7-8-2010

Stochastic and State Space Models of Carcinogenesis Under Complex Situation

Xiaowei Yan

To the Graduate Council:

I am submitting herewith a dissertation written by Xiaowei Yan entitled "Stochastic and State Space Models of Carcinogenesis under Complex Situation." I have examined the final copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy with a major in Mathematical Sciences.

_____
Wai-Yuan Tan, Ph.D.
Major Professor

We have read this dissertation and
recommend its acceptance:

_____
Lih-Yuan Deng, Ph.D.

_____
Seok P. Wong, Ph.D.

_____
Xiaoping Xiong, Ph.D.

Accepted for the Council:

_____
Karen D. Weddle-West, Ph.D.
Vice Provost for Graduate Programs

STOCHASTIC AND STATE SPACE MODELS OF CARCINOGENESIS UNDER COMPLEX SITUATION

by

Xiaowei Yan

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Major: Mathematical Sciences

The University of Memphis

August 2010

ACKNOWLEDGEMENTS

I would like to sincerely thank my advisor and mentor, Dr. Wai-Yuan Tan, for his wonderful guidance, supervision, support, profound patience and encouragement, during my graduate studies at the University of Memphis.

I am also very grateful for having outstanding doctoral committee and wish to thank Dr. Seok P. Wong, Dr. Lih-Yuan Deng, and Dr. Xiaoping Xiong for their support, suggestions and encouragements. I also want to thank Dr. Xiaoping Xiong and Dr. Jianrong Wu for their wonderful supervision and continual support, during my two-year intern at the St. Jude Children's Research Hospital.

I would like to thank Dr. Jason Schroeder, who is a special friend, for his suggestions, instruction, and support. I am very grateful for having him around to discuss career development and research, as well as books, music we appreciate.

I also wish to thank my best friends, Shafiqa Fakir, Wen Feng, Chong (Katy) Wang, Huiqing Zhang and Heng Yang for their encouragements, support, and friendship. Their friendship gives me new meaning of everyday-life and motivates me to continue to pursue new goals. I would not have been happy without their friendship.

I extend many thanks to my colleagues in the St. Jude Children's Research Hospital and friends in the University of Memphis, especially Mehmet Kocak, Chenghong Li, Wei Liu, Quan Tang, Yanan Wu etc., their support and friendship represent most significant finding of this endeavor.

Finally, I am especially grateful to my families, husband, parents and sister, for their absolute unselfishness and hearty support. My studies and life cannot be in the proper order and balance without them.

ABSTRACT

Yan, Xiaowei. Ph.D. The University of Memphis, August, 2010. Stochastic and State Space Models of Carcinogenesis under Complex Situation. Major Professor: Wai-Yuan Tan, Ph. D.

With more and more biological mechanisms of cancer development being discovered, in order to improve cancer control and prevention, it becomes necessary to develop effective and efficient mathematical and statistical models and methods to incorporate the biological information, and to identify critical events in the process of carcinogenesis. In this dissertation, the complex nature of carcinogenesis has been represented by stochastic system model; combining this model with information from observations and prior knowledge, we have developed state space models to evaluate cancer gene mutations and cell proliferation at different cancer development stages. Also, we have proposed a generalized Bayesian method via multi-level Gibbs sampling procedure to predict state (stage) variables of the models.

In this dissertation, stochastic models have been proposed for initiation, promotion and complete carcinomas experiments; these experiments are most commonly performed in cancer risk assessment of environmental agents. These stochastic models are simple multi-pathway models which are constructed based on biological mechanisms. The estimates we obtained from the models have provided quantitative evaluation of dose related mutation rates of major genes and cells proliferation rates; these results could be used to assess the risk of developing malignant tumor in the environment we live.

More complicated stochastic and state space models have been developed for sporadic human colon cancer and for hereditary and non-hereditary human liver cancer. We have utilized the proposed models to fit to Surveillance Epidemiology and End

Results (SEER) data. The results imply that our models have effectively incorporated biological information and observations; these models fitted the data very well and the inferences based on estimate were very consistent with biological findings. Furthermore, the models reflected the complex nature of carcinogenesis. We notice that many cancers are developed through multiple-stage multiple-pathway. Our analyses of colon cancer and liver cancer have showed that some pathways are more devastated than others. This suggests thus it would be more efficient to intervene or treat the critical events in the more devastated pathways.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Recent studies by cancer molecular biologists and cancer geneticists have shown that in most practical situations, the process of carcinogenesis is very complex.

First, the process of carcinogenesis may not be time homogeneous; this is especially true in cancer initiation and promotion experiments and in animal carcinogenicity experiments as those conducted by scientists at EPA. (Many other biological and epidemiological evidences can be found in papers in the NCI Symposium (1985) and in Tan and Chen (1995)).

Second, in most cases the same cancer is developed by several different pathways (Tan, 1991; Tan & Chen, 1991, 1998, 2000, 2005; Weinberg, 2007); this includes, for example, colorectal cancer, liver cancer, lung cancer, and melanoma among many others.

Third, inheritance and heredity is an important factor in carcinogenesis. For example, family cancer history may predict cancer development speed among breast cancers in women, and FAP and HNPCC are well known inherited cancer cases in human colon cancers. Also, pediatric cancers give cancer cases for new born babies, contributing to pediatric cancer cases. Examples of pediatric cancer cases include medulloblastoma, retinoblastoma, hepatoblastoma, pediatric lung cancer and rhabdomyosarcoma.

Fourth, it is well recognized that for most of the cancers, the process of carcinogenesis involves a large number of cancer genes (5 to 10 genes in most cases (Hopkin, 1996), but as many as 200 genes may have been involved in the origin of all leukemia; see Greaves, 1997). It follows that in many cases, three or more stages may be more appropriate to represent the true biologically supported stochastic models of carcinogenesis. For example, as demonstrated by Little and his associates (Little, 1995,

1996; Little, Murihead, Boice, & Kleinerman, 1995; Little, Muirhead, & Stiller, 1996) for cancers from Japanese atomic bomb survivors and radiation-induced leukemia, carcinogenesis is better described by three-stage models than by two-stage models although the same data can be fitted equally well by a three stage model as by a two-stage model.

The purpose of the dissertation is to develop stochastic and state space models for cancer development under complex situation.

In Chapter 1, we summarize the traditional stochastic models for carcinogenesis, and discuss the difficulties and disadvantages of the traditional stochastic models and the traditional approaches. Then we briefly introduce the biologically supported stochastic and state space models under realistic complex conditions. Next, some generalized Bayesian procedures are discussed for estimating parameters and for predicting state variables.

In Chapter 2, a set of stochastic models are proposed for initiation-promotion experiments, which are commonly used by researchers to assess cancer risk of environmental agents. It this chapter we propose a simple two pathways model involving a generalized one stage model and a generalized two-stage model for the generation of papillomas and carcinomas from these experiments.

In Chapter 3, the stochastic and state space models are proposed for human colorectal cancer. These models involve 2 pathways, with one pathway being a 4-stage model and the other pathway being a 5-stage model. Through fitting of the SEER data, we showed that the model we proposed was more appropriate than existing single-pathway models.

In Chapter 4, we discuss some stochastic and state space models for adult liver cancer (Hepatocarcinoma, HCC) and for pediatric liver cancer (hepatoblastoma), the latter of which incorporating hereditary segregation of genes.

In Chapter 5, we give a brief summary of our work and propose some future research.

Quantitative Carcinogenesis Modeling

The first sets of mathematical carcinogensis models, reflecting essential biological processes via pathway from a normal stem cells to a cancer cells, were proposed in 1950s (Nordling, 1953). Amitage-Doll model (Amitage & Doll, 1954) may be the best known model among them, which gave the age-specific incidence of many carcinomas, and also connected rate-limiting steps to the mathematical model. The model assumes that cancer cells are developed from a single stem cell through a series irreversible, heritable events, which are mutations related. However, the model completely ignores cell proliferation and differentiation of intermediate cells. In 1979 Moolgavkar and Venzon and in 1981 Moolgavkar and Knudson proposed a 2-stage model, called MVK-model, in which a cancer tumor is developed from a single normal stem cell by clonal expansion, and carcinogensis is taken as the end result of two discrete, heritable and irreversible events. This model became the most commonly used to fit different type of carcinomas since then (Moolgavkar, Dewanji, & Venzon, 1988; Moolgarvkar & Luebeck, 1989; Portier & Bailer, 1989). The major limits of the model are:

(1) It assumes that tumor is developed instantaneously from initiated cells at the last stage, which usually is not true. If this assumption is violated, then the number of cancer tumors is no longer Markov because it depends on the time when the last-stage initiated

cells are generated. In this case, Markov theories cannot be applied to cancer tumor (Tan & Chen, 1998).

(2) It has been well accepted that most of cancer tumors are developed through more than two stages, some cancer tumors are even compilcated, developing through multiple-pathway, and each pathway contains mutilple stages. So that two-stage model is obviously too simple to describe carcinogenesis (Tan, 2002).

(3) It is extremely hard to adopt the model to multiple-pathway multiple-stage carcinogenesis because the mathematical results become too complicated (Tan, 1991).

(4) Not all parameters could be identified from the incidence function and the probability distribution of time to tumors, because function of parameters instead of separate parameters is involved in the incidence function and the probability distribution.

To overcome the limits and difficulties of the MVK-model, we proposed an alternative approach, in which stochastic equations were used to represent the biological and genetic information of carcinogenesis, and a statistical model was used to represent observation. By combining these models and information from other sources, we were able to estimate all parameters and state variables, and also provide biologically reliable interpretation.

Many biological evidences have been reported to support cancer cells are developed through multiple pathways (Weinberg, 2007). Before discussing the detail of the new model we proposed and applied in multiple-pathway carcinogenesis, we would like to briefly review MVK model by applying it for a simplest multiple-pathway carcinogensis, which can be taken as an extended MVK two-stage model. Through the assumptions and model construction, the limits and difficulties of MVK model were revealed. Then we

introduced our approach to see the new approach has many advantages over the extended MVK model.

We started with a simple multiple-pathway model (called extended MVK model). The regular MVK two-stage model of carcinogenesis assumes three types of cancer cells: The normal stem cells, the intermediate cells and the tumor cells. It also assumes that with probability one, a tumor cell will develop into malignant tumor. We denote three types of cancer cells by N(t), I(t) and T(t) at time t, respectively. In the extended MVK model, instead, two intermediate cells I_1 and I_2 cells are present.

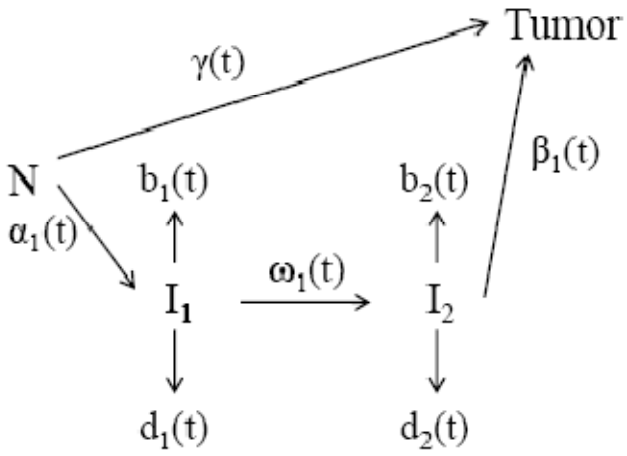


Figure 1. A Simple Multiple-Pathway Model

As shown in Figure 1, the extended MVK model involves a one-stage model and a three-stage model. In this model, a cancer tumor develops either by normal cells N to tumor cells T directly, or by normal cells N to tumor cells T via I_1 and I_2 cells. Note that because mutation rates are usually very small, in order that tumor cells can develop through three-stage pathway, the one-stage pathway must be a rare event too, which is actually consistent with a lot observations (Land, Parada, & Weinberg, 1983; Schwab, Varmus, & Bishop, 1985; Tan, 2001; Tan & Chen, 1998; Yancopoulos et al., 1984).

Let $N(t)$, $I_1(t)$, $I_2(t)$ and $T(t)$ be the number of normal stem cells, I_1 cells, I_2 cells and cancer tumor cells at time t, respectively. To derive mathematical procedures for the model shown in Figure 1, we make the following assumptions. Most of the assumptions are the same as for MVK two-stage model.

1. At the time t_0 of birth, $N(t_0) = N_0$ is very large ($N_0 \leq 10^8$). This assumption usually holds for most of tissues (Moolgarkar & Knudson, 1981).

2. The I_j, $j=1, 2$ cells follow a nonhomogeneous Feller-Arley birth-death process for cell proliferation and cell differentiation with birth rate $b_j(t)$ and death rate $d_j(t)$.

3. Given that the cancer tumors developed from normal stem cells by one-stage pathway, ($N \rightarrow T$), we assume that the probability that a normal stem cell at time t produces one normal stem cells and one tumor cell at time $t + \Delta t$ is $\gamma(t)\,\Delta t + o(\Delta t)$. Similarly, given the three-stage pathway $N \rightarrow I_1 \rightarrow I_2 \rightarrow T$, we assume that the probability of a normal cell producing one normal cell and one I_1 cell at time $t + \Delta t$ is $\alpha_1(t)\Delta t + o(\Delta t)$. And the probability that an I_1 cell at time t produces one I_1 cell and one I_2 cell at time $t + \Delta t$ is $\omega_1(t)\Delta t + o(\Delta t)$. The probability of I_2 cell yields one I_2 cell and one tumor cell at time $t + \Delta t$ is $\beta_1(t)\Delta t + o(\Delta t)$.

4. As in Moolgavkar and Knudson (1981) and Tan and Brown (1987), we assume that the time require for the development of tumors from tumor cells is very short compared with the time cancer tumors developed from normal stem cells. This implies that with probability one tumor cells grow into tumors and that random variation for the time between the initiated tumor cells and cancer tumors is ignored.

5. The birth-death processes and the mutation processes are independent of one another and each cell goes through the above processes independently of other cells.

Given above assumptions, it follows that:

(1) The processes is basically a Markov process

(2) Since the mutation rates are very small ($10^{-8} \sim 10^{-6}$), also because in most cases the birth rates are greater than death rates for normal cells (Buick & Pollack, 1984), the number of normal cells is very large for all $t \geq t_0$. In that case, we assume the number of normal cells ($N(t)$) is a deterministic function. And because of mutation rates are very small so that $N(t)\gamma(t)$ and $N(t)\alpha_1(t)$ are expected to be finite. If $N(t)\gamma(t)$ and $N(t)\alpha_1(t)$ are finite for all $t \geq t_0$, it is reasonable to assume that during $[t, t + \Delta t)$ the mutation processes from normal cells to tumor cells (one-stage pathway), to I_1 cells follow Poisson processes with parameters $N(t)\gamma(t)\Delta t + o(\Delta t)$ and $N(t)\alpha_1(t)\Delta t + o(\Delta t)$.

(3) As shown in Tan and Brown (1987), if the normal stem cells follow nonhomogeneous Feller-Arley Birth-death processes with birth rate $b_N(t)$ and death rate $d_N(t)$, then from above assumption (2)-(5), $\{N(t), I_i(t), i = 1, 2, T(t)\}$ form a continuous time multiple branching process, with rates of $b_N(t_j)\Delta t + o(\Delta t)$, $d_N(t_j)\Delta t + o(\Delta t)$ for normal stem cells; $b_i(t_j)\Delta t + o(\Delta t)$, $b_i(t_j)\Delta t + o(\Delta t)$, in which $\{i = 1, 2; j = 0, 1, 2, \ldots, n\}$, for I_i cells; and $1 + o(\Delta t)$ for tumor cells.

In MVK model, the traditional procedures to obtain the incidence function are:

First, derive the probability generating function (PGF) of the number of intermediate initiated cells and cancer tumors through Kolmogorov forward equation.

Second, use the PGF to obtain cancer incidence function and the probability distribution of the number of tumors.

We will show the traditional procedures for the simple two-pathway model.

Let:

$P_1(I, j_u, u=1,2, k; t)$ be the conditional probability of $[N(t) = I, I_u(t) = j_u, T(t) = k \mid N(t_0) = N_0]$,

$P_2(j_1,j_2, k; t)$ the conditional probability of $[I_1(t) = j_1, I_2(t) = j_2, T(t) = k \mid I_1(t_0) = 1]$. and $P_3(j, k; t)$ the conditional probability of $[I_2(t_0) = j, T(t) = k \mid I_2(t_0) = 1]$.

Let $\eta(x, y_1, y_2, z; t_0, t) = \eta(t_0, t)$ be the PGF of $[N(t), I_1(t), I_2(t), T(t) \mid N(t_0) = N_0]$;

$\eta(1, y_1, y_2, z; t_0, t) = \psi(x, y_1, y_2, z; t_0, t) = \psi(t_0, t)$ the PGF of $[I_1(t), I_2(t), T(t) \mid N(t_0)=N_0]$; and $g(z; t_0, t) = \psi(1, 1, 1, z; t_0, t)$ the PGF of $[T(t) \mid N(t_0)= N_0]$.

Let $\zeta_1(y_1,y_2, z; t_0, t) = \zeta_1(t_0, t)$ be the PGF of $[I_1(t), I_2(t), T(t) \mid I_1(t_0)=1]$, and $\phi(y_2, z; t_0, t) = \phi(t_0, t)$ the PGF of $[I_2(t), T(t) \mid I_2(t_0) = 1]$.

Then

$$\eta(t_0, t) = \sum_i \sum_{j_1} \sum_{j_2} \sum_k x^i y_1^{j_1} y_2^{j_2} z^k P_1(i, j_1, j_2, k; t) \tag{1.1}$$

$$\zeta_1(t_0, t) = \sum_{j_1} \sum_{j_2} \sum_k y_1^{j_1} y_2^{j_2} z^k P_2(j_1, j_2, k; t) \tag{1.2}$$

and

$$\phi(t_0, t) = \sum_j \sum_k y^j z^k P_3(j, k; t) \tag{1.3}$$

Because the processes are Markov processes, we can write down the Kolmogorov forward equations for the above conditional probabilities by using standard procedures. Using transition probability given in Table 1, the Kolmogorov forward equations are:

$$\frac{d}{dt} P_1(i, j_1, j_2, k; t) = (i-1) b_N(t) P_1(i-1, j_1, j_2, k; t) + (i+1) d_N(t) P_1(i+1, j_1, j_1, k; t) + \sum_{u=1}^{2} \{(j_u - 1) b_u(t) P_1(j_u - 1, j_v, u \neq v, k; t) + \big((j_u + 1) b_u(t) P_1(j_u + 1, j_v, u \neq v, k; t)\big)\} + i\alpha_1(t) P_1(i, j_1 - 1, j_2, k; t) + j_1 \omega_1(t) P_1(i, j_1, j_2, k; t) + [i\gamma(t) + j_2 \beta_1(t)] P_1(i, j_1, j_2, k; t) - \{i[b_N(t) + d_N(t) + \gamma(t) + \alpha_1(t)] + j_1 \omega_1(t) + j_2[b_2(t) + d_2(t) + \beta_1(t)]\} * P_1(i, j_1, j_2, k; t) \tag{1.4}$$

and

$$\frac{d}{dt}P_2(j_1,j_2,k;t) = (j_1-1)b_1(t)P_2(j_1-1,j_2,k;t) + (j_1+1)d_1(t)P_2(j_1+1,j_2,k;t) +$$

$$j_2\omega_1(t)P_2(j_1,j_2-1,k;t) + (j_2-1)b_2(t)P_2(j_1,j_2-1,k;t) + (j_2+1)d_2(t)P_2(j_1,j_2+$$

$$1,k;t) + j_2\beta_1(t)P_2(j_1,j_2,k-1;t) - \{j_1[b_1(t)+d_1(t)+\omega_1(t)] + j_2[b_2(t)+d_2(t)+$$

$$\beta_1(t)]\}P_2(j_1,j_2,k;t) \qquad (1.5)$$

and

$$\frac{d}{dt}P_3(j,k;t) = (j-1)b_2(t)P_3(j-1,k;t) + (j+1)d_2(t)P_3(j+1,k;t) +$$

$$j\beta_1(t)P_3(j,k-1;t) - j_2[b_2(t)+d_2(t)+\beta_1(t)]P_3(j_1,j_2,k;t) \qquad (1.6)$$

The initial conditions are:

$P_1(i,j_1,j_2,k;t_0) = \delta_{iN_0}$; *i, j_1, j_2, k = 0, 1, 2…*

$P_2(j_1,j_2,k;t_0) = \delta_{1j_i}$, *i = 1, 2 and j_1, j_2, k = 0, 1, 2…*

$P_3(j,k;t_0) = \delta_{ij}$; *j, k = 0, 1, 2…*

Under the assumption (2)-(5), Tan (1991) has shown that $\zeta_1(t_0,t)$ and $\phi(t_0,t)$ satisfy the following first order partial differential equations:

$$\frac{\partial}{\partial t}\zeta_1(t_0,t) = \{(y_1-1)[y_1b_1(t)-d_1(t)] + y_1(y_2-1)\omega_1(t)\}\frac{\partial}{\partial y_1}\zeta_1(t_0,t)$$

$$+\{(y_2-1)[y_2b_2(t)-d_2(t)] + y_2(z-1)\beta_1(t)\}\frac{\partial}{\partial y_2}\zeta_1(t_0,t) \qquad (1.7)$$

with initial condition $\zeta_1(t_0,t_0) = y_1$

and

Table 1

Transition Probabilities of Simple Two-pathway Model

Cells at time t_j	Cells at time $t_j+\Delta t$	Probability
1 N cell	2 N Cells	$b_N(t_j)\Delta t+o(\Delta t)$
	0 Cells	$d_N(t_j)\Delta t+o(\Delta t)$
	1 N cell and 1 T cell	$\gamma(t_j)\Delta t+o(\Delta t)$
	1 N cell and 1 I_1 cell	$\alpha_1(t_j)\Delta t+o(\Delta t)$
	1 N cell	$1-[\, b_N(t_j)+ d_N(t_j)+ \gamma(t_j)+ \alpha_1(t_j)]\Delta t+o(\Delta t)$
1 I_1 cell	2 I_1 Cells	$b_1(t_j)\Delta t+o(\Delta t)$
	0 Cells	$d_1(t_j)\Delta t+o(\Delta t)$
	1 I_1 cell and 1 I_2 cell	$\omega_1(t_j)\Delta t+o(\Delta t)$
	1 I_1 cell	$1-[\, b_1(t_j)+ d_1(t_j)+ \omega_1(t_j)]\Delta t+o(\Delta t)$
1 I_2 cell	2 I_2 Cells	$b_2(t_j)\Delta t+o(\Delta t)$
	0 Cells	$d_2(t_j)\Delta t+o(\Delta t)$
	1 I_2 cell and 1 T cell	$\beta_1(t_j)\Delta t+o(\Delta t)$
	1 I_2 cell	$1-[\, b_2(t_j)+ d_2(t_j)+ \beta_1(t_j)]\Delta t+o(\Delta t)$
1 Tumor	1 Tumor	$1+ o(\Delta t)$
	Other Cases	$o(\Delta t)$

$$\frac{\partial}{\partial t}\phi(t_0,t) = \{(y_2-1)[y_2 b_2(t) - d_2(t)] + y_2(z-1)\beta_1(t)\}\frac{\partial}{\partial y_2}\phi(t_0,t) \quad (1.8)$$

with initial condition $\phi(t_0,t_0) = y_2$.

If the normal stem cells follow nonhomogeneous Feller-Arley birth-death process with birth rate $b_N(t)$ and death rate $d_N(t)$, then $\psi(t_0, t)$ satisfies the following first-order partial differential equation:

$$\frac{\partial}{\partial t}\eta(t_0,t) =$$

$$\{(x-1)[xb_N(t)-d_N(t)] + x(z-1)\gamma_1(t) + x(y_1-1)\alpha_1(t)\}\frac{\partial}{\partial x}\eta(t_0,t) +$$

$$\{(y_1-1)[y_1b_1(t)-d_1(t)] + y_1(y_2-1)\omega_1(t)\}\frac{\partial}{\partial y_1}\eta(t_0,t) + \{(y_2-1)[y_2b_2(t) -$$

$$d_2(t)] + y_2(z-1)\beta_1(t)\}\frac{\partial}{\partial y_2}\eta(t_0,t) \qquad (1.9)$$

with initial condition $\eta(t_0, t_0) = x^{N_0}$

The solution of partial differential equation of $\phi(t_0, t)$ is available for most of important special cases. As shown in Tan (1991), if $\beta_1(t) = \beta_1, b_2(t) = b_2, d_2(t) = d_2$ and *N(t)* is a deterministic function, in this case, $\phi(t_0, t) = \phi(t_0 - t)$, if we set $t_0 = 0$, then $\phi(y_2, z; 0, t) = \phi(t)$ satisfies the following Ricatti equation with initial condition $\phi(0) = y_2$:

$$\frac{d}{dt}\phi(t) = b_2\phi^2(t) + [\beta_1 z - (b_2 + d_2 + \beta_1)]\phi(t) + d_2 \qquad (1.10)$$

Then with an assumption that the probability of tumor cell developing to cancer tumor is one, then solution of above differential equation can be readily solved and the solution is:

$$\phi(t) = \{x_2(y_2 - x_1) + x_1(x_2 - y_2)\exp[b_2(x_2 - x_1)]\,t\} * \{(y_2 - x_1) +$$

$$(x_2 - y_2)\exp[b_2(x_2 - x_1)]\,t\}^{-1} \qquad (1.11)$$

where $x_2 > x_1$ are given by $2b_2x_2 = (b_2 + d_2 + \beta_1 - \beta_1 z) + [(b_2 + d_2 + \beta_1 - \beta_1 z)^2 - 4b_2d_2]^{1/2}$ and $2b_2x_1 = (b_2 + d_2 + \beta_1 - \beta_1 z) - [(b_2 + d_2 + \beta_1 - \beta_1 z)^2 - 4b_2d_2]^{1/2}$.

Though solution of $\phi(t_0, t)$ is available, the solution of $\zeta_1(t_0, t)$ and $\psi(t_0, t)$ are extremely difficult even in homogenous cases. We have to add more assumptions to find the solution. If N(t) is very large, and γ(t) and α_1(t) are very small, so that both N(t)γ(t) and N(t)α_1(t) are finite for all $t \geq t_0$, we can assume Poisson distributions for the number of mutations from normal stem cells to tumor cells and to I_1 cells and also assume $\omega_1(t) = \omega_1$. Then as shown in Tan (1991) the solution for $\psi(t_0, t) = \eta(1, y_1, y_2, z; t_0, t)$ is as following:

$$\psi(t_0, t) = \exp\{(z-1)\int_{t_0}^{t} N(u)\gamma(u)du + \int_{t_0}^{t} N(u)\alpha_1(u)[\zeta_1(u, t) - 1]du\} \tag{1.12}$$

In order to find $\psi(t_0, t)$, we have to find solution for $\zeta_1(u, t)$. Similar as to derive Ricatti equation for $\phi(t)$, the $\zeta_1(t_0, t) = \zeta_1(t - t_0)$ and set $t_0 = 0$, we have following Ricatti equation:

$\frac{d}{dt}\zeta_1(t) = b_1{\zeta_1}^2(t) + [\omega_1\phi(t) - (b_1 + d_1 + \omega_1)]\zeta_1(t) + d_1$, with initial condition $\zeta_1(0) = y_1$.

In general it is very difficult to solve above differential equation because $\phi(t)$ is not linear, thus the whole function is not linear. In order to find close form solution for above differential equation, we have to add further assumptions or find appropriate approximation. Moolgavkar and Venzon (1979) assumed $d_1 = 0$, then above differential equation of $\zeta_1(t)$ becomes:

$\frac{d}{dt}\zeta_1(t) = b_1{\zeta_1}^2(t) + [\omega_1\phi(t) - (b_1 + \omega_1)]\zeta_1(t)$, with initial condition $\zeta_1(0) = y_1$.

By using linear approximation of $\phi(t)$ (Taylor expansion), the solution of $\zeta_1(t)$ is:

$$\zeta_1(t) = y_1\theta_1(t)\{1 - y_1 b_1 U_1(t)\}^{-1} \tag{1.13}$$

where

$\theta_1(t) = exp\{-tb_1 + \omega_1 \int_0^t [\phi(x) - 1]dx\}$ and $U_1(t) = \int_0^t \theta_1(y)dy$.

Cancer Incidence Function λ(t)

After obtaining solutions for $\zeta_1(t)$, we can further obtain incidence rate λ(t) of tumor at time t. It has been shown by Tan (1991) that:

$$\lambda(t) = -\psi'(1,1,0; t_0, t)/\psi(1,1,0; t_0, t) \quad (1.14)$$

where

$$\psi'(1,1,0; t_0, t) = \frac{d}{dt}\psi(1,1,0; t_0, t)$$

To the order $O(N_0^{-1})$:

$$\psi'(1,1,0; t_0, t) = \psi(1,1,0; t_0, t)\left\{-N(t)\gamma(t) + N(t)\alpha_1(t)[\zeta_1(1,1,0; t, t) - 1] + \int_{t_0}^{t} N(u)\alpha_1(u)\upsilon_1(u, t)du\right\}$$

where $\upsilon_1(u, t) = \frac{\partial}{\partial t}\zeta_1(1,1,0; u, t)$

Since $\zeta_1(y_1, y_2, z; t_0, t_0) = y_1$, so

$$\lambda(t) = N(t)\gamma(t) + \int_{t_0}^{t} N(u)\alpha_1(u)[-\upsilon_1(u, t)]du \quad (1.15)$$

and $[-\upsilon_1(u, t)] = \beta_1(t)\zeta_{10}(u, t)E[I_2(t)|T(t) = 0, I_1(u) = 1]$,

where $\zeta_{10}(u, t) = \zeta_1(1,1,0; u, t)$ and $E[I_2(t)|T(t) = 0, I_1(u) = 1]$ is the conditional expected number of I_2 cells at time t given $T(t) = 0$ and $I_1(u) = 1$, $t \geq u$.

It follows that

$$\lambda(t) = N(t)\gamma(t) + \int_{t_0}^{t} N(u)\alpha_1(u)\beta_1(t)\zeta_{10}(u, t)E[I_2(t)|T(t) = 0, I_1(u) = 1]du \quad (1.16)$$

That is, incidence rate is the function of conditional expected number of the last stage initiated cells.

From above incidence function, we notice that the product of some mutation rates is involved in the incidence function, such as $\alpha_1(u)\beta_1(t)$, it is not possible to estimate these parameters separately. So that even if this model could be applied to estimate parameters, not all parameters are identifiable.

As we have shown above, three major drawbacks of the MVK model or extended MVK model are:

Many biological evidences indicated that in most of cases cancer tumor developed from multiple pathways rather than a single pathway, and each pathway contains more than two stages. Even we can extend the MVK two-stage to multiple-pathway multiple-event model, it would be extremely difficult to solve partial differential equations involved in the model, and it is almost not possible to obtain close form of incidence rate.

Even a model is as simple as above, the incidence function has been very difficult to obtain. We have also shown that in order to find close form of the solutions to differential equations, we have to have more assumptions, some assumptions are not realistic. For example, in order to obtain the solution for $\zeta_1(t)$, Moolgavkar and Venzon have assumed $d_1 = 0$.

Even if the close form of incidence function could be derived, we have shown that many parameters cannot be estimated separately.

The MVK model is based on Markov process, which implies that with probability 1 each cancer cell instantaneously develops into a cancer tumor. This assumption ignores

tumor progression (birth-death process), which is usually not true in reality (Fakir, Tan, Hlatky, Hahnfeldt, & Sachs, 2009).

In order to overcome these difficulties and drawbacks, we introduce stochastic and state space model developed for complex system, which usually contains multiple pathways, multiple-stage.

Stochastic Equations and State Space Model

Biologists and geneticists have discovered that the same cancer can be developed through more than one pathway, and that each pathway contains more than two stages. Along single pathway models, realizing the complex nature of carcinogenesis, Chu (1985) and Little (1995) have developed generalized multistage stochastic model of carcinogenesis, extending the two-stage MVK model to models with multiple stages. These models and the methods were based on Markov process which would need to assume that the last stage cells grow instantaneously into cancer tumors as soon as they are generated; in this case, one may identify the last stage cell as cancer tumors (i.e., ignoring cancer progression). Unfortunately, in many practical situations, the Markov assumption is not realistic and gives confusing results (Fakir et al., 2009; Yakovlev & Tsodikov, 1996; Yang & Chen, 1991), especially in mouse models and in radiation carcinogenesis. Furthermore, even if the Markov assumption is valid, because it is often impossible to obtain analytical solutions for partial differential equations of the probability generating function of stage variables and cancer tumors derived from Kolmogorov's forward/backward equations, the methods would need to make many additional unrealistic assumptions and to drive approximations. Following Yang and Chen (1991) to postulate that cancer tumor developed by clonal expansion from a

primary cancer cell to account for cancer progression, Tan and Chen (1998) have developed an extended k-stage model and have proposed a stochastic equations approach to develop mathematical analysis for these models. This extended k-stage model is the most general model for single pathway and the stochastic equations method is very powerful and can be used to more complex models involving multiple pathways.

Because many cancers involve multiple pathways, in this section we will propose a more generalized model involving multiple-pathways. We will illustrate how to construct this new model and discuss some advantages of the new model. As we will show in Chapters 3 and 4, these models can be used to fit cancer incidence of human colon cancer and liver cancer.

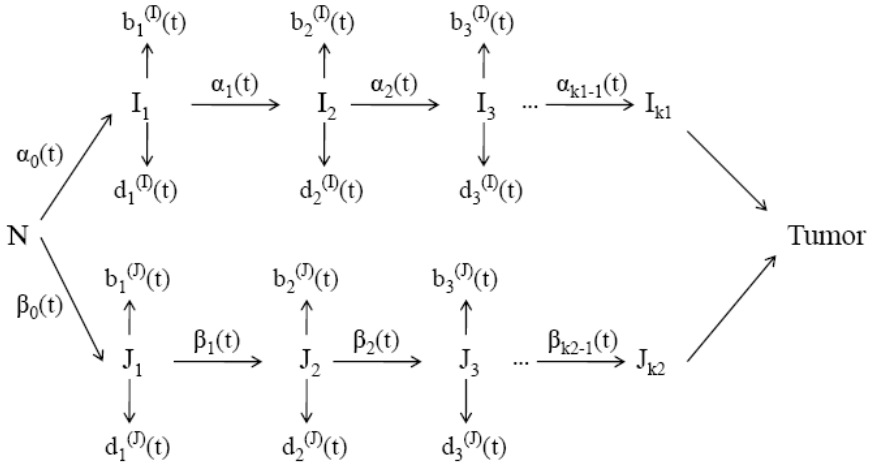


Figure 2. Multiple Pathways

In Figure 2, cancer tumor is developed from two pathways; one pathway, called I pathway, contains k_1 stages, and another pathways, called J pathway, contains k_2 stages. We define the I_{k1} and J_{k2} cells as primary cancer cells, which arise directly from I_{k1-1} and J_{k2-1} cells by mutation or genetic changes. Thus, I_{k1} and J_{k2} arising from other I_{k1} and J_{k2}

cells by cell proliferation are called secondary I_{k1} andJ_{k2} cells. Therefore, cancer tumors either derive from primary I_{k1} cells or from primary J_{k2} cells by stochastic birth-death process.

We define intermediate stages involved in the multiple-pathway model as discrete, heritable and irreversible events, denoted by I_u, u = 1, …, k_1-1and J_v, v=1, …, k_2-1. I_u and J_v initiated cells arise from (u-1)th stage in I pathway and (v-1)th stage in J pathway by mutation or genetic changes. Let N denote normal stem cells, and T denote cancer tumors arising from primary I_{k1} andJ_{k2} cells. Then the model assumes that the two pathways are exclusive, and assumes $N \rightarrow I_1 \rightarrow I_2 \rightarrow \ldots \rightarrow I_{k1}$ and $N \rightarrow J_1 \rightarrow J_2 \rightarrow \ldots \rightarrow J_{k2}$ with I_u and J_v cells being subject to stochastic birth-death process. Let $I_u(t)$, u = 1, 2, …, k_1-1 denote the number of I_u cells at time t and $J_v(t)$, v = 1, 2, …, k_2-1 denote the number of J_v cells at time t, and T(t) denote the number of malignant cancer tumor at time t. We use $\mathcal{U}(t)$ = { $I_u(t)$, u = 1, 2, …, k_1-1, $J_v(t)$, v = 1, 2, …, k_2-1, T(t)} to denote the set of state variables, and use $\{b_u(t), d_u(t), \alpha_u(t)\}$ represent the birth rate, death rate and mutation rate from $I_u \rightarrow I_{u+1}$ at time t. Similarly, we use $\{b_v(t), d_v(t), \beta_u(t)\}$ denote the birth rate, death rate and mutation rate from $J_v \rightarrow J_{v+1}$ at time t. That is, at small time interval [t, t+Δt), the probability that I_u cell will produce two I_u cells, zero I_u cells, and one I_u cell and one I_{u+1} cell with probability $b_u(t)\Delta t + o(\Delta t)$, $d_u(t)\Delta t + o(\Delta t)$, and $\alpha_u(t)\Delta t + o(\Delta t)$, respectively. The same to J_v cells. At the last stage, the malignant cancer tumors developed from a primary I_{k1} or J_{k2} cell through stochastic birth-death process with birth rate $b_{k1}(s_1,t)$ and death rate $d_{k1}(s_1,t)$ for I pathway, and $b_{k2}(s_2,t)$ for birth rate and $d_{k2}(s_2,t)$ for J pathway, where s_1 and s_2 are onset time of I_{k1} or J_{k2} cell, respectively.

To derive the stochastic differential equations for above processes, let $\{B_u^I(t), D_u^I(t), M_u^I(t)\}$ denote the number of birth, death, and mutated cells from $I_u \rightarrow I_{u+1}$ at time t, and $\{B_v^J(t), D_v^J(t), M_v^J(t)\}$ denote the number of birth, death, and mutated cells from $J_v \rightarrow J_{v+1}$ at time t. Tan (2002) has shown that the conditional probability distribution of $\{B_u^I(t), D_u^I(t), M_u^I(t)|I_u(t)\}$ is multinomial with parameters *{$I_u(t)$; $b_u(t)\Delta t$, $d_u(t)\Delta t$, $\alpha_u(t)\Delta t$}*. Similarly,

$\{B_v^J(t), D_v^J(t), M_v^J(t)|J_v(t)\} \sim ML\{J_v(t), b_v(t)\Delta t, d_v(t)\Delta t, \beta_v\Delta t\}$, v = 1, 2, …, k_2-1.

Then the differential equations for $I_u(t)$, u = 1, 2, …, k_1-1 and for $J_v(t)$, v = 1, 2, …, k_2-1 could be obtained from above distributions:

$$dN(t) = N(t+\Delta t) - N(t) = B_0(t) - D_0(t) = N(t)\gamma_0(t)\Delta t + \varepsilon_0(t) \qquad (1.17)$$

$dI_u(t) = I_u(t+\Delta t) - I_u(t) = M_{u-1}(t) + B_u(t) - D_u(t) = \{I_{u-1}(t)\alpha_{u-1}\Delta t + I_u(t)\gamma_u(t)\Delta t\}$, where $\gamma_u(t) = b_u(t) - d_u(t)$, u = 0, 1,…, k_1-1.

The random noises are given by:

$$\varepsilon_0(t) = [B_0(t) - N(t)b_0(t)\Delta t] - [D_0(t) - N(t)d_0(t)\Delta t] \qquad (1.18)$$

and

$$\varepsilon_u(t) = [M_{u-1}(t) - I_{u-1}(t)\alpha_{u-1}(t)\Delta t] + [B_u(t) - I_u(t)b_u(t)\Delta t] - [D_u(t) - I_u(t)d_u(t)\Delta t] \qquad (1.19)$$

Based on the distribution, the random noises $\{\varepsilon_u(t), u = 0, 1, \ldots, k_1 - 1\}$ have condistional expectation zero. It is also easy to see $\varepsilon_u(t)$ are uncorrelated with state variables. The covariance of $\varepsilon_u(t)$'s are also easy to derive, that is, to order o(Δt), the covariance is:

$$Cov(\varepsilon_u(t)\Delta t, \varepsilon_u(\tau)\Delta t) = \delta(t-\tau)E\{[1-\delta_{u0}]I_{u-1}(t)\alpha_{u-1}(t) + I_u(t)[b_u(t) + d_u(t)]\}\Delta t + o(\Delta t), \text{ and } Cov(\varepsilon_u(t)\Delta t, \varepsilon_w(\tau)\Delta t) = 0 \text{ for } u \neq w \qquad (1.20)$$

Similar differential equations can be derived for $J_v(t)$ cells, v = 1, 2, …, k_2-1.

We also can develop the probability distribution of T(t) by using the I pathway for illustration . We use $P_T(s_1,t)$ to represent the probability a primary I_u cell arising at time s_1 developing a detectable caner tumor at time t. Assume that cancer tumor is detectable only if it contains at least N_T cancer tumor cells. As shown in Tan (2002), given an I_u cell arising from I_{u-1} cell at time s_1, the probability that this I_u cell will produce j I_u cells at time t is given by:

$$P_M(j) = \begin{cases} 1 - (h(t-s_1) + g(t-s_1))^{-1} \; if \;\; j = 1 \\ (\frac{g(t-s_1)}{h(t-s_1)+g(t-s_1)})^{j-1} \frac{h(t-s_1)}{(h(t-s_1)+g(t-s_1))^2} \; if \; j > 1 \end{cases} \tag{1.21}$$

where

$$h(t-s_1) = exp\,\{-\int_{s_1}^{t} [b_u(y-s_1) - d_u(y-s_1)]\, dy\}$$

and

$$g(t-s_1) = \int_{s_1}^{t} b_u(y-s_1)h(t-s_1)dy$$

Then $P_T(s_1, t)$ is given by:

$$\mathrm{P_T(s_1, t)} = \sum_{\mathrm{j=N_T}}^{\infty} \mathrm{P_M(j)} = \frac{1}{\mathrm{h(t-s_1)+g(t-s_1)}} (\frac{\mathrm{g(t-s_1)}}{\mathrm{h(t-s_1)+g(t-s_1)}})^{\mathrm{N_T-1}} \tag{1.22}$$

Given this probability, the conditional probability distribution of T(t) given $\{I_{k1-1}(\tau), \tau \leq t\}$ is a Poisson with parameter:

$$\lambda^I(t) = \int_{t_0}^{t} I_{k_1-1}(x)\alpha_{k_1-1}(x)P_T(x,t)dx \tag{1.23}$$

That is,

$$\{T(t) \mid I_{k1-1}(\tau),\ t_0 \leq \tau \leq t\} \sim Poisson\ \{\ \lambda^I(t)\ \} \tag{1.24}$$

Similarly, we can derive the conditional probability distribution of $T(t)$ given $\{J_{k2\text{-}1}(\tau),\ \tau \le t\}$.

Then, if tumor is developed from I pathway alone, the conditional probability of producing a detectable cancer tumor during $[t_{j-1}, t_j)$ given $\{I_{k1\text{-}1}(\tau),\ t_0 \le \tau \le t_j\}$ is:

$$Q_T^I(t_{j-1}, t_j) =$$

$$exp\left\{-\int_{t_0}^{t_{j-1}} I_{k_1-1}(x)\alpha_{k_1-1}(x)P_T(x,t)dx\right\} - exp\left\{-\int_{t_0}^{t_j} I_{k_1-1}(x)\alpha_{k_1-1}(x)P_T(x,t)dx\right\} \quad (1.25)$$

Similarly, we can derive the probability of detectable tumor developed from J pathway alone during $[t_{j-1}, t_j)$.

Since the cancer tumor is developed from at least one pathway, during $[t_{j-1}, t_j)$, the probability of detectable tumor is developed from at least one pathway is:

$$Q_T(t_{j-1}, t_j) = 1 - \left(1 - Q_T^I(t_{j-1}, t_j)\right)\left(1 - Q_T^J(t_{j-1}, t_j)\right) = Q_T^I(t_{j-1}, t_j) +$$

$$Q_T^J(t_{j-1}, t_j) - Q_T^I(t_{j-1}, t_j) * Q_T^J(t_{j-1}, t_j) \quad (1.26)$$

We have illustrated how to derive the conditional probability of state variables and cancer incidence function for two pathways. The conditional probability of state variables and cancer incidence depend on the model structure, for example, single multiple-stage model, or multiple-pathway multiple-event. No matter what the model structure is utilized, cancer incidence and conditional probability of state variables can be readily derived. Tan and Chen (1998) have shown that the model through stochastic differential equations are equivalent to the classical Markov theory method, that is, intermediate initiated cells and cancer tumors satisfy the same set of partial differential equations. Above model has also taken tumor progression of the last stage into consideration, thus the model is more appropriate in practice.

The observation model, on the other hand, depends on available data. For example, in cancer risk assessment studies of environmental agents, the number of animals with papillomas during $[t_{i,j-1}, t_{i,j})$ is observed, denoted by $M(j)$; starting with n animals with no papillomas at t_0, then we can construct observation model as following:

$$M(j) \sim Binomial(n; Q_T(t_{j-1}, t_j)) \tag{1.27}$$

where $Q_T(t_{j-1}, t_j)$ is the probability of developing papillomas during $[t_{i,j-1}, t_{i,j})$.

The observation model, as shown above, integrates the stochastic model with the observation. By incorporating differential equations with observations, and information from other sources, Tan and Chen (1998) proposed state space models of carcinogenesis for multi-stage models. It combines the basic mechanism and random variation, represented by stochastic differential equations, with observations (via observation model). Tan and Zhang (2007) and Tan and Yan (2009) applied the state space models in several types of cancers to study cancer tumor development. The applications have shown that state space models have many advantages over traditional stochastic model or statistical model used alone:

Biologists and geneticists have discovered more and more mechanisms of carcinogenesis, which provide a rich source in carcinogenesis modeling. The state space model combines information from different sources, and are easier to adapt to study different types of cancers.

The state space model can provide reasonable estimation of parameters which usually cannot be identified by the stochastic model or statistical model alone.

Through Gibbs sampling the state space model can provide an optimal procedure to estimate unknown parameters and state variables simultaneously.

The state space model can provide a new avenue to predict cancer incidence in the future.

We have constructed the state space models for different types of cancers, the details are shown in the following chapters.

Pathways of Carcinogenesis

Stochastic differential equations are constructed based on pathways involved in cancer development. In this section, we will briefly introduce pathways of carcinogenesis.

It has been commonly accepted that carcinogenesis is a multiple-step random process, and all steps reflect genetic and/or epigenetic changes, which initiate many mutations and promote proliferation of mutated cells, leading to cancer malignancy (Nettesheim & Barrett, 1985; Tan, 1991). However, because genetic changes are rare events and can only occur during cell cycle, it is statistically nearly impossible that all genetic changes occur during one cell cycle. On the other hand, though genetic changes can take place in any cell cycle, only certain order of genetic changes can lead to the completion of cascade of carcinogenesis to generate cancer tumor. The major genetic changes and their order form genetic pathways of carcinogenesis. For example, in human colon cancer, a cell with mutated ras oncogene but no other genetic changes would eventually be eliminated. Thus in colon cancer, mutation of ras gene has never been observed as an initiated early event. Molecular biological explanation is that though mutated ras enables the cell to enter cell cycle without growth factor and can also evoke MAPK pathway and the PI3K-Akt pathway to increase transcription of many genes in nucleus (Osada & Takahashi, 2002; Weinberg, 2007); however, it can also induce the

suppressor gene $p14^{ARF}$ to activate p53 gene via ARF-MDM2-p53 pathway, leading to apoptosis of cell. In contrast, if APC gene is mutated or inactivated, it would generate chromosome instability and LOH (loss of Heterozygosity) because the APC gene affect the G_2 checking point during mitosis stage by interfering with microtubule and centrosome coursing aberrant chromosomal segregation, the daughter cells become aneuploidy or polyploidy (Fodde, Smit, & Clevers, 2001; Green & Kaplan, 2003), which would increase the fitness of the cells and also speed up mutations or inactivation of other genes.

Micro-array analyses have indicated that in most human cancers, a large number of cancer genes are involved. However, only a few of the genes are stage and rate limiting, leading to a finite number of stages in the multi-stage model of carcinogenesis (Renan, 1993). The age-dependent cancer incidence data for many human cancers imply four to seven rate-limiting stages from normal stem cells to malignant cancer tumors in most of human cancers (Renan, 1993). These stages are reflected by observable pathological lesions and the transition from one stage to the next higher stage may involve several genetic changes and/or epigenetic changes.

In this section, we gave brief summary of cancer genetics, and some well-recognized carcinogenesis pathways, such as TGF-β pathway, p53 pathway, and so on.

Usually, three types of genes contribute to cancer phenotype: oncogenes (dominant cancer genes), suppressor genes (recessive genes) and mis-match repair genes (MMR). Cancer can be initiated either by the activation of an oncogene, or by inactivating or silencing a suppressor gene, or a MMR gene.

If an oncogene is mutated or expressed at high levels, normal growth of a normal cell is unleashed, leading to continuous proliferating. Most oncogenes interact with many other genes to promote proliferation or abrogate differentiation and inhibition effects of many other protection devises.

Tumor suppressor gene is a gene whose action is required for normal development. When this gene is mutated or inactivated, the cell can progress to cancer, usually by combination with other genetic changes. Because tumor suppressor genes play more active roles in cell cycle, for example, tumor suppressor genes either have a dampening or repressive effect on the regulation of the cell cycle (by controlling the gap stage (G_1 and G_2)) or promote apoptosis, or control the activation of an oncogene. Stage and rate limiting genes are usually tumor suppressor genes (Osada & Takahashi, 2002; Weinberg, 2007).

Mismatch repair (MMR) genes are actually tumor suppressor genes but with special functions that they recognize and repair erroneous insertion, deletion and mis-incorporation of base that arise during DNA replication and recombination, as well as repairing some forms of DNA damage. Mutation or deletion of MMR genes lead to microsatellite repeats and create a mutator phenotype, resulting in genetic instability and increasing mutation rates of many relevant cancer genes.

Carcinogenesis pathway is usually represented by rate-limiting genes and their order. It has been observed that the same type of cancer may be developed through more than one pathway. For example, lung cancer may involve as many as 4 pathways, such as Wnt, Akt, Hedgedog, and p53 pathways. Similarly, colon cancer is also developed through many pathways, TGF-β, ras, MAPK, PI3k-Akt and p53 pathways. Because

different pathway function differently, some pathways may sequentially appear in the cancer development, for example, p53 pathway always appear at the last stage of cancer development since mutation of p53 will arrogate apoptosis and lead cells to unregulated proliferation, while Wnt pathway usually appears in cancer early development because several proteins involved in this pathway have been associated with the ability of the cell surface Wnt-activated Wnt receptor complex to bind axin and disassemble the axin/GSK3 complex, and lead to transcript some oncogenes (Nusse, 2005).

To date, more than 40 pathways have been identified, and more will be discovered in the future. We gave brief summary of several pathways which are related to cancer development.

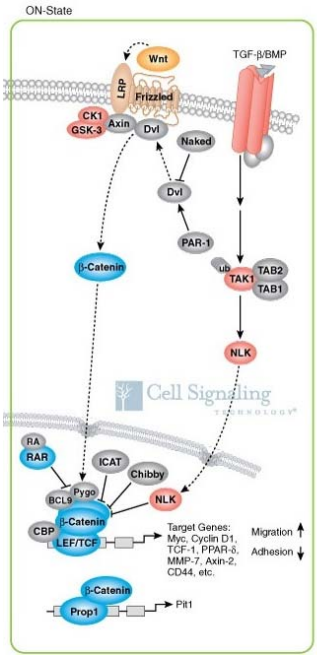


Figure 3. Wnt/β-Catenin Pathway
(Copyright Reserved by "Cell Signaling Technology")

Wnt/β-Catenin Pathway

As shown in Figure 3, the Wnt is a secreted glycoprotein that binds to Frizzled receptors, which triggers a cascade resulting in displacement of the multifunctional kinase GSK-3β from the APC/Axin/GSK-3β-complex. In the presence of Wnt binding (On-state), Dishevelled (Dsh) is activated, seemingly at least in part by phosphorylation, which in turn recruits GSK-3β away from the degradation complex. This allows for stabilization of β-catenin levels, nuclear import and recruitment to the LEF/TCF DNA-binding factors where it acts as an activator for transcription by displacement of Groucho-HDAC co-repressors. Importantly, some human cancers harbor point-mutations in β-catenin leading to its deregulated stabilization, and APC as well as axin mutations have also been documented, underscoring the involvement of abnormal activation of this pathway in human tumors. During development the Wnt/β-catenin pathway integrates signals from many other pathways including FGF, TGF-β and BMP in many different cell-types and tissues (Bienz, 2000; Gordon & Nusse, 2006; Willert & Jones, 2006).

TGF-β Pathway

As shown in Figure 4, transforming growth factor-β (TGF-β) signaling plays a critical role in the regulation of cell growth, differentiation, and development in a wide range of biological systems. In general, signaling is initiated with ligand-induced oligomerization of serine/threonine receptor kinases and phosphorylation of the cytoplasmic signaling molecules Smad2 and Smad3 for the TGF-β pathway, or Smad1/5/8 for the bone morphogenetic protein (BMP) pathway. Phosphorylation of Smads by activated receptors results in their partnering with the common signaling transducer Smad4, and translocation to the nucleus. Activated Smads regulate diverse

biological effects by partnering with transcription factors resulting in cell-state specific modulation of transcription. TGF-β signaling is also documented to affect Smad-independent pathways, including Erk, SAPK/JNK and p38 MAPK pathways. Activation of Smad-independent pathways through TGF-β signaling is also common. Rho GTPase (RhoA) activates downstream target proteins, such as mDia and ROCK, to prompt rearrangement of the cytoskeletal elements associated with cell spreading, cell growth regulation, and cytokinesis (Herpin & Cunningham, 2007; Kitisin et al., 2007; Schmierer & Hill, 2007).

Akt Pathway

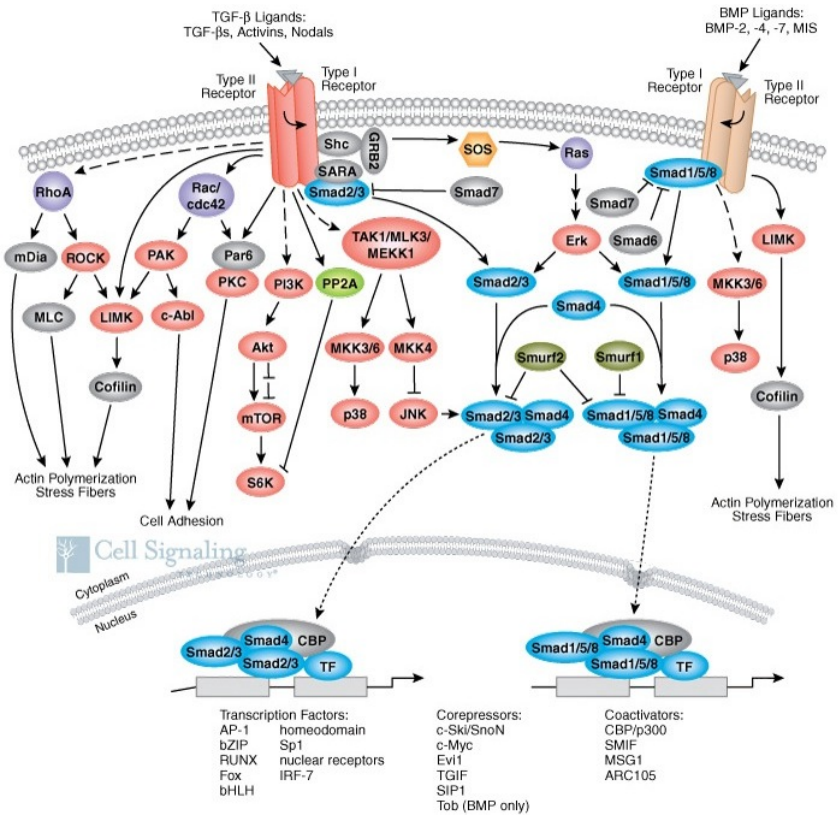


Figure 4. TGF-β Pathway
(Copyright Reserved by "Cell Signaling Technology")

The Akt (also known as protein kinase B or PKB) has become a major focus of attention because of its critical regulatory role in diverse cellular processes, including cancer progression and insulin metabolism. As shown in Figure 5, the Akt cascade is

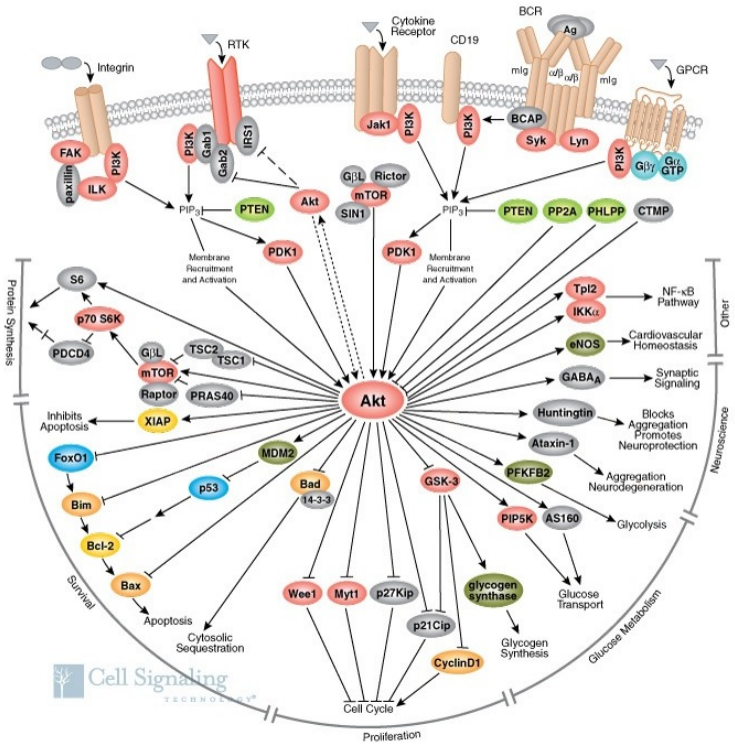


Figure 5. Akt Pathway
(Copyright reserved by "Cell Signaling Technology")

activated by receptor tyrosine kinases, integrins, B and T cell receptors, cytokine receptors, G-protein coupled receptors, and other stimuli. The three Akt isoforms (Akt1, Akt2, and Akt3) mediate many of the downstream events regulated by PI3K. For example, Akt is a major regulator of insulin signaling and glucose metabolism, with genetic studies in mice revealing a central role for Akt2 in these processes. Akt regulates cell growth through its effects on the mTOR and p70 S6 kinase pathways, as well as cell cycle and cell proliferation through its direct action on the CDK inhibitors p21 and p27, and its indirect effect on the levels of cyclin D1 and p53. Akt is a major mediator of cell

survival through direct inhibition of pro-apoptotic signals such as Bad and the Forkhead family of transcription factors. Recently, Akt has been demonstrated to interact with Smad molecules to regulate TGFβ signaling (Bhaskar & Hay, 2007; Brugge, Hung, & Mills, 2007; Carnero, Blanco-Aparicio, Renner, Link, & Leal, 2008).

p53 Pathway

As shown in Figure 6, p53 is a tumor suppressor protein that regulates the expression of a wide variety of genes involved in Apoptosis, Growth arrest, Inhibition of cell cycle progression, Differentiation and accelerated DNA repair or Senescence in response to genotoxic or cellular Stress. When the cell is confronted with stress like DNA damage, hypoxia, cytokines, metabolic changes and viral infection, p53 is activated (Francoz et al., 2006; Hanazono et al., 2006). In addition, p53 can transcriptionally activate PTEN (Phosphatase and Tensin Homolog), which may further inhibit Akt activity. Therefore, inhibition of Akt by the inhibitors may trigger a positive feedback with perhaps additional anti-tumor effects.

Cell cycle inhibition takes place when there is a block in cell-cycle division. p53 does this by stimulating the expression of p21/WAF1/CIP1 (Cyclin Dependent Kinase Inhibitor-p21). This protein is an inhibitor of CDKs (Cyclin-Dependent Kinases) that regulate the cell cycle via perturbation of their partner cyclin. Since p21/WAF1/CIP1 inhibits CDKs it results in inhibition of both G1-to-S and G2-to-mitosis transitions by causing hypophosphorylation of Rb (Retinoblastoma) and preventing the release of E2F. Additionally p53 can stimulate 14-3-3, a protein that sequesters Cyclin B1-CDK1 complexes out of the nucleus. This results in a G2 block. Activated p53 may also initiate

apoptosis and stop cell proliferation (Akhtar, Geng, Klocke, & Roth, 2006; Zheng, Ma, Zhu, Zhang, & Tong, 2006).

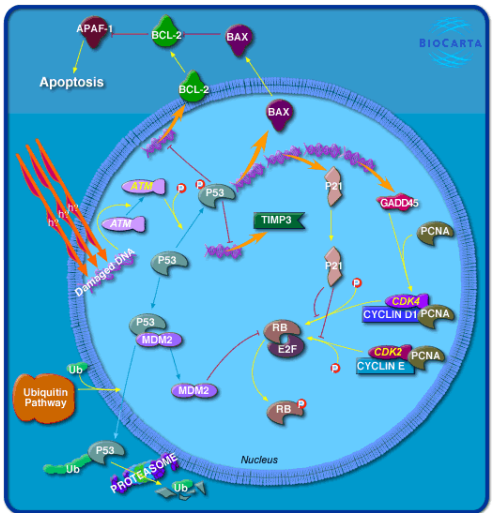


Figure 6. p53 Pathway
(Copyright Reserved by Biocarta)

Mutations in p53 are associated with genomic instability and increased susceptibility to cancer. It is the most frequently mutated protein in all cancer with an estimated 60% of all cancers having mutated forms that affect its growth suppressing activities. However some common tumors have a higher incidence such that 90% of cervical and 70% of colorectal are found to have p53 mutations. The p53 protein can be inactivated in several ways, including inherited mutations that result in a higher incidence of certain familial cancers such as Li-Fraumeni syndrome (Francoz et al., 2006; Gomes & Andrade, 2006; Jiang et al., 2006).

Parameters Estimation

We have shown in section 1.2 that a state space model for a complex system usually contains many parameters, depending on the number of pathways and the number of

stage in each pathway. In order to estimate the parameters simultaneously, Tan and Ye (2000) have developed general Bayesian procedures based on state space model, while the state variables are estimated by multi-level Gibbs sampling and weighted bootstrap procedures. In this section, we briefly introduced these procedures.

Information coming from three sources makes these procedures possible and reliable:

Prior information comes from previous studies or experiences, which give reasonable range or distribution of parameters.

Biological mechanisms, such as pathways we discussed in previous section, provide the information for constructing stochastic differential equations and conditional distribution.

Observation is represented by statistical model, which connects the biological system by incidence function, and also controls the randomness of observation.

We used two-pathway model described in section 1.1 as an example to illustrate the procedures. Let $\boldsymbol{X}$ be state variables $\{N(t), I_1(t), I_2(t), T(t)\}$, $\boldsymbol{\Theta}$ the set of all unknown parameters $\{\alpha_1, \alpha_2, \omega_1, \beta_1, b_1, d_1, b_2, d_2\}$ and $\boldsymbol{Y}$ the observed data, say cancer incidence for each age group. Let $P(\theta)$ be the prior distribution of the parameters, constructed by previous studies. If the prior information is vague, then a uniform or non-informative prior is applied. Let $P(\boldsymbol{X}|\boldsymbol{\Theta})$ be the conditional probability density of $\boldsymbol{X}$ given the parameters $\boldsymbol{\Theta}$ and $P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\Theta})$ the conditional probability density of $\boldsymbol{Y}$ given $\boldsymbol{X}$ and $\boldsymbol{\Theta}$. $P(\boldsymbol{X}|\boldsymbol{\Theta})$ is derived from the stochastic system model and $P(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\Theta})$ is derived from the observation model and is usually taken as likelihood function. By combining three

distributions, we can obtain the joint distribution of $\{\boldsymbol{X, Y, \Theta}\}$: $P(\boldsymbol{X, Y, \Theta}) = P(\boldsymbol{\Theta})\, P(\boldsymbol{X}|\boldsymbol{\Theta})\, P(\boldsymbol{Y}|\boldsymbol{X, \Theta})$. Then conditional distributions $P(\boldsymbol{X}|\boldsymbol{Y, \Theta})$ and $P(\boldsymbol{\Theta}|\boldsymbol{X, Y})$ are derived by:

$$\boldsymbol{P}(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{\Theta}) \propto \boldsymbol{P}(\boldsymbol{X}|\boldsymbol{\Theta})\boldsymbol{P}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\Theta})$$

$$\boldsymbol{P}(\boldsymbol{\Theta}|\boldsymbol{X}, \boldsymbol{Y}) \propto \boldsymbol{P}(\boldsymbol{\Theta})\boldsymbol{P}(\boldsymbol{X}|\boldsymbol{\Theta})\boldsymbol{P}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\Theta})$$

Using these conditional probability distributions, one can estimate the unknown parameters and predict the state variables by using the multi-level Gibbs sampling procedures (Liu & Chen 1998; Shephard, 1994, Tan, Zhang, Chen, & Zhu, 2008a; Tan, Zhang, Chen, & Zhu, 2008b). The algorithm is as following:

(1) Given $\boldsymbol{Y}$ and initial value of $\boldsymbol{\Theta}^{(0)}$, generated $\boldsymbol{X}^{(*)}$ from $P(\boldsymbol{X}|\boldsymbol{Y, \Theta}^{(0)})$.

(2) Generate $\boldsymbol{\Theta}^{(*)}$ from $P(\boldsymbol{\Theta}|\, \boldsymbol{X}^{(*)}, \boldsymbol{Y})$

(3) Back to step (1) to generate new $\boldsymbol{X}^{(*)}$ from updated $P(\boldsymbol{X}\,|\,\boldsymbol{Y, \Theta}^{(*)})$

(4) Repeat (1)-(3) till convergence.

The convergence of the above algorithm has been proved by Tan (2002). At convergence, one can generate a large set of $\boldsymbol{X}$ from $P(\boldsymbol{X}\,|\,\boldsymbol{Y})$, independent on $\boldsymbol{\Theta}$, and generate a large set of $\boldsymbol{\Theta}$ from posterior marginal distribution of $P(\boldsymbol{\Theta}|\,\boldsymbol{Y})$, independent on $\boldsymbol{X}$. Then we can use the sample means of $\boldsymbol{X}$ and $\boldsymbol{\Theta}$ as estimates of $\boldsymbol{X}$ and $\boldsymbol{\Theta}$ respectively, and use the sample variances as variances of the estimates. Alternatively, one can use also Efron's bootstrap method (Efron, 1982) to estimate the standard errors of estimates.

To implement the above procedures, one difficulty is that $P(\boldsymbol{X}\,|\,\boldsymbol{Y, \Theta})$ is hard to derive in practice. In order to generate $\boldsymbol{X}$ from conditional distribution of $P(\boldsymbol{X}\,|\,\boldsymbol{Y, \Theta})$, Tan and Chen (1998) have developed an indirect method based on the weighted bootstrap method (Smith & Gelfand, 1992). The algorithm used in weighted bootstrap method follows:

(1) Given $\boldsymbol{\Theta}^{(0)}$ and $\widetilde{\mathrm{X}}(l)$ $(0 \leq l \leq \mathrm{j})$, generate a large number of random sample of size N from $P(\tilde{X}(j+1)|\tilde{X}(j))$ (from the stochastic system model), denoted them by $\{\tilde{X}^{(1)}(j+1), \ldots, \tilde{X}^{(N)}(j+1)\}$.

(2) Compute $\Omega_k = P(\tilde{Y}(j+1)|\tilde{X}(s), s = 0,1, \ldots, j, \tilde{X}^{(k)}(j+1), \Theta^{(0)})$ and $\mathrm{p_k} = \Omega_\mathrm{k} / \sum_{\mathrm{i}=1}^{\mathrm{N}} \Omega_\mathrm{i}$ for k = 1, …, N.

(3) Construct a population π with element $(\mathrm{E}_1, \ldots, \mathrm{E}_\mathrm{N})$ and with $\mathrm{P}(\mathrm{E_k}) = \mathrm{p_k}$. Draw an element randomly from π, if the outcome is $\mathrm{E_k}$, then $\widetilde{\mathrm{X}}^{(\mathrm{k})}(\mathrm{j}+1)$ is the element $\widetilde{\mathrm{X}}(\mathrm{j}+1)$ from the conditional $\mathrm{P}(\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{\Theta})$.

(4) Repeat (1)-(3) until $\mathrm{j} = \mathrm{T_M}$, which is largest time point for discrete time in stochastic system.

2. STOCHASTIC MODELS OF CARCINOGENESIS FOR INITIATION-PROMOTION BIOASSAY AND APPLICATIONS

Introduction

To assess cancer risk of environmental agents, a common approach is to conduct initiation-promotion experiments using the mouse skin bioassay system to test if the agent can initiate cancer and/or promote cancer (Misfeld, 1980; Waters, Sandhu, Huisingh, Claxton, & Nesnow, 1981). For example, during the 1980's, the US EPA had conducted extensive initiation-promotion experiments to test a wide range of environmental risk agents. These environmental agents include among others, benzo(a)pyrene, topside coke oven extract, coke oven main extract, Nissan extract, roofing tar extract, Oldsmobile extract, Mercedes extract, Caterpillar extract, residential extract, and Mustang extract. For each of these environmental agents, valuable data concerning the number of papillomas and carcinomas had been generated. Many of these data were summarized in the paper by Nesnow et al. (1982); to date these useful data have only been used to construct empirical dose- response curves without any input from biologically supported model of carcinogenesis. Tan, Chen, and Wang (2001) have used the average number of papillomas per animal over different time points from Nissan extract to develop a state space model for the generation of papillomas. In this paper we will develop a comprehensive stochastic model of carcinogenesis for these experiments. We will show that even with some summary data as given in Nesnow et al. (1982), by using these models we will be able to extract extensive information than are possible by statistical methods.

In Section 2.2, we will develop a stochastic model for the initiation-promotion experiments using mouse skin bioassay. In Section 3, we will derive stochastic equations and probability distributions of state variables in these experiments. Using results from Section 2.3, in Section 2.4 we will develop statistical models and probability distributions for data on papillomas and carcinomas from these experiments. In Section 2.5, by using results from the above sections, we will develop some statistical inference procedures to estimate unknown genetic parameters, to validate the model and to predict future cases. In Section 2.6, we will apply the models and methods of this paper to analyze some summary data in Nesnow et al. (1982). We will show that even with some summary data we will be able to derive many useful results and extract some important information from the system. To further confirm the usefulness of our model and methods, in Section 2.7 we will generate some Monte Carlo data to illustrate the model and methods, will discuss the usefulness of the model and methods.

A Stochastic Model for Initiation-Promotion Bioassay

In initiation-promotion bioassay experiments, mouse skin is shaved and is exposed to a chemical called "Initiator" for a short time period (normally a few days or a week); immediately following initiation the exposed mice are treated by a chemical called "Promoter" weekly or twice weekly until sacrifice or termination of experiment. From studies of molecular biology of skin cancer (DiGiovanni 1992; Hennings et al., 1993; Missero, D'Errico, Dotto, & Dogliotti, 2002; Weinberg, 2007, pp. 435-439; Yuspa, 1994), carcinogenesis for this bioassay experiment to generate papillomas and carcinomas can best be described by a multiple pathway model involving a two-stage

model and an one-stage model; see **Remark 1.** This is represented schematically in Figure 7.

In Figure 7, the animal is in stage I_1 if the H-ras oncogene in a skin stem cell of the animal has been mutated (Brown, Buchmann, & Balmain, 1990; Missero et al., 2002; Weinberg, 2007, P.439). The I_1 stage animal is in stage I_2 if the p53 gene in an I_1 cell of this animal has been mutated or deleted or inactivated (Missero et al., 2002; Weinberg, 2007; Ruggeri et al., 1991); the animal is in stage I_2 immediately after treatment if the agent can induce mutations of both the H-ras gene and the p53 gene simultaneously. Thus, an agent is an initiator if the agent can induce H-ras mutation in skin stem cells in animals and is a strong initiator if it can induce mutation of both the H-ras gene and the

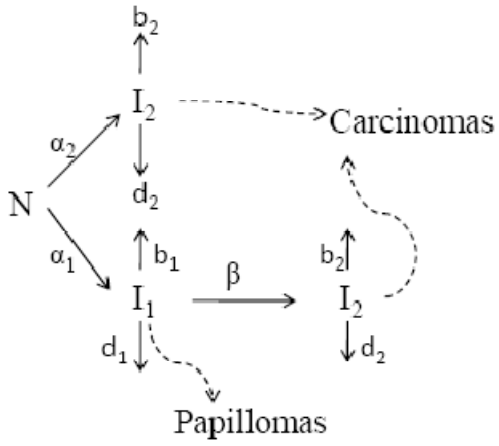


Figure 7. Multiple-pathway for Initiation and Promotion

p53 gene at the same time in some skin stem cells of mice. Promoters such as TPA (12-0 tetradecanoylphorbol-13-acetate) cannot induce genetic changes but only promote cell proliferation of initiated cells (Hennings et al., 1993; Weinberg, 2007). Karin, Liu, and Zandi (1997), Saez et al. (1995), and Weinberg (2007, P.439) have shown that promoters

such as TPA act synergistically with the H-ras gene protein to activate kinase Cα (PKC-α) which in turn activates the NK-κB, the AP-1 (heterodimer of c-jun and c-fos oncogenes) and the MAPK signal pathways to drive cell proliferation; Missero, Ramon, Cajal, and Dotto (1991) and Missero et al. (2002) have also noticed that promoters may facilitate cell proliferation of I_1 cells by inhibiting TGF-β (Transforming Growth Factor β) activities.

In the model in Figure 7, the I_r ($r = 1, 2$) cells are subjected to stochastic birth and death and each cell proceeds forward independently of other cells. Further, papillomas develop from primary I_1 cells through clonal expansion (i.e., stochastic birth-death) of these cells, where primary I_1 cells are I_1 cells generated directly by normal stem cells. Similarly, carcinomas develop from primary I_2 cells by clonal expansion (stochastic birth and death) of these cells, where primary I_2 cells are I_2 cells generated directly by mutation from I_1 cells. From this it is clear that the two-stage model in Figure 7 is not the classical MVK two-stage model because the MVK model assumes that I_2 cells grow instantaneously into carcinomas as soon as they are generated ignoring completely the stochastic birth and death of I_2 cells; see Tan (1991, Chapter 3).

Let t_0 be the time to start the experiment and $t_0^{(+)}$ the time to end initiation. For the model in Figure 7, the response variables (state variables) are the number $I_r(t)$ ($r = 1, 2$) of I_r cells at time t ($t \geq t_0$), the number $Y_P(t)$ of papillomas at time t and the number $Y_C(t)$ of carcinomas at time t.

In the next section, we will derive probability distributions of these variables under different bioassay experiments.

Remark 2.1. Because the promoter cannot induce genetic changes, the data in Table 2 implied that carcinomas could only be derived from I_2 cells generated by the initiator during the short initiation period.

The Stochastic Equations and Probability Distribution of State Variables in Initiation-Promotion Experiments

Let $\tilde{X}(t) = \{I_r(t),\ r = 1,\ 2\}$. To derive probability distribution of the state variables, observe that $\{X(t),\ t \geq t_0\}$ is a two-dimensional Markov process although $Y_P(t)$ and $Y_C(t)$ may not be Markov. For simplicity, notice further that we can make the following two assumptions: (a) Since the number of normal stem cells is very large (i.e., $10^8 \sim 10^9$) and the size of organ in adult mice is stable, it is reasonable to assume that the number $N(t)$ of normal stem cells at time t is a deterministic function of t and is a constant (i.e., $N(t) = N_0$). (b) Because the spontaneous mutation rates from $N \rightarrow I_r$ ($r = 1,\ 2$) and from $I_1 \rightarrow I_2$ are very small, in practice one can ignore spontaneous mutations from $N \rightarrow I_r$ ($r = 1,\ 2$) and from $I_1 \rightarrow I_2$; thus we assume that $N \rightarrow I_r$ ($r = 1,\ 2$) mutations can only be generated by initiators and $I_1 \rightarrow I_2$ can only be generated by initiators or mutagen (not by promoters; see Hennings et al., 1983).

Experiments in Which the Agent Is Used as an Initiator

In these experiments, similar animals are treated by the testing agent at time t_0 for a very short period with k different dose levels ($u_i,\ i = 1, \ldots, k$); then the treated animals are promoted by a well-known promoter such as TPA with fixed dose level weekly or twice weekly until sacrifice or the end of experiment. Let s_j ($j = 1, \ldots, m$) be the time to start the $j-th$ round of promotion with promotion period $s_{j+1} - s_j = 7$ days when promotion is applied weekly or 3 and half days when promotion is applied twice weekly.

If the agent is an initiator, it would induce mutation of the H-ras oncogene in some normal skin stem cells in mice to generate I_1 cells during the initiation period $(t_0, t_0^+]$ and hence papillomas at later times. If the agent is a strong carcinogen, with positive probability it may also induce mutations of both the H-ras gene and the p53 gene simultaneously in some skin stem cells in animals to generate I_2 cells during $(t_0, t_0^+]$ and hence carcinomas at later times; see data Table 2 where both papillomas and carcinomas are generated by the initiator.

Let $\alpha_r(i) > 0$ $(r = 1, 2)$ be the mutation rate from $N \rightarrow I_r$ induced by the agent with dose level u_i during the initiation period, where $(\alpha_1(i) > \alpha_2(i) > 0)$. Let $X_r(i)$ $(r = 1, 2)$ be the number of I_r cells generated during $(t_0, t_0^+]$ in the animal treated by the agent with dose level u_i. Because the $\alpha_r(i)$ $(r = 1, 2)$'s are very small, the $X_r(i)$ $(r = 1, 2, i = 1, \ldots, k)$ are distributed as Poisson random variables with mean $\lambda_r(i)$ respectively, where $\lambda_r(i) = \alpha_r(i)N_0$ $(r = 1, 2, i = 1, \ldots, k)$. That is,

$$X_r(i) \sim Poisson\{\lambda_r(i)\}. \tag{2.1}$$

The Stochastic Equations and Probability Distribution of State Variables

Let $\{B_r(j), D_r(j)\}$ $(r = 1, 2)$ denote the number of birth and the number of death of the I_r cells respectively in the animal during $(s_j, s_{j+1}]$ $(j = 1, \ldots, m)$. Because the promoter would not induce genetic changes but only increase the proliferation rates of I_r cells, by the conservation law we have the following stochastic equations for $I_r(t)$ $(r = 1, 2)$:

$$I_r(s_{j+1}) = I_r(s_j) + B_r(j) - D_r(j), j = 1, \ldots, m, \tag{2.2}$$

$$I_r(t) = I_r(t_0^{(+)}) = X_r(i) \text{ for } t_0^{(+)} \le t \le s_1, i = 1, \ldots, k.$$

In the above equations, notice that the $\{B_r(j), D_r(j)\}$ are independent of the agents and hence independent of the dose levels u_i of the agent. Notice also that only the

initiated number of I_r cells at $t_0^{(+)}$ is dependent on the agent and hence the dose levels of the agent.

To derive probability distributions of the $I_r(t)$ ($r = 1, 2$), let $b_r(j)$ and $d_r(j)$ denote the probability of birth and the probability of death of the I_r cells during the time period $(s_j, s_{j+1}]$ respectively. (Because $s_{j+1} - s_j$ is a constant, one may assume $(b_r(j) = b_r, d_r(j) = d_r)$.) These probabilities depend only on the promoter and hence are independent of the agent and the dose level of the agent. It is shown in Tan (2002, Chapter 8) that the conditional probability distribution of $\{B_r(j), D_r(j)\}$ given $I_r(s_j)$ is multinomial with parameters $\{I_r(s_j), b_r, d_r\}$. That is,

$$\{B_r(j), D_r(j)\} \mid I_r(s_j) \sim Multinomial\{I_r(s_j); b_r, d_r\},\ r = 1, 2,\ j = 1, \ldots, m. \tag{2.3}$$

Let $f(x; N, p)$ denote the density at x of $X \sim Binomial\{N, p\}$. Using equation (2.2) and the above conditional probability distribution of $\{B_r(j), D_r(j)\}$ given $I_r(s_j)$, we obtain for ($r = 1, 2, j = 1, \ldots, m, i = 1, \ldots, k$):

$$P\{I_r(s_{j+1}) = n2 \mid I_r(s_j) = n_1\} = \sum_{u=0}^{n_1} f(u; n_1, b_r)\delta(n_1 - n_2 + u) \times f\{n_1 - n_2 + u; n_1 - u, \frac{d_r}{1-b_r}\}, \tag{2.4}$$

where $I_r(s_1) = X_r(i)$, $r = 1, 2$ under treatment with dose level u_i and where $\delta(n_1 - n_2 + u) = 1$ if $n_1 - n_2 + u \geq 0$ and $= 0$ if $n_1 - n_2 + u < 0$.

Let $h(x; \lambda_r(i))$ denote the density at x of the Poisson distribution $X_r(i) \sim Poisson\{\lambda_r(i)\}$. Then, for $r = 1, 2$ and $i = 1, \ldots, k$, the probability $P\{I_r(s_2) = n; i\}$ of $I_r(s_2) = n$ under treatment with dose level u_i is:

$$P\{I_r(s_2) = n; i\} = \sum_{u=0}^{\infty} h\{u; \lambda_r(i)\} \sum_{v=0}^{u} f\ (v; u, b_r)\delta(u - n + v) \times f(u - n + v; u - v, \frac{d_r}{1-b_r}). \tag{2.5}$$

The Probability Distribution of the Number of Papillomas

To derive probability distribution of the number of papillomas, observe that papillomas are produced by clonal expansion (i.e., stochastic birth and death process) of I_1 cells generated by the initiator during the initiation period and that the I_1 cells can only proliferate under promotion by the promoter after time s_1. Let $P_I(s_1, t) = P_I(t - s_1)$ denote the probability that a primary I_1 cell at $s1$ develops into a detectable papillomas by time t ($t \geq s_1 > t_0^{(+)}$). Then, as proved in Tan (2002, Chapter 8),

$$P_I(s_1, t) = P_I(t - s_1) = \left(\frac{\theta_I[1-e^{-\gamma_1(t-s_1)}]}{\theta_I+(1-\theta_I)e^{-\gamma_1(t-s_1)}}\right)^{N_1-1} \times \frac{1}{\theta_I+(1-\theta_I)e^{-\gamma_1(t-s_1)}}, \quad (2.6)$$

where $\gamma_1 = b_1 - d_1$, $\theta_I = \frac{b_1}{\gamma_1}$ and where N_I is the number of I_1 cells in the papillomas for the papillomas to be detectable.

Let $Y_P(t; i)$ be the total number of papillomas by time t derived from the $X_1(i)$ I_1 cells which are generated by the initiator with dose level u_i during $(t_0, t_0^+]$. Then

$Y_P(t; i) | X_1(i) \sim Binomial\{X_1(i), P_I(t - s_1)\}$. Since $X_1(i) \sim Poisson\{\lambda_1(i)\}$, so,

$P(Y_P(t; i) = j) = \sum_{u=0}^{\infty} h(u; \lambda_1(i)) f(j; u, P_I(t - s_1))$

$= \sum_{u=0}^{\infty} \frac{1}{u!} e^{-\lambda_1(i)} [\lambda_1(i)]^u \binom{u}{j} [P_I(t - s_1)]^j [1 - P_I(t - s_1)]^{u-j}$

$= \frac{1}{j!} e^{-\lambda_1(i)} [\lambda_1(i) P_I(t - s_1)]^j \sum_{u=j}^{\infty} \frac{1}{(u-j)!} \{\lambda_1(i)[1 - P_I(t - s_1)]\}^{u-j}$

$= \frac{1}{j!} e^{-\lambda_1(i)} [\lambda_1(i) P_I(t - s_1)]^j e^{\lambda_1(i)[1-P_I(t-s_1)]}$

$= \frac{1}{j!} e^{-\lambda_1(i) P_I(t-s_1)} [\lambda_1(i) P_I(t - s_1)]^j.$ (2.7)

It follows that,

$Y_P(t; i) \sim Poisson\{\lambda_1(i) P_I(t - s_1)\}.$ (2.8)

The Probability Distribution of Carcinomas

To derive probability distribution of the number of carcinomas, observe that carcinomas are produced by clonal expansion (i.e., stochastic birth and death process) of I_2 cells generated by the initiator. Let $P_C(s_1, t)$ denote the probability that an I_2 cell at s_1 develops into a detectable carcinomas by time t ($t \geq s_1 > t_0^+$). Then, as shown in Tan (2002, Chapter 8),

$$P_C(s_1, t) = P_C(t - s_1) = \left(\frac{\theta_c[1-e^{-\gamma_2(t-s_1)}]}{\theta_c + (1-\theta_c)e^{-\gamma_2(t-s_1)}}\right)^{N_c-1} \times \left(\frac{1}{\theta_c + (1-\theta_c)e^{-\gamma_2(t-s_1)}}\right), \quad (2.9)$$

where $\gamma_2 = b_2 - d_2$, $\theta_C = \frac{b_2}{\gamma_2}$and where N_C is the number of I_2 cells in the carcinoma for the carcinoma to be detectable.

Let $Y_C(t; i)$ be the total number of carcinomas by time t produced by the $X_2(i)$, I_2 cells which are generated by the initiator with dose level u_i during $(t_0, t_0^+]$. Then

$Y_C(t; i) | X_2(i) \sim Binomial\{X_2(i), P_C(t - s_1)\}$. Since $X_2(i) \sim Poisson\{\lambda_2(i)\}$, so,

$$Y_C(t; i) \sim Poisson\{\lambda_2(i)P_C(t - s_1)\},\ i = 1, \ldots, k. \quad (2.10)$$

Experiments in Which the Agent Is Used as a Promoter

When the agent is an initiator, the next step is to test if the agent is also a promoter. In the initiation-promotion experiment, one then uses a well-known initiator such as B(a)P (benzo[a]pyrene) as the initiator with fixed dose level but use the testing agent as the promoter with k dose levels u_i ($i = 1, \ldots, k$). Because the I_1 cells can proliferate only under promotion, if the agent is not a promoter, the number of I_1 cells are very small and hence papillomas and carcinomas would not be generated. On the other hand, if the agent is a promoter, then, papillomas and possibly carcinomas will be generated at latter times.

In this experiment, when both the initiator and the testing agent (i.e., the promoter) are strong carcinogens, the I_r ($r = 1, 2$) cells are generated not only by the initiator during the initiation period $(t_0, t_0^{(+)}]$, but also by the testing agent during promotion periods $(s_j, s_{j+1}]$ $(j = 1, \ldots, m)$.

The Stochastic Equations and Probability Distribution of State Variables

Let $X_r^{(I)}(r = 1, 2)$ denote the number of I_r cells generated by the initiator during the initiation period and $X_r(i, j)$ $(r = 1, 2, i = 1, \ldots, k, j = 1, \ldots, m)$ the number of I_r cells generated by the testing agent with dose level u_i during the promotion period $(s_j, s_{j+1}]$. Let v_r $(r = 1, 2)$ be the mutation rate of $N \rightarrow I_r$ by the initiator. ($v_r \neq \alpha_r$ because the initiator used is not the testing agent.) Then, as in Section (2.3.1), we have for $(r = 1, 2)$:

$$X_r^{(I)} \sim Poisson\{\omega_r\}, \text{ with } \omega_r = v_r N_0,$$

$$X_r(i, j) \sim Poisson\{\lambda_r(i)\}, \text{ independently for } i = 1, \ldots, k, j = 1, \ldots, m. \quad (2.11)$$

To derive stochastic equations for the state variables, let $\{B_r(i, j), D_r(i, j)\}$ $(r = 1, 2, j = 1, \ldots, m, i = 1, \ldots, k)$ denote the number of birth and the number of death of the I_r cells respectively in the animal during $(s_{j-1}, s_j]$ under promotion by the agent with dose level u_i. Let $M_1(i, j)$ $(j = 1, \ldots, m)$ denote the number of mutation from $I_1 \rightarrow I_2$ in the animal during $(s_j, s_{j+1}]$ under promotion by the agent with dose level u_i. By the conservation law we have the following stochastic equations for $I_r(t)$ $(r = 1, 2)$:

$$I_1(s_{j+1}) = I_1(s_j) + X_1(i, j) + B_1(i, j) - D_1(i, j), j = 1, \ldots, m,$$

$$i = 1, \ldots, k \text{ with } I_1(t) = I_1(t_0^{(+)}) = X_1^{(I)} \text{ for } t_0^{(+)} \leq t \leq s_1; \quad (2.12)$$

$$I_2(s_{j+1}) = I_2(s_j) + M_1(i, j) + X_2(i, j) + B_2(i, j) - D_2(i, j), j = 1, \ldots, m,$$

$$i = 1, \ldots, k, \text{ with } I_2(t) = I_2(t_0^{(+)}) = X_2^{(I)} \text{ for } t_0^{(+)} \leq t \leq s_1. \quad (2.13)$$

To derive probability distribution of the state variables, let $b_r(j; i) = b_r(i)$ and $d_r(j; i) = d_r(i)$ denote the probability of birth and the probability of death of the I_r ($r = 1, 2$) cells during the interval $(s_j, s_{j+1}]$ under promotion with dose level u_i respectively. Let $\beta_1(j; i) = \beta_1(i)$ be the probability of mutation from $I_1 \rightarrow I_2$ during the promotion period $(s_j, s_{j+1}]$ under promotion by the agent with dose level u_i. Then as in the previous section, we have:

$\{B_r(i, j), D_r(i, j)\} | I_r(s_j) \sim Multinomial\{I_r(s_j), b_r(i), d_r(i)\}$, independently for $r = 1, 2, j = 1, \ldots, m, i = 1, \ldots, k$; (2.14)

$M_1(i, j) | I_1(s_j) \sim Poisson\{I_1(s_j)\beta_1(i)\}$, independent of $B_r(i, j)$, $D_r(i, j)$, $X_r(i, j)$, $X_2^{(I)}$, $r = 1, 2, j = 1, \ldots, m, i = \ldots, k.$ (2.15)

Let $g(x, y; N, p, q)$ denote the density at (x, y) of a multinomial distribution $(X, Y) \sim Multinomial\{N; p, q\}$. Using equations (2.12)-(2.13) and the probability distributions in equations (2.14)-(2.15), we obtain:

$$P\{I_1(s_{j+1}) = n_1 \mid I_1(s_j) = r_1\} = \sum_{u_1=0}^{r_1} \sum_{v_1=0}^{r_1-u_1} g\{u_1, v_1; r_1, b_1(i), d_1(i)\}$$
$$\times h\{n_1 - r_1 - u_1 + v_1; \lambda_1(i)\}\, \delta(n_1 - r_1 - u_1 + v_1) \quad (2.16)$$

$$P\{I_2(s_{j+1}) = n_2 \mid I_1(s_j) = r_1, I_2(s_j) = r_2\} = \sum_{l=0}^{\infty} \sum_{u_2=0}^{r_2} \sum_{v_2=0}^{r_2-u_2} h\{l; \lambda_2(i)\}$$
$$\times g\{u_2, v_2; r_2, b_2(i), d_2(i)\}\, h(n_2 - r_2 - l - u_2 + v_2; r_1\beta_1(i))$$
$$\times \delta(n_2 - r_2 - l - u_2 + v_2) \quad (2.17)$$

where $\delta(x) = 1$ if $x \geq 0$ and $= 0$ if $x < 0$.

To derive probability distributions of papillomas and carcinomas, notice that papillomas are generated from $X_1^{(I)}$ and $X_1(i, j)$ cells whereas carcinomas are generated from $X_2^{(I)}$ cells, $X_2(i, j)$ cells and $M_1(i, j)$ cells.

The Probability Distribution of Papillomas

To derive probability distribution of the number of papillomas, observe that each papilloma is derived by clonal expansion (i.e., stochastic birth and death process) from a single primary I_1 cell; $X_1^{(I)}$ are primary I_1 cells generated by the initiator during the initiation period and $X_1(i, j)$ are primary I_1 cells generated by the testing agent during promotion periods $(s_j, s_{j+1}]$ $(j = 1, \ldots, m)$. Let $Y_P^{(I)}(t; i)$ denote the number of papillomas by time t derived from $X_1^{(I)}$ cells and $Y_P(t; i, j)$ the number of papillomas by time t derived from $X_1(i, j)$ cells. Then these variables (i.e., $\{Y_P^{(I)}(t; i), Y_P(t; i, j), i = 1, \ldots, k, j = 1, \ldots, m\}$) are independently distributed of one another. As in the previous section, the probability distributions of these variables are:

$$Y_P^{(I)}(t; i) \sim Poisson\ \{\omega_1 P_I(t - s_1; i)\},$$

$$Y_P(t; i, j) \sim Poisson\ \{\lambda_1(i) P_I(t - s_{j+1}; i)\},$$

independently for $i = 1, \ldots, k, j = 1, \ldots, m,$ (2.18)

where

$$P_I(t - s_j; i) = \left(\frac{\theta_I[1-e^{-\gamma_1(i)(t-s_j)}]}{\theta_I(i)+(1-\theta_I(i))e^{-\gamma_1(t-s_j)}}\right)^{N_I-1} \times \left(\frac{1}{\theta_I(i)+(1-\theta_I(i))e^{-\gamma_1(i)(t-s_j)}}\right),$$

where $\gamma_1(i) = b_1(i) - d_1(i)$, $\theta_I(i) = \frac{b_1(i)}{\gamma_1(i)}$, $i = 1, \ldots, k.$ (2.19)

Let $Y_P(t; i)$ be the total number of papillomas by time t generated by the initiator and by the agent with dose level u_i. Then $Y_P(t; i) = Y_P^{(I)}(t; i) + \sum_{j=1}^{m} Y_P(t; i, j)$. It follows that $Y_P(t; i)$ is distributed as Poisson with mean $\psi_P(t; i) = \omega_1 P_I(t - s_i; i) + \sum_{j=1}^{m} \lambda_1(i) P_I(t - s_{j+1}; i)$. That is,

$$Y_P(t; i) \sim Poisson\{\psi_P(t; i)\}.$$ (2.20)

The Probability Distribution of Carcinomas

To derive probability distribution of the number of carcinomas, observe that each carcinoma is derived by clonal expansion from a single primary I_2 cell. Hence, carcinomas are derived by clonal expansion from $X_2^{(I)}$ cells, $X_2(i, j)$ cells and from $M_1(i, j)$ cells as these cells are primary I_2 cells. Let $Y_C^{(I)}(t; i)$ denote the number of carcinomas by time t derived from $X_2^{(I)}$cells, $Y_C(t; i, j)$ the number of carcinomas by time t derived from $X_2(i, j)$ cells and $Y_C^{(I)}(t; i, j)$ the number of carcinomas by time t derived from $M_1(i, j)$ cells. Then the variables $\{Y_C^{(I)}(t; i),\ Y_C(t; i, j),\ Y_C^{(I)}(t; i, j),\ i = 1, \ldots, k, j = 1, \ldots, n\}$ are independently distributed of one another. As in the previous section, the probability distributions of these variables are:

$$Y_C^{(I)}(t; i) \sim Poisson\ \{\omega_2 P_C(t - s_1; i)\},$$

$$Y_C(t; i, j) \sim Poisson\ \{\lambda_2(i) P_C(t - s_{j+1}; i)\},$$

independently for $i = 1, \ldots, k, j = 1, \ldots, m,$

where

$$P_C(t - s_j ; i) = \left(\frac{\theta_c(i)[1 - e^{-\gamma_2(i)(t-s_j)}]}{\theta_c(i) + (1 - \theta_c(i)) e^{-\gamma_2(t-s_j)}}\right)^{N_c - 1} \times \left(\frac{1}{\theta_c(i) + (1 - \theta_c(i)) e^{-\gamma_2(i)(t-s_j)}}\right), \qquad (2.21)$$

where $\gamma_2(i) = b_2(i) - d_2(i)$, $\theta_C(i) = \frac{b_2(i)}{d_2(i)}$, $i = 1, \ldots, k.$

To derive the probability distribution of $Y_c^{(I)}(t; i, j)$, observe that the conditional distribution of $Y_c^{(J)}(t; i, j)$ given $M_1(i, j)$ is binomial with parameters $\{M_1(i, j),\ P_C(t - s_{j+1}; i)\}$ and the conditional distribution of $M_1(i, j)$ given $I_1(s_j)$ cells is Poisson with mean $I_1(s_j)\beta_1(i)$. Hence the conditional distribution of $Y_c^{(I)}(t; i, j)$ given $I_1(s_j)$ is Poisson with

mean $I_1(s_j)\beta_1(i)P_C(t - s_{j+1}; i)$. It follows that to the order of $o(\beta_1(i))$, the distribution of $Y_c^{(I)}(t; i, j)$ is Poisson with mean $E[I_1(s_j)]\beta_1(i)P_C(t - s_{j+1}; i)$.

Let $Y_C(t; i)$ be the total number of carcinomas by time t generated by the initiator and by the agent with dose level u_i. Then $Y_C(t; i) = Y_C^{(I)}(t; i) + \sum_{j=1}^{m}[Y_C(t; i, j) + Y_C^{(J)}(t; i, j)]$. It follows that to the order of $o(\beta_1(i))$, $Y_C(t; i)$ is distributed as Poisson with mean

$\psi_C(t; i) = \omega_2 P_C(t - s_1; i) + \sum_{j=1}^{m}[\lambda_2(i) + E(I_1(s_j))\beta_1(i)]P_C(t - s_{j+1}; i)$. That is,

$$Y_C(t; i) \sim Poisson\{\psi_C(t; i)\}. \quad (2.22)$$

Experiments in Which the Agent Is Used as a Complete Carcinogen

The testing agent is a complete carcinogen if it is both an initiator and a promoter. To test if the environmental agent is a complete carcinogen, the agent is used both as an initiator and a promoter with dose levels $u_i (i = 1, \ldots, k)$. Thus, the experiment is similar to the experiment in Section 2.3.2 except that the initiator is now the testing agent with dose level u_i $(i = 1, \ldots, k)$. In these experiments, if papillomas are not generated, then either the agent is not an initiator, or the agent is not a promoter; the agent is a complete carcinogen if papillomas are generated in the experiment.

Stochastic Equations and Probability Distribution of the State Variables

Let $X_r(i; 0)$ $(r = 1, 2, i = 1, \ldots, k)$ be the I_r cells generated by the agent with dose level ui during the initiation period. Because the length of initiation period is very short and is approximately equal to the promotion period $s_{j+1} - s_j$, $X_r(i; 0) \sim Poisson\{\lambda_r(i)\}$ independently of $X_r(i; j)$ $(j = 1, \ldots, m)$, where $X_r(i; j)$ is defined in Section 2.3.2. Obviously, except with $I_r(t) = X_r(i; 0)$ $(r = 1, 2)$ for $t_0^{(+)} \le t \le s_1$ under the initiator with dose level u_i, the stochastic equations for $I_r(t)$ are exactly the same as given by equations

(2.12)- (2.13) respectively. Similarly the conditional probability distributions of $\{B_r(i; j), D_r(i; j)\}$ given $I_r(s_j)$ and the conditional probability distribution of $M_1(i, j)$ given $I_1(s_j)$ are given by equations (2.14)-(2.15) respectively. It follows that the conditional probability distribution of $\tilde{X}(s_{j+1})$ given $\tilde{X}(s_j)$ are given by equations (2.16)-(2.17) respectively.

The Probability Distributions of Papillomas and Carcinomas

Let $Y_P(t; i, 0)$ ($Y_C(t; i, 0)$) denote the number of papillomas (carcinomas) by time t generated by the initiator with dose level ui during the initiation period respectively. Then, with notations from Section 2.3.2, $Y_P(t; i) = \sum_{j=0}^{m} Y_P(t; i, j)$ and $Y_C(t; i) = \sum_{j=0}^{m}[Y_C(t; i, j) + Y_C^{(J)}(t; i, j)]$. Further,

$$Y_P(t; i, j) \sim Poisson\ \{\lambda_1(i)P_I(t - s_1; i)\}, j = 0, 1, \ldots, m \quad (2.23)$$

$$Y_C(t; i, j) \sim Poisson\{\lambda_2(i)P_C(t - s_{j+1}; i)\}, j = 0, 1, \ldots, m; \quad (2.24)$$

and to the order of $o(\beta_1(i))$,

$$Y_C^{(J)}(t; i, j) \sim Poisson\{E[I_1(s_j)]\beta_1(i)P_C(t - s_{j+1}; i)\}, j = 1, \ldots, m.$$

Hence,

$$Y_P(t; i) \sim Poisson\{\eta_P(t; i)\}. \quad (2.25)$$

$$Y_C(t; i) \sim Poisson\{\eta_C(t; i)\}, \quad (2.26)$$

where $\eta_P(t; i) = \sum_{j=0}^{m} \lambda_1(i)P_I(t - s_{j+1}; i)$ and

$$\eta_C(t; i) = \lambda_2(i)P_C(t - s_1; i) + \sum_{j=0}^{m}[\lambda_2(i) + E(I_1(s_j))\beta_1(i)]P_C(t - s_{j+1}; i).$$

Statistical Models and Probability Distributions

In initiation-promotion experiments, similar mice are randomized into k groups. In the i-th group, n_i ($i = 1, \ldots, k$) animals are treated by the testing agent with dose level u_i;

among these n_i animals, $n_i(l)$ animals are sacrificed at time t_l when autopsies are performed over these sacrificed animals and the numbers of papillomas and carcinomas are counted. The observed data available are the number $m_P(i)$ of animals with papillomas among the $n_i(l)$ animals sacrificed at time t_l in the $i-$th treatment group, the average number $Y_P(t_l, i)$ of papillomas per animal in the $i-$th treatment group over the $n_i(l)$ animals sacrificed at time t_l, the number $m_C(i)$ of animals with carcinomas in the $i-$th treatment group among the $n_i(l)$ animals sacrificed at time t_l and the average number $Y_C(t_l, i)$ of carcinomas per animal in the $i-$th treatment group among the $n_i(l)$ animals sacrificed at time t_l. Given in Tables 2-4 are these observed numbers in percentages with t_l equal to six months or one year. To develop statistical models for these observed data, in what follows we let $Y_P(v; t_l, i)$ be the number of papillomas by time t_l in the v-th mouse in the $i-th$ treatment group among the $n_i(l)$ mice sacrificed at time t_l; similarly, we let $Y_C(v; t_l, i)$ be the number of carcinomas by time t_l in the v-th mouse in the $i-th$ treatment group among the $n_i(l)$ mice sacrificed at time t_l.

Using results from Section 2.3, in this section we will develop statistical models for these observed variables and derive probability distributions generating these observed data under different initiation-promotion bioassay experiments.

Experiments in Which the Testing Agent Is Used as an Initiator

From results in Section (2.3.1.2) and equation (2.8), the probability that the animal in the $i-$th treatment group would develop at least one detectable papillomas by time t_l is

$$Q_P^{(1)}(t_l; i) = 1 - exp\{-\lambda_1(i)P_1(t_l - s_1)\}.$$

It follows that

$$m_P(i) \mid n_i(l) \sim Binomial\{n_i(l), Q_P^{(1)}(t_l; i)\}, \text{ independently for } i = 1, \ldots, k. \qquad (2.27)$$

The deviance $Dev_P^{(1)}(m; i)$ of the density in (2.27) is

$$Dev_P^{(1)}(m; i) = 2\{A_p(i) - m_P\ (i) \log\ [Q_P^{(1)}(t_l; i)] - [n_i(l) - m_P(i)] \log[1 - Q_P^{(1)}(t_l; i)]\},$$

(2.28)

where $A_P(i) = m_P(i) \log[m_P(i)] + [n_i(l) - m_P(i)] \log[n_i(l) - m_P(i)] - n_i(l) \log[n_i(l)]$.

By equation (2.8) in Section (3.1.2), we have also that

$Y_P(v; t_l, i) \sim Poisson\{\lambda_1(i)P_I(t_1 - s_1)\}$, independently for $v = 1, \ldots, n_i(l)$, $i = 1, \ldots, k$.

(2.29)

It follows that

$Y_P(t_l, i) = \frac{1}{n_i(l)} \sum_{v=1}^{n_i(l)} Y_P(v; t_l, i) = \lambda_1(i)P_I(t_l - s_1) + e_P^{(1)}(i)$, independently for $i = 1, \ldots, k$. (2.30)

In the above equation, $e_P^{(1)}(i)$ has expected value 0 and variance $\frac{1}{n_i(l)} \lambda_1(i)P_I(t_l - s_1)$.

When $n_i(l)$ is not small, by the central limit theorem one may practically assume that the $e_P^{(1)}(i)$ are independently distributed as normal random variables.

Using this distribution result, the deviance $Dev_P^{(1)}(Y\ ; i)$ of the density of $Y_P(t_i, i)$ is

$$Dev\ _P^{(1)}(Y\ ; i) = \log\{\lambda_1(i)P_1\ (t_l - s_1)\} - \log\ n_i(l) + \frac{n_i(l)}{\lambda_1(i)P_I\ (t_l - s_1)} \times (Y_P(t_l, i)\ - \lambda_1(i)P_I(t_l\ -\ s_1))^2 \quad (2.31)$$

Let $Q_C^{(1)}(t_l; i)$ be the probability that the animal in the i−th treatment group would develop at least one detectable carcinomas by time t_l. From results in Section (2.3.1.3) and equation (2.10), we obtain $Q_C^{(1)}(t_l; i) = 1 - exp\{-\lambda_2(i)P_C(t_l - s_1)\}$ and

$$m_C(i) \mid n_i(l) \sim Binomial\{n_i(l), Q_C^{(1)}(t_l; i)\}. \quad (2.32)$$

The deviance $Dev_C^{(1)}(m; i)$ of the density in (2.32) is

$$Dev_C^{(1)}(m;\, i) = 2\{A_C(i) - m_P(i) \log [Q_C^{(1)}(t_l;\, i)] - [n_i(l) - m_C(i)] \log[1 - Q_C^{(1)}(t_l;\, i)]\},$$

(2.33)

where $A_C(i) = A_C(i) \log[m_C(i)] + [n_i(l) - m_C(i)] \log[n_i(l) - m_C(i)] - n_i(l) \log[n_i(l)]$.

Similarly, by equation (2.10) in Section (3.1.3), we have that

$Y_C(v;\, t_l,\, i) \sim Poisson\{\lambda_2(i) P_C(t_l - s_1)\}$, independently for $v = 1, \ldots, n_i(l)$, $i = 1, \ldots, k$.

(2.34)

It follows that

$Y_C(t_l,\, i) = \frac{1}{n_i(l)} \sum_{v=1}^{n_i(l)} Y_C\ (v;\, t_l,\, i) = \lambda_1(i) P_l(t_l - s_1) + e_C^{(1)}(i)$, independently for $i = 1, \ldots, k$

(2.35)

By the central limit theorem, the $e_C^{(1)}(i)$ are distributed independently as normal variables with expected value 0 and variances $\frac{1}{n_i(l)} \lambda_2(i) P_c(t_l - s_1)$ approximately.

Using this distribution result, the deviance $Dev_c^{(1)}(Y;\, i)$ of the density of $Y_C(t_i,\, i)$ is

$$Dev_C^{(1)}(Y;\, i) = \log\{\lambda_2(i) P_C(t_l - s_1)\} - \log n_i(l) + \frac{n_i(l)}{\lambda_2(i) P_C(t_l - s_1)} \times (Y_C(t_l,\, i) - \lambda_2(i) P_C(t_l - s_1))^2.$$

(2.36)

Experiments in Which the Testing Agent Is Used as a Promoter

As in Section (2.3.2), assume that a well-known chemical (e.g., B(a)P) with certain fixed dose level is used as an initiator but the testing agent is used as a promoter over k dose levels (u_i, $i = 1, \ldots, k$).

Let $Q_P^{(2)}(t_l;\, i)$ be the probability that the animal in the i−th treatment group would develop at least one detectable papillomas by time t_l; and $Q_c^{(2)}(t_l;\, i)$ the probability that

the animal in the i−th treatment group would develop at least one detectable carcinomas by time t_l. Then, we have:

$Q_C^{(2)}(t_l; i) = 1 - exp\{-\psi_P(t_l; i)\}$, with $\psi_P(t; i) = \omega_1 P_1(t_1 - s_1; i) + \sum_{j=1}^{m} \lambda_1(i) P_1(t_l - s_{j+1}; i)$;

$Q_C^{(2)}(t_l; i) = 1 - exp\{-\psi_c(t_l; i)\}$, with $\psi_C(t; i) = \omega_2 P_c(t_1 - s_1; i) + \sum_{j=1}^{m} [\lambda_2(i) + E(I_1(s_j))\beta_1(i)] P_C(t_l - s_{j+1}; i)$,

where $P_1(t_l - s_j ; i)$ and $P_C(t_l - s_j ; i)$ are given in equations (2.18) and (2.22) respectively.

It follows that

$$m_P(i)|n_i(l) \sim Binomial\{n_i(l), Q_P^{(2)}(t_l; i)\}, \tag{2.37}$$

$$m_C(i)|n_i(l) \sim Binomial\{n_i(l), Q_P^{(2)}(t_l; i)\}, \tag{2.38}$$

independently for $i = 1, \ldots, k$.

The deviance $Dev_P^{(2)}(m; i)$ of the density in (2.37) and the deviance $Dev_C^{(2)}(m; i)$ of the density in (2.38) are given respectively by:

$$Dev_P^{(2)}(m; i) = 2\{A_P(i) - m_P(i) \log[Q_P^{(2)}(t_l; i)] - [n_i(l) - m_P(i)] \log[1 - Q_P^{(2)}(t_l; i)]\}, \tag{2.39}$$

$$Dev_C^{(2)}(m; i) = 2\{A_C(i) - m_P(i) \log[Q_C^{(2)}(t_l; i)] - [n_i(l) - m_C(i)] \log[1 - Q_C^{(2)}(t_l; i)]\}. \tag{2.40}$$

Similarly, from equations (2.18) and (2.22) in Section (2.3.2),

$Y_P(v; t_l, i) \sim Poisson\{\psi_P(t_l; i))\}$; $Y_C(v; t_l, i) \sim Poisson\{\psi_C(t_l; i))\}$, independently for $v = 1, \ldots, n_i(l)$, $i = 1, \ldots, k$.

It follows that

$$Y_P(t_l, i) = \frac{1}{n_i(l)} \sum_{v=1}^{n_i(l)} Y_P(v; t_l, i) = \psi_P(t_l; i) + e_P^{(2)}(i), \tag{2.41}$$

$$Y_C(t_l, i) = \frac{1}{n_i(l)} \sum_{v=1}^{n_i(l)} Y_P(v; t_l, i) = \psi_c(t_l; i) + e_c^{(2)}(i), \quad (2.42)$$

independently for $i = 1, \ldots, k$.

By the central limit theorem, the $e_P^{(2)}(i)$ and the $e_c^{(2)}(i)$ are approximately normally distributed with means 0 and variances $\frac{1}{n_i(l)}\psi_P(t_l; i)$ and $\frac{1}{n_i(l)}\psi_c(t_l; i)$ respectively.

The deviance $Dev_P^{(2)}(Y; i)$ of the density in (2.41) and the deviance $Dev_C^{(2)}(Y; i)$ of the density in (2.42) are given respectively by:

$$Dev_P^{(2)}(Y; i) = log\{\psi_1(t_l; i)\} - log\, n_i(l) + \frac{n_i(l)}{\psi_1(t_l; i)} \times (Y_P(t_l, i) - \psi_1(t_l; i))^2. \quad (2.43)$$

$$Dev_C^{(2)}(Y; i) = log\{\psi_2(t_l; i)\} - log\, n_i(l) + \frac{n_i(l)}{\psi_2(t_l; i)} \times (Y_C(t_l, i) - \psi_2(t_l; i))^2. \quad (2.44)$$

Experiments in Which the Testing Agent Is Used as a Complete Carcinogen

Let $Q_P^{(3)}(t_l; i)$ be the probability that the animal in the i−th treatment group would develop at least one detectable papillomas by time t_l; and $Q_P^{(3)}(t_l; i)$ the probability that the animal in the i−th treatment group would develop at least one detectable carcinomas by time t_l. Then, we have:

$Q_P^{(3)}(t_l; i) = 1 - exp\{-\eta_P(t_l; i)\}$, with $\eta_P(t; i) = \sum_{j=1}^{m} \lambda_1(i) P_I(t_l - s_{j+1}; i)$;

$Q_C^{(3)}(t_l; i) = 1 - exp\{-\eta_C(t_l; i)\}$, with

$\eta_C(t; i) = \lambda_2(i) P_C(t_l - s_1; i) + \sum_{j=1}^{m} [\lambda_2(i) + E(I_1(s_j))\beta_1(i)] P_C(t_l - s_{j+1}; i)$,

where $P_I(t_l - s_j; i)$ and $P_C(t_l - s_{j+1}; i)$ are given in equations (2.18) and (2.21) respectively.

It follows that

$$m_P(i) \mid n_i(l) \sim Binomial\{n_i(l), Q_P^{(3)}(t_l; i)\}; \quad (2.45)$$

$$m_C(i) \mid n_i(l) \sim Binomial\{n_i(l),\ Q_c^{(3)}(t_l;\ i)\},\tag{2.46}$$

independently for $i = 1, \ldots, k$.

The deviance $Dev_P^{(3)}(m;\ i)$ of the density in (2.45) and the deviance $Dev_c^{(3)}(m;\ i)$ of the density in (2.46) are given respectively by:

$$Dev_P^{(3)}(m;\ i) = 2\{A_P(i) - m_P(i) \log [Q_P^{(3)}(t_l;\ i)] - [n_i(l) - m_P(i)] \log [1 - Q_P^{(3)}(t_l;\ i)]\},\tag{2.47}$$

$$Dev_c^{(3)}(m;\ i) = 2\{A_C(i) - m_P(i) \log [Q_c^{(3)}(t_l;\ i)] - [n_i(l) - m_C(i)] \log [1 - Q_c^{(3)}(t_l;\ i)]\}.\tag{2.48}$$

Similarly, from equations (2.25) and (2.26) in Section (2.3.2),

$$Y_P\ (v;\ t_l,\ i) \sim Poisson\{\eta_P(t_l;\ i))\};\ Y_C(v;\ t_l,\ i) \sim Poisson\{\eta_c(t_l;\ i))\},$$

independently for $v = 1, \ldots, n_i(l)$, $i = 1, \ldots, k$.

It follows that

$$Y_P(t_l,\ i) = \frac{1}{n_i(l)} \sum_{v=1}^{n_i(l)} Y_P(v;\ t_l,\ i) = \eta_P(t_l;\ i) + e_P^{(3)}(i),\tag{2.49}$$

$$Y_C(t_l,\ i) = \frac{1}{n_i(l)} \sum_{v=1}^{n_i(l)} Y_c(v;\ t_l,\ i) = \eta_c(t_l;\ i) + e_c^{(3)}(i),\tag{2.50}$$

independently for $i = 1, \ldots, k$.

By the central limit theorem, the $e_P^{(3)}(i)$ and the $e_c^{(3)}(i)$ are approximately normally distributed with means 0 and variance $\frac{1}{n_i(l)}\eta_P(t_l;\ i)$ and $\frac{1}{n_i(l)}\eta_c(t_l;\ i)$, respectively.

The deviance $Dev_P^{(3)}(Y;\ i)$ of the density in (2.49) and the deviance $Dev_c^{(3)}(Y;\ i)$ of the density in (2.50) are given respectively by:

$$Dev_P^{(3)}(Y;\ i) = log\{\eta_1(t_l;\ i)\} - log\ n_i(l) + \frac{n_i(l)}{\eta_1(t_l;\ i)} \times (Y_P\ (t_l,\ i) - \eta_1(t_l;\ i))^2.\tag{2.51}$$

$$Dev_c^{(3)}(Y;\ i) = log\{\eta_2(t_l;\ i)\} - log\ n_i(l) + \frac{n_i(l)}{\eta_2(t_l;\ i)} \times (Y_C(t_l,\ i) - \eta_2(t_l;\ i))^2.\tag{2.52}$$

Statistical Inference Procedures to Estimate Parameters

To estimate the unknown parameters in initiation-promotion experiments, the data available are the number of animals with papillomas, the number of animals with carcinomas, the average number of papillomas per mouse and the average number of carcinomas per mouse. While the probability distribution of each of these variables is readily available (see Section above), the joint probability distribution of these variables are very complicated and remain to be derived. To make use of all data to estimate the unknown parameters in the Bayesian approach, in this chapter we will employ two different approaches: One approach is through the weighted average of the log of the posterior distribution from different data sets, the other approach is sequential using estimates from one data set as prior information for the posterior distribution using the other data sets and vice versa; see Remark 2.2.

Remark 2.2. As shown in the next section, the fitting of some actual data from some initiation-promotion experiments indicated that these two approaches yielded very similar results.

The Bayesian Approach

In the Bayesian approach, because prior information about the genetic parameters are vague and imprecise but previous biological information and studies provide some lower bounds and upper bounds for the parameters, we will assume a partially non-informative prior for these parameters. That is, we assume the prior distribution as proportional to positive constants within the bounded region but are zeros for otherwise. We will derive the posterior modes as estimates. For easy of computation, we will use the deviance of the posterior distribution which is the negative of the sum of the log of the

prior and the log of the standardized likelihood function. Under non-informative prior or partially non- informative prior (i.e., non-informative prior but with lower and upper bounds for the parameters), the deviance of the posterior distribution is equivalent numerically to the deviance function of the likelihood as defined in generalized linear models (Nelder, 1989). It follows that deriving the posterior model is equivalent to minimize the deviance function of the posterior distribution. Notice that under non-informative prior or partially non- informative prior, the estimates through deriving the posterior mode are equivalent numerically to the maximum likelihood estimates under constraints in the classical sampling theory approach.

Obviously, the traditional Newton-Raphson optimization would not work and is very inefficient to minimize the deviance under constraints, because the deviances are nonlinear and contain many unknown parameters. We would employ the genetic algorithm (GA) to derive the optimal values of parameters simultaneously. The genetic algorithm is an optimization process based on evolution principles involving gene mutation or chromosomal aberrations, mating types (referred to as crossing over in GA), as well as fitness-based selection.

In this thesis, we have used the genetic algorithm package "rgenoud" in "R", which was developed by Walter R. Mebane, Jr., and Jasjeet S. Sekhon. The advantages of using "R" are:

(1) The control arguments are clearly defined and easy to implement.

(2) If the gradients of parameters in an objective function are easy to derive and the estimating function is generally concave, the algorithm takes advantage of steepest descent method to locate the neighborhood region of optima.

(3) It takes advantage of pure genetic algorithm that global optima is searched based on fitness of selection procedure; this procedure avoids being trapped in a local optimum in the event that multiple local optima exist.

We will use this genetic algorithm in combination with the stochastic model of carcinogenesis and the Bayesian procedures to derive estimates of the unknown parameters (mutation rates, birth rates and death rates, etc.). In the Bayesian procedure, we will use a partially non-informative prior as the parameters satisfy prior constraints imposed by the cancer biological process.

Procedures to Estimate Parameters

We have used the above approach to estimate parameters in each experiment described in Section 2.3. The procedures of estimating the parameters are given as following:

(1) Data Augmentation:

Given parameter values, we use the stochastic equations shown in section 3 (for each experiment) and the associated probability distributions to generate a large sample of Z(j) = {X(j), U(j), j = 1, 2, …, T_M}, where T_M is the maximum time point under consideration, and where X(j)'s are state variables and U(j)'s are augmented variables. Next we combine these sample with the probably distribution of data given state variables and parameters and use the weighted Bootstrap method due to Smith and Gelfant (1992) to select a sample from this large sample, say $\hat{Z}(j)$. This $\hat{Z}(j)$ is then a sample of size one from the conditional probability distribution of Z(j) given data and given parameter values. Numerically this is equivalent to the E-step of the E-M algorithm in the sampling theory framework.

(2) Estimation of parameters given $\hat{Z}(j)$:

Taken $\hat{Z}(j)$ as mean values of the state variables and augmented variables from the stochastic system model, we apply genetic algorithm (GA) to minimize the deviance function given in section 2.4. These posterior modes are taken as estimates of the parameters $\hat{\Theta}$.

(3) Back to step (1) with $\{\hat{\Theta}\}$. and continue until convergence.

Table 2

Data from Initiation Experiment

Agent	Dose (µg/mouse)	No. Mice Surviving	Mice with Papillom-as[a] (%)	Papillomas per mouse[a]	Mice with Carcino-mas[b] (%)	Carcino mas per mouse[b]
B(a)P	0	37	8	0.08	5	.05
	2.52	40	45	0.5	5	0.07
	12.6	40	73	1.8	20	0.2
	50.5	39	100	5.8	25	0.25
	101	38	95	10.2	30	0.33
Coke Oven Main	100	38	50	0.63	10	0.1
	500	39	90	3.7	54	0.59
	1000	39	87	3.3	53	0.53
	2000	40	78	3.1	48	0.48

[a] Scored at 6 month; [b] Cumulative score at 12 month

An Illustrative Example

In this section we use the actually data from Nesnow et al. (1982) to illustrate how to estimate parameters in initiation-promotion experiments. Before stepping into parameter estimation, we describe the experiments from which the data were generated.

Observations from Initiation-Promotion Experiments

Initiation Experiment. In this experiment, B(a)P and coke oven main are taken as initiator. For each agent, 200 similar sencar mice were randomly assigned to 5 dose groups, 40 mice of each group. A single application of B(a)P (or coke oven main) to each mice is followed one week later by multiple application of the tumor promoter TPA. The data is shown in Table 2.

In this experiments, clearly both the B(a)P and the coke over main are strong carcinogens because both papillomas and carcinomas were observed. Thus these agents can not only induce mutation of Ras gene to generate I_1 cells but also generate simultaneous mutation of both the Ras-gene and the p-53 gene to generate I_2 cells. Notice that both of these mutations are dose depend; however, because TPA has only one dose level, the birth rates and death rates of the I_1 and I_2 cells are independent of dose levels.

Promotion Experiment. In this experiment, the Coke oven main was used as the promoter and was applied weekly to sencar mice. The mice were initiated with a single dose of 50.5 μg/mouse of B(a)P (benzo(a)pyrene) and are scored at 34 weeks. The observed data are given in Table 3. Notice that in this experiment, data on carcinomas were not recorded.

In this experiment, because the coke oven main is also a strong carcinogen, during the promotion periods both the mutation rates from N to I_1 and from N to I_2 are dose

Table 3

Data from Promotion Experiment

Dose (µg/mouse)	Mice with Papillomas[a] (%)	Papillomas per mouse[a]
100	3	0.02
500	26	0.44
1000	53	1.2
2000	84	2.5
4000	100	8.2

[a] Scored at 6 month; [b] Cumulative score at 12 month

dependent. Similarly, both the birth rates and the death rates of the I_1 cells and the I_2 cells are dose dependent.

Complete Experiment. In this experiment, the coke oven is used as both an initiator and a promoter. After initiation by coke oven main in the first week, mice were treated weekly with coke oven main at different dose levels; at the end of one year data on carcinomas were measured and is summarized in Table 4.

In this experiment, both the mutation rates from N to I_1 and from N to I_2 are dose dependent. Similarly, both the birth rates and the death rates of the I_1 cells and the I_2 cells are dose dependent.

Table 4

Data from Complete Carcinomas

Dose (μg/mouse)	Mice with Carcinomas[a] (%)	Carcinomas per mouse[a]
100	5	0.05
500	36	0.36
1000	48	0.55
2000	82	1.0
4000	98	0.98

[a] Scored at 6 month; [b] Cumulative score at 12 month

Estimation Scheme

Using data from the above three experiments, we will proceed to estimate genetic parameters relevant to coke oven main, which include mutation rates, birth rates and death rates. The basic reason we use all of the above experiments instead of just one complete experiment is that none of a single experiment contain data on both papillomas and carcinomas. We will proceed sequentially. That is, we use first experiment data to estimate part of all parameters and then apply these estimated values to the next experiment.

(a) Initiation Experiment. In above initiation experiment, the unknown parameters are: mutation rate from normal cells to I_1 cells (initiated by B(a)P or coke oven main), mutation rate from normal cells to I_2 cells for both agents, birth rate and death rate for I_1

and I_2 cells (due to TPA). We will the estimated mutation rates of B(a)P in this experiment as prior information in the promotion experiment and use the estimated mutation rates of coke oven main as prior information in the promotion and complete experiments. In this experiment, because the promoter is TPA, the birth rate and the death rates of I_1 and I_2 cells are independent of dose levels and are not completely independent of the birth rate and death rates of these cells in promotion and complete experiments in which coke oven main is used as promoter. In this experiment, as in Tan et al. (2001), the dose related mutation rate from the normal cell to I_1 and from N to I_2 are represented respectively by:

$$\lambda_1(u_i) = \alpha_{10}^{Initiator} exp(\alpha_{11}^{Initiator} \log(1 + u_i)) \tag{2.53}$$

and

$$\lambda_2(u_i) = \alpha_{20}^{Initiator} exp(\alpha_{21}^{Initiator} \log(1 + u_i)) \tag{2.54}$$

From (2.53)-(2.54), obviously, it is only possible to estimate $\alpha_{10}^{Initiator} P_I(t - t_0)$ and $\alpha_{11}^{Initiator}$ from papillomas data and only $\alpha_{20}^{Initiator} P_c\ (t - t_0)$and $\alpha_{21}^{Initiator}$ from carcinomas data; from these it follows that only $\alpha_{11}^{Initiator}$ can be used as prior information in the promotion experiment and only $\alpha_{21}^{Initiator}$ can be used as prior information in the complete experiment.

From these we also have:

$$Y_P(t, i) \sim Poison\{\alpha_{10}^{Initiator} P_I(t - t_0) * exp(\alpha_{11}^{Initiator} \log(1 + u_i))\}$$

$$Y_c(t, i) \sim Poison\{\alpha_{20}^{Initiator} P_c(t - t_0) * exp(\alpha_{21}^{Initiator} \log(1 + u_i))\} \tag{2.55}$$

(b) Promotion experiment. In promotion experiment, only papillomas response is recorded (Table 3). Coke oven main is not only a promoter but also a carcinogen; therefore papillomas are produced not only by initiator (B(a)P) at the initiation period, but also by the coke oven main during the promotion periods. In this experiment, let ω_r be the mutation rates from theinitiator, and $\lambda_r(u_i)$ from the coke oven main. Notice that ω_r is dose independent and $\lambda_r(u_i)$ are dose level dependent. These rates are represented respectively by:

$$\varpi_1(50.5) = \varpi_{10} * exp(\alpha_{11}^{B(a)P} \log(1 + 50.5)) \quad (2.56)$$

$$\lambda_r(u_r) = \alpha_{10}^{coke} * exp(\alpha_{11}^{coke} \log(1 + u_r)) \quad (2.57)$$

where u_r is the dose level of coke oven main, and where $\beta_{11}^{B(a)P}$ and β_{11}^{coke} have been estimated from the initiator experiment.

Equation (2.18) shows that the probability to develop detectable papillomas depends on dose level through γ_1 and θ_1. The birth rate and death rate are represented as follow:

$$b_1(u_i) = b_{10} * exp(\delta_1 \log(1 + u_i))$$

$$d_1(u_i) = d_{10} * exp(\delta_1 \log(1 + u_i))$$

It follows that $\gamma_1(u_i) = b_1(u_i) - d_1(u_i)$, that is,

$$(b_{10} - d_{10}) * exp(\delta_1 \log(1 + u_i)) = \gamma_{10} * exp(\delta_1 \log(1 + u_i)) \quad (2.58)$$

and $\theta_1(u_i) = \frac{b_1(u_i)}{\gamma_1(u_i)} = \frac{b_{10}}{b_{10} - d_{10}}$, which is not a function of dose level.

Five parameters, ω_{10}, α_{10}^{coke}, γ_{10}, δ_1 and θ_1 can be estimated from the promotion experiment.

(c) Complete Experiment. In the complete carcinomas experiment, carcinomas are produced from two pathways: One is from normal cells directly, where $\alpha_{21}^{Initiator}$ has been estimated from initiation experiment; the other one is produced through I_1 cells, where all parameters related to I_1 cells are estimated from above two experiments. The rest of the parameters will be estimated from the complete carcinomas experiment; here α_{20}^{coke} in (2.52) is a parameter related to mutation rate from normal cells to I_2 cells, β_{10} and β_{11} the mutation rates from I_1 cells to I_2 cells, γ_{20}, and δ_2 the parameters which are related to proliferation rate, and θ_2.

Parameters Estimation

All of the above parameters are estimated by using procedures described in section 2.5. Prior information is collected to provide the lower bound and upper bound for each parameter. Table 5 gives the lower and upper bounds of parameters (prior information). Two deviance functions will be used: One is for the number of mice with papillomas or carcinomas, and the other is for the average number of papillomas or carcinomas per mouse. These deviances are minimized iteratively as described in the estimate procedure. Two approaches will be employed: One is the weighted deviance from different data using equal weight. The other is sequential starting with one type of data by assuming given initial values of the unknown parameters; the estimated parameters which minimize one deviance are taken as initial value to minimize the second deviance and repeat and continue until convergence. The estimated parameters and standard errors are also shown in table 5. We used bootstrap method proposed by Efron (1982) to obtain standard error for each parameter.

Table 5

Lower Bound, Upper Bound and Estimates of Parameters

Parameters	Estimate	Standard Error	Lower Bound	Upper Bound
α_{10}	201.83229819	26.42	100	1000
α_{11}	0.3374434	0.12	0.1	0.5
γ_{10}	0.06664222	0.0083	0.01	0.5
θ_1	3.38206134	0.22	1	10
δ_1	0.05147995	0.01	0.01	0.2
α_{20}	8.222657628	0.348	1	100
α_{21}	0.25125966	0.095	0.1	0.5
β_{10}	0.001593745	0.00034	1.0E-05	0.01
β_{11}	0.523923978	0.019	0.1	1
γ_{20}	0.059995162	0.0027	0.01	0.5
θ_2	4.937317031	0.894	1	10
δ_2	0.004955142	0.000397	0.001	0.1

Table 6

Generated Number of Mice with Papillomas

Data		Mice with Papillomas[a] (%)	Papillomas per mouse[a]	Mice with Cacinomas[a] (%)	Cacinomas per mouse[a]
Dose level	100	6	0.02	8	0.08
	500	16	0.44	20	0.22
	1000	58	1.09	43	0.45
	2000	80	2.74	87	0.88
	4000	100	3.07	92	1.02

[a] Data is collected at 34th week; [b] Data is collected at 52th week

Simulation Study

To assess the usefulness of our models and methods, in this section we generate some Monte Carlo data by assuming some parameters values. Then we use these generated data as the observations to estimate parameters. Complete carcinomas experiment contains both initiation and promotion procedures. The data we generated are the number of mice with papillomas and the average number of papillomas per mouse at t_1, the number of mice with carcinomas and the average number of carcinomas per mouse at t_2 ($t_2 > t_1$).

Generating Simulated Data

Table 7

Estimates of Parameters and Standard Error

Parameters	Original Parameters	Estimate	Standard Error
α_{10}	201	234.3	63.26
α_{11}	0.34	0.37	0.1
γ_{10}	0.07	0.062	0.0059
θ_1	3.38	3.88	0.199
δ_1	0.05	0.057	0.014
α_{20}	8.2	8.39	0.73
α_{21}	0.25	0.2	0.08
β_{10}	0.0002	0.00017	0.000056
β_{11}	0.227	0.336	0.018
γ_{20}	0.06	0.059	0.0036
θ_2	4.94	4.89	0.95
δ_2	0.005	0.0035	0.0005

Papillomas are developed as a result of single mutation, and the mutated cells are proliferated to a detectable size. Therefore, papillomas are generated only from single step of a single pathway.

The parameter values are given as following:

$\alpha_{10} = 201,\ \alpha_{11} = 0.34,\ \gamma_{10} = 0.07,\ \theta_1 = 3.38,\ \delta_1 = 0.05,\ N_p = 698$ (the number of cells in a detectable papillomas), and $\beta_{20} = 8.2$, $\beta_{21} = 0.25$, $\alpha_{10} = 0.0002$, $\alpha_{11} = 0.227$, $\gamma_{20} = 0.06$, $\theta_{20} = 4.94,\ \delta_2 = 0.005$, $Nc = 698$. The dose levels are 100, 500, 1000, 2000, and 4000. We collected papillomas data at the 34th week and collect carcinomas data at the 52th week for each dose level. The number of mice with papillomas is generated from a binomial distribution and is given in Table 6. The average number of papillomas per mouse is generated from a normal distribution with Poisson means and Poisson varainces. Similarly, the carcinomas data are generated.

Table 8

Sensitivity Test

Parameters	Original Parameters	Average Change of Responses (%)*
α_{10}	201	5.005
α_{11}	0.34	12.6
γ_{10}	0.07	2.02
θ_1	3.38	4.19E-07
δ_1	0.05	0.61
α_{20}	8.2	2.0
α_{21}	0.25	4.85
β_{10}	0.0002	5.0
β_{11}	0.227	8.2
γ_{20}	0.06	269.6
θ_2	4.94	78.9
δ_2	0.005	7.31

Parameter Estimation

We use the same estimation procedures as above to estimate the unknown parameters. The estimated values and standard errors are given in Table 7.

From results in Table 7, clearly, the estimates are very close to true values implying that the models and methods we proposed are quite reliable.

Sensitivity Test

To test sensitivity of the model to a parameter, in the simulation we increase the value of the parameter by 10% at a time and simulate responses (the number of mice with carcinomas) of the experiment. By comparing the responses with those from the original settings, we can assess the model's sensitivity to each parameter; in this way we can assess the reliability of the estimates. The result is given in Table 8.

Table 8 shows that the model is very sensitive to γ_{20} (the difference between birth rate and death rate) and θ_2. This implies the estimates of γ_{20} and θ_2 from the model are very reliable. On the other hand, it appears that the model is quite insensitive to θ_1; therefore, the response would not change significantly even if the value of θ_1 assumes any value in a relatively wide range. Thus it is practically impossible to estimate θ_1 accurately by using the model.

3. A NEW STOCHASTIC AND STATE SPACE MODEL OF HUMAN COLON CANCER: INCORPORATING MULTIPLE PATHWAYS

Introduction

In the past 15 years, molecular biologists and geneticists have revealed the basic molecular and genetic mechanisms for human colon cancer. These mechanisms have been linked to two avenues: The chromosomal instability (CIN) involving chromosomal aberrations and loss of heterozygosity (LOH), and the micro-satellite instability (MSI) involving mis-match repair genes and the creation of mutator phenotype (Chapelle, 2004; Fodde et al., 2001; Fodde et al., 2001; Green & Kaplan, 2003; Hawkins & Ward, 2001; Hisamuddin & Yang, 2004; Sparks, Morin, Vogelstein, & Kinzler, 1998; Peltonmaki, 2001). The pathway of the CIN avenue (also referred to as LOH pathway) involves inactivation through genetic and/or epigenetic mechanisms, or loss, or mutation of the suppressor APC gene in chromosome 5q (about 85% of all human colon cancers) whereas the pathway of the MSI avenue involves mutation or epigenetic inactivation of the mis-match repair suppressor genes (about 15% of all colon cancers). This leads to multiple pathways for the generation of human colon cancer tumors with each pathway following a stochastic multi-stage model and with intermediate transformed cells subjecting to stochastic proliferation (birth) and differentiation (death). The goal of this paper is to develop a stochastic model for human colon cancer to incorporate these biological information and pathways. This paper is an extension of Tan and Zhang (2008), Little and Wright (2003), and Little, Vineis and Li (2008). We note that besides the multiple pathways considered above, Little and Wright (2003), Little (2008) and Little et al. (2008) have also included mixture type of multiple pathways; however,

because the mutation rates are very small, the chance of mixture type of pathways will be extremely small in which case the Little model is equivalent to the model in Section 3.3.

For developing biologically supported stochastic model of carcinogenesis, in Section 3.2 we present the most recent cancer biology of human colon cancer. Using results from Section 3.2, we develop in Section 3.3 a stochastic model for carcinogenesis of human colon cancer involving multiple pathways. In Section 3.4 we derive a statistical model for cancer incidence data of human colon cancer. By combining models from Sections 3.3 and 3.4, in Section 3.5 we develop a state space model for human colon cancer. In Section 3.6, by using the state space model in Section 3.5, we develop a generalized Bayesian inference procedure to estimate unknown parameters and to predict state variables. To illustrate the applications of the model and methods, in Section 3.7 we apply the model and methods to the colon cancer incidence data from SEER. Finally in Section 3.8, we discuss the usefulness of the model and methods and provide some conclusions from model and results.

A Brief Summary of Colon Cancer Biology

As discussed in the introduction, genetic studies have indicated that there are two major avenues by means of which human colon cancer is derived: The Chromosomal Instability (CIN) and the Micro-Satellite Instability (MSI). The first avenue is associated with the LOH pathway involving the APC gene in chromosome 5q and the latter associated with the micro-satellite pathway involving mis-match repair genes. The most important oncogene is the β-Catenin gene in chromosome 3p22.

The CIN (LOH) Pathway of Human Colon Cancer (The APC- β– catenin −Tcf − myc pathway)

The CIN pathway involves loss or inactivation of the tumor suppressor genes - the APC gene in chromosome 5q, the Smad-4 gene in chromosome 18q and the p53 gene in chromosome 17p; see **Remark 3.1**. This pathway accounts for about 85% of all colon cancers. It has been referred to as the LOH pathway because it is characterized by aneuploidy /or loss of chromosome segments (chromosomal instability); see **Remark 3.2**. This pathway has also been referred to as APC− *β -catenin − Tcf − myc* pathway because it involves the destruction complex GSK-3*β* − Axin− APC which phosphorylates the *β*-Catenin protein leading to its degradation; when both copies of the APC gene are inactivated or mutated, the destruction complex is then inactive leading to accumulation of free *β*-Catenin proteins in the cytoplasm which move to the nucleus to complex with Tcf/Lef transcription factor to activate and transcript oncogenes myc, cyclin D and CD44. (Free *β* -Catenin protein in the cytoplasm also binds with E-cadherin and *β* -Catenin to disrupt the gap junction between cells, leading to migration and metastasis of cancer tumors.)

Morphological studies have indicated that inactivation, or loss or mutation of APC creates dysplastic aberrant crypt foci (ACF) which grow into dysplastic adenomas. These adenomas grow to a maximum size of about 10 mm^3; further growth and malignancy require the abrogation of differentiation, cell cycle inhibition and apoptosis which are facilitated by the inactivation, or mutation or loss of Smad-4 gene in 18q and the p53 gene in 17p. The mutation or activation of the oncogene H-ras in chromosome 11p and /or mutation and/or activation of the oncogene src in chromosome 20q would speed up

these transitions by promoting the proliferation rates of the respective intermediate initiated cells (Jessup, Garllic, & Liu, 2002). This pathway is represented schematically by Figure 8. The model in Figure 8 is a 6-stage model. However, because of the haplo-insufficiency of the Smad4 gene (see Alberici, Jagmohan-Changur, & De Pater, 2006) and the haplo-insufficiency of the p53 gene (Lynch & Milner, 2006), one may reduce this 6-stage model into a 4-stage model by combining the third stage and the fourth stage into one stage and by combining the fifth stage and the sixth stage into one stage. This may help explain why for single pathway models, the 4-stage model fits the human colon cancer better than other single pathway multi-stage models (Leubeck & Moolgaokar, 2002). Recent biological studies by Green and Kaplan (2003) and others have also shown that the inactivation or deletion or mutation of one copy of the APC gene in chromosome 5 can cause defects in microtubule plus-end attachment during mitosis dominantly, leading to aneuploidy and chromosome instability. This would speed up the mutation or inactivation of the second copy of the APC gene and increase fitness of the APC-carrying cells in the micro-evolution process of cancer progression. This could also help explain why the APC LOH pathway is more frequent than other pathways.

Remark 3.1. As observed by Sparks et al. (1998), instead of the APC gene, this pathway can also be initiated mutation of the oncogene β -catenin gene; however, the proportion of human colon cancer due to mutation of β -catenin is very small (less than 1%) as compared to the APC gene, due presumably to the contribution of the APC on chromosome instability (Green & Kaplan, 2003). Similarly, the destruction complex can become inactive either by the inhibition of GSK-3 β through the Wnt signaling pathway (see Sparks et al. 1998) or the inactivation or mutation of the Axin protein, leading to

accumulation of the β - Catenin proteins in the cytoplasm; but the proportion of colon cancer caused by inhibition of GSK-3 β is also very small as compared to the colon cancer cases caused by the CIN and the MSI pathways.

Remark 3.2. The APC gene in chromosome 5q acts both as a tumor suppressor gene and an oncogene in initiating and promoting colon carcinogenesis. As an oncogene, the APC gene acts dominantly in regulating microtubule plus-end attachment during mitosis (Green & Kaplan, 2003). Thus, the inactivation or deletion or mutation of one copy of the APC gene in chromosome 5 can cause defects in microtubule plus-end attachment during

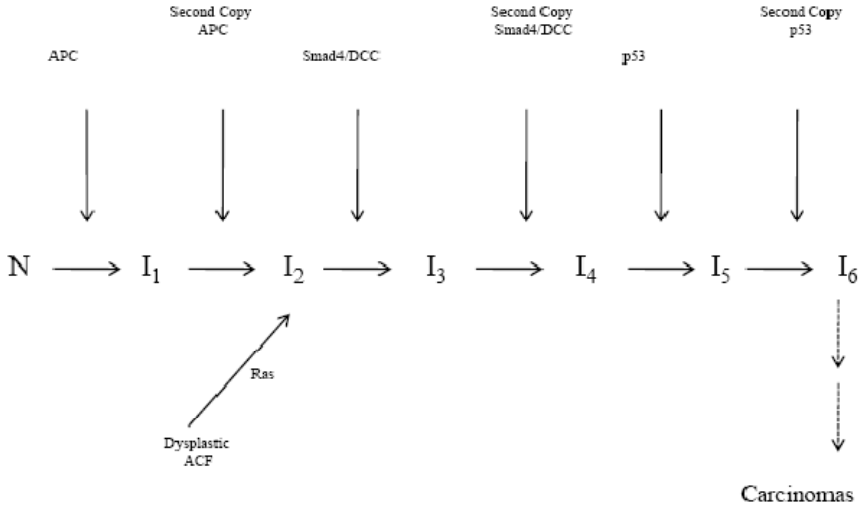


Figure 8. The CIN Pathway of Human Colon Cancer

mitosis, leading to aneuploidy and chromosome instability. This would speed up the mutation or inactivation of the second copy of the APC gene and increase fitness of the APC-carrying cells in the micro-evolution process of cancer progression. This could also help explain why the APC LOH pathway is more frequent than other pathways.

The MSI (Micro-Satellite Instability) Pathway of Human Colon Cancer

This pathway accounts for about 15% of all colon cancers and appears mostly in the right colon. It has been referred to as the MSI pathway or the mutator phenotype pathway because it is initiated by the mutations or epigenetic methylation of the mis-match repair genes (mostly hMLH1 in chromosome 3p21 and hMSH2 in chromosome 2p16) creating a mutator phenotype to significantly increase the mutation rate of many critical genes 10 to 1000 times. Normally these critical genes are TGF-*β RII*, Bax (The X protein of bcl-2 gene), IGF2R, or CDX-2. The mis-match repair genes are hMLH1, hMSH2, hPMS1, hPMS2, hMSH6 and hMSH3; mostly hMLH1 (50%) and hMSH2 (40%). This pathway is represented schematically by Figure 9. As in the LOH pathway, assuming haplo-insufficiency of tumor suppressor genes, one may approximate this pathway by a 5-stage model.

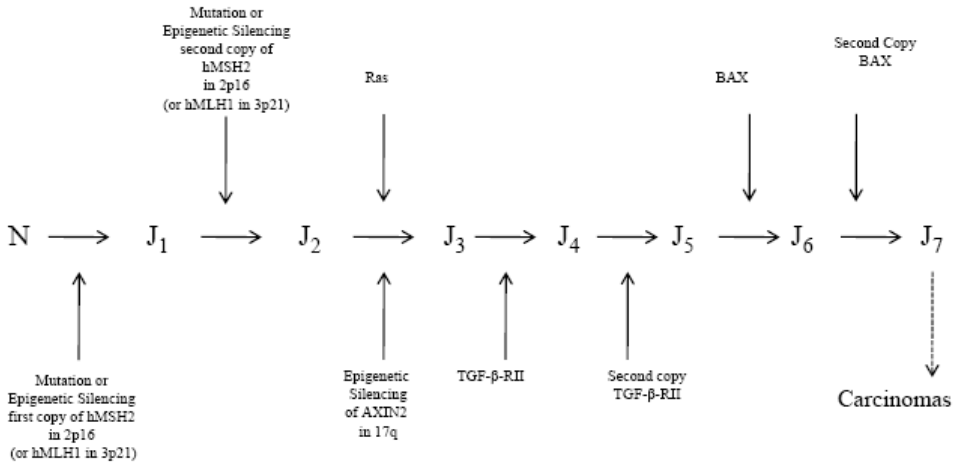


Figure 9. The MSI Pathway of Human Colon Cancer

Morphologically, mutation or methylation silencing of the MMR gene hMLH1 or hMSH2 generates hyperplastic polyps which lead to the generation of serrated adenomas.

These adenomas develop into carcinomas after the inactivation, or loss or mutations of the TGF-β RII gene and the Bax gene, thus abrogating differentiation and apoptosis. (Bax is an anti-apoptosis gene.) In what follows, we let N denote the normal stem cells, Ji the i−th stage cells in the MSI pathways. Then for sporadic MSI, the model is $N \rightarrow J_1 \rightarrow J_2 \rightarrow J_3 \rightarrow J_4 \rightarrow J_5 \rightarrow$ cancer tumor.

The Major Signaling Pathways for Human Colon Cancer

Recent biological studies (Baylin & Ohm, 2006; Koinuma et al., 2006) have shown that both the CIN and the MSI pathways involve the Wnt signaling pathway and the destruction complex (this complex is a downstream of the Wnt signaling pathway), the TGF-β inhibiting signaling pathway and the p53-Bax apoptosis signaling pathway, but different genes in the CIN and MSI pathways are affected in these signaling processes. In the CIN pathway, the affected gene is the APC gene in the Wnt signaling, the Smad4 in the TGF- β signaling and the p53 gene in the p53-Bax signaling; on the other hand, in the MSI pathway, the affected gene is the Axin 2 gene in the Wnt signaling, the TGF- β - Receptor II in the TGF- β signaling and the Bax gene in the p53-Bax signaling. Because the probability of point mutation or genetic changes of genes are in general very small compared to epigenetic changes, one may speculate that colon cancer may actually be initiated by some epigenetic mechanisms (Baylin & Ohm, 2006; Breivik & Gaudernack, 1999; Jones & Baylin, 2002). In fact, Breivik and Gaudernack (1999) showed that in human colon cancer, either methylating carcinogens or hyper-methylation at *C*p*G* islands would lead to G/T mismatch which in turn leads to Mis-match Repair (MMR) gene deficiency or epigenetic silencing of the MMR genes and hence MSI (Micro-satellite Instability); alternatively, either hypo-methylation, or bulky-adduct forming (BAF)

carcinogens such as alkylating agents, UV radiation and oxygen species promote chromosomal rearrangement via activation of mitotic check points (MCP), thus promoting CIN (Chromosomal Instability). A recent review by Baylin & Ohm (2006) have demonstrated that epigenetic events may lead to LOH and mutations of many genes which may further underline the importance of epigenetic mechanisms in cancer initiation and progression.

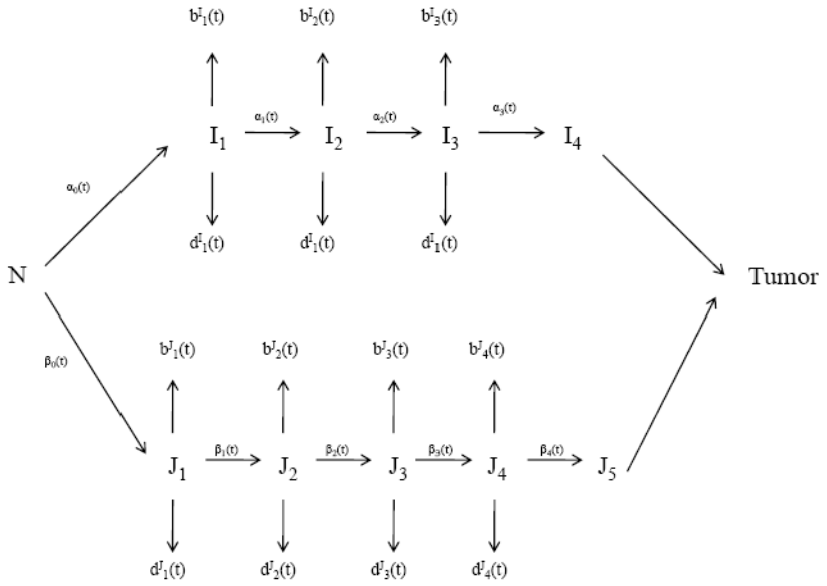


Figure 10. The Multiple Pathways of Human Colon Cancer

Based on the above biological studies, in this chapter we thus postulate that the incidence data of human colon cancer are described and generated by a multi-stage model involving 2 pathways as defined above. In this chapter, because of haploid-insufficiency of the tumor suppressor genes $\{Smad4, p53, Axin, Bax, TGF-\beta-Receptor\ II\}$, the number of stages for the CIN pathway and MSI are assumed as 4 and 5 respectively.

Methods

Stochastic Multi-Stage Model of Carcinogenesis for Human Colon Cancer Involving Multiple Pathways

From results of Section 3.2, it follows that the stochastic multi-stage model for human colon cancer can be represented schematically by Figure 10.

In Figure 10, the model assumes that cancer tumors are generated by two pathways with pathway 1 as a k_1-stage multi-stage model involving I_l ($l = 1, \ldots, k_1$) cells and with pathway 2 as a k_2-stage multi-stage model involving J_r ($r = 1, \ldots, k_2$) cells. (For human

Table 9

Transition Rates and Transition Probabilities for Human Colon Carcinogenesis

Transition	Transition Probability
$1N \rightarrow 1N, 1I_1$	$\alpha_0(t)\Delta t$
$1N \rightarrow 1N, 1J_1$	$\beta_0(t)\Delta t$
$1\ I_l \rightarrow 2I_l$	$b_l^{(I)}(t)\Delta t$
$1\ I_l \rightarrow$ Death	$d_l^{(I)}(t)\Delta t$
$1\ I_l \rightarrow 1\ I_l, 1\ I_{l+1}$	$\alpha_l(t)\Delta t$
$l = 1,2\ldots, k_1\text{-}1$	
$1\ J_r \rightarrow 2J_r$	$b_r^{(J)}(t)\Delta t$
$1\ J_r \rightarrow$ Death	$d_r^{(J)}(t)\Delta t$
$1\ J_r \rightarrow 1\ J_r, 1\ J_{r+1}$	$B_r(t)\Delta t$
$r = 1,2\ldots, k_2\text{-}1$	

colon cancer, $k_1 = 4$, $k_2 = 5$.) The state variables are then $\tilde{X}(t) = \{I_l(t),\ l = 1, \ldots, k_1 - 1,\ J_r(t),\ r = 1, \ldots, k_2 - 1\}$ and $T(t)$, where $T(t)$ denotes the number of cancer tumors at time t and where $I_l(t)$ ($J_r(t)$) denote the number of the I_l (J_r) initiated cells for $\{l = 1, \ldots, k_1 - 1$ ($r = 1, \ldots, k_2 - 1$)$\}$ respectively. Notice that because cell proliferation, cell differentiation and apoptosis, mutation or genetic changes all occur during cell division and cell division cycle, and because $\tilde{X}(t)$ $(t + \Delta t)$ develop from $\tilde{X}(t)$ through cell divisions during $(t,\ t + \Delta t]$, one may practically assume that $(X(t),\ t \geq t_0)$ is a Markov process with continuous time, where $t0$ represents time at birth; one the other hand, $T(t + \Delta t)$ may derive from I_{k1} (J_{k2}) cells before time t, $T(t)$ is in general not Markov (Fakir & Tan, 2009; Yakovlev & Tsodikov, 1996). If one assumes that the I_{k1} and J_{k2} cells grow instantaneously into cancer tumors as soon as they are generated, then one may also assume the $T(t)$ as Markov. In this case, as illustrated in Tan (1991), one may use standard Markov theory to derive the probability generating function (pgf) of the probabilities of these variables and hence the probability distribution of these variables. Let $\psi(x_l, l = 1, \ldots, k_1 - 1, y_r, r = 1, \ldots, k_2 - 1, z; t_0, t) = \psi(\tilde{x}, \tilde{y}, z;\ t_0, t)$ denote the pgf of $\{\tilde{X}(t), T(t)\}$. Let $\{\alpha_l(t), \beta r\ (t), b_l^{(I)}, d_l^{(I)}, b_r^{(J)}, d_r^{(J)}\}$ denote the mutation rates, the birth rates and the death rates of $\{I_l, J_r\}$ cells as given in Table 9 respectively. If $T(t)$ is Markov, then by using the method of Kolmogorov forward equation of these variables (Tan, 1991), it can readily be shown that $\psi(\tilde{x}, \tilde{y}, z;\ t_0, t)$ satisfies the following partial differential equation (pde):

$$\frac{\partial}{\partial t}\psi(\tilde{x}, \tilde{y}, z;\ t_0, t) =$$

$$\left[\lambda_I(t)(x_1 - 1) + \lambda_J(t)(y_1 - 1)\right]\psi(\tilde{x}, \tilde{y}, z;\ t_0, t) +$$

$$\sum_{l=1}^{k_1-1} g_1(x_l, x_{l+1}; t)\frac{\partial}{\partial x_l}\psi(\tilde{x}, \tilde{y}, z;\ t_0, t) + \sum_{r=1}^{k_2-1} g_1(y_r, y_{r+1}; t)\frac{\partial}{\partial y_r}\psi(\tilde{x}, \tilde{y}, z;\ t_0, t)$$

(3.1)

where $\lambda_l(t) = N(t)\alpha_0(t)$, $\lambda_J(t) = N(t)\beta_0(t)$,

$$g_1(x_l, x_{l+1}; t) = x_l(x_l - 1)b_l^{(I)}(t) - (x_l - 1)d_l^{(I)}(t) + x_l(x_{l+1} - 1)\alpha_l(t),$$

$$g_2(y_r, y_{r+1}; t) = y_r(y_r - 1)b_r^{(J)}(t) - (y_r - 1)d_r^{(J)}(t) + y_r(y_{r+1} - 1)\beta_r(t) \qquad (3.2)$$

and the initial condition is $\psi(\tilde{x}, \tilde{y}, z;\ t_0, t_0) = 1$ given normal individuals at risk at time t_0.

The above pde is in general very difficult to solve; further, even if the solution of this equation can be derived, the results are very difficult to apply to estimate the unknown parameters and to predict future cancer cases. Most importantly, $T(t)$ may not be Markov so that this theory is not applicable (Fakir, 2009; Yakovlev & Tsodikov, 1996). In this chapter, we will thus propose an alternative approach through stochastic equations. It can easily be shown through the method of pgf that if $T(t)$ is Markov, then the stochastic equation method is equivalent to the method of Markov theory; as we shall see, however, the stochastic equation method is more powerful and does not need to assume Markov for $T(t)$.

The Stochastic Equation for State Variables

To derive stochastic equations for the state variables, let $B_l^{(I)}(t)(B_r^{(J)}(t))$ be the number of births of the I_l (J_r) initiated cells during $(t, t + \Delta t]$ $\{l = 1, \ldots, k_1 - 1$ $(r = 1, \ldots, k_2 - 1)\}$, $D_l^{(I)}(t)(D_r^{(J)}(t))$ the number of deaths of the I_l (J_r) initiated cells during $(t, t + \Delta t]$ $\{l = 1, \ldots, k_1 - 1$ $(r = 1, \ldots, k_2 - 1)\}$ and $M_l^{(I)}(t)(M_r^{(J)}(t))$ the number of mutation ($I_l \rightarrow I_{l+1}$) ($J_r \rightarrow J_{r+1}$) of I_l (J_r) cells during $(t, t + \Delta t]$ $\{l = 1, \ldots, k_1 - 1$ $(r = 1, \ldots, k_2 - 1)\}$. Also let $M_0^{(I)}(t)(M_0^{(J)}(t))$ be the number of mutation of $N \rightarrow I_1$ ($N \rightarrow J_1$) during $(t, t + \Delta t]$.

Taking into account of all possible input and output of relevant cells, we have the following stochastic equations for the state variables:

$$I_l(t+\Delta t) = I_l(t) + M_{l-1}^{(I)}(t) + B_l^{(I)}(t) - D_l^{(I)}(t), l = 1, \ldots, k_1 - 1 \quad (3.3)$$

$$J_r(t+\Delta t) = J_r(t) + M_{r-1}^{(J)}(t) + B_r^{(J)}(t) - D_r^{(J)}(t), r = 1, \ldots, k_2 - 1 \quad (3.4)$$

Because the transition variables $\{ M_l^{(I)}(t),\ M_r^{(J)}(t), B_l^{(I)}(t), D_l^{(I)}(t), B_r^{(J)}(t), D_r^{(J)}(t)\}$ are random variables, the above equations are stochastic equations. With the transition rates as given in Table 9, it can readily be shown that to the order of $o(\Delta t)$, the conditional probability distributions of $M_0^{(I)}(t)$ and $M_0^{(J)}(t)$ given $N(t)$ are Poisson with means $\lambda_I(t)\Delta t$ and $\lambda_J(t)\Delta t$ respectively whereas the conditional probability distributions of the numbers of births and deaths given the staging variables (i.e., the $I_l(t)$ and $J_r(t)$) follow multinomial distributions independently. That is,

$$M_0^{(I)}(t) | N(t) \sim \text{Poisson}\{\lambda_I(t)\Delta t\}, \text{ independent of } M_0^{(J)}(t) \quad (3.5)$$

$$M_0^{(J)}(t) | N(t) \sim \text{Poisson}\{\lambda_J(t)\Delta t\}, \text{ independent of } M_0^{(I)}(t) \quad (3.6)$$

for $l = 1, 2, \ldots, k_1\text{-}1$,

$$\{B_l^{(I)}(t), D_l^{(I)}(t) | N(t) \sim \text{Multinomial}\{I_l(t); b_l^{(I)}(t)\Delta t, d_l^{(I)}(t)\Delta t\} \quad (3.7)$$

for $r = 1, 2, \ldots, k_2\text{-}1$,

$$\{B_r^{(J)}(t), D_r^{(J)}(t) | N(t) \sim \text{Multinomial}\{J_r(t); b_r^{(J)}(t)\Delta t, d_r^{(J)}(t)\Delta t\} \quad (3.8)$$

where $\lambda_I(t) = N(t)\alpha_0(t)$, $\lambda_J(t) = N(t)\beta_0(t)$.

Because the number of mutations of the I_l cells would not affect the size of the I_l population but only increase the number of I_{l+1} cells and because the mutation rate of I_l cells is very small ($10^{-8} \sim 10^{-5}$), it can readily be shown that to the order of $o(\Delta t)$, the

conditional probability distribution of $M_l^{(I)}(t)$ given $I_l(t)$ I_l cells at time t is Poisson with mean $I_l(t)\alpha_l(t)\Delta t$ independently of $\{B_l^{(I)}(t), D_l^{(I)}(t)\}$ and other transition variables. That is,

$$M_l^{(I)}(t)|\, N(t) \sim \text{Poisson}\,\{I_l(t)\alpha_l(t)\Delta t\}, l = 1, 2, \dots, k_1 - 1 \tag{3.9}$$

independently of $\{B_l^{(I)}(t), D_l^{(I)}(t)\}$ and other transition variables.

Similarly, we have that to the order of $o(\Delta t)$,

$$M_r^{(J)}(t)|\, N(t) \sim \text{Poisson}\left\{J_r(t)\beta_r(t)\Delta t\right\}, r = 1, 2, \dots, k_2 - 1 \tag{3.10}$$

independently of $\{B_r^{(J)}(t), D_r^{(J)}(t)\}$ and other transition variables.

Using the probability distributions given by equations (3.5)-(3.10) and by subtracting from the transition variables the conditional expected values respectively, we have the following stochastic differential equations for the staging state variables:

$$\begin{aligned} dI_l(t) &= I_l(t+\Delta t) - I_l(t) = M_{l-1}^{(I)}(t) + B_l^{(I)}(t) - D_l^{(I)}(t) \\ &= \left\{I_{l-1}(t)\alpha_{l-1}(t) + I_l(t)\gamma_l^{(I)}(t)\right\}\Delta t + e_l^{(I)}(t)\Delta t \end{aligned}$$

$$l = 1, \dots, k_1 - 1 \tag{3.11}$$

$$\begin{aligned} dJ_r(t) &= J_r(t+\Delta t) - J_r(t) = M_{r-1}^{(J)}(t) + B_r^{(J)}(t) - D_r^{(J)}(t) \\ &= \left\{J_{r-1}(t)\beta_{r-1}(t) + J_r(t)\gamma_r^{(J)}(t)\right\}\Delta t + e_r^{(J)}(t)\Delta t \end{aligned}$$

$$r = 1, \dots, k_2 - 1 \tag{3.12}$$

where $\gamma_l^{(I)}(t) = b_l^{(I)}(t) - d_l^{(I)}(t)$, $\gamma_r^{(J)}(t) = b_r^{(J)}(t) - d_r^{(J)}(t)$

In the above equations, the random noises $\{e_l^{(I)}(t)\Delta t, e_r^{(J)}(t)\Delta t\}$ are derived by subtracting the conditional expected numbers from the random transition variables respectively. Obviously, these random noises are linear combinations of Poisson and

multinomial random variables. These random noises have expected value zero and are un-correlated with the state variables $\{I_l(t),\ l = 1, \ldots, k_1 - 1,\ J_r(t),\ r = 1, \ldots, k_2 - 1\}$. It can also be shown that to the order of $o(\Delta t)$, these random noises are uncorrelated with one another and have variances given by:

$$\mathrm{Var}\left\{e_l^{(I)}(t)\Delta t\right\} = EI_{l-1}(t)\alpha_{l-1}(t)\Delta t + EI_l(t)\left[b_l^{(I)}(t) + d_l^{(I)}(t)\right]\Delta t + o(\Delta t)$$

for $l = 1, \ldots, k_1 - 1$, (3.13)

$$\mathrm{Var}\left\{e_r^{(J)}(t)\Delta t\right\} = EJ_{r-1}(t)\beta_{r-1}(t)\Delta t + EJ_r(t)\left[b_r^{(J)}(t) + d_r^{(J)}(t)\right]\Delta t + o(\Delta t)$$

for $r = 1, \ldots, k_r - 1$, (3.14)

The Expected Numbers

Let $u_I(l, t) = E[I_l(t)]$ and $u_J(r, t) = E[J_r(t)]$ denote the expected numbers of $I_l(t)$ and $J_r(t)$ respectively and write $u_I(0,t) = u_J(0,t) = N(t)$. Using equations (3.11)-(3.12), we have the following differential equations for these expected numbers:

$$du_I(l,t) = u_I(l,t)\gamma_l^{(I)}(t) + u_I(l-1,t)\alpha_{l-1}(t) \quad (3.15)$$

$l = 1, \ldots, k_1 - 1$,

$$du_J(r,t) = u_J(r,t)\gamma_r^{(J)}(t) + u_J(r-1,t)\beta_{r-1}(t) \quad (3.16)$$

$r = 1, \ldots, k_2 - 1$.

The solutions of the above equations are:

$$u_I(1,t) = \int_{t_0}^{t} \lambda_I(x)\, e^{\int_x^t \lambda_1^{(I)}(z)dz} dx$$

$$u_J(1,t) = \int_{t_0}^{t} \lambda_J(x)\, e^{\int_x^t \lambda_1^{(J)}(z)dz} dx$$

$$u_I(l,t) = \int_{t_0}^{t} u_I(l-1,x)\, e^{\int_x^t \lambda_l^{(I)}(z)dz} dx, \text{ for } l = 2, \ldots, k_1 - 1,$$

$$u_J(r,t) = \int_{t_0}^{t} u_J(r-1,x)\, e^{\int_x^t \lambda_r^{(J)}(z)dz} dx, \quad \text{for } r = 2, \ldots, k_2 - 1.$$

If the model is time homogeneous, then $\lambda_I(t) = \lambda_I$, $\lambda_J(t) = \lambda_J$, $\alpha_l(t) = \alpha_l$ and $\gamma_l^{(I)}(t) = \gamma_l^{(I)}$ for $l = 1, \ldots, k_1 - 1$ and $\beta_r(t) = \beta_r$ and $\gamma_r^{(J)}(t) = \gamma_r^{(J)}$ for $r = 1, \ldots, k_2 - 1$. If the proliferation rates are not zero and if $\gamma_l^{(I)} \neq \gamma_u^{(I)} \neq \gamma_r^{(J)} \neq \gamma_v^{(J)}$ for all $l \neq u$ and $r \neq v$, then the above solutions reduce to:

$$u_I(1,t) = \frac{\lambda_I}{\gamma_1^{(I)}}[e^{\gamma_1^{(I)}} - 1], \qquad u_J(1,t) = \frac{\lambda_J}{\gamma_1^{(J)}}[e^{\gamma_1^{(J)}} - 1]$$

$$u_I(l,t) = \sum_{u=1}^{l} A_l(u)\, e^{\gamma_u^{(I)} t} \quad \text{for } l = 1, \ldots, k_1 - 1;$$

$$u_J(r,t) = \sum_{u=1}^{r} B_r(u)\, e^{\gamma_u^{(J)} t} \quad \text{for } r = 1, \ldots, k_2 - 1$$

where $A_l(u) = \lambda_I \prod_{u=1}^{l-1}\{\prod_{\substack{v=1 \\ v\neq u}}^{l}(\gamma_u^{(I)} - \gamma_v^{(I)})\}^{-1}$, $B_r(u) = \lambda_J \prod_{u=1}^{r-1}\{\prod_{\substack{v=1 \\ v\neq u}}^{r}(\gamma_u^{(J)} - \gamma_v^{(J)})\}^{-1}$

The Probability Distribution of State Variables and Transition Variables

Although $T(t)$ is not Markov, the random vector $\{\tilde{X}(t),\ t \geq t_0\}$ is Markov with continuous time. To derive the transition probability of this process, denote by $f(x, y : N, p_1, p_2)$ the density at (x, y) of the multinomial distribution $ML(N; p_1, p_2)$ with parameters $(N; p_1, p_2)$ and $h(x; \lambda)$ the density at x of the Poisson distribution with mean λ. Then, using the probability distributions given by equations (3.5)-(3.10), the transition probability of this Markov process is, to order of $o(\Delta t)$:

$$P\{\tilde{X}(t+\Delta t) \mid \tilde{X}(t)\} = \prod_{u=1}^{k_1-1}\{\sum_{l_u=0}^{I_u(t)} \sum_{i_u=0}^{I_u(t)-l_u} h[a(l_u, i_u; t); I_{r-1}(t)\alpha_{u-1}(t)\Delta t] \times$$

$$f[l_u, i_u; I_u(t), b_u^{(I)}(t)\Delta t, d_u^{(I)}(t)\Delta t]\} \times$$

$$\prod_{v=1}^{k_2-1}\{\sum_{m_v=0}^{J_v(t)} \sum_{j_v=0}^{J_v(t)-m_v} h[b(m_v, J_v; t); J_{v-1}(t)\beta_{v-1}(t)\Delta t] \times$$

$$f[m_v, j_v; J_v(t), b_v^{(J)}(t)\Delta t, d_v^{(J)}(t)\Delta t]\}$$

where $I_0(t) = J_0(t) = N(t)$, $a(l_u, i_u; t) = I_u(t + \Delta t) - I_u(t) - l_u + i_u$, $u = 1, \ldots, k_1 - 1$ and $b(m_v, j_v; t) = J_v(t + \Delta t) - J_v(t) - m_v + j_v$, $v = 1, \ldots, k_2 - 1$.

The above transition probability and hence the probability distribution of $\tilde{X}(t)$ is too complicated to be f much use. For implementing the Gibbs sampling procedure to estimate parameters and to predict state variables, we use data augmentation method to expand the model. Thus, we define the augmented variables $\tilde{U}(t) = \{B_l^{(I)}(t), D_l^{(I)}(t), l = 1, \ldots, k_1 - 1, B_r^{(J)}(t), D_r^{(J)}(t), r = 1, \ldots, k_2 - 1\}$. (In what follows we will refer these variables as the transition variables, unless otherwise stated.) Put $\tilde{Z}(t) = \{\tilde{X}(t)', \tilde{U}(t - \Delta t)'\}'$. Then $\{\tilde{Z}(t), t \geq t_0\}$ is Markov with continuous time. Using the probability distributions of the transition random variables given by equations (3.5)-(3.10), the transition probability is $P\{\tilde{Z}(t + \Delta t)|\tilde{Z}(t)\}$ is:

$$P\{\tilde{Z}(t + \Delta t)\big|\tilde{Z}(t)\} = P\{\tilde{X}(t + \Delta t)\big|\tilde{X}(t), \tilde{U}(t)\} \times P\{\tilde{U}(t)|\tilde{X}(t)\} \tag{3.17}$$

where

$$P\{\tilde{U}(t)\big|\tilde{X}(t)\} =$$

$$\prod_{l=1}^{k_1-1} f\{B_l^{(I)}(t), D_l^{(I)}(t); I_l(t), b_l^{(I)}(t)\Delta t, d_l^{(I)}(t)\Delta t\} \times$$

$$\prod_{r=1}^{k_2-1} f\{B_r^{(J)}(t), D_r^{(J)}(t); J_r(t), b_r^{(J)}(t)\Delta t, d_r^{(J)}(t)\Delta t\} \tag{3.18}$$

and

$$P\{\tilde{X}(t + \Delta t)\big|\tilde{X}(t), \tilde{U}(t)\} =$$

$$\prod_{l=1}^{k_1-1} h\{u_I(l, t); I_l(t)\alpha_l(t)\Delta t\} \times \prod_{r=1}^{k_2-1} h\{u_J(r, t); J_r(t)\beta_r(t)\Delta t\} \tag{3.19}$$

where $u_I(l, t) = I_l(t + \Delta t) - I_l(t) - B_l^{(I)}(t) + D_l^{(I)}(t)$ for $l = 1, \ldots, k_1 - 1$ and

$u_J(r, t) = J_r(t + \Delta t) - J_r(t) - B_r^{(J)}(t) + D_r^{(J)}(t)$ for $r = 1, \ldots, k_2 - 1$.

The probability distribution given by equation (3.17) will be used to derive estimates and predicted numbers of state variables. This is discussed in Section 3.6.

A Statistical Model and the Probability Distribution of the Number of Detectable Tumors

The data available for modeling carcinogenesis are usually cancer incidence over different time periods. For example, the SEER data of NCI/NIH for human cancers are given by $\{(y_j, n_j), j = 1, \ldots, n\}$, where y_j is the observed number of cancer cases during the $j-$th age group and n_j is the number of normal people who are at risk for cancer and from whom y_j of them have developed cancer during the age group. Given in Table 10 are the SEER data for human colon cancer adjusted for genetic cancer cases.

The Probability Distribution of the Number of Detectable Tumors for Colon Cancer

To derive the probability distribution of time to tumors, one needs the probability distribution of $T(t)$. For deriving this probability distribution, we observe that malignant cancer tumors arise by clonal expansion from primary I_{k1} cells and primary J_{k2} cells, where primary I_{k1} cells are I_{k1} cells derived from $I_{k1}-1$ cells by mutation of $I_{k1}-1$ cells and primary J_{k2} cells are J_{k2} cells derived from $J_{k2}-1$ cells by mutation of $J_{k2}-1$ cells.

Let $P_T^{(I)}(s,t)$ ($P_T^{(J)}(s,t)$) be the probability that a primary I_{k1} (J_{k2}) cancer cell at time s develops into a detectable cancer tumor at time t. Let $T_i(t)$ be the number of cancer tumors derived from the $i-$th pathway. Then, to order of $o(\Delta t)$, the conditional probability distribution of $T_1(t)$ given $\{I_{k1}-1(s), s \leq t\}$ is Poisson with mean $\omega_1(t)$ independently of $T_2(t)$, where

$\omega_1(t) = \int_{t_0}^{t} I_{k_1-1}(x)\alpha_{k_1-1}(x)P_T^{(I)}(x,t)dx$. Similarly, to order of $o(\Delta t)$, the conditional probability distribution of $T_2(t)$ given $\{J_{k2}-1(s),\ s \le t\}$ is Poisson with mean $\omega_2(t)$ independently of $T_1(t)$, where

$$\omega_2(t) = \int_{t_0}^{t} J_{k_2-1}(x)\beta_{k_2-1}(x)P_T^{(J)}(x,t)dx.$$

Let $Q_i(j)$ ($i = 1, 2$) be defined by:

$$Q_i(j) = E\left\{e^{-\omega_i(t_{j-1})} - e^{-\omega_i(t_j)}\right\} = E\left\{e^{-\omega_i(t_{j-1})}[1 - e^{R_i(t_{j-1},t_j)}]\right\}$$

where $R_i(t_{j-1},t_j)=\omega_i(t_{j-1})-\omega_i(t_j)$.

Then $Q_i(j)$ is the probability that cancer tumors would develop during the $j-$th age group by the $i-$th pathway. Since cancer tumors develop if and only if at least one of the two pathways yield cancer tumors, the probability that each normal person at time t_0 will develop cancer tumors during $(t_{j-1}, t_j]$ is given by $Q_T(j)$, where

$$Q_T(j) = 1-[1 - Q_1(j)][1 - Q_2(j)]= Q_1(j) + Q_2(j) - Q_1(j)Q_2(j).$$

For practical applications, we observe that to order of $o(\alpha_{k_1-1}(t))$ and $o(\beta_{k_2-1}(t))$ respectively, the $\omega_i(t)$ in $Q_i(j)$ are approximated by

$$\omega_1(t) \sim \int_{t_0}^{t} E\left[I_{k_1-1}(s)\right]\alpha_{k_1-1}(s)P_T^{(I)}(s,t)ds,\ \ \omega_2(t) \sim \int_{t_0}^{t} E\left[J_{k_2-1}(s)\right]\beta_{k_2-1}(s)P_T^{(J)}(s,t)ds.$$

Similarly, it can readily be shown that to the order of $\text{Min}\{o\left(\alpha_{k_1-1}(t)\right), o\left(\beta_{k_2-1}(t)\right)\}$, $Q_T(t) \sim Q_1(t)+Q_2(t)$. To further simplify the calculation of $Q_T(j)$, we observe that in studying human cancers, one time unit (i.e., $\Delta t = 1$) is usually assumed to be 3 months or 6 months or longer. In these cases, one may practically assume $P_T^{(I)}(s,t) \approx 1$ and $P_T^{(J)}(s,t) \approx 1$ if $t - s \ge 1$.

A Statistical Model for Cancer Incidence Data

Let y_j be the observed number of the number of cancer cases Y_j developed during $(t_{j-1}, t_j]$ given n_j people at risk for cancer, who are normal at birth (t_0). We assume that each individual develops colon cancer tumor by the same mechanism independently of one another. Then for each person who is normal at birth (t_0), the probability that this individual would develop colon cancer tumor during the j-th age group $(t_{j-1}, t_j]$ is given by $Q_T(j)$. It follows that the probability distribution of Y_j given that n_j is:

$$Y_j \sim \text{Binomial}\{n_j, Q_T(j)\}. \tag{3.20}$$

Because n_j is very large and $Q_T(j)$ is very small, approximately Y_j is Possion with mean $\tau_j = n_j Q_T(j)$. Notice that to the order of $\text{Max}\{o\left(\alpha_{k_1-1}(t)\right), o\left(\beta_{k_2-1}(t)\right)\}$, τ_j (and hence the probability distribution of Y_j) depends on the stochastic model of colon carcinogenesis through the expected number $\{E[I_{k1-1}(t)], E[J_{k2-1}(t)]\}$ of $\{I_{k1-1}(t), J_{k2-1}(t)\}$ and the parameters $\{\alpha_{k_1-1}(t), \beta_{k_2-1}(t)\}$ over the time period $(t_{j-1}, t_j]$.

The State Space Model of Human Colon Cancer

State space model is a stochastic model which consists of two sub-models: The stochastic system model which is the stochastic model of the system and the observation model which is a statistical model based on available observed data from the system. Thus, the state space model of a system takes into account the basic mechanisms of the system and the random variation of the system through its stochastic system model and incorporates all these into the observed data from the system; furthermore, it validates and upgrades the stochastic model through its observation model and the observed data of the system. As illustrated in Tan (2002), the state space model has many advantages over

both the stochastic model and the statistical model when used alone since it combines information and advantages from both of these models.

For human colon cancer, the stochastic system model of the state space model is the stochastic model consisting of 2 pathways with each pathway following a multi-stage model as described in Section 3.3; the observation model of this state space model is a statistical model based on the observed number of colon cancer cases as described in Section 3.4.

The Stochastic System Model and the State Variables

Putting $\Delta t = 1$ for some fixed small interval, then the staging variables are

$\mathbf{X} = \{\widetilde{\mathrm{X}}(\mathrm{t}), \mathrm{t} = \mathrm{t}_0, \mathrm{t}_0 + 1, \ldots, \mathrm{t}_\mathrm{M}\}$ and the transition variables are $\mathbf{U} = \{\widetilde{\mathrm{U}}(\mathrm{t}), \mathrm{t} = \mathrm{t}_0, \mathrm{t}_0 + 1, \ldots, \mathrm{t}_\mathrm{M} - 1\}$.

From results in Section 3.3, the joint probability distribution of $\{\mathbf{X}, \mathbf{U}\}$ given the parameters Θ is:

$$\mathrm{P}(\mathrm{X}, \mathrm{U}|\Theta) = \prod_{\mathrm{t}=\mathrm{t}_0+1}^{\mathrm{t}_\mathrm{M}} \mathrm{P}\{\widetilde{\mathrm{X}}(\mathrm{t})|\widetilde{\mathrm{X}}(\mathrm{t}-1), \widetilde{\mathrm{U}}(\mathrm{t}-1)\} \times \mathrm{P}\{\widetilde{\mathrm{U}}(\mathrm{t}-1)|\widetilde{\mathrm{X}}(\mathrm{t}-1)\} \qquad (3.21)$$

where $\mathrm{P}\{\widetilde{\mathrm{U}}(\mathrm{t}-1)|\widetilde{\mathrm{X}}(\mathrm{t}-1)\}$ and $\mathrm{P}\{\widetilde{\mathrm{X}}(\mathrm{t})|\widetilde{\mathrm{X}}(\mathrm{t}-1), \widetilde{\mathrm{U}}(\mathrm{t}-1)\}$ are given by equations (3.16) and (3.17) respectively and where

$\Theta = \{\lambda_\mathrm{I}, \lambda_\mathrm{J}, \alpha_\mathrm{l}(\mathrm{t}), \beta_\mathrm{r}(\mathrm{t}), \mathrm{b}_\mathrm{l}^{(\mathrm{I})}(\mathrm{t}), \mathrm{d}_\mathrm{l}^{(\mathrm{I})}(\mathrm{t}), \mathrm{b}_\mathrm{r}^{(\mathrm{J})}(\mathrm{t}), \mathrm{d}_\mathrm{r}^{(\mathrm{J})}(\mathrm{t}), l = 1, \ldots, k_1 - 1, r = 1, \ldots, k_2 - 1\}$.

Notice that this probability distribution is basically a product of Poisson distributions and multinomial distributions.

The Observation Model Using SEER Data

Put $\mathbf{Y} = (Y_j, j = 1, \ldots, m)$ and $y = (y_j, j = 1, \ldots, m)$. By the probability distribution given by equation (3.18), the conditional probability density of $\mathbf{Y}$ given $\{\mathbf{X}, \mathbf{U}, \Theta\}$ is approximately given by:

$$P(X, U|\Theta) = \prod_{i=1}^{m} h(Y_j; \tau_j) \quad (3.22)$$

where $h(Y_j; \tau_j)$ is the density at Y_j of the Poisson distribution with mean τ_j. Then the likelihood function of Θ given $(\mathbf{X}, \mathbf{U})$ is $L(\Theta|\tilde{y}, X, U) = \prod_{j=1}^{m} h(Y_j; \tau_j)$. It follows that the deviance from this density is:

$$Dev = -2\{\log L(\Theta|\tilde{y}, X, U) - \log L(\widehat{\Theta}|\tilde{y}, X, U)\} = \sum_{j=1}^{m}\{\tau_j - y_j - y_j \log\frac{\tau_j}{y_j}\} \quad (3.23)$$

where $\widehat{\Theta} = \hat{\tau}_j, j = 1, \ldots, m$ and $\hat{\tau}_j = y_j$ is the maximum likelihood estimate of τ_j. From equations (3.19)-(3.20), we have for the joint density of $(\mathbf{X}, \mathbf{U}, \mathbf{Y})$ given Θ:

$$P\{\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{Y}\,|\Theta\} = P\{\boldsymbol{X}, \boldsymbol{U}|\Theta\}P\{\boldsymbol{Y}\,|\boldsymbol{X}, \boldsymbol{U}, \Theta\} \quad (3.24)$$

To apply the above distribution to estimate unknown parameters and to fit real data, we also make the following assumptions: (a) From biological observations (Chapelle, 2004; Fodde et al., 2001; Fodde et al., 2001; Green & Kaplan, 2003, Hawkins & Ward, 2001, Hisamuddin & Yang, 2004; Peltonmaki, 2001; Sparks et al., 1998; Ward et al. 2001), one may practically assume that $\{\alpha_l(t) = \alpha_l, l = 0, 1, 2, 3;\ \beta_r(t) = \beta_r,\ r = 0, 1, 2, 3, 4;\ b_3^{(I)}(t) = b_3^{(I)}, d_3^{(I)}(t) = d_3^{(I)}, b_4^{(J)}(t) = b_4^{(J)}, d_4^{(J)}(t) = d_4^{(J)}\}$. (b) Because the colon polyps are generated by proliferation of I_2 cells and J_3 cells and because the polyps can only grow to a maximum size of about 10 mm^3, we assume that $\{b_2^{(I)}(t) = b_2^{(I)}e^{-\delta_1 t}, d_2^{(I)}(t) = d_2^{(I)}e^{-\delta_1 t}\}$ and $\{b_3^{(J)}(t) = b_3^{(J)}e^{-\delta_2 t}, d_3^{(J)}(t) = d_3^{(J)}e^{-\delta_2 t}\}$ for some small ($\delta_i > 0$, $i = 1, 2$). (c) Because colon cell divisions are mainly due to action of the β-

Catenin gene, one may also assume $\{\gamma_1^{(I)}(t) = \gamma_1^{(J)}(t) = 0, j = 1, 2\}$. In this case, one has approximately $I_1(t + \Delta t) = I_1(t) + M_0^{(I)}(t)$ and $J_r(t + \Delta t) = J_r(t) + M_{r-1}^{(J)}(t), r = 1, 2$. Under these assumptions, the unknown parameters of interest are $\Theta = \{\Theta_1, \Theta_2\}$, where $\Theta_1 = \{\lambda_I, \lambda_J, \alpha_i, \beta_3, \beta_j, b_{i+1}^{(I)}, d_{i+1}^{(I)}, b_{j+2}^{(J)}, d_{j+2}^{(J)}, i = 1, 2, j = 1, 2, \delta_l, l = 1, 2\}$ and $\Theta_2 = (\alpha_3, \beta_4)$.

The Generalized Bayesian Method and the Gibbs Sampling Procedure

The generalized Bayesian inference is based on the posterior distribution $P(\Theta|\tilde{y}, \boldsymbol{X}, \boldsymbol{U})$ of Θ given $\{\mathbf{X}, \mathbf{U}, \mathbf{Y} = \tilde{y}\}$. This posterior distribution is derived by combining the prior distribution $P\{\Theta\}$ of Θ with the probability distribution $P\{\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{Y}|\Theta\}$ given by equation (3.20) with $\mathbf{Y}$ being replaced by $\tilde{y}$. It follows that this inference procedure would combine information from three sources:

(1) Previous information and experiences about the parameters in terms of the prior distribution $P\{\Theta\}$; (2) Biological information represented by the stochastic system equations of the stochastic system ($P\{\boldsymbol{X}, \boldsymbol{U}|\Theta\}$); (3) Information from observed data, represented by the statistical model through the conditional likelihood $L(\Theta|\, \tilde{y}, \boldsymbol{X}, \boldsymbol{U})$. Because of additional information from the stochastic system model, this inference procedure is advantageous over the standard Bayesian procedure in that it can avoid the identifiability problems associated with standard Bayesian method. For example, we have shown that to the order of $\text{Max}\{o(\alpha_3(t)), o(\beta_4(t))\}$ the probability distribution of the Y_j's depends on the stochastic model through the expected numbers of $I_3(t)$ and $J_4(t)$, which depend on the birth rates and death rates only through the difference of these rates. It follows that it is not possible to estimate the birth rates and death rates separately by

the traditional Bayesian method. Most importantly, the number of parameters is very large and the number of data points is limited. Thus, without information from the stochastic system model, it is virtually impossible to estimate all unknown parameters; for more examples, see Tan (2000, 2002).

The Prior Distribution of the Parameters

For the prior distributions of Θ, because biological information have suggested some lower bounds and upper bounds for the mutation rates and for the proliferation rates, we assume

$$P(\Theta) \propto c \ (c > 0) \tag{3.25}$$

where c is a positive constant if these parameters satisfy some biologically specified constraints; and equal to zero otherwise. These biological constraints are:

(i) For the mutation rates of the I_i cells in the LOH pathway, $1 < \lambda_I < 1000$ ($N \rightarrow I_1$), $10^{-6} < \alpha_i < 10^{-4}$, $i = 1, 2, 3$. For the proliferation rates of I_i cells in the LOH pathway, $\gamma_1(t) = 0$, $0 < b_i^{(I)} < 0.5$, $i = 2, 3$, $\gamma_2^{(I)}(t) = \gamma_2^{(I)} e^{-\delta_1 t}$, $10^{-2} < \gamma_2 < 2 * 10^{-2}$, $10^{-5} < \delta_I < 5 *10^{-3}$, $10^{-2} < \gamma_2 < 0.5$.

(ii) For the mutation rates in the MSI pathway, $1 < \lambda_J < 1000$ ($N \rightarrow J_1$), $10^{-8} < \beta_1 < 10^{-5}$, $10^{-6} < \beta_j < 10^{-2}$, $j = 2, 3, 4$. For the proliferation rates in the MSI pathway, $\gamma_i^{(J)}(t) = 0, i = 1, 2$, $\gamma_3^{(J)}(t) = \gamma_3^{(J)} e^{-\delta_2 t}$, $10^{-3} < \gamma_3^{(J)} < 0.5$, $j = 3, 4$, $10^{-6} < \delta_2 < 5 *10^{-4}$, $0 < b_j^{(J)} < 0.5$, $j = 3, 4$.

We will refer the above prior as a partially informative prior which may be considered as an extension of the traditional non-informative prior given in Box and Tiao (1973).

The Posterior Distribution of the Parameters Given {Y = ỹ, X, U}

Combining the prior distribution given in (3.6.1) with the density of $P\{\boldsymbol{X},\boldsymbol{U},\boldsymbol{Y}\,|\Theta\}$ given in equation (3.20), one can readily derive the conditional posterior distribution of Θ given $\{\boldsymbol{X},\boldsymbol{U},\boldsymbol{Y}=\tilde{y}\}$. For ($l$ = 2, 3), denote by: $N_{lI}=\sum_{t=1}^{t_M} I_l(t)$, $B_{lI}=\sum_{t=1}^{t_M} B_l^{(I)}(t)$, $D_{lI}=\sum_{t=1}^{t_M} D_l^{(I)}(t)$; similarly, for r = 3, 4, we define $\{B_{rJ}, D_{rJ}, N_{rJ}\}$ by replacing $(I_l(t), B_l^{(I)}(t), D_l^{(I)}(t))$ by $(J_r(t), B_r^{(J)}(t), D_r^{(J)}(t))$ respectively. Then, we have the following results for the conditional posterior distributions:

(i) The conditional posterior distributions of $\Theta_1(1) = \{\lambda_I, \lambda_J, \alpha_l, l = 1, 2, \beta_r, r = 1, 2, 3\}$ given

$\{\mathbf{X},\mathbf{U},\mathbf{Y}=\tilde{y}\}$ is:

$$P\{\Theta_1(1)|X,U,\tilde{y}\}$$

$$\propto P\{\Theta_1(1)\}e^{-(\lambda_I+\lambda_J)(t_M-t_0)}\,[\lambda_I]^{\sum_{t=t_0}^{t_M} M_0^{(I)}(t)}$$

$$\times\ [\lambda_J]^{\sum_{t=t_0}^{t_M} M_0^{(J)}(t)} \prod_{l=1}^{2} e^{-N_{lI}\alpha_l}[\alpha_l]^{\sum_{r=1}^{m} y_r} \times \prod_{r=1}^{3} e^{-N_{rJ}\beta_r}[\beta_r]^{\sum_{r=1}^{m} y_r}$$

(ii) The conditional posterior distributions of $\Theta_1(2)=\left\{b_3^{(I)}, d_3^{(I)}, b_4^{(J)}, d_4^{(J)}\right\}$ given $\{\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{Y}=\tilde{y}\}$ is:

$$P(\Theta_1(2)|\boldsymbol{X},\boldsymbol{U},\tilde{y}) \propto P\{\Theta_1(2)\}f\left(B_{3I},D_{3I},N_{3I},b_3^{(I)},d_3^{(I)}\right)f(B_{4J},D_{4J},N_{4J},b_4^{(J)},d_4^{(J)})$$

(iii) The conditional posterior distribution of $\{\alpha_3, \beta_4\}$ given $\{\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{Y}=\tilde{y}\}$ is:

$$P\{\alpha_3,\beta_4|X,U,\tilde{y}\} \propto P\{\alpha_3,\beta_4\}\prod_{j=1}^{m} e^{-\tau_j}(\tau_j)^{y_j}$$

(vi) The conditional posterior distribution of $\{b_2^{(I)}, d_2^{(I)}, \delta_l\}$ given $\{\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{Y}=\tilde{y}\}$ and the conditional posterior distribution of $\{b_3^{(J)}, d_3^{(J)}, \delta_2\}$ given $\{\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{Y}=\tilde{y}\}$ are represented respectively by:

$$P\left\{b_2^{(I)}, d_2^{(I)}, \delta_1 \middle| X,U,\tilde{y}\right\} \propto P\{b_2^{(I)}, d_2^{(I)}, \delta_1\}\, f(B_{2I},D_{2I},N_{2I},b_2^{(I)},d_2^{(I)}) \times \prod_{t=1}^{t_M}\left(1-\frac{\left(b_2^{(I)}+d_2^{(I)}\right)(1-e^{-\delta_1 t})}{1-b_2^{(I)}-d_2^{(I)}}\right)^{I_2(t)-B_2^{(I)}(t)-D_2^{(I)}(t)},$$

$$P\left\{b_3^{(J)}, d_3^{(J)}, \delta_2 \middle| X, U, \tilde{y}\right\} \propto P\{b_3^{(J)}, d_3^{(J)}, \delta_2\} f(B_{3J}, D_{3J}, N_{3J}, b_3^{(J)}, d_3^{(J)}) \times \prod_{t=1}^{t_M} (1 - \frac{\left(b_3^{(J)} + d_3^{(IJ)}\right)(1-e^{-\delta_2 t})}{1-b_3^{(J)} - d_3^{(J)}})^{J_3(t) - B_3^{(J)}(t) - D_3^{(J)}(t)}$$

The Multi-level Gibbs Sampling Procedure For Estimating Parameters

Given the above probability distributions, the multi-level Gibbs sampling procedure for deriving estimates of the unknown parameters are given by:

(a) Step 1: Generating (**X**,**U**) Given (**Y** = ỹ, Θ) (The Data-Augmentation Step):

Given **Y** = ỹ and given Θ, use the stochastic equations (3.3)-(3.4) and the probability distributions given by equations (3.5)-(3.10) in Section 3.3 to generate a large sample of (**X**,**U**). Then, by combining this sample with $P\{\mathbf{Y} = \tilde{y} \mid \mathbf{X},\mathbf{U},\Theta\}$ to select (**X**,**U**) through the weighted bootstrap method due to Smith and Gelfant (1992). This selected (**X**,**U**) is then a sample from $P\{\mathbf{X},\mathbf{U}|\mathbf{Y} = \tilde{y}, \Theta\}$ even though the latter is unknown.

(For proof, see Tan (2002), Chapter 3.) Call the generated sample $(\hat{X}, \hat{U})$.

(b) Step 2: Estimation of $\Theta = \{\Theta_1, \Theta_2\}$ Given {**Y** = ỹ, **X**, **U**}:

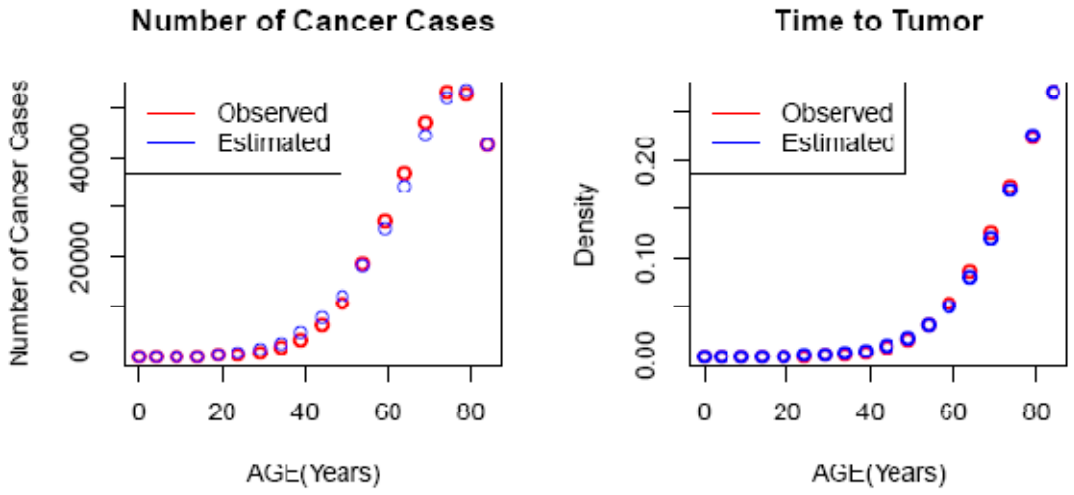


Figure 11. Estimated and Observed Colon Cancer Cases and Density of Time to Tumor (2006 SEER data)

Given $\mathbf{Y} = \tilde{y}$ and given $(\mathbf{X}, \mathbf{U}) = (\hat{X}, \hat{U})$ from Step 1, derive the posterior mode of the parameters by maximizing the conditional posterior distribution $P\{\Theta|\hat{X}, \hat{U}, \hat{y}\}$. Denote the generated mode as $\hat{\Theta}$.

(c) Step 3: Recycling Step.

With $\{(\mathbf{X},\mathbf{U}) = (\hat{X}, \hat{U}), \Theta = \hat{\Theta}\}$ given above, go back to Step (a) and continue until convergence.

The convergence of the above steps can be proved using procedure given in Tan (2002, Chapter 3). At convergence, the $\hat{\Theta}$ are the generated values from the posterior distribution of Θ given $\mathbf{Y} = \tilde{y}$ independently of $(\mathbf{X},\mathbf{U})$ (for proof, see Tan (2002), Chapter 3). Repeat the above procedures one then generates a random sample of Θ from the posterior distribution of Θ given $\mathbf{Y} = \tilde{y}$; then one uses the sample mean as the estimates of Θ and use the sample variances and covariances as estimates of the variances and covariances of these estimates.

Application to Fit the SEER Data

In this section, we will apply the above model to the NCI/NIH colon cancer 1996,

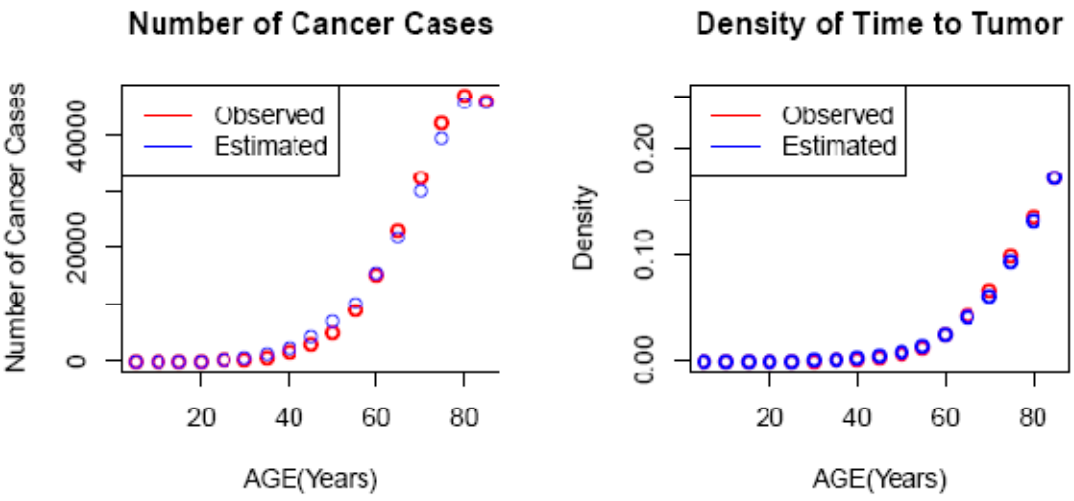


Figure 12. Estimated and Observed Colon Cancer Cases and Density of Time to Tumor (2001 SEER data)

2001, and 2006 data from the SEER project. Given in Table 10-12 are the numbers of people at risk and colon cancer cases in the age groups together with the predicted cases from the model for above three years, respectively. There are 18 age groups with each group spanning over 5 years. To fit the data, we have assumed that $\gamma_1^{(I)} = \gamma_1^{(J)} = 0$ for j = 1, 2 because of the observation that uncontrolled cell division of colon stem cells is mainly initiated by the oncogene β-Catenin in 3p22. Given in Table 13-15 are the estimates of the mutation rates, the birth rates and the death rates of the I_i cells and J_j cells. Given in Figure 11-13 are plots of number of cancer cases probability density of time to tumors.

From these results, we have made the following observations:

(a) As shown by results in Table 10-12 and Figure 11-13, the predicted number of cancer cases is very close to the observed cases in all age groups for all three datasets. This indicates that the model fits the data well and the mode is quite stable. Thus one can safely assume that the human colon cancer can be described by a model of 2 pathways. The AIC (Akaike Information Criterion) and the BIC (Bayesian Information Criterion) from the model are 55.96 and 81.30 obtained from SEER 2006 data which are smaller than the AIC of 816.0667 and the BIC value of 827.1513 from a single pathway 4-stage model respectively (Luebeck & Moolgavkar, 2002). The AIC and BIC computed from SEER 1996 and 2001 are also much smaller than those from single 4-stage pathway model. This shows that the multiple pathway model fits better than the single pathway 4-stage model as proposed by Luebeck and Moolgavkar (2002).

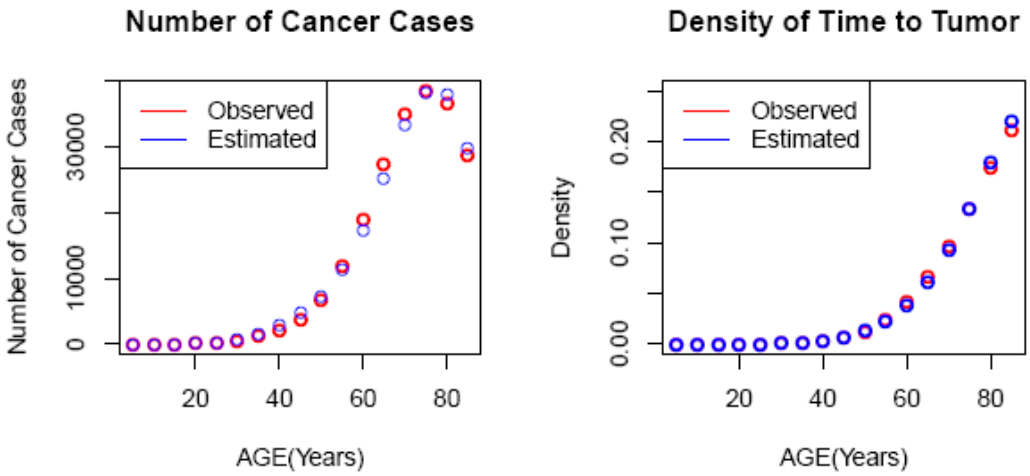


Figure 13. Estimated and Observed Colon Cancer Cases and Density of Time to Tumor (1996 SEER data)

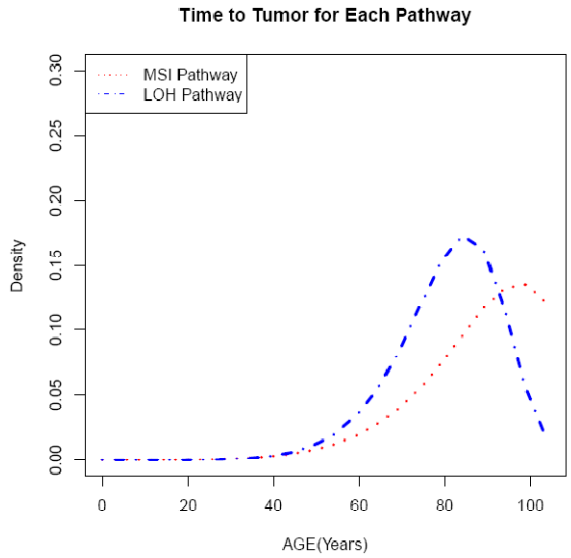


Figure 14. Time to Tumor for Each Pathway (2006 SEER)

(b) From Table 10-12 and Figure 14-16, it is observed that the largest number of cancer cases is in the age group between 70 and 75 years old. Comparing the values of $Q_i(j)$ between the CIN pathway ($i = 1$) and the MSI pathway ($i = 2$), it appears that the largest cancer cases is between the age group 80 and 85 years old for the CIN pathway

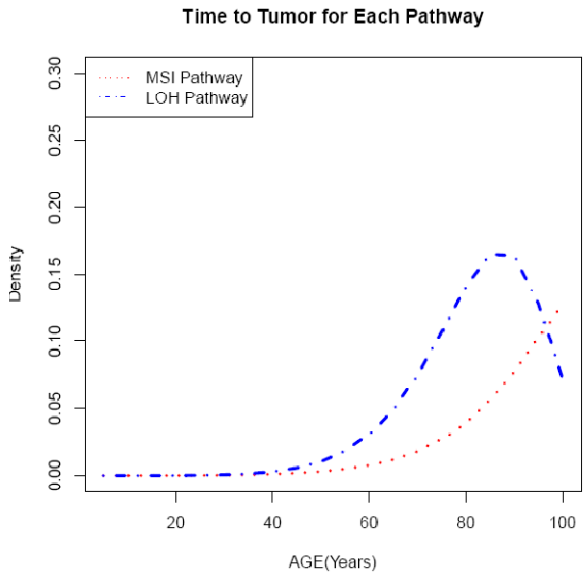


Figure 16. Time to Tumor for Each Pathway (1996 SEER)

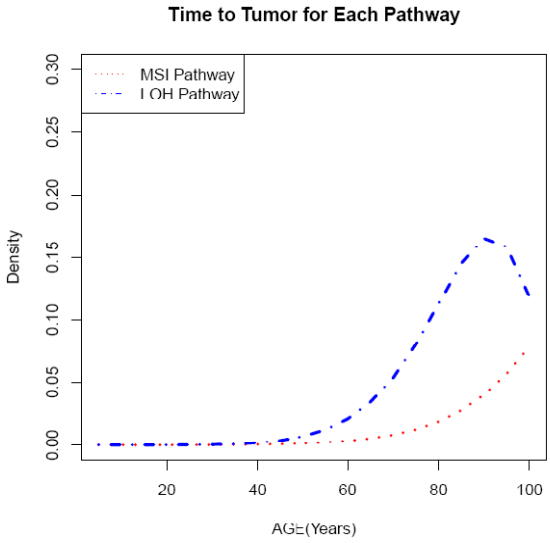


Figure 15. Time to Tumor for Each Pathway (2001 SEER)

and is greater than 85 years old for the MSI pathways. Presumably this might be due to the fact that the MSI pathway has one more stage than the CIN pathway.

(c) Reflecting the contribution of the APC gene on chromosomal instability, results in Table 13-15 showed that the mutation rates of the I_r cells from $I_1 \rightarrow I_2$ and from $I_2 \rightarrow I_3$

had increased about 100 times and 1000 times respectively than the mutation rate from $N \rightarrow I_1$ cells. Similarly, due to the contribution to genomic instability by the mis-match repair genes, the mutation rates from $J_1 \rightarrow J_2$, from $J_2 \rightarrow J_3$ and $J_3 \rightarrow J_4$ had increased about 5×10^2, 0.5×10^4 and 10^4 times respectively than the mutation rate from $N \rightarrow J_1$. Notice also from Table 13-15 that the mutation rates from $J_1 \rightarrow J_2 \rightarrow J_3 \rightarrow J_4$ are about 2 to 3 times of those from $I_1 \rightarrow I_2 \rightarrow I_3$. As shown in probability plots (not shown here), these increases have speeded up the time to cancer in the MSI pathway by about 5-10 years.

(d) Results in Table 13-15 showed that the mutation rates from $I_3 \rightarrow I_4$ and from $J_4 \rightarrow J_5$ are of the order 10^{-6} which were about $10^2 \rightarrow 10^3$times smaller than the mutation rates from $I_1 \rightarrow I_2 \rightarrow I_3$ and from $J_1 \rightarrow J_2 \rightarrow J_3 \rightarrow J_4$. These results might be the consequence that we had ignored the stages of vascular carcinogenesis (i.e., angiogenesis and metastasis; see Hanahan & Weinberg (2000) and Weinberg, 2007) by merging these stages into the last stage. From Weinberg (2007, Chapters 13-14), notice that the angiogenesis and metastasis are also multi-stage processes.

(e) Results in Table 13-15 showed that the proliferation rates (birth rate - death rate) of the I_3 cells and the J_4 cells are of order 10^{-2}which are much larger than the proliferation rates of the I_2 cells and the J_3 cells, due presumably to the effects of the silencing or inactivation of the cell cycle inhibition genes (Smad4 and TGF-β-RII) and the apoptosis inhibition genes (p53 and Bax). Notice from Table 13-15 that the estimates of the proliferation rates of the I_2 and I_3 cells are approximately equal to those of the J_3 and J_4 cells respectively. These results seemed to suggest that the genomic instabilities had little effects on cell proliferations.

Table 10

Colon Cancer Data from SEER 2006 (Overall Population)

Age Group	Number of People at Risk	Observed Colon Cancer Cases	Predicted Colon Cancer Cases*
0	9934747	1	0
0−4	38690768	2	0
5−9	48506058	2	6
10−14	49881935	35	44
15−19	50447512	104	164
20−24	51612785	337	370
25−29	54071811	847	965
30−34	54194486	1829	2080
35−39	50363957	3420	3534
40−44	46029771	6174	6698
45−49	40674188	10950	11072
50−54	36070434	18716	18256
55−59	31084543	27438	25875
60−64	26507762	37155	34867
65−69	22772688	47202	45156
70−74	18785224	53190	52810
75−79	14592602	52887	53479
80−84	9751212	42589	41517

* The predicted numbers were generated by the model with unknown parameters being substituted by the estimates respectively.

Table 11

Colon Cancer Data from SEER 2001 (Overall Population)

Age Group	Number of People at Risk	Observed Colon Cancer Cases	Predicted Colon Cancer Cases*
0	10133168	0	0
0−4	39407080	2	2
5−9	49548675	1	5
10−14	50738796	31	42
15−19	51758589	90	92
20−24	53154897	268	293
25−29	56279465	704	764
30−34	55641828	1532	1662
35−39	50895387	2859	3161
40−44	45268107	5065	5778
45−49	39358389	8954	9345
50−54	34911012	15309	15511
55−59	30414133	23349	22020
60−64	26715010	32760	31318
65−69	23364715	42119	41720
70−74	19125836	47209	46012
75−79	14548483	46122	45963
80−84	9492871	36432	35672

*The predicted numbers were generated by the model with unknown parameters being substituted by the estimates respectively.

Table 12

Colon Cancer Data from SEER 1996 (Overall Population)

Age Group	Number of People at Risk	Observed Colon Cancer Cases	Predicted Colon Cancer Cases*
0−4	40207272	1	0
5−9	39633757	1	3
10−14	41046534	24	25
15−19	42500387	71	94
20−24	44400349	211	262
25−29	46227557	562	701
30−34	44888803	1205	1311
35−39	39700589	2181	2239
40−44	34570896	3839	3981
45−49	29704698	6753	7193
50−54	26561424	11973	11443
55−59	24278604	19111	17366
60−64	22008619	27362	26103
65−69	19243853	35062	34461
70−74	15271418	38503	38322
75−79	11245708	36605	37660
80−84	7237028	28710	29170

*The predicted numbers were generated by the model with unknown parameters being substituted by the estimates respectively.

Table 13

Estimates of Parameters for Each Pathway (SEER 2006)

LOH Pathway				
	I_0	I_1	I_2	I_3
Mutation Rate	1.4E-06 ±1.69E-08	2.2E-04 ±1.32E-05	3.2E-03 ±3.33E-04	1.2E-06 ±2.06E-07
Proliferation Rate	0 N/A	0 N/A	3.6E-03 ±1.12E-03	1.6E-02 ±4.78E-04
Birth Rate Para.	0 N/A	0 N/A	7.4E-03 ±1.03E-03	1.9E-02 ±4.08E-04
Growth Limiting Para.	N/A	N/A	8.3E-05 ±1.4E-05	N/A

MSI Pathway					
	J_0	J_1	J_2	J_3	J_4
Mutation Rate	8.3E-07 ±1.38E-08	3.5E-04 ±1.89E-05	1.4E-03 ±8.57E-05	9.3E-03 ±1.22E-03	7.7E-06 ±1.79E-06
Proliferation Rate	0 N/A	0 N/A	0 N/A	2.8E-03 ±7.01E-04	2.0E-02 ±3.31E-04
Birth Rate	0 N/A	0 N/A	0 N/A	9.6E-03 ±6.08E-04	2.6E-02 ±2.88E-04
Growth Limiting Para.	N/A	N/A	N/A	1.6E-03 ±3.7E-04	N/A

Table 14

Estimates of Parameters for Each Pathway (SEER 2001)

LOH Pathway				
	I_0	I_1	I_2	I_3
Mutation Rate	1.38E-06 ±6.53E-09	1.6E-04 ±3.59E-06	1.6E-03 ±2.6E-05	1.4E-06 ±1.39E-07
Proliferation Rate	0 N/A	0 N/A	4.6E-03 ±3.2E-04	1.9E-02 ±9.83E-05
Birth Rate Para.	0 N/A	0 N/A	6.2E-03 ±3.2E-04	2.2E-02 ±9.49E-05
Growth Limiting Para.	N/A	N/A	8.31E-05 ±5.0E-05	N/A

MSI Pathway					
	J_0	J_1	J_2	J_3	J_4
Mutation Rate	8.42E-07 ±5.48E-09	2.8E-04 ±6.01E-06	1.1E-03 ±1.38E-05	5.4E-03 ±7.29E-05	9.67E-06 ±3.93E-06
Proliferation Rate	0 N/A	0 N/A	0 N/A	1.3E-02 ±1.95E-03	1.8E-02 ±6.33E-05
Birth Rate	0 N/A	0 N/A	0 N/A	2.4E-02 ±1.77E-03	1.9E-02 ±5.64E-04
Growth Limiting Para.	N/A	N/A	N/A	6.5E-05 ±1.5E-04	N/A

Table 15

Estimates of Parameters for Each Pathway (SEER 1996)

LOH Pathway				
	I_0	I_1	I_2	I_3
Mutation Rate	1.4E-06 ±6.4E-09	1.5E-04 ±5.05E-06	8.0E-04 ±3.84E-05	1.9E-06 ±2.48E-07
Proliferation Rate	0 N/A	0 N/A	4.9E-03 ±3.54E-04	2.0E-02 ±1.63E-04
Birth Rate Para.	0 N/A	0 N/A	5.7E-03 ±3.5E-04	2.1E-02 ±1.46E-04
Growth Limiting Para.	N/A	N/A	8.3E-05 ±1.4E-05	N/A

MSI Pathway					
	J_0	J_1	J_2	J_3	J_4
Mutation Rate	8.5E-07 ±4.97E-08	2.6E-04 ±5.72E-06	1.1E-03 ±2.52E-05	5.2E-03 ±1.34E-04	1.7E-05 ±6.46E-06
Proliferation Rate	0 N/A	0 N/A	0 N/A	1.8E-02 ±3.1E-03	1.9E-02 ±3.1E-04
Birth Rate	0 N/A	0 N/A	0 N/A	1.9E-02 ±2.9E-03	2.0E-02 ±2.6E-04
Growth Limiting Para.	N/A	N/A	N/A	3.6E-04 ±2.7E-04	N/A

Conclusions and Discussion

Recent studies of cancer molecular biology have indicated very clearly that human colon cancer is developed through multiple pathways (Chapelle, 2004; Fodde et al, 2001; Fodde et al., 2001; Green & Kaplan, 2003, Hawkins & Ward, 2001, Hisamuddin & Yang, 2004; Peltonmaki, 2001; Sparks et al., 1998; Ward et al., 2001). This indicates that single pathway models are not realistic and hence may lead to incorrect prediction and confusing results. For developing efficient prevention and controlling procedures for human colon cancer and for prediction of future human colon cancer, in this chapter we have developed a stochastic model and a state space model for carcinogenesis of human colon cancer involving multiple pathways with each pathway being a multi-stage model. Using this model, we have derived for the first time the probability distribution of the numbers of initiated cells and the probability distribution of time to cancer tumors. Such derivation by the traditional approach is extremely difficult and had not been attempted previously for multiple pathway models. Based on the state space model of colon cancer, we have developed a generalized Bayesian procedure to estimate the unknown parameters and to predict future cancer cases. This approach combines information from three sources: The stochastic system model via $P\{\mathbf{X}, \mathbf{U} \mid \Theta\}$, the prior information via $P\{\Theta\}$ and information from data via $L\{\Theta|\tilde{y}, \mathbf{X}, \mathbf{U}\}$. Because of additional information from the stochastic system model, our procedure is advantageous over the standard Bayesian procedure and the sampling theory procedure. Notice that there are a large number of unknown parameters in the model and only a limited amount of data are available. Without this additional information, it is then not possible to estimate all unknown parameters. Notice also that through the stochastic system model, one can

incorporate biological mechanism into the model. Because the number of stages and the mutation rates of intermediate cells in different pathways are different and different drugs may affect different pathways, we believe that this is important and necessary.

We have applied these models and procedure to three different NCI SEER data (up to November, 2007). Our results showed that the proposed multiple pathways model was quite reliable and fitted better than the single pathway 4-stage model as proposed by Luebeck and Moolgavkar (2002). (The respective AIC and BIC for the multiple pathways model are 55.96 and 81.30 which are 10 times smaller than those of the AIC (816.0667) and BIC (827.1513) respectively of the single pathway 4-stage model.)

In this preliminary study, we have not yet compared the multiple pathways model with the single pathway model regarding prediction of future cancer cases and evaluation of treatment protocols for human colon cancer. This will be our future research, and we will not go any further here.

4. A STOCHASTIC AND STATE SPACE MODEL OF HUMAN LIVER CANCER-MULTIPLE-PATHWAY MODEL INVOLVING BOTH HEREDITARY AND NON-HEREDITARY CANCER

Introduction

It is well documented that each cancer tumor is derived from a single stem cell which has sustained a finite number of genetic and epigenetic changes and with intermediate cells subjecting to stochastic cell proliferation and differentiation. That is, carcinogenesis is a stochastic multi-stage process involving genetic and epigenetic changes and stochastic cell proliferation and cell differentiation. Recent studies by molecular biologists have also shown clearly that for many human cancers (colon cancer, liver cancer, lung cancer and melanoma, etc.; see Little (2008), Tan (1991), Tan et al. 2008a, 2008b), the same cancer can be derived by multiple pathways with each pathway being a multi-stage model. In Tan et al. (2008a) and Tan and Yan (2009), we have developed multiple pathway models for human colon cancer. The major stages for human liver cancer development, on the other hand, have been identified by histopathological evidence though genetic signaling pathways have not been well established. In this chapter we develop a multiple-pathway model for human liver cancer, including non-hereditary hepatocelluler carcinoma and hereditary hepatoblastoma.

For developing biologically supported stochastic model of carcinogenesis, in Section 4.2 we present the most recent cancer biology of human liver cancer. Using results from Section 4.2, in section 4.3 we develop a multi-stage model for carcinogenesis of non-hereditary human liver cancer, involving multiple pathways. In Section 4.4 we derive a statistical model for cancer incidence data of non-hereditary human liver cancer. And

then we extend the multiple-pathway multiple-stage model to incorporate inherited liver cancer, which is shown in Section 4.5. In Section 4.6 we derive a statistical model for cancer incidence data of hereditary human liver cancer. By combining models from Sections 4.3 - 4.6, in Section 4.7 we develop a state space model for human liver cancer. In Section 4.8, by using the state space model in Section 4.7, we develop a generalized Bayesian inference procedure to estimate unknown parameters and to predict state variables. To illustrate the applications of the model and methods, in Section 4.9 we apply the model and methods to the liver cancer incidence data from SEER. Finally in Section 4.10, we discuss the usefulness of the model and methods and provide some conclusions.

A Brief Summary of Liver Cancer Biology

Hereditary Liver Cancer

The molecular biologists and clinicians have shown that mechanisms of developing embyronal liver cancer (hepatoblastoma, or HBL) are quite different from these for sporadic liver cancer (Dufour & Clavien, 2010; Grisham, 2002; Hirschman & Tomlinson, 2005). HBL occurs almost exclusively in infants and children 4 years of age or younger, especially in children with a family history of Familial adenomatous polyposis (FAP). Clinical studies indicated no association between HBL and hepatitis infection, maternal estrogen exposure and cigarette smoking (Grisham, 2002). The major genetic mutation in FAP is mutation of adenomatous polyposis coli (APC) gene (Hirschman & Tomlinson, 2005). According to this mechanism for HBL, the individual is in the first stage (A_1 stage) if one copy of the APC gene has been lost or mutated or inactivated, in the second stage (A_2 stage) if both copies have been lost or mutated or inactivated.

Non-hereditary Liver Cancer

For sporadic liver cancer, the primary hepatocellular neoplasms, on the other hand, occur at all ages and consist of hepatocellular adenoma (HCA), hepatocellular carcinoma (HCC). HCA is benign but HCC is malignant. HCA is extremely rare occurring mainly in hepatocyte which lack both chronic hepatitis and cirrhosis. From these observations, to model sporadic human liver cancer (i.e., after age 15 years old), one may practically ignore HCA but only concern HCC. We give a brief summary of the mechanisms of HCC.

Risk Factors. HCC is a fatal malignant cancer with the majority of cases (90-95%) occurring in hepatitis B virus (HBV, over 65%) and hepatitis C virus (HCV, over 20%) carriers (Wands, 2004). Other important risk factors are food contamination with aflatoxin (AFB1), alcohol consumption, smoking, exposure to arsenic or vinyl chloride, high intake of iron and low intake of antioxidant vitamins and selenium. HCC is rare in the United States and Western Europe because of the low infection rate of HBV in these areas, but it is one of the most common malignancies in Taiwan, Southern China (Qidong, China) and Africa, due presumably to the high incidence of HBV, HCV and AFB1 contamination in these areas. HCC is rising in US due to rampant spread of HCV among US population between 1970 and 1990 via contaminated blood product, needle sharing, etc.; see DeFrances (2005).

The Multi-stage Model and Multiple Pathways for HCC. Non-diseased liver hepatocyte (denoted by NL) are long-lived cells with little turnover. When infected by HBV or HCV, the liver develops into a liver with chronic hepatitis hepatocyte (denoted by CH), which can further develop into cirrhosis (denoted by CC), a diffuse form of

hepatic fibrosis. (Cirrhosis may also be induced by alcohol intake or other factors than HBV and HCV, but the incidence of HCC in cirrhosis without hepatitis infection is low (< 15%) when compared to the incidence of HCC in cirrhosis with HBV /or HCV, see Craig (2003). HCC can either develop from cirrhosis (60-80%) by following the multi-stage model: $NL \rightarrow CH \rightarrow CC \rightarrow PAH$ (*denoted by* PAH_1) $\rightarrow DH$ (*denoted by* DH_1) $\rightarrow$ *HCC*, or from chronic hepatitis hepatocyte (20-40%) by following the multistage model: $NL \rightarrow CH \rightarrow PAH$ (*denoted by* PAH_2) $\rightarrow DH$ (*denoted by* DH_2) $\rightarrow$ *HCC*; see Figure 17 as proposed by Thorgeirsson and Grisham (2002). In these multi-stage models, the intermediate lesions of pre-neoplastic hepatocyte are PAH (Foci of Phenotypically Altered Hepatocyte), and DH (Dysplastic Hepatocyte). Notice that the PAH and DH from the first pathway are different from those of the second pathway because the former cells carry more genetic and epigenetic changes than those from the latter pathway respectively (Grisham, 2002; Thorgeirsson & Grisham, 2002, Yeh et al., 2001). Also, because the first pathway accounts for about 60-80% of all HCC, the proliferation rates and mutations rates of the PAH_1 cells and DH_1 cells are considerably greater than those of the PAH_2 cells and DH_2 cells respectively.

Some Relevant Biological Information for Modeling HCC. To develop stochastic models of HCC along the above pathways, the following observations are useful information which needs to be taken into account. In this chapter, we will integrate this information into the models through partial informative prior as proposed by Tan et al. (2008a).

All lesions (CC, FAH, DH) and HCC are clonal deriving from a single hepatocyte which has sustained irreversible genetic or epigenetic changes with the latter stage lesions

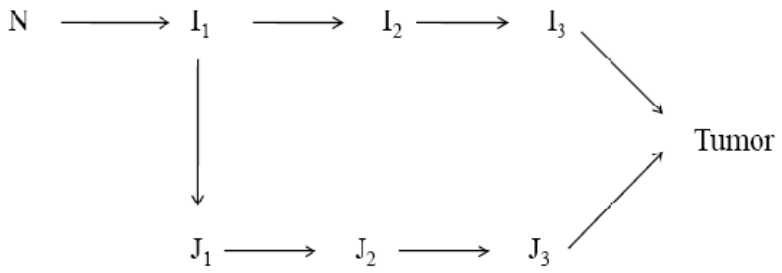
accumulate more changes (Breuhahn, Longerich, & Schirmacher, 2006; Buendia, 2000; Chen & Chen, 2002; Grisham, 2002; Moradpour & Wands, 2003; Pogribny, Rusyn, & Beland, 2008; Shih et al., 2006; Thorgeirsson & Grisham, 2002; Villanueva, Newell, Chiang, Scott, & Llovet, 2007; Yeh et al., 2001;). Genomic alterations appear to be developed randomly beginning in pre-neoplastic stages CH and CC, and escalating in PAH, DH, and HCC. Thus, all important genetic changes may appear in any of CH, CC, PAH, DH, and HCC but the frequencies of different genetic changes may be very different between these stages, due to different cell micro-environment in the liver at different stages and interactions with hepatitis infection. For example, for the percentage of hepatocyte with activated telomerase, there are 4% in CH, 8% in CC, 55% in DH and 84% in HCC respectively; for aberrant methylation, 10% in CH, 16% in CC, 15% in DH, and 40% in HCC respectively.

In normal liver cells and normal hepatocyte (NL cells), both the proliferation rate and the death rate are very small and the birth rate is equal to the death rate. On the other hand, both of these rates are increased in CH, CC, the intermediate lesions and in HCC and the birth rate in these cells is always greater than the death rate respectively. Further these rates of cells in more advanced stages are always greater than those of cells in less advanced stages. This is probably the consequence of biological results that advanced stages have accumulated more genetic and epigenetic changes; see Grisham (2002) Pogribny et al. (2008), and Villanueva et al. (2007).

A Multi-Stage Model of Carcinogenesis for HCC

From results of Section 4.2, it follows that the stochastic multi-stage model for HCC can be represented schematically by Figure 17.

In Figure 17, the model assumes that cancer tumors are generated by two pathways with pathway 1 as a k_1-stage multi-stage model ($N \rightarrow I_1 \rightarrow \ldots \rightarrow I_{k_1} \rightarrow Tumor$) involving normal stem cell N and I_i ($i = 1, \ldots, k_1$) cells and with pathway 2 as a ($k_2 + 1$)-stage multi-stage model ($I_1 \rightarrow J_1 \rightarrow \ldots \rightarrow J_{k_2} \rightarrow Tumor$) involving I_1 and J_j ($j = 1, \ldots, k_2 - 1$) cells. (For HCC, $k_1 = k_2 = 3$.) The state variables are then $\tilde{X}(t)$= $\{I_i(t), i = 1, \ldots, k_1 - 1, J_j(t), j = 1, \ldots, k_2 - 1\}$, $\{I_{k_1}(t), J_{k_2}(t)\}$ and $T(t)$, where $T(t)$ denote the number of cancer tumors at time t and where $I_i(t)$ ($J_j(t)$) denote the number of the I_i (J_j) initiated cells for $\{i = 1, \ldots, k_1 - 1$ ($j = 1, \ldots, k_2 - 1$)$\}$ respectively. Notice that while the $\{\tilde{X}(t), t \geq t_0\}$ is a Markov process with continuous time, $T(t)$ in general may not be



I_1= Chronic Infected Hepatocyte Stage
J_1= Cirrhosis Stage
$I_2(J_2)$ = PAH (Phenotype of Adhered Hepatocyte)
$I_3(J_3)$ = DH (Dysplastic Hepatocyte)

Figure 17. A Two-Pathway Model for HCC

Markov; see **Remark 4.1**. If one assumes that the I_{k_1} and J_{k_2} cells grow instantaneously into cancer tumors as soon as they are generated, then one may also assume the $T(t)$ as Markov and identify $\{I_{k_1}(t), J_{k_2}(t)$ as T(t). In this case, as illustrated in Tan (1991), one may use standard Markov theory to derive the probability generating function (pgf) of the

probabilities of the these variables and hence the probability distribution of these variables (i.e., $\tilde{X}(t)$, $T(t)$). Because $T(t)$ may not be Markov (Yakovlev & Tsodikov, 1996) and because cancer progression from $\{I_{k_1}, J_{k_2}\}$ cells to cancer tumors may have some important impacts on tumor development (Fakir, Tan, Hlatky, Hahnfeldt, & Sachs, 2009), in this paper, we will thus propose an alternative approach through stochastic equations. It can easily be shown through the method of pgf that if $T(t)$ is Markov, then the stochastic equation method is equivalent to the method of Markov theory; as we shall see, however, the stochastic equation method is more powerful and does not need to assume Markov for $T(t)$.

Remark 4.1. Because cell proliferation, cell differentiation and apoptosis, mutation orgenetic changes all occur during cell division and cell division cycle, and because $\tilde{X}(t + \Delta t)$ develop from $\tilde{X}(t)$through cell divisions during $(t, t + \Delta t]$, one may practically assume that $(\tilde{X}(t), t \geq t_0)$ is a Markov process with continous time, where t_0 represents time at birth; on the other hand, $T(t + \Delta t)$ may derive from $I_{k_1}(J_{k_2})$ cells before time t, $T(t)$ is in general not Markov (Fakir et al., 2009; Yakovlev & Tsodikov, 1996).

The Stochastic Equation for State Variables

To derive stochastic equations for the state variables, let $B_i^{(I)}(t)$ $(B_j^{(J)}(t))$ be the number of birth of the I_i (J_j) initiated cells during $(t, t + \Delta t]$ $\{i = 1, \ldots, k_1 - 1$ $(j = 1, \ldots, k_2 - 1)\}$, $D_i^{(I)}(t)$ $(D_j^{(J)}(t))$ the number of death of the I_i (J_j) initiated cells during $(t, t + \Delta t]$ $\{i = 1, \ldots, k_1 - 1$ $(j = 1, \ldots, k_2 - 1)\}$ and $M_i^{(I)}(t)$ $(M_j^{(J)}(t))$ the number of mutation $(I_i \rightarrow I_i + 1)$ $(J_j \rightarrow J_j + 1)$ of I_i (J_j) cells during $(t, t+\Delta t]$ $\{i = 1, \ldots, k_1 - 1$ $(j = 1, \ldots, k_2 - 1)\}$. Also let $M_0^{(I)}(t)$ $(M_0^{(J)}(t))$ be the number of mutation of $N \rightarrow I_1$ $(I_1 \rightarrow J_1)$ during

(t, t+Δt]. By the conservation law, we have then the following stochastic equations for the state variables:

$$I_i\ (t+\Delta t) = I_i(t) + M_{i-1}^{(I)}\ (t) + B_i^{(I)}(t) - D_i^{(I)}(t),\ i = 1, \ldots, k_1 - 1, \tag{4.1}$$

$$J_j(t+\Delta t) = J_j(t) + M_{j-1}^{(J)}(t) + B_j^{(J)}(t) - D_j^{(J)}(t),\ j = 1, \ldots, k_2 - 1, \tag{4.2}$$

Because the transition variables $\{M_i^{(I)}(t),\ M_j^{(J)}(t),\ B_i^{(I)}(t),\ B_j^{(J)}(t),\ B_j^{(J)}(t),\ D_j^{(J)}(t)\}$ are random variables, the above equations are stochastic equations. To derive the probability distributions of these transition variables and hence the probability distribution of state variables, let the transition rates (mutation rates, birth rates and death rates) of the state variables as given in Table 16. Then, as shown in Tan et al. (Tan, Zhang, & Chen, 2004), we have that to the order of $o(\Delta t)$, the conditional probability distributions of $M_0^{(I)}(t)$ given $N(t)$ and of $M_0^{(J)}\ (t)$ given $I_1(t)$ are Poisson with means $\lambda_I(t)\Delta t$ and $\lambda_J(t)\Delta t$ respectively, where $\lambda_I(t) = N(t)\ \alpha_0\ (t)$ and $\lambda_J(t) = I_1(t)\ \beta_0(\ (t)$.

Similarly, it is shown in Tan et al. (2004) that the conditional probability distributions of the numbers of births and deaths given the staging variables (i.e., the $I_i(t)$ and $J_j(t)$) follow multinomial distributions independently. That is,

$$M_0^{(I)}(t)|N(t) \sim \text{Poisson}\{\lambda_I(t)\Delta t\},\ \text{independently of } M_0^{(J)}(t); \tag{4.3}$$

$$M_0^{(J)}(t)|I_I(t) \sim \text{Poisson}\{I_1(t)\beta_0(t)\Delta t\},\ \text{independently of } M_0^{(I)}(t); \tag{4.4}$$

for $i = 1, 2, \ldots, k_1 - 1$,

$$\{B_i^{(I)}(t),\ D_i^{(I)}(t)\}|\ I_i(t) \sim \text{Multinomial}\{I_i(t);\ b_i^{(I)}(t)\Delta t,\ d_i^{(I)}(t)\Delta t\}; \tag{4.5}$$

for $j = 1, \ldots, k_2 - 1$,

$$\{B_j^{(J)}(t),\ D_j^{(J)}(t)\}\,|\,J_j(t) \sim \text{Multinomial}\{J_j(t);\ b_j^{(J)}(t)\Delta t,\ d_j^{(J)}(t)\Delta t\}. \quad (4.6)$$

Table 16

Transition Rates and Transition Probabilities for Human Liver Carcinogenesis

1 N	1 N, 1 I_1	$\alpha_0(t)\Delta t + o(\Delta t)$
1 I_1	1 I_1, 1 J_1	$\beta_0(t)\ \Delta t + o(\Delta t)$
1 I_i	2 I_i	$b_i^{(I)}(t)\Delta t + o(\Delta t)$
1 I_i	death	$d_i^{(I)}(t)\Delta t + o(\Delta t)$
1 I_i	1 I_i, 1 I_{i+1}	$\alpha_i(t)\Delta t + o(\Delta t)$
i = 1, …, k_1– 1		
1 J_j	2 J_j	$b_j^{(J)}(t)\Delta t + o(\Delta t)$
1 J_j	death	$d_j^{(J)}(t)\Delta t + o(\Delta t)$
1 J_j	1 J_j, 1 J_{j+1}	$\beta_j(t)\Delta t + o(\Delta t)$
j = 1, …, k_2– 1		

Because the number of mutations of the I_i cells would not affect the size of the I_i population but only increase the number of I_{i+1} cells and because the mutation rate of I_i cells is very small ($10^{-5} \sim 10^{-8}$), it can readily be shown that to the order of $o(\Delta t)$, the conditional probability distribution of $M_i^{(I)}(t)$ given I_i cells at time t is Poison with mean $I_i(t)\alpha_i(t)\Delta t$ independently of $\{B_i^{(I)}(t),\ D_i^{(I)}(t)\}$ and other transition variables. That is,

$$M_i^{(I)}(t)|\ I_i(t) \sim Poisson\{I_i(t)\alpha_i(t)\Delta t\ \},\ i = 1, \ldots, k_1 - 1, \quad (4.7)$$

independently of $\{B_i^{(I)}(t),\ D_i^{(I)}(t)\}$ and other transition variables.

Similarly, we have that to the order of $o(\Delta t)$,

$$M_j^{(J)}(t)|\, J_j(t) \sim Poisson\{J_j(t)\,\beta_j(t)\Delta t\},\ j = 1, \ldots, k_2 - 1, \tag{4.8}$$

Independently of $\{B_j^{(J)}(t),\ D_j^{(J)}(t)\}$ and other transition variables.

Using the probability distributions given by equations (4.3)-(4.8) and by subtracting from the transition variables the conditional expected values respectively, we have the following stochastic differential equations for the staging state variables:

$$dI_i(t) = I_i(t+\Delta t) - I_i(t) = M_{i-1}^{(I)}(t) + B_i^{(I)}(t) - D_i^{(I)}(t) = \Big\{I_i(t)\alpha_{i-1}(t) + I_i(t)\gamma_i^{(I)}(t)\Big\}\Delta t + e_i^{(I)}(t)\Delta t, \quad i = 1, \ldots, k_1 - 1 \tag{4.9}$$

$$dJ_1(t) = J_1(t+\Delta t) - J_1(t) = M_0^{(J)}(t) + B_1^{(J)}(t) - D_1^{(J)}(t) = \Big\{I_1(t)\beta_0(t) + J_1(t)\gamma_1^{(J)}(t)\Big\}\Delta t + e_1^{(J)}(t)\Delta t \tag{4.10}$$

$$dJ_j(t) = J_j(t+\Delta t) - J_j(t) = M_{j-1}^{(J)}(t) + B_j^{(J)}(t) - D_j^{(J)}(t) = \Big\{J_{j-1}(t)\beta_{j-1}(t) + J_j(t)\gamma_j^{(J)}(t)\Big\}\Delta t + e_j^{(J)}(t)\Delta t, \quad j = 1, \ldots, k_2 - 1 \tag{4.11}$$

where $\gamma_i^{(I)}(t) = b_i^{(I)}(t) - d_i^{(I)}(t)$, $\gamma_j^{(J)}(t) = b_j^{(J)}(t) - d_j^{(J)}(t)$.

In the above equations, the random noises $\{e_i^{(I)}(t)\Delta t,\ e_j^{(J)}(t)\Delta t\}$ are derived by subtracting the conditional expected numbers from the random transition variables respectively. Obviously, these random noises are linear combinations of Poisson and multinomial random variables. These random noises have expected value zero and are uncorrelated with the state variables $\{I_i(t),\ i = 1, \ldots, k_1 - 1,\ J_j(t),\ j = 1, \ldots, k_2 - 1\}$. It can also be shown that to the order of $o(\Delta t)$, these random noises are un-correlated with one another and have variances given by:

$$Var\left\{e_i^{(I)}(t)\Delta t\right\} = EI_{i-1}(t)\alpha_{i-1}(t)\Delta t + EI_i(t)\left[b_i^{(I)}(t) + d_i^{(I)}(t)\right]\Delta t + o(\Delta t)$$

for $i = 1, \ldots, k_1 - 1$,

$$Var\left\{e_1^{(J)}(t)\Delta t\right\} = EI_1(t)\beta_0(t)\Delta t + EJ_1(t)\left[b_1^{(J)}(t) + d_1^{(J)}(t)\right]\Delta t + o(\Delta t)$$

$$Var\left\{e_j^{(J)}(t)\Delta t\right\} = EJ_{j-1}(t)\beta_{j-1}(t)\Delta t + EJ_j(t)\left[b_j^{(J)}(t) + d_j^{(J)}(t)\right]\Delta t + o(\Delta t)$$

for $j = 2, \ldots, k_2 - 1$,

where $I_0(t) = N(t)$.

The Expected Numbers

Let $u_I(i, t) = E[I_i(t)]$ and $u_J(j, t) = E[J_j(t)]$ denote the expected numbers of $I_i(t)$ and $J_j(t)$ respectively and write $u_I(0, t) = u_J(0, t) = N(t)$. Using equations (4.9)-(4.11), we have the following differential equations for these expected numbers:

$$\frac{d}{dt}u_I(i,t) = u_I(i,t)\gamma_i^{(I)}(t) + u_I(i-1,t)\alpha_{i-1}(t), \quad i = 1, \ldots, k_1 - 1$$

$$\frac{d}{dt}u_J(1,t) = u_J(1,t)\gamma_1^{(J)}(t) + u_J(1,t)\beta_0(t)$$

$$\frac{d}{dt}u_J(j,t) = u_J(j,t)\gamma_j^{(J)}(t) + u_J(j-1,t)\beta_{j-1}(t), \quad j = 1, \ldots, k_2 - 1$$

The solutions of the above equations are:

$$u_I(1,\mathrm{t}) = \int_{t_0}^{t} \lambda_I(x) e^{\int_x^t \gamma_1^{(I)}(z)dz} dx, \; u_J(1,\mathrm{t}) = \int_{t_0}^{t} \lambda_J(x) e^{\int_x^t \gamma_1^{(J)}(z)dz} dx \qquad (4.12)$$

$$u_I(\mathrm{i},\mathrm{t}) = \int_{t_0}^{t} u_I(i-1,x) e^{\int_x^t \gamma_i^{(I)}(z)dz} dx, \quad i = 2, \ldots, k_1 - 1 \qquad (4.13)$$

$$u_J(\mathrm{i},\mathrm{t}) = \int_{t_0}^{t} u_J(j-1,x) e^{\int_x^t \gamma_j^{(J)}(z)dz} dx, \quad j = 2, \ldots, k_2 - 1 \qquad (4.14)$$

The Probability Distribution of State Variables and Augmented State Variables

Although $T(t)$ is not Markov, the random vector $\tilde{X}(t)\{, t \geq t_0\}$ is Markov with continuous time. To derive the transition probability of this process, denote by $f(x, y : N,$

p_1, p_2) is the density at (x, y) of the multinomial distribution $ML(N; p_1, p_2)$ with parameters ($N; p_1, p_2$) and $h(x : \lambda)$ the density at x of the Poisson distribution with mean λ. Then, using the probability distributions given by equations (4.3)-(4.8), to order of $o(\Delta t)$ the transition probability of this Markov process is:

$$P\{\tilde{X}(t+\Delta t)\big|\tilde{X}(t)\} = \prod_{u=1}^{k_1-1}\Big\{\sum_{m_u=0}^{I_u(t)}\sum_{i_u=0}^{I_u(t)-m_u} h[a(m_u, i_u; t); I_{u-1}(t)\alpha_{u-1}(t)\Delta t] \times$$

$$f[m_u, i_u; I_u(t), b_u^{(I)}(t)\Delta t, d_u^{(I)}(t)\Delta t]\Big\} \times \Big\{\sum_{r_1=0}^{J_1(t)}\sum_{j_1=0}^{J_1(t)-r_1} h[b(r_1, j_1; t); I_1(t)\beta_0(t)\Delta t] \times$$

$$f[r_1, j_1; J_1(t), b_1^{(J)}(t)\Delta t, d_1^{(J)}(t)\Delta t]\Big\} \times$$

$$\prod_{v=2}^{k_2-1}\Big\{\sum_{r_v=0}^{J_v(t)}\sum_{j_v=0}^{J_v(t)-r_v} h[b(r_v, j_v; t); J_{v-1}(t)\beta_{v-1}(t)\Delta t] \times$$

$$f[r_v, j_v; J_{uv}(t), b_v^{(J)}(t)\Delta t, d_v^{(J)}(t)\Delta t]\Big\} \qquad (4.15)$$

where $I_0(t) = N(t)$, $a(m_u, i_u; t) = I_u(t+\Delta t) - I_u(t) - m_u + i_u$, $u = 1, \ldots, k_1 - 1$ and where

$b(r_v, j_v; t) = J_v(t+\Delta t) - J_v(t) - r_v + j_v$, $v = 1, \ldots, k_2 - 1$.

The above transition probability and hence the probability distribution of $\tilde{X}(t)$ is too complicated to be of much use. For implementing the Gibbs sampling procedure to estimate parameters and to predict state variables, we use data augmentation method to expand the model. Thus, we define the augmented state variables $\tilde{U}(t) = \{B_i^{(I)}(t), D_i^{(I)}(t), i = 1, \ldots, k_1 - 1;\ B_j^{(J)}(t), D_j^{(J)}(t), j = 1, \ldots, k_2 - 1\}$ (In what follows we will refer these variables as the transition variables, unless otherwise stated.)

Put $\tilde{Z}(t) = \{\tilde{X}(t)', \tilde{U}(t)'\}'$, then $\{\tilde{Z}(t), t \geq t_0\}$ is Markov with continuous time. Using the probability distributions of the transition random variables given by equations (4.3)-(4.8), the transition probability

$$P\{\tilde{Z}(t+\Delta t)\big|\tilde{Z}(t)\} = P\big\{\tilde{X}(t+\Delta t)\big|\tilde{X}(t), \tilde{U}(t)\big\} \times \big\{\tilde{U}(t)\big|\tilde{X}(t)\big\} \qquad (4.16)$$

where $P\{\tilde{U}(t)|\tilde{X}(t)\} = \prod_{i=1}^{k_1-1} f\{B_i^{(I)}(t), D_i^{(I)}(t); I_i(t), b_i^{(I)}(t)\Delta t, d_i^{(I)}(t)\Delta t\} \times$

$\prod_{j=1}^{k_2-1} f\{B_j^{(J)}(t), D_j^{(J)}(t); J_j(t), b_j^{(J)}(t)\Delta t, d_j^{(J)}(t)\Delta t\}$ (4.17)

and

$$P\{\tilde{X}(t+\Delta t)|\tilde{X}(t), \tilde{U}(t)\} =$$

$\prod_{i=1}^{k_1-1} h\{u_I(i,t); I_{i-1}(t)\alpha_{i-1}(t)\Delta t\} \times h\{u_J(1,t); I_1(t)\beta_0(t)\Delta t\} \times$

$\prod_{j=2}^{k_2-1} h\{u_J(j,t); J_{j-1}(t)\beta_{j-1}(t)\Delta t\}$ (4.18)

where $u_I(i,t) = I_i(t+\Delta t) - I_i(t) - B_i^{(I)}(t) + D_i^{(I)}(t)$ for $i = 1, \ldots, k_1 - 1$ and $u_J(i,t) = J_j(t+\Delta t) - J_j(t) - B_j^{(J)}(t) + D_j^{(J)}(t)$ for $j = 1, \ldots, k_2 - 1$.

The probability distribution given by equation (4.15) will be used to derive estimates and predicted numbers of state variables. This is discussed in Section 4.6.

A Statistical Model and the Probability Distribution of the Number of Detectable Tumors for HCC

The data available for modeling carcinogenesis are usually cancer incidence over different time periods. In this section, we will derive the cancer incidence function for two pathways discussed in Section 4.3.

The Probability Distribution of the Number of Detectable Tumors for HCC

To derive the probability distribution of time to tumors, one needs to find the probability distribution of $T(t)$. For deriving this probability distribution, we observe that malignant cancer tumors arise by clonal expansion from primary I_{k_1} cells and primary J_{k_2} cells, where primary I_{k_1} cells are I_{k_1} cells derived from I_{k_1-1}cells by mutation of I_{k_1-1} cells and primary J_{k_2} cells are J_{k_2} cells derived from J_{k_2}cells by mutation of J_{k_2} cells.

Let $P_T^{(I)}(s,t)$ ($P_T^{(J)}(s,t)$) be the probability that a primary $I_{k_1}(J_{k_2})$ cancer cell at time s develops into a detectable cancer tumor at time t. Let $T_i(t)$ be the number of cancer tumors derived from the i−th pathway. Then, to order of $o(\Delta t)$, the conditional probability distribution of $T_1(t)$ given $I_{k_1-1}(s)$, $\{s \le t\}$ is Poisson with mean $\mu_1(t)$ independently of $T_2(t)$, where

$$\mu_1(t) = \int_{t_0}^{t} I_{k_1-1}(s)\alpha_{k_1-1}(s)P_T^{(I)}(s,t)\,ds \tag{4.19}$$

Similarly, to order of $o(\Delta t)$, the conditional probability distribution of $T_2(t)$ given $J_{k_2-1}(s)$, $\{s \le t\}$ is Poisson with mean $\mu_2(t)$ independently of $T_1(t)$, where

$$\mu_2(t) = \int_{t_0}^{t} J_{k_2-1}(s)\beta_{k_2-1}(s)P_T^{(J)}(s,t)\,ds \tag{4.20}$$

Let $Q_i(j)$ (i = 1, 2) be defined by:

$$Q_i(j) = E\left\{e^{-\mu_i(t_{j-1})} - e^{-\mu_i(t_j)}\right\} = E\{e^{-\mu_i(t_{j-1})}(1 - e^{-R_i(t_{j-1},t_j)})\} \tag{4.21}$$

where $R_i(t_{j-1},\ t_j) = \mu_i(t_{j-1}) - \mu_i(t_j)$.

Then $Q_i(j)$ is the probability that cancer tumors would develop during the j−th age group by the i−th pathway. Since cancer tumors develop if and only if at least one of the two pathways yield cancer tumors, the probability that a normal person at time t_0 will develop cancer tumors during (t_{j-1}, t_j] is given by $Q_T(j)$, where

$$Q_T(j) = 1 - [1 - Q_1(j)][1 - Q_2(j)] = Q_1(j) + Q_2(j) - Q_1(j)Q_2(j) \tag{4.22}$$

For practical applications, we observe that to order of $o(\alpha_{k_1-1}(t))$ and $o(\beta_{k_2-1}(t))$ respectively, the $\mu_i(t)$ in $Q_i(j)$ are approximated by

$$\mu_1(t) \sim \int_{t_0}^{t} E\left[I_{k_1-1}(s)\right]\alpha_{k_1-1}(s)P_T^{(I)}(s,t)ds \tag{4.23}$$

$$\mu_2(t) \sim \int_{t_0}^{t} E\left[J_{k_2-1}(s)\right]\beta_{k_2-1}(s)P_T^{(J)}(s,t)ds \tag{4.24}$$

Similarly, it can readily be shown that to the order of $Min\{o(\alpha_{k_1-1}(t)),\ o(\beta_{k_2-1}(t))\}$, $Q_T(j) = Q_1(j) + Q_2(j)$.

To further simplify the calculation of $Q_T(j)$, we observe that in studying human cancers, one time unit (i.e., $\Delta t = 1$) is usually assumed to be 3 months or 6 months or longer. In these cases, one may practically assume $P_T^{(I)}(s,t) \sim 1$ and $P_T^{(I)}(s,t) \sim 1$ if $t - s \geq 1$.

A Statistical Model for Cancer Incidence Data

Let y_j be the observed number of cancer cases during $(t_{j-1},\ t_j]$ given n_j normal people at risk at t_0 for cancer. We assume that each individual develops colon cancer tumor by the same mechanism independently of one another. Then for each normal person at time t_0, the probability that this individual would develop liver cancer tumor during the j-th age group $(t_{j-1},\ t_j\ ]$ is given by $Q_T(j)$. It follows that the probability distribution of y_j given that n_j people are at risk for liver cancer at time t_0 is:

$$y_j \sim Poisson\{\tau_j\} \tag{4.25}$$

where $\tau_j = n_j Q_T(j)$.

Notice that to the order of $Max\{o(\alpha_{k_1-1}(t)),\ o(\beta_{k_2-1}(t))\}$, τ_j (and hence y_j) depends on the stochastic model of liver carcinogenesis through the expected number $\{E[I_{k_1-1}(t)],\ E[J_{k_2-1}(t)]\}$ of $\{I_{k_1-1}(t),\ J_{k_2-1}(t)\}$ and the parameters $\{\alpha_{k_1-1}(t),\ \beta_{k_2-1}(t)\}$ over the time period $(t_{j-1},\ t_j]$.

A Stochastic Model for Hereditary Liver Cancer (HBL)

Hirschman, Pollock, and Tomlinson (2005) has discovered the mutated APC gene is a HBL-associated germ line mutation, it follows that the germ line cells (eggs and sperms)

may carry the mutant APC allele so that liver cancer may occur at birth or at a young age.

Table 17

Liver Cancer Data from SEER (2008)

Age Group	Number of People at Risk	Observed Liver Cancer Cases	Predicted Liver Cancer Cases
0	12,069,564	126	126
0-4	46,971,230	182	169
5-9	58,891,282	49	58
10−14	60,616,195	54	48
15−19	61321678	62	52
20−24	62,699,020	98	91
25−29	65,530,352	148	151
30−34	65,677,509	220	238
35−39	61,351,859	406	440
40−44	56,337,973	814	871
45−49	50,065,865	1720	1688
50−54	44,430,452	2665	2601
55−59	38,337,957	3083	2894
60−64	32,444,455	3415	3383
65−69	27,712,647	3725	3762
70−74	22,840,369	3948	3770
75−79	17,813,869	3541	3397
80−84	11,981,840	2275	2256

Table 17

Liver Cancer Data From SEER (2008) (continued)

aa Genotype	Aa Genotype	AA Genotype	I pathway	J pathway
126	0	0	0	0
0	169	0	0	0
0	57	0	0	1
0	32	1	2	13
0	11	1	5	35
0	4	2	11	74
0	1	3	23	124
0	0	4	39	195
0	0	5	58	377
0	0	6	81	784
0	0	7	104	1577
0	0	7	130	2464
0	0	7	153	2734
0	0	8	172	3203
0	0	8	192	3562
0	0	8	203	3559
0	0	7	200	3190
0	0	5	168	2083

As shown in Table 17, it is clear shown the liver cancer incidence data from NCI/NIH SEER data gives very high cancer rates at birth and before 5 years old.

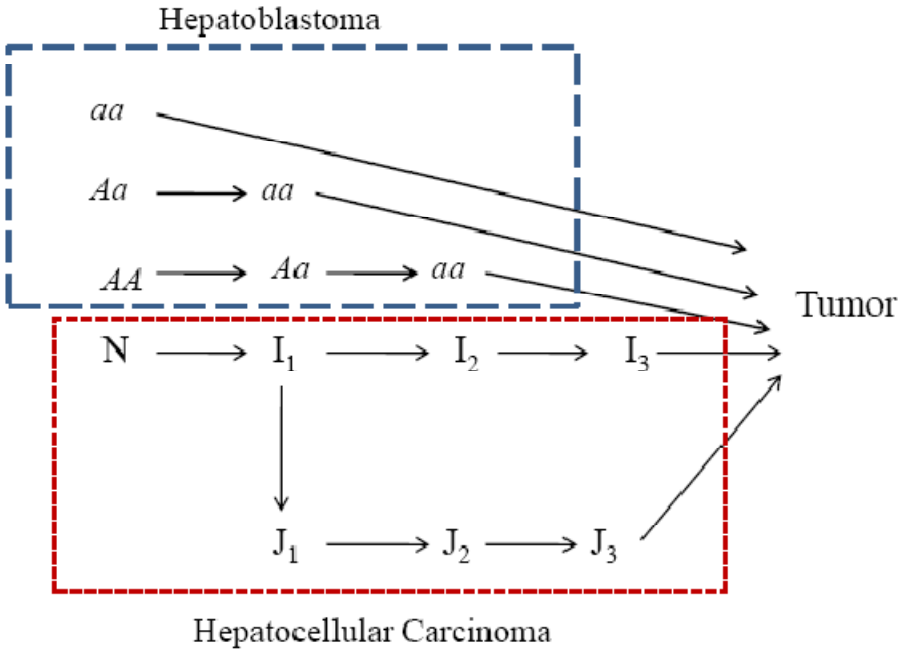


Figure 18. A Multiple-Pathway Model for HBL and HCC

To account for inherited cancer cases in the stochastic model for liver cancer, we take hereditary segregation into consideration for the population. Let p be the frequency of the mutated APC gene (denoted by a) in the population, so that $q = 1-p$ is the frequency of the normal APC allele in the population (denoted by A). Assume that population is very large and that mating (marriage) between people is random respect to the HBL-related APC locus. Then by the Hardy-Weinberg law (Crow & Kimura, 1970; Tan, 2002) the frequency of individuals with genotype aa, Aa and AA at the embryo stage in the population are given by p^2, $2pq$ and q^2, respectively. The individuals carrying aa (both APC alleles are mutated) genotype will developed liver cancer (HBL) at birth. Similarly, the individuals carrying one mutated APC allele (Aa genetype) require mutation $A \rightarrow a$

during pregnancy to become individuals with genotype *aa* at birth. We use α to represent the probability of $A \rightarrow a$ transition. It follows that the probability of individuals with genotype *Aa* at birth is $1-\alpha$. α is usually small ($10^{-3} \sim 10^{-2}$). On the other hand, for the individuals with both normal APC alleles (*AA* genotype), the probability of two $A \rightarrow a$ transitions to *aa* genotype is very small, we ignore those transitions in our model, and assume that individuals with *AA* alleles are still normal at birth. However, HBL can be developed through mutation of APC gene after born, therefore, there are two pathways associated with HBL: for individuals born with *Aa* genotype, the HBL is developed through one stage ($Aa \rightarrow aa$); for individuals born with *AA* genotype, the HBL is developed through two stages ($AA \rightarrow Aa \rightarrow aa$). The overall pathways of developing liver cancer, including inherited HBL and non-hereditary liver cancer, are schematically shown in Figure 18.

Because the number of stages and hence the probability distribution of time to cancer tumors for each individual depend on the genotype of the individual at the embryo stage, we let $A_j^{(i)}(t)$ (j = 1, 2) denote the number of A_j cells at time *t* in people who have genotype *i* at the embryo state, with *i* =1for genotype *aa*, *i* = 2 for genotype *Aa*, and *i* =3 for genotype *AA*, respectively. Since the number of stages, through which HBL is developed, depends on genotype at the embryo state, it follows that j = 1 for $A_j^{(2)}(t)$, and j = 1, 2 for $A_j^{(3)}(t)$. Similarly, we write $\{B_j^{A^{(i)}}(t), D_j^{A^{(i)}}(t), M_j^{A^{(i)}}(t),\ j = 1, 2, T^{A^{(i)}}(t)\}$ for number of new proliferated $A_j^{(i)}$ cells, number of deaths, number of mutants of $A_j^{(i)}$ cells at time t for genotype *i*, and number of tumors developed from genotype *i*.

The One-Stage Model and Mathematical Analysis

As shown in Figure 18, if an individual has genotype *Aa* at embryo stage, then one stage is required to developed HBL (*aa* genotype). This is a one stage model given by $A_1^{(2)}$ (*Aa* genotype) → $A_2^{(2)}$ (*aa* genotype) → Tumor. For this person, the probability is α that he/she would develop tumor at birth, the probability that he/she will remain *Aa* type at birth is $1-\alpha$.

Because all stem cells at embryo stage are $A_1^{(2)}$ (with *Aa* genotype) cells in this person, the number of $A_1^{(2)}$ cells can only be generated by stochastic birth and death of these cells. It follows that the stochastic equation and the stochastic difference equation for the stage variable $A_1^{(2)}(\mathrm{t})$ are given respectively by:

$$A_1^{(2)}(t+\Delta t) = A_1^{(2)}(t) + B_1^{A^{(2)}}(t) - D_1^{A^{(2)}}(t) \tag{4.26}$$

$$dA_1^{(2)}(t) = A_1^{(2)}(t+\Delta t) - A_1^{(2)}(t) = B_1^{A^{(2)}}(t) - D_1^{A^{(2)}}(t) = A_1^{(2)}(t)\gamma^{A^{(2)}}(t)\Delta t + e^{A^{(2)}}(t)\Delta t \tag{4.27}$$

where $e^{A^{(2)}}(t)\Delta t = \left[B_1^{A^{(2)}}(t) - A_1^{(2)}(t)b_1^{A^{(2)}}(t)\Delta t\right] - [D_1^{A^{(2)}}(t) - A_1^{(2)}(t)d_1^{A^{(2)}}(t)\Delta t]$.

In the above equations, the conditional distributions of $\{B_1^{A^{(2)}}(t), D_1^{A^{(2)}}(t)\}$ given $A_1^{(2)}(t)$ is multinomial. It follows that $E\left[e^{A^{(2)}}(t)\right] = 0$, and $Var\left[e^{A^{(2)}}(t)\right] = E\left[A_1^{(2)}(t)\right]\left\{b_1^{A^{(2)}}(t) + d_1^{A^{(2)}}(t)\right\}\Delta t + o(\Delta t)$, and $e^{A^{(2)}}(t)$ is uncorrelated with $\{A_1^{(2)}(t), T^{A^{(2)}}(t)\}$.

Let *f(x; N, p)* denote the density of binomial distribution with parameters *(N,p)*. From equation (4.26), we obtain the transition probability of the Markov process $A_1^{(2)}(t)$ as, to the order of $o(\Delta t)$:

$$P\left\{A_1^{(2)}(t+\Delta t)=j_2 \middle| A_1^{(2)}(t)=i_2\right\}=\sum_{u=0}^{i_2} f(u;i_2,b_1^{A^{(2)}}(t)\Delta t)\times f(i_2-j_2+u,i_2-u,\frac{d_1^{A^{(2)}}(t)\Delta t}{1-b_1^{A^{(2)}}(t)\Delta t}) \quad (4.28)$$

With $A_1^{(2)}(t)$ generated by the above stochastic equation, the number of tumor at time *t* given number of primary A_1 cells has Poisson distribution as follow:

$$T^{A^{(2)}}(t)|\left\{A_1^{(2)}(s),\ t_0<s<t\right\}\sim Poisson\{\xi_1(t)\} \quad (4.29)$$

where $\xi_1(t)=\int_{t_0}^{t} A_1^{(2)}(x)\omega_1(x)P_T(x,t)dx$.

Let $Q_{Aa}(j)$ denote the probability that an individual with genotype *Aa* at the embryo stage develops cancer tumor during period $(t_{j-1}, t_j]$ (j > 1). For individuals with genotype *Aa* at the embryo stage, let $P_{Aa}(j)$ be the probability that this individual develops cancer tumor during the period $(t_{j-1}, t_j]$ (j > 1) via genotype *Aa* at birth. Then from equation (4.29) and Figure 18, and

$$Q_{Aa}(j)=(1-\alpha)P_{Aa}(j) \quad (4.30)$$

where $P_{Aa}(j)$ is given by:

$P_{Aa}(j)=E\{e^{-\xi_1(t_{j-1})}-e^{-\xi_1(t_j)}\}$, where $\xi_1(t)$ is given above.

Let $u_1^{A^{(2)}}(t)=E[A_1^{(2)}(t)|Aa\ at\ t_0]$ be the expected number of $A_1^{(2)}(t)$ for people with genotype *Aa* at the embryo stage. Assuming $\omega_1(t)=\omega_1$ is very small, then $P_{Aa}(j)\approx e^{-\omega_1\eta_1(t_{j-1})}-e^{-\omega_1\eta_1(t_j)}$, where $\eta_1(t)=\int_{t_0}^{t} u_1^{A^{(2)}}(x)P_T(x,t)dx$.

From equation (4.27), the differential equation for $u_1^{A^{(2)}}(t)$ ($t > t_0$) is $\frac{d}{dt}u_1^{A^{(2)}}(t)=u_1^{A^{(2)}}(t)\gamma^{A^{(2)}}(t)$. So that $u_1^{A^{(2)}}(t)=u_1^{A^{(2)}}(t_0)e^{\int_{t_0}^{t}\gamma^{A^{(2)}}(x)dx}$.

If $\gamma^{A^{(2)}}(t)=\gamma^{A^{(2)}}$, then $u_1^{A^{(2)}}(t)=u_1^{A^{(2)}}(t_0)\exp\{\gamma^{A^{(2)}}(t-t_0)\}$, and hence

$$\eta_1(t) = u_1^{A^{(2)}}(t_0) \int_{t_0}^{t} \exp\{\gamma^{A^{(2)}}(x - t_0)\} P_T(x,t)dx$$

and

$$P_{Aa}(j) \approx e^{-\omega_1 u_1^{A^{(2)}}(t_0)\varphi_1(t_{j-1})} - e^{-\omega_1 u_1^{A^{(2)}}(t_0)\varphi_1(t_j)} \tag{4.31}$$

where $\varphi_1(t) = \int_{t_0}^{t} \exp\{\gamma^{A^{(2)}}(x - t_0)\} P_T(x,t)dx$. If $P_T(x,t) = 1$for $t > x$, then $\varphi_1(t) = \frac{1}{\gamma^{A^{(2)}}}[e^{\gamma^{A^{(2)}}(t-t_0)} - 1]$ if $\gamma^{A^{(2)}} \neq 0$.

The Two-Stage Model and Mathematical Analysis

Assume that an individual has genotype AA at the embryo stage (a normal person). Then for this individual all stem cells are normal cells at the embryo stage in which case cancer tumors are derived by a two-stage model: $A_0^{(3)}$ (*AA genotype)* → $A_1^{(3)}$*(Aa genotype)* → $A_2^{(3)}$ *(aa genotype)* → *Tumor* (as shown in Figure 18). The mutation rate between *AA* and *Aa* is $\omega_0(t)$, and the mutation rate between *Aa* and *aa* is $\omega_1(t)$.

For this model, we use $\{B_j^{A^{(3)}}(t), D_j^{A^{(3)}}(t),\ M_{j-1}^{A^{(3)}}(t),\ A_1^{(3)}(t),\ T^{A^{(3)}}(t)\}$ to represent random variables of number of birth, death, mutants, A_j cells and number of cancer tumors developed from this two-stage pathway. The stochastic equations and the stochastic difference equation for $A_1^{(3)}$ cells are given as follow:

$$A_1^{(3)}(t + \Delta t) = A_1^{(3)}(t) + M_0^{A^{(3)}}(t) + B_1^{A^{(3)}}(t) - D_1^{A^{(3)}}(t) \tag{4.32}$$

$$dA_1^{(3)}(t) = A_1^{(3)}(t + \Delta t) - A_1^{(3)}(t) = M_0^{A^{(3)}}(t) + B_1^{A^{(3)}}(t) - D_1^{A^{(3)}}(t) =$$

$$A_1^{(3)}(t)\gamma^{A^{(3)}}(t)\Delta t + e^{A^{(3)}}(t)\Delta t \tag{4.33}$$

where $e^{A_1^{(3)}}(t)\Delta t = \left[B_1^{A^{(3)}}(t) - A_1^{(3)}(t) b_1^{A^{(3)}}(t)\Delta t\right] - [D_1\ ^{A^{(3)}}(t) - A_1^{(3)}(t) d_1^{A^{(3)}}(t)\Delta t]$.

The normal cells (with genotype *AA*) increase or decrease only by stochastic birth-death process. The stochastic equation and difference equation are the same as (4.26) and (4.27). Similarly, the conditional distributions of the transition variables $\{B_1^{A^{(3)}}(t), D_1^{A^{(3)}}(t), M_0^{A^{(3)}}(t)\}$ given $A_j^{(3)}(t)$ are as follow:

$$M_0^{A^{(3)}}(t) \sim Poisson\{N(t)\omega_0(t)\Delta t\} \tag{4.34}$$

$$\left\{B_1^{A^{(3)}}(t), D_1^{A^{(3)}}(t)\right\} \mid A_1^{(3)}(t) \sim Multinomial\ \{A_1^{(3)}(t);\ b_1^{A^{(3)}}(t)\Delta t, d_1^{A^{(3)}}(t)\Delta t\} \tag{4.35}$$

Then it follows that $E\left[e^{A_1^{(3)}}(t)\Delta t\right] = 0$, and $cov\left\{e^{A_1^{(3)}}(t)\Delta t, e^{A_2^{(3)}}(t)\Delta t\right\} = o(\Delta t)$ and $Var\left[e^{A_1^{(3)}}(t)\Delta t\right] = E\left[A_1^{(3)}(t)\right]\left\{b_1^{A^{(3)}}(t) + d_1^{A^{(3)}}(t)\right\}\Delta t + \delta_1 E\left[A_1^{(3)}(t)\right]\omega_1(t)\Delta t + o(\Delta t)$. (4.36)

Similarly, let *f(x; N, p)* denote the density at *x* of a binomial distribution with parameter (*N,p*). Using the stochastic equation given by (4.26) and (4.32) and the probability distributions given by equations (4.27) and (4.33), we obtain, to the order of *o(Δt)*:

$$P\left\{A_1^{(3)}(t+\Delta t) = j_1 \middle| A_1^{(3)}(t) = i_1\right\} =$$

$$\sum_{u_1=0}^{i_1}\sum_{v_1=0}^{i_1-u_1} g\left(u_1, v_1; i_1, b_1^{A^{(3)}}(t)\Delta t, d_1^{A^{(3)}}(t)\Delta t\right) \times h(j_1 - i_1 - u_1 + v_1; \omega_0(t)\Delta t) \tag{4.37}$$

As in last section, we have:

$$T^{A^{(3)}}(t) \left| \left\{A_1^{(3)}(s),\ t_0 < s < t\right\} \sim Poisson\{\xi_2(t)\}\right. \tag{4.38}$$

where $\xi_2(t) = \int_{t_0}^{t} A_1^{(3)}(x)\omega_1(x)P_T(x,t)dx$.

Let $Q_{AA}(j)$ denote the probability that an individual with genotype *AA* at the embryo stage develops cancer tumor during period $(t_{j-1}, t_j]$ $(j > 1)$. For individuals with genotype *AA* at the embryo stage, let $P_{AA}(j)$ be the probability that this individual develops cancer tumor during the period $(t_{j-1}, t_j]$ $(j > 1)$ via genotype *Aa* or *AA* at birth, and α_{AA} be the probability of individual born with *Aa* genotype but with *AA* genotype at embryo stage. Similarly as in Section 4.4.1, $Q_{AA}(j) = \alpha_{AA}P_{Aa}(j) + (1 - \alpha_{AA})P_{AA}(j)$, where we usually assume $\alpha_{AA} \approx 0$, and $P_{Aa}(j)$ is given in (4.30) and $P_{AA}(j)$ is given by:

$$P_{AA}(j) = E\{e^{-\xi_2(t_{j-1})} - e^{-\xi_2(t_j)}\} \tag{4.39}$$

Let $u_1^{A^{(3)}}(t|u), u = 1,2$ be the expected number of $A_1^{(3)}(t)$ for people with genotype *Aa* (u=1) an *AA* (u=2) at birth in people who have genotype *AA* at the embryo stage, respectively. If $\omega_1(t) = \omega_1$ is very small, then, as in the section 4.4.1, we have, to the order of $o(\omega_1)$:

$P_{AA}(j) \approx e^{-\omega_1\eta_2(t_{j-1})} - e^{-\omega_1\eta_2(t_j)}$, where $\eta_2(t) = \int_{t_0}^{t} u_1^{A^{(3)}}(x)P_T(x,t)dx$.

Let's derive $u_1^{A^{(3)}}(t)$ as shown in the last section, we have following equations:

$$u_1^{A^{(3)}}(t|1) = u_1^{A^{(3)}}(t_0|1)\mathrm{e}^{\int_{t_0}^{t}\gamma_1^{A^{(3)}}(x)\mathrm{dx}} \tag{4.40}$$

$$u_1^{A^{(3)}}(t|2) = u_1^{A^{(3)}}(t_0|2)\int_{t_0}^{t}\mathrm{e}^{\int_{t_0}^{x}\gamma_1^{A^{(3)}}(y)\mathrm{dy}}\,\omega_0(\mathrm{x})\mathrm{dx} \tag{4.41}$$

where $u_1^{A^{(3)}}(t_0|i) = E\left[A_1^{(3)}(t_0)\middle| i = 1\ for\ Aa\ at\ t_0, i = 2\ for\ AA\ at\ t_0\right], i = 1,2.$

(Notice that $u_1^{A^{(3)}}(t_0|2) = E\left[A_1^{(3)}(t_0)\middle|AA\ at\ t_0\right] \gg u_1^{A^{(3)}}(t_0|1) = E\left[A_2^{(3)}(t_0)\middle|Aa\ at\ t_0\right]$, and $u_1^{A^{(3)}}(t_0|1) = 0$) if $\alpha_{AA} = 0$).

If $\{\omega_0(t) = \omega_0, \omega_1(t) = \omega_1, \gamma_1^{A^{(3)}}(t) = \gamma_1^{A^{(3)}}\}$, then (4.40) and (4.41) reduce to:

$$u_1^{A^{(3)}}(t|1) = u_1^{A^{(3)}}(t_0|1)\mathrm{e}^{\gamma_1^{A^{(3)}}(t-t_0)} \tag{4.42}$$

$$u_1^{A^{(3)}}(t|2) = u_1^{A^{(3)}}(t_0|2)\omega_0 \int_{t_0}^{t} e^{\gamma_1^{A^{(3)}}(x-t_0)}\,dx = u_1^{A^{(3)}}(t_0|2)e^{\gamma_1^{A^{(3)}}(x-t_0)}\omega_0/\gamma_1^{A^{(3)}}.$$

(4.43)

Now $\eta_2(t)$ becomes $\varphi_2(t)$, where $\varphi_2(t) = \int_{t_0}^{t} \phi_1(t)P_T(x,t)dx$ and $\phi_1(t)$ is given

$$\phi_1(t) = \alpha_{AA}u_1^{A^{(3)}}(t|1) + (1-\alpha_{AA})u_1^{A^{(3)}}(t|2) \tag{4.44}$$

Then to the order of $o(\omega_1)$:

$$P_{AA}^{(3)}(j) \approx e^{-\omega_1\phi_2(t_{j-1})} - e^{-\omega_1\phi_2(t_j)} \tag{4.45}$$

Notice that if $P_T(x,t) = 1$ for $t > x$, then $\phi_2(t)$ reduces to $\psi_2(t)$:

$$\psi_2(t) = \frac{\alpha_1}{\gamma_1^{A^{(3)}}}u_1^{A^{(3)}}(t_0|1)\{e^{\gamma_1^{A^{(3)}}(t-t_0)} - 1\} + \frac{(1-\alpha_1)}{\left(\gamma_1^{A^{(3)}}\right)^2}u_1^{A^{(3)}}(t_0|2)\{e^{\gamma_1^{A^{(3)}}(t-t_0)} - 1\}.$$

(4.46)

Statistical Model and the Probability Distribution of the Number of Detectable Tumors for Hereditary Liver Cancer

The data available for modeling carcinogenesis are usually cancer incidence over different time periods. For example, the SEER data of NCI/NIH for human cancers are given by $\{(y_j, n_j), j = 1, \ldots, n\}$, where y_j is the number of cancer cases during the j−th age group and n_j is the number of normal people who are at risk for cancer and from whom y_j of them have developed cancer during the age group. Given in Table 17 are the SEER data for human liver cancer. From this data set, notice that there are a large number of cancer cases before 10 year-old, which implies a large number of inherited cancer cases. In this section, we will develop a statistical model for the data set.

The Probability Distribution of the Number of Detectable Tumors for Different Genotypes

To incorporate inherited cancer cases, among the n_j people at risk for liver cancer, let n_{1j} be the number of individuals who have genotype *aa* at the embryo stage, n_{2j} be the number of individuals with genotype *Aa* at the embryo stage and n_{3j} be the number of individuals with genotype *aa* at the embryo stage. Then, from results in Section 4.5, the conditional probability distribution of (n_{2j}, n_{3j}) given n_j is multinomial with parameters $\{n_j; 2pq, q^2\}$. It follows that $n_{2j} \mid n_j \sim Binomial\ \{n_j, 2pq\}$.

Because y_0 is the number of cancer cases at birth, which derive from individuals who have genotype *aa* at the embryo stage, and we consider all individuals who have genotype *aa* at the embryo stage will develop HBL at birth, so that expect number of liver cancer cases at birth with genotype *aa* at the embryo stage is $n_0 p^2$.

To derive the probability distribution of y_j ($j \geq 1$) in the j-th age group, let y_{2j} be number of cancer cases generated with genotype *Aa* at embryo stage. The population who have *Aa* genotype at embryo stage is n_{2j}, and n_{2j} given n_j is binomial with parameter $\{n_j, 2pq\}$, and as we have shown in section 4.5.1, each individual with *Aa* genotype at embryo stage develops liver cancer though one-stage pathway with probability $Q_{Aa}(j)$ (see (4.30)). Then we can have:

$$Y_{2j} | n_j \sim Poisson\ \{\chi_0\} \tag{4.47}$$

where $\chi_0 = 2pqn_jQ_{Aa}(j)$.

Then $y_{3j} = y_j - y_{1j} - y_{2j}$ is the number of cancer case generated by the $n_{3j} = n_j - n_{1j} - n_{2j}$ people with genotype *AA* at the embryo stage. As we discussed in the section 4.5.1, individuals who have genotype *AA* at the embryo stage usually born with

normal, liver cancer developed either through two-stage $AA \rightarrow Aa \rightarrow aa \rightarrow Tumor$, or three-stage $N \rightarrow I_1 \rightarrow I_2 \rightarrow I_3 \rightarrow Tumor$, or four-stage $N \rightarrow I_1 \rightarrow J_1 \rightarrow J_2 \rightarrow J_3 \rightarrow Tumor$ as shown in Figure 18. The probability of an individual with genotype AA at the embryo stage develops cancer tumor at j-th age group is $Q_{AA}(j)$ through two-stage pathway, $Q_1(j)$ through three stage pathway, and $Q_2(j)$ through 4-stage pathway, as shown in (4.40) and (4.21), respectively. Then the probability of an individual developing liver cancer through at least one of above three pathways is:

$$Q_T^{AA}(j) = 1 - \big(1 - Q_{AA}(j)\big)\big(1 - Q_1(j)\big)\big(1 - Q_2(j)\big) \tag{4.48}$$

Similar as in section 4.4.1, it can readily be shown that to the order of $Min\{o(\omega_1(t)), o(\alpha_{k_1-1}(t)), o(\beta_{k_2-1}(t))\}$, $Q_T^{AA}(j) = Q_{AA}(j) + Q_1(j) + Q_2(j)$.

Combine all genotypes together, the conditional distribution of y_j given $\{n_{ij}, i = 2, 3; n_j\}$ is Poisson with mean $Q_T^{ALL}(j) = 2pqn_jQ_{Aa}(j) + q^2Q_T^{AA}(j)$. It follows that the probability distribution of y_j given n_j is:

$$P\big(y_j|n_j\big) = \sum_{n_{2j}}^{n_j} \sum_{n_{3j}}^{n_j - n_{2j}} g\big(n_{2j}, n_{3j}; n_j, 2pq, q^2\big)\, h\big(y_j, Q_T^{ALL}(j)\big) \tag{4.49}$$

where $g\big(n_{2j}, n_{3j}; n_j, 2pq, q^2\big)$ is the multinomial density of $(n_{2j}, n_{3j})|\, n_j \sim Multinomial\ (n_j;\ 2pq, q^2)$ and $h(y_j, Q_T^{ALL}(j))$ the Poisson density of $y_j|\, n_{2j}, n_{3j}, n_j \sim Poisson\ \{Q_T^{ALL}(j)\}$.

The probability given by (4.49) is a mixture of Poisson distribution with mixing probability distribution given by the multinomial distribution of $\{n_{2j}, n_{3j}\}$ given n_j. This mixture distribution represents individuals with different genotypes at the embryo stage in the population.

The Probability Distribution of the Mixture Model

Let Θ be the set of all unknown parameters in the mixture model (4.49). Based on data (y_j, $j = 0,1, \ldots, k$), the likelihood function of Θ is $L\{\Theta \,|y_j, j = 0,1, \ldots, k\} = \prod_{j=0}^{k} P(y_j|n_j)$.

In order to make inference about unknown parameters, we expand the model to include the unobservable variables $\{n_{2j}, n_{3j}, y_{2j}, y_{3j}\}$. To derive the joint probability distribution of these variables, observe that for j ≥ 1, the conditional probability of $\{y_{2j}, y_{3j}\}$ given $\{n_{ij}, i = 2,3, n_j, y_j\}$ is multinomial with parameters $\{y_j, \frac{2pqn_jQ_{Aa}(j)}{Q_T^{ALL}(j)}, \frac{q^2Q_T^{AA}(j)}{Q_T^{ALL}(j)}\}$. That is,

$$P\{y_{2j}, y_{3j} \mid n_{ij}, i = 2,3,\ n_j, y_j\} \sim Multinomial\ \{y_j, \frac{2pqn_jQ_{Aa}(j)}{Q_T^{ALL}(j)}, \frac{q^2Q_T^{AA}(j)}{Q_T^{ALL}(j)}\} \quad (4.50)$$

for j ≥ 1.

Hence for j ≥ 1, the joint distribution of $\{n_{ij}, y_{ij}, i = 2,3, y_j\}$ given n_j is

$$P\{n_{ij}, y_{ij}, i = 2,3, y_j, j = 0,1, \ldots, k|n_j, \Theta\} =$$

$$g(n_{2j}, n_{3j}; n_j, 2pq, q^2)h(y_0, \alpha p^2)h\left(y_{2j}, 2pqn_jQ_{Aa}(j)\right)h(y_{3j}, q^2Q_T^{AA}(j)). \quad (4.51)$$

Put $\boldsymbol{Y} = (y_{ij},\ i = 2,3, j = 1,\ldots,k)$, $\boldsymbol{N} = (n_{ij},\ i = 2,3, j = 1,\ldots,k)$, $\tilde{y} = (y_j, j = 0,1, \ldots, k)$ and $\tilde{n} = (n_j, j = 0,1, \ldots, k)$. For the SEER data, the joint density $P\{Y, \tilde{y}, N \mid \tilde{n}, \Theta\}$ given $\{\tilde{n}, \Theta\}$ is:

$$P\{Y, \tilde{y}, N \mid \tilde{n}, \Theta\} =$$

$$h(y_0, \alpha p^2)\ \prod_{j=1}^{k} g(n_{2j}, n_{3j}; n_j, 2pq, q^2)\ h\left(y_{2j}, 2pqn_jQ_{Aa}(j)\right)h(y_{3j}, q^2Q_T^{AA}(j)) \quad (4.52)$$

Notice that the above distribution is a product of multinomial distributions and Poisson distributions. For this join distribution, the deviances $Dev = -2\left\{P\{Y, \tilde{y}, N \mid \tilde{n}, \Theta\} - P\{Y, \tilde{y}, N \mid \tilde{n}, \widehat{\Theta}\}\right\}$ is:

$$Dev = D_0 + Dev(p) + \sum_{j=1}^{k} D_j \tag{4.53}$$

where

$$Dev(p) = 2\sum_{j=1}^{k}\{n_{1j}\log(\frac{n_{1j}}{n_j p^2}) + n_{2j}\log\left(\frac{n_{2j}}{2n_j p(1-p)}\right) + n_{3j}\log(\frac{n_{3j}}{n_j(1-p)^2})\}$$

$$D_j = 2\left\{2pqn_j Q_{Aa}(j) - y_{2j} - y_{2j}\log\left(\frac{n_{2j}\times 2pqn_j Q_{Aa}(j)}{n_j}\right)\right\} + 2\left\{q^2 Q_T^{AA}(j) - y_{3j} - y_{3j}\log\left(\frac{n_{3j}\times q^2 Q_T^{AA}(j)}{n_j}\right)\right\} \tag{4.54}$$

The joint density $P\{Y, \tilde{y}, N \mid \tilde{n}, \Theta\}$ given by (4.52) will be used as the kernel for the Bayesian method to estimate the unknown parameters and to predict the state variables.

State Space Model and Estimation of Unknown Parameters

Unknown Parameters and Fitting of the Model by Cancer Incidence Data

In the above model, the unknown parameters are $\{p, \alpha, \omega_0(t), \omega_1(t), \alpha_i(t), i = 0,1,2, \beta_j(t), j = 0,1,2,\ b_1^{(AA)}(t),\ d_1^{(AA)}(t),\ b_i^{(I)}(t),\ d_i^{(I)}(t),\ b_j^{(J)}(t),\ d_j^{(J)}(t)\}$. Because the mutation rates are very small, it is reasonable to assume all mutation rate are homogeneous. Since the proliferation rate (i.e., birth rate – death rate) of normal stem cells in individuals after birth is expected to be very small (Weinberg, 2007), so that we assume that the number of stem cells of individual is a constant (~10^8).

To fit the SEER liver cancer data, we let one time unit (i.e., Δt=1) be 6 months after birth. Then because the growth of last stage cells is very rapid, during a six months period

one may practically assume $P_T(s,t) \approx 1$ for $t - s \geq 1$. Then using this approximation, we summarize probability of developing liver cancer from each pathway.

For individuals with *aa* genotype at embryo stage, the probability of developing HBL at birth is 1.

For individuals with *Aa* genotype at embryo stage, as shown in section 4.5.1, the probability of developing HBL during period (t_{j-1}, t_j] is:

$$P_{Aa}(j) \approx e^{-\omega_1 u_1^{A^{(2)}}(t_0)\varphi_1(t_{j-1})} - e^{-\omega_1 u_1^{A^{(2)}}(t_0)\varphi_1(t_j)} \tag{4.55}$$

where $u_1^{A^{(2)}}(t) = u_1^{A^{(2)}}(t_0)e^{\int_{t_0}^{t}\gamma^{A^{(2)}}(x)dx} = u_1^{A^{(2)}}(t_0)e^{\gamma^{A^{(2)}}(t-t_0)}$ and $\varphi_1(t) = \frac{1}{\gamma^{A^{(2)}}}[e^{\gamma^{A^{(2)}}(t-t_0)} - 1]$.

For individuals with *AA* genotype at embryo stage, as shown (4.45) and (4.46) in section 4.5.2, the probability of developing HBL during period (t_{j-1}, t_j] is:

$$P_{AA}^{(3)}(j) \approx e^{-\omega_1\varphi_2(t_{j-1})} - e^{-\omega_1\varphi_2(t_j)} \tag{4.56}$$

where $\varphi_2(t_{j-1})$ is given by (4.46).

For individuals with AA genotype (normal person) at birth, as shown in section 4.4.1, the probability of developing HCC during period (t_{j-1}, t_j] through I-pathway (3 stages) and J-pathway (4 stages) are:

$$Q_i(j) = E\left\{e^{-\mu_i(t_{j-1})} - e^{-\mu_i(t_j)}\right\} \tag{4.57}$$

where

$$\mu_1(t) \sim \int_{t_0}^{t} E\left[I_{k_1-1}(s)\right]\alpha_{k_1-1}(s)ds \tag{4.58}$$

$$\mu_2(t) \sim \int_{t_0}^{t} E\left[J_{k_2-1}(s)\right]\beta_{k_2-1}(s)ds \tag{4.59}$$

where $E\left[I_{k_1-1}(s)\right]$ and $E\left[J_{k_2-1}(s)\right]$ can be readily obtained, but $E\left[J_{k_2-1}(s)\right]$ is very complicate, we will use numerical way to get the approximation, which will be discussed in the next section. Here we only show how to derive $E\left[I_{k_1-1}(s)\right]$, where $k_1 = 3$ for I-pathway. Let $u_2^{(I)}(t)$ represent $E[I_2(t+1)]$ and $u_1^{(I)}(t)$ represent $E[I_1(t+1)]$,

$$u_2^{(I)}(t+1) = E[I_2(t+1)] = u_2^{(I)}(t)\left(1+\gamma_2^{(I)}\right) + u_1^{(I)}(t)\alpha_1 = u_2^{(I)}(t-1)\left(1+\gamma_2^{(I)}\right)^2 + \alpha_1\left[u_1^{(I)}(t-1)\left(1+\gamma_2^{(I)}\right) + u_1^{(I)}(t)\right] = \cdots = u_2^{(I)}(t-1)\left(1+\gamma_2^{(I)}\right)^{t-t_0+1} + \alpha_1\sum_{j=1}^{t-t_0} u_1^{(I)}(t_0+j)\left(1+\gamma_2^{(I)}\right)^{t-t_0-j} = \frac{\alpha_0\alpha_1}{\gamma_1^{(I)}}\left(1+\gamma_2^{(I)}\right)^{t-t_0}\sum_{j=1}^{t-t_0}\left[\left(1+\gamma_1^{(I)}\right)^j - 1\right](1+\gamma_2^{(I)})^{-j} \quad (4.60)$$

From all above results, it follows that the probability of developing liver cancer depends on parameters through some functions of parameters, such as $\alpha_0\alpha_1\alpha_2$, $\alpha_0\beta_0\beta_1\beta_2$. And also only proliferation rates $\gamma_1^{(I)}, \gamma_2^{(I)}, \gamma_1^{(J)}, \gamma_2^{(J)}, \gamma_1^{A^{(3)}}$ could be estimated. However, since we assume the number of stem cells at birth for normal person (genotype *AA*) and for person with *Aa* genotype are constant (~10^8), that is, $N(t_0)\sim 10^8, u_1^{A^{(3)}}(t_0)\sim 10^8,\ u_1^{A^{(2)}}(t_0)\sim 10^8$, we can use Generalized Gibbs Sampling method to obtain estimates for the mutation rates separately. The method used to estimate parameters will be discussed in the next section. To simplify the notation, from now on, we use AA to represent the $A^{(3)}$ in the superscribe of notations, and Aa to represent the $A^{(2)}$, for example, $u_1^{A^{(3)}}(t_0)$ is represented by $u_1^{AA}(t_0)$.

State Space Model of Human Liver Cancer

State space model is a stochastic model which consists of two sub-models: The stochastic system model which is the stochastic model of the system and the observation

model which is a statistical model based on available observed data from the system. Thus, the state space model of a system takes into account the basic mechanisms of the system and the random variation of the system through its stochastic system model and incorporates all these into the observed data from the system; furthermore, it validates and upgrades the stochastic model through its observation model and the observed data of the system. As illustrated in Tan (2002, Chapters 8-9), the state space model has many advantages over both the stochastic model and the statistical model when used alone since it combines information and advantages from both of these models.

For human liver cancer (HBL and HCC), the stochastic system model of the state space model is the stochastic multiple pathways model consisting of 5 pathways with each pathway following a single-stage or a multi-stage model as described in Section 4.4-4.5; the observation model of this state space model is a statistic model based on the observed number of liver cancer cases as described in Section 4.6.

The Stochastic System Model and the State Variables

Putting $\Delta t = 1$ for some fixed small interval, then the staging variables are $\boldsymbol{X} = \{\tilde{X}(t), t = t_0, t_0 + 1, \ldots, t_M\}$, the transition variables are $\boldsymbol{U} = \{\tilde{U}(t), t = t_0, t_0 + 1, \ldots, t_M\}$ and population segregation variables are $\boldsymbol{N} = (n_{ij},\ i = 2,3,\ j = 1,\ldots,k)$. From results in section 4.6, the joint probability distribution of $\{\boldsymbol{X}, \boldsymbol{U}\}$ given the parameters Θ and $\boldsymbol{N}$ is:

$$P\{X, U|N, \Theta\} = \prod_{t=t_0+1}^{t_M} P(\tilde{X}(t)|\tilde{X}(t-1), \tilde{U}(t-1), N, \Theta) \times P\{\tilde{U}(t-1)|\tilde{X}(t-1), N, \Theta\} \quad (4.61)$$

where $P\{\tilde{U}(t-1)|\tilde{X}(t-1), N, \Theta\}$ and $P(\tilde{X}(t)|\tilde{X}(t-1), \tilde{U}(t-1), N, \Theta)$ are given by equations (4.18), (4.28) and (4.38) respectively, where $\Theta = \{p,\ \omega_0, \omega_1, \alpha_i, i =$

$0,1,2,\beta_j, j = 0,1,2, b_i^{(I)}, d_i^{(I)}, i = 1,2,\ b_j^{(J)}, d_j^{(J)}, j = 1,2,\ \gamma_1^{AA}(= \gamma_1^{Aa})\}$. We assume the model is time homogeneous, then all parameters are independent of time so that Θ = $\{p, \Theta_1, \Theta_2, \Theta_3\}$, where $\Theta_1 = \{\ \omega_0, \alpha_i, i = 0,1, \beta_j, j = 0,1\}$, $\Theta_2 = \{b_i^{(I)}, d_i^{(I)}, i = 1,2,$ $b_j^{(J)}, d_j^{(J)}, j = 1,2, b_1^{AA}(= b_1^{Aa}), d_1^{AA}(= d_1^{Aa})\}$, $\Theta_3 = \{\omega_1, \alpha_2, \beta_2\}$.

Notice that this probability distribution is basically a product of Poisson distributions and multinomial distributions.

The Observation Model Using SEER Data

Put **Y** = (y_j, j = 1, . . . ,m) and ***N*** = (n_{ij}, i = 2,3, j = 1,...,k), by the probability distribution given by equation (4.52), the conditional joint probability density of **Y**, **N** given {**X, U,** Θ} is:

$$P(Y, N|X, U, \Theta) = P\{Y, \tilde{y}, N \mid \tilde{n}, \Theta\} \tag{4.62}$$

And above distribution is a product of multinomial distributions and Poisson distributions. The deviance from above density has been given in (4.53).

From equations (4.61) and (4.62), we have following join density of (**X**, **U**, **Y**, **N**) given Θ:

$$P(X, U, Y, N|\Theta) = P(X, U, N|\Theta) \times P(Y, N|X, U, \Theta) \times P(N|\Theta) \tag{4.63}$$

where $P(N|\Theta)$ is a multinomial distribution.

The Generalized Bayesian Method and the Gibbs Sampling Procedure

To fit the model to the data and to valid the model, one would need to estimate the unknown parameters and to predict the state variable. We propose generalized Bayesian inference procedures to achieve those purposes.

The generalized Bayesian inference is based on the posterior distribution $P\{\Theta|\mathbf{X}, \mathbf{U}, \mathbf{Y}, \mathbf{N}\}$ of Θ given $\{\mathbf{X}, \mathbf{U}, \mathbf{Y}, \mathrm{N}\}$. This posterior distribution is derived by combining the prior distribution $P\{\Theta\}$ of Θ with the probability distribution $P\{\mathbf{X}, \mathbf{U}, \mathbf{Y}, \mathbf{N} \mid \Theta\}$ given by equation (4.63). It follows that this inference procedure would combine information from three sources:

(1) Previous information and experiences about the parameters in terms of the prior distribution $P\{\Theta\}$ of the parameters.

(2) Biological information via the stochastic system equations of the stochastic system ($P\{\mathbf{X}, \mathbf{U} \mid \mathrm{N}, \Theta\}$) .

(3) Information from observed data via the statistical model from the system ($P\{\mathbf{Y}, \mathbf{N} |\mathbf{X}, \mathbf{U}, \Theta\}$).

Because of additional information from the stochastic system model, this inference procedure is advantageous over the standard Bayesian procedure in that it can avoid the identifiability problems associated with standard Bayesian method. For example, we have shown that to the order of $Max\{o(\alpha_2(t)), o(\beta_2(t)), o(\omega_1(t))\}$ the probability distribution of the y_j's depends on the stochastic model through the expected numbers of $I_2(t)$, $J_2(t)$, $A_1^{(2)}$and $A_2^{(3)}$, which depend on function of mutation rates (as shown in section 4.7.1) as well as the birth rates and death rates through the difference of these rates. It follows that it is not possible to estimate the birth rates, death rates and mutation rates separately by the traditional Bayesian method. Most importantly, the number of parameters is very large and the number of data points is limited. Thus, without information from the stochastic system model, it is virtually impossible to estimate all unknown parameters; for more examples, see Tan (2000, 2002). Notice that if one uses the standard Bayesian

inference procedure by combining the prior with the density $P\{\mathbf{Y}, \mathbf{N} | \mathbf{X}, \mathbf{U}, \Theta\}$, then because the density $P\{\mathbf{Y}, \mathbf{N} | \mathbf{X}, \mathbf{U}, \Theta\}$ depends on the birth rates and the death rates only through the differences of these rates, it is not possible to estimate the birth rates $(b_i^{(I)}, b_i^{(J)}, i = 1,2)$ and death rates $(b_i^{(I)}, b_i^{(J)}, i = 1,2)$ separately but only the proliferation rates (the difference between birth rate and death rate).

The Prior Distribution of the Parameters

For the prior distributions of Θ, because biological information have suggested some lower bounds and upper bounds for the mutation rates and for the proliferation rates, we assume

$$P(\Theta) \propto c \ (c > 0) \tag{4.64}$$

where c is a positive constant if these parameters satisfy some biologically specified constraints; and equal to zero for otherwise. These biological constraints are:

(i) To ease the problem, we assume individuals with genotype *aa* at embryo stage develop HBL with probability 1, and almost all of HBL cases at birth are from individuals with genotype *aa* at embryo stage. Then the frequency of individuals with genotype *aa* at the embryo stage in the population p^2 can be approximately determined by rate of HBL cases among all population at birth.

(ii) For the mutation rates of the I_i cells in the first pathway (i.e., the $N \rightarrow I_1 \rightarrow \ldots \rightarrow I_{k_1} \rightarrow$ *tumor* pathway), $1 < N_0 * \alpha_0 < 1000$ $(N \rightarrow I_1)$, $10^{-8} < \alpha_i < 10^{-3}$, $i = 1, 2$, where $N_0 = 10^8$. For the proliferation rates of I_i cells in the I-pathway, $0 < b_i^{(I)} < 0.5$, $i = 1, 2$, $10^{-4} < \gamma_i^{(I)} = b_i^{(I)} - d_i^{(I)} < 0.2$, $b_1^{(I)} < b_2^{(I)}$.

(iii) For the mutation rates in the J-pathway (i.e., the $N \rightarrow I_1 \rightarrow J_1 \rightarrow \ldots \rightarrow J_{k_2} \rightarrow$ *tumor* pathway), $10^{-8} < \beta_j < 10^{-3}$, $j = 0, 1, 2$. For the proliferation rates in the second pathway, $0 < b_i^{(J)} < 0.5$, $10^{-4} < (\gamma_j^{(J)} = b_j^{(J)} - d_j^{(J)} < 0.2, j = 1, 2$, $b_1^{(J)} < b_2^{(J)}$.

(iv)For the mutation rates in the AA-pathway (i.e., the $AA \rightarrow Aa \rightarrow aa \rightarrow$ *tumor* pathway), $10^{-10} < \omega_k < 10^{-4}$, k = 0, 1. For the proliferation rates of Aa cells, $0 < b_2^{(AA)} = b_1^{(Aa)} < 10^{-3}$, and $10^{-6} < \gamma_1^{(Aa)} = b_1^{(Aa)} - d_1^{(Aa)} = b_1^{(AA)} - d_1^{(AA)} = \gamma_1^{(AA)} < 10^{-3}$.

(v) From information from Section 4.2, $\alpha_i < \beta_i$, $\gamma_i^{(I)} < \gamma_i^{(J)}$, $i = 1, 2$.We will refer the above prior as a partially informative prior which may be considered as an extension of the traditional non- informative prior given in Box and Tiao (1973).

The Posterior Distribution of the Parameters Given {Y,X, U, N}

Combining the prior distribution given in section 4.8.1 with the density of $P\{\mathbf{X}, \mathbf{U}, \mathbf{Y}, \mathbf{N} | \Theta\}$ given in equation (4.63), one can readily derive the conditional posterior distribution of Θ given $\{\mathbf{X}, \mathbf{U}, \mathbf{Y}, \mathbf{N}\}$. For ($i = 1, 2$), denote by: $N_{iI} = \sum_{t=1}^{t_M} I_i(t)$, $B_{iI} = \sum_{t=1}^{t_M} B_i^{(I)}(t)$ and $D_{iI} = \sum_{t=1}^{t_M} D_i^{(I)}(t)$; similarly, for $j = 1, 2$, we define $\{N_{jJ}, B_{jJ}, D_{jJ}\}$ by replacing $(I_i(t), B_i^{(I)}(t), D_i^{(I)}(t))$ by $(J_j(t), B_j^{(J)}(t), D_j^{(J)}(t))$, respectively. We also define $A = \sum_{t=1}^{t_M} Aa(t)$, $B_{1Aa} = \sum_{t=1}^{t_M} B_1^{(Aa)}(t)$ and $D_{1Aa} = \sum_{t=1}^{t_M} D_1^{(Aa)}(t)$. For $i = 1, 2$, put $S_i^{(I)} = I_i(t_M) - I_i(t_0) - B_{iI} + D_{iI}$, $S_i^{(J)} = J_i(t_M) - J_i(t_0) - B_{iJ} + D_{iJ}$ and $S_i^{(Aa)} = A(t_M) - A(t_0) - B_{1Aa} + D_{1Aa}$. Then, we have the following results for the conditional posterior distributions:

(i) The conditional posterior distributions of $\Theta_1 = \{\omega_0, \alpha_i, i = 0,1, \beta_j, j = 0, 1\}$ given $\{\mathbf{X}, \mathbf{U}, \mathbf{Y}, \mathbf{N}\}$ is:

$$P\{\Theta_1|X,U,Y,N\} \propto$$

$$P\{\Theta_1\}e^{-N_0\alpha_0(t_M-t_0)-\beta_0 N_{1I}-\omega_0(t_M-t_0)} \times (N_0\alpha_0)^{S_1^{(I)}}(\beta_0)^{S_1^{(J)}}(\omega_0)^{S_1^{(Aa)}} \times$$

$$e^{-\alpha_1 N_{1I}}(\alpha_1)^{S_2^{(I)}} e^{-\beta_1 N_{1J}}(\beta_1)^{S_2^{(J)}} \quad (4.65)$$

(ii) The conditional posterior distributions of $\Theta_2=\{b_i^{(I)}, d_i^{(I)}, i = 1,2,\ b_j^{(J)}, d_j^{(J)}, j = 1,2, b_1^{AA}, d_1^{AA}\}$ given {**X**, **U**, **Y**, N } is:

$$P\{\Theta_2|X,U,Y,N\} \propto$$

$$P\{\Theta_2\}f(B_{1Aa}, D_{1Aa}; A, b_1^{AA}, d_1^{AA}) \prod_{i=1}^{2}\{f(B_{iI}, D_{iI}; N_{iI}, b_i^{(I)}, d_i^{(I)})f(B_{ij}, D_{iJ}; N_{iJ}, b_i^{(J)}, d_i^{(J)})\} \quad (4.66)$$

where *f(x, y; N, p, q)* is a multinomial density with parameters (*N, p, q*).

(iii) The conditional posterior distribution of $\Theta_3 = \{\omega_1, \alpha_2, \beta_2\}$ given {**X**, **U**, **Y**, **N** } is:

$$P\{\omega_1, \alpha_2, \beta_2|X,U,Y,N\} \propto P\{\Theta_3\} \prod_{j=1}^{m} e^{-\tau_j}(\tau_j)^{y_j} \quad (4.67)$$

where *m* is the number of time units between t_0 to t_M.

The Multi-level Gibbs Sampling Procedure for Estimating Parameters

Given the above probability distributions, the multi-level Gibbs sampling procedure for deriving estimates of the unknown parameters are given by:

(a) Step 1: Generating (**X**,**U**, **N**) Given (**Y** , **Θ**) (The Data-Augmentation Step):

Use computed $\hat{p}$ (*p* be the frequency of the mutated APC gene (denoted by *a*) in the population) to segregate the population into three groups: *AA* group consists people who are normal at birth, *Aa* group consists of all individuals with *Aa* genotype at birth, and *aa* group consists of individuals with *aa* genotype at birth. Given **Y** and **Θ**, use the stochastic equations (4.1), (4.2), (4.32) and the probability distributions given by equations (4.3) -

(4.8) and (4.34) - (4.36) in Section 4.3 and 4.4 to generate a large sample of (**X**,**U**). Then by combining the sample with $P\{\mathbf{Y}\,|\mathbf{X},\,\mathbf{U},\,\boldsymbol{\Theta},\,\mathrm{N}\}$, (**X**, **U**) are selected through the weighted bootstrap method proposed by Smith and Gelfand (1992). This selected (**X**, **U**) is then a sample from $P\{\mathbf{X},\mathbf{U}|\mathbf{Y}\,,\,\boldsymbol{\Theta}\,\}$ even though the latter is unknown. (For proof, see Tan (2002), Chapter 3.) Call the generated sample $(\widehat{\boldsymbol{X}},\widehat{\boldsymbol{U}})$.

(b) Step 2: Estimation of $\boldsymbol{\Theta} = \{\Theta_1, \Theta_2, \Theta_3\}$ Given {**Y** ,**X**, **U**}:

Given **Y** and given (**X**,**U**) = $(\widehat{\boldsymbol{X}},\widehat{\boldsymbol{U}}\,)$ from Step 1, derive the posterior mode of the parameters by maximizing the conditional posterior distribution $P\{\,\boldsymbol{\Theta}\mid\widehat{\boldsymbol{X}},\widehat{\boldsymbol{U}},\boldsymbol{Y},\hat{p}\}$. Denote the generated mode as $\widehat{\boldsymbol{\Theta}}$.

(c) Step 3: Iterative Step.

With $\{(\mathbf{X},\,\mathbf{U}) = (\widehat{\boldsymbol{X}},\widehat{\boldsymbol{U}}),\,\boldsymbol{\Theta} = \widehat{\boldsymbol{\Theta}}\}$ given above, go back to Step (a) and continue until convergence.

The proof of convergence of the above steps can be proved using procedure given in Tan (2002) (Chapter 3). At convergence, the $\widehat{\boldsymbol{\Theta}}$ are the generated values from the posterior distribution of **Θ** given **Y** independently of (**X**,**U**) (for proof, see Tan (2002, Chapter 3). Repeat the above procedures one then generates a random sample of **Θ** from the posterior distribution of **Θ** given **Y** ; then one uses the sample mean as the estimates of **Θ** and use the sample variances and covaraiances as estimates of the variances and covariances of these estimates.

Application to Fit the SEER Data

In this section, we will apply the above model to the NCI/NIH liver cancer data from the SEER project. Given in Table 17 are the numbers of people at risk and liver cancer cases in all age groups, as well as predicted cancer cases by using our model. The cancer

Table 18

Estimates of Parameters for Each Pathway

p = 3.48E-03 ± 8.51E-04, α = 1.12 E-05 ± 1.03E-05			
$AA \rightarrow Aa \rightarrow aa \rightarrow tumor$ pathway			
	$AA \rightarrow Aa$ (ω_0)	$Aa \rightarrow aa$ (ω_1)	
Mutation rate	3.98E-08 ± 1.11E-09	1.27E-05 ± 6.96E-08	
	Aa	*aa*	
Birth Rate	4.33E-04 ± 1.89E-05	N/A	
Death Rate	3.92E-04 ± 3.07E-05	N/A	
I-Pathway ($N \rightarrow I_1 \rightarrow I_2 \rightarrow I_3 \rightarrow tumor$)			
	$N \rightarrow I_1$ ($N_0\alpha_0$)	$I_1 \rightarrow I_2$	$I_2 \rightarrow I_3$
Mutation rate	3.88 ± 0.14	4.36 E-03 ± 9.99E-04	1.16E-06 ± 1.33E-07
	I_1	I_2	I_3
Birth Rate	7.48E-03± 2.24 E-03	1.2 E-02 ± 1.26E-03	N/A
Death Rate	4.43E-03± 2.55 E-03	9.07E-03 ± 6.12 E-03	N/A
J-Pathway ($N \rightarrow I_1 \rightarrow J_1 \rightarrow J_2 \rightarrow J_3 \rightarrow tumor$)			
	$I_1 \rightarrow J_1$	$J_1 \rightarrow J_2$	$J_1 \rightarrow J_2$
Mutation rate	4.37E-03 ± 1.35E-03	3.67E-03 ± 1.82E-03	2.80 E-05 ± 7.41E-07
	J_1	J_2	J_3
Birth Rate	2.14E-02 ± 2.83E-03	2.86E-02 ± 1.48E-02	N/A
Death Rate	1.38E-02 ± 3.39E-03	4.37E-03 ± 5.43E-03	N/A

cases developed from each pathway are also predicted and show in Table 17. There are 18 (m = 18) age groups with each group spanning over 5 years except at birth.

Given in Table 18 are the estimates of the mutation rates, the birth and death rates of the I_i cells, J_j cells AA cells and Aa cells. Given in Figure 19 is the plot of the observed and predicted cancer incidence and Figure 20 the probability distributions of time to tumors from each pathway.

From these results, we have made the following observations:

(a) As shown by results in Table 17, the predicted number of cancer cases are very close to the observed cases. The AIC (Akaike Information Criteria) and the BIC (Schwarz Bayesian Information Criteria) are given by 66.23 and 87.94 respectively. This indicates that the model fits the data well and that one can safely assume that the adult human liver cancer can be described by a model of multiple pathways as given in Figure 18. It fits also considerably better than single pathway models. The AIC and the BIC of the best fitted single pathway model (i.e., the 4-stage single pathway model) are given by

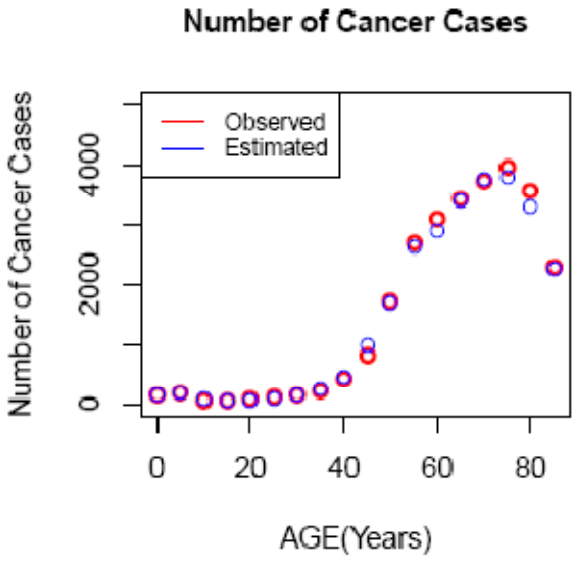


Figure 19. Observed and Predicted Liver Cancer Cases

316.0184 and 326.2786 respectively.

(b) Most of individuals with *Aa* genotype at embryo stage develop liver cancer before 20 year-old, which is consistent with clinical results (Hirschman, Pollock, & Tomlinson, 2005). If individuals are born with normal, most of liver cancer is developed through I-pathway and J-pathway, that is, environmental effect play the most important role for adult liver cancer.

(c) From Table 17, it is observed that the largest number of cancer cases is in the age group between 65 and 75 years old. Most of liver cancer cases develop liver cancer through J-pathway though one more pathway is needed for J-pathway than the I-pathway. As shown parameters estimate in Table 18, the higher mutation rates for $I_1 \rightarrow J_1$ and $J_1 \rightarrow J_2$, as well as higher proliferation rates for J_1 and J_2 account for the difference.

(d) Results in Table 18 showed that the mutation rates from $I_1 \rightarrow J_1$, from $J_1 \rightarrow J_2$

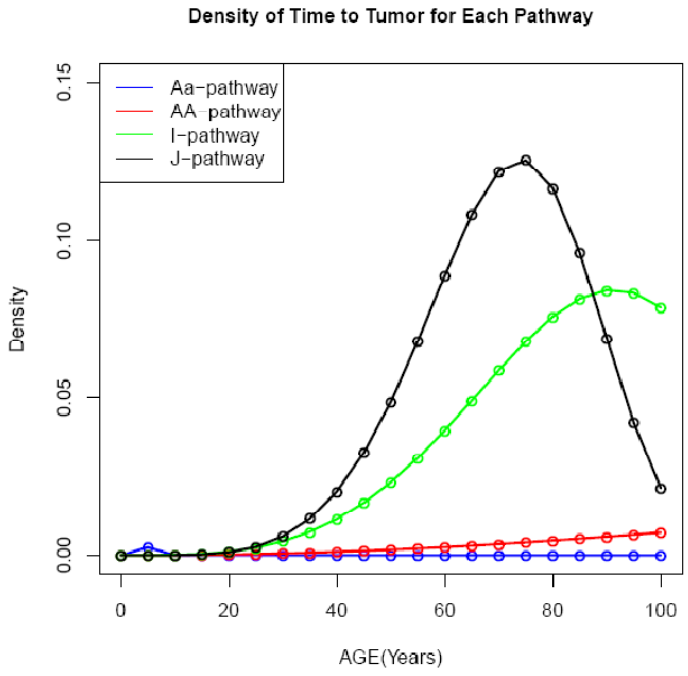


Figure 20. Density of Time to Tumor for Each Pathway

and from $J_1 \rightarrow J_2$ are e about 10 times greater than the mutation rates from $I_1 \rightarrow I_2 \rightarrow I_3$. These results might be the consequence that the J-pathway has accumulated more genetic and epigenetic changes (see Pogribny, Rusyn, & Beland, 2008; Villanueva et al., 2007).

(e) Results in Table 18 showed that the estimates of the birth rates of J_2 cells were about 10 times greater than those of I_2 cells. The estimate of proliferation rate (birth rate - death rate) of the J_2 cells is around 4 times of that of J_1 cells. These estimates are about 10 times greater the estimates of the proliferation rates of I_2 cells (2.42E-02 and 0.3E-02, respectively). Notice that the estimate of the birth rate of I_2 cells is 0.012 which is greater than the estimate 0.00748 of I_1 cells but the estimate of death rate of I_2 cells is 0.00907 which is much greater than the estimate 0.00443 of I_1 cells, due presumably to effects of apoptosis (Notice that apoptosis is usually a late event in carcinogenesis.). This may help explain why the proliferation rate of I_2 cells is about equal to (or a little smaller than) that of I_1 cells. Notice also that the estimates of the proliferation rates of J_i ($i = 1, 2$) cells are of order 10^{-2} (proliferation rate for J_1 is 0.0076, which is slightly smaller than 10^{-2}) whereas those of I_i ($i = 1, 2$) cells are of order 10^{-3}. These results clearly reflect the biological observation that more genetic and epigenetic changes have accumulated in the J-pathway than in the I-pathway (Pogribny et al., 2008; Villanueva et al., 2007).

Conclusion and Discussion

Recent studies of cancer molecular biology have indicated very clearly that human liver cancer is developed through multiple pathways (Grisham, 2002), and the cancer mechanisms for pediatric liver cancer (HBL) are different from adult liver cancer (HCC) (Hirschman et al. 2005). This indicates that single pathway models are not realistic and hence may lead to incorrect prediction and confusing results. For developing efficient

prevention and controlling procedures for human liver cancer and for prediction of future human liver cancer, in this chapter we have developed a stochastic model and a state space model for carcinogenesis of human liver cancer involving multiple pathways incorporating hereditary liver cancer, with each pathway being a multi-stage model. Using this model, we have derived for the first time the probability distribution of the numbers of initiated cells and the probability distribution of time to cancer tumors for each pathway. Such derivation by the traditional approach is extremely difficult and had not been attempted previously for liver cancer involving single pathway models and multiple pathway models.

Based on the state space model of liver cancer, we have developed a generalized Bayesian procedure to estimate the unknown parameters and to predict future cancer cases. This approach combines information from three sources: The stochastic system model via $P\{\mathbf{X}, \mathbf{U}, \mathbf{N} \mid \boldsymbol{\Theta}\}$, the prior information via $P\{\boldsymbol{\Theta}\}$ and information from data via $P\{\mathbf{Y} \mid \mathbf{N}, \mathbf{X}, \mathbf{U}, \boldsymbol{\Theta}\}$. Because of additional information from the stochastic system model, our procedure is advantageous over the standard Bayesian procedure and the sampling theory procedure. For example, for the first time we can estimate the birth rates and death rates separately for all pathways which are not possible by the classical Bayesian methods or sampling theory methods. Notice that there are a large number of unknown parameters in the model and only a limited amount of data are available. Without this additional information, it is then not possible to estimate all unknown parameters. Notice also that through the stochastic system model, one can incorporate biological mechanism into the model. Because the number of stages and the mutation rates of intermediate cells

in different pathways are different and different drugs may affect different pathways, we believe that this is important and necessary.

We have applied this model and methods to the NCI SEER data (up-to November, 2008). Our results showed that the proposed multiple-pathway model fit the data remarkably well. The estimates from the model are strongly consistent with biological findings.

5. SUMMARY AND FUTURE RESEARCH

We have developed stochastic models and statistical models and hence state space model for carcinogenesis under complex situations (i.e., multiple-pathway multiple-stage model).

In chapter 1, we have used a simple two-pathway model with one pathway involving a single stage model and the pathway involving a three stages model to assess the classical two-stage Markov model (MVK model) for carcinogenesis; we have derived the incidence function and have revealed many difficulties and drawbacks of the traditional MVK model. To overcome these difficulties we have introduced stochastic difference equation method and have used the generalized Bayesian procedure to estimate unknown parameters. In this chapter we have briefly revealed the complex nature of carcinogenesis.

In chapter 2, we have developed stochastic models for the three commonly used experiments in bioassay in the area of cancer risk assessment of environmental agents: The initiation, the promotion and the complete experiments. We notice that the stochastic process for each experiment is unique and each can be represented by an unique multiple-pathway model. Most of these stochastic models are simple processes. For initiation experiment, the model is a two-pathway model with one stage for each pathway whereas for promotion and for complete experiments, the models are two-pathway model with one stage for one pathway and with two stages for the other pathway. A generalized Bayesian procedure was developed to estimate unknown parameters. Simulation study

shows that the models are quite reliable and estimates are very close to true (given) parameter values.

In chapter 3, based on biological information, we have developed a two-pathway model for sporadic human colon cancer; one pathway has 4 stages, and the other has 5 stages. Using this as the stochastic system model, we have developed a state space model with the observation model being a statistic model based on cancer incidence data relating the model to the system. In this model, biological information are effectively captured by a stochastic system model and cancer incidence are effectively linked to the system through the observation model. We have fitted the model to the NIH/NCI SEER data. The fitting results have revealed that the multiple-pathways model not only fit the observed incidence data better than the single 4-stage model, but also are consistent with biological findings.

In chapter 4, we have further extended the state space model to handle more complicated multiple-pathway carcinogenesis of human liver cancer incorporating hereditary and non-hereditary cancer incidence. Cancer heredity has been well known to account for pediatric liver cancer, also known as hepatoblastoma (HBL). The mechanism of developing HBL is different from cancer for adult, usually known as hepatocellular carcinoma (HCC). For HBL, the major mutated gene is APC, and for HCC, liver cancer is developed through two pathways with multiple stages for each pathway. We segregated the population by using the frequency of a major mutated gene (APC gene) into three subgroups, and for each subgroup, we constructed stochastic model accordingly. In combination of generalized Bayesian

approach using multi-level Gibbs sampling procedures, we have estimated the segregation frequency as well as mutation rates, birth rates and death rates for all stages. The model predicted cancer incidences of all age groups very well, and through the parameters estimated from the model we have also predicted the number of cancer cases for each cancer subgroups for each age group, as well as density of time to tumor for each pathway.

Overall, we have developed innovative stochastic and state space models for multiple-pathway carcinogenesis. We have developed models to assess bioassay's effect in tumor development, models to evaluate tumor development for colon cancer, and models for inherited and non-inherited liver cancer. The stochastic and state space models we have developed have many advantages over the traditional Markov multistage models, and are more consistent with biological information; these models are based on biological information and hence are more realistic and applicable in practice.

So far our study in risk assessment has only concentrated on time to tumor (papillomas and/or carcinomas); we have not extended the models to cover progression of papillomas and carcinomas. For future research, we will proceed to develop models to cover tumor progression; for these purposes we need to consider correlation between cases at two time points during tumor progression. Thus, for the initiation-promotion bioassay, our future work will cover developing state space model for longitudinal observations involving tumor progression.

The stochastic and state space models for carcinogenesis are still exploratory, and the models we developed have been applied for overall population, not for

small cohort. With more biological information discovered in molecular biology, cancer genetics, and cancer clinic, we will develop stochastic and state space models for more specific population group. In the future, we will also incorporate treatment protocols into models for colon cancer, liver cancer or other types of cancers to provide quantitative evaluation of treatment effects.

We have not predicted the cancer cases for the future time points. Thus, another future research subject is to develop predictive inference for cancer incidence and progression. We will access the predictive inference through two ways: Bayesian prediction and Kalman filter prediction. The Bayesian prediction is basically updating the posterior distribution of parameters by introducing the new sets of observations and newly updated prior distribution whereas the Kalman fitter prediction is based on construction of smoothing prediction function. We will compare these two predictive inferences to give more insight of prediction precision as well as to validate the models.

REFERENCES

Amitage, P., & Doll, R. (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis. *British Journal of Cancer*, *8*, 1-12.

Akhtar, R.S, Geng, Y., Klocke, B. J., & Roth, K.A. (2006). Neural precursor cells possess multiple p53-dependent apoptotic pathways. *Cell Death and Differentiation*, *13(10)*, 1727-1739.

Alberici, P., Jagmohan-Changur S., & De Pater, E. (2006). Smad4 haplo-insufficiency in mouse models for intestinal cancer. *Oncogene*, *25*, 1841-1851.

Barker, N., & Clevers, H. (2006) Mining the Wnt pathway for cancer therapeutics. *Nature Reviews Drug Discovery, 5(12)*, 997–1014.

Baylin, S.B., & Ohm, J.E. (2006). Epigenetic silencing in cancer-a mechanism for early oncogenic pathway addiction. *Nature Reviews (Cancer)*, *6*, 107-116.

Bhaskar, P. T., & Hay, N. (2007). The two TORCs and Akt. *Developmental Cell*, *12(4)*, 487–502.

Box, G.E.P., & Tiao, G.C. (1973). Bayesian Inference in Statistical Analysis. Reading, MA: Addison-Wesley.

Breivik, J., & Gaudernack, G. (1999). Genomic instability, DNA methylation, and natural selection in colorectal carcinogenesis, *Seminars in Cancer Biology*, *9*, 245-254.

Breuhahn, K., Longerich, T., & Schirmacher, P. (2006). Dysregulation of growth factor signaling in human hepatocellular carcinoma, *Oncogene*, *25*, 3787-3800.

Brown, K., Buchmann, A., & Balmain, A. (1990). Carcinogen-induced mutations in mouse c-Ha-Ras gene provide evidence of multiple pathways for tumor progression, *Proceedings of the National Academy of Sciences*, *87*, 538-542.

Brugge, J., Hung, M.C., & Mills, G.B. (2007). A new mutational AKTivation in the PI3K pathway. *Cancer Cell*, *12(2)*, 104–107.

Buendia, M.A. (2000). Genetics of hepatocellular carcinoma. *Seminar in cancer Biology*, *10*, 185-200.

Buick, R.N., & Pollak, M.N. (1984). Perspectives on clonogenic tumor cells, stem cells, and Oncogenes. *Cancer Research*, *44*, 4909-4918.

Carnero, A., Blanco-Aparicio C., Renner, O., Link W., & Leal J.F., (2008) The PTEN/PI3K/AKT signalling pathway in cancer, therapeutic implications, *Current Cancer Drug Targets*, *8(3)*, 187–98.

Chapelle, A. (2004). Genetic predisposition to colorectal cancer. *Nature Review (Cancer), 4*, 769-780.

Chen, C. J., & Chen, D. S. (2002). Interaction of hepatitis B virus, chemical carcinogen, and genetic susceptibility: Multistage hepatocarcinogenesis with mutifactorial etiology, *Hepatology*, *36*, 1046-1049.

Clevers, H. (2006). Wnt/beta-catenin signaling in development and disease. *Cell*, *127(3)*, 469–80.

DiGiovanni, J. (1992). Multistage carcinogenesis in mouse skin, *Pharmacology & Therapeutics*, *54*, 63-128.

DeFrances, M.C. (2005). Molecular mechanisms of hepatocellular carcinoma. In Brian I. Carr (Ed.), *Hepatocellular Cancer*. Totowa, NJ: Humana Press

Dufour, J. F., & Clavien, P. A., (2010). *Signaling Pathways in Liver Diseases*, New York, NY: Springer.

Fakir, H., Tan W.Y., Hlatky, L., Hahnfeldt, P., & Sachs, R.K., (2009). Stochastic population dynamic effects for lung cancer progression. *Radiation Research*, *172*, 383-393.

Fodde, R., Kuipers, J., Rosenberg, C., Smits, R., Kielman, M., Gaspar, C., van Es, J.H., Breukel, C., Wiegant, J., Giles, R. H., & Clevers, H. (2001). Mutations in the APC tumor suppressor gene cause chromosomal instability. *Nature Cell Biology*, *3*, 433-438.

Fodde, R., Smit, R., & Clevers, H (2001). APC, signal transduction and genetic instability in colorectal cancer. *Nature Review (Cancer)*, *1*, 55- 67.

Francoz, S., Froment, P., Bogaerts, S., De Clercq, S., Maetens, M., Doumont, G., Bellefroid, E., & Marine, J.C. (2006). Mdm4 and Mdm2 cooperate to inhibit p53 activity in proliferating and quiescent cells in vivo. *Proceedings of the National Academy of Sciences*, *103 (9)*, 3232-3237.

Gordon, M.D., & Nusse, R. (2006), Wnt signaling: multiple pathways, multiple receptors, and multiple transcription factors. *Journal of Biological Chemistry*, *281(32)*, 22429–22433.

Green R.A., & Kaplan K.B. (2003). Chromosomal instability in colorectal tumor cells is associated with defects in microtubule plus-end attachments caused by a dominant mutation in APC. *The Journal of Cell Biology*, *163*, 949-961.

Gomes, C. P., & Andrade, L.A. (2006). PTEN and p53 expression in primary ovarian carcinomas: immunohistochemical study and discussion of pathogenetic mechanisms. *International Journal of Gynecological Cancer*, *16 (1)*, 254-258.

Grisham, J.W. (2002). Molecular genetic alterations in primary hepatocellular neoplasms. In Coleman, W.B., & Tsongalis, G.J. (Ed.), *The Molecular Basis of Human Cancer*, Totowa, NJ: Humana Press

Greaves, M.F. (1997). Aetiology of acute leukaemia. *Lancet*, *349*, 344-349.

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

Hanazono, K., Natsugoe, S., Stein, H. J., Aikou, T., Hoefler, H., & Siewert, J.R. (2006) Distribution of p53 mutations in esophageal and gastric carcinomas and the relationship with p53 expression. *Oncology Rep*ort, *15(4)*, 821-824.

Hawkins, N. J., & Ward, R. L.(2001). Sporadic colorectal cancers with micro-satellite instability and their possible origin in hyperplastic polyps and serrated adenomas. *Journal of National Cancer Institute*, *93*, 1307-1313.

Herrero-Jimenez, P., Thilly, G., Southam, P. J., Tomita-Mitchell, A., Morgenthaler, S., Furth, E.E., et al. (1998). Mutation, cell kinetics, and subpopulations at risk for colon cancer in the United States, *Mutation Research*, *400*, 553-578.

Hennings, H., Glick, A.B., Greenhalph, D.A, Morgan, D.L., Strickland, J.E., Tennenbaum, T., & Yuspa, S. H. (1993). Critical aspects of initiation, promotion and progression in multistage experimental carcinogenesis. *Proceedings of the Society for Experimental Biology and Medicine*, *202*, 1-8.

Herpin, A., & Cunningham, C. (2007). Cross-talk between the bone morphogenetic protein pathway and other major signaling pathways results in tightly regulated cell-specific outcomes. *FEBS Journal*, *274(12)*, 2977–85.

Hisamuddin, I. M., & Yang, V.W. (2004). Genetics of colorectal cancer. *Medscape General Medicine*, *6 (3)*, 13.

Hirschman, A.B., Pollock, H.B., & Tomlinson, E. G. (2005). The spectrum of APC Mutations in Children with Hepatoblastoma from Familial Adenomatous Polyposis Kindreds, *Journal of Pediatrics*, *147*, .263-266.

Hopkin, K. (1996). Tumor evolution: survival of the fittest cells, *Journal of NIH Research*, *8*, 37-41.

Jiang, P. H., Motoo, Y., Garcia, S., Iovanna, J. L., Pebusque, M..J., & Sawabu, N. (2006). Down-expression of tumor protein p53-induced nuclear protein 1 in human gastric cancer. *World Journal* of *Gastroenterology*, *12(5)*, 691-696.

Jones, P. A., & Baylin, S.B. (2003). The fundamental role of epigenetic events in cancer. *Nature Reviews Genetics*, *3*, 415-428.

Karin, M., Liu, Z.G., & Zandi, E. (1997). AP-1 function and regulation. *Current Opinion in Cell Biology*, *9*, 240-246.

Kitisin, K., Saha, T., Blake, T., Golestaneh, N., Deng, M., Kim, C., Tang, Y., Shetty, K., Mishra, B., & Mishra, L. (2007). TGF-βeta signaling in development. *Science Signaling*, *399*, cm1

Koinuma, K., Yamashita, Y., Liu, W., Hatanaka, H., Kurashina, K., Wada, T., Takada, S., Kaneda, R., Choi, Y.L., Fujiwara, S.I., Miyakura, Y., Nagai, H., & Mano, H. (2006). Epigenetic silencing of AXIN2 in colorectal carcinoma with microsatellite instability. *Oncogene*, *25*, 139-146.

Land, H., Parada, L.F., & Weinberg, R.A. (1983). Tumorigenic conversion of primary embryo fibroblasts requires at least two cooperating oncogenes. *Nature*, *304*, 596-602.

Little, M. P. (1995). Are two mutations sufficient to cause cancer? Some generalizations of the two-mutation model of carcinogenesis of Moolgavkar, Venson and Knudson, and of the multistage model of Armitage and Doll. *Biometrics*, *51*, 1278-1291.

Little, M. P. (1996). Generalizations of the two-mutation and classical multi-stage models of carcinogenesis fitted to the Japanese atomic bomb survivor data, *Journal of Radiological Protection*, *16*, 7-24.

Little, M. P. (2008). Cancer models, ionization and genomic instability: A review. In Tan W.Y. & Hanin L. (Ed.), *Handbook of Cancer Models with Applications*, Singapore and River Edge, NJ: World Scientific.

Little, M.P., Muirhead, C. R., Boice, J.D., & Kleinerman, R.A. (1995). Using multistage models to describe radiation-induced leukaemia. *Journal of Radiological Protection*, *15*, 315-334.

Little, M.P., Muirhead, C.R., & Stiller, C.A. (1996). Modeling lymphocytic leukaemia incidence in England and Wales using generalizations of the two-mutation model of carcinogenesis of MVK. *Statistics in Medicine*, *15*, 1003-1022.

Little, M. P., Vineis, P., & Li, G. (2008). A stochastic carcinogenesis model incorporating multiple types of genomic instability fitted to colon cancer data. *Journal of Theoretical Biology*, *254*, 229-238.

Little, M. P., & Wright, E. G. (2003). A stochastic carcinogenesis model incorporating genomic instability fitted to colon cancer data. *Mathematical Biosciences*, *183*, 111-134.

Lynch, C. J., & Milner, J. (2006). Loss of one p53v allele results in four-fold reduction in p53 mRNA and protein: A basis for p53 haplo-insufficiency. *Oncogene*, *25*, 3463-3470.

Luebeck, E. G., & Moolgavkar, S. H. (2002). Multistage carcinogenesis and colorectal cancer incidence in SEER. *Proceedings of the National Academy of Sciences*, *99*, 15095-15100.

Ma, H., Nguyen, C., Lee, K. S., & Kahn, M., (2005). Differential roles for the coactivators CBP and p300 on TCF/beta-catenin-mediated survivin gene expression, *Oncogene*, *24(22)*, 3619-3631.

Manning, B. D., & Cantley, L.C. (2007). AKT/PKB signaling: navigating downstream. *Cell*, *129(7)*, 1261–74.

Misfeld, J. (1984). The tumor-producing effects of automobile exhaust condensate and of diesel exhaust condensate: mathematical-statistical evaluation of test results. In Pepelko, W.E., Danner, R.M., & Clarke, N.A. (Ed.), *Health Effects of Diesel Engine Emissions. Proceedings of an International Symposium* (Vol.2, pp.1012-1025). Washington, DC: U.S. Govt. Printing Office

Missero, C., D'Errico, M., Dotto, G.P., & Dogliotti, E. (2002). The molecular basis of skin carcinogenesis. In W.B. Coleman, & G.J. Tsongalis (Ed.), *The Molecular Basis of Human Cancer*, Totowa, NJ: Humana Press

Missero, C., Ramon, Y., Cajal, S., & Dotto, G. P. (1991). Escape from transforming growth factor control and oncogene cooperation in skin tumor development. *Proceedings of the National Academy of Sciences*, *88*, 9613-9617.

Moolgovkar, S. H., Dewanji, A., & Venzon, D.J. (1988). A stochastic two-stage for cancer risk assessment: the hazard function and the probability of tumor, *Risk Analysis*, *3*, 383-392.

Moolgavkar, S.H., Dewanji, A., & Luebeck, G. (1989). Cigarette smoking and lung cancer: reanalysis of the British doctors' data, *Journal of the National Cancer Institute*, *81 (6)*, 415-20.

Moolgovkar, S.H., & Knudson, A.G. (1981). Mutation and cancer: a model for human carcinogenesis. *Journal of the National Cancer Institute*, *66*, 1037-1052.

Moolgovkar, S.H., & Venzon, D.J. (1979). Two-event models for carcinogenesis: incidence curve for childhood and adult tumors, *Mathematical Biosciences*, *47*, 55-77.

Moradpour, D., & Wands, J. R. (2003). Molecular pathogenesis of hepatocellular carcinoma. In Zakim D., & Boyer T.D. (Ed.), *Hepatology: A textbook of the Liver Disease* (4th ed., pp. 1333-1354). Philadelphia, PA: Saunders.

Nelder, J., & Wedderburn, R. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General), 135 (3)*, 370–384.

Nelson, W. J., & Nusse, R. (2004). Convergence of Wnt, beta-catenin, and cadherin pathways. *Science, 303(5663)*, 1483–1487.

Nesnow, S., Evans, C., Stead, A., Creason, J., Slaga, T.J., & Triplett, L. L. (1982). Skin carcinogenesis studies of emission extracts. *Developments in Toxicology & Environmental Science*, *10*, 295–320.

Nordling, C.O. (1953). A new theory on the cancer induction mechanism, *British Journal of Cancer*, *7*, 68-72.

Nusse, R. (2005). The Wnt gene, http://www.stanford.edu/~rnusse/wntwindow.html

Peltomaki, P. (2001). Deficient DNA mismatch repair: A common etiologic factor for colon cancer. *Human Molecular Genetics*, *10*, 735-740.

Pogribny, I.P., Rusyn, I., & Beland, F.A. (2008). Epigeneticaspects of genotoxic and nongenotoxic hepatocarcinogenesis: Studies in rodents. *Environmental and Molecular Mutagenesis*, *49*, 9-15.

Portier C.J., & Bailer A.J. (1989). Testing for increased carcinogenicity using a survival-adjusted quantal response test. *Fundament and Applied Toxicology*, *12*, 731–737.

Rak, J., Milsom, C., May, L., Klement, P., & Yu, J. (2006). Tissue Factor in Cancer and Angiogenesis: The Molecular Link between Genetic Tumor Progression, Tumor Neovascularization and Cancer Coagulopathy. *Seminars in Thrombsis and Hemostasis*, *32(1)*, 54-70.

Ruggeri, B., Caamano, J., Goodrow, T., DiRado, M., Bianchi, A., Trono, D., Conti, C.J., & Klein-Szanto, A.J. (1991). Alterations of the p53 tumor suppressor gene during mouse skin tumor progression. *Cancer Research*, *51*, 6615-6621.

Saez,E., Rutberg, S.E., Mueller, E., Oppenheim, H., Smoluk, J., Yuspa, S.H., & Spiegelman, B.M. (1995). c-fos is reguired for malignant progression of skin tumors. *Cell*, *82*, 721-732.

Salmena, L., Carracedo, A., & Pandolfi, P. P. (2008). Tenets of PTEN tumor suppression. *Cell*, *133(3)*, 403–414.

Schulte, G., & Bryja, V. (2007). The Frizzled family of unconventional G-protein-coupled receptors. *Trends in Pharmacological Sciences*, *28(10)*, 518–25.

Schmierer, B., & Hill, C.S. (2007). TGFbeta-SMAD signal transduction: molecular specificity and functional flexibility. *Nature Review (Molecular Cell Biology), 8(12*), 970–82.

Schwab, M., Varmus H.E., & Bishop, J.M., (1985). Human N-myc gene contributes to neoplastic transformation of mammalian cells in culture. *Nature*, *316*, 160-162.

Shih, W-L, Yu, M-W, Chen, P-J, Yeh, S-H, Lo, M.T., Chang, H.C., Liaw, Y.F., Lin, S.M., Liu, C.J., Lee, S.D., Lin, C.L., Hsiao, C.K., Yang, S.Y., & Chen, C.J. (2006). Localization of a susceptibility locus for hepatocellular carcinoma to chromosome 4q in a hepatitis B hyperendemic area, *Oncogene*, *25*, 3219-3224.

Smith, A.F.M., & Gelfant, A. E. (1992). Bayesian statistics without tears: A sampling-resampling perspective. *American Statistician*, *46*, 84-88.

Sparks, A. B., Morin, P. J., Vogelstein, B., & Kinzler, K. W. (1998). Mutational analysis of the APC/beta-catenin/Tcf pathway in colorectal cancer. *Cancer Research*, *58*, 1130-1134.

Stoick-Cooper, C. L., Moon, R. T., & Weidinger, G. (2007). Advances in signaling in vertebrate regeneration as a prelude to regenerative medicine. *Genes and Development*, *21(11)*, 1292–315.

Tan, W.Y. (1991). *Stochastic Models of Carcinogenesis*. New York, NY: Marcel Dekker.

Tan, W.Y. (2000). Stochastic Modeling of AIDS Epidemiology and HIV Pathogenesis. Singapore and River Edge, NJ: World Scientific.

Tan, W.Y. (2002). Stochastic Models with Applications to Genetics, Cancers, AIDS and Other Biomedical Systems. Singapore and River Edge, NJ: World Scientific.

Tan, W. Y., & Chen, C. W. (1991). A multiple pathway model of carcinogenesis involving one stage model and two-stage model, in Arino, O., Axelrod, D.E., Kimmel, M., & DekkerMath, M (Ed.), *Mathematical Population Dynamics* (pp. 469-482), New York, NY: Malcel Dekker.

Tan, W. Y., & Chen, C. W. (1995). A bivariate stochastic model of carcinogenesis involving two cancer tumors. In Proceeding of the 9th international conference on mathematics and computer modeling. University of California, Berkeley, CA.

Tan, W. Y., & Chen, C. W. (1998). Stochastic modeling of carcinogenesis: some new insight. *Mathematical Computation Modeling*, *28*, 49-71.

Tan, W. Y., & Chen, C. W. (2000). Assessing effects of changing environment by a multiple-pathway model of carcinogenesis. *Mathematical Computation Modeling*, *32*, 229-250.

Tan, W. Y., & Chen, C. W. (2005). Cacner stochastic models. *Encyclopedia of statistical sciences*, New York, NY: John Wiley and Sons.

Tan, W. Y., Chen, C. W., & Wang, W. (2000). A generalized stage space model of carcinoneisis. In Paper presented at the 2000 International Biometric Conference. UC Berkeley, CA.

Tan, W. Y., Chen, C. W., & Wang, W. (2001). Stochastic modeling of carcinogenesis by state space models: A new approach. *Mathematical Computation Modeling*, *33*, 1323-1345.

Tan, W. Y., & Gastardo, M.T.C. (1985). On the assessment of effects of environmental agents on cancer tumor development by a two-stage model of carcinogenesis, *Mathematical Biosciences*, *73*, 143-155.

Tan, W.Y., & Yan, X.W. (2010). A new stochastic and state space model of human colon cancer incorporating multiple pathways. *Biology Direct Theramic Series: Mathematics and Evolution of Cancer*, *5:26*.

Tan, W. Y., & Ye, Z. Z. (2000). Estimation of HIV and HIV incubation via state space models. *Mathematical Biosciences*, *167*, 31-40.

Tan, W.Y., Zhang, L.J., Chen ,W., & Zhu, J.M. (2008a). A stochastic model of human colon cancer involving multiple pathways. In Tan W.Y., & Hanin L., *Handbook of Cancer Models with Applications* (Ed.), Singapore and River Edge, NJ: World Scientific.

Tan, W.Y., Zhang, L.J., Chen, W., & Zhu, J.M. (2008b). A stochastic model of human colon cancer involving multiple pathways. In Tan W.Y., & Hanin L., *Handbook of Cancer Models with Applications* (Ed.), Singapore and River Edge, NJ: World Scientific.

Thorgeirsson, S. S., & Grisham, J.W. (2002). Molecular pathogenesis of human hepatocellular carcinoma. *Nature Genetics*, *31*, 339-346.

Verheyen, E. M. (2007). Opposing effects of Wnt and MAPK on BMP/Smad signal duration. *Developmental Cell, 13(6)*, 755–756.

Villanueva, A., Newell, P., Chiang, D.Y. Scott, L.F., & Llovet, J.M. (2007). Genomics and signaling pathways in hepatocellular carcinoma. *Seminars in Liver Disease*, *27*, 55-76.

Wands, J.R. (2004). Prevention of hepatocellular carcinoma. *New England Journal of Medicine*, *351*, 1567-1570.

Ward, R., Meagher, A., Tomlinson, I., O'Connor, T., Norriea, M., Wu, R., & Hawkins, N. (2001). Microsatellite instability and the clinicopathological features of sporadic colorectal cancer. *Gut*, *48*, 821-829.

Waters, M.D., Sandhu, S. S., Huisingh, J.L., Claxton, L., & Nesnow, S. (1981). *Application of Short-term bioassays to the Analysis of Complex Environmental Mixtures,* New York, NY: Plenum Press.

Weinberg R.A. (2007). *The Biology of Cancer*. New York, NY: Garland Sciences, Taylor and Frances Group.

Willert, K., & Jones, K. A. (2006). Wnt signaling: is the party in the nucleus? *Genes and Development*, *20(11)*, 1394–1404.

Xiao, Y. T., Xiang, L. X., & Shao, J. Z. (2007). Bone morphogenetic protein. *Biochemical and Biophysical Research Communications, 362(3)*, 550–553.

Yakovlev A.Y., & Tsodikov A.D. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. Singapore and River Edge, NJ: World Scientific.

Yancopoulos, G. D., Desiderio, S. V., Paskind, M., Kearney, J. F., Baltimore, D. & Alt, F. W. (1984). *Nature, 311*, 727-*733.*

Yeh, S.H., Chen, P.J., Shau, W.Y., Chen, Y.W., Lee, P.H., Chen, J.T., & Chen D.S. (2001). Chromosomal allelic inbalance evolving from liver cirrhosis to hepatocellular carcinoma. *Gastroenterology*, *121*, 699-709.

Yuspa, S.H. (1994). The pathogenesis of squamous cell cancer: Lessons learned from studies of skin carcinogenesis. *Cancer Research*, *54*, 1178-1189.

Zheng, Q. H., Ma, L.W., Zhu, W.G., Zhang, Z. Y., & Tong, T. J. (2006). p21(Waf1/Cip1) plays a critical role in modulating senescence through changes of DNA methylation. *Journal of Cell Biochemistry*, *98(5)*, 1230-1248.