# Statistical Significance of Patterns from Small Dataset Phylogenetic Studies

## 1. Introduction

Understanding biology on a microscopic scale is a feat that we have made more trivial through advancements in lab techniques and developments in stable environments and sterilization that allow us to derive biological processes with fewer confounding variables. However, the few confounding variables we have yet to solve in biology can have a sizable impact on the significance of biological processes observed in wet lab settings. One such process in the pathway of normal cells to cancerous cells, a process which has long been somewhat indiscernible as wet lab experimentation can add stress to the cells of interest which could have resulted in a myriad of the cell states supposedly part of this pathway. Some have tried a computational approach to this problem however such studies are purely theoretical models which do not directly interact with the biological process of interest, and therefore can only have so much bearing on the canonical understanding of this process. As such, we would like to utilize phylogenetics, the study of deriving evolutionary relationships throughout time and a combination of wet lab and computational studies, to aid in identifying cell states within this cell pathway. Particularly, we plan to use single-cell phylogenetics, which uses segments nucleic information from individual cells as the extant lineages for the phylogeny, coupled with hidden states to derive the evolution of cells within a singular microenvironment whilst derived the ideal number of intermediate cell states as per the phylogeny.

While phylogenetics can produce fruitful and statistically significant patterns, it does employ simplifications of deeply complex mechanisms of biological change. However, we do not believe this severely detracts from the validity of this study.

## 2. Background

Phylogenetics traces back the evolution of various lineages using genomic data with lineages representing species, genuses, individual cells, subclones, and more depending on the context of the study. Deriving these relationships necessitates understanding the probability of deviations in genetic sequences between different lineages. The probability of such shifts is often represented using a matrix, in the case of genetic mutations this table is a four by four table as shown below:

$$
P(dt) = \begin{array}{cc}
 & \begin{array}{cccc} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\
\begin{array}{c} \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array} &
\begin{bmatrix}
1-(\alpha\pi_C+\beta\pi_A+\beta\pi_G)dt & \alpha\pi_C dt & \beta\pi_A dt & \beta\pi_G dt \\
\alpha\pi_T dt & 1-(\alpha\pi_T+\beta\pi_A+\beta\pi_G)dt & \beta\pi_A dt & \beta\pi_G dt \\
\beta\pi_T dt & \beta\pi_C dt & 1-(\alpha\pi_G+\beta\pi_T+\beta\pi_C)dt & \alpha\pi_G dt \\
\beta\pi_T dt & \beta\pi_C dt & \alpha\pi_A dt & 1-(\alpha\pi_A+\beta\pi_T+\beta\pi_C)dt
\end{bmatrix}
\end{array}
$$

The matrix represented above is the HKY85 [1], the most recent iteration of nucleotide substitution probability matrices. It should be noted that as per current phylogenetic techniques it is assumed that all genetic deviations occur due to substitution as there are no current techniques to model inserts and deletions for evolutionary processes. Such matrices are used in functions known as likelihood functions, with the solution to these functions being the phylogeny. Recent phylogenetic techniques employ Bayes theorem, a powerful statistical theory that uses expert knowledge in the form of priors to mitigate impact of skewed data results.

Tracing nucleotide changes is the basis of relationships established between different lineages in phylogenetics and allows us to create a framework to extrapolate other information such as traits. We define traits as “states” within phylogenies and use the framework of the phylogeny to overlay the states on the tree. These states have their own properties, specifically birth-rate, death rate, and diversification rate. States can be divided into two separate groups: conventional states and hidden states. Conventional states are the default when referring to states and consist of traits or characteristics that can be observed in the data (also known as the extant lineages) from which the phylogeny is derived and require a fixed number of states. Hidden states are traits or characteristics that cannot be determined from the data from which the phylogeny is created and do not require a set number of states. Such state types are particularly useful in single-cell phylogenetic trait studies as some cell states are only easily identifiable using transcriptome or morphological data which can be difficult to score as they are dependent on the environment in which the cell properties are obtained [2].

In this study, we will employ hidden states to discern not only the number of derived states within this cellular process but also the characteristics of these individual states. While this representation of this process will be discretized due to the nature of states within phylogenetics, it should be noted that these cellular processes are likely spectrum-like processes and therefore individual cell states will not convey the full picture of this intricate process. Current phylogenetic techniques do have the capability to model spectrum-like processes such as OU and Brownian; however, such processes assume that the evolutionary process modeled has a semblance to linear evolution but such is not the case with cellular evolution [3][4]. Using discretized forms in this phylogenetic analysis allows us to circumvent this assumption and allow the cellular process to return to similar cell states as before and therefore be able to better determine the overarching structure of this process.

## 2.A. Cancer biology

As per current scientific understanding, the cellular transition of normal cells to cancerous cells contains three distinct cell types: mitotically-overactive cells (which we have dubbed exposed cells for reasons stated later in this paper), senescent cells, and crisis cells. Identifying the previously listed intermediate states would require transcriptomic data [5][6][7], which we have already stated is dependent on the environment in which the data is collected therefore not standardized amongst different studies and also is difficult to obtain in conjunction with genomic data which is needed to construct a phylogeny [8].

To mitigate this issue, we propose two discrete observed traits, normal and cancer, with hidden states with the normal phenotype composed of the hidden states normal and exposed and the cancer phenotype composed of the hidden states senescent, crisis, and cancerous [9]. Each of the hidden states will have their own respective birth and death rates which we will derive from the phylogenetic analysis using strong priors based on the known potential characteristics of the birth and death rates for hypothetical phenotypes part of this process. However, obtaining hidden state transition patterns and their individual properties requires large sets of data typically spanning 500 to 1000 samples []. It should be noted that this paper was written on July 25, 2024 under the assumption that such a large genomic data set with normal and cancerous cells can feasibly be found as multiple studies providing transcriptomic data from 1000+ on this particular topic have been found.

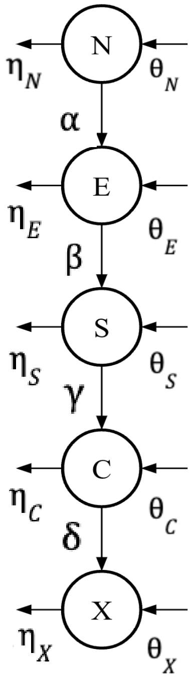
The transition from normal to cancerous cells is a “viral” process as cancer exosomes, vesicles composed of miRNA and proteins,  infect normal cells leading to a cascade of changes, represented through the different states the cell goes through, to potentially turn into a cancerous cell. Through this “viral” mechanism, cancerous cells can alter the transcriptome of the noncancerous cells surrounding the cancerous cell therefore changing the noncancerous cell’s phenotypic presentation, resulting in cancerous behavior [10]. This process is facilitated by Dicer among other proteins needed for miRNA in cancer exosomes to silence the PTEN and HOX10 genes, both of which regulate mitotic and differentiation activity, within normal neighboring cells [5]. When these cells were injected into mice they resulted in tumors, suggesting the normal cells had become cancerous through exposure to cancer cell exosomes. The typical process of cancer cell development starts with a normal cell with a mutation that results in rapid proliferation which shortens telomeres, a sequence of buffering DNA used to prevent coding DNA loss during DNA replication, to a critical state [11]. This critical state results in senescence, a stage in which the cell does not undergo birth or death processes [12]. In some instances, these cells can escape into crisis, a cell state characterized by rapid proliferation and rapid cell death that entails chromosome end fusion which results in chromosome instability [13]. Throughout this process, cells can undergo mutations that can result in a cancerous cell type [11].

## 3. Model

We are proposing a modified SIR model, a model typically used to represent the epidemiological evolution of viruses, coupled with a birth-death model to account for the multifaceted nature of cancer evolution within a singular site. As mentioned, much is known about this process past the exposed cell type so we propose a NE model which includes the transition from normal to exposed mirroring the transition from susceptible to infected, but from there the hidden states within the model determine the derive trajectory of this pathway.

Below we have represented an idea of what a derived trajectory would look like under the assumption that the derived pathway is the following: normal, exposed, senescent, crisis, and cancerous (which we will dub as NESCC). Please note in a conventional SIR model, it is assumed the population is stagnant which does not reflect the cancer microenvironment and as such we added birth and death events within the differential equations for each population and including the BDS model within our likelihood function [14].

The markov chain is as follows:



$$\frac{dN(t)}{dt} = \frac{-\alpha N(t)X(t)}{n} + N(t) * (\theta_N - \eta_N) \qquad \text{(Equation 1)}$$

$$\frac{dE(t)}{dt} = \frac{\alpha N(t)X(t)}{n} - \beta^{\theta_E}E(t) + E(t) * (\theta_E - \eta_E) \qquad \text{(Equation 2)}$$

$$\frac{dS(t)}{dt} = \beta^{\theta_E}E(t) - \gamma S(t) + S(t) * (\theta_S - \eta_S) \qquad \text{(Equation 3)}$$

$$\frac{dC(t)}{dt} = \gamma S(t) - \delta C(t) + C(t) * (\theta_C - \eta_C) \qquad \text{(Equation 4)}$$

$$\frac{dX(t)}{dt} = \delta C(t) + X(t) * (\theta_X - \eta_X) \qquad \text{(Equation 5)}$$

Figure 1. The following figure depicts the markov chain in the model with N for noncancerous cells, E for exposed cells, S for senescent cells, C for crisis cells, and X for cancerous cells. θ and η represent the birth and death rates respectively and are used in the differential equations to derive the true birth rate which is used to determine the change in each cell population. In Equations 2 and 3 the transition rate β is taken to the power of $\theta_E$ as the transition from exposed to senescent cells is dependent on the number of birth events undergone by the cell. In Equations 1 and 2 the expression $\frac{X(t)}{n}$, with $n$ referring to the number of total cells in the system, can be assumed to be 1 as we make the assumption that a significant number of the cells in the microenvironment are cancer cells.

## 4. Simulation

While phylogenetics can prove to be a successful tool in uncovering the process of normal to cancerous cells, it is important we acknowledge some of the pitfalls of this tool, most notably its possibility for error or inability to derive such complex processes. For this reason, we want to use CompuCell [15], an intracellular interaction simulator, to simulate the transition from normal to cancerous cells under the NESCC pathway and run through our phylogenetic analysis to see if we can pick up the NESCC pattern. This will provide us a frame of reference for the accuracy of phylogenetic analyses in this context and validate or disregard the process recovered by our phylogenetic analysis of real data.

The simulation will start with a random number of normal cells primarily using biological accurate information since we do not need to worry about the same computational limitations here as we do in our phylogenetic model.

In this representation of the cancer microenvironment, we have the Dicer exosome act as a factor that "infects" the normal cell. After the normal cell is exposed to the Dicer exosome, the cell will then transition to the exposed state. As the exposed state cells rapidly divide, they lose base pairs in their telomeres resulting in a senescent state. To record the nuances of this transition we added a new attribute to the cells in CompuCell dubbed "telomere_length" which monitors the length of the telomeres of each cell in the simulation. The telomere length starts at 3000 base pairs and decreases by approximately 50 to 200 base pairs per division [16]. Then the transition from the senescent to crisis states are at random using randint as the mechanism behind this transition is still unknown. We also used the same approach for the transition between the crisis and cancerous states.

To construct a phylogeny from the simulated data we also added another attribute to the cells in CompuCell dubbed "mutations" which records the mutations in a 1000 base long part of the cell's genetic material that result from each cell division. Although phylogenetic techniques only account for substitutions, we will also include common point mutations such as deletions and insertions into the simulation to best mimic the behavior of real cells undergoing the pathway of interest. Alexandrov et. al characterized mutational signals from most types of cancer and found 79,793,266 somatic single-base substitutions, 814,191 doublet-base substitutions and 4,122,233 deletions and insertions and we will base the frequency of these mutations events in the simulation based on their findings [17]

While we are modeling this pathway as discrete in the simulation, we acknowledge that phenotypes can be somewhat fluid and do not adhere to cellular behavior in that manner. We also opted to disregard the characteristics, notably volume discrepancies and secretions of factors other than the Dicer exosome, of the cell types in the simulation as we wanted to probe the pathway of normal to cancerous in a broader scope.