Number of segregating sites in a sample of genes under the genetic instability hypothesis

Mathieu Emily and Olivier François

Laboratoire TIMC-IMAG
Institut d'Ingénierie de l'Information de Santé
Faculté de Médecine
38706 La Tronche cedex, France
(e-mail: mathieu.emily@imag.fr, olivier.francois@imag.fr)

Abstract. Early detection of (pre)tumor is a priority in the understanding of cancer development in tissues. Several hypotheses have been proposed to explain tumorigenesis. One of them, the *mutator phenotype*, postulates that the loss of mismatch repair (MMR) generates a raise in the mutation rate. Under this assumption estimating the increase in the mutation rate is a key step for detecting a tumor. In this paper an estimator of the raised mutation rate based on the number of segregating sites in a sample of cells is proposed. The bias and the mean squared error of this estimator have been assessed through a simulation study.
Keywords: Tumorigenesis, Genetic instability, Mutator phenotype, Coalescent theory, Number of segregating sites.

1 Introduction

Cancer is known to be a very complex phenomenon. Since early in the 20th century [Boveri, 1929], it is widely assumed that a normal cell is converted to a tumoral cell by a succession of genetic events. More precisely the genetic equilibrium of a cell is disrupted by an initiating event and then because of a cascade process the cell becomes tumoral. This is the so-called *genetic instability* hypothesis for tumorigenesis.

Three major competing hypotheses have been formulated concerning the initial event of tumorigenesis. The first one [Tomlinson and Bodmer, 1999] [Cairns, 1975]explains that a cell must exhibit a selective advantage to be converted into a pretumoral cell. Then by a selective clonal expansion the cell becomes malignant. The second hypothesis is based on the experimental results that most of tumoral cells are victims of aneuploidy [Duesberg *et al.*, 1998]. This chromosomal instability may be responsible for the multistep process that leads to cancer [Duesberg and Rasnick, 2000]. The third hypothesis is called the *mutator phenotype* [Loeb and Springgate, 1974]. Considering the high fidelity of DNA replication in normal cells and the large number of genetic alterations that are observable in cancer cells, it postulates that the initial event in tumorigenesis is a particular mutation. This mutation should

take place in genes that control the fidelity of DNA replication and the efficacity of DNA repair. These genes are directly responsible for the genetic stability of a cell. An alteration of their functions called loss of Mismatch Repair (*loss of MMR*) may generate a deregulation of the apoptosis or a reduction of the cell cycle duration. As a result of loss of MMR, the mutation rate will be raised in all cells that are descendants of the cell affected by loss of MMR. It has been suggested that the loss of MMR is required to initiate tumorigenesis [Loeb, 1991].

It is still a matter of debate to know exactly which event is the initiating event of tumorigenesis. Several mathematical models have been studied to understand the *mutator phenotype* hypothesis. Some of them argued that selection prevails on the raise of the mutation rate [Tomlinson and Bodmer, 1995]. Other models study the effect of the loss of MMR and how it hastens tumorigenesis [Plotkin and Nowak, 2002, Michor *et al.*, 2003]. Evolutionary models had been developed to infer the age of the loss of MMR [Tsao *et al.*, 2000] [Calabrese *et al.*, 2004]. It is widely assumed ([Shibata *et al.*, 1994] and [Bhattacharyya *et al.*, 1994]) that after the loss of MMR the mutation rate increases 10^2- to 10^3-fold. The need for deeper mathematical studies has been formulated in a recent review [Michor *et al.*, 2004] to better understand the influence of the three hypothesis (selection, aneuploidy and mutator phenotype) in the evolution of a cell. So far no mathematical model has been developed to estimate the raised mutation rate, and this is the focus of this article. A classical method in population genetics for estimating a mutation rate consists in counting the number of segregration sites in a sample of genes. In this article we propose a correction of this estimator in the context of genetic instability based on the coalescent theory.

2 Model Description

We consider a sample of n copies of a particular gene taken from a (pre)tumoral tissue, and assume that the loss of MMR occurred once in the sample history. However, the date and place at which this event occurred are unknown. Loss of MMR can be considered as a particular deleterious mutation of a mismatch repair gene. We denote by μ_{LMMR} the rate of this particular mutation, and we assume that the rate of this event is very small (μ_{LMMR} goes to 0).

The sample is divided in two random subsamples $\mathcal{B}$ and $\mathcal{C}$ where $\mathcal{B}$ denotes the subset of descendants of the mutation and $\mathcal{C}$ its complement. Given the number $B = b$ of genes in $\mathcal{B}$, the number of genes in $\mathcal{C}$ is then equal to $n - b$ (see Figure 1). Genes are characterized by their DNA sequences. For instance, such data may arise from the FISH (Fluorescence In Situ Hybridation) technology [Pinkel *et al.*, 1986]. In our model, the evolution of genes is described by a two-rates model. We denote by $\mu_{\mathcal{C}}$ the *normal rate*, ie the

mutation rate per base per generation in $\mathcal{C}$. On the other hand, we denote by $\mu_{\mathcal{B}}$ the *raised mutation rate* in $\mathcal{B}$.

Conditional on $B = b$, the genealogy of the n genes can be described by the so-called *conditional coalescent* [Wiuf and Donnelly, 1999] for which the genes in $\mathcal{B}$ share a common ancestor before any of them shares an ancestor with $\mathcal{C}$. In addition, loss of MMR occurred between the time of the most recent common ancestor (MRCA) of the subsample $\mathcal{B}$ and the time at which $\mathcal{B}$ coalesces with $\mathcal{C}$ (see Figure 1). The coalescent approximation was introduced by Kingman in the 80's [Kingman, 1982]. It is similar to the diffusion approximation in population genetics. Time is measured in units of N generations where N is the total population size. In this setting, mutation rates are rescaled as $\theta_{\mathcal{B}}/2 = 2N\mu_{\mathcal{B}}$ and $\theta_{\mathcal{C}}/2 = 2N\mu_{\mathcal{C}}$.

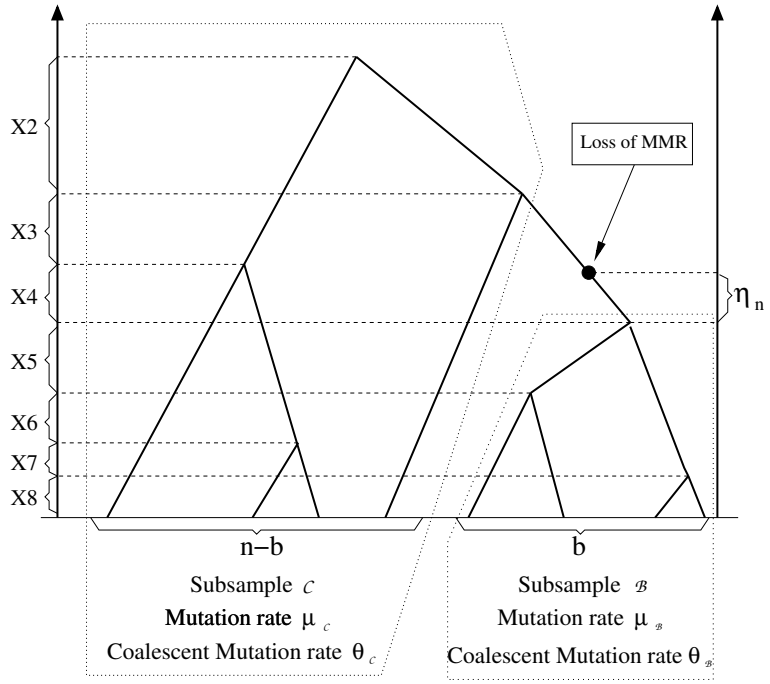

Fig. 1. A coalescent tree of size $n = 8$ conditional on $B = 4$. Time is represented backward and the loss of MMR event is indicated.

In the coalescent, mutations occur according to independent Poisson processes of rate $\theta/2$ along the branches of the tree. Among the various models that describe the mutation types, the *infinitely-many sites* model may be one of the most appropriate [Watterson, 1975]. In this model, each DNA sequence consists of completely linked sites (ie, no recombination occurs).

Each mutation occurs at a site of the DNA sequence that had not been mutated before, so that a new segregating site arises. The number of segregating sites corresponds to the number of substitutions of ancestral bases since the MRCA.

3 Theoretical analysis

3.1 Background

In this section, we recall well-known results about the number of segregating sites under the infinitely-many sites model of mutation. These results are valid when loss of MMR do not occur which means that there is only one mutation rate , written θ [Watterson, 1975]. In the neutral coalescent, the gene lineages coalesce at random, and the times separating the coalescence events X_i, $i = 2 \cdots n$ are independent exponential random variables of parameter $i(i-1)/2$. The tree has total length $L_n = \sum_{i=2}^{n} iX_i$ of expectation

$$I\!E[L_n] = 2H_{n-1} \approx 2 \log n$$

and variance

$$Var[L_n] = 4 \sum_{i=1}^{n-1} \frac{1}{i^2} \approx \frac{2\pi^2}{3}$$

where H_n is the n^{th} harmonic number $H_n = \sum_{i=1}^{n} 1/i$.

The number of segregating sites $\hat{\theta} = S_n/H_{n-1}$ is frequently used as an unbiased estimator of the mutation rate θ. Using that

$$Var[S_n] = \sum_{i=1}^{n-1} \left(\frac{\theta^2}{i^2} + \frac{\theta}{i} \right)$$

we see that $\hat{\theta}$ converges to θ at a logarithmic rate.

3.2 Number of sites of segregation in the two-rates model

In the mutator phenotype hypothesis, a rare mutation is responsible for an increase in the DNA mutation rate from $\theta_{\mathcal{C}}$ to $\theta_{\mathcal{B}}$. In this section, we build an approximately unbiased estimator of $\theta_{\mathcal{B}}$.

First of all, the number B of genes that carry the mutator phenotype (the *frequency spectrum*) has a Yule distribution [Stephens, 2000]

$$P(B = b) = \frac{1}{bH_{n-1}}, \quad b = 1, \ldots, n-1 \tag{1}$$

Given $B = b$, the total length $\widetilde{L_n}$ of the genealogy of the subsample $\mathcal{B}$ has an expected value equal to [Griffiths and Tavaré, 2003]

$$I\!E[\widetilde{L_n}|B = b] = L_{n,b} = \binom{n-1}{b}^{-1} \sum_{j=2}^{n-b+1} \binom{n-j}{b-1} \sum_{k=j+1}^{n} \frac{2}{k(k-1)} c_{jk} \tag{2}$$

where $b = 2, \cdots, n-1$, and

$$c_{jk} = b - (b-1)\frac{n-k}{n-j} - \frac{(n-k)!(n-j-b+1)!}{(n-j)!(n-k-b+1)!} \tag{3}$$

Consider the time η_n that separates the MRCA of the $\mathcal{B}$ sample from the loss of MMR event. Wiuf and Donnelly [Wiuf and Donnelly, 1999] showed that

$$I\!E[\eta_n | B = b] = 2\binom{n-1}{b}^{-1} \sum_{j=2}^{n-b+1} \frac{1}{j}\binom{n-j}{b-1} \quad b = 1, \cdots, n-1 \tag{4}$$

Now, consider the total length $\tilde{L}_n + \eta_n$ (see Figure 1), and take the expectation. We set

$$\beta_n = I\!E[\tilde{L}_n + \eta_n]/2$$

The average number of mutations in descendants of the loss of MMR event is given by

$$I\!E[S_n^{\mathcal{B}}] = \beta_n \theta_{\mathcal{B}}$$

In addition, the average number of mutations in the subsample $\mathcal{B}$ is

$$I\!E[S_n^{\mathcal{C}}] = \gamma_n \theta_{\mathcal{C}}$$

where

$$\gamma_n \approx H_{n-1} - \beta_n \tag{5}$$

Finally, consider the total number S_n of segregating sites. We obtain that

$$I\!E[S_n] = \beta_n \theta_{\mathcal{B}} + \gamma_n \theta_{\mathcal{C}} \tag{6}$$

An unbiased estimator of the raised mutation rate can be proposed as follows

$$\hat{\theta}_{\mathcal{B}} = \frac{S_n - \gamma_n \theta_{\mathcal{C}}}{\beta_n} \tag{7}$$

4 Results and discussion

In this section, we study the behaviour of the estimator given in equation (7) through simulations. Data were simulated as follows. The first step was the determination of $\mathcal{B}$ using the *frequency spectrum* distribution described in equation (1). Then, we built a conditional coalescent tree given $B = b$, with biased inter-coalescence times. We computed both the total length L_n of the tree and the length $\tilde{L}_n$ of the b-subtree, and we simulated the random variable η_n. Finally, we simulated the random variable S_n as a Poisson distributed variable of rate $\hat{\beta}_n \theta_{\mathcal{B}} + \hat{\gamma}_n \theta_{\mathcal{C}}$ where $\hat{\beta}_n = (\tilde{L}_n + \eta_n)/2$ and $\hat{\gamma}_n = L_n/2 - \hat{\beta}_n$. Biased inter-coalescence times were obtained from a rejection algorithm. The

simulation procedure was validated by recovering various known quantities (such as $L_{n,b}$).

The experimental results regarding the estimator $\hat{\theta}_{\mathcal{B}}$ are presented in Table 1. These results were obtained from the procedure above described using the following experimental design. The parameter $\theta_{\mathcal{C}}$ was set equal to a small value $\theta_{\mathcal{C}} = 0.01$. This corresponds to the rough value of a mutation rate $\mu_{\mathcal{C}} \approx 10^{-10}$, the total number of cells $N \approx 10^8$. Four different values for the raised mutation rate $\theta_{\mathcal{B}} = 0.1$, 0.2, 1 and 10 were considered. Sample sizes of $n = 10$, $n = 20$ and $n = 50$ cells were considered. For each simulation we took two configurations of the mutation rate θ_{LMMR}, and observed that this had a weak influence on the result.

	$\theta_{\mathcal{B}} = 0.1$			$\theta_{\mathcal{B}} = 0.2$		
	$\mathbb{E}[\hat{\theta_{\mathcal{B}}}]$	$SD[\hat{\theta_{\mathcal{B}}}]$	$\sqrt{MSE[\hat{\theta_{\mathcal{B}}}]}$	$\mathbb{E}[\hat{\theta_{\mathcal{B}}}]$	$SD[\hat{\theta_{\mathcal{B}}}]$	$\sqrt{MSE[\hat{\theta_{\mathcal{B}}}]}$
$n = 10$						
	0.085	0.48	0.48	0.20	0.60	0.60
$n = 20$						
	0.057	0.38	0.38	0.19	0.71	0.71
$n = 50$						
	0.18	0.58	0.59	0.21	0.66	0.66

	$\theta_{\mathcal{B}} = 1$			$\theta_{\mathcal{B}} = 10$		
	$\mathbb{E}[\hat{\theta_{\mathcal{B}}}]$	$SD[\hat{\theta_{\mathcal{B}}}]$	$\sqrt{MSE[\hat{\theta_{\mathcal{B}}}]}$	$\mathbb{E}[\hat{\theta_{\mathcal{B}}}]$	$SD[\hat{\theta_{\mathcal{B}}}]$	$\sqrt{MSE[\hat{\theta_{\mathcal{B}}}]}$
$n = 10$						
	0.93	1.43	1.42	8.92	11.42	11.44
$n = 20$						
	0.94	1.56	1.56	6.80	11.43	11.84
$n = 50$						
	0.75	1.82	1.84	9.65	16.15	16.11

Table 1. Results of our estimator $\hat{\theta_{\mathcal{B}}}$ on simulations. This table summarises results obtained under various conditions. Simulations were made for a population of $n = 10$, $n = 20$ and $n = 50$ cells in total. For each n, 500 simulations were performed in each 4 cases : $\theta_{\mathcal{B}} = 0.1$ $\theta_{\mathcal{B}} = 0.2$, $\theta_{\mathcal{B}} = 1$ and $\theta_{\mathcal{B}} = 10$.

Table 1 gives the bias, variance and mean squared error estimated over 500 simulations. The results show that $\hat{\theta}_{\mathcal{B}}$ is indeed weakly biased. The major source of bias was the limit of a null mutation rate μ_{LMMR} considered in the theoretical analysis. Nevertheless, the mean squared error is very high, and the distribution of the estimator appeared to be positively skewed. In addition, the variance did not decrease with the sample size. This might be

due the dependence of data within the $\mathcal{B}$ subsample, and the fact that the MRCA of the subsample is expected to be recent.

Although it is crucial in the fight against cancer, detection of the disease at a pretumoral stage is a very difficult issue. In this paper we showed that estimating the raised mutation rate (posterior to loss of MMR) based on the number of segregating sites may not be an efficient method while the use of this estimator is widely spread in more classical population genetics studies.

As well, we observed that the variance of the estimator did not decrease as the sample size increased (from $n = 10$ to $n = 50$). Consequently if DNA analyses of a (supposed tumoral) tissue are necessary, collecting a large number of DNA sequences may not be the best approach for inferring the raised mutation rate. This issue may be overcome by considering several chromosomal loci instead of a single locus as we did. Nevertheless the fact that empirical distributions of the estimator are positively skewed indicates that statistical testing using the number of segregating sites might be lacking power.

References

[Bhattacharyya *et al.*, 1994]N.P. Bhattacharyya, A. Skandalis, A. Ganesh, J. Groden, and M. Meuth. Mutator phenotypes in human colorectarl carcinoma cell lines. *Proc. Nat. Acad. Sc.*, 91:6319–6323, 1994.

[Boveri, 1929]T. Boveri. *Origin of the Malignant Tumors.* Williams & Williams Publising Co., 1929.

[Cairns, 1975]J. Cairns. Mutation selection and the natural history of cancer. *Nature*, 255:197–200, 1975.

[Calabrese *et al.*, 2004]P. Calabrese, J.L. Tsao, Y. Yatabe, R. Salovaara, J.P. Mecklin, H.J. Järvinen, L.A. Aaltonen, S. Tavaré, and Shibata D. Colorectal pretumor progression before and after loss of DNA mismatch repair. *Am. J. Pathol.*, 164:1447–1453, 2004.

[Duesberg and Rasnick, 2000]P. Duesberg and D. Rasnick. Aneuploidy, the somatic mutation that makes cancer species of its own. *Cell Motil Cytoskeleton*, 47:81–107, 2000.

[Duesberg *et al.*, 1998]P. Duesberg, C. Raush, D. Rasnik, and R. Hehlmann. Genetic instability of cancer cells is proportional to their degree of aneuploidy. *Proc. Nat. Acad Sci.*, 95:13692–13697, 1998.

[Griffiths and Tavaré, 2003]R.C. Griffiths and S. Tavaré. The genealogy of a neutral mutation. In P. Green, N. Hjort, and S. Richardson, editors, *Highly Structured Stochastic Systems*, pages 393–412, 2003.

[Kingman, 1982]J.F.C. Kingman. The coalescent. *Stoch. Process. Appl.*, pages 235–248, 1982.

[Loeb and Springgate, 1974]L.A. Loeb and Battula N. Springgate, C.F. and. Errors in DNA replication as a basis of malignant changes. *Cancer Res.*, pages 238–242, 1974.

[Loeb, 1991]L.A. Loeb. Mutator phenotype may be required for multistage carcinogenesis. *Cancer Res.*, 51:3075–3079, 1991.

[Michor *et al.*, 2003]F. Michor, M.A. Nowak, S.A. Franck, and Y. Iwasa. Stochastic elimination of cancer cells. *Proc. R. Soc. Lond.*, 270:2017–2024, 2003.

[Michor *et al.*, 2004]F. Michor, Y. Iwasa, and M.A. Nowak. Dynamics of cancer progression. *Nature Reviews*, 4:197, 2004.

[Pinkel *et al.*, 1986]D. Pinkel, T. Straume, and J.W. Gray. Cytogenetic analysis using quantitative, high sensitivity, fluorescence hybridation. *Proc. Nat. Acad. Sci.*, pages 2934–2938, 1986.

[Plotkin and Nowak, 2002]J.B. Plotkin and M.A. Nowak. The differnet effects of apoptosis and DNA repair on tumorigenesis. *J. Theor. Biol.*, 214:453–467, 2002.

[Shibata *et al.*, 1994]D. Shibata, M.A. Peinado, Y. Ionov, S. Malkhosyan, and M. Perucho. Genomic instability in repeated sequences is an early somatic event in colorectal tumorigenesis that persists after transformation. *Nat. Genet.*, 6:273–281, 1994.

[Stephens, 2000]M. Stephens. Time on trees and the age of an allele. *Theo. Pop. Biol.*, pages 109–119, 2000.

[Tomlinson and Bodmer, 1995]I.P.M. Tomlinson and W.F. Bodmer. Failure of programmed cell death and differentiation as causes of tumors : some simple mathematical models. *Proc. Nat. Acad. Sci.*, 92:11130–11134, 1995.

[Tomlinson and Bodmer, 1999]I.P.M. Tomlinson and W.F. Bodmer. Selection, the mutation rate and cancer: Ensuring that the tail does not wag the dog. *Nat. Med.*, 5(1):11–12, 1999.

[Tsao *et al.*, 2000]J.L. Tsao, Y. Yatabe, R. Salovaara, J.P. Mecklin, H.J. Järvinen, L.A. Aaltonen, S. Tavaré, and Shibata D. Genetic reconstruction of individual colorectal tumor histories. *Proc. Nat. Acad. Sc.*, 97(3):1236–1241, 2000.

[Watterson, 1975]G.A. Watterson. On the number of segregating sites in genetical models without recombination. *Theo. Pop. Biol.*, pages 256–276, 1975.

[Wiuf and Donnelly, 1999]C. Wiuf and P. Donnelly. Conditionnal genealogies and the age of a neutral mutant. *Theo. Pop. Biol.*, pages 183–201, 1999.