A gentle Introduction to Probabilistic Evolutionary Models

Tal Pupko, Itay Mayrose

Chapter 1.1 A gentle Introduction to Probabilistic Evolutionary Models

Tal Pupko
School of Molecular Cell Biology & Biotechnology, Tel Aviv University, Ramat Aviv, Tel-Aviv 69978, Israel
TalP@tauex.tau.ac.il

Itay Mayrose
School of Plant Sciences and Food security, Tel Aviv University, Ramat Aviv, Tel-Aviv 69978, Israel
itaymay@post.tau.ac.il

Abstract

A large body of research is dedicated to model sequence evolutionary dynamics. The evolutionary process may vary within groups of genes, among sites within a gene, between populations and among diverged species. Evolutionary models aiming to describe these dynamics must account for base pair substitutions as well as insertion and deletion (indel) events. Here, we explain the fundamental of continuous time Markov models used to describe sequence evolution. We begin by describing discrete Markov models, and slowly progress towards more realistic and more computationally complicated continuous time Markov models. Among other topics, we discuss nucleotide, amino acid, and codon models, among site rate variation, model reversibility, stationary distributions, rate matrix normalization, mixture models, indel models, and models of gene family evolution. Understanding the concepts presented here is vital for various phylogenomics analyses such as the inference of positive selection, alignment and phylogeny reconstruction, ancestral sequence reconstruction, and molecular dating.

How to cite: Tal Pupko and Itay Mayrose (2020). A gentle Introduction to Probabilistic Evolutionary Models. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 1.1, pp. 1.1:1–1.1:21. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

1 General outline – the great importance of evolutionary models

The past decade has witnessed a revolution in evolutionary biology, driven by advances in high-throughput sequencing, functional genomics, and computational technologies. The genomic revolution opened up possibilities for rapidly generating large-scale sequencing data from non-model organisms at a reasonable cost. However, while new methodologies have been devised to handle and assemble these data (Ekblom and Galindo, 2011), methodological advances to convert these data into meaningful biological knowledge are still lagging behind. Indeed, one of the main challenges for the decade ahead will be to unravel the connection between genomic changes and the diversity of phenotypes seen in nature and to decipher the evolutionary forces responsible for this diversification. Such research efforts will lead to a more explanatory theory of evolution, with implications for all branches of the life sciences, from agriculture to ecology to medicine. Discovering these linkages is a difficult task that requires novel biologically-inspired computational methodologies that will identify candidate loci under selection and will correlate these loci to phenotypic differences. In the post-genomic era, probabilistic models of molecular evolution are the powerhouse of

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 1.1; pp. 1.1:1–1.1:21

A book completely handled by researchers.
No publisher has been paid.

data analyses. Such models form the backbone of various bioinformatics applications such as phylogenetic reconstruction (see Chapters 1.2 [Stamatakis and Kozlov 2020] and 1.4 [Lartillot 2020]), sequence alignment (see Chapter 2.2 [Ranwez and Chantret 2020]), molecular dating (see Chapter 5.1 [Pett and Heath 2020]), and functional sites prediction (Anisimova et al., 2013; Yang, 2014). These models are an attempt to represent the stochastic nature of evolution, built within a robust statistical framework of inference. Constructing evolutionary models enables us to mathematically describe various biological phenomena, estimate relevant parameters such as the strength of selection (see Chapter 4.5 [Lowe and Rodrigue 2020]) and rate of evolutionary events (see Chapter 4.4 [Bromham 2020]), and test different hypotheses, e.g., to determine which of several alternative evolutionary pathways is more likely. A main challenge in designing a model is to capture the key elements of the biological process at hand without over-parameterizing the model, which will render it inadequate. In molecular phylogenetics, Markov models, a specific class of stochastic models, are intensively used to analyze sequence data and to quantify the evolutionary dynamics of genes and genomes. In this chapter we introduce the theoretical foundations of these probabilistic models, starting with the simplest ones and progressing towards richer and more realistic models.

2 Discrete Time Markov chains

It is often helpful to start thinking of probabilistic evolutionary models in terms of simulations, i.e., to describe how one would mimic the evolutionary process using a computer. Specifically, we are interested in describing how sequences change through time and which parameters control this evolutionary process. For simplicity, we describe the evolution of only a single sequence site; assuming that all sites evolve independently, this process can be repeated N times to generate a sequence of length N. We start by drawing the identity of this position in the initial generation – the ancestor. Assuming that the probabilities of all nucleotides are equal, this is equivalent to rolling a dice with four possible outcomes (A, C, G, or T). Next, we would like to decide on the fate of this position in the next generation. It can either mutate or not. If it mutates, it can change to each one of the other nucleotides with equal probabilities. We can put this model into a matrix form:

$$M_{ij} = \begin{pmatrix} 1-3\epsilon & \epsilon & \epsilon & \epsilon \\ \epsilon & 1-3\epsilon & \epsilon & \epsilon \\ \epsilon & \epsilon & 1-3\epsilon & \epsilon \\ \epsilon & \epsilon & \epsilon & 1-3\epsilon \end{pmatrix} \tag{1}$$

This matrix represents a simple discrete-time Markov model with four states. The model is called discrete, because t, the time that corresponds to the number of generations, can only take discrete values $(0, 1, 2, \cdots)$. M_{ij} is the probability to start with nucleotide i and end in nucleotide j after one generation. Because each row includes all possible events, the sum of each row must equal one. There is one free parameter in this model, $0 \leq \epsilon \leq \frac{1}{3}$, which dictates the rate of evolution. If we run the simulations for many generations, a fewer mutations are expected for low values of ϵ compared to larger values. In fact, in each generation there are two possibilities: either a mutation has occurred, with probability 3ϵ or the position remained the same, with probability $1-3\epsilon$. Thus, the expectation of the number of changes for a single generation per position is 3ϵ.

Suppose now that we use this model to simulate 1,000 generations and that we are interested in calculating the probability that we end up with the nucleotide C in generation

1,000, given that we started with nucleotide A at generation zero. A naïve approach would be to simulate the evolution process for 1,000 generations according to the matrix M and record the final nucleotide (note that we always start the simulation with an A at generation zero). This represents one possible evolutionary outcome. In order to estimate the desired probability we repeat this process multiple times (say 1,000,000). The frequency of the outcome "ending with C" serves as a good estimation for the desired probability.

However, we can use Markov chain theory to analytically compute this probability. We first compute (rather than simulate) the probability of the event starting with A and after two generations ending with C. We denote this probability $P_{A,C}(2)$. The course of this event can be described as "A will change to some unknown character X after one generation" ($P_{A,X}(1)$) and then X will change to C in one generation ($P_{X,C}(1)$). By the law of total probability, we can express this probability by summing over all possible values of X:

$$P_{A,C}(2) = P_{A,A}(1)P_{A,C}(1) + P_{A,C}(1)P_{C,C}(1) + P_{A,G}(1)P_{G,C}(1) + P_{A,T}(1)P_{T,C}(1) \quad (2)$$

We note that we implicitly assume that the probability to change from A in generation 0 to X in generation 1, is the same as the probability to change from A in generation 1 to X in generation 2. More formally, we assumed that the Markov chain is time homogenous. In probabilistic terms, let $X(k)$ represent the nucleotide after k generations. Matrix M represents the transition probability from generation g to generation $g+1$ for every g and using this notation, we can express M as

$$M_{ij} = P(X(k+1) = j | X(k) = i) = P_{i,j}(1) \quad (3)$$

Note that this equality directly stems from the time homogeneity assumption. In matrix notations, we will hence consider M as $P(1)$, i.e., the Markov matrix representing the probabilities of changes after one generation. Note that we also implicitly assume that the transition probability from A in generation 1000 to C in generation 1001 is only dictated by M, and does not depend on the history of events that have occurred in generations 0 to 1000. This is an important property of the Markovian process, which is called memorylessness.

Returning to Eq. 2, one could notice that this computation is identical to the dot product of row A (first row) and column C (second column) in M. Let us denote by $P(2)$ the matrix of transition probabilities between each pair of nucleotides after two generations (and in general, $P(k)$ the matrix after k generations). The computation of each entry is performed in an identical way to Eq. 2:

$$P(2) = \begin{pmatrix} P_{AA}(2) & P_{AC}(2) & P_{AG}(2) & P_{AT}(2) \\ P_{CA}(2) & P_{CC}(2) & P_{CG}(2) & P_{CT}(2) \\ P_{GA}(2) & P_{GC}(2) & P_{GG}(2) & P_{GT}(2) \\ P_{TA}(2) & P_{TC}(2) & P_{TG}(2) & P_{TT}(2) \end{pmatrix} \quad (4)$$

From the above consideration, $P(2) = P(1)^2 = M^2$ and in general $P(k) = P(1)^k$ and $P(n+m) = P(n)P(m)$. The importance of the last equality will become clearer in the next section. Thus, in a discrete time process, by knowing the matrix M, the transition probabilities $P(k)$ can be derived for all possible k values. Hence, $M = P(1)$ is sometimes called the generator matrix for discrete-time Markov chains. Putting it all together, we obtain:

$$P_{ij}(k) = P(X(k) = j | X(0) = i) = \left[M^k\right]_{i,j} \quad (5)$$

3 Continuous Time Markov chains

A natural generalization of the above discrete time Markov chain is to replace the number of generations with a continuous parameter t that measures time. Indeed, such a process in which t can have any value in the interval $[0, \infty)$ is termed a continuous time Markov chain. Similar to discrete processes, $X(t)$ represents the state of a specific site at time t. Here, we would like to compute not only $P(1)$ and $P(2)$ but also, say, $P(2.335)$. To understand how such a matrix can be computed, we recall that for the discrete Markov process we had $P(t_1+t_2) = P(t_1)P(t_2) = P(t_2)P(t_1)$. This is also true for continuous time Markov chains, i.e., it is true for every two time points t_1 and t_2. From this, we obtain:

$$P(t_1+t_2) - P(t_1) = P(t_1)P(t_2) - P(t_1) = P(t_1)(P(t_2) - I) \tag{6}$$

where I is the identity matrix. Note that $P(0) = I$ since $P_{ii}(0) = P(X(0) = i|X(0) = i) = 1$ and similarly, $P_{ij}(0) = 0$ for all $i \neq j$. The above equations imply that for any $t_2 \neq 0$:

$$\frac{P(t_1+t_2) - P(t_1)}{t_2} = \frac{P(t_1)(P(t_2) - I)}{t_2} \tag{7}$$

We can then write t instead of t_1 and Δt instead of t_2. When Δt approaches zero, we obtain:

$$\lim_{\Delta t \to 0} \frac{P(t+\Delta t) - P(t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{P(t)(P(\Delta t) - P(0))}{\Delta t} = \lim_{\Delta t \to 0} P(t)\frac{P(\Delta t) - P(0)}{\Delta t} = P(t)P'(0) \tag{8}$$

where the last equality stems from the definition of the derivative at time zero. The left side of the above equation is the definition of the derivative at time t, and thus:

$$P'(t) = P(t)P'(0) \tag{9}$$

We note that in the equation above we made use of the concept of matrix derivative: if for a matrix $P(t)$ each element of a matrix is a function of t, then in $P'(t)$ each element is the derivative of the corresponding entry of $P(t)$. $P'(0)$ is a fixed matrix called the instantaneous rate matrix, and is often denoted as $Q \equiv P'(0)$.

As will be detailed below, given Q we can compute the transition probability matrix $P(t)$ for any time interval t, and thus Q is often termed the generator matrix of the continuous Markov process. The elements of each row in $P(t)$ sum to 1:

$$P_{i,1}(t) + P_{i,2}(t) + P_{i,3}(t) + \cdots + P_{i,n}(t) = 1 \tag{10}$$

where n is the number of possible states (four for nucleotides). Deriving both sides of this equation with respect to t, we obtain:

$$P'_{i,1}(t) + P'_{i,2}(t) + P'_{i,3}(t) + \cdots + P'_{i,n}(t) = 0 \tag{11}$$

Specifically, for $t = 0$, we obtain:

$$P'_{i,1}(0) + P'_{i,2}(0) + P'_{i,3}(0) + \cdots + P'_{i,n}(0) = 0 \tag{12}$$

Since $Q = P'(0)$ we obtain

$$Q_{i,1} + Q_{i,2} + Q_{i,3} + \cdots + Q_{i,n} = 0 \tag{13}$$

Thus, the elements in each row of Q sum to zero. For $i \neq j$, $P_{i,j}(t)$ is an increasing function of t: it is zero for $t = 0$ and non-negative for $t > 0$. Thus, each non-diagonal element, $Q_{i,j} = P'_{i,j}(0)$ is non-negative and represents the instantaneous transition rate from state i to state j, while each diagonal element, $Q_{i,i}$ is negative and equals to the negative sum of all other elements in that row. $Q_{i,i}$ represents the total instantaneous transition rate away from state i.

Eq. 9 above is in fact a differential equation in a matrix form:

$$\frac{dP(t)}{dt} = P(t)Q \tag{14}$$

The solution to this equation, subjected to the boundary condition $P(0) = I$ is

$$P(t) = e^{Qt} = I + Qt + \frac{(Qt)^2}{2!} + \frac{(Qt)^3}{3!} + \cdots \tag{15}$$

We will note that this power series always converges. Computing matrix exponentials is a well-studied topic in numerical analysis and will not be discussed here. However, for certain types of simple models, a closed-form solution for $P(t)$ can be obtained directly. Below we derive the $P(t)$ matrix for the simplest nucleotide model, the Jukes and Cantor (JC) model.

4 The Jukes and Cantor model

The simplest continuous time Markov model for nucleotides assumes that the transition probabilities between each two different nucleotides is the same: $f(t)$. In a matrix form, $P(t)$ is:

$$P(t) = \begin{pmatrix} 1-3f(t) & f(t) & f(t) & f(t) \\ f(t) & 1-3f(t) & f(t) & f(t) \\ f(t) & f(t) & 1-3f(t) & f(t) \\ f(t) & f(t) & f(t) & 1-3f(t) \end{pmatrix} \tag{16}$$

By the definition of Q for this model we obtain:

$$Q = P'(0) = \begin{pmatrix} -3f'(0) & f'(0) & f'(0) & f'(0) \\ f'(0) & -3f'(0) & f'(0) & f'(0) \\ f'(0) & f'(0) & -3f'(0) & f'(0) \\ f'(0) & f'(0) & f'(0) & -3f'(0) \end{pmatrix} \tag{17}$$

If we denote $f'(0)$ as α we obtain:

$$Q = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix} \tag{18}$$

The stochastic process generated by this matrix (and assuming equal frequencies for all nucleotides) is called the Jukes and Cantor model (Jukes and Cantor, 1969). Applying Eq. 14 to the JC model, we obtain:

$$\frac{d}{dt}\begin{pmatrix} 1-3f(t) & f(t) & f(t) & f(t) \\ f(t) & 1-3f(t) & f(t) & f(t) \\ f(t) & f(t) & 1-3f(t) & f(t) \\ f(t) & f(t) & f(t) & 1-3f(t) \end{pmatrix} =$$

$$= \begin{pmatrix} 1-3f(t) & f(t) & f(t) & f(t) \\ f(t) & 1-3f(t) & f(t) & f(t) \\ f(t) & f(t) & 1-3f(t) & f(t) \\ f(t) & f(t) & f(t) & 1-3f(t) \end{pmatrix} \cdot \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix} \quad (19)$$

If we focus only on the element in the first row and second column of the left side of this equation (and marking all the other elements as "*") and use the definition of matrix multiplication we obtain:

$$\begin{pmatrix} * & f'(t) & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{pmatrix} = \begin{pmatrix} 1-3f(t) & f(t) & f(t) & f(t) \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{pmatrix} \cdot \begin{pmatrix} * & \alpha & * & * \\ * & -3\alpha & * & * \\ * & \alpha & * & * \\ * & \alpha & * & * \end{pmatrix} \quad (20)$$

which yields:

$$f'(t) = (1-3f(t))\alpha + f(t)(-3\alpha) + f(t)\alpha + f(t)\alpha \quad (21)$$

Which, after simplification, results in the following regular differential equation:

$$\frac{df(t)}{dt} = \alpha(1-4f(t)) \quad (22)$$

This equation can informally be written as:

$$\frac{df(t)}{1-4f(t)} = \alpha 4dt \quad (23)$$

Integrating both sides yields:

$$\frac{ln(1-4f(t)}{-4}) = \alpha t + C \quad (24)$$

Isolating the term $f(t)$, we obtain:

$$f(t) = \frac{1}{4} - \frac{e^{-4\alpha t - 4C}}{4} \quad (25)$$

Recall that $f(t)$ is in fact $P_{i,j}(t)$ for $i \neq j$. Given that $f(0) = 0$ (because $P(0) = I$), we obtain that $C = 0$ and thus

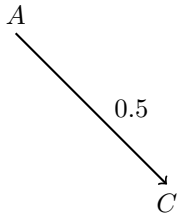
$$f(t) = \frac{1}{4} - \frac{e^{-4\alpha t}}{4} \quad (26)$$

Similarly, since $P_{i,i}(t)$ is $1 - 3f(t)$ we obtain:

$$P_{i,j}(t) = \begin{cases} \frac{1}{4} - \frac{e^{-4\alpha t}}{4} & \text{if } i \neq j, \\ \frac{1}{4} + \frac{3e^{-4\alpha t}}{4} & \text{if } i = j \end{cases} \quad (27)$$

5 Likelihood functions and matrix normalization

In the JC model, given the α and t parameters, we can simulate an initial sequence and let it evolve, *in silico*. We can also write explicit probabilistic expressions for evolutionary scenarios. We start with a nucleotide sequence composed of a single position, in which the character A is observed. The probability of starting with A according to the JC model is $P(A) = 0.25$ since the JC model assumes equal probabilities for all nucleotides. Of note, this implies that out of all simulations starting with a sequence of a single position, on average 1/4 would start with the character A. Now, consider the case of a position in which the ancestral character is A and the descendant character, after time $t = 0.5$, is C. We graphically denote this scenario by:



The probability of simulating such a scenario under the model is $P(A) \cdot P_{A,C}(0.5)$. If we consider the two sequences as our data, D, and the model as M, we can denote this probability by $P(D|M)$, i.e., the probability of the data given the model and its parameters. This is called the likelihood of the model. Consider now the more complicated case in which the sequence ACGGT evolved to ACGAC, again along time $t = 0.5$. The likelihood of the model is the product of the probabilities of each position in the data:

$$(P(A)\cdot P_{A,A}(0.5))\cdot(P(C)\cdot P_{C,C}(0.5))\cdot(P(G)\cdot P_{G,G}(0.5))\cdot(P(G)\cdot P_{G,A}(0.5))\cdot(P(T)\cdot P_{T,C}(0.5)) \tag{28}$$

Note that the product of probabilities reflects the assumption that given the model parameters, the tree and its branches, positions evolve independently of one another. This assumption is a clear oversimplification of the biological reality, but it is included in the vast majority of models employed in molecular evolution since it allows rapid computation of the likelihood function. Under the JC model, the likelihood depends on both t and α (each $P_{i,j}(t)$ term depends on both t and α). We can thus term it $L(t, \alpha)$.

When examining natural biological sequences, we do not know the values of the parameters that gave rise to them. As common in many statistical applications, inferring these values often involves finding the parameter values that maximize the likelihood function. These values are called MLEs: maximum-likelihood estimates. Note however that there are infinitely many (t, α) pairs that maximize the likelihood function. This is because in all these expressions, t and α always appear together as a single term, αt. Thus, we can always multiply α by a constant value, and divide t by this same constant value and the likelihood would remain unchanged. This is a well-known phenomenon in statistics and such models are termed non-identifiable, i.e., there is no one-to-one mapping between the likelihood function and a set of model parameter values. Note that this non-identifiability is not restricted to the JC model; from Eq. 15, we see that $P(t)$ depends on Qt. Thus, we can always multiple Q by a factor and divide t by that same factor, and the $P(t)$ function (and consequently the likelihood function) would remain unchanged. However, in spite the infinite number of MLE (t, α) pairs, we can uniquely infer their product αt. For the above example, and using the JC model, $\alpha t = 0.19$ maximizes the likelihood function.

We will now see that the αt product is also related to the average number of substitutions when sequences evolve for t units of time.

For continuous time evolutionary Markov processes, d, the expected number of substitutions along a branch of length t, can be computed by the following expression:

$$d = -t \sum_i P(i) \cdot Q_{i,i} \tag{29}$$

We do not show the derivation of this formula here. For the above JC model, this yields $d = 3\alpha t$. This suggests that we can choose any one of the pairs (t, α) that maximizes the likelihood function, compute d based on this pair, and obtain the expected number of substitutions along the branch associated with the MLEs. Alternatively, we can set $\alpha = 1/3$, and that would imply that $d = t$. The resulting JC model is:

$$P_{i,j}(t) = \begin{cases} \frac{1}{4} - \frac{e^{-4t}}{4} & \text{if } i \neq j, \\ \frac{1}{4} + \frac{3e^{-4t}}{4} & \text{if } i = j \end{cases} \tag{30}$$

This can be generalized beyond the JC model as according to Eq. 29, d would be equal to t if

$$\sum_i P(i) \cdot Q_{i,i} = -1 \tag{31}$$

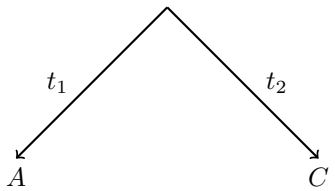
We can always multiply the Q matrix by a fixed factor so that Eq. 31 holds. This is called matrix normalization and, following this procedure, estimates of t would in fact correspond to the expected number of substitutions along a branch of length t, rather than to time.

6 Reversibility, homogeneity and stationarity

Time reversibility of a Markov process imposes the following equation:

$$P(i) \cdot P_{i,j}(t) = P(j) \cdot P_{j,i}(t) \tag{32}$$

This suggests that the probability of starting with character A and ending up with character C along a branch of length t is the same as the probability of starting with character C and ending up with character A (along the same branch). More generally, it suggests that when we have two sequences and we assume that the first is the ancestor and the second is the descendent, the probability of obtaining the two sequences would have been the same if we had rather assumed that the first is the descendent and the second the ancestor. Even more generally, we can choose the "oldest" (ancestral root) point anywhere along the branch t. This point divides the branch t into two segments t_1 and $t_2 = (t - t_1)$, each of which connects the sequences to that new root. Reversibility enforces that the likelihood function becomes invariant to the position of the root, i.e., it would remain the same regardless of our choice of t_1. We show it here for the very simple case of two single character sequences:



The division of the branch between the sequences defines the simplest bifurcating tree. The likelihood computation along the tree is given by summing over all possible states x of the root:

$$\sum_x P(x) \cdot P_{x,A}(t_1) \cdot P_{x,C}(t_2) \tag{33}$$

Using the reversibility condition for the first two terms we obtain:

$$\sum_x P(A) \cdot P_{A,x}(t_1) \cdot P_{x,C}(t_2) \tag{34}$$

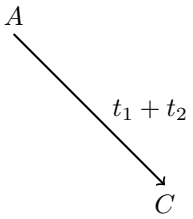
We can take the first term outside the summation:

$$P(A) \sum_x \cdot P_{A,x}(t_1) \cdot P_{x,C}(t_2) \tag{35}$$

Using the Markovian property, we can simplify the equation to:

$$P(A) P_{A,C}(t_1 + t_2) \tag{36}$$

We thus see that the likelihood depends on the total length of two branches, suggesting that one can move the root position as long as the total length remains fixed. This means that we can set one branch to zero and the other one to $t_1 + t_2$ and the likelihood would remain unchanged, i.e., one sequence can be considered ancestral to the other one:



For an unrooted tree with three taxa, it is possible to place the root at any point along each one of the three branches. The same argument that shows that all root positions results in the same likelihood for two sequences holds also for this case, and in fact, the reversibility condition makes all rooted trees that are derived from the same unrooted tree equally probable. This suggests that when assuming a reversible model, it is impossible to infer the location of the tree root. However, reversibility allows very efficient algorithms for inferring branch lengths and thus, for searching for the tree with the highest likelihood, by using the so called pulley principle (Felsenstein, 1981).

If we look at a discrete (or continuous) time Markov chain, we are often interested in the probabilities of being at a specific state after the chain has been running for a long time. In some cases, we find that as t approaches infinity, $P(t)$ converges to a specific form in which all rows are identical, such that $P(\infty) = \pi \cdot I$, and π is a row vector of size n, in which π_i is the probability of reaching character i after infinite amount of time. Moreover, the vector of probabilities π is unchanged by the transition matrix $P(1)$, i.e., $\pi \cdot P(1) = \pi$. Such cases are called stationary chains and π is a unique vector of stationary probabilities. Once the chain reaches stationarity, it remains in stationarity. The JC model, for example, has stationary probabilities of exactly 0.25 for each character.

Homogeneity means that a single Q matrix characterizes the entire evolutionary process. In contrast, branch heterogeneous models assume that different branches of a tree have different Q matrices.

Currently, the vast majority of phylogenetic studies assume models that are reversible, stationary, and homogenous. However, it is clear that such models are oversimplification of reality in many cases. For example, bacterial genomes vary substantially in their GC composition across homologous regions, in clear violation of these assumptions (Galtier and Lobry, 1997).

7 Basic models of nucleotide substitutions

As stated above, in the JC model, which is represented by the simplest nucleotide transition probabilities (see Eq. 27), it is assumed that the instantaneous substitution rates between all pairs of different nucleotides are identical, which implies that the stationary frequencies equal 0.25. There is, however, ample empirical evidence that substitution probabilities vary among nucleotide pairs. A plethora of extensions to this basic modeling scheme have been developed over the last 50 years, each implying different hypotheses regarding the pattern of nucleotide evolution. The Kimura two parameters model (K2P, Kimura, 1980) alleviates the often unrealistic assumption that transitions (a substitution between two pairs of purines or between two pairs of pyrimidines) and transversions (a substitutions between a purine and a pyrimidine or vice versa) are equiprobable. This has resulted is the following Q matrix, where α is the rate of transitions and β is the rate of transversions:

$$Q[K2P] = \begin{pmatrix} -\alpha-2\beta & \beta & \alpha & \beta \\ \beta & -\alpha-2\beta & \beta & \alpha \\ \alpha & \beta & -\alpha-2\beta & \beta \\ \beta & \alpha & \beta & -\alpha-2\beta \end{pmatrix} \tag{37}$$

As for the JC mode, matrix normalization introduces a constraint on the matrix, i.e., given a specific α value, one can compute the β value. In fact, it is possible to rewrite the K2P matrix, following normalization, using a single parameter: the transition-transversion rate ratio. Which of the infinite many possible transition-transversion rate ratios should be used? The one that maximizes the likelihood function. We call this parameter "free", because it is estimated using likelihood based on the analyzed data. While other types of parameterizations are possible, each capturing a different biological aspect, the richest model possible, under the assumption that substitution rates are symmetrical (i.e,, $Q_{i,j} = Q_{j,i}$ for all i and j) is captured by a matrix with six parameters (five free parameters) and is denoted as the SYM model (Zharkikh, 1994):

$$Q[SYM] = \begin{pmatrix} -\alpha-\beta-\gamma & \beta & \alpha & \gamma \\ \beta & -\beta-\delta-\epsilon & \delta & \epsilon \\ \alpha & \delta & -\alpha-\delta-\eta & \eta \\ \gamma & \epsilon & \eta & -\gamma-\epsilon-\eta \end{pmatrix} \tag{38}$$

Because the JC, K2P, and SYM matrices are symmetrical ($Q_{i,j} = Q_{j,i}$ for all i and j) the stationary frequencies of all nucleotides are equal to 0.25. As stated above, biological dataset are often characterized with biased nucleotide frequencies. An important extension to such models relies on non-symmetrical matrices that can results in any desired stationary nucleotide frequencies, denoted by π. To this end, three additional free parameters are added to the model: π_A, π_C, and π_G, representing the frequency of A, C, and G (the frequency of T, π_T is constrained such that the sum of all probabilities is 1). Incorporating these parameters

into the JC, K2P, and SYM models resulted in an expanded set of models, termed F81 (Felsenstein, 1981), HKY (Hasegawa et al., 1985), and the General Time Reversible model (GTR, Tavaré, 1986), respectively, and in general, models that incorporate this possibility are usually denoted by "+F". To retain the time reversibility property, such models take the following general form

$$Q[GTR] = \begin{pmatrix} -\dots & \beta\pi_C & \alpha\pi_G & \gamma\pi_T \\ \beta\pi_A & -\dots & \delta\pi_G & \epsilon\pi_T \\ \alpha\pi_A & \delta\pi_C & -\dots & \eta\pi_T \\ \gamma\pi_A & \epsilon\pi_C & \eta\pi_G & -\dots \end{pmatrix} \tag{39}$$

Note that the model is time reversible because it satisfies $\pi_i Q_{i,j} = \pi_j Q_{j,i}$ for all i and j.

8 Basic codon models

The models detailed above all assumed that the only possible characters are the four DNA nucleotides, i.e., we use an alphabet of size four. When protein-coding nucleotide sequences are analyzed, the transition rates are highly influenced by the effect they exert on the encoded protein. In this case, transition rates between codons I and J may depend on the type of nucleotide substitution, the type of amino-acid replacement, and by codon usage (i.e., the frequency of the target codon). The use of codon models was originally developed around the same time by Goldman and Yang (1994) and Muse and Gaut (1994), see also Chapter 4.5 (Lowe and Rodrigue, 2020). The model is generally represented by a 61×61 codon rate matrix, which describes the instantaneous substitution rate from codon $I = i_1 i_2 i_3$ to codon $J = j_1 j_2 j_3$. This matrix assigns different rates to nucleotide substitutions at the three codon sites according to their type and effect on the coded amino acid:

$$Q_{I,J} = \begin{cases} \alpha_{i_k j_k}\pi_J & I \text{ and } J \text{ differ by one synonymous substitution at codon site } k \\ r(A_I, A_J)\alpha_{i_k j_k}\pi_J & I \text{ and } J \text{ differ by one nonsynonymous substitution at codon site } k \\ 0 & I \text{ and } J \text{ differ by more than one nucleotide} \end{cases} \tag{40}$$

Here, π_J is the frequency of codon J and $\alpha_{i_k j_k}$ is the substitution rate between nucleotide i_k to nucleotide j_k, which can be parameterized based on any time reversible nucleotide substitution model, such as GTR or HKY. $r(A_I, A_J)$ is the replacement rate between the amino acids A_I and A_J encoded by codon I and J, respectively. Note that to maintain the reversibility property, $r(A_I, A_J)$ should be equal to $r(A_J, A_I)$. In the model originally suggested Goldman and Yang (1994), $r(A_I, A_J)$ was dependent on two components: (1) the physiochemical distance between the amino acids A_I and A_J, as measured by the matrix provided by Grantham (1974), and (2) on a free parameter that accounts for the intensity of selection acting on the encoded protein. In the MG model of Muse and Gaut (1994), $r(A_J, A_I)$ was represented simply by the selection-intensity parameter ω, which specifies the nonsynonymous to synonymous substitution rate ratio:

$$Q_{I,J} = \begin{cases} \alpha_{i_k j_k}\pi_J & I \text{ and } J \text{ differ by one synonymous substitution at codon site } k \\ \omega\alpha_{i_k j_k}\pi_J & I \text{ and } J \text{ differ by one nonsynonymous substitution at codon site } k \\ 0 & I \text{ and } J \text{ differ by more than one nucleotide} \end{cases} \tag{41}$$

9 Basic amino-acid models

The analysis of sequences at the amino acid level requires an instantaneous rate matrix of dimension 20 over 20. If the substitution rate between each pair of amino acids is considered a free parameter, 190 parameters need to be estimated from the analyzed data (189 after matrix normalization), which is both computationally demanding and requires large amounts of data for accurate inference. Thus, for protein sequences, a pre-computed matrix is usually used for which the parameters were inferred based on a very large protein dataset. For example, Dayhoff et al. (1978) curated an atlas of all the available protein sequences at that time, and estimated the substitution rates based on a maximum-parsimony-like procedure. When more data became available and when inference procedures improved, updated matrices were inferred, e.g., the JTT, the WAG, and the LG matrices (Jones et al., 1992; Le and Gascuel, 2008; Whelan and Goldman, 2001). In addition, matrices for specific datasets were also introduced: mtREV for mitochondrially encoded proteins (Adachi and Hasegawa, 1996), cpREV for chloroplast encoded proteins (Adachi et al., 2000), and similarly, matrices for different secondary structures or surface accessibility (Koshi and Goldstein, 1995). These matrices are considered empirical because they are based on averaging substitution rates across many datasets. This is in contrast to the above mechanistic models, in which the model parameters were chosen to reflect certain assumptions regarding the substitution pattern and are estimated for each dataset. Of note, these matrices can be decomposed into two components:

$$Q = S \cdot \Pi \tag{42}$$

where S is a symmetrical matrix describing the amino-acid exchangeability component and the diagonal matrix Π represents the amino-acid stationary frequencies. These 20 amino acid frequencies can be estimated from each analyzed data (the "+F" option), and thus, amino-acid models are often a mix between a component estimated based on a very large sequence compendium (the "S" component) and a dataset-specific component (the stationary amino acid frequencies).

10 Among-site-rate-variation

The above models assume the exact same stochastic process at each sequence site. This should not be taken to mean that all positions would experience the same number of substitutions. Due to the stochastic nature of the evolutionary process, given a phylogenetic tree and a specific model, by chance alone, some sequence sites experience more substitutions than others. These differences in the number of substitutions per site follow a Poisson distribution (Yang, 1996). However, it is now well established that the distribution of the number of substitutions per site in real biological data is substantially different from what can be expected by chance alone, i.e., different from a scenario in which all positions evolve exactly under the same stochastic model. For example, for a given dataset, there is often an excess of invariant positions (i.e, they are fully conserved and experience no substitutions along the phylogeny), compared to data simulated assuming a constant model across sites. Branch lengths reflect the number of substitutions averaged over all sites. Thus, if the average number of substitutions per site across the entire tree is high, sites that experience no substitutions are expected to be extremely rare. However, in protein sequences, for example, there are many more invariant sites than the expected number. These sites are usually those that are critical for maintaining the function or structure of the protein,

so that purifying selection removes most of the mutations that appear at these positions. Models that aim to capture the observed variability of substitution rates among different sites are called among-site-rate-variation (ASRV) models. To understand how they work, we first introduce the concept of site-specific evolutionary rate. Consider a site that evolves along a branch of length t, say under the JC model. Because the branch lengths are indicative of the average number of substitutions per site, a site that evolves along a branch of length t is expected to accumulate half the number of substitutions compared to a site that evolves along a branch of length $2t$. Consider the case, in which in a specific position the ancestor character is A and the descent character is C. The likelihood of these data given a branch length of $2t$ is $P_{A,C}(2t)$. An equivalent way to think of this case is of a character evolving along a branch of length t, with a site-specific rate of 2. In general, the likelihood of a scenario in which sequence $A_1, A_2, \ldots, A_N$ evolves into the sequence $B_1, B_2, \ldots, B_N$ along a branch of length t and with site specific rates $r_1, r_2, \ldots, r_N$ is:

$$\prod_{i=1}^{N} P_{A_i B_i}(r_i \cdot t) \tag{43}$$

In general, the site-specific rates are unknown. One possibility would be to assign each site with its own rate parameter. This, however, would result in inevitably large number of parameters that have to be estimated from the data, and would generally result in inferior inferences (Mayrose et al., 2004). Instead, we can assume that these rates are taken from a limited set of values $r_1, r_2, \ldots, r_k$, with corresponding probabilities $w_1, w_2, \ldots, w_k$. Thus, these rates are sampled from a discrete distribution. We can then compute the likelihood while averaging over all possible rates:

$$\prod_{i=1}^{N} \sum_{j=1}^{k} P_{A_i B_i}(r_j \cdot t) \cdot w_j \tag{44}$$

The question then becomes how to choose these rates and their probabilities. Tamura and Nei (1993) were the first to suggest that, for pairwise sequences, rates are gamma distributed. Yang (1993) showed how to compute the likelihood of a tree assuming the rates are sampled from a continuous gamma distribution. Later, Yang (1994) showed how assuming the rates are sampled from a discretized version of the gamma distribution can speed up computations. By far, the discrete gamma distribution is the most widely used ASRV distribution. The gamma distribution with parameters α and β has mean α/β and variance α/β^2. Usually, the unit 1 gamma distribution is used such that $\alpha = \beta$ and the mean rate over all sites is 1. Models incorporating this possibility thus include a single additional parameter, α, and are usually denoted by "+G". In the discretized version of these models, it is common to choose the representative rates, such that each has equal probability $1/k$ (Yang, 1994). While alternative approaches for the discretization of the gamma distribution, based on Laguerre quadrature for example, were suggested (Felsenstein, 2001; Mayrose et al., 2005), these are rarely used. Additionally, several alternatives to the gamma distribution were proposed. First, an additional category, specific for invariant sites was proposed, generating the G+I model (Gu et al., 1995). Second, Kosakovsky Pond and Frost (2005) suggested to discretize the gamma distribution into categories based on quantiles estimated using a discretized beta distribution. Third, Mayrose et al. (2005) suggested that a mixture of several discrete gamma distributions better captures ASRV compared to using a single gamma distribution. Finally, Yang (1995) suggested a free parameter distribution, in which both the rates and their probabilities are parameters which are estimated using maximum likelihood. While

this distribution is highly flexible and can well approximate data-specific characteristics, it is also very parameter rich. Introducing ASRV into probabilistic evolutionary models was shown to increase the accuracy of tree inference and branch lengths as well as many downstream applications that rely on evolutionary models such as molecular dating (Yang, 1996) and quantifying site specific conservation scores (Mayrose et al., 2004). For a review regarding modeling ASRV, see Yang (1996) and Pupko and Mayrose (2010).

11 Mixture models

The above models assume that a single Q matrix represents the evolutionary dynamics across all sites. However, the selective forces and possibly the mutation processes may vary among sites, and in this sense, such models may be an oversimplification of the evolutionary process. It is possible to alleviate this restriction by assuming that there are several possible instantaneous rate matrices Q_1, Q_2, ..., Q_N and that each site is associated with one of these matrices. In case we do not know the matrix that is associated with each site, we compute the likelihood by averaging over all possible matrix assignments, weighted by their probabilities:

$$P(D|M) = \sum_{j=1}^{k} P(D|M_j)P(M_j) \tag{45}$$

where M_j is the model defined by using the matrix Q_j. This is very similar to likelihood computations that incorporate site-specific rates , as detailed in the previous section, where here we average over the various Q matrices rather than averaging over the possible rate categories. Both of these cases are considered mixture models (Zhang and Huang, 2015). Mixture models are widely used to describe codon evolution. While in the MG codon model described in Eq. 41 a single ω parameter is assumed for all sites, it is clear that the type and intensity of this selection coefficient vary, with some sites experiencing purifying selection ($\omega < 1$) while some sites evolve neutrally ($\omega = 1$) or possibly, under positive selection ($\omega > 1$). Mixture models allow modeling this variability in selection intensity directly. A set of possible ω values is assumed, which results in a set of Q matrices. The likelihood is then computed using the above formula for mixture models. Various mixture models were suggested for codon models, in which the ω values vary according to either a free, a gamma or a beta distribution (Yang et al., 2000). Using these models it is possible to test for the presence of positive selection, by contrasting the likelihoods of a null mixture model that allows only Q matrices with $\omega \leq 1$ and an alternative mixture model that also includes a Q matrix with $\omega > 1$. Such models are also often used to infer the posterior estimates of ω for each site, thus revealing sites that evolve under specific selection regimes (Cannarozzi and Schneider, 2012).

12 Gene family models

The above models consist of continuous time Markov chains of nucleotide, amino acid and codon sequences. However, Markov models were also developed to describe the evolutionary dynamics of gene families. The simplest form of such models allows the analysis of phyletic patterns, in which the dataset is a matrix representation of gene family presence and absence across a set of genomes, in which each row represents a genome, each column represents a gene family, and the i, j entry in the matrix is 1 if a member of gene family j is present in genome i and 0 otherwise. The states of such a model are binary $\{0, 1\}$, and thus, a two by

two Q matrix is used to model the evolution of these characters. In such a matrix, $\lambda = Q_{0,1}$ is the gain rate and $\mu = Q_{1,0}$ is the loss rate:

$$Q = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix} \tag{46}$$

We note that a gain event may reflect either de-novo emergence of a gene family or cases of gains by obtaining a copy of a gene via horizontal gene transfer. We also note that a '1' to '0' transition reflects the loss of all copies of a given gene family. Using this rate matrix, it is possible to derive explicit formulae for $P_{ij}(t)$ (Ross, 1996). Imposing reversibility on such a matrix, i.e., $\pi_0\lambda = \pi_1\mu$ results in the condition $\mu = \lambda\pi_0/\pi_1$ and we obtain an alternative representation:

$$Q = \begin{bmatrix} -\lambda & \lambda \\ \lambda\pi_0/\pi_1 & -\lambda\pi_0/\pi_1 \end{bmatrix} = \frac{\lambda}{\pi_1}\begin{bmatrix} -\pi_1 & \pi_1 \\ \pi_0 & -\pi_0 \end{bmatrix} = r_Q \begin{bmatrix} -\pi_1 & \pi_1 \\ \pi_0 & -\pi_0 \end{bmatrix} \tag{47}$$

Here, r_Q is a scaling parameter. Modifying its value would only change the total number of transitions along the phylogenetic tree but not the relative number of gains and losses. If we impose the constraint that branch lengths reflect average number of substitutions along the tree, we should enforce $\sum_i \pi_i Q_{i,i} = -1$, which for this case yields $r_Q = 1/(2\pi_0\pi_1)$ and since $\pi_0 + \pi_1 = 1$, this model has only a single free parameter: π_0. The above model was used and extended in various studies. Hao and Golding (2006) assumed that the gain and loss rates are equal, thus ensuring that the size of the genome does not vary much during evolution. Cohen et al. (2008) extended this model by allowing unequal gain and loss rates as well as introducing rate variation among gene families using a discrete gamma distribution. In such a model, the rate varies but the gain and loss rate ratio is assumed to be identical among all gene families. Subsequent development alleviated this assumption of fixed gain and to loss ratio by introducing mixture models (Cohen and Pupko, 2010; Spencer and Sangaralingam, 2009). One interesting aspect of such models is the need to correct for unobservable data when analyzing empirical data. Consider a case where the ancestral gene family was lost in all its descendants. In the phyletic pattern, this corresponds to a column of zero, reflecting a gene family that is absent from all present-day genomes. However, in empirical data, columns of zero are never observed because phyletic patterns are constructed by homology searches among present-day genomes. Felsenstein (1992) already noted and suggested a solution for this problem when analyzing restriction site presence-absence data. Specifically, the likelihood of the observed data for a specific gene family (D) is in fact conditioned on not being a column of zeros. Let C_0 and C_+ be the events "a column of zeros" and "not a column of zeros", respectively. We obtain:

$$P(D|C_+) = \frac{P(D\&C_+)}{P(C_+)} = \frac{P(D)}{1 - P(C_0)} \tag{48}$$

Thus, the desired probability, $P(D)$ is the product of the conditional probability and the probability of a column not made entirely of zeros. The latter can be easily estimated by computing the probability of the complementary event–a column of zeros. In the above continuous time Markov models, the state "1" represents one or more copies. Instead, it is possible to extend such models to explicitly account for the number of copies in each gene family. In theory, the number of states is infinite but in practice a pre-defined upper bound, M, is used to transfer the state space into a finite state Markov chain, such that the last state includes all values equal or above that number. In this case, the data are coded over the alphabet 0, 1, 2,..., M. The rate matrix in this case is a variant of a birth death model

(Ross, 1996), where the birth rate, λ, and the death rate, δ, are the rates of a single gene gain or loss:

$$Q_{i,j} = \begin{cases} \lambda & \text{j=i+1} \\ \delta & \text{j=i-1} \\ 0 & \text{otherwise} \end{cases} \tag{49}$$

Here too, several extensions have been proposed that allowed for: (i) de novo emergence of a gene family, as specified by the birth-death-innovation model (Csurös, 2010; Karev et al., 2003; Librado et al., 2012; Spencer et al., 2006), (ii) the possibility of whole genome duplication (Rabier et al., 2014), (iii) dependence of the gain and loss rates on the number of gene families (Spencer et al., 2006), and (iv) accounting for differential sequencing coverage and quality of annotations across genomes (Han et al., 2013).

13 Indel models

While probabilistic substitution models are now routinely used when reconstructing phylogenetic trees or searching for positive selection, for the inference of multiple sequence alignments, ad-hoc methods are commonly employed. This difference stems from the fact that modeling insertion and deletion (indel) events is more challenging compared to modeling substitutions, mainly because incorporating indel events within the likelihood function violates the assumption that different sites evolve independently, which is required for efficient likelihood computations. In a breakthrough paper, Thorne et al. (1991) developed the first probabilistic model that includes indel events (the TKF91 model). Unlike substitution models, in which the number of states in the Q matrix is 4, 20 or 61 (for nucleotides, amino acids, or codons), in TKF91, the sequences need to be considered as whole and cannot be seen as "independent columns". Thus, the number of states is exponential in the length of the sequence. This large number of possible states makes direct exponentiation of an explicit Q matrix impossible. To reduce the complexity of the model, the TKF91 model assumes that $Q_{i,j} = 0$ if the length difference between sequences i and j is greater than one. This assumption implies that longer length differences are the outcome of several indel events of length one. While this assumption is clearly unrealistic from what we know of indel mutations, it was made in order to make computations with this model feasible. The introduction of Bayesian integration techniques in TKF91 as well as advanced dynamic programming algorithms further enhanced the ability to compute with such a probabilistic indel model. The TKF91 was extended by the same group of researchers to allow longer indels (TKF92, Thorne et al., 1992). However, to overcome computational challenges, it was assumed that overlapping indels never occur through evolution, which is again biologically unrealistic. A full long-indel model was developed by Miklos et al. (2004). While this model allows indels of any size and overlaps, it is extremely computationally intensive and cannot be applied for more than a handful of sequences. Recently, Levy Karin et al. (2018) introduced significant accelerations to this long-indel model and demonstrated that using such an indel model results in more accurate pairwise alignments compared to widely-used alignment programs, such as MAFFT (Katoh and Standley, 2013), which do not rely on an explicit probabilistic model. Statistical alignment algorithms aim at the joint inference of trees and alignments using probabilistic models (Steel and Hein, 2001). Current tools for statistical alignments such as BaliPhy (Redelings and Suchard, 2005) rely on hidden Markov Models (HMMs), which are also probabilistic-based indel models. However, implicitly, in HMM-based models there is a distinct Markov model for each tree branch. In contrast,

the indel models presented above describe a single process over the whole tree, in which a single set of model parameters is shared among all tree branches. While indel-based models currently lag behind substitution-only models, their development holds the promise to dramatically change various molecular-evolution applications, such as sequence alignment algorithms, the characterization of indel evolutionary dynamics in various genes and lineages (Chen et al., 2009; Levy Karin et al., 2017a; Lunter, 2007), algorithms for simulating sequences (Cartwright, 2005; Fletcher and Yang, 2009) as well as downstream analyses such as tree inference and molecular dating.

14 More sophisticated models

With the increased availability of sequence data and the increased computational resources, the development of more sophisticated inference procedures of sequence evolution has undergone accelerated evolution in itself. A very partial list of influential directions in model development includes:

1. models that account for variation of the process among tree branches. Such model include, for example, codon models that allow for positive selection only on a subset of tree branches (see Yang et al. (2002); Chapter 4.5 [Lowe and Rodrigue 2020]) and models that allow the rate of evolution to change along the tree, e.g., the covarion models (Galtier, 2001);
2. more sophisticated mixture models, which allow averaging over a set of empirical amino-acid matrices (Quang et al., 2008) or sampling amino acid matrices using a Bayesian approach (see also Lartillot and Philippe (2004); Chapter 1.4 [Lartillot 2020]);
3. models that integrate protein structure information with sequence evolution (Choi et al., 2007);
4. models that integrate trait information with sequence evolution (Lartillot and Poujol, 2011; Levy Karin et al., 2017b);
5. models in which the substitution rate continuously evolves (Lartillot and Poujol, 2011);
6. models that allow different partitions of the datasets to evolve under different sets of parameters (Nylander et al., 2004; Lanfear et al., 2016).

References

Adachi, J. and Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *Journal of molecular evolution*, 42(4):459–68.

Adachi, J., Waddell, P. J., Martin, W., and Hasegawa, M. (2000). Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *Journal of molecular evolution*, 50(4):348–58.

Anisimova, M., Liberles, D. A., Philippe, H., Provan, J., Pupko, T., and von Haeseler, A. (2013). State-of the art methodologies dictate new standards for phylogenetic analysis. *BMC evolutionary biology*, 13(1):161.

Bromham, L. (2020). Substitution rate analysis and molecular evolution. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.4, pages 4.4:1–4.4:21. No commercial publisher | Authors open access book.

Cannarozzi, G. M. and Schneider, A. (2012). *Codon Evolution: Mechanisms and Models*. Oxford University Press.

Cartwright, R. A. (2005). DNA assembly with gaps (Dawg): simulating sequence evolution. *Bioinformatics (Oxford, England)*, 21 Suppl 3(Suppl 3):iii31–8.

1.1:18 REFERENCES

Chen, J. Q., Wu, Y., Yang, H., Bergelson, J., Kreitman, M., and Tian, D. (2009). Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Molecular Biology and Evolution*, 26(7):1523–1531.

Choi, S. C., Hobolth, A., Robinson, D. M., Kishino, H., and Thorne, J. L. (2007). Quantifying the impact of protein tertiary structure on molecular evolution. *Molecular Biology and Evolution*, 24(8):1769–1782.

Cohen, O. and Pupko, T. (2010). Inference and characterization of horizontally transferred gene families using stochastic mapping. *Molecular biology and evolution*, 27(3):703–13.

Cohen, O., Rubinstein, N. D., Stern, A., Gophna, U., and Pupko, T. (2008). A likelihood framework to analyse phyletic patterns. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 363(1512):3903–11.

Csurös, M. (2010). Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics (Oxford, England)*, 26(15):1910–2.

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978). A model of evolutionary change in proteins. In Dayhoff, editor, *Atlas of protein sequence and structure*, volume 5 suppleme, pages 345–352. National Biomedical Research Foundation, Washington.

Ekblom, R. and Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, 107(1):1–15.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol*, 17(6):368–76.

Felsenstein, J. (1992). Phylogenies From Restriction Sites: A Maximum-Likelihood Approach. *Evolution*, 46(1):159–173.

Felsenstein, J. (2001). Taking variation of evolutionary rates between sites into account in inferring phylogenies. *Journal of Molecular Evolution*, 53(4-5):447–455.

Fletcher, W. and Yang, Z. (2009). INDELible: A flexible simulator of biological sequence evolution. *Molecular Biology and Evolution*, 26(8):1879–1888.

Galtier, N. (2001). Maximum-likelihood phylogenetic analysis under a covarion-like model. *Molecular Biology and Evolution*, 18(5):866–873.

Galtier, N. and Lobry, J. (1997). Relationships Between Genomic G+C Content, RNA Secondary Structures, and Optimal Growth Temperature in Prokaryotes. *Journal of Molecular Evolution*, 44(6):632–636.

Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, 11(5):725–736.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science (New York, N.Y.)*, 185(4154):862–4.

Gu, X., Fu, Y. X., and Li, W. H. (1995). Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Molecular Biology and Evolution*, 12(4):546–57.

Han, M. V., Thomas, G. W. C., Lugo-Martinez, J., and Hahn, M. W. (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Molecular biology and evolution*, 30(8):1987–97.

Hao, W. and Golding, G. B. (2006). The fate of laterally transferred genes: Life in the fast lane to adaptation or death. *Genome Research*, 16(5):636–643.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, 22(2):160–174.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer applications in the biosciences : CABIOS*, 8(3):275–82.

Jukes, T. H. and Cantor, C. R. (1969). Evolution of protein molecules. *Academy Press*, pages p. 21–132.

Karev, G. P. G., Wolf, Y. Y. I., and Koonin, E. E. V. (2003). Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve? *Bioinformatics*, 19(15):1889–1900.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, 16(2):111–20.

Kosakovsky Pond, S. L. L. and Frost, S. D. D. (2005). A simple hierarchical approach to modeling distributions of substitution rates. *Mol Biol Evol*, 22(2):223–234.

Koshi, J. M. and Goldstein, R. A. (1995). Context-dependent optimal substitution matrices. *Protein engineering*, 8(7):641–5.

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., and Calcott, B. (2016). PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses. *Molecular biology and evolution*, page msw260.

Lartillot, N. (2020). The bayesian approach to molecular phylogeny. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.4, pages 1.4:1–1.4:17. No commercial publisher | Authors open access book.

Lartillot, N. and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21(6):1095–1109.

Lartillot, N. and Poujol, R. (2011). A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Mol Biol Evol*, 28(1):729–744.

Le, S. Q. and Gascuel, O. (2008). An Improved General Amino Acid Replacement Matrix. *Molecular Biology and Evolution*, 25(7):1307–1320.

Levy Karin, E., Ashkenazy, H., Hein, J., and Pupko, T. (2018). A simulation-based approach to statistical alignment. *Systematic Biology*.

Levy Karin, E., Shkedy, D., Ashkenazy, H., Cartwright, R. A., and Pupko, T. (2017a). Inferring rates and length-distributions of indels using approximate Bayesian computation. *Genome Biology and Evolution*, 9(5):1280–1294.

Levy Karin, E., Wicke, S., Pupko, T., and Mayrose, I. (2017b). An Integrated Model of Phenotypic Trait Changes and Site-Specific Sequence Evolution. *Systematic Biology*, 66(6).

Librado, P., Vieira, F. G., and Rozas, J. (2012). BadiRate: estimating family turnover rates by likelihood-based methods. *Bioinformatics (Oxford, England)*, 28(2):279–81.

Lowe, C. and Rodrigue, N. (2020). Detecting adaptation from multi-species protein-coding dna sequence alignments alignments. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.5, pages 4.5:1–4.5:18. No commercial publisher | Authors open access book.

Lunter, G. (2007). Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes. *Bioinformatics*, 23(13).

Mayrose, I., Friedman, N., and Pupko, T. (2005). A gamma mixture model better accounts for among site rate heterogeneity. *Bioinformatics*, 21 Suppl 2:ii151–8.

Mayrose, I., Graur, D., Ben-Tal, N., and Pupko, T. (2004). Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol Biol Evol*, 21(9):1781–1791.

Miklos, I., Lunter, G. A., and Holmes, I. (2004). A "Long Indel" Model for Evolutionary Sequence Alignment. *Molecular Biology and Evolution*, 21(3):529–540.

Muse, S. and Gaut, B. (1994). A likelihood approach for comparing synonymous and non-synonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, 11(5):715–724.

Nylander, J., Ronquist, F., Huelsenbeck, J., and Nieves-Aldrey, J. (2004). Bayesian Phylogenetic Analysis of Combined Data. *Systematic Biology*, 53(1):47–67.

Pett, W. and Heath, T. A. (2020). Inferring the timescale of phylogenetic trees from fossil data. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.1, pages 5.1:1–5.1:18. No commercial publisher | Authors open access book.

Pupko, T. and Mayrose, I. (2010). Probabilistic Methods and Rate Heterogeneity. In *Elements of Computational Systems Biology*. Wiley Online Library.

Quang, L. S., Gascuel, O., and Lartillot, N. (2008). Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics*, 24(20):2317–2323.

Rabier, C.-E., Ta, T., and Ané, C. (2014). Detecting and locating whole genome duplications on a phylogeny: a probabilistic approach. *Molecular biology and evolution*, 31(3):750–62.

Ranwez, V. and Chantret, N. (2020). Strengths and limits of multiple sequence alignment and filtering methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.2, pages 2.2:1–2.2:36. No commercial publisher | Authors open access book.

Redelings, B. D. and Suchard, M. A. (2005). Joint bayesian estimation of alignment and phylogeny. *Systematic Biology*, 54(3):401–418.

Ross, S. M. (1996). *Stochastic Processes*. John Wiley & Sons, New York, NY, 2nd edition.

Spencer, M. and Sangaralingam, A. (2009). A Phylogenetic Mixture Model for Gene Family Loss in Parasitic Bacteria. *Molecular Biology and Evolution*, 26(8):1901–1908.

Spencer, M., Susko, E., and Roger, A. J. (2006). Modelling prokaryote gene content. *Evolutionary bioinformatics online*, 2:157–78.

Stamatakis, A. and Kozlov, A. M. (2020). Efficient maximum likelihood tree building methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.2, pages 1.2:1–1.2:18. No commercial publisher | Authors open access book.

Steel, M. and Hein, J. (2001). Applying the Thorne-Kishino-Felsenstein model to sequence evolution on a star-shaped tree. *Applied Mathematics Letters*, 14(6):679–684.

Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10(3):512–26.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences. Volume 17*, pages 57–86.

Thorne, J. L., Kishino, H., and Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of molecular evolution*, 33(2):114–24.

Thorne, J. L., Kishino, H., and Felsenstein, J. (1992). Inching toward reality: an improved likelihood model of sequence evolution. *Journal of molecular evolution*, 34(1):3–16.

Whelan, S. and Goldman, N. (2001). A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Molecular Biology and Evolution*, 18(5):691–699.

Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular biology and evolution*, 10(6):1396–401.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J Mol Evol*, 39(3):306–14.

Yang, Z. (1995). A space-time process model for the evolution of dna sequences. *Genetics*, 139(2):993–1005.

Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends in ecology & evolution*, 11(9):367–72.

Yang, Z. (2014). *Molecular evolution: a statistical approach*. Oxford University Press.

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A. M. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155(1):431–449.

Yang, Z., Nielsent, R., and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*, 19(6):908–917.

Zhang, H. and Huang, Y. (2015). Finite Mixture Models and Their Applications: A Review. *Austin Biom and Biostat. Austin Biom and Biostat*, 2(2):1013–1.

Zharkikh, A. (1994). Estimation of evolutionary distances between nucleotide sequences. *Journal of molecular evolution*, 39(3):315–29.