Phylogenetic Inference Project: Group 2

Background/Introduction

In organisms, hereditary information in the form of DNA undergoes random mutations due to exposure to mutagens and also as the result of spontaneous processes such as hydrolysis. These alterations of the genetic code, formed by the four nitrogenous bases adenine, guanine, cytosine, and thymine, can occasionally induce changes in phenotype and also allow for the analysis of evolutionary relationships between extant members of a common ancestor.

By formulating a notion of phylogenetic distance, a measure of how evolutionarily distant two groups are, one can classify organisms into a phylogenetic tree that visualizes branching points in history and reveals the relative relationships between groups.

In our project, we apply the neighbor joining technique to 1) a simulated dataset created through the Jukes-Cantor and Kimura models of evolution and 2) a HIV sequence dataset obtained from Florida.

Modeling: Basic Assumptions

In order to simulate the evolutionary history of a random sequence, we take several assumptions axiomatically:

1. The mutation rate at each site is independent of that of any other site.
2. Mutations can be modeled as a Markov chain.
3. For the Juke-Cantor model, the types of possible mutations at a single point all have the same probability, α.
4. The mutation rate is constant across time.

It must be noted that these assumptions do not necessarily hold true.

Mutation rates differ due to a variety factors. Mutations in coding regions, especially those critical for basic function, are more likely to have negative effects than those in noncoding regions. Furthermore, mutations at the third codon position in DNA (individual letters are read in groups of three by protein/RNA complexes known as ribosomes) are less likely to cause catastrophic changes due to a redundancy caused by "wobble pairing." Even for those mutations that do result in changes to a protein's amino acid sequence, a mutation to a non-critical part of the structure (as opposed to a residue located within the active site) is less likely to be harmful and to a lesser degree.

Mutation probabilities are not independent of base type. Purines (adenine and guanine) and pyrimidines (cytosine and thymine), structurally more similar to each other than the other class, are more likely to mutate to a base type within their own group. Maintaining a purine-pyrimidine base pairing also helps hide the mutation from DNA repair machinery that can sense distortions in the DNA backbone (it should be noted that the Kimura 2-parameter model accounts for these points). In addition, certain mutation types are more likely occur due to differences in transition

state energies. Finally, we do not take into account base conversions to uracil, a nucleotide used in RNA in place of thymine, or hypoxanthine.

Mutation frequency also changes with time. Organisms may experience higher rates of mutation as factors with global reach (e.g. increased UV light exposure due to fluctuations in solar irradiance) naturally change or even as local environments shift.

Nevertheless, these assumptions allow the construction of two relatively powerful techniques known as the Jukes-Cantor and Kimura 2-Parameter models.

Jukes-Cantor Model

The first part of our project uses the Jukes-Cantor model to simulate evolutionary history. This model is known as a Markov chain, a sequence of events with transition probabilities between them dependent only upon the prior state. More rigorously stated a Markov chain exhibits the following property:

$$p(X_m = y \mid X_{m-1} = y_{m-1}, X_{m-2} = y_{m-2}, \ldots, X_1 = y_1) = p(X_m = y \mid X_{m-1} = y_{m-1}) \quad \forall y, ym\text{-}1, \ldots y1 \in T$$

Where all possible values of y_i form the state space of the model (T). The probability from one state to another is known as a *transition probability* and, as mentioned, is dependent only on the prior state of the system.

Using these probabilities, we can construct a *Markov matrix* where all entries are nonnegative real numbers and the columns sum to 1. The dimensions of such a matrix are N x N, where N = | T |. Each entry A_{ij} in the matrix represents the transition probability $j \rightarrow i$ or, more precisely, $p(X_m = i \mid X_{m-1} = j)$. Below is such a matrix for a two-state system:

$$\begin{bmatrix} p(X_2 = 1 \mid X_1 = 1) & p(X_2 = 1 \mid X_2 = 2) \\ p(X_2 = 2 \mid X_1 = 1) & p(X_2 = 2 \mid X_2 = 2) \end{bmatrix}$$

Right multiplication by an initial probability vector p_0, with components representing the probability of being in a given state, yields the probability distribution of possible states after one time step. Using our two-state system described earlier, the first component of the resultant vector Ap_0 is:

$$[\, p(X_2 = 1 \mid X_1 = 1) * p(X_1 = 1) + p(X_2 = 1 \mid X_2 = 2) * (pX_1 = 2\,)] = p(X_2 = 1)$$

Hence each component of our new vector describes the probability of the system being in the corresponding state after one time step.

In order to simulate evolution, we assign to each letter in our sequence an initial probability distribution vector p_0 that contains a 1 in the position corresponding to the identity of the letter

(since the probability of the nucleotide being itself is 1). We then perform matrix multiplication on each letter using our Markov matrix and randomly choose a new base identity weighted by the resultant probability distribution. We repeat this process *t* times in order to cover the necessary amount of time.

Returning to our example of Ap_0 earlier, simple arithmetic and probability rules will produce the same results for a second time step (i.e. $p(X_2 = i$ for component i of the resultant vector). Hence $A^t * p_0$ generates the probability distribution of bases after *t* time steps.

Since we assume all base transition probabilities are the same, our Markov matrix takes the following form:

$$\begin{bmatrix} 1-\alpha & \alpha/3 & \alpha/3 & \alpha/3 \\ \alpha/3 & 1-\alpha & \alpha/3 & \alpha/3 \\ \alpha/3 & \alpha/3 & 1-\alpha & \alpha/3 \\ \alpha/3 & \alpha/3 & \alpha/3 & 1-\alpha \end{bmatrix}$$

The i[th] column and row correspond to the same base type.

Note that α is the probability that a mutation occurs for a given base, but that it can also be interpreted as a rate of observable mutations per time step *t*.

Motivated by a desire to quickly calculate powers of this matrix, we can solve for an eigenvector basis and form a diagonal matrix. Solving the characteristic equation generated by:

$\det(M - \lambda I) = 0$ *where M is our Markov matrix*

[illegible], we obtain the following eigenvalue-eigenvector pairs:

$\lambda_1 = 1$ $\quad$ $\lambda_3 = \frac{1}{3}(3 - 4\alpha)$,
$V_1 = (1, 1, 1, 1)$ $\quad$ $V_3 = (1, -1, 1, -1)$

$\lambda_2 = \frac{1}{3}(3 - 4\alpha)$ $\quad$ $\lambda_4 = \frac{1}{3}(3 - 4\alpha)$,
$V_2 = (1, 1, -1, -1)$ $\quad$ $V_4 = (1, -1, -1, 1)$

We can extract the first column of M^t by multiplying by (1, 0, 0, 0) which is equivalently $(\frac{1}{4}\lambda_1 + \frac{1}{4}\lambda_2 + \frac{1}{4}\lambda_3 + \frac{1}{4}\lambda_4)$ when written in our new eigenvector basis. Since we now have an eigenvector representation for this vector and consequently our first column, we know:

$$M^t \lambda_1 = \lambda_1{}^t \lambda_1$$

Performing this procedure for each column, we obtain the following matrix representation of M^t:

$$\begin{bmatrix} \frac{1}{4} + \frac{3}{4}(1 - \frac{4}{3}\alpha)^t & \frac{1}{4} - \frac{1}{4}(1 - \frac{4}{3}\alpha)^t & \frac{1}{4} - \frac{1}{4}(1 - \frac{4}{3}\alpha)^t & \frac{1}{4} - \frac{1}{4}(1 - \frac{4}{3}\alpha)^t \\ \frac{1}{4} - \frac{1}{4}(1 - \frac{4}{3}\alpha)^t & \frac{1}{4} + \frac{3}{4}(1 - \frac{4}{3}\alpha)^t & \frac{1}{4} - \frac{1}{4}(1 - \frac{4}{3}\alpha)^t & \frac{1}{4} - \frac{1}{4}(1 - \frac{4}{3}\alpha)^t \\ \frac{1}{4} - \frac{1}{4}(1 - \frac{4}{3}\alpha)^t & \frac{1}{4} - \frac{1}{4}(1 - \frac{4}{3}\alpha)^t & \frac{1}{4} + \frac{3}{4}(1 - \frac{4}{3}\alpha)^t & \frac{1}{4} - \frac{1}{4}(1 - \frac{4}{3}\alpha)^t \\ \frac{1}{4} - \frac{1}{4}(1 - \frac{4}{3}\alpha)^t & \frac{1}{4} - \frac{1}{4}(1 - \frac{4}{3}\alpha)^t & \frac{1}{4} - \frac{1}{4}(1 - \frac{4}{3}\alpha)^t & \frac{1}{4} + \frac{3}{4}(1 - \frac{4}{3}\alpha)^t \end{bmatrix}$$

Hence the expected number of mutated sites after a time *t* is $□(□) = \frac{3}{4} - \frac{3}{4}(1 - \frac{4}{3}□)^{□}$

In order to estimate p, we can examine the fraction of changed sites between two sequences. Plugging in *p* as a constant and applying a Taylor series approximation that $□□ (1 + □) \approx □$ for small values of X, we obtain the following formula:

$$□□ = -\frac{3}{4}□□ (1 - \frac{4}{3}□) \approx □$$

This $d(S_i, S_j)$ is the *Jukes-Cantor distance* between the compared sequences and provides a rough estimate of "evolutionary distance" between the two. Assuming the molecular clock assumptions holds true, this distance value has an additive property such that:

$□□ □_3 = □_1 + □_2$ □□□□ $□□_3 = □□_1 + □□_2$

$$\rightarrow □(□_1, □_3) \approx □(□_1, □_2) + □(□_2, □_3) \qquad \text{approximating } □□_□$$

Kimura Model

The Kimura 2-Parameter model is a more powerful version of the above approach. While it suffers from most of the drawbacks iterated earlier, the Kimura model uses two variables in order to account for the difference in transversions and transitions rates.

Mutation rates β and γ are used for transitions and transversions respectively. Hence the transition matrix takes the form of:

$$\begin{pmatrix} 1-□-2□ & □ & □ & □ \\ □ & 1-□-2□ & □ & □ \\ □ & □ & 1-□-2□ & □ \\ □ & □ & □ & 1-□-2□ \end{pmatrix}$$

It is important to note that the ordering of bases along the rows and columns above is A,G,C, and T. This ordering ensures that matrix entries correspond correctly to a transition or transversion (if the ordering was different, so would the location of the □′□ □□□ □′□.

In order to solve for the form of the Kimura matrix raised to a given power *t*, we perform the same procedure used earlier for the Jukes-Cantor model. Solving for eigenvectors and then pulling out matrix values, we find that:

$$□' = \frac{1}{4} + \frac{1}{4}(1 - 2□ - 2□)^{□} - \frac{1}{4}(1 - 2□ - 2□)^{□} - \frac{1}{4}(1 - 2□ - 2□)^{□}$$

$$□' = \frac{1}{4} - \frac{1}{4}(1 - 2□ - 2□)^{□} + \frac{1}{4}(1 - 2□ - 2□)^{□} - \frac{1}{4}(1 - 2□ - 2□)^{□}$$

$$□' = \frac{1}{4} - \frac{1}{4}(1 - 2□ - 2□)^{□} - \frac{1}{4}(1 - 2□ - 2□)^{□} + \frac{1}{4}(1 - 2□ - 2□)^{□}$$

Where β', γ', and δ' are the rate values associated with the exponentiated Kimura matrix, M^t (i.e. $M(\beta, \gamma, \delta)^t = M(\beta', \gamma', \delta')$ for some *t*). Through substitution and algebraic manipulation, it can be shown that:

$$1 - 2\square' - 2\square' = (1 - 2\square - 2\square)^{\square}$$

$$1 - 2\square' - 2\square' = (1 - 2\square - 2\square)^{\square}$$

$$1 - 2\square' - 2\square' = (1 - 2\square - 2\square)^{\square}$$

Using the Taylor approximation $\square\square\ (1 + \square) \approx \square$ used earlier, we find that:

$$\square\square\ (1 - 2\square' - 2\square') + \square\square\ (1 - 2\square' - 2\square') + \square\square\ (1 - 2\square' - 2\square') \approx -4\square(\square + \square + \square)$$

Solving for $(\square + \square + \square)$, the substitution rate, we obtain the following formula for the 3-parameter condition:

$$\square_{\square 3} = -\frac{1}{4}(\square\square(1 - 2\square - 2\square) + \square\square(1 - 2\square - 2\square) + \square\square(1 - 2\square - 2\square))$$

The 2-parameter model is a single case of this formula where γ = δ. In order to estimate the substitution rates, we can evaluate the number of transversions and transitions between two sequences (let these fractions be P_1 and P_2) as rough probability estimates. We then obtain the following formula:

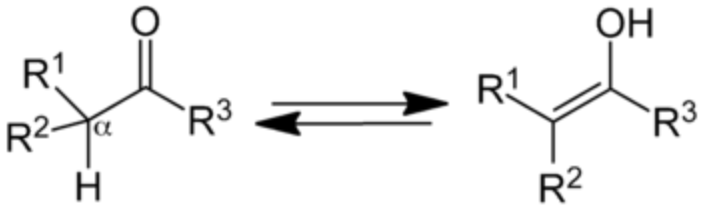
$$\square_{\square 2} = -\frac{1}{2}\square\square\ (1 - 2\square_1 - \square_2) - \frac{1}{4}\square\square\ (1 - 2\square_2)$$

Biological Motivation for the Kimura Model

The impetus for the development of the Kimura model was the observation that base substitution rates were distinct for specific nucleotides. In other words, the mutation rate for DNA is not independent of base type.

For the Kimura 2-parameter model that we implemented, the specific biological emphasis is that transition and transversion rates are not equal. In fact, evidence shows that transitions are twice as common as transversions. We can put forward two hypotheses for this difference: 1) transitions are not corrected as easily as transversions or 2) transitions are chemically easier.

Transitions are enabled through a process called tautomerization in which a double bond is traded for a proton.



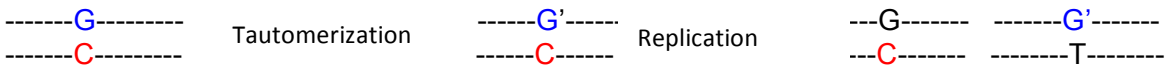
The above molecules are examples of a keto-enol tautomerism (on the left and right respectively). This interconversion between the two can be catalyzed by acidic and basic environments but both processes yield the same pattern of double bond for hydrogen exchange.

Note that this exchange flips the role of the oxygen in hydrogen bonding, a form of attraction where hydrogen bonded to an electronegative atom feels an attraction toward another electronegative atom (this is because electrons are unevenly shared between the hydrogen and its neighbor, resulting in an accumulation of positive charge on the hydrogen). The hydrogen is referred to as a hydrogen bond donor and the electronegative atom it feels an attractive force toward is a hydrogen bond acceptor.

In neutral environments, the keto form is generally favored as it is more energetically stable. The hydrogen in the hydroxyl group (OH) is far more acidic than the hydrogen attached to the α atom so the chemical equilibrium shifts in favor of the keto form. The human body maintains a pH close to neutral in order to avoid damaging itself and the nucleus is similarly regulated. One measurement of a human cell line puts the pH of the nucleus at 7.35[1], which is quite close to a neutral environment, and hence the keto form of nucleotides is favored.

DNA base pairing is dictated by hydrogen bonding between nucleotides. In other words, each nucleotide pairing (A/T and G/C) is due to complementary bonding patterns on molecules. When a molecule undergoes a tautomerism between its enol and ketone forms, the role of the oxygen as either a donor or acceptor is flipped. Similarly, the roles are reversed for the α atom if it is sufficiently electronegative (e.g. if α were a nitrogen atom). Hence base pairs exhibit swapped bonding patterns when they undergo a tautomerization (specifically between amino and imino forms, the oxygen is now a nitrogen in this tautomerism) which in turn dictates base pairing preferences.

Therefore a nucleotide that undergoes a tautomerization will present a different hydrogen bonding pattern that causes it to pair with the incorrect base pair of opposite type (a purine will pair with an incorrect pyrimidine and vice versa). After one round of DNA replication, this will result in a point mutation at the site. In a diagrammatic fashion:



Since tautomerizations are not complex chemica ly involve a redistribution of electrons (aside from the movement of a proton) itaneous and occur relatively

[1] http://bionumbers.hms.harvard.edu//bionumber.aspx

frequently. One estimate[2] places the fraction of bases in the imino form at 10^{-4} which, given an estimated human mutation rate of approximately $1 * 10^{-8}$ per site[3], is quite high.

The activation energy for base tautomerizations is estimated to be roughly 20-25 kcal/mol for double stranded DNA[4], which corresponds to roughly 80-100 kJ/mol. By contrast, the process of a transversion by chemical modification of a base, say a purine to pyrimidine, would involve breaking a carbon-carbon bond, which would require 346 kJ/mol[5] when not factoring in the extra stabilization in the molecule due to a phenomenon called aromaticity.

The advantage of the tautomerization is that it involves breaking one bond in the double bond, not a single bond (save for the bonding between the hydrogen and another molecule although this bond dissociation energy is quite low). The specific bond broken is called a π bond and is weaker than the σ bond that is found in single bonds due to the orientation of electron orbitals participating in the π bond having less overlap than those in the σ bond.

Hence transversions are almost always the product of an incorrect base pairing. For instance, a purine is matched incorrectly with another purine. This in turn results in the placement of a pyrimidine in one of the new DNA molecules in the position of the original purine. Yet for this situation to arise, the mismatch must be able to hide from DNA repair machinery long enough for replication to occur.

Mismatch errors in DNA replication are quite rare due to the fidelity of DNA polymerases. Polymerases are highly accurate to begin with but also contain a secondary measure against mutation called a "proofreading" ability, conferred by an exonuclease activity on the enzyme that allows it to cut out incorrectly paired bases as replication is occurring. In total, these two factors generate an error rate of roughly 4.4×10^{-5} for human polymerase ε (the main error correction polymerase for polymerase δ, which does most of the replication heavy work).[6] Interestingly, both pol δ and pol ε display the greatest error bias toward introducing specific transversion types.[7] When averaged over categories however, the overall error rate for transitions is higher than that of transversions[8], indicating that eukaryotic DNA polymerases more easily fix transversions than transitions.

DNA repair pathways themselves exhibit biases toward easy detection of purine-purine and pyrimidine-pyrimidine mispairings due to the noticeable strain and resulting distortions of the molecule. These distortions can act as intracellular signals for repair machinery activation. In nucleotide excision repair (NER), a repair pathway used for DNA lesions such as thymine dimers from UV exposure, a protein called HR23B forms a complex with another protein called XPC in order to detect DNA damage. The XPC-hHR23B complex, although used for larger lesions resulting from base damage, is highly adept at detecting even local distortions from base

[2] http://www.ncbi.nlm.nih.gov/books/NBK22525/

[3] http://www.nature.com/ng/journal/v43/n7/full/ng.862.html

[4] http://www.ncbi.nlm.nih.gov/pubmed/16955739

[5] http://www.ncbi.nlm.nih.gov/pubmed/16955739

[6] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3061053/

[7] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3061053/figure/F3/

[8] http://genome.cshlp.org/content/24/11/1751.full

mispairing.[9] Although these base mispairings are not individually corrected *in vivo*, they do contribute to XPC-hHR23B response to nearby damage. In other words, XPC-hHR23B is far more likely to correct lesions that contain pyrimidine-pyrimidine or purine-purine mispairings, due to the increased distortion of the backbone, than it is to respond to lesions containing purine-pyrimidine mispairings.[10] This differential response has the effect of 1) driving down the probability of transversions and 2) reducing even further the probability of a transversion occurring due to chemical modification of the base, as many forms of energy intense enough to break the necessary bonds would likely damage the local region.

In DNA mismatch repair (MMR), a MSH2-MSH6 protein complex binds to DNA through DNA binding domains I and IV respectively along the minor groove (a non-specific groove that does not have unique hydrogen bond patterning for nucleotides) and scans along the molecule looking for signs of mismatch, specifically bent DNA.[11] MSH6 alone engages in binding with the mismatch site through an evolutionarily conserved motif in domain IV that interacts with the mispaired base, a task more easily accomplished by a distorted and bent local structure.[12] This observation is correlated by site-directed mutagenesis of the residues within the motif that leads to lowered mismatch detection and increased mutation rates in yeast.[13]

Applicability of Kimura Model to a HIV Dataset

The contributing role of cellular repair pathways in creating a difference in transition and transversion rates begs the questions of whether or not the Kimura model is equally applicable to other scenarios, specifically our dataset obtained from HIV.

HIV exhibits a very high transition:transversion bias (κ). For HIV-1, the major virus type largely responsible for the AIDS pandemic and also the isolated type from our Florida dataset, estimated κ values range from 2.1 to 13.0 depending upon the codon position (as mentioned earlier, the third codon position displays higher mutation rates due to redundancies in translation)[14]. Our aligned sequences were obtained from the V3 encoding region of the *env* gene, which participates in production of a membrane protein that mediates HIV host invasion. Hence the appropriate κ value ranges from 3.1 to 6.6 using the previously cited source. Regardless of codon position, these values show a clear difference in transition and transversion rates that is even more pronounced than that in humans.

This bias agrees with evidence regarding HIV-1 reverse transcriptase's fidelity. HIV is a retrovirus since its genetic information is stored in RNA. In order to incorporate itself into the host genome, HIV must transcribe its RNA into a DNA format that can then be inserted. This process is mediated by a protein called reverse transcriptase. For HIV-1, reverse transcriptase exhibits a high mutation rate and a bias toward transition mutations.[15]

---

[9] http://www.nature.com/nrm/journal/v15/n7/full/nrm3822.html
[10] http://www.nature.com/nrm/journal/v15/n7/full/nrm3822.html
[11] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3325129/#bib1
[12] Ibid.
[13] Ibid.
[14] http://www.biochemsoctrans.org/content/ppbiost/28/2/275.full.pdf
[15] http://onlinelibrary.wiley.com/doi/10.1016/0014-5793%2892%2980988-S/pdf

Neighbor Joining Algorithm

For the algorithm, 3-point formulas are used to solve systems of linear equations representing distances in a base case scenario. This simple situation is used to construct the larger tree.

Consider a set of nodes {A, B, C}, joined at a central node V, with values d_{ij} representing distances between them (i, j unique). Furthermore, let {X, Y, Z} be the distances between V and A, B, and C respectively.

Then:

$$\begin{aligned} d_{AB} &= X + Y & X &= (d_{AC} + d_{AB} - d_{BC})/2 \\ d_{AC} &= X + Z & \Rightarrow\ Y &= (d_{AC} + d_{BC} - d_{AB})/2 \\ d_{BC} &= Y + Z & Z &= (d_{AB} + d_{BC} - d_{AC})/2 \end{aligned}$$

The formulas on the left are collectively referred to as the 3-point formulas.

For our neighbor criterion, we use the fact that if S_i and S_j are neighbors then:

Let R_i denote $\sum_{k \neq i} d(i,k)$ for fixed i and $i \neq k$

$$\begin{aligned} d(i,j) &< d(i,k) + d(j,k) \text{ for fixed } i,j,k \text{ and } \forall k \\ &\rightarrow (n-2)d(i,j) - R_i - R_j \\ &< (n-2)d(i,k) - R_i - R_k \text{ for all } k \neq i \end{aligned}$$

For simplicity's sake, let the derived criterion be labeled M_{ij}. It can also be shown that if M_{ij} is a global minimum then S_i and S_j are neighbors. A proof of this criterion is given in a paper by Saitou & Nei[16]. The algorithm is constructed as follows:

1) Compute M_{ij} for all sequences i, j unique
2) Pick M_{nm} a global minimum (the sequences are guaranteed neighbors)
3) Join S_n, S_m
4) Compute distances from S_n, S_m to their parent node V as follows:
    a. Collapse all other sequences into a temporary node, G
    b. Compute $d(k,n)$ and $d(k,m)$ for average from S_n, S_m to all members of G
    c. Compute $d(k,n)$, $d(k,m)$ via 3-point formulas (remarked upon earlier).
5) $\forall k \in G$, compute $d(k,V)$
6) Drop S_i, S_j from consideration. V acts in place of these sequences now.

[16] http://mbe.oxfordjournals.org/content/4/4/406.long

7) Repeat steps 1-6 so long as $\square \geq 4$.
    a. If N-1 = 3, use 3-point formulas to finish.

This algorithm guarantees a joining of true neighbors since, at each step, the selection of M_{ij} a global minimum ensures that S_i and S_j are neighbors. Hence the resulting tree has fully labeled distances and known connectivity.

Results

In order to test our neighbor joining algorithm, we tested the procedure on two datasets: trees generated according to the Jukes-Cantor and Kimura models of evolution and a HIV dataset.

For our Kimura model of HIV, we used a κ obtained by averaging the transition:transversion bias across the three codon positions as our model did not account for differential substitution rates at each site. Using a per site mutation rate of 1.62×10^{-2} per year,[17] we solved for β and γ values of 0.013 and 0.0032 respectively. For Jukes-Cantor, we used

[17] http://www.ncbi.nlm.nih.gov/pubmed/11264414