DNA Sequence Inheritance in Parasite Populations

Aiden Lewis

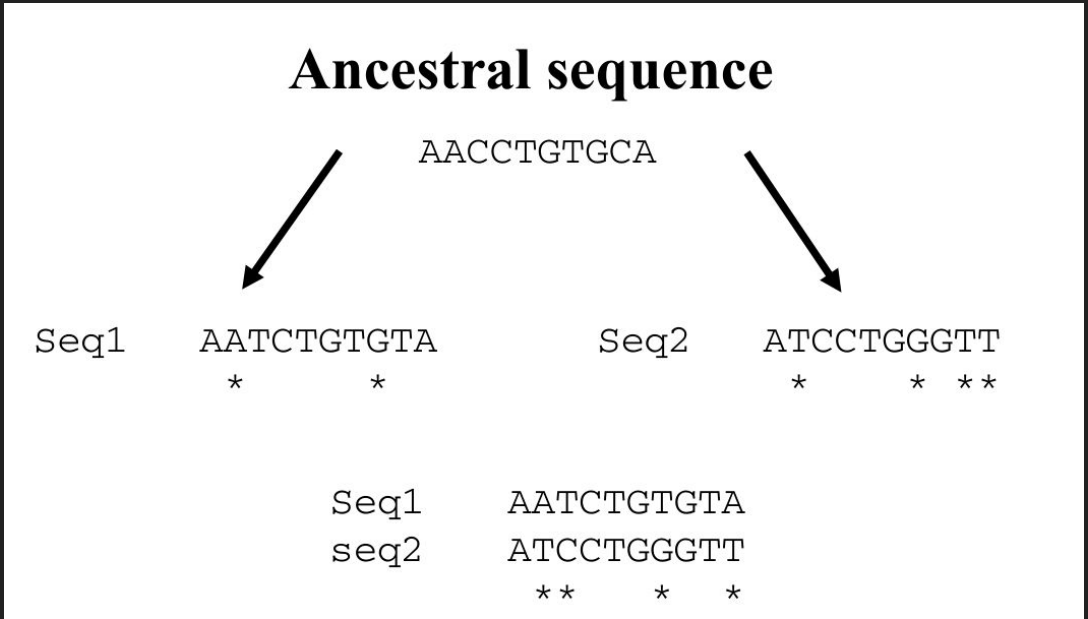
Introduction: DNA basics

- DNA encodes a cell’s genetics as sequences of nucleotide bases (adenine, cytosine, guanine, thymine)
- During cell replication, DNA polymerase reads these sequences and copies them. Errors in this process are called mutations
- Mutations drive evolution: they can give individuals advantageous traits that then spread across the whole population
- Two separate populations of a common species will eventually diverge from one another, becoming distinct species

Introduction: modelling mutations

- During transcription, each nucleotide base has a certain probability of undergoing a transition to another base. This allows us to represent the process as a Markov chain
- These are called substitution models, and they describe the evolution of a base sequence over time
- We can compare two different base sequences sharing a common ancestor to estimate the genetic distance between them
- We shall use a toy model of *Plasmodium*, the genus of parasites that causes malaria, to evaluate six different common substitution models

Introduction: modelling mutations



Continuous-time Markov chains

- A Markov chain is a stochastic process where a variable goes from state *A* to state *B* with a probability described by a transition matrix *P*
- In the stationary, continuous case, we represent this matrix as:

$$P = e^{Qt},$$

  where *Q* is called the rate matrix. Each substitution model uses a different *Q*, depending on its purpose
- Since measuring genetic distance is much easier when you have the full substitution history, we have modelled it iteratively, choosing *t* such that 1 substitution occurs per generation on average

Substitution models: JC69

- Developed by Jukes and Cantor in 1969
- The simplest possible model – all base frequencies are equal, and all substitutions happen at equal rates
- The only free parameter is the overall substitution rate μ

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

Substitution models: K80

- Developed by Kimura in 1980
- Expands upon JC69 by adding the ratio of transition and transversion rates κ
  - Transitions are purine-purine/pyrimidine-pyrimidine substitutions (A and G are purines, C and T are pyrimidines); transversions are purine-pyrimidine (and vice versa) substitutions
  - Transitions are more likely than transversions

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

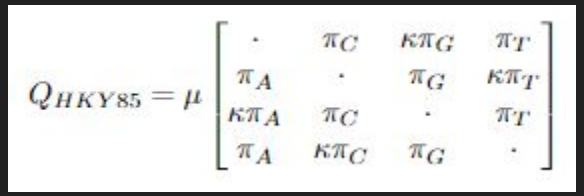
Substitution models: F81

- Developed by Felsenstein in 1981
- Expands upon JC69 by allowing for different base frequencies

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

Substitution models: HKY85

- Developed by Hasegawa, Kishino, and Yano in 1985
- Combines K80 and F81, adding both the transition/transversion rate ratio and different base frequencies



Substitution models: TN93

- Developed by Tamura and Nei in 1993
- Expands upon HKY85 by allowing for different rates of pyrimidine and purine transitions, defined by their ratio *γ*

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

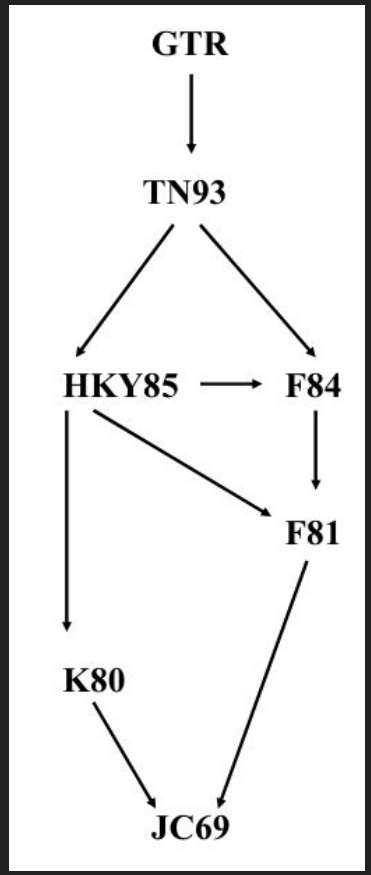
$$\tau = \frac{2\kappa}{1+\gamma}$$

Substitution models: GTR

- “General Time-Reversible”
- Developed by Tavaré in 1986
- Includes all possible free parameters
  - All other substitution models are essentially GTR with different parameter sets
- Full generality is not necessarily a good thing, though

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

Substitution models

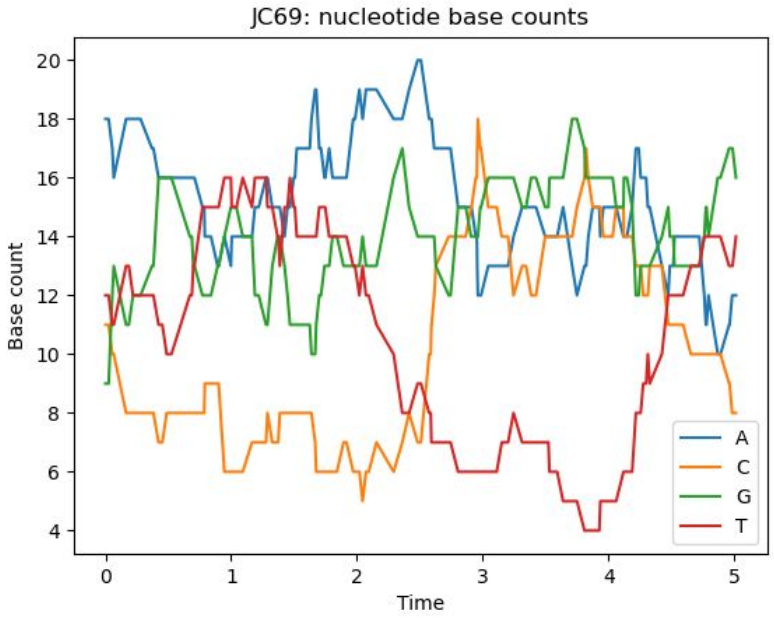


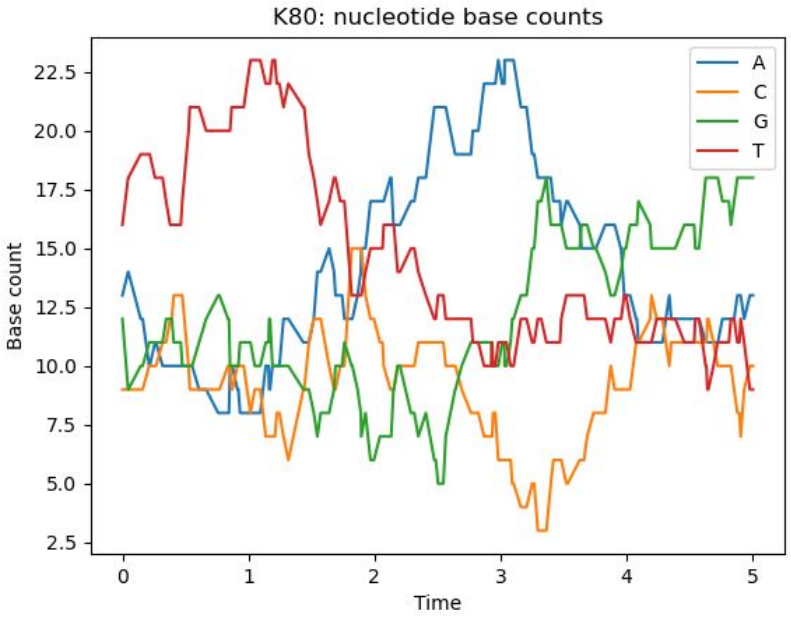
Methods

- Object-oriented representation in Python
  - Allowed for intuitive representations of biological and mathematical concepts: e.g., the “Sequence” object contained a method for finding the genetic distance between two base sequences, and the “Model” object stored a specific model’s rate matrix
- Keeps the code flexible and versatile by avoiding unnecessary constraints

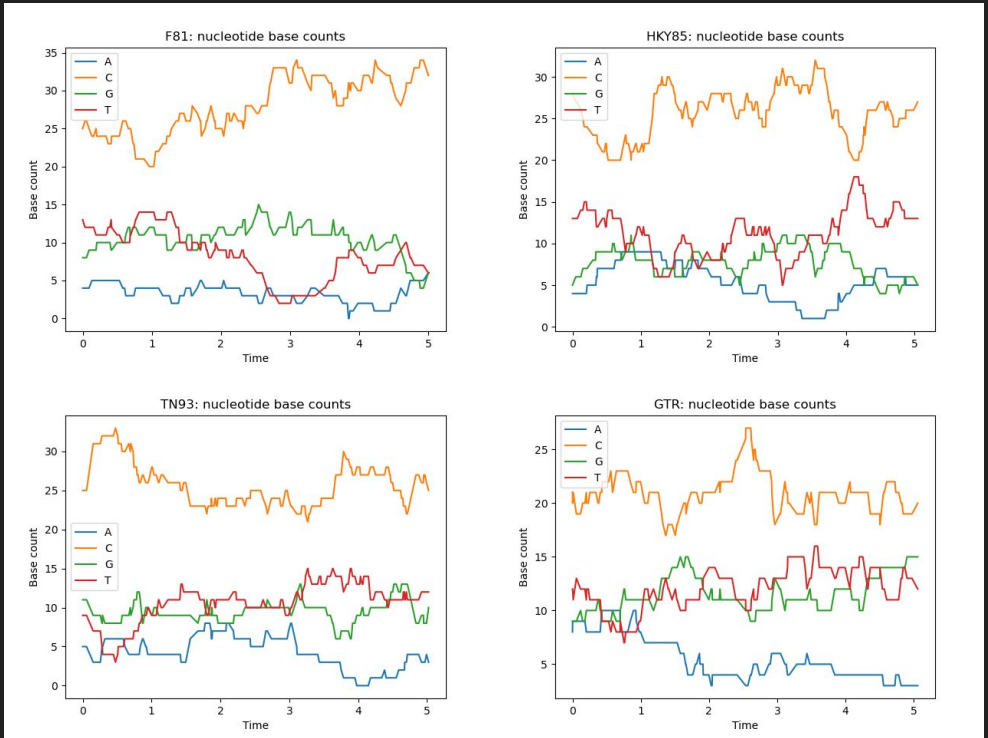
```
m = 1.0
JC69 = Model('JC69') # Jukes & Cantor 1969
JC69.addRM(np.array([[N, 1, 1, 1],
                     [1, N, 1, 1],
                     [1, 1, N, 1],
                     [1, 1, 1, N]])*m)

k = 2.0 # Transition-transversion ratio
K80 = Model('K80') # Kimura 1980
K80.addRM(np.array([[N, 1, k, 1],
                    [1, N, 1, k],
                    [k, 1, N, 1],
                    [1, k, 1, N]])*m)

bfs = {A: 0.1, C: 0.5, G: 0.2} # Base frequencies (T implicit, as they must sum to 1)
F81 = Model('F81', bfs) # Felsenstein 1981
F81.addRM(JC69.rate_mat*4)

HKY85 = Model('HKY85', bfs) # Hasegawa, Kishino, Yano 1985
HKY85.addRM(K80.rate_mat*4)
```

Results: relative base frequencies





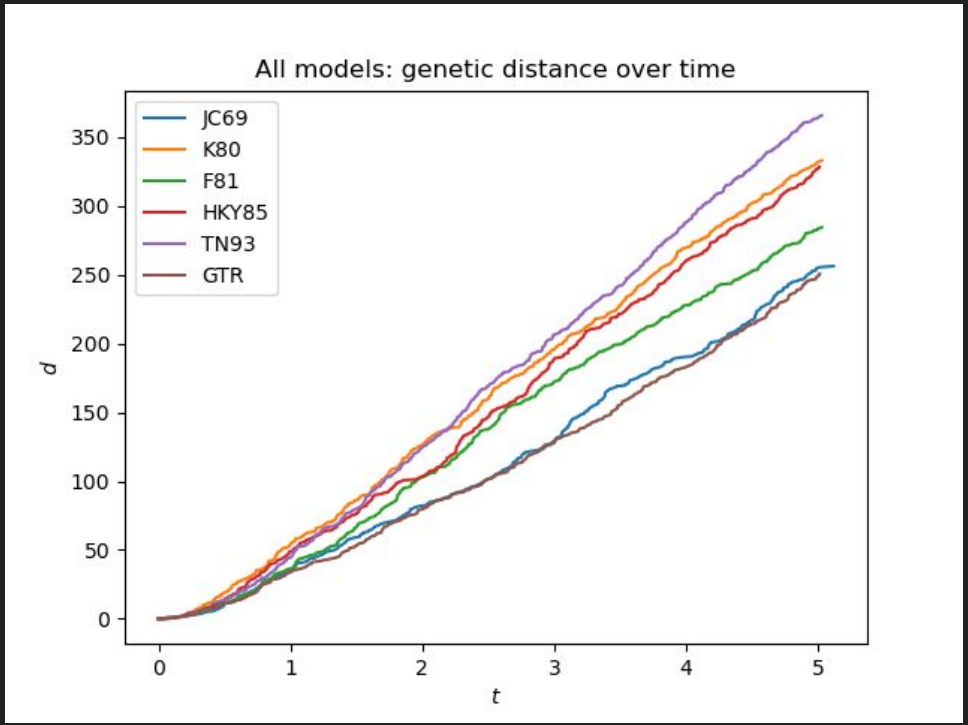
Results: relative base frequencies



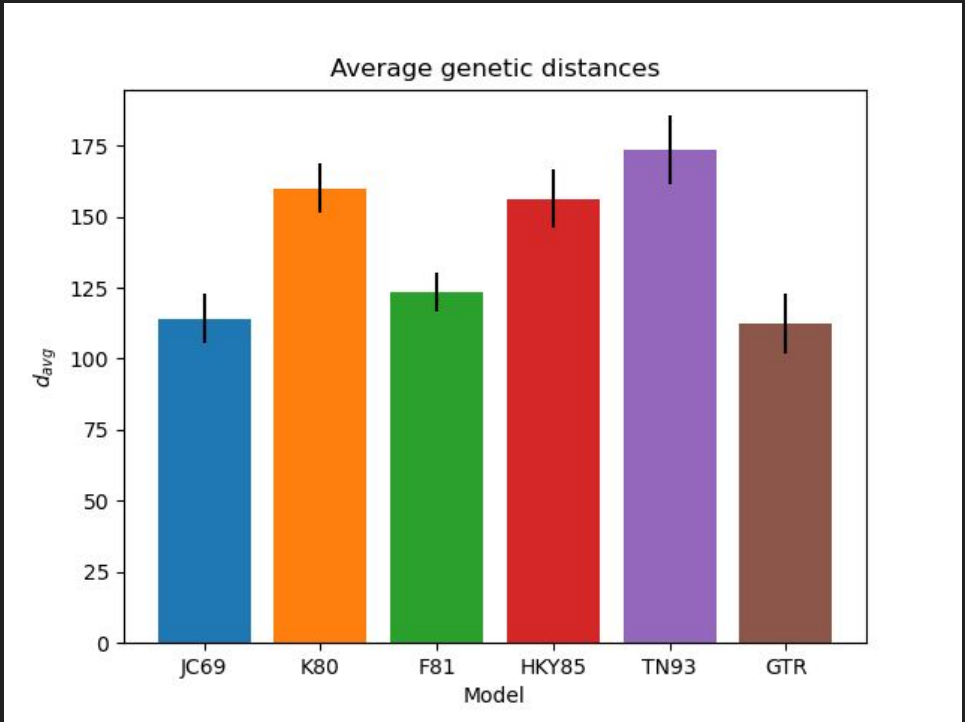
Results

- As all the models with non-equal base frequencies follow the same stationary distribution π, looking at the relative base frequencies over time can only tell us so much
- It is instead more useful to look at genetic distances

Results: genetic distances



Results: average genetic distances



Discussion

- All the models produce genetic distance estimates that are clearly and significantly different from one another
- JC69 and F81, the two models that assume equal transition/transversion rates, produce the lowest estimates of genetic distance
- K80, HKY85, and TN93, by contrast, assume that transitions are more likely than transversions, and they produce the highest estimates of *d*
    - This suggests that differences in transition and transversion rates play a more significant role in genetic drift than simple base frequencies
- Due to its generality, GTR includes no inherent biological assumptions, and so its results are of limited meaning on their own

Conclusions

- There is no one “best” substitution model – each one fills a different niche, and they all have different strengths and weaknesses
    - For instance, it is often much easier to collect data on relative base frequencies than transition or transversion rates; in such cases, F81 might be the most useful model
    - On the other hand, if those rates *are* known to a reasonable degree of accuracy, then HKY85 would likely be the most useful model
- Nonetheless, from a biological standpoint, there are certainly better and worse models
    - K80’s assumptions are sufficiently unreasonable that it is of little use in practical contexts, and doubly so for JC69
    - HKY85 balances biological inference and data flexibility fairly well, rendering it relatively versatile

References

- 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.
- N. Lanchier, *Stochastic Modeling* (2017), Springer, p. 101-124.
- P. Lemey et al., *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing* (2009), Cambridge University Press, p. 111-123.