

**RESEARCH**

**Open Access**

# A new stochastic and state space model of human colon cancer incorporating multiple pathways

Wai Y Tan<sup>\*</sup>, Xiao W Yan<sup>\*</sup>

## Abstract

**Background and Purpose:** Studies by molecular biologists and geneticists have shown that tumors of human colon cancer are developed from colon stem cells through two mechanisms: The chromosomal instability and the micro-satellite instability. The purpose of this paper is therefore to develop a new stochastic and state space model for carcinogenesis of human colon cancer incorporating these biological mechanisms.

**Results:** Based on recent biological studies, in this paper we have developed a state space model for human colon cancer. In this state space model, the stochastic system is represented by a stochastic model, involving 2 different pathways—the chromosomal instability pathway and the micro-satellite instability pathway; the observation, cancer incidence data, is represented by a statistical model. Based on this model we have developed a generalized Bayesian approach to estimate the parameters through the posterior modes of the parameters via Gibbs sampling procedures. We have applied this model to fit and analyze the SEER data of human colon cancers from NCI/NIH.

**Conclusions:** Our results indicate that the model not only provides a logical avenue to incorporate biological information but also fits the data much better than other models including the 4-stage single pathway model. This model not only would provide more insights into human colon cancer but also would provide useful guidance for its prevention and control and for prediction of future cancer cases.

**Reviewers:** This article was reviewed by M.P. Little and M. Kimmel

## Background

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 ([1-9]). 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 et al. [10], Little and Wright [11] and Little et al. [12]. We note that besides the multiple pathways considered above, Little and Wright [11], Little et al. [12] and Little [13] 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.

For developing biologically supported stochastic model of carcinogenesis, in Section 2 we present the most

<sup>\*</sup> Correspondence: waitan@memphis.edu; xyan@memphis.edu  
Department of Mathematical Sciences, The University of Memphis, Memphis, TN 38152-6429, USA

recent cancer biology of human colon cancer. Using results from Section 2, we develop in Section 3 a stochastic model for carcinogenesis of human colon cancer involving multiple pathways. In Section 4 we derive a statistical model for cancer incidence data of human colon cancer. By combining models from Sections 3 and 4, in Section 5 we develop a state space model for human colon cancer. In Section 6, by using the state space model in Section 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 7 we apply the model and methods to the colon cancer incidence data from SEER. Finally in Section 8, we discuss the usefulness of the model and methods and provide some conclusions.

### 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  $\beta$ -catenin gene in chromosome 3p22.

#### The CIN (LOH) Pathway of Human Colon Cancer (The APC- $\beta$ -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 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 2**. This pathway has also been referred to as APC- $\beta$  - catenin - Tcf - myc pathway because it involves the destruction complex GSK-3 $\beta$  - Axin-APC which phosphorylates the  $\beta$ -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  $\beta$ -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  $\beta$ -catenin protein in the cytoplasm also binds with E-cadherin and  $\alpha$ -catenin to disrupt the gap junction between cells, leading to migration and metastasis of cancer tumors.)

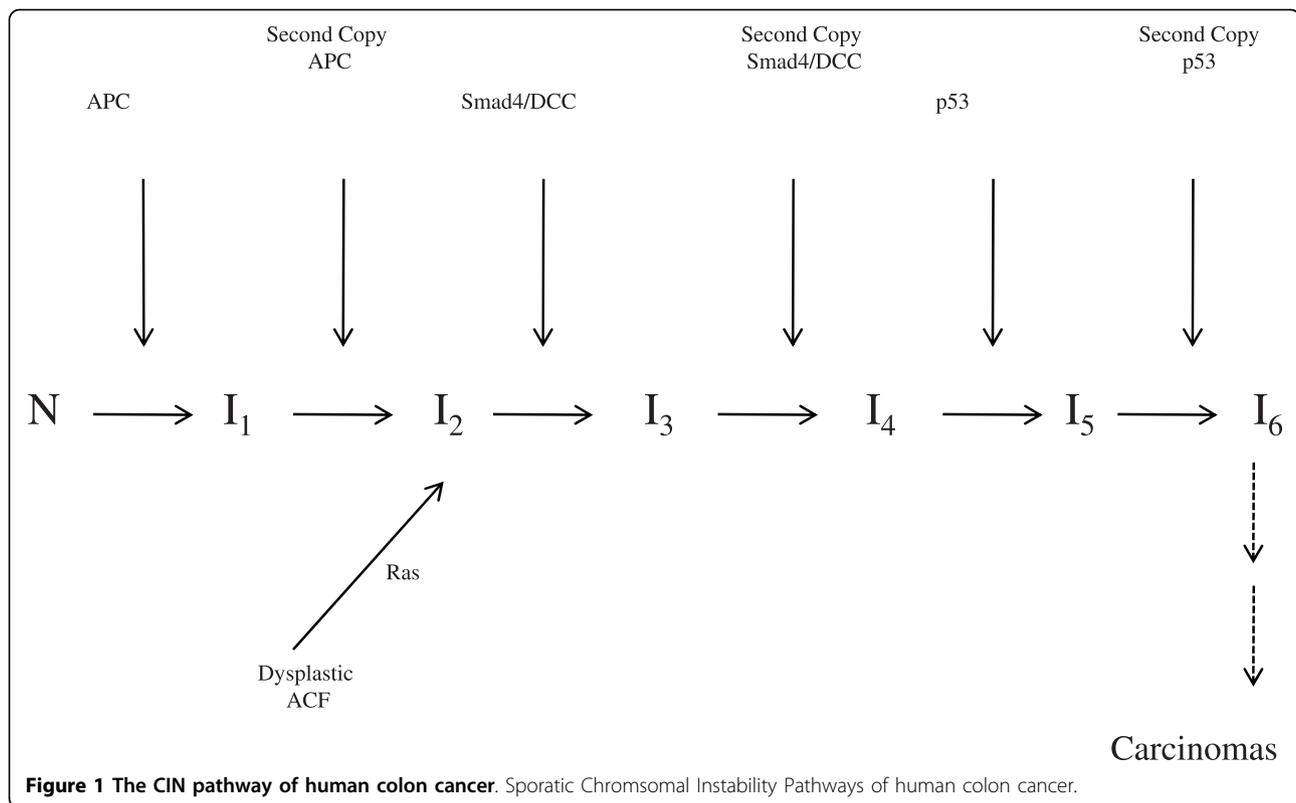
Morphological studies have indicated that inactivation, or loss or mutation of APC creates dysplastic aberrant crypt foci (ACF) which grow into dysplastic adenomas.

These adenomas grow to a maximum size of about 10 mm<sup>3</sup>; 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 [14]. This pathway is represented schematically by Figure 1.

The model in Figure 1 is a 6-stage model. However, because of the haplo-insufficiency of the Smad4 gene (see Alberici et al.[15]) and the haplo-insufficiency of the p53 gene ([16]), 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 ([17]). Recent biological studies by Green and Kaplan [4] 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 1:** As observed by Sparks et al. [8], instead of the APC gene, this pathway can also be initiated by mutation of the oncogene  $\beta$ -catenin gene; however, the proportion of human colon cancer due to mutation of  $\beta$ -catenin is very small (less than 1%) as compared to the APC gene, due presumably to the contribution of the APC on chromosome instability ([4]). Similarly, the destruction complex can become inactive either by the inhibition of GSK-3 $\beta$  through the Wnt signalling pathway (see [18]) or the inactivation or mutation of the Axin protein, leading to accumulation of the  $\beta$ -Catenin proteins in the cytoplasm; but the proportion of colon cancer caused by inhibition of GSK-3 $\beta$  is also very small as compared to the colon cancer cases caused by the CIN and the MSI pathways.

**Remark 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 ([4]). 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 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

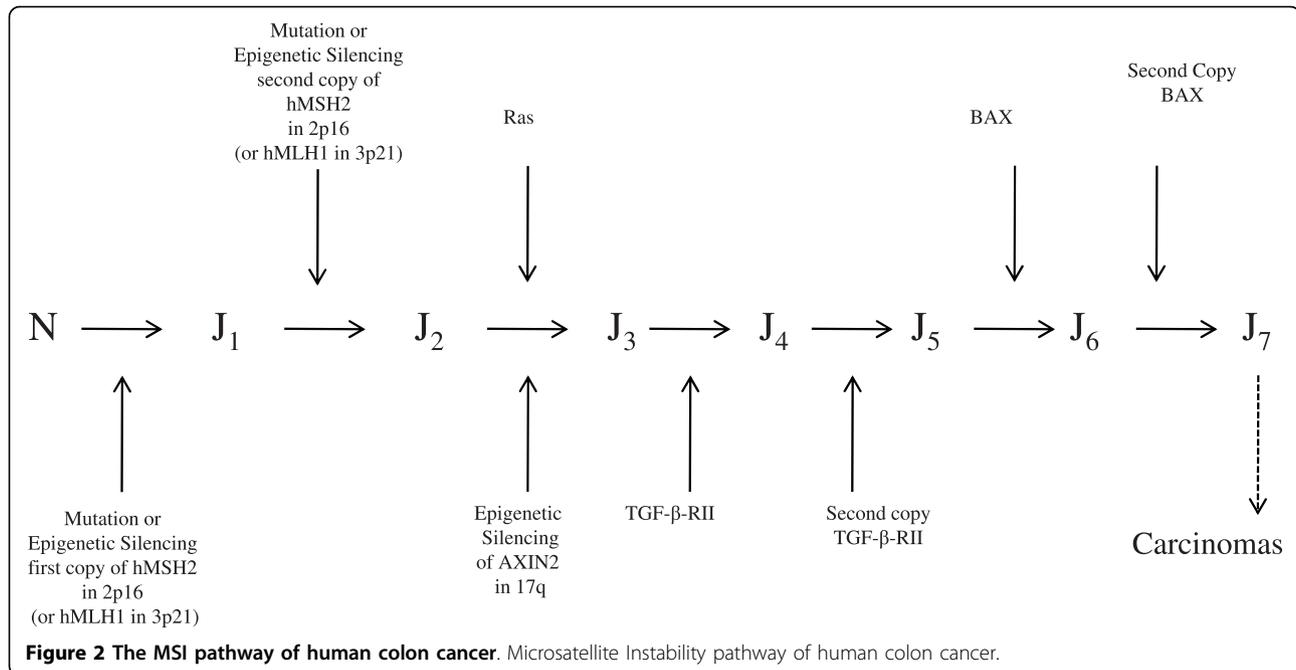
This pathway accounts for about 15% of all colon cancers and appears mostly in the right colon. It has been referred to as the MSI pathway or the mutator phenotype pathway because it is initiated by the mutations or epigenetic methylation of the mis-match repair genes (mostly hMLH1 in chromosome 3p21 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- $\beta$  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 2. As in the LOH pathway, assuming haplo-insufficiency of tumor

suppressor genes, one may approximate this pathway by a 5-stage model.

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- $\beta$  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,  $I_i$  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 Signalling Pathways for Human Colon Cancer

Recent biological studies ([18,19]) have shown that both the CIN and the MSI pathways involve the Wnt signalling pathway and the destruction complex (this complex is a downstream of the Wnt signalling pathway), the TGF- $\beta$  inhibiting signalling pathway and the p53-Bax apoptosis signalling pathway, but different genes in the CIN and MSI pathways are affected in these signalling processes. In the CIN pathway, the affected gene is the APC gene in the Wnt signalling, the Smad4 in the TGF- $\beta$  signalling and the p53 gene in the p53-Bax signalling; on the other hand, in the MSI pathway, the affected gene is the Axin 2 gene in the Wnt signalling, the



TGF- $\beta$  -Receptor II in the TGF- $\beta$  signalling and the Bax gene in the p53-Bax signalling.

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 ([18,20,21]). In fact, Breivik and Gaudernack [20] showed that in human colon cancer, either methylating carcinogens or hyper-methylation at  $C_pG$  islands would lead to G/T mismatch which in turn leads to Mismatch 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 checkpoints (MCP), thus promoting CIN (Chromosomal Instability). A recent review by Baylin and Ohm [18] 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.

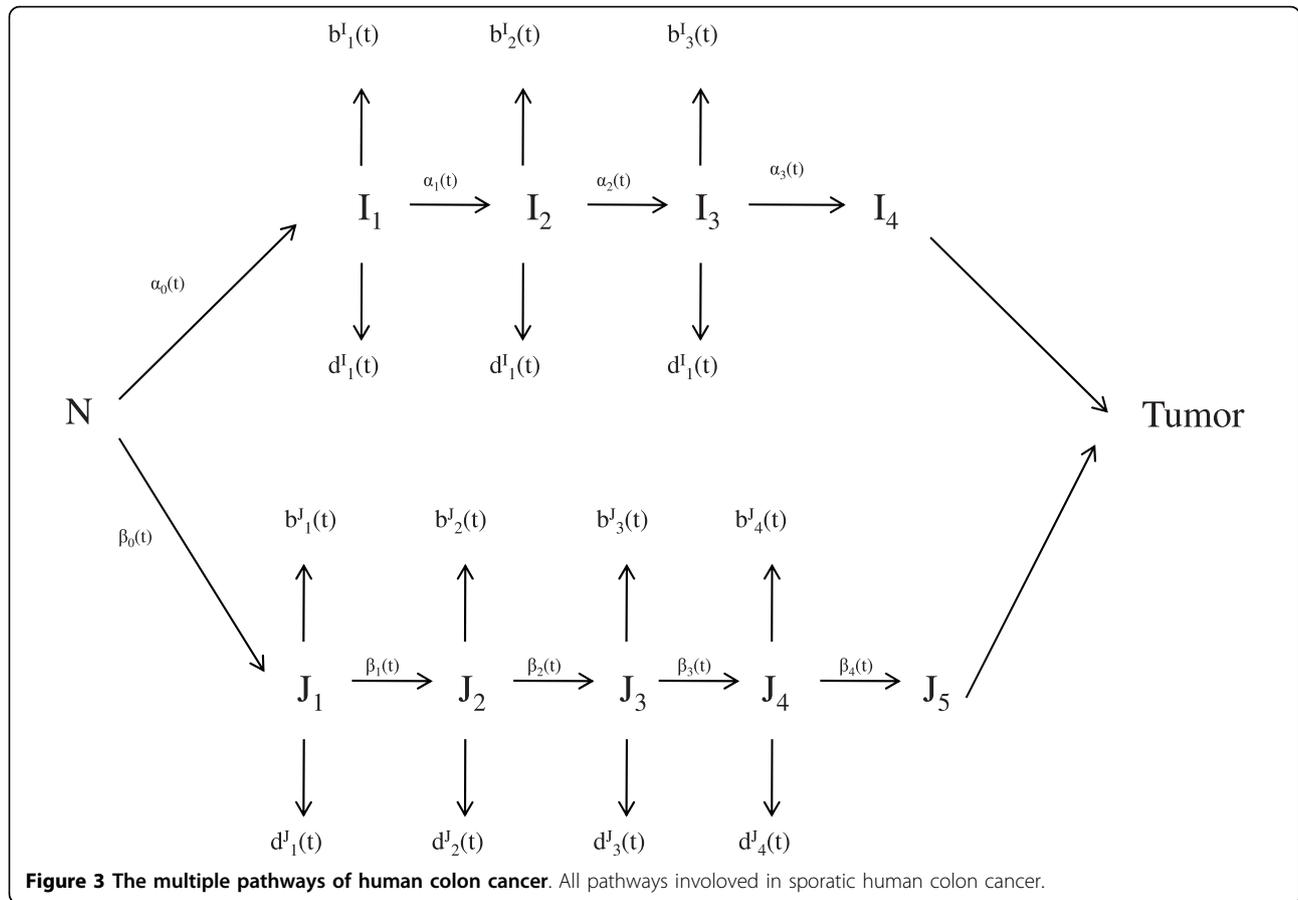
Based on the above biological studies, in this paper we thus postulate that the incidence data of human colon cancer are described and generated by a multi-stage model involving 2 pathways as defined above. In this paper, because of haploid-insufficiency of the tumor suppressor genes {*Smad4*, *p53*, *Axin*, *Bax*, *TGF -  $\beta$  - ReceptorII*}, 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 2, it follows that the stochastic multi-stage model for human colon cancer can be represented schematically by Figure 3.

In Figure 3, the model assumes that cancer tumors are generated by two pathways with pathway 1 as a  $k_1$ -stage multi-stage model involving  $I_l$  ( $l = 1, \dots, k_1$ ) cells and with pathway 2 as a  $k_2$ -stage multi-stage model involving  $J_r$  ( $r = 1, \dots, k_2$ ) cells. (For human colon cancer,  $k_1 = 4$ ,  $k_2 = 5$ .) The state variables are then  $\underline{X}(t) = \{I_l(t), l = 1, \dots, k_1 - 1, J_r(t), r = 1, \dots, k_2 - 1\}$  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  $\{l = 1, \dots, k_1 - 1$  ( $r = 1, \dots, k_2 - 1$ ) 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  $\underline{X}(t + \Delta t)$  develop from  $\underline{X}(t)$  through cell divisions during  $(t, t + \Delta t]$ , one may practically assume that  $(\underline{X}(t), t \geq t_0)$  is a Markov process with continuous time, where  $t_0$  represents time at birth; one 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 ([22,23]). 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. In this case, as illustrated in Tan [24], one may use standard Markov theory to derive the probability



**Figure 3** The multiple pathways of human colon cancer. All pathways involved in sporadic human colon cancer.

generating function (pgf) of the probabilities of these variables and hence the probability distribution of these variables. Let  $\psi(x_l, l = 1, \dots, k_1 - 1, y_r, r = 1, \dots, k_2 - 1, z; t_0, t) = \psi(\underline{x}, \underline{y}, z; t_0, t)$  denote the pgf of  $\{X_l(t), T(t)\}$ . Let  $\{\alpha_l(t), \beta_r(t), b_l^{(l)}(t), b_r^{(r)}(t), d_l^{(l)}(t), d_r^{(r)}(t)\}$  denote the mutation rates, the birth rates and the death rates of  $\{I_l, J_r\}$  cells as given in Table 1 respectively.

If  $T(t)$  is Markov, then by using the method of Kolmogorov forward equation of these variables (Tan [24]), it

can readily be shown that  $\psi(\underline{x}, \underline{y}, z; t_0, t)$  satisfies the following partial differential equation (pde):

$$\begin{aligned} & \frac{\partial}{\partial t} \psi(\underline{x}, \underline{y}, z; t_0, t) \\ &= [\lambda_I(t)(x_1 - 1) + \lambda_J(t)(y_1 - 1)] \psi(\underline{x}, \underline{y}, z; t_0, t) \\ &+ \sum_{l=1}^{k_1-1} g_1(x_l, x_{l+1}; t) \frac{\partial}{\partial x_l} \psi(\underline{x}, \underline{y}, z; t_0, t) \\ &+ \sum_{r=1}^{k_2-1} g_2(y_r, y_{r+1}; t) \frac{\partial}{\partial y_r} \psi(\underline{x}, \underline{y}, z; t_0, t), \end{aligned} \quad (1)$$

**Table 1** Transition rates and transition probabilities for human colon carcinogenesis

1 N	→	1 N, 1 I <sub>1</sub>	$\alpha_0(t)\Delta t$
1 N	→	1 N, 1 J <sub>1</sub>	$\beta_0(t)\Delta t$
1 I <sub>l</sub>	→	2 I <sub>l</sub>	$b_l^{(l)}(t)\Delta t$
1 I <sub>l</sub>	→	death	$d_l^{(l)}(t)\Delta t$
1 I <sub>l</sub>	→	1 I <sub>l</sub> , 1 I <sub>l+1</sub>	$a_l(t)\Delta t$
	$l = 1, \dots, k_1 - 1$		
1 J <sub>r</sub>	→	2 J <sub>r</sub>	$b_r^{(r)}(t)\Delta t$
1 J <sub>r</sub>	→	death	$d_r^{(r)}(t)\Delta t$
1 J <sub>r</sub>	→	1 J <sub>r</sub> , 1 J <sub>r+1</sub>	$\beta_r(t)\Delta t$
	$r = 1, \dots, k_2 - 1$		

where  $\lambda_I(t) = N(t)\alpha_0(t)$ ,  $\lambda_J(t) = N(t)\beta_0(t)$ ,  $g_2(y_r, y_{r+1}; t) = y_r(y_r - 1)b_r^{(r)}(t) - (y_r - 1)d_r^{(r)}(t) + y_r(y_{r+1} - 1)\beta_r(t)$ ,  $g_1(x_l, x_{l+1}; t) = x_l(x_l - 1)b_l^{(l)}(t) - (x_l - 1)d_l^{(l)}(t) + x_l(x_{l+1} - 1)\alpha_l(t)$ , and the initial condition is  $\psi(\underline{x}, \underline{y}, z; t_0, t_0) = 1$  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 et al.[22,23]). 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  $T(t)$ .

### 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  $(t, t + \Delta t]$   $\{l = 1, \dots, k_1 - 1$  ( $r = 1, \dots, k_2 - 1$ ) $\}$ ,  $D_l^{(I)}(t)$  ( $D_r^{(J)}(t)$ ) the number of deaths of the  $I_l$  ( $J_r$ ) initiated cells during  $(t, t + \Delta t]$   $\{l = 1, \dots, k_1 - 1$  ( $r = 1, \dots, k_2 - 1$ ) $\}$  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  $(t, t + \Delta t]$   $\{l = 1, \dots, k_1 - 1$  ( $r = 1, \dots, k_2 - 1$ ) $\}$ .

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:

$$I_l(t + \Delta t) = I_l(t) + M_{l-1}^{(I)}(t) + B_l^{(I)}(t) - D_l^{(I)}(t), \quad (2)$$

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

$$J_r(t + \Delta t) = J_r(t) + M_{r-1}^{(J)}(t) + B_r^{(J)}(t) - D_r^{(J)}(t), \quad (3)$$

$$r = 1, \dots, k_2 - 1,$$

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 1, 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,

$$M_0^{(I)}(t) | N(t) \sim \text{Poisson}\{\lambda_I(t)\Delta t\}, \quad (4)$$

independently of  $M_0^{(J)}(t)$ ;

$$M_0^{(J)}(t) | N(t) \sim \text{Poisson}\{\lambda_J(t)\Delta t\}, \quad (5)$$

independently of  $M_0^{(I)}(t)$ ;

for  $l = 1, 2, \dots, k_1 - 1$ ,

$$\{B_l^{(I)}(t), D_l^{(I)}(t)\} | I_l(t) \sim \text{ML}\{I_l(t); b_l^{(I)}(t)\Delta t, d_l^{(I)}(t)\Delta t\}; \quad (6)$$

for  $r = 1, \dots, k_2 - 1$ ,

$$\{B_r^{(J)}(t), D_r^{(J)}(t)\} | J_r(t) \sim \text{ML}\{J_r(t); b_r^{(J)}(t)\Delta t, d_r^{(J)}(t)\Delta t\}; \quad (7)$$

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 ( $10^{-5} \sim 10^{-8}$ ), 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,

$$M_l^{(I)}(t) | I_l(t) \sim \text{Poisson}\{I_l(t)\alpha_l(t)\Delta t\}, \quad (8)$$

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

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

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

$$M_r^{(J)}(t) | J_r(t) \sim \text{Poisson}\{J_r(t)\beta_r(t)\Delta t\}, \quad (9)$$

$$r = 1, \dots, k_2 - 1,$$

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

Using the probability distributions given by equations (5)-(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:

$$\begin{aligned} dI_l(t) &= I_l(t + \Delta t) - I_l(t) \\ &= M_{l-1}^{(I)}(t) + B_l^{(I)}(t) - D_l^{(I)}(t) \\ &= \{I_{l-1}(t)\alpha_{l-1}(t) + I_l(t)\gamma_l^{(I)}(t)\}\Delta t \\ &\quad + e_l^{(I)}(t)\Delta t, l = 1, \dots, k_1 - 1, \end{aligned} \quad (10)$$

$$\begin{aligned} dJ_r(t) &= J_r(t + \Delta t) - J_r(t) \\ &= M_{r-1}^{(J)}(t) + B_r^{(J)}(t) - D_r^{(J)}(t) \\ &= \{J_{r-1}(t)\beta_{r-1}(t) + J_r(t)\gamma_r^{(J)}(t)\}\Delta t \\ &\quad + e_r^{(J)}(t)\Delta t, r = 1, \dots, k_2 - 1, \end{aligned} \quad (11)$$

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  $\{e_l^{(l)}(t)\Delta t, e_r^{(r)}(t)\Delta t\}$  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  $\{I_l(t), l = 1, \dots, k_1 - 1, J_r(t), r = 1, \dots, k_2 - 1\}$ . 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:

$$\begin{aligned} \text{Var}\{e_l^{(l)}(t)\Delta t\} &= EI_{l-1}(t)\alpha_{l-1}(t)\Delta t \\ &+ EI_l(t)[b_l^{(l)}(t) + d_l^{(l)}(t)]\Delta t \\ &+ o(\Delta t), \text{ for } l = 1, \dots, k_1 - 1, \\ \text{Var}\{e_r^{(r)}(t)\Delta t\} &= EJ_{r-1}(t)\beta_{r-1}(t)\Delta t \\ &+ EJ_r(t)[b_r^{(r)}(t) + d_r^{(r)}(t)]\Delta t \\ &+ o(\Delta t), \text{ for } r = 1, \dots, k_2 - 1, \end{aligned}$$

where  $I_0(t) = J_0(t) = N(t)$ .

### The Expected Numbers

Let  $u_l(l, t) = E [I_l(t)]$  and  $u_j(r, t) = E [J_r(t)]$  denote the expected numbers of  $I_l(t)$  and  $J_r(t)$  respectively and write  $u_l(0, t) = u_j(0, t) = N(t)$ . Using equations (11)-(12), we have the following differential equations for these expected numbers:

$$\frac{d}{dt} u_l(l, t) = u_l(l, t)\gamma_l^{(l)}(t) + u_l(l-1, t)\alpha_{l-1}(t), \quad (12)$$

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

$$\frac{d}{dt} u_j(r, t) = u_j(r, t)\gamma_r^{(r)}(t) + u_j(r-1, t)\beta_{r-1}(t), \quad (13)$$

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

The solution of the above equations are:

$$\begin{aligned} u_l(1, t) &= \int_{t_0}^t \lambda_l(x) e^{\int_x^t \gamma_1^{(1)}(z) dz} dx, \\ u_j(1, t) &= \int_{t_0}^t \lambda_j(x) e^{\int_x^t \gamma_1^{(1)}(z) dz} dx, \\ u_l(l, t) &= \int_{t_0}^t u_l(l-1, x) e^{\int_x^t \gamma_u^{(l)}(z) dz} dx, \\ &\text{for } l = 2, \dots, k_1 - 1, \\ u_j(r, t) &= \int_{t_0}^t u_j(r-1, x) e^{\int_x^t \gamma_r^{(r)}(z) dz} dx, \\ &\text{for } r = 2, \dots, k_2 - 1. \end{aligned}$$

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

$$\begin{aligned} u_l(1, t) &= \frac{\lambda_l}{\gamma_1^{(1)}} [e^{\gamma_1^{(1)} t} - 1], \\ u_j(1, t) &= \frac{\lambda_j}{\gamma_1^{(j)}} [e^{\gamma_1^{(j)} t} - 1], \\ u_l(l, t) &= \sum_{u=1}^l A_l(u) e^{\gamma_u^{(l)} t}, \\ &\text{for } l = 1, \dots, k_1 - 1; \\ u_j(r, t) &= \sum_{u=1}^r B_r(u) e^{\gamma_u^{(r)} t}, \\ &\text{for } r = 1, \dots, k_2 - 1 \end{aligned}$$

$$A_l(u) = \lambda_l \left\{ \prod_{u=1}^{l-1} \left\{ \prod_{v=1, v \neq u}^l (\gamma_u^{(l)} - \gamma_v^{(l)}) \right\}^{-1} \cdot B_r(u) = \right.$$

where

$$\lambda_j \left\{ \prod_{u=1}^{r-1} \left\{ \prod_{v=1, v \neq u}^r (\gamma_u^{(j)} - \gamma_v^{(j)}) \right\}^{-1} \right.$$

### The Probability Distribution of State Variables and Transition Variables

Although  $T(t)$  is not Markov, the random vector  $\{\underline{X}(t), t \geq t_0\}$  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  $\lambda$ . Then, using the probability distributions given by equations (5)-(10), the transition probability of this Markov process is, to order of  $o(\Delta t)$ :

$$\begin{aligned} &P\{\underline{X}(t + \Delta t) | \underline{X}(t)\} \\ &= \prod_{u=1}^{k_1-1} \left\{ \sum_{l_u=0}^{I_u(t)} \sum_{i_u=0}^{I_u(t)-l_u} h[a(l_u, i_u; t); I_{r-1}(t)\alpha_{u-1}(t)\Delta t] \right. \\ &\times f[l_u, i_u; I_u(t), b_u^{(l)}(t)\Delta t, d_u^{(l)}(t)\Delta t] \} \\ &\times \prod_{v=1}^{k_2-1} \left\{ \sum_{m_v=0}^{J_v(t)} \sum_{j_v=0}^{J_v(t)-m_v} h[b(m_v, j_v; t); J_{v-1}(t)\beta_{v-1}(t)\Delta t] \right. \\ &\times f[m_v, j_v; I_v(t), b_v^{(j)}(t)\Delta t, d_v^{(j)}(t)\Delta t] \}, \end{aligned}$$

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

The above transition probability and hence the probability distribution of  $\underline{X}(t)$  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 variables  $\underline{U}(t) = \{B_l^{(l)}(t), D_l^{(l)}(t), l = 1, \dots, k_1 - 1, B_r^{(r)}(t), D_r^{(r)}(t), r = 1, \dots, k_2 - 1\}$ . (In what follows we will refer these variables as the transition variables, unless otherwise stated.)

Put  $\underline{Z}(t) = \{\underline{X}(t)', \underline{U}(t - \Delta t)'\}'$ . Then  $\{\underline{Z}(t), t \geq t_0\}$  is Markov with continuous time. Using the probability distributions of the transition random variables given by equations (5)-(10), the transition probability  $P\{\underline{Z}(t + \Delta t) | \underline{Z}(t)\}$  is

$$P\{\underline{Z}(t + \Delta t) | \underline{Z}(t)\} = P\{\underline{X}(t + \Delta t) | \underline{X}(t), \underline{U}(t)\} \times P\{\underline{U}(t) | \underline{X}(t)\}, \quad (14)$$

where

$$P\{\underline{U}(t) | \underline{X}(t)\} = \prod_{l=1}^{k_1-1} f\{B_l^{(l)}(t), D_l^{(l)}(t); I_l(t), b_l^{(l)}(t)\Delta t, d_l^{(l)}(t)\Delta t\} \times \prod_{r=1}^{k_2-1} f\{B_r^{(r)}(t), D_r^{(r)}(t); J_r(t), b_r^{(r)}(t)\Delta t, d_r^{(r)}(t)\Delta t\}; \quad (15)$$

and

$$P\{\underline{X}(t + \Delta t) | \underline{U}(t), \underline{X}(t)\} = \prod_{l=1}^{k_1-1} h\{u_l(l, t); I_l(t), \alpha_l(t)\Delta t\} \times \prod_{r=1}^{k_2-1} h\{u_r(r, t); J_r(t), \beta_r(t)\Delta t\}, \quad (16)$$

where  $u_l(l, t) = I_l(t + \Delta t) - I_l(t) - B_l^{(l)}(t) + D_l^{(l)}(t)$  for  $l = 1, \dots, k_1 - 1$  and  $u_r(r, t) = J_r(t + \Delta t) - J_r(t) - B_r^{(r)}(t) + D_r^{(r)}(t)$  for  $r = 1, \dots, k_2 - 1$ .

The probability distribution given by equation (15) will be used to derive estimates and predicted numbers of state variables. This is discussed in Section 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  $\{(y_j, n_j), j = 1, \dots, n\}$ , 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

**Table 2 Colon Cancer Data from SEER(overall population)**

Age Group	Number of People at Risk	Observed Colon Cancer Cases	Total Predicted Colon Cancer
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.

during the age group. Given in Table 2 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  $I_{k_1}$  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  $J_{k_2-1}$  cells by mutation of  $J_{k_2-1}$  cells.

Let  $P_T^{(l)}(s, t)$  ( $P_T^{(j)}(s, t)$ ) 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  $\{I_{k_1-1}(s), s \leq t\}$  is Poisson with mean  $\omega_1(t)$  independently of  $T_2(t)$ , where

$$\omega_1(t) = \int_{t_0}^t I_{k_1-1}(x) \alpha_{k_1-1}(x) P_T^{(l)}(x, t) dx.$$

Similarly, to order of  $o(\Delta t)$ , the conditional probability distribution of  $T_2(t)$  given  $\{J_{k_2-1}(s), s \leq t\}$  is Poisson with mean  $\omega_2(t)$  independently of  $T_1(t)$ , where

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

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

$$\begin{aligned} Q_i(j) &= E\{e^{-\omega_i(t_{j-1})} - e^{-\omega_i(t_j)}\} \\ &= E\{e^{-\omega_i(t_{j-1})}(1 - e^{R_i(t_{j-1}, t_j)})\}, \end{aligned}$$

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

$$\begin{aligned} Q_T(j) &= 1 - [1 - Q_1(j)][1 - Q_2(j)] \\ &= Q_1(j) + Q_2(j) - Q_1(j)Q_2(j). \end{aligned}$$

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

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

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

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  $P_T^{(I)}(s,t) \sim 1$  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 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:

$$Y_j \sim \text{Binomial}\{n_j, Q_T(j)\}. \quad (17)$$

Because  $n_j$  is very large and  $Q_T(j)$  is very small, approximately  $Y_j$  is Poisson with mean  $\tau_j = n_j Q_T(j)$ .

Notice that to the order of  $Max\{o(\alpha_{k_1-1}(t)), o(\beta_{k_2-1}(t))\}$ ,  $\tau_j$  (and hence the probability distribution of  $Y_j$ ) depends on the stochastic model of colon 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  $(t_{j-1}, t_j]$ .

#### 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 ([25], 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 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; 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 4.

#### The Stochastic System Model and the State Variables

Putting  $\Delta t = 1$  for some fixed small interval, then the staging variables are  $\mathbf{X} = \{\underline{X}(t), t = t_0, t_0 + 1, \dots, t_M\}$  and the transition variables are  $\mathbf{U} = \{\underline{U}(t), t = t_0, t_0 + 1, \dots, t_M - 1\}$ . From results in Section (3.3), the joint probability distribution of  $\{\mathbf{X}, \mathbf{U}\}$  given the parameters  $\Theta$  is:

$$\begin{aligned} &P\{\mathbf{X}, \mathbf{U} \mid \Theta\} \\ &= \prod_{t=t_0+1}^{t_M} P\{\underline{X}(t) \mid \underline{X}(t-1), \underline{U}(t-1)\} \\ &\times P\{\underline{U}(t-1) \mid \underline{X}(t-1)\}, \end{aligned} \quad (18)$$

where  $P\{\underline{U}(t-1) \mid \underline{X}(t-1)\}$  and  $P\{\underline{X}(t) \mid \underline{X}(t-1), \underline{U}(t-1)\}$  are given by equations (16) and (17) respectively and where  $\Theta = \{\lambda_l, \lambda_r, \alpha_l(t), \beta_r(y), b_l^{(I)}(t), d_l(t)^{(I)}(t), b_r^{(I)}(t), d_r(t)^{(I)}(t), l = 1, \dots, k_1 - 1, r = 1, \dots, k_2 - 1\}$ .

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

### The Observation Model Using SEER Data

Put  $Y = (Y_j, j = 1, \dots, m)$  and  $\underline{y} = (y_j, j = 1, \dots, m)$ . By the probability distribution given by equation (18), the conditional probability density of  $Y$  given  $\{X, \mathbf{U}, \Theta\}$  is approximately given by:

$$P\{Y | X, \mathbf{U}, \Theta\} = \prod_{j=1}^m h(Y_j; \tau_j), \quad (19)$$

where  $h(Y_j; \tau_j)$  is the density at  $Y_j$  of the Poisson distribution with mean  $\tau_j$ .

Then the likelihood function of  $\Theta$  given  $(X, \mathbf{U})$  is  $L(\Theta | \underline{y}, X, \mathbf{U}) = \prod_{j=1}^m h(y_j; \tau_j)$ . It follows that the deviance from this density is:

$$\begin{aligned} Dev &= -2\{\log L(\Theta | \underline{y}, X, \mathbf{U}) \\ &- \log L(\hat{\Theta} | \underline{y}, X, \mathbf{U})\} \\ &= \sum_{j=1}^m \left\{ \tau_j - y_j - y_j \log \frac{\tau_j}{y_j} \right\}, \end{aligned} \quad (20)$$

where  $\hat{\Theta} = (\hat{\tau}_j, j = 1, \dots, m)$  and  $\hat{\tau}_j = y_j$  is the maximum likelihood estimate of  $\tau_j$ .

From equations (19)-(20), we have for the joint density of  $(X, \mathbf{U}, Y)$  given  $\Theta$ :

$$P\{X, \mathbf{U}, Y | \Theta\} = P\{X, \mathbf{U} | \Theta\} P\{Y | X, \mathbf{U}, \Theta\}. \quad (21)$$

To apply the above distribution to estimate unknown parameters and to fit real data, we also make the following assumptions: (a) From biological observations ([1-9]), one may practically assume that  $\{\alpha_l(t) = \alpha_l, l = 0, 1, 2, 3; \beta_r(t) = \beta_r, r = 0, 1, 2, 3, 4, b_3^{(I)}(t) = b_3^{(I)}, d_3^{(I)}(t) = d_3^{(I)}, b_4^{(I)}(t) = b_4^{(I)}, d_4^{(I)}(t) = d_4^{(I)}\}$ . (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  $10 \text{ mm}^3$ , 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  $\{b_3^{(I)}(t) = b_3^{(I)} e^{-\delta_2 t}, d_3^{(I)}(t) = d_3^{(I)} e^{-\delta_2 t}\}$  for some small  $(\delta_i > 0, i = 1, 2)$ . (c) Because colon cell divisions are mainly due to action of the  $\beta$ -catenin gene, one may also assume  $\{g_1^{(I)}(t) = g_j^{(I)}(t) = 0, j = 1, 2\}$ . In this case, one has approximately  $I_1(t + \Delta t) = I_1(t) + M_0^{(I)}(t)$  and  $J_r(t + \Delta t) = J_r(t) + M_{r-1}^{(I)}(t), r = 1, 2$ . Under these assumptions, the unknown parameters of interest are  $\Theta = \{\Theta_1, \Theta_2\}$ , where  $\Theta_1 = \{\lambda_1, \lambda_j, \beta_3, \alpha_i, \beta_j, b_{i+1}^{(I)}, d_{i+1}^{(I)}, b_{j+2}^{(I)}, d_{j+2}^{(I)}, i = 1, 2, j = 1, 2, \delta_l, l = 1, 2\}$  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  $P\{\Theta | X, \mathbf{U}, \underline{y}\}$  of  $\Theta$  given  $\{X, \mathbf{U}, Y = \underline{y}\}$ . This posterior distribution is derived by combining the prior distribution  $P\{\Theta\}$  of  $\Theta$  with the probability distribution  $P\{X, \mathbf{U}, Y | \Theta\}$  given by equation (20) with  $Y$  being replaced by  $\underline{y}$ . 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 ( $P\{X, \mathbf{U} | \Theta\}$ ) and (3) Information from observed data, represented by the statistical model through the conditional likelihood  $L(\Theta | \underline{y}, X, \mathbf{U})$ .

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  $Max\{o(\alpha_3(t)), o(\beta_4(t))\}$  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 ([25,26]).

### The Prior Distribution of the Parameters

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

$$P(\Theta) \propto c(c > 0) \quad (22)$$

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,  $\gamma_1(t) = 0, 0 < b_i^{(I)} < 0.5, i = 2, 3, \gamma_2(t) = \gamma_2 e^{-\delta_1 t}, 10^{-4} < \gamma_2 < 2 * 10^{-2}, 10^{-5} < \delta_1 < 5 * 10^{-3}, 10^{-2} < \gamma_3 < 0.5$ .

(ii) For the mutation rates in the MSI pathway,  $1 < \lambda_j < 1000$  ( $N \rightarrow I_1$ ),  $10^{-8} < \beta_1 < 10^{-5}$ ,  $10^{-6} < \beta_j < 10^{-2}$ ,  $j = 2, 3, 4$ . For the proliferation rates in the MSI pathway,  $\gamma_i^{(l)}(t) = 0, i = 1, 2, \gamma_i^{(l)}(t) = \gamma_i^{(l)} e^{-\delta_i t}, 10^{-3} < \gamma_i^{(l)} < 0.5, j = 3, 4, 10^{-6} < \delta_2 < 10^{-4}, 0 < b_j^{(l)} < 0.5, j = 3, 4$ .

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 [27].

### The Posterior Distribution of the Parameters Given $\{Y = \underline{y}, X, U\}$

Combining the prior distribution given in (6.1) with the density of  $P\{X, U, Y | \Theta\}$  given in equation (20), one can readily derive the conditional posterior distribution of  $\Theta$  given  $\{X, U, Y = \underline{y}\}$ . For  $(l = 2, 3)$ , denote by:

$N_{ll} = \sum_{t=1}^{t_M} I_l(t), B_{ll} = \sum_{t=1}^{t_M} B_l^{(l)}(t)$  and  $D_{ll} = \sum_{t=1}^{t_M} D_l^{(l)}(t)$ ; similarly, for  $r = 3, 4$ , we define  $\{B_{rr}, D_{rr}, N_{rr}\}$  by replacing  $\{I_l(t), B_l^{(l)}(t), D_l^{(l)}(t)\}$  by  $\{J_r(t), B_r^{(r)}(t), D_r^{(r)}(t)\}$  respectively. Then, we have the following results for the conditional posterior distributions:

(i) The conditional posterior distributions of  $\Theta_1(1) = \{\lambda_b, \lambda_j, \alpha_b, l = 1, 2, \beta_r, r = 1, 2, 3\}$  given  $\{X, U, Y = \underline{y}\}$  is:

$$\begin{aligned} & P\{\Theta_1(1) | X, U, \underline{y}\} \\ & \propto P\{\Theta_1(1)\} e^{-(L_1+L_2)(t_M-t_0)} [L_1] \sum_{t=t_0}^{t_M} M_0^{(1)}(t) \\ & \times [L_1] \sum_{t=t_0}^{t_M} M_0^{(1)}(t) \prod_{l=1}^2 e^{-N_{ll} a_l} [a_l] \sum_{r=1}^m \gamma_r \\ & \times \prod_{r=1}^3 e^{-N_{rr} b_r} [b_r] \sum_{r=1}^m \gamma_r. \end{aligned}$$

(ii) The conditional posterior distributions of  $\Theta_1(2) = \{b_3^{(l)}, d_3^{(l)}, b_4^{(l)}, d_4^{(l)}\}$  given  $\{X, U, Y = \underline{y}\}$  is:

$$\begin{aligned} & P\{\Theta_1(2) | X, U, \underline{y}\} \\ & \propto P\{\Theta_1(2)\} f(B_{3l}, D_{3l}; N_{3l}, b_3^{(l)}, d_3^{(l)}) \\ & \times f(B_{4j}, D_{4j}; N_{4j}, b_4^{(j)}, d_4^{(j)}). \end{aligned}$$

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

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

(vi) The conditional posterior distribution of  $\{b_2^{(l)}, d_2^{(l)}, \delta_1\}$  given  $\{X, U, Y = \underline{y}\}$  and the conditional

posterior distribution of  $\{b_3^{(l)}, d_3^{(l)}, \delta_2\}$  given  $\{X, U, Y = \underline{y}\}$  are represented respectively by:

$$\begin{aligned} & P\{b_2^{(l)}, d_2^{(l)}, \delta_1 | X, U, \underline{y}\} \\ & \propto P(b_2^{(l)}, d_2^{(l)}, \delta_1) f(B_{2l}, D_{2l}; N_{2l}, b_2^{(l)}, d_2^{(l)}) \\ & \times \prod_{t=1}^{t_m} \left( 1 - \frac{(b_2^{(l)} + d_2^{(l)})(1 - e^{-\delta_1 t})}{1 - b_2^{(l)} - d_2^{(l)}} \right) \\ & P\{b_3^{(l)}, d_3^{(l)}, \delta_2 | X, U, \underline{y}\} \\ & \propto P(b_3^{(l)}, d_3^{(l)}, \delta_2) f(B_{3j}, D_{3j}; N_{3j}, b_3^{(l)}, d_3^{(l)}) \\ & \times \prod_{t=1}^{t_M} \left( 1 - \frac{(b_3^{(j)} + d_3^{(j)})(1 - e^{-\delta_2 t})}{1 - b_3^{(j)} - d_3^{(j)}} \right) \end{aligned}$$

### 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  $(Y = \underline{y}, \Theta)$  (The Data-Augmentation Step):

Given  $Y = \underline{y}$  and given  $\Theta$ , use the stochastic equations (3)-(4) and the probability distributions given by equations (5)-(10) in Section 3 to generate a large sample of  $(X, U)$ . Then, by combining this sample with  $P\{Y = \underline{y} | X, U, \Theta\}$  to select  $(X, U)$  through the weighted bootstrap method due to Smith and Gelfand [28]. This selected  $(X, U)$  is then a sample from  $P\{X, U | Y = \underline{y}, \Theta\}$  even though the latter is unknown. (For proof, see Tan [25], Chapter 3.) Call the generated sample  $(\hat{X}, \hat{U})$ .

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

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

(c) Step 3: Recycling Step.

With  $\{(X, U) = (\hat{X}, \hat{U}), \Theta = \hat{\Theta}\}$  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 ([25], Chapter 3). At convergence, the  $\hat{\Theta}$  are the generated values from the posterior distribution of  $\Theta$  given  $Y = \underline{y}$  independently of  $(X, U)$  (for proof, see Tan [25], Chapter 3). Repeat the above procedures one then generates a random sample of  $\Theta$  from the posterior distribution of  $\Theta$  given  $Y = \underline{y}$ ; then one uses the sample mean as the estimates of  $(\Theta)$  and use the sample

**Table 3 Estimates of Parameters for Each Pathway**

LOH Pathway					
	$I_0$	$I_1$	$I_2$	$I_3$	
Mutation Rate	1.4E-06 $\pm 1.69E-08$	2.2E-04 $\pm 1.32E-05$	3.2E-03 $\pm 3.33E-04$	1.2E-06 $\pm 2.06E-07$	
Proliferation Rate	0 N/A	0 N/A	3.6E-03 $\pm 1.12E-03$	1.6E-02 $\pm 4.78E-04$	
Birth Rate Para.	0 N/A	0 N/A	7.4E-03 $\pm 1.03E-03$	1.9E-02 $\pm 4.08E-04$	
Growth Limiting Para.	N/A N/A	N/A N/A	8.3E-05 $\pm 1.4E-05$	N/A N/A	
MSI Pathway					
	$J_0$	$J_1$	$J_2$	$J_3$	$J_4$
Mutation Rate	8.3E-07 $\pm 1.38E-08$	3.5E-04 $\pm 1.89E-05$	1.4E-03 $\pm 8.57E-05$	9.3E-03 $\pm 1.22E-03$	7.7E-06 $\pm 1.79E-06$
Proliferation Rate	0 N/A	0 N/A	0 N/A	2.8E-03 $\pm 7.01E-04$	2.0E-02 $\pm 3.31E-04$
Birth Rate	0 N/A	0 N/A	0 N/A	9.6E-03 $\pm 6.08E-04$	2.6E-02 $\pm 2.88E-04$
Growth Limiting Para.	N/A N/A	N/A N/A	N/A N/A	1.6E-03 $\pm 3.7E-04$	N/A N/A

variances and covarainces as estimates of the variances and covariances of these estimates.

## Results

### Application to Fit the SEER Data

In this section, we will apply the above model to the NCI/NIH colon cancer data from the SEER project. Given in Table 2 are the numbers of people at risk and colon cancer cases in the age groups together with the predicted cases from the model. There are 18 age groups with each group spanning over 5 years. To fit the data, we have assumed that  $\gamma_1^{(j)} = \gamma_j^{(j)} = 0$  for  $j = 1, 2$  because of the observation that uncontrolled cell division of colon stem cells is mainly initiated by the oncogene  $\beta$ -catenin in 3p22. Given in Table 3 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 3 is the plot of probability density of time to tumors.

From these results, we have made the following observations:

(a) As shown by results in Table 2, the predicted number of cancer cases are very close to the observed cases in all age groups. This indicates that the model fits the data well and that 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 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 and

Moolgavkar [17]). This shows that the multiple pathway model fits better than the single pathway 4-stage model as proposed by Luebeck and Moolgavkar [17].

(b) From Table 2, 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 65 and 70 years old for the CIN pathway and is between 85 and 90 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 3 showed that the mutation rates of the  $I_i$  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 3 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 3 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 and Weinberg [29] and Weinberg [30]) by merging these stages into the last stage. From Weinberg ([30], Chapters 13-14), notice that the angiogenesis and metastasis are also multi-stage processes.

(e) Results in Table 3 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- $\beta$ -RII) and the apoptosis inhibition genes (p53 and Bax). Notice from Table 3 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.

## Conclusions and Discussion

Recent studies of cancer molecular biology have indicated very clearly that human colon cancer is developed through multiple pathways ([1-9]). 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 paper 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  $P\{X, U|\Theta\}$ , the prior information via  $P\{\Theta\}$  and information from data via  $L\{\Theta|\mathcal{Y}, X, U\}$ . 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 the NCI SEER data (upto November, 2007). Our results showed that the proposed multiple pathways model fitted better than the single pathway 4-stage model as proposed by Luebeck and Moolgavkar [17]. (The respective AIC and BIC for the multiple pathways model are 55.96 and 81.30 which are ten 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, we will not go any further here.

## Reviewers' comments

### Reviewer 1 (M.P. Little)

#### General comments

This is a generally well-written paper, describing a model very similar to that recently developed by Little et al. [12], generalising the model of Little and Wright ([11]). Arguably this latest model should be referenced, with discussion of mathematical differences between the model outlined here and that one. There could also be discussion of the somewhat different conclusions reached in fits to more or less the same SEER colon cancer data.

**Response:** The paper by Little et al. has been added to the reference; see Little et al. [12].

#### Specific comments (page/line)

(1) The 6-stage model is suddenly drawn from the hat here, but is then almost immediately reduced to a 4-stage model! Why does the haplo-insufficiency of Smad4 and p53 justify combining the stages in this way? Is there evidence that Smad4 and p53 are without function at half gene dose? Also, what is the evidence for mutations being in the order given in Figure 1?

**Response:** References documenting haplo-insufficiency of p53 and smad4 have been given in the paper. The reason why we can combine the two-stages involving P53 into one-stage is based on these papers. To illustrate, let B denote the P53 gene and b the mutant of p53. Then, under haplo-insufficiency of P53, the level and effects of the P53 protein has been reduced significantly (at least 4-fold or more) so that the phenotype of genotype B/b (or B/-) is closely approximated by that of genotype b/b.

(2) How would a model incorporating epigenetic effects differ from the (DNA mutational) models already outlined? I suspect that mathematically the formalism would be exactly the same, although the implied “mutation rates” would be very much higher for epigenetic events.

**Response:** From modeling viewpoint, it is difficult to tell the difference between epigenetic changes and mutation except that the former is much more frequent and very often reversible. However, epigenetic changes can help the modeler to incorporate biological information into the biological process. Many biological papers (just during 2008, there are hundreds of biological papers published) support the epigenetic changes and the view that epigenetic changes are the driving force for cancer initiation, progression and metastasis, more important than gene mutations in cancer initiation and progression. Epigenetic changes include methylation (hypo- and hyper-methylation), micro-RNA (non-coding RNA), loss of imprinting, Histone acetylation, HDAC, tissue disorganization and gap junction disruption, etc.; epigenetic changes can also lead to gene mutations; for a brief review, see Tan and Hanin [31] (Chapter 3).

(3) This is slightly confusing. I assume that the two pathways referred to are CIN and MSI, but coming after discussion of epigenetic effects, perhaps this was meant as well or instead.

**Response:** The genetic sequence of the CIN and MSI pathways were determined by molecular biology of colon cancer and have been published in cancer journals. (There are a large number of biological papers documenting this. I have just listed a few of them in our paper. I can provide many more published papers from cancer journals if one wishes.) This sequence appears to be logical from biological mechanism. Notice that the APC- $\beta$ -Catenin activates myc and cyclin D to push the cell into cell cycle; for the cell cycle to progress, the inhibition effects of p15, p16, p18, p19, p27 have to be abrogated through the inhibition of the TGF- $\beta$  signaling pathway (epigenetic silencing or inactivation or mutation or deletion of smad2/sm4 (CIN pathway) or TGF- $\beta$  Receptor II (MSI pathway)). When the number of cells has increased to certain level, then the apoptosis p53-Bax pathway is activated. For the progression of carcinogenesis, the p53 (CIN pathway) or the Bax gene (MSI pathway) have to be epigenetically silenced or inactivated or mutated or deleted. This is the reason why only the late stage involves silencing or inactivation of the gene p53. This is illustrated in our paper; see also Tan and Hanin [31] (Chapter 3, Chapter 11 and Chapter 12). There are no biological supports single-stage or two stage models. Hence it is not logical to accept single stage or two stage models for colon cancer for people who are born normal. Also, it is extremely difficult for

me to accept that there are no proliferation for  $I_2$  cells as assumed in some of the cancer model papers simply because of the observation of polyps in colon which are derived from proliferation of second-stage cells and is the basis for colon cancer screening procedure “Colonoscopy” practiced by medical doctors.

(4) I assume that asymmetric mutations are assumed, in which a cell produces one normal daughter cell and one mutant daughter cell, as assumed by Little et al. [12] and Little and Wright [11] and many others, but this could perhaps be clarified. It might be useful to derive this Kolmogorov forward equation (2) in an Appendix. Clearly this forward equation (2) is in general intractable. I suspect that as in the papers of Little et al. [12] and Little and Wright [11] it is much easier to attempt to solve the Kolmogorov backward equations.

**Response:** As in Little, I assume that under genetic changes, a normal (or one  $I_j$  cell) cell will give rise to one normal cell (or one  $I_j$  cell) and one mutant cell (or one  $I_{j+1}$  cell) (asymmetric change or mutation). This is logical because mutation or genetic changes occur during cell division.

(5) These conclusions are somewhat at odds with those of Little et al. [12] and Little and Li [32] who fitted models to very similar SEER colon cancer data and demonstrated that there was little evidence of better fit of models that allowed for genomic instability compared with those that did not. In particular the 4-stage model of Luebeck and Moolgavkar [16] provided as good fit as models that allowed for genomic instability [32], or as here multiple types of instability (CIN and MSI) [12]. The authors may care to discuss this.

**Response:** On November 13, 2009, Dr. Little had sent me his comments about our paper and a copy of his 2008 paper which I was not aware of before I wrote our paper. Dr. Little claimed that their fitting of the data could not differentiate between many different models. When I examined Dr. Little’s paper, I found out that our estimation approach is very different from Dr. Little’s. While Dr. Little’s approach used the classical sampling theory through maximum likelihood estimates; our approach is the state space modeling and the generalized Bayesian inference incorporating information from three different sources: (1) The biological information from cancer molecular biology, (2) prior information from epigenetic and genetic mechanism, and (3) information from the likelihood function of observed data involving cancer incidence. Furthermore, because cancer incidence are derived only from people who are less than 100 years old (his/her life time), we first change the SEER data from  $(n_j, y_j)$  to  $(m_j, y_j)$ , where  $m_j$  is the number of people who can develop colon cancer during his life time (i.e., less than 100 years old). Notice that  $n_j$  is of the order  $10^7$  while  $m_j$  is of the order  $10^5$ . This is described in detail in our paper in Tan

and Hanin [31] (Chapter 11). Because of this and because of additional information from biological mechanism and prior distribution, we do not have the problem of identification of parameters which statisticians usually encounter in analyzing data without using information from biological mechanism. Thus, using classical approach, the parameters are not identifiable so that one has to make some assumptions such as that growth rates (birth rates) of different stage cancer initiated cells are equal which can hardly be realistic in carcinogenesis. (Biological studies by cancer biologists clearly demonstrated that the proliferation rates of cancer initiated cells with different genetic changes are very different in most of the cases.)

Finally, I want to emphasize that many models can fit the data but some fit better than others. Finally I like to emphasize what Dr. Van Ryzin had concluded 20 years ago that for cancer risk assessment, many models can fit the data but only models which are biologically supported can give correct results for cancer prediction and cancer risk assessment. It is important to list the predicted numbers along with the observed numbers in cancer modeling research. EPA has revised the guidelines to require that cancer risk assessment models should be biologically supported.

#### Reviewer 2 (M. Kimmel)

The paper by Tan and Yan, proposes a new stochastic model of colon cancer progression, involving the chromosomal instability pathway and the micro-satellite instability pathway. This model not only might provide more insights into human colon cancer but also might provide useful guidance for its prevention and control and for prediction of future cancer cases. One interesting question is as to whether Tan and Yang model might help in deciding about the value of early detection of colon cancer by screening examination. Although colonoscopy is an accepted procedure, there are questions that linger concerning the impact of early detection on mortality reduction.

#### Response to Reviewer 2

Thank you very much for your comments and suggestion. We are currently collecting data on screening by colonoscopy. We will apply the model to estimate cancer incidence under screening and will examine if colonoscopy will help reduce cancer incidence. We will do computer simulation to find this out. This is our next research on colon cancer.

#### Abbreviations

CIN: Chromosomal Instability; LOH: Loss of Heterozygosity; MSI: Micro-Satellite Instability; APC: Adenomatous Polyposis Coli; MMR: Mis-Match Repair; ML:

Multinomial; Dev: Deviance; AIC: Akaike Information Criterion; BIC: Bayesian Information Criterion; SEER: Surveillance Epidemiology and End Results.

#### Acknowledgements

The research of this paper is supported by an NCI/NIH grant, R15 CA113347-01 and an NSF grant, DMS-0805829. As a one-time exception to the publishing policy of Biology Direct the articles in this series are being published with two reviewers.

#### Authors' contributions

Both authors contributed equally.

#### Competing interests

The results of this paper have not been published in other places.

Received: 19 January 2010 Accepted: 20 April 2010

Published: 20 April 2010

#### References

1. de la Chapelle A: Genetic predisposition to colorectal cancer. *Nature Review (Cancer)* 2004, **4**:769-780.
2. Fodde R, Smit R, Clevers H: APC, signal transduction and genetic instability in colorectal cancer. *Nature Review (Cancer)* 2001, **1**:55-67.
3. Fodde R, Kuipers J, Rosenberg C, et al: Mutations in the APC tumor suppressor gene cause chromosomal instability. *Nature Cell Biology* 2001, **3**:433-438.
4. Green RA, Kaplan KB: 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* 2003, **163**:949-961.
5. Hawkins NJ, Ward RL: Sporadic colorectal cancers with micro-satellite instability and their possible origin in hyperplastic polyps and serrated adenomas. *J Natl Cancer Institute* 2001, **93**:1307-1313.
6. Hisamuddin IM, Yang VW: Genetics of colorectal cancer. *Medscape General Medicine* 2004, **6**(3):2004.
7. Peltomaki P: Deficient DNA mismatch repair: A common etiologic factor for colon cancer. *Human Molecular Genetics* 2001, **10**:735-740.
8. Sparks AB, Morin PJ, Vogelstein B, Kinzler KW: Mutational analysis of the APC/beta-catenin/Tcf pathway in colorectal cancer. *Cancer Res* 1998, **58**:1130-1134.
9. Ward R, Meagher A, Tomlinson I, O'Connor T, et al: Microsatellite instability and the clinicopathological features of sporadic colorectal cancer. *Gut* 2001, **48**:821-829.
10. Tan WY, Zhang LJ, Chen W, Zhu JM: A stochastic model of human colon cancer involving multiple pathways. "Handbook of Cancer Models with Applications" World Scientific, Singapore and River Edge, NJTan WY, Hanin L 2008, Chapter 3.
11. Little MP, Wright EG: A stochastic carcinogenesis model incorporating genomic instability fitted to colon cancer data. *Math Biosci* 2003, **183**:111-134.
12. Little MP, Vineis P, Li G: A stochastic carcinogenesis model incorporating multiple types of genomic instability fitted to colon cancer data. *J Theor Biol* 2008, **254**:229-238.
13. Little MP: Cancer models, ionization and genomic instability: A review. "Handbook of Cancer Models with Applications" World Scientific, Singapore and River Edge, NJTan WY, Hanin L 2008, Chapter 5.
14. Jessup JM, Gallic GG, Liu B: The molecular biology of colorectal carcinoma: Importance of the Wg/Wnt signal transduction pathway. "The Molecular Basis of Human Cancer" Humana Press, TotowaColeman WB, Tsongalis GJ 2002, Chapter 13:251-268.
15. Alberici P, Jagmohan-Changur S, De Pater E, et al: Smad4 haplo-insufficiency in mouse models for intestinal cancer. *Oncogene* 2006, **25**:1841-1851.
16. Lynch CJ, Milner J: Loss of one p53v allele results in four-fold reduction in p53 mRNA and protein: A basis for p53 haplo-insufficiency. *Oncogene* 2006, **25**:3463-3470.
17. Luebeck EG, Moolgavkar SH: Multistage carcinogenesis and colorectal cancer incidence in SEER. *Proc Natl Acad Sci USA* 2002, **99**:15095-15100.
18. Baylin SB, Ohm JE: Epigenetic silencing in cancer-a mechanism for early oncogenic pathway addition. *Nature Rev Cancer* 2006, **6**:107-116.

19. Koinuma K, Yamashita Y, Liu W, Hatanaka H, Kurashina K, Wada T, *et al*: **Epigenetic silencing of AXIN2 in colorectal carcinoma with microsatellite instability.** *Oncogene* 2006, **25**:139-146.
20. Breivik J, Gaudernack G: **Genomic instability, DNA methylation, and natural selection in colorectal carcinogenesis.** *Semi Cancer Biol* 1999, **9**:245-254.
21. Jones PA, Baylin SB: **The fundamental role of epigenetic events in cancer.** *Nat Rev Genet* 2002, **3**:415-428.
22. Fakir H, Tan WY, *et al*: **Stochastic population dynamic effects for lung cancer progression.** *Radiation Research* 2009, **172**:383-393.
23. Yakovlev AY, Tsodikov AD: **Stochastic Models of Tumor Latency and Their Biostatistical Applications.** World Scientific, Singapore and River Edge, New Jersey 1996.
24. Tan WY: **Stochastic Models of Carcinogenesis.** Marcel Dekker, New York 1991.
25. Tan WY: **Stochastic Models With Applications to Genetics, Cancers, AIDS and Other Biomedical Systems.** World Scientific, Singapore and River Edge, New Jersey 2002.
26. Tan WY: **Stochastic Modeling of AIDS Epidemiology and HIV Pathogenesis.** World Scientific, Singapore and River Edge, New Jersey 2000.
27. Box GEP, Tiao GC: **Bayesian Inference in Statistical Analysis.** Addison-Wesley, Reading, MA 1973.
28. Smith AFM, Gelfand AE: **Bayesian statistics without tears: A sampling-resampling perspective.** *American Statistician* 1992, **46**:84-88.
29. Hanahan D, Weinberg RA: **The hallmarks of cancer.** *Cell* 2000, **100**:57-70.
30. Weinberg RA: **The Biology of Cancer.** Garland Sciences, Taylor and Frances Group, New York 2007.
31. Tan WY, Hanin L: **Handbook of Cancer Models with Applications.** World Scientific Publishing Co. Pte. Ltd 2008, **9**.
32. Little MP, Li G: **Stochastic modelling of colon cancer: is there a role for genomic instability?** *Carcinogenesis* 2007, **28**:479-487.

doi:10.1186/1745-6150-5-26

**Cite this article as:** Tan and Yan: A new stochastic and state space model of human colon cancer incorporating multiple pathways. *Biology Direct* 2010 **5**:26.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

