DNA Sequence Inheritance in Parasite Populations

Aiden Lewis

April 2024

1 Introduction

DNA sequences are the fundamental building blocks of the genetics of all living organisms: they contain the instructions required for cells to function, encoded as long strings of molecules called nucleotides. The four nucleotide bases are adenine, cytosine, guanine, and thymine, commonly referred to as A, C, G, and T in shorthand. During the cell replication process, an enzyme called DNA polymerase reads and copies the base sequence; it is meant to be an exact copy, but the enzyme is not infallible, and it can make errors during transcription. These are called mutations, and they are sometimes passed down from an individual to its offspring. At a macroscopic level, this process is the driving force behind evolution – certain mutations can offer a competitive advantage, allowing them to spread across a population and become "standard." Over time, two separate populations of the same original species will gradually diverge more and more, until they become sufficiently different to be considered distinct species. This is generally true for all kinds of organisms, but we shall be focusing specifically on *Plasmodium* – the genus of parasites that causes malaria – due to its short reproductive cycle.

We can describe the mutation process using nucleotide base substitution models, which use continuous-time Markov chain algorithms to simulate the evolution of a base sequence. By comparing corresponding base sequences in two individuals, we can then estimate the genetic distance between them, and this serves as a measure of the extent to which the populations they belong to have diverged from their most recent common ancestor. Different substitution models will produce different genetic distance values, and this paper seeks to compare and contrast them accordingly using a toy model of *Plasmodium*.

2 Mathematics

2.1 Continuous-time Markov chains

A Markov chain is a stochastic process in which a variable X_n experiences a transition from step n to step $n+1$: $X_n \to X_{n+1} = A \to B$, $A, B \in S$, where S is the state space of the process. Governing it is the transition matrix P, where the elements $P_{AB} = P(X_{n+1} = B | X_n = A)$. The time steps are either discrete or continuous – in the former case, dt and P are both constant across the entire process. In the latter, however, the time step is dynamic: dt is instead an exponential random variable representing the length of time the next event took to occur, and P changes accordingly. For a stationary process such as this one, dt can be replaced with t, thereby eliminating the iterative aspect. It is represented as:

$$P = e^{Qt}, \tag{1}$$

where Q is the rate matrix of the process and t is the total time. With regards to substitution models specifically, they differ largely in the rate matrices they use and the number of free parameters governing them.

One of these parameters, and the only one common to all substitution models, is the overall substitution rate μ (normalised to 1). This is the parameter of the exponential distribution dt follows, and as such:

$$\langle dt \rangle = \frac{1}{\mu}.$$

A more useful variable for our purposes, however, is the number of substitutions per site between iterations ν:

$$\nu = -\delta\mu dt, \tag{2}$$

where δ is the value of Q's diagonal elements with (all the same, definitionally chosen such that the rows of Q sum to 0 – a requirement of a stationary process such as this). Though the process is not inherently iterative, we are nonetheless simulating it that way in order to make measuring genetic distance easier. As each iteration is meant to represent a mutation spreading throughout the population, we would ideally like the expected number of substitutions per time step to be 1. For each site, we thus have:

$$\langle\nu\rangle = \frac{1}{n},$$

where n is the total length of the sequence. Rewriting this in terms of t:

$$\langle\nu\rangle = -\delta\mu\langle t\rangle \rightarrow \langle t\rangle = -\frac{\langle\nu\rangle}{\delta\mu} = -\frac{1}{\delta\mu n}.$$

Thus, it makes more sense to scale the parameter of the exponential distribution by a factor of $-\delta n$.

2.2 Genetic distance

The easiest and most intuitive way to measure genetic distance d is by directly comparing two base sequences and finding the number of differences per site – the term for this is observed distance, or p-distance. As it fails to account for any substitutions that may have occurred in prior generations, it is of limited use on its own when directly simulating a future generation. We can, however, get around this by instead explicitly simulating each successive generation as outlined in the previous section: in that case, we can sum the p-distances for every generation to get the total distance d. While this does provide a fairly robust and reliable method for estimating d, the iterative simulation is also significantly more computationally expensive than the direct simulation, rendering it impractical for use in solving sufficiently large-scale problems.

2.3 Substitution models

2.3.1 JC69

The first substitution model was developed in 1969 by Jukes and Cantor, and was thus dubbed JC69. It is the simplest possible model, assuming that all base frequencies are equal, and its rate matrix is:

$$Q_{JC69} = \frac{\mu}{4}\begin{bmatrix} \cdot & 1 & 1 & 1 \\ 1 & \cdot & 1 & 1 \\ 1 & 1 & \cdot & 1 \\ 1 & 1 & 1 & \cdot \end{bmatrix} \tag{3}$$

The diagonal elements $\cdot$ are chosen such that each row sums to zero.

2.3.2 K80

The Kimura 1980 model expands upon JC69 by allowing for different rates of transitions and transversions, their ratio being κ:

$$Q_{K80} = \frac{\mu}{4}\begin{bmatrix} \cdot & 1 & \kappa & 1 \\ 1 & \cdot & 1 & \kappa \\ \kappa & 1 & \cdot & 1 \\ 1 & \kappa & 1 & \cdot \end{bmatrix} \tag{4}$$

Transitions are purine-purine and pyrimidine-pyrimidine substitutions (the purines being A and G, and the pyrimidines being C and T), while transversions are purine-pyrimidine (and vice versa) substitutions. Due to the molecular structures of the nucleotide bases, transitions are more likely to occur than transversions. As such, we should expect that κ is greater than 1 in general.

2.3.3 F81

The Felsenstein 1981 model expands upon JC69 by allowing for different base frequencies:

$$Q_{F81} = \mu \begin{bmatrix} \cdot & \pi_C & \pi_G & \pi_T \\ \pi_A & \cdot & \pi_G & \pi_T \\ \pi_A & \pi_C & \cdot & \pi_T \\ \pi_A & \pi_C & \pi_G & \cdot \end{bmatrix} \tag{5}$$

2.3.4 HKY85

The Hasegawa, Kishino, Yano 1985 model combines the additions of K80 and F81, allowing for both different base frequencies and different rates of transitions and transversions:

$$Q_{HKY85} = \mu \begin{bmatrix} \cdot & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & \cdot & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & \cdot & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & \cdot \end{bmatrix} \tag{6}$$

2.3.5 TN93

The Tamura and Nei 1993 model expands upon HKY85 by allowing for different rates of pyrimidine and purine transitions, their ratio being γ:

$$Q_{TN93} = \mu \begin{bmatrix} \cdot & \pi_C & \tau\pi_G & \pi_T \\ \pi_A & \cdot & \pi_G & \gamma\tau\pi_T \\ \tau\pi_A & \pi_C & \cdot & \pi_T \\ \pi_A & \gamma\tau\pi_C & \pi_G & \cdot \end{bmatrix} \tag{7}$$

$$\tau = \frac{2\kappa}{1+\gamma}$$

2.3.6 GTR

The general time-reversible model was developed by Tavaré in 1986, and it is the most fundamental of the substitution models – all the others are essentially GTR with different parameter sets. Its rate matrix is:

$$Q_{GTR} = \mu \begin{bmatrix} \cdot & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & \cdot & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & \cdot & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & \cdot \end{bmatrix} \tag{8}$$

Despite its generality, GTR is not always the best model to use for a given situation; we shall get further into this in a later section.

3 Methods

We implemented our model in the programming language Python, used for its object-oriented nature and strong selection of scientific computing-oriented packages. The two object types we created were Sequences and Models – a Sequence was a representation of a nucleotide base sequence, with methods to find data of interest like total base counts and genetic distance, and a Model essentially assigned linked a model's name with the corresponding rate matrix. Each Model object was pre-defined in the code, whereas Sequences were dynamically generated.

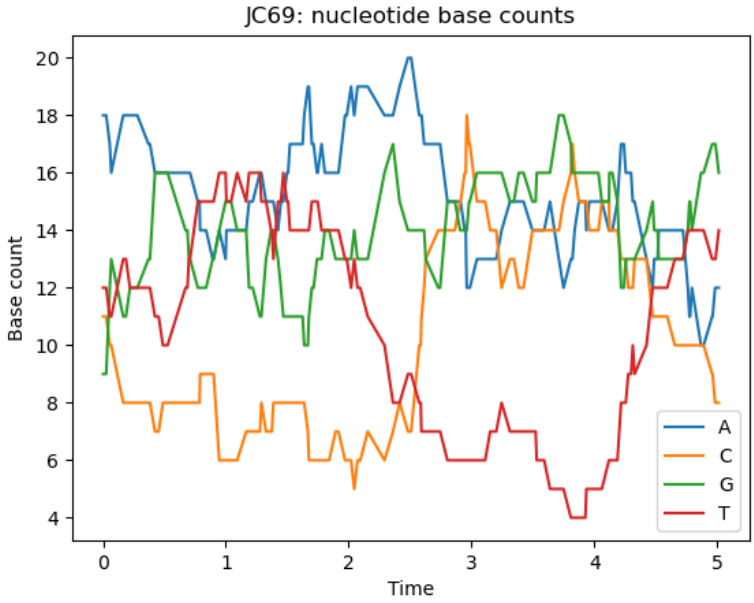
Two methods featured particularly prominently in the simulation process: the base simulation method itself and the genetic distance calculation method. The former modelled the evolution of two different base sequences over time according to the continuous-time Markov chain described previously; the latter,

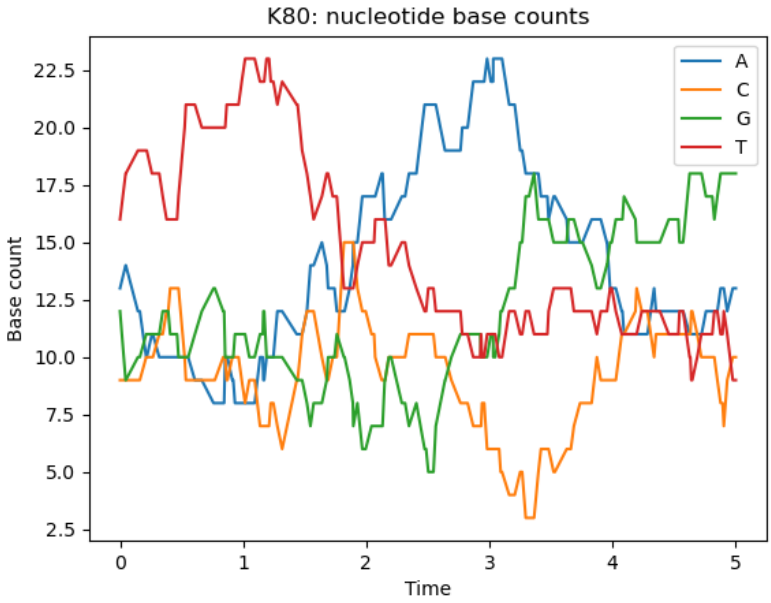
meanwhile, was somewhat more complex. One of the issues introduced by comparing the outputs of two continuous-time Markov chains is that the times visited by each simulation will in general all be different from one another, and so a direct element-wise comparison is impossible. To solve that problem, we "duplicated" each sequence at the times visited by the other sequence, and that allowed us to then do the necessary element-wise comparison.

4 Results and Discussion

4.1 Base frequencies

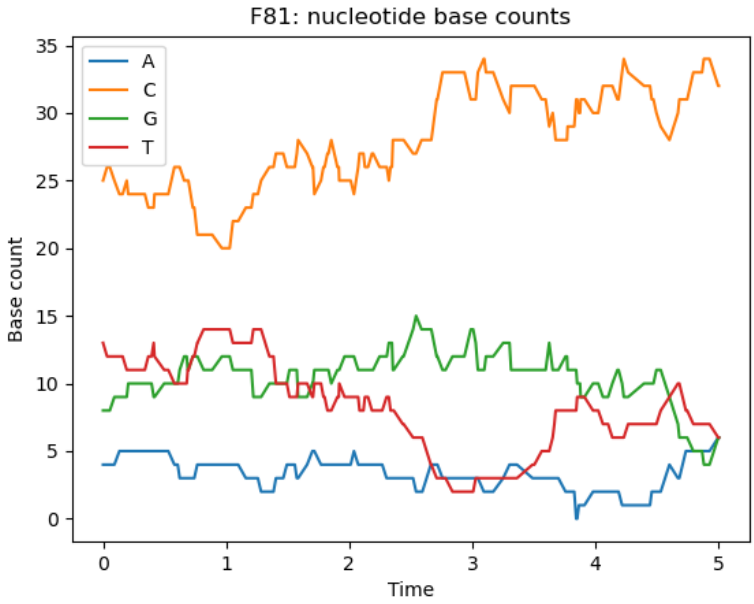
4.1.1 Equal base frequency models (JC69, K80)

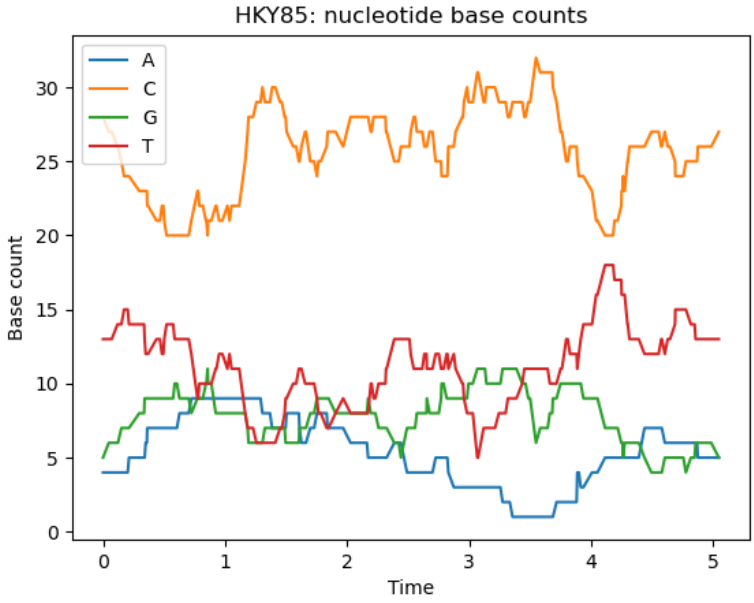


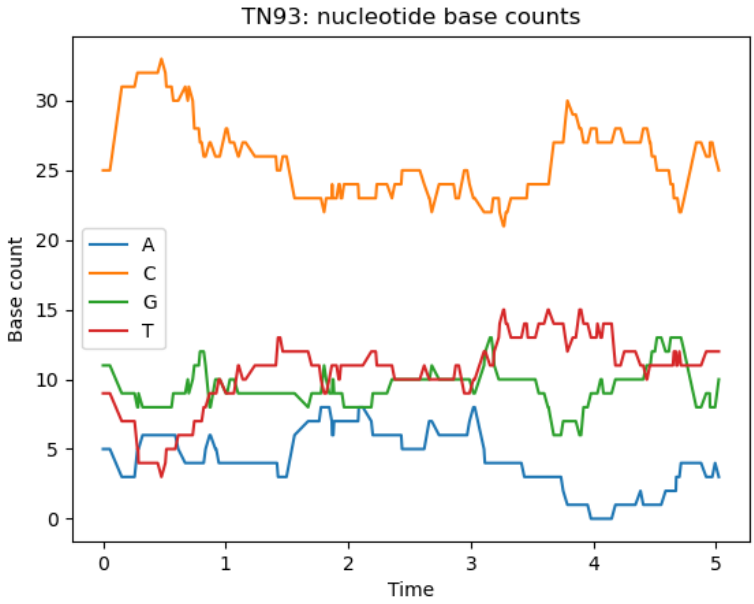


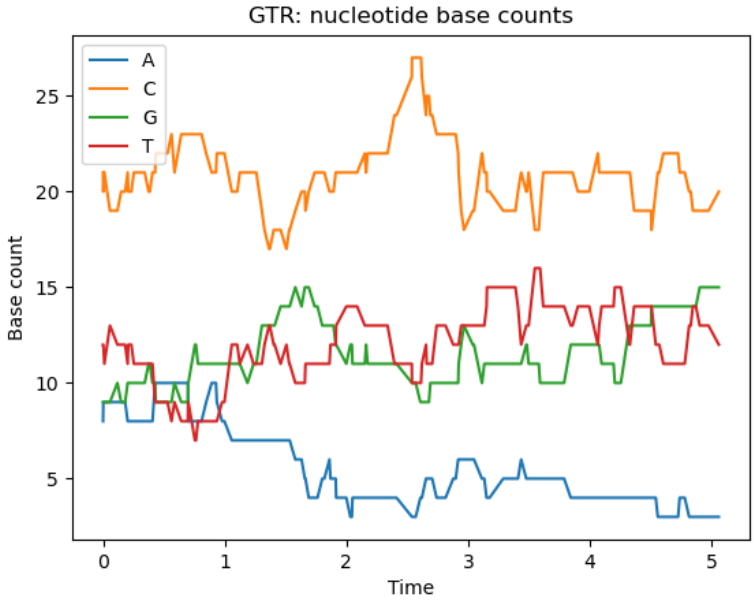
We modelled a sequence of 50 bases over a total evolutionary timespan of 5. There is, evidently, no pattern followed by the base counts over time; their values are effectively random. As these two models set the base frequencies to be equal, this is to be expected.

4.1.2 Different base frequency models (F81, HKY85, TN93, GTR)



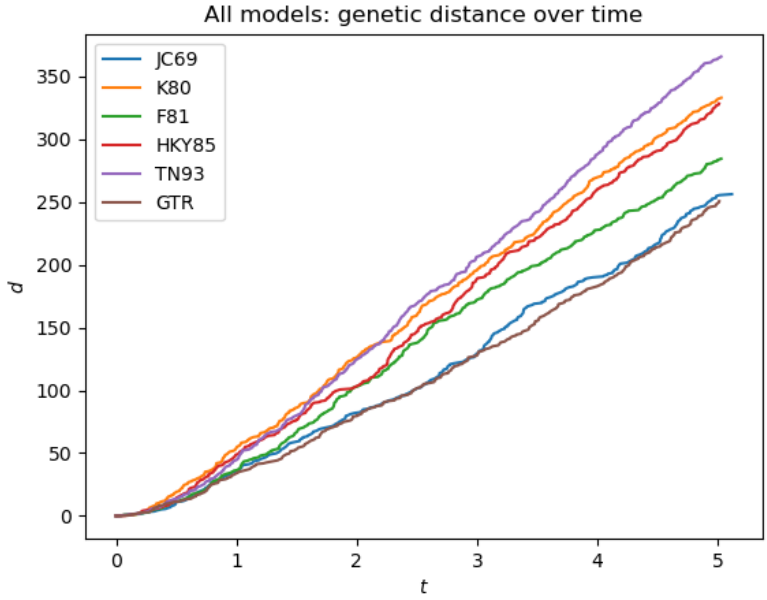




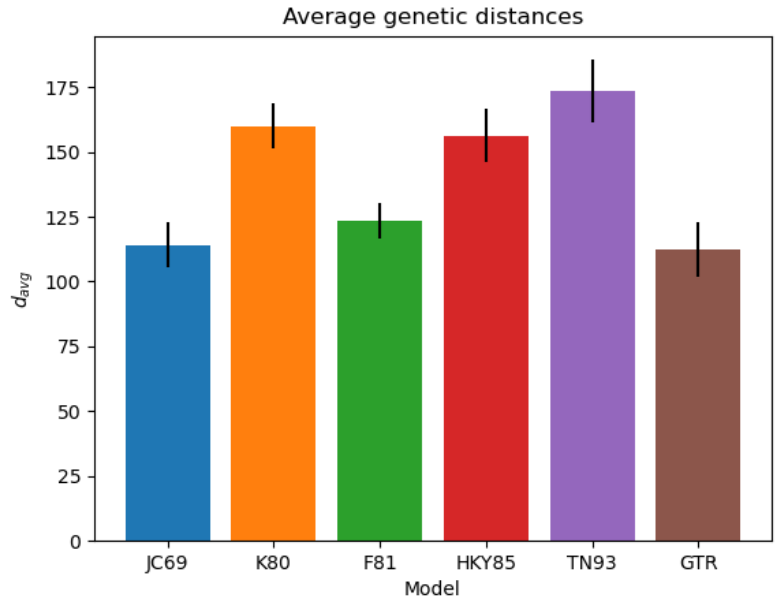


We used a base frequency distribution of $\pi = [0.1, 0.5, 0.2, 0.2]$, corresponding to A, C, G, and T respectively, and that is clearly visible in the above graphs: on average, each base is present at roughly the rates given by π. Considering that π is meant to be a stationary distribution, this serves as a reasonable way of verifying that our model is performing as expected; since each model obeys the same stationary distribution, however, it is not very useful as a point of comparison.

4.2 Genetic distance



Above is the output of a single simulation of genetic distances for each model; some evidently diverge faster than others, and all exhibit an approximately linear relationship between d and t (as we should expect them to). For a stochastic process such as this, though, it is best to look at the average of many simulations, rather than just one.



The genetic distances were averaged over a total of 20 separate simulations; more simulations can be used to improve the data, but the improvement is sufficiently marginal as to render the added computational expense unjustified. The black lines represent the standard deviations of the models' datasets. Clearly, the results of each model are meaningfully different from one another – JC69 and F81 produce the lowest genetic distance estimates, while K80, HKY85, and TN93 produce the highest estimates. Interestingly, all of these

models include one particular parameter that the other two do not: κ, the transition/transversion rate ratio. This would suggest that κ has a stronger influence on the rate of genetic drift than does π, which is a very intriguing result indeed.

Regarding GTR, its emphasis on generality means that it is bound by no biology-based constraints; it can be made to produce results of any sort, including those of all the other models, and so it is not worth studying much on its own (at least, not for a toy model like ours).

5 Conclusions

Overall, there is no singular 'best' substitution model: each of them chooses a set of parameters to suit a specific purpose, and their strengths and weaknesses differ depending on the context they are being evaluated in. It is often preferable to limit the number of free parameters and infer the rest from biological principles; each free parameter needs to be estimated from the data, and so any errors in the data will have an increased effect on the quality of the results depending on how many free parameters there are. As an example, it is typically easier to collect data on nucleotide base frequencies than on transition/transversion rate ratios – in such cases, F81 may prove most effective. On the other hand, if κ is known with reasonable accuracy, HKY85 may instead be the superior model.

That said, there *are* still quality differences between the models – K80 and, in particular, JC69 are the worst, as their assumptions are quite biologically unreasonable. HKY85 strikes a fair balance between biologically reasonable assumptions and data flexibility, making it relatively versatile, and TN93 allows for more transition and transversion rate information to be included in the model.

6 References

1. H. Honma et al., "Mutation tendency of mutator *Plasmodium berghei* with proofreading-deficient DNA polymerase δ" (2016), Nature. Retrieved from `https://www.nature.com/articles/srep36971`.

2. N. Lanchier, *Stochastic Modeling* (2017), Springer, p. 101-124.

3. P. Lemey et al., *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing* (2009), Cambridge University Press, p. 111-123.

All code can be found on Github: `https://github.com/aclewis242/APM530_final.git`