

Preferred Sequences of Genetic Events in Carcinogenesis: Quantitative Aspects of the Problem

ANIKO SZABO, ANDREI YAKOVLEV

*Huntsman Cancer Institute and Department of Oncological Sciences,
University of Utah, 2000 Circle of Hope,
Salt Lake City, UT 84112-5550*

ABSTRACT

In this paper we discuss some natural limitations in quantitative inference about the frequency, correlation and ordering of genetic events occurring in the course of tumor development. We consider a simple, yet frequently used experimental design, under which independent tumors are examined once for the presence/absence of specific mutations of interest. The most typical factors that affect the inference on the chronological order of genetic events are: a possible dependence of mutation rates, the sampling bias that arises from the observation process and small sample sizes. Our results clearly indicate that just these three factors alone may dramatically distort the outcome of data analysis, thereby leading to estimates of limited utility as an underpinning for mechanistic models of carcinogenesis.

Keywords: cancer, tumorigenesis, mutation, order, sampling bias

1. Introduction

Carcinogenesis is believed to be a multistep process that proceeds through a series of genetic alterations which are, for the most part, represented by somatic mutations. It is also commonly believed that genetic changes observed in tumors produce phenotypic changes associated with tumor progression, i.e. these changes are thought to cause malignant properties of a tumor such as rapid growth, invasiveness and ability to metastasize [17]. Colorectal cancers in humans are considered to be an excellent system in which to study the timing and interrelations of genetic alterations occurring in tumorigenesis [16]. Fearon and

Author for correspondence: Aniko Szabo, Huntsman Cancer Institute, University of Utah, 2000 Circle of Hope, Salt Lake City, UT 84112-5550, aniko.szabo@hci.utah.edu .

Condensed title: Estimating sequences of genetic events.

Vogelstein [6] summarized experimental findings in a genetic model of colorectal carcinogenesis as early as in 1990. The model describes colorectal carcinogenesis as a (linear) sequence of genetic alterations involving oncogenes and tumor suppressor genes. The authors derived their conclusion based on observed frequencies of occurrence of certain genetic alterations in tumors of various histopathological stages (early, intermediate, late adenomas, carcinoma). The seminal paper by Fearon and Vogelstein [6] attracted much attention in the literature and has since been repeatedly discussed by numerous authors. In a recent paper [8], Kinzler and Vogelstein proposed a more elaborate biological interpretation of the observed genetic alterations in terms of putative regulatory "gates" to be passed in the process of tumorigenesis, and "gatekeeper" and "caretaker" genes that control the gateways.

Despite the plethora of genetic abnormalities (activation of oncogenes, inactivation of tumor suppressor genes, chromosome losses, aneuploidy, etc.) discovered in carcinogenesis in general and in the development of colorectal cancer in particular, the biological role of specific sequences of genetic changes is not well understood [14]. The observed diversity of genetic abnormalities in preinvasive disease is thought to manifest widespread genetic instability characterized by an increase in the intrinsic mutation rate following certain genetic events. Numerous studies have suggested that the development of genetic diversity in right-sided (located predominantly in the proximal colon) colon tumors can be attributed to microsatellite instability in diploid or near-diploid tumor cells. The second mechanism of genetic instability is associated with the development of p53 abnormalities and aneuploidy; these changes are most frequently seen in tumors located in the distal colon. The concept of genetic instability has been questioned by Tomlinson et al. [15]. Using computer simulations, the authors demonstrated that raised mutation rates are more likely to occur in late rather than early-onset tumors. Thus, while some cancers may acquire a mutator phenotype, this event is not necessary for carcinogenesis. Duesberg et al. [4] advocate that an abnormal balance of chromosomes, i.e. aneuploidy of cancer cells, is sufficient to explain genetic instability. Their hypothesis introduces aneuploidy as a general, unique mechanism both of genetic instability and of tumor progression, independent of somatic mutations. The controversy extends even further: some authors believe that cancer is largely an epigenetic disease, while mutations are postulated to be secondary genetic changes that accumulate in inactive sites of the genome [12].

Yet another reason for casting doubts upon the proposed schemes of multiple pathways in carcinogenesis is a lack of rigorous quantitative methods for identifying preferred sequences of genetic abnormalities from the data generated by the most typical study designs. While the model by Fearon and Vogelstein [6] and its more recent version by Kinzler and Vogelstein [8] are very appealing from the biological point of view, the more fundamental question still remains: how strong is a foundation for these and other similar qualitative models (see [14] for review) in the context of real data analysis? In this paper we address quantitative aspects of the problem.

Proceeding from simple probabilistic considerations and using computer sim-

ulations we make an attempt to understand what can and what cannot be inferred from the commonly available data on sequences of genetic alterations in carcinogenesis. In doing so, we confine ourselves to a typical study design where tumors are examined in patients of various ages with various degrees of advancement of the disease, and the presence/absence of certain genetic abnormalities (mutations) in each tumor is recorded. This design implies that each patient is examined only once. While some information on the disease advancement, like tumor grade, is available, the age of the tumor is unknown.

The purpose of data analysis is to provide quantitative insight into the chronological order of mutations which presumably drive the processes of carcinogenesis and tumor progression. In Section 2, we consider the situation when no staging information is available. In Section 3, we incorporate information on tumor grade and investigate the impact of some restrictions that may occur when collecting data on genetic events in tumors.

There are experimental studies designed to yield multiple measurements for each tumor [14]. These designs are likely to provide more quantitative information on the processes under study than the traditional one with the proportion of tumors bearing a given mutation as the only endpoint. However, it is not an easy task to utilize this information because of lack of adequate statistical methods for this type of data. Our choice of the traditional experimental design was inspired by clear statistical properties of associated estimation methods. We believe that our main conclusions are all the more relevant to complex study designs that involve additional confounding factors and call for additional parametric assumptions in order for desired statistical inference to be tractable. Every new experimental setting invites a special simulation study, like the one reported here, which might be an interesting avenue for future research.

It should be emphasized that we do not propose a new methodology of data analysis. The aim of this paper is much more modest - to pinpoint the key obstacles that hinder reliable biological inferences from the existing data.

2. Inference on the order of mutations in the absence of stage information

Suppose there are k mutations M_1, M_2, \dots, M_k that may occur in a cell during its neoplastic transformation or in the tumor itself. Consider n independent specimens of the tissue under examination; for simplicity sake we will call them "tumors". Each specimen (tumor) is obtained (biopsied or removed) at an unknown time t_j and the presence/absence of the mutations is recorded as a binary vector $\mathbf{x}_j = (x_{j1}, x_{j2}, \dots, x_{jk})$, where

$$x_{jl} = \begin{cases} 0, & \text{if the } l^{\text{th}} \text{ mutation is absent in the } j^{\text{th}} \text{ tumor} \\ 1, & \text{if the } l^{\text{th}} \text{ mutation is present in the } j^{\text{th}} \text{ tumor} \end{cases},$$

$j = 1, 2, \dots, n; l = 1, 2, \dots, k$.

A question of particular interest is the determination of the preferred pathways (sequences of genetic changes) of tumorigenesis. It is commonly believed

that there may be several alternative pathways that occur with different probabilities. We assume that given enough time each mutation would occur, so it is natural to consider the $k!$ pathways consisting of all the k mutations. The i^{th} pathway $M_{\pi_i(1)} \rightarrow M_{\pi_i(2)} \rightarrow \dots \rightarrow M_{\pi_i(k)}$, where π_i is one of the $k!$ different permutations of $1, 2, \dots, k$, occurring with probability p_i . Our goal is to make inference about the values of p_i based on a sample of n observations: $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$.

First we explore what one can go about estimating the probabilities p_i within the non-parametric approach. We use the method of maximum likelihood for this purpose. Since the time of observation measured from the (random) time of tumor initiation is unavailable, we have to proceed from a marginal likelihood function. A specific observation informs us that there is a pathway in which the mutations present in the tumor under examination come before the mutations that are absent. For example, a vector like $(0, 0, 1, 1, 0, 1)$ comes from a pathway that starts with M_3, M_4, M_6 (in any order) and ends with M_1, M_2, M_5 (in any order). Let D_j denote the set of the pathways "compatible" with \mathbf{x}_j , i.e. the set of pathways that could have generated a given observation. Then the contribution of the observation \mathbf{x}_j to the likelihood is $p(D_j) = \sum_{i \in D_j} p_i$. To obtain the maximum likelihood estimators, we have to maximize

$$L = \prod_{j=1}^n p(D_j),$$

$$\text{subject to } \sum_i p_i = 1$$

In the Appendix we show that the maximum likelihood estimates under this experimental design can be obtained only in the case of two mutations; the model becomes non-identifiable when more mutations are considered. This means that the data under consideration do not provide enough information for reconstructing preferred sequences. Nonetheless, pairwise comparisons between mutations can be made and a preferred relative position of any two mutations can be identified. It should be noted that in this multiple testing situation pertinent adjustment procedures should be used to control the type I error [7]. However, it is important to remember that the pairwise ordering of mutations cannot be extended even to triples: for example, if mutation M_1 precedes M_2 in 70% of the cases and M_2 precedes M_3 in 70% of the cases, it is still possible that M_3 precedes M_1 with probability 60%.

3. Inference using stage information

3.1. General remarks

The observed tumors are often classified into one of several stages based on histopathologic features (e.g a colon tumor might be classified as adenoma, carcinoma, etc.). While in different types of cancer this staging or grading might be based on different principles, from the point of view of inference they all

provide similar information about the timing of the observation in the tumor's life. These stages (grades) usually form an ordered sequence; we assign higher numbers to represent more advanced tumors. In this section, we discuss inference procedures about the distribution of the occurrence of mutations across these morphologically recognizable stages of tumor development. For simplicity, we assume that tumors are homogeneous so that there is no branching point that determines different pathways leading either to adenomas or to the more advanced stages of tumor progression.

The estimation of the distribution of a mutation, say A , by itself is fairly straightforward:

$$P(A \text{ occurs in the } i^{\text{th}} \text{ stage}) = P(A \text{ occurs by the end of the } i^{\text{th}} \text{ stage}) - P(A \text{ occurs by the end of the } (i-1)^{\text{st}} \text{ stage})$$

The probabilities on the right side can be estimated from the data by the corresponding proportions; these estimators are unbiased, of course. Using the above formulas in some cases the difference of the probabilities might happen to be negative (for example, if a certain mutation is observed in 20% of adenomas, but only in 15% of carcinomas). This may be due to the inherent variability of the sampling process and does not necessarily mean that a given mutation can disappear after it has occurred in a tumor. Should this be the case, the maximum likelihood estimates can be obtained using isotonic regression (see, e.g. [13]) and they coincide with the usual proportions when there are no order reversals.

While the estimated marginal distribution of the mutation occurrence can be easily obtained, the joint distribution of two or more generally cannot be estimated. Consider two mutations, say A and B . The probability that both mutations (or none, or only A , or only B) have occurred by the end of the i^{th} stage can be estimated from the data by the corresponding proportion using isotonic regression if necessary to ensure that the estimates do not decrease with advancing stages. However, the more specific terms, like $P(A \text{ occurs in the } i^{\text{th}} \text{ stage, } B \text{ occurs in the } j^{\text{th}} \text{ stage})$ are not identifiable (there are more unknowns than equations).

A special case where the joint distribution can be derived is when the mutations A and B occur independently in the sense that the events " A occurs in the i^{th} stage" and " B occurs in the j^{th} stage" are independent for any pair of stages i and j . In this case the joint distribution is simply the product of the corresponding marginal distributions. Unfortunately, since the joint distribution is generally not identifiable, the hypothesis of independence cannot be tested directly. Some measure of the relationship of the two mutations can be obtained by considering the "correlation" of the events " A has occurred by the end of the i^{th} stage" and " B has occurred by the end of the i^{th} stage", defined as the correlation of the corresponding indicator variables. If the mutations are independent in the sense defined above, these events are also independent (and hence uncorrelated), however the converse is not true, so only dependence can be inferred.

3.2. *Estimating the order of mutations in the presence of stage in-*

formation

In this section, we evaluate the possibility of estimating the order of mutations based on an ideal “complete” data set, in which tumors are observed at the end of a stage, so no new mutation can occur in the same stage after the moment observation. A quantity of interest may be the odds ratio

$$R = \frac{P(A \text{ precedes } B)}{P(B \text{ precedes } A)} = \frac{\sum_{i < j} P(A_i B_j)}{\sum_{i > j} P(A_i B_j)}$$

As discussed in Section 3.1, the joint probabilities in the above expression cannot be estimated unless independence of the corresponding events is assumed. In the following we investigate the effect of such an assumption on the estimate of the odds ratio when the events are, in fact, dependent.

Suppose that the events A_i and B_j are not independent. Then their joint distribution can be written as

$$P(A_i B_i) = P(A_i)P(B_i) + \xi_i$$

$$P(A_i B_j) = P(A_i)P(B_j) - \eta_{ij}, \quad i \neq j$$

so that positive values of ξ_i represent positive correlation, or synergism between the mutation rates and negative values represent antagonism.

Let a and b denote the numerator and denominator of the “naive” estimate that ignores the correlation structure:

$$R_0 = \frac{\sum_{i < j} P(A_i)P(B_j)}{\sum_{i > j} P(A_i)P(B_j)} = \frac{a}{b}.$$

Then the “true” value of R is

$$R = \frac{a - \Delta_1}{b - \Delta_2},$$

where $\Delta_1 = \sum_{i < j} \eta_{ij}$ and $\Delta_2 = \sum_{i > j} \eta_{ij}$. From the above formulas one can see that the two estimates differ from each other unless $R = \Delta_1/\Delta_2$. More specific relations are summarized in Table 1.

To model a variety of situations, we consider mutation data with 4 different patterns of the marginal distribution over four stages. The pattern numbers in Table 2 are used in the rest of this paper in all the tables presenting simulation results. For the purposes of this illustration we assume that $\Delta_1 = \Delta_2 = \Delta$, so that only a real odds ratio of 1 will be always correctly estimated; the results were similar when we chose other values as correctly estimable. Then it can be easily shown that $\Delta = \frac{1}{2} \sum \xi_i$, so it is enough to specify the values for ξ_i . To introduce synergism (or antagonism) between two independent mutations, we increase (or decrease) the joint probabilities $P(A_i)P(B_i)$ by a certain fraction, that is we choose $\xi_i = \gamma P(A_i)P(B_i)/100$, that is $P(A_i B_i) = P(A_i)P(B_i)(1 + \frac{\gamma}{100})$. The true values of the odds ratio R are tabulated against the estimate R_0 for various combinations of the marginal distributions and values of the percentage increase γ in Table 3.

The examination of the table shows that the same estimate R_0 of the odds ratio can be obtained for a wide range of values of the underlying odds. One has to be careful in accepting the estimate, especially whenever one has a reason to believe that the mutations are strongly correlated and there actually is a preferred order of occurrence.

4. Estimating stage specific probabilities and correlations of genetic events: A simulation study

4.1. The simulation model

The estimators described above were derived under the following two implicit assumptions:

1. all observations occur at the end of the corresponding stage, i.e. no new mutations can arise in the same stage after the observation has been carried out (no *sampling bias*);
2. the number of independent observations is large, so the estimates are close to their real values.

The reason for invoking these assumptions was to elucidate possible difficulties that emerge if mutation rates are stochastically dependent. Unfortunately, none of the conditions formulated above is met in practical applications. Furthermore, one cannot even expect the first of these conditions to be satisfied. To assess the behavior of the estimators in a realistic setting, we have conducted a simulation study.

Our model does not attempt to simulate the process of tumor development in depth, but rather to generate data with realistic properties. The effects that we will demonstrate do not depend intrinsically on the specific distributional assumptions of the model; the same qualitative conclusions would result using different distributions to generate the data. The key component of our model is the observation (data gathering) process, which follows a commonly used scheme. We use Monte Carlo techniques described later in detail to obtain empirical estimates of the quantities of interest. These are compared with the exact values they are meant to estimate. To balance complexity and generality, we have chosen to investigate the situation with two mutations (A and B) and three stages (I , II and III). We model separately the distribution of each mutation across the stages, and a possible correlation between the two mutations.

Let L_I, L_{II} and L_{III} denote the lengths of the stages. We assume that the lengths are independent random variables and each follows a gamma distribution with shape parameter α_i and scale parameter β_i . The main reason for our choice is that this distribution has been successfully used in similar context: incorporating the gamma distribution for the time of tumor promotion, i.e. the time it takes for an initiated cell to undergo malignant transformation, into a stochastic model of tumor latency yields an excellent fit to experimental data on radiation and chemical carcinogenesis [1, 2, 18].

Let M_A and M_B denote the time of occurrence of the mutations A and B respectively (time 0 is the start of stage I). The time to the occurrence of an event is modeled by the exponential distribution. This distribution was chosen to reflect a purely random nature of mutagenesis. In particular, it is common practice to proceed from a Poisson process of initiation events when developing stochastic two-stage models of carcinogenesis (see for example [9–11, 19]). To incorporate dependence between the mutations into the model, we assume that the random variables M_A and M_B follow a Farlie-Morgenstern distribution with exponential marginal distributions, i.e. the joint pdf is

$$w(t_1, t_2) = \delta_A e^{-\delta_A t_1} \delta_B e^{-\delta_B t_2} [1 + \xi(2e^{-\delta_A t_1} - 1)(2e^{-\delta_B t_2} - 1)], \quad t_1, t_2 \geq 0.$$

Here the parameter $-1 \leq \xi \leq 1$ determines the correlation of the marginal variables: $\text{Corr}(M_A, M_B) = \xi/4$. This approach is commonly used to investigate the effect of correlation when the marginal distributions are given (e.g. [5]).

We assume that the time of occurrence of the mutations is independent of the length of the stages. There are two reasons for the latter assumption:

1. It is just natural to proceed from the hypothesis of independence in face of the absence of any relevant experimental evidence. Since this hypothesis has to do with unobservable variables it is not testable in a direct experiment. In such an event, the independence hypothesis is the first one to be explored by means of mathematical modeling.
2. While the dependence of mutation rates discussed in Section 3.2 has a direct bearing on the problem of identifiability, the type of dependence under discussion herein is of much lesser importance to the net results of our study. It is very unlikely that our basic conclusions (Sections 4.2, 4.3) regarding the bias and variability of the resultant estimates would have been different if we had proceeded from a more sophisticated model structure.

The variables introduced above generate the underlying process, but their values cannot be observed. The only observable information is the tumor status at examination. We assume that the time, T , of observation is independent of the stage of tumor development or occurrence of mutations. Also, since there is no reason to believe that the observation is more likely to occur during a particular period within a stage, we assume that T is uniformly distributed over an interval that is large enough to contain all three stages.

One simulation run consisted of generating the stage lengths $\ell_I, \ell_{II}, \ell_{III}$, then the correlated times of occurrence of the mutations A and B (m_A, m_B), and finally a random observation time t was chosen. The current stage and the presence/absence of the mutations A and B by this time were recorded as one

observation, that is

$$stage = \begin{cases} I, & \text{if } t \leq \ell_I \\ II, & \text{if } \ell_I < t \leq \ell_I + \ell_{II} \\ III, & \text{if } \ell_I + \ell_{II} < t \leq \ell_I + \ell_{II} + \ell_{III} \\ \text{after } III, & \text{if } t > \ell_I + \ell_{II} + \ell_{III} \end{cases}$$

A is present, if $m_A \leq t$

B is present, if $m_B \leq t$

The specific values of the simulation parameters were chosen to yield the prescribed stage-specific mutation probabilities $P(A_i)$, $i = I, II, III$ for various patterns given in Table 2. Based on the model assumptions the stage-specific probabilities can be expressed in terms of the parameters (similar formula holds for mutation B):

$$P(A_I) = 1 - q_I, \quad P(A_{II}) = q_I(1 - q_{II}), \quad P(A_{III}) = q_I q_{II}(1 - q_{III}),$$

where

$$q_i = \frac{\beta_i^{\alpha_i}}{(\beta_i + \delta_A)^{\alpha_i}}.$$

For the marginal results we specified the stage-specific probabilities according to Table 2, set $\alpha_i = 2$, $i = I, II, III$, $\delta_A = 0.1$ (this only defines the unit of time) and used the above expression to obtain the appropriate values for β_i . For the joint results we have specified everything required for the marginal results for mutation A and additionally $P(B_I)$. As the stage length is shared by all mutations, the value for δ_B and the rest of the stage-specific probabilities of B can be calculated.

4.2. The effect of a biased sampling

We first study the net effect of the sampling bias by using large sample estimates. This bias arises unavoidably from the obvious inconsistency between the quantities of interest and the observation process: we wish to estimate the probabilities of mutation occurrence between the *beginning* and the *end* of each stage, but we observe a tumor only at an unknown *in-between* timepoint. Thus a mutation that occurs later in the stage is less likely to be observed before the tumor has progressed to the next stage than a mutation that occurred earlier. As a result of this a certain portion of stage I occurrences will be attributed to stage II . Similarly a portion of stage II occurrences will be attributed to stage III , etc. Hence the probability of a mutation occurring during stage I is underestimated and the probability of its occurring after the last stage (i.e. not occurring) is overestimated, while for the rest of the stages the over- and underestimation effects might balance each other to give reasonable estimates.

Parameter values were chosen as described in the previous section. The $n = 20,000$ observations thus generated were used for estimating the quantities of interest as described in Section 3.1 to be compared with their prescribed ("true") values. The results of the first simulation study aimed at testing the estimates of P_i are presented in Table 4. As a measure of closeness of the estimated values to the theoretical ones, we use a χ^2 type quantity: $MSE = \sum \frac{(observed - expected)^2}{observed}$, where the summation is over all stages.

From Table 4 we can see that the sampling bias has a great impact on the nonparametric estimates of P_i , $i = I, II, III$. As expected, this effect is especially strong with respect to the probability of the mutation occurrence in stage I – this probability is consistently underestimated by 20 – 25%. Despite this bias, some general qualitative characteristics, like the mode or shape of the distribution across stages are discernible in the simulated data (Table 4).

Similarly to the stage specific probabilities of mutation occurrence, the theoretical values for the correlation coefficients of the indicator variables $\text{Corr}(A, B|S = i)$, $i = I, II, III$ can be obtained (they are straightforward to derive, but too involved to include here; numerical integration with Mathematica was used to evaluate the specific values). Table 5 provides a comparison of the theoretical and estimated values for a large sample size ($n = 20,000$) and various values of the correlation between mutations. The simulation results do not show any systematic bias in the estimation and with a few exceptions, the large sample estimates are reasonably close to their expected values. It is interesting to note that independence of the times to the occurrence of mutations (when their correlation is equal to 0) does not entail independence of the events $A|S = i$ and $B|S = i$. Hence, when referring to independent rates of mutations, one must specify exactly what random events or variables satisfy this property.

4.3. The effect of small sample size

In actual applications only a limited number of independent observations is available. To evaluate the effect of sample size, we have conducted 500 simulations using a small sample size $n = 100$ and for each of these simulations we estimated the stage-specific probabilities and correlations. The results are summarized in Tables 6 and 7; the mean and the coefficient of variation (the ratio of the standard deviation and the mean) or the standard deviation for the estimated parameters of interest are presented. Table 6 also shows the percentage of simulations where an order reversal occurred in estimating the stage probabilities. With large sample sizes, we have seen that, while the probability estimates are biased, some characteristics are fairly robust. For example, the stage with the highest frequency of occurrence (we term this stage the mode of the distribution among stages) could be identified from the estimates. To see whether this observation still holds for smaller sample sizes, we counted the percentage of simulations in which a given stage was the mode of the distribution. The results are shown in Table 6.

As expected, the small sample estimates of the first stage probability also seriously underestimate the true value. More importantly, the coefficient of

variation is generally very high, which means that the individual estimates using $n = 100$ observations vary widely, so any one of them is extremely inaccurate. This wide variation makes it difficult to draw any conclusions even about the general shape of the distribution: for example, the mode is incorrectly estimated in a large proportion of the repeated samples. The percentage of data sets with order reversal is fairly high, especially when there are consecutive stages in which the mutation is not likely to be observed. So with small sample sizes one can expect to encounter many cases when the mutations seem to disappear from one stage to another.

The situation is similar with small sample estimates of the correlations. The standard deviations of the estimates are very high, so that a 95% confidence interval would cover a significant portion of a possible range for the values of the correlations. This means that with sample sizes the hypothesis of independence of mutation rates is hardly practicable for testing.

5. Discussion

We have considered the problem of analyzing mutation occurrence data resulted from a commonly used experimental design: for each patient a tumor is examined once with the information on histological stage/grade and the presence/absence of certain genetic changes being recorded. The goal of our study was to estimate the distribution of the mutations across the putative stages of tumor development and through this to elicit the occurrence order of the mutations. From our results it follows that the most typical study design does not provide enough information to reconstruct complete sequences of genetic events. However, using the likelihood approach it is possible to make pairwise comparisons for different genetic abnormalities.

We have considered three most obvious factors that may affect the inference procedures: a possible dependence of mutation rates, the sampling bias that arises from the observation process, and small sample sizes. All these factors seriously influence quantitative inference by introducing a heavy bias or/and a large variation. There are many more known phenomena that further complicate the problem; some of these are the uncertainty in stage boundaries, the spatial heterogeneity of tumors or large number of the genetic changes under consideration.

The only way to overcome the above mentioned difficulties is through obtaining more data by increasing both the number of independent observations and useful information on each tumor. One fruitful idea is to look at the incidence of mutations in various stages of tumor progression within the same individual tumor [6]. Yet another possibility could be invoking tumor size measurements as a surrogate endpoint to represent the time elapsed from the tumor onset (the event of initiation). The latter possibility is based on the assumption that the tumor size is a monotonically increasing function of time. However, these and other ways of utilizing additional sources of biological information should first be explored in theoretical and simulation studies.

A very interesting paper by Desper et al. [3] suggests modeling sequences of genetic events using a tree structure instead of a linear pathway. The authors

are able to reconstruct the underlying structure from the observed data provided that the observations are generated by one such tree. However, there may be several distinct ways in which malignancies develop, so that each observation follows one of possible developmental models. In this paper we show that the probabilities of these possible models cannot be identified even when only linear pathways are assumed. Allowing for more general tree structures further complicates the issue.

6. Acknowledgements

This work was supported by NCI Cancer Center Support Grant 5P30 CA42014. The authors are very grateful to Drs. R. White, K. Boucher, and A. Tsodikov for fruitful and stimulating discussions. We would like to thank the editor and the referees whose comments helped to greatly improve this paper.

Appendix A. The maximum likelihood inference

It is worth noting that if in a particular tumor one observes all or none of the mutations, i.e. $\mathbf{x}_j = (0, 0, \dots, 0)$ or $(1, 1, \dots, 1)$, then all possible pathways are compatible with this observation. Since $p(D_j) = 1$ and such observations do not contribute to the likelihood, they can be dropped from the original data.

Special case of two mutations ($k = 2$) In the case of two mutations of interest, there are only 2 possible pathways:

$$M_1 \rightarrow M_2, \quad \text{with probability } p_1$$

$$M_2 \rightarrow M_1, \quad \text{with probability } p_2.$$

There are two possible observations that contribute information on the order of mutations; these are

obs.	freq.
$(1, 0),$	a_1
$(0, 1),$	$a_2.$

Since each of these observations uniquely determines the underlying pathway, the likelihood function is of the form

$$L = p_1^{a_1} p_2^{a_2},$$

and is maximized by

$$\hat{p}_1 = \frac{a_1}{a_1 + a_2},$$

$$\hat{p}_2 = \frac{a_2}{a_1 + a_2}.$$

To determine whether there is a preferred order of the mutations, the hypothesis $\hat{p}_1 = 0.5$ should be tested. This can be done either using a normal

approximation or, when the sample size is small, by performing an exact permutation test.

Special case of three mutations ($k = 3$) In this case, all possible pathways can be described as

$$M_1 \rightarrow M_2 \rightarrow M_3, \quad \text{with probability } p_1$$

$$M_1 \rightarrow M_3 \rightarrow M_2, \quad \text{with probability } p_2$$

$$M_2 \rightarrow M_1 \rightarrow M_3, \quad \text{with probability } p_3$$

$$M_2 \rightarrow M_3 \rightarrow M_1, \quad \text{with probability } p_4$$

$$M_3 \rightarrow M_1 \rightarrow M_2, \quad \text{with probability } p_5$$

$$M_3 \rightarrow M_2 \rightarrow M_1, \quad \text{with probability } p_6$$

and the data are of the form

obs.	freq.	prob.
$(1, 0, 0)$,	a_1	$p_1 + p_2$
$(0, 1, 0)$,	a_2	$p_3 + p_4$
$(0, 0, 1)$,	a_3	$p_5 + p_6$
$(1, 1, 0)$,	a_4	$p_1 + p_3$
$(1, 0, 1)$,	a_5	$p_2 + p_5$
$(0, 1, 1)$,	a_6	$p_4 + p_6$

Hence the likelihood function is given by

$$L = (p_1 + p_2)^{a_1} (p_3 + p_4)^{a_2} (p_5 + p_6)^{a_3} (p_1 + p_3)^{a_4} (p_2 + p_5)^{a_5} (p_4 + p_6)^{a_6}.$$

If $a_j \neq 0$ for any $j = 1, 2, \dots, 6$, then the maximum is achieved on a line segment:

$$\begin{aligned} \hat{p}_1 &= \hat{p}_1 \\ \hat{p}_2 &= \frac{a_1}{n_1} - \hat{p}_1 \\ \hat{p}_3 &= \frac{a_4}{n_2} - \hat{p}_1 \\ \hat{p}_4 &= \frac{a_5}{n_2} - \frac{a_1}{n_1} + \hat{p}_1 \\ \hat{p}_5 &= \frac{a_2}{n_1} - \frac{a_4}{n_2} + \hat{p}_1 \\ \hat{p}_6 &= \frac{a_6}{n_2} - \frac{a_2}{n_1} + \frac{a_4}{n_2} - \hat{p}_1. \end{aligned}$$

Hence the model is generally nonidentifiable in the case of $k = 3$. With the larger number of mutations the nonidentifiability aspect of the problem becomes even more evident because a maximum of the likelihood will be described by high-dimensional regions. In order for the above solution to be admissible, all the estimates should fall between 0 and 1. This restriction results in the following bounds for \hat{p}_1 :

$$\max\left(0, \frac{a_1}{n_1} - \frac{a_5}{n_2}, \frac{a_4}{n_2} - \frac{a_2}{n_1}\right) \leq \hat{p}_1 \leq \min\left(\frac{a_1}{n_1}, \frac{a_4}{n_2}, 1 - \frac{a_5}{n_2} - \frac{a_2}{n_1}\right).$$

From this expression it is to derive similar bounds for the remaining stage probabilities. In some cases these boundaries may coincide, resulting in a unique maximum likelihood estimate, or they may overlap, resulting in an empty set. In the latter case a maximum of the likelihood is achieved on the boundary of the parameter space and has to be found by some other method. In the case where some of possible combinations are not observed, i.e. $a_j = 0$ for some a_j , the maximum is always achieved on the boundary; this maximum is either unique or independent of the data.

References

- [1] Boucher K., Pavlova L.V., and Yakovlev A.Yu. A model of multiple tumorigenesis allowing for cell death: Quantitative insight into biological effects of urethane. *Mathematical Biosciences*, **150** (1998) pp. 63–82.
- [2] Boucher K. and Yakovlev A.Y., Estimating the probability of initiated cell death before tumor induction. *Proc. Natl. Acad. Sci. USA* **94** (1997) pp. 12776–12779.
- [3] Desper R., Jiang F., Kallioniemi O.P., Moch H., Papadimitriou C.H., Schäffer A.A. Inferring Tree Models for Oncogenesis from Comparative Genome Hybridization Data. *J. Comp. Biol.* **6**(1) (1999) pp. 37–51.
- [4] Duesberg P., Rausch C., Rasnick D., Hehlmann R., Genetic instability of cancer cells is proportional to their degree of aneuploidy. *Proc. Natl. Acad. Sci.* **95** (1998) pp. 13692–13697.
- [5] Farlie D. J. G. The performance of some correlation coefficients for a general bivariate distribution. *Biometrika* **47** (1960) pp. 307–323.
- [6] Fearon R., Vogelstein B., A genetic model for colorectal tumorigenesis. *Cell* **61** (1990) pp. 759–767.
- [7] Hochberg Y. and Tamhane A.C., *Multiple Comparison Procedures* (John Wiley & Sons, New York) 1987.
- [8] Kinzler K.W., Vogelstein B., Gatekeepers and caretakers. *Nature* **386** (1997) pp. 761–763.
- [9] Moolgavkar S.H. and Knudson A.G. Mutation and cancer: a model for human carcinogenesis. *J. Natl. Cancer Inst.* **66** (1981) pp. 1037–1052.
- [10] Moolgavkar S.H. and Luebeck E.G. Two-event model for carcinogenesis: Biological, mathematical and statistical considerations. *Risk Anal.* **10** (1990) pp. 323–341.
- [11] Moolgavkar S.H. and Venzon D.J. Two event model for carcinogenesis: Incidence curves for childhood and adult tumors. *Math. Biosci.* **47** (1979) pp. 55–77.
- [12] Prehn R.T., Cancers beget mutations *versus* mutations beget cancers. *Cancer Research* **54** (1994) pp. 5296–5300.
- [13] Robertson T., Wright F.T., Dykstra R.L., *Order Restricted Statistical Inference* (Wiley, London) 1988.

- [14] Shackney S.E., Shankey T.V., Common patterns of genetic evolution in human solid tumors. *Cytometry* **29** (1997) pp. 1–27.
- [15] Tomlinson I.P.M., Novelli M.R., Bodmer W.F., The mutation rate and cancer. *Proc. Natl. Acad. Sci. USA* **93** (1996) pp. 14800–14803.
- [16] Vogelstein B., Fearon E.R., Hamilton S.R., Kern S.E., Preisinger A.C. Leppert, M., Nakamura Y., White R., Smits A.M.M., Bos J.L., Genetic alterations during colorectal tumor development. *N. Engl. J. Med.* **319** (1988) pp. 525–532.
- [17] Vogelstein B., Kinzler K.W., The multistep nature of cancer. *Trends in Genetics* **9** (1993) pp. 138–141.
- [18] Yakovlev A.Y., Müller W. A., Pavlova L.V. and Polig E. Do cells repair precancerous lesions induced by radiation? *Mathematical Biosciences*, **142** (1997) pp. 107–117.
- [19] Yakovlev A.Y. and Polig E. A diversity of responses displayed by a stochastic model of radiation carcinogenesis allowing for cell death. *Mathematical Biosciences*, **132** (1996) pp. 1–33.

	synergism	antagonism
$R > \Delta_1/\Delta_2$	underestimates	overestimates
$R = \Delta_1/\Delta_2$	equal	equal
$R < \Delta_1/\Delta_2$	overestimates	underestimates

Table 1. The property of the estimate R_0 under various conditions.

pattern number	probability of the mutation occurring			
	in stage I	in stage II	in stage III	after stage III
<u>E</u> qual	0.300	0.300	0.300	0.100
<u>F</u> irst	0.500	0.200	0.200	0.100
<u>M</u> iddle	0.200	0.500	0.200	0.100
<u>L</u> ast	0.200	0.200	0.500	0.100

Table 2. Notation for the marginal distribution patterns.

marginal pattern	estimate R_0	percent increase γ				
		-50%	-20%	20%	50%	100%
L-F	0.389	0.452	0.416	0.359	0.309	0.204
M-F, L-M	0.563	0.613	0.584	0.538	0.497	0.408
L-E, E-F	0.600	0.654	0.623	0.573	0.526	0.419
M-E, E-M	1.000	1.000	1.000	1.000	1.000	1.000
F-E, E-L	1.667	1.529	1.604	1.744	1.900	2.384
F-M, M-L	1.778	1.632	1.712	1.857	2.012	2.448
F-L	2.571	2.211	2.404	2.784	3.237	4.882

Table 3. The real value of the odds ratio R under various conditions.

pattern		stage		
		I	II	III
E	theoretical	0.300	0.300	0.300
	simulated	0.246	0.292	0.308
	MSE	0.0123		
F	theoretical	0.500	0.200	0.200
	simulated	0.384	0.274	0.199
	MSE	0.0549		
M	theoretical	0.200	0.500	0.200
	simulated	0.164	0.437	0.253
	MSE	0.0283		
L	theoretical	0.200	0.200	0.500
	simulated	0.174	0.156	0.494
	MSE	0.0166		

Table 4. Large sample estimates of the stage probabilities.

pattern		correlation								
		-0.25			0.00			0.25		
E-E	theoretical	-0.047	-0.061	0.005	0.129	0.129	0.125	0.306	0.319	0.244
	simulated	0.012	-0.022	0.027	0.156	0.169	0.096	0.297	0.339	0.312
	MSE	0.3776			0.0229			0.0163		
F-F	theoretical	0.024	-0.033	-0.010	0.196	0.141	0.108	0.369	0.314	0.225
	simulated	0.107	0.015	-0.061	0.206	0.091	0.096	0.401	0.302	0.298
	MSE	0.2606			0.0295			0.0209		
M-M	theoretical	-0.067	0.013	0.003	0.091	0.178	0.122	0.248	0.343	0.241
	simulated	0.036	0.070	0.004	0.095	0.408	0.255	0.095	0.408	0.255
	MSE	0.3414			0.1992			0.2575		
L-L	theoretical	-0.066	-0.113	0.044	0.091	0.091	0.168	0.148	0.195	0.292
	simulated	-0.080	-0.161	0.060	0.261	0.127	0.236	0.246	0.286	0.361
	MSE	0.0210			0.1405			0.0811		

Table 5. Large sample estimates of correlations.

pattern		stage			order
		I	II	III	reversal
E	theoretical	0.300	0.300	0.300	22.8%
	mean	0.214	0.321	0.314	
	coeff. of var.	0.752	0.629	0.527	
	mode	28%	36%	36%	
F	theoretical	0.500	0.200	0.200	33.4%
	mean	0.384	0.264	0.207	
	coeff. of var.	0.292	0.601	0.670	
	mode	63%	24%	13%	
M	theoretical	0.200	0.500	0.200	7.4%
	mean	0.147	0.459	0.247	
	coeff. of var.	1.155	0.407	0.432	
	mode	13%	74%	13%	
L	theoretical	0.200	0.200	0.500	53.0%
	mean	0.119	0.269	0.446	
	coeff. of var.	1.526	0.992	0.587	
	mode	11%	27%	62%	

Table 6. Small sample estimates of the stage probabilities.

pattern		correlation								
		-0.25			0.00			0.25		
E-E	theoretical	-0.047	-0.061	0.005	0.129	0.129	0.125	0.306	0.319	0.244
	mean	0.010	-0.029	0.027	0.147	0.115	0.154	0.378	0.363	0.285
	standard dev.	0.442	0.343	0.190	0.485	0.343	0.235	0.467	0.318	0.235
F-F	theoretical	0.024	-0.033	-0.010	0.196	0.141	0.108	0.369	0.314	0.225
	mean	0.096	-0.012	-0.005	0.257	0.182	0.106	0.393	0.336	0.256
	standard dev.	0.243	0.312	0.190	0.240	0.310	0.224	0.234	0.290	0.232
M-M	theoretical	-0.067	0.013	0.003	0.091	0.178	0.122	0.248	0.343	0.241
	mean	0.002	0.056	0.007	0.164	0.217	0.142	0.323	0.379	0.272
	standard dev.	0.581	0.200	0.187	0.523	0.190	0.234	0.515	0.184	0.249
L-L	theoretical	-0.066	-0.113	0.044	0.091	0.091	0.168	0.148	0.195	0.292
	mean	0.144	-0.144	0.074	0.344	0.119	0.216	0.364	0.380	0.362
	standard dev.	0.632	0.582	0.193	0.674	0.595	0.230	0.658	0.551	0.225

Table 7. Small sample estimates of the correlations.