Cancer incidence model

Vibishan B.

March 27, 2017

1 Introduction

Accumulation of somatic mutations is a central process in carcinogenesis, and at least five distinct mutations are required to initiate a potential malignancy. The average somatic mutation rate is of the order of 10^{-6} per cell per generation [9, 5, 8], making it highly unlikley that any five cancerous mutations co-occur in the cell by random chance ($(10^{-6})^5 = 10^{-30}$). Given this remoteness, it follows that cancer must be a relatively rare occurence, with a fixed threshold age above which cancer is certain to occur. Both these predictions can be proven wrong based on actual incidence of cancer in the human population in recent years [3]. An elevated mutation rate alone is therefore insufficient to completely explain cancer incidence.

It is more realistic for mutations to accumulate sequentially through a process of clonal evolution, in which a mutation with a positive growth advantage would allow the mutant cells to grow and replace non-mutants. With sufficient expansion, this mutant population could then similarly acquire enough mutations for cancer to occur. This is the current paradigm of **clonal selection** [6], and significant progress has been made to develop this basic view of intra-cellular competition into a more substantive evolutionary account of cancer processes [1, 2, 4]. These perspectives have greatly improved the study of cancer, and opened new avenues of research focused on applying existing evolutionary theory to cancer biology.

However, the clonal selection theory and subsequent developments have consistently ignored an important aspect of cancer incidence that concerns population heterogeneity. By assuming clonality between individuals, the hypothesis concludes that any given somatic mutation must have the same all-time advantage wherever it occurs, leading inevitably to cancer. But this assumption is not necessary from a physiological viewpoint; tissues and micro-environmental factors could vary widely between individuals of the same species, depending on genetic and life style differences. This implies that the exact outcome of an identical mutation need not necessarily be identical amongst members of a population. Selection on somatic mutations could thus be conditional on factors that reach *beyond* the mutation itself, to the local micro-environment that sets the stage for cellular competition. The extra-cellular matrix (ECM), for instance, is a complex combination of proteins and carbohydrates that is secreted by almost all somatic cells. By nature, it is highly dynamic, and has been shown to regulate a diverse array of cellular functions through feedback signaling. More importantly, the exact composition of the ECM differs between members of the same species, mostly in response to environmental factors, which could potentially result in a heterogeneous population. Given these confounding observations, it remains unclear whether or not the assumptions of the clonal selection theory can be accepted without a doubt. In this report, we argue that a complete model of carcinogenesis must explicitly account for such heterogeneity in the population, and posit that our theory of **conditional selection** is capable of doing so.

To demonstrate this, we developed a simple simulation of somatic mutagenesis, in which cells within an individual grow and accumulate mutations sequentially, ultimately leading to cancer. Since the argument here concerns incidence of cancer at the population level, we scaled up our individual-level process to 100000 individuals to generate predictions that can be compared with epidemiological data [3]. Through this comparison, we are able to show that an elevated mutation rate and clonal selection are insufficient to completely explain cancer incidence. We also demonstrate that the conditional selection theory can produce better agreement with this data, while also giving a more realistic relationship of cancer incidence with age than the canonical power law equations [7].

2 Methods

2.1 The process

We begin by considering a group of 100000 individuals, each with a stable poulation of N stem cells. If p is the probability that a single cell in a given generation acquires a single mutation, then the probability that at least one mutation occurs per generation in the population of N cells, based on simple probability, is given by:

$$p_{mut} = 1 - (1-p)^N \tag{1}$$

Since mutations in nature occur randomly, we use the above probability to simulate a stochastic mutation process; if the value p_{mut} exceeds a randomly chosen value between 0 and 1, a mutation is considered to have occurred. For this, we use the *random_sample* function from Python's *numpy.random* package, which generates random numbers between 0 and 1. This gives rise to a new mutant population of size $m = 1$, which continues to grow logistically at a rate g relative to the normal stem cells and replaces them in the tissue. We discuss how the value of g is determined in the simulation later in Section 2.3. Once a mutant population is initiated, we replace N with m in equation 1, and calculate the value of p_{mut} iteratively for the mutant population as it undergoes logistic growth. At some point, one of the cells in this population stochastically acquires a second mutation, which creates a new doubly mutant population. We follow the growth of these cells by setting the value of m back to 1, and allowing them to grow at a rate, g. It must be noted here that this growth rate is now relative to the single mutation population rather than the normal stem cell population. Five such mutation events are considered to constitute a case of cancer. We kill off the current individual upon occurrence of cancer, or at the end of 100 cellular generations, whichever occurs first. We then repeat the entire process for the next, starting from the zeroth generation. We record each instance of cancer and the corresponding age at which it occurs, to get an age-wise count of cancer incidence.

For the purpose of this simulation, we define cancer incidence as the number of cases of cancer in the population per 100000 individuals.

2.2 Effects of p and N on incidence

Each mutation is characterized by both p_{mut} and g. It follows then that different combinations of values of p and N would give different p_{mut} values and therefore, different levels and trends of incidence. We carried out the entire simulation for a range of values of p and N chosen based on existing literature on physiological mutation rates and cell numbers[9, 8]. For each combination of p and N, we calculate normalized cancer incidence, with and without the effect of age on incidence, as detailed above. We also consider separately the relationship that emerges, of total cancer incidence in the population with p and N.

For subsequent exploration, we choose intermediate values of $p = 2 * 10^{-8}$ and $N = 10^7$.

2.3 Determination of g

Since we use g as the growth rate of mutant cells relative to their non-mutant neighbours, it is essentially a proxy for selection acting on mutations in an individual at the population level. The way in which g is determined can therefore vary based on the corresponding hypothesis of cancer incidence, as follows:

Clonal selection Under this model, individuals in the population are clonal, and similar mutations would hence have the same selection advantage. Therefore, in any given run of the simulation, we set the first mutation in each individual to have the same value of g, across all 100000 individuals. As each individual starts off with the same N and p, the first mutation ocurring in each individual is essentially identical. Each subsequent mutation has the same relative growth rate given by g, as mentioned above. Earlier, we determined by trial and error that a growth rate less than 0.75 fails to produce any cancer in the population (not shown here). We therefore use $g = 0.75$ for the clonal selection model.

Conditional selection In this case, individuals still have the same values of N and p (and hence, p_{mut}), but we model g as a normally-distributed random variable, as this view explicitly accounts for a heterogeneous population. For each individual, a different value of g is chosen randomly based on the underlying distribution. Therefore, mutations with identical p_{mut} values can still differ in terms of their selective advantage. We use existing functions from Python's *numpy.random* package to generate this set of random numbers. At this stage, we choose to maintain the same relative growth rates for subsequent mutations occurring within an individual, without considering intra-tumour heterogeneity. Although this would be a simple modification to the simulation, it is not directly relevant to our argument of conditional selection at the population level, and we do not expect it to change our conclusions significantly. We also used the uniform and triangular distributions for g to verify that the qualitative relationship between cancer incidence and age is not sensitive to the exact shape of the distribution of g.

2.4 Normalized cancer incidence

Depending on the distribution of g and the sequence in which the values occur, the fraction of the population surviving to a particular age could vary between subsequent runs. Moreover, absolute count of cases always under-estimates actual incidence at higher age groups as the suriving population size is continuously decreasing. To correct for this, we calculate the normalized cancer incidence for each age by diving the number of observed cases by the size of the suriving population at that age. This gives us the expected fraction of cancer for each age, and by multiplying this fraction by 100000, we obtain the expected number of cancer cases per 100000 for that age. As available epidemiological data is in the form of age group distributions, we calculate the incidence for the given age groups by adding all the expected fractions in that range and multiplying the sum by 100000. We consider this number to be the model's prediction of cancer incidence in the population.

2.5 Effect of age on g

It has been argued that somatic maintenance declines with age[7], and from a life history perspective, this does not seem entirely implausible. Evolutionary theories of ageing describe a steady decrease in selection pressure in the later stages of life beyond the age of peak fecundity. The "carcinogencitiy" of mutations could therefore increase with age as suppression mechanisms decline in efficiency. We model this by adding various monotonously increasing functions to the initial value of g at each age. Therefore, the value of g increases continuously within the lifespan of every individual starting from the initial value obtained from the random number generator, as dictated by the age function.

Codes for all simulations were written in Python, and the data were exported to Excel for plotting and analysis.

3 Results and Discussion

- Clonal selection predicts a discrete threshold age above which cancer is a certain occurence. It is therefore unacceptable as a plausible mechanism of carcinogenesis.
- Conditional selection is in better agreement with actual data.
- Four parameters-p, N, and the mean and variance of the distribution of g. We could discuss the effects of these parameters.
- Is it possible to understand differences in cancer incidence between various ethnic and geographic groups through changes in these four parameters?
- We get a unimodal curve for cancer incidence with age, as a direct result of the exact combination of mean and variance for g. Do we interpret this curve, or the parameters that gave rise to it?