, region II was causal for region I and III, but there was no causal relationship between Region III and Region I. Region II was the earliest region, region III was the second, and region I was the third. Tree models showed inconsistencies among each region, i.e. in region II, the orders of 61, 62, 63 and 66 differed in the tree models.

[1] http://www.ensembl.org/Rattus_norvegicus

However, subtrees showed high consistency among tree models (Table 2). Region I occurred 18 times in 27 tree models; region II occurred in 25 out of 27, region III 21/27, region IV 9/27, and region V 6/27.

Table 2. Consistency in subtrees of tree models. Subtrees were included in 27 tree models built by three algorithm and 9 data sets.

	SPRDCu3xBDII Backcross			BNxBDII Backcross			TOTAL		
	ST	TC	TOTAL	ST	TC	TOTAL	ST	TC	TOTAL
MWB	I II III	II	I II III	II III	II	I II III IV	I II III V	I II III IV	I II III IV V
NJ	I II III	II	I III	II III	II	I II III IV	I II III V	I II III IV	I II III IV V
FW	I II III	II	I III	II III	II	I II III IV	I II III V	I II III IV	I II III IV V

Figure 4 also shows all known cancer-related genes located in the studied positions. N-methylpurine-DNA glycosylase (Mpg) and Tsc1 (tuberous sclerosis 1) are located in position 25; LOC363557, similar to p19, is located in position 26; early B-cell factor (Ebf) and interleukin 12b (Il12b) are located in position 46; LOC303073, similar to Tp53 inducible protein, is located in position 48; the LOC303076 gene is located in position 49; tumor protein p53 (Tp53) is located between position 62 and 63; and neurofibromatosis 1 (Nf1) is located in position 66.

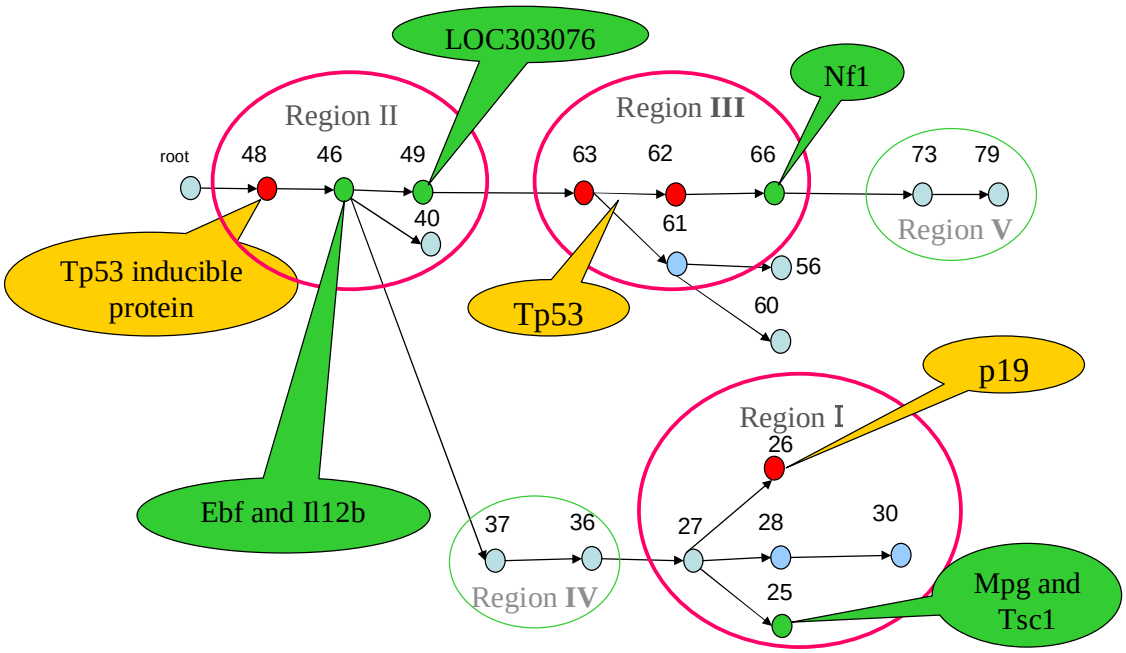

Figure 4. Branching tree built on whole data set. Occurrences of events generating arrows increase the probability of occurrence of events to which the arrows point. Events which are closer to the root are more likely to be the earlier events during tumorigenesis. 5 regions are illuminated. The locations of known cancer-related genes indicated by filled callouts.

Tree models derived from FM and NJ were quite similar. There was only one inconsistency regarding position 48 in the BNxBDII backcross genetic data set, i.e. that

the distance from the root to position 48 was shorter in trees derived by the NJ algorithm than in those derived by the FM algorithm.

The bootstrapping tree model based on the total data set of NJ showed that there was 76.9% support for 48 and the root occurring together, 65.8% support for positions 79 and 73 occurring together, and 59.9% support for positions 25, 26, 27, 28 and 30 occurring together. This result was consistent with the tree models derived by the other algorithms.

4 Discussion

In some cases, there were some differences between tree models. When looking into all the inconsistencies, the reason seemed to be that the information carried by the frequency matrices was not sufficient to distinguish them from each other. For example, in the distance-based tree models (Total and ST) based on the BNxBDII backcross, there was an inconsistency regarding positions 46 and 48. Position 48 occurs earlier than 46 in NJ trees, while the order between these events was reversed in FM trees.

As mentioned previously, a cut-off AIR value of 0.6 was used in this thesis based on the previous study (Behboudi et al., 2001). Events with AIR values above 0.6 were represented by 0 (no AI occurs), while in contrast events with AIR values below 0.6 were represented by 1 (AI occurs). When using 0 and 1 in place of real values between 0 and 1, which reflect the degree of chromosomal aberration, the information content is radically reduced.

Because this thesis is based on previous studies (Nordlander et al., in press), it was important to compare the current results with the previous ones. Four of the same recurrent AI regions (region I (25~28), region II (46~48), region III (61~66) and region IV (79~100)) were found in both the previous and these current results. But in this thesis, more detailed information of such regions was produced. For the whole data set, the same four regions were found. But for TC data sets, only Region II was found, while the other three regions were missing. For the total data set and the ST data set based on the SPRD backcross, the four regions were found. For the total data set based on the BN backcross, three regions were found, while region IV was missing. For the ST data set based on the BN backcross, only two regions were found, while region I and IV were missing.

The reason for there being a smaller number of genetic events in TC than in ST is an intrinsic characteristic of TC. The process of tissue culture produces a more homogenous cell population, because one or a few sub-clones with better growth capacity in vitro gradually overwhelm the others (Macoska et al., 2000). Genetic events may thus be more complex in ST than in TC because of the heterogeneity of the tumor cell population in the solid tumors or because of admixture of normal cells (Hanahan & Weinberg, 2000).

Hence, the genetic events in TC may be more clear and closer to the genetic hypothesis of cancer development than in ST. The phenomenon that Region II always occurs in spite of different data sets supports that Region II was the most important region on chromosome 10 during the EAC process.

Cancer-related genes located in only three main regions constitute supportive biological evidence for the theoretic feasibility of genetic events with high frequency being more important. The relationships among p19, Tp53 and Tp53 inducible protein provided additional supportive evidence for the reliability of the results.

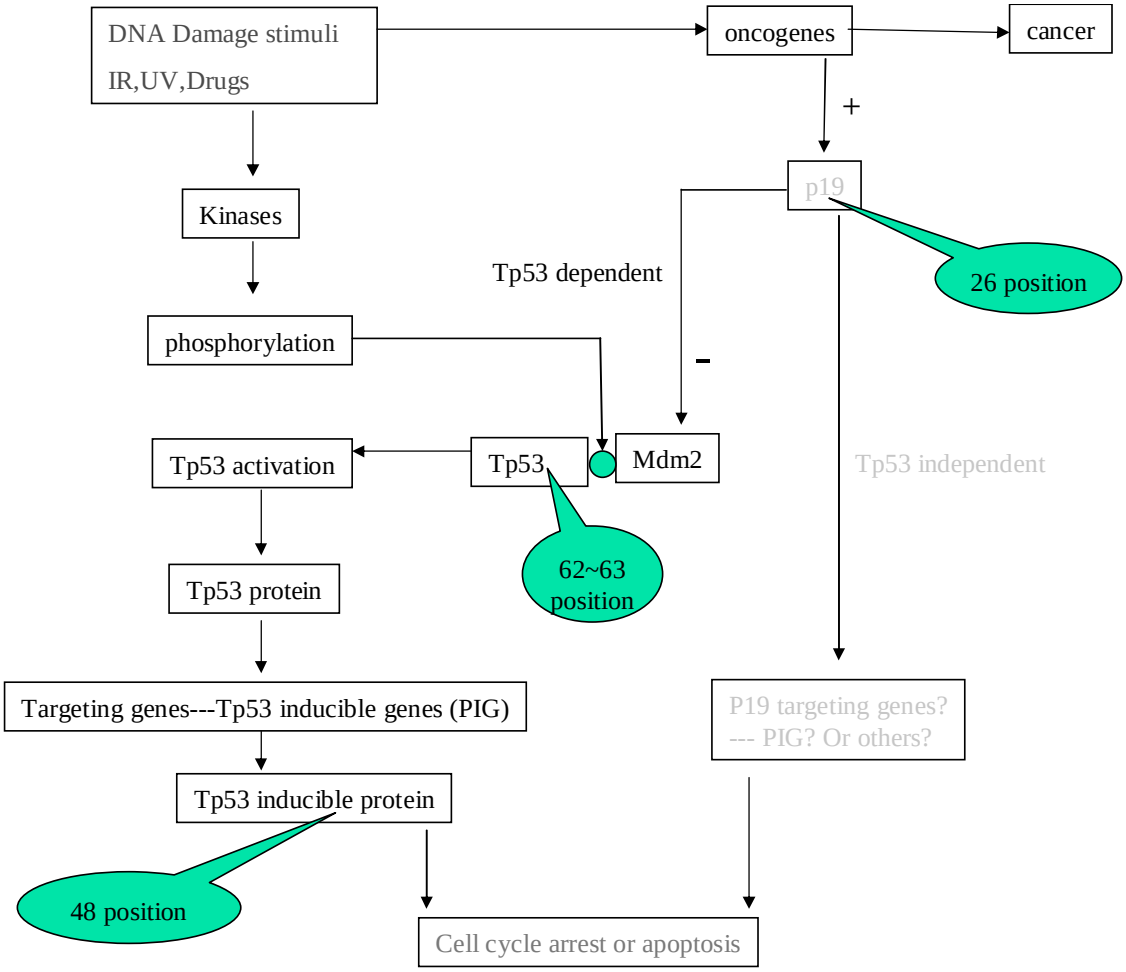

Figure 5. The relationship between the Tp53, Tp53 inducible protein and P19. DNA damaging agents (like radiation) induce the activation of kinases (such as ATM and DNA-PK) that can phosphorylate a critical serine residue in the Mdm2-binding domain of Tp53. When Tp53 is phosphorylated here, it can no longer bind to Mdm2. Tp53 recognizes when the cell has sustained DNA damage and halts the cell cycle so that the cell can repair the damage or in many cases just directs the cell to die (apoptosis). Another mechanism (Dlamini et al., 2004) to inhibit Mdm2 is by oncogenes, constitutively active mutant proteins that continually tell the cell to grow (such as Ras). Oncogenes inhibit Mdm2 by inducing expression of a tumor suppressor protein called p19 ARF. P19 ARF is a key regulator of Tp53 stability and activation. In addition, another study (Weber et al., 2000) demonstrated that p19ARF can act independently of the Mdm2-Tp53 axis in tumor surveillance. p19ARF can arrest the cell division cycle in the G1 phase without Tp53 and Mdm2 genes. Thus, in the absence of Mdm2, p19ARF interacts with other targets to inhibit cell proliferation.

In position 48, estimated gene LOC303073 was found. Regarding the function of LOC303073, it is similar to the Tp53 inducible protein. Dlamini et al., (2004) reported that Tp53 inducible protein is a Tp53 apoptosis effector with a death domain (DD). DD-containing Tp53 -inducible protein with a DD (PIDD) is likely to participate in the apoptotic signal pathway. Tumor protein Tp53 inducible proteins are a group of Tp53

target genes, also referred to as Tp53 inducible genes (PIG). The PIGs are direct downstream targets that can be activated by Tp53 and trigger the apoptosis in the Tp53-dependent apoptosis pathway. Mutations that inactivate the Tp53 tumor suppressor protein are the most common genetic aberrations known to occur in human cancers. The best-described biological functions of Tp53 are the induction of cell cycle arrest and apoptosis in response to cellular stresses. It is known that Tp53 is a transcription factor that binds to specific sequences in DNA and activates the transcription of the target gene. In recent years, a number of Tp53 target genes have been identified that mediate apoptosis including Fas, DR5/Killer, Bax, and so on. The process of cell apoptosis initiated by Tp53 will be stopped by either aberration of Tp53 or PIGs (Lin et al., 2000; Liang et al., 2004; Polyak et al., 1997), so genetic aberration in this region may damage the apoptosis signal pathway mediated by Tp53 (Liang et al., 2004). The relationship between Tp53 and PIGs (Figure 5) clearly explains why there is inconsistency regarding the order between position 48 and 62 and 63 in tree models, where some tree models treat position 48 harboring the gene of Tp53 inducible protein as earlier events than position 62 and 63 harboring Tp53, while the order is reversed in other tree models. The loss of PIGs can result in oncogenesis without the activation from Tp53, this can be observed in tree models derived from TC data sets.

In position 49, the cancer-related gene LOC303076 was found, which is similar to nuclear protein. Although the function of LOC303076 is not clear, most nuclear proteins can influence the activities of DNA in the cell nucleus.

Ebf and Il12b genes were found at position 46. The early B-cell factor (Ebf)-associated zinc-finger protein (EBFAZ) binds to and negatively regulates Ebf, which is a basic helix-loop-helix transcription factor required for B-cell lineage commitment and development. EBFAZ is a frequent target of retroviral integration in B-cell lymphomas. Integrations at Ebfaz and Evi3 are mutually exclusive, suggesting that they function in the same tumor pathway. Lymphomas with integrations at Ebfaz or Evi3 express the pre-B-cell-specific marker immunoglobulin lambda chain 5, and contain immunoglobulin heavy-chain rearrangements, suggesting that they are blocked at an early B-cell stage (Warming et al, 2004).

IL-12, a p35 and p40 heterodimer, initially defined as natural killer-stimulating factor (NKSF), is now known to drive multiple integrate steps in the immune response. IL-12 stimulates enhanced cytotoxicity by NK cells and T cells and induces IFNy release by NK cells and T cells. In vivo mouse studies indicate that IL-12 facilitates the proliferation of T cells of Thl type and inhibits T cells of the Th2 type. These properties suggest that IL-12 might be able to activate Thl-dependent local help and support the generation of a tumor-specific immune response. Previous work has demonstrated that in vivo-administered IL-12 has a potent antitumor effect against various kinds of transplantable tumor cells (Brunda and Gately, 1994; Imagawa et al., 2004).

Tp53 was found between positions 62 and 63. In the cell cycle, the Tp53 signal pathway is responsible for the repairing of cell DNA damage. DNA damage (e.g. by radiation), can induce the activation of kinases (such as ATM and DNA-PK) that can phosphorylate

a critical serine residue in a binding domain of Tp53. When Tp53 is phosphorylated, it is able to relieve the inhibition of Tp53. Tp53 recognizes when the cell has sustained DNA damage and halts the cell cycle so the cell can repair the damage or in many cases just directs the cell to die (apoptosis). If Tp53 was lost, or inhibited by oncogenes, constitutively active mutant proteins continually tell the cell to grow, and carcinoma may occur (May et al., 1999; Oren, 1999; El-deiry, 1998).

In position 66, the Nf1 gene was found. Neurofibromas are common tumors that arise from elements in the peripheral nervous system. The association of neurofibromatosis with gynecologic malignancies is quite uncommon, with only a few cases reported. Nf1 represents also a major risk factor in the development of several malignancies such as malignant peripheral nerve sheath tumors, optic gliomas and leukemias (Murat et al, 2004).

Mpg and Tsc1 genes were found in position 25. Mpg (N-methylpurine-DNA glycosylase), is an enzyme that cleaves 3-methyladenine and 7-methylguanine residues from DNA, and takes part in the biological process of DNA repair and base-excision repair. If DNA damage can not be repaired, it may lead to carcinogenesis, cell death, and aging. N-methylpurine-DNA glycosylase (MPG) is a DNA repair enzyme that removes N-alkylpurine damage, in normal, malignant, and immortalized breast epithelial cells, and breast cancer cell lines. The investigation for the expression of MPG in human breast cancer showed that the N-methylpurine-DNA glycosylase messenger RNA was overexpressed up to 24-fold in breast cancer as compared to normal primary breast epithelium. MPG protein expression and glycosylase activity were evaluated in the breast cancer cell lines (Cerdaa et al., 1998).

Loss of heterozygosity (LOH) in the tuberous sclerosis 1 (TSC1)-gene-associated region was frequently observed in adenocarcinoma (AC) of the lung and its putative precursor lesion, atypical adenomatous hyperplasia (AAH) (Takamochi et al., 2004). Some novel tumor suppressor gene for AC of the lung may be present in this region, and the TSC1 gene located on chromosome 9q34 in human. The TSC1 and the TSC2 genes are known to act as Knudson-type tumor suppressor genes in the pathogenesis of TSC-related hamartomas. If the TSC1 gene itself is also responsible for the development of AC of the lung, biallelic inactivation of the TSC1 gene, allelic loss and mutation, would frequently be seen in ACs (Kajino and Hino, 1999).

LOC363557, similar to p19, is located in position 26. Two proteins encoded by the *INK4a/ARF* locus, $p16^{INK4a}$ and $p19^{ARF}$, functionally interact with the retinoblastoma protein (RB) and Tp53 tumor suppressors (Figure 5), respectively. These four proteins are among the most frequently disrupted genes in human cancer (Ko and Prives, 1996). The $p19^{ARF}$ tumor suppressor acts as a sensor of hyperproliferative signals emanating from oncoproteins and inducers of S-phase entry, such as Myc, Ras, and E2F-1. In turn, $p19^{ARF}$ triggers Tp53 -dependent growth arrest in the G1 and G2 phases of the cell cycle or, in the presence of appropriate collateral signals, sensitizes cells to apoptosis. ARF binds directly to Mdm2, sequestering it in the nucleolus and enabling transcriptionally

active Tp53 to accumulate in the nucleoplasm. ARF also limits the ability of Mdm2 to ubiquitinate Tp53 in vitro and in vivo (Kalinichenko et al., 2004).

Weber et al. (2000), proposed that the p19 (ARF) tumor suppressor has Tp53 independent functions (see figure 5). Individual TKO (triple knock out) mice nullizygous for ARF, p53, and Mdm2 develop multiple tumors at a frequency greater than those observed in animals lacking both Tp53 and Mdm2 or Tp53 alone, demonstrating that p19ARF can act independently of the Mdm2-Tp53 axis in tumor surveillance. Weber et al found that reintroduction of ARF into TKO mouse embryo fibroblasts (MEFs), but not into those lacking both Tp53 and ARF, arrested the cell division cycle in the G1 phase. Inhibition of the retinoblastoma protein had no effect on the ability of ARF to arrest TKO MEFs. Thus, in the absence of Mdm2, p19ARF interacts with other targets to inhibit cell proliferation (Figure 5).

The fact that p19 has Tp53-independent function is supportive evidence for the conclusion drawn from the tree models that there was no causality between region I and region III.

5 Conclusions

The high degree of consistency of results among the different methods, previous studies and previous knowledge about tumorigenesis constitute supportive evidence for the reliability of tree models.

Although the existence of inconsistencies in tree models prevents us from drawing clearer conclusions about each position, there are still were common features in tree models: Firstly, positions that are nearby on the chromosome also appear together in the trees, regardless of which method was used. Secondly, there were three major regions: region I (25, 26, 27, 28 and 30), region II (40, 46, 48 and 49), and region III (61, 62, 63 and 66), and two subordinate regions: region IV (36 and 37) and region V (73 and 79). Region I and III branch from region I, in other words, region II was causal for region I and III, but there was no causal relationship between region III and region I. Region II was the earliest region, region III was the second, and region I was the third. Thirdly, there was a high degree of consistency between subtrees in tree models.

The results drawn from tree models show directions for further biological experiments. The functions of some genes (such as LOC303076 and LOC303073) harbored in important positions and the relationships among them were not clear. On the other hand, not much information was found in some important position (for example, 28, 40 and 61).

6 Future work

The results presented in this thesis indicate that the future for applying pathway models in marker data of EAC is very promising, but there is still much work to be done in developing and optimizing the method.

6.1 Further refining the results.

As discussed previously, information useful for distinguishing between different genetic events was radically reduced by using a cut-off AIR value of 0.6. (Behboudi et al., 2001). To improve the performance of tree models, we can refine the current results by using the real AIR value in future work. (Mention if this requires modification of algorithms).

6.2 Improving the algorithm

Traditional algorithms were used in this thesis, but other algorithms may also be suitable for inferring models of tumorigenesis. For example, Markov chains seem suitable for mimicking the different tumor stages. The more different angles the cancerous process is observed from, the more of the truth of EAC tumorigenesis will we come to know.

6.3 Further biology experiment

Firstly, maybe, it is good to study further the shifting importance of regions with respect to different genetic backgrounds. The regions shift when data sets change. Models derived from the ST data set of SPRD include one distal region IV, while models derived from the ST data set of BN lack region I and region IV. In addition, models derived from TC data sets only have region II. These phenomena give hints for further study.

Secondly, on the one hand, only focusing on chromosome 10 made the problem of tree building easier to solve; on the other hand, the information about interrelationships between different chromosomes is omitted. Consequently, the contribution is limited by only working on chromosome 10. If we work on more available data from the whole genome, it is possible that we will get a panoramic picture in the future.

Last but not least, the hypothesis given by tree models give hints to biologists, at the same time, it needs to be verified by further biological experiments. For example, the function of some genes harbored in some important positions given by tree models is not clear, such as LOC303076 which is an unknown nuclear protein and LOC303073 which is similar to Tp53 inducible protein. Once we know more functions of the genes and the interactions between them, the hypothesis may be verified.

Acknowledgments

I would like to express my sincerest gratefulness to my thesis adviser, Prof. Bjorn Olsson and Prof. Karin Klinga Levan. During whole thesis, they have been giving me valuable academic guidance. I have benefited and will still benefit significantly from their sharp vision on scientific research and dedication to the profession.

References

Balmain A. (2002) Cancer as a complex genetic trait: tumor susceptibility in humans and mouse models. Cell 108: 145-152.

Balmain A. and Nagase H. (1998) Trends in genetics : TIG. 14: 139-144.

Behboudi A., Levan G., Hedrich HJ., Klinga-Levan K. (2001) High-density marker loss of heterozygosity analysis of rat chromosome 10 in endometrial adenocarcinoma. Genes chromosomes cancer 32(4):330-341.

Brodeur GM., Tsiatis AA., Williams DL., Luthardt FW., Green AA. (1982) Statistical analysis of cytogenetic abnormalities in human cancer cells. Cancer genetics and cytogenetics. 7(2):137-152.

Broman KW., Weber JL. (2000) Characterization of human crossover interference. American journal of human genetics 66(6):1911-1926.

Brunda MJ. and Gately MK. Antitumor Activity of Interleukin-12. (1994) Clinical Immunology and Immunopathology 71(3):253-255.

Cawkwell L., Bell SM., Lewis FA., Dixon MF., Taylor GR., Quirke P. (1993) Rapid detection of allele loss in colorectal tumours using microsatellites and fluorescent DNA technology. British journal of cancer 67(6):1262-1267.

Cavanagh D., Fiorica JV., Hoffman MS., Durfee J., Nicosia SV. (1999) Adenocarcinoma of the endometrium: an institutional review. Cancer Control JMCC 6:354-360.

Cerdaa SR., Turka PW., Thorb AD., Weitzman SA.. (1998) Altered expression of the DNA repair protein, N-methylpurine-DNA glycosylase (MPG), in breast cancer. FEBS Letters 431(1):12-18.

Chen KS., Shepel LA., Haag JD., Heil GM., Gould MN. (1996) Cloning, genetic mapping and expression studies of the rat Brca1 gene. Carcinogenesis 17(8):1561-1566.

Deerberg F. and Kaspareit J. (1987) Endometrial carcinoma in BD II/Han rats: model of a spontaneous hormone-dependent tumor. Journal of the National Cancer Institute 78(6):1245-1251.

Denise Casey. Primer on Molecular Genetics. Human Genome 1989–90 Program Report. U.S. Department of Energy, Office of Energy Research, Office of Health and Environmental Research. Washington, DC 20585. 1992, p7

Desper R., Jiang F., Kallioniemi OP, Moch H, Papadimitriou CH, Schaffer AA. (1999) Inferring tree models for oncogenesis from comparative genome hybridization data. Journal of computational biology : a journal of computational molecular cell biology. 6(1):37-51.

Desper R., Jiang F., Kallioniemi OP, Moch H, Papadimitriou CH, Schaffer AA. (2000) Distance-based reconstruction of tree models for oncogenesis. J Comput Biology 7(6):789-803.

Dlamini Z., Mbita Z. and Zungu M.(2004) Genealogy, expression, and molecular mechanisms in apoptosis. Pharmacology & Therapeutics 101(1):1-15

Edmonds J.(1967) “Optimum Branchings,”]ournal of Research of National Bureau of Standards, 71B:233-240.

El-Deiry W. and Semin S..(1998) Regulation of Tp53 downstream genes. Seminars In Cancer Biology 8(5):345-357

Efron B. and Tibshirani R. J. (1993). An introduction to the Bootstrap. Chapman and Hall.

Efron B., Halloran E. and Holmes S. (1996) Bootstrap confidence levels for phylogenetic trees. Proceedings of the National Academy of Sciences of the USA 93:13429-13434.

Esteller M., Xercavins J. and Reventos J. (1999). Advances in the molecular genetics of endometrial cancer (Review). Oncology Reports 6(6):1377-1382.

Felsenstein J. (1989) PHYLIP-phylogeny Inference Package (version 3.2). Cladistics : the international journal of the Willi Hennig Society 5:164-166.

Fitch W.M., and Margoliash, E. (1967) Construction of phylogenetic trees. Science 155: 279~284. http://ecoevo.bio.uci.edu/Faculty/Fitch/Fitch.html

Hanahan D., and Weinberg R.A. (2000) The hallmarks of cancer. Cell 100: 57-70.

Hansen M. F., and Cavenee W. K. (1987) Genetics of cancer predisposition. Cancer Research 47: 5518-5527.

Helou K., Walentinsson A., Beckman B., Samuelson E., Hedrich H. J., Szpirer C., Klinga-levan K., and Levan G. (2000). Comparative genome hybridzation (CGH) analysis in rat uterine endometrial carcinoma. Rat Genome 6: 78.

Helou K., Walentinsson A., Beckmann B., Johansson A., Hedrich HJ., Szpirer C., Klinga-Levan K., Levan G. (2001) Analysis of genetic changes in rat endometrial carcinomas by means of comparative genomic hybridization. Cancer genetics and cytogenetics. 127(2):118-127.

Imagawa Y., Satake K., Kato Y., Tahara H. and Tsukuda M. (2004) Antitumor and antiangiogenic effects of interleukin 12 gene therapy in murine head and neck carcinoma model. Auris Nasus Larynx. 31(3):239-245

Kalinichenko VVl, Michael L., Major ML., Wang X., Petrovic V., Kuechle J., Yoder HM., Dennewitz MB., Shin B., Datta A., Raychaudhuri P., Costa RH.. (2004) Foxm1b transcription factor is essential for development of hepatocellular carcinomas and is negatively regulated by the p19ARF tumor suppressor. genes & development 18:830-850.

Kajino K. and Hino O. TSC1 and TSC2 gene mutations in human kidney tumors. (1999) Contrib Nephrol. 128:45-50.

Karp R.M. (1971) A simple derivation of Edmonds' algorithm for optimum branching. Networks 1: 265~272.

Kaspareit-Rittinghausen J., Deerberg F., Rapp K. (1987) Mortality and incidence of spontaneous neoplasms in BDII/Han rats. Zeitschrift fur Versuchstierkunde 30(5-6):209-216.

Koelsch BU. (1998) Assignment of Erbb2 to rat chromosome band 10q32.1 by in situ hybridization. Cytogenetics and Cell Genetics 81(3-4):182.

Ko LJ., and Prives C. p53: puzzle and paradigm.Genes Development. 10(9):1054-1072.

Kuukasjarvi T., Karhu R., Tanner M., Kahkonen M., Schaffer A., Nupponen N., Pennanen S., Kallioniemi A., Kallioniemi OP., Isola J. (1997) Genetic heterogeneity and clonal evolution underlying development of asynchronous metastasis in human breast cancer. Cancer Research. 57(8):1597-1604.

Liang XQ., Cao EH., Zhang Y., and Qin JF. (2004) A P53 target gene, PIG11, contributes to chemosensitivity of cells to arsenic trioxide. FEBS Letters 569(1-3):94-98

Lin Y., Ma W., and Benchimol S. (2000) Pidd, a new death-domain-containing protein, is induced by Tp53 and promotes apoptosis. Nature genetics 26(1):122-127.

Lothe RA. and Blomhoff HK. (1998) Tumor suppressors--genes and proteins. Tidsskrift for den Norske laegeforening 118(12):1887-1892.

Macoska JA., Beheshtid B., Rhime JS., Hukku B., Lehr J., Pienta KJ., Squire JA. (2000) Genetic Characterization of Immortalized Human Prostate Epithelial Cell Cultures--- Evidence for Structural Rearrangements of Chromosome 8 and i(8q) Chromosome Formation in Primary Tumor-Derived Cells. Cancer Genetics and Cytogenetics 120(1):50-57

May P. and May E. (1999) Twenty years of Tp53 research: structural and functional aspects of the Tp53 protein. Oncogene 18(53):7621-7636

Melendez B., Diaz-Uriarte R., Cuadros M, Martinez-Ramirez A, Fernandez-Piqueras J, Dopazo A, Cigudosa JC, Rivas C, Dopazo J, Martinez-Delgado B, Benitez J. (2004) Gene expression analysis of chromosomal regions with gain or loss of genetic material detected by comparative genomic hybridization. Genes, chromosomes & cancer 41(4):353-365.

Radmacher Michael D., Simon Richard, Desper Richard, Taetle Raymond, Schaffer Alejandero A. and Nelson Mark A.. (2001) Graph Models of Oncogenesis with an Application to Melanoma. Journal of theoretical biology 212:535-548.

Moch H. and Mihatsch MJ. (2002) Genetic progression of renal cell carcinoma. Virchows Archiv : an international journal of pathology 441(4):320-327.

Murat A., Kansizb F., Kabakusc N., Kazez A., Ozercan R. (2004) Neurofibroma of the breast in a boy with neurofibromatosis type 1. Clinical Imaging 28(6):415-417

Nagase H., Mao J.H. and Balmain A. (1999) Proceedings of the National Academy of Sciences of the United States of America 96: 15032-15037.

Nordlander C., Behboudi A., Levan G and Klinga Levan K. (in press) Four segments show allelic imbalance on chromosome 10 in rat endometrial adenocarcinomas.

Noumoff JS. and Faruqi S. (1993) Endometrial adenocarcinoma. Microscopy research and technique.. 25(3):246-254.

Olah E. (1999) Hereditary neoplastic diseases (genetic predisposition and cancer syndromes). Orvosi hetilap 140(9):451-466.

Oren M. Regulation of the Tp53 tumor suppressor protein. (1999) The Journal Of Biological Chemistry 274(51):36031-36034

Pathak S. (1999) Cytogenetic abnormalities in cancer: with special emphasis on tumor heterogeneity. Cancer and Metastasis Review 8:299-318.

Polyak K., Xia Y., Zweier JL., Kinzler KW., Vogelstein B. (1997) A model for p53-induced apoptosis. Nature 389(6648):300-305

Saitou N,. and Nei M. (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular biology and evolution 4: 406~424.

Subramanian S, Madgula VM, George R, Mishra RK, Pandit MW, Kumar CS, Singh L.(2002) MRD: a microsatellite repeats database for prokaryotic and eukaryotic genomes. Genome Biology 3(12):PREPRINT0011.

Suzuki A., Fukushige S., Nagase S., Ohuchi N., Satomi S., Horii A.. (1997) Frequent gains on chromosome arms 1q and/or 8q in human endometrial cancer. Human Genetics 100(5-6):629-636.

Szpirer C., Riviere M., Szpirer J., Hanson C., Levan G., Hendy GN. (1991) Assignment of the rat parathyroid hormone-like peptide gene (PTHLH) to chromosome 4: evidence for conserved synteny between human chromosome 12, mouse

chromosome 6, and rat chromosome 4. Cytogenetics and Cell Genetics. 56(3-4):193-195.

Takamochi K, Ogura T, Yokose T, Ochiai A, Nagai K, Nishiwaki Y, Suzuki K, Esumi H. (2004) Molecular analysis of the TSC1 gene in adenocarcinoma of the lung. Lung Cancer. 46(3):271-281.

Tomlinson I, Sasieni P, Bodmer W. (2002) How many mutations in a cancer? American journal of pathology 160(3):755-758.

Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, Leppert M, Nakamura Y, Wmutatione R, Smits AM, Bos JL. (1988) Genetic alterations during colorectal-tumor development. The New England journal of medicine 319(9):525-532.

Vogelstein B, Kinzler KW. (1993) The multistep nature of cancer. Trends in genetics : TIG 9(4):138-141.

Vogelstein B, Lane D, Levine AJ. (2000) Surfing the Tp53 network. Nature. 408:307-310.

Warming S., Suzuki T., Yamaguchi TP., Jenkins NA., Copeland NG. (2004) Early B-cell factor-associated zinc-finger gene is a frequent target of retroviral integration in murine B-cell lymphomas. Oncogene. 8;23(15):2727-2731.

Weber JD., Jeffers JR., Rehg JE., Randle DH., Lozano G., Roussel MF., Sherr CJ., Zambetti GP. (2000) p53 -independent functions of the p19(ARF) tumor suppressor. Genes Development. 14(18):2358-2365.