Topic 5a: Continuous Time/Discrete-Space Stochastic Pocesses

Learning Objectives

1. Use a transition rate matrix to describe a continuous-time discrete-space Markov Chain (CTDS-MC)
   a. Describe the properties of the transition rate matrix
   b. Derive transition probabilities from a transition rate matrix
2. Propose a justify a CTDS-MC for a biological process
   1. Describe 2 different models of molecular evolution.
   2. Describe a phylogenetic tree and the derivation of the tree likelihood under JC69
3. Analyze a CTDS-MC and use these analyses to draw biological conclusions using:
   a. Simulate a CTDS MC
   b. Derive and analyze the stationary distribution of a CTDS-MC
   c. Derive and analyze the Master equations of a CTDS-MC
   d. Derive and analyze an Ensemble Moment Approximation of a CTDS-MC
   e. Derive and analyze the diffusion approximation to a CTDS-MC
4. Define reversibility of a CTDS-MC, assess whether a given CTDS is reversible, and give an example of a biological process that is modelled as reversible and why.

Lecture 5.1: Continuous-Time Discrete-Space Stochastic Processes

Review Discrete-Time Discrete-Space Stochastic Processes

Discussion: What are some examples of DTDS stochastic processes?

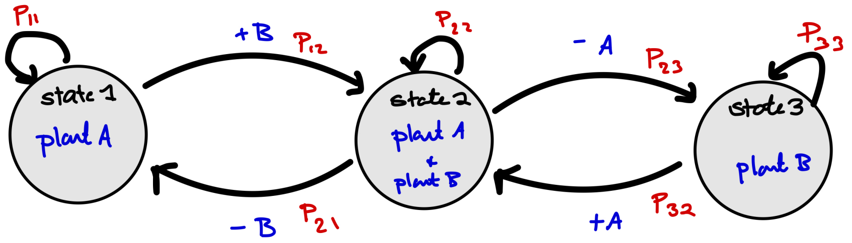
1. We can draw a diagram of DTDS stochastic processes
2. We can represent DTDS processes with a **transition probability matrix**
3. To analyze DTDS processes we can use: **stationary distributions**, **hitting times and probabilities**, **simulation**

Example: 5.1 Species coexistence (revisted)

In ecology, one common application of Markov chains is in modelling the (co)occurrence of species in a community. Let's consider a simple example of a Markov chain modelling the coexistence of two hypothetical species Plant A and Plant B in an ecological reserve.

State 1: Plant A only
State 2: Plant A and B
State 3: Plant B only

1. Draw this stochastic process assuming only one or zero species can be lost/gained in a given year.



2. Let's assume the following transition rate matrix

$$P = \begin{bmatrix} 0.7 & 0.3 & 0 \\ 0.4 & 0.5 & 0.1 \\ 0 & 0.3 & 0.7 \end{bmatrix}$$

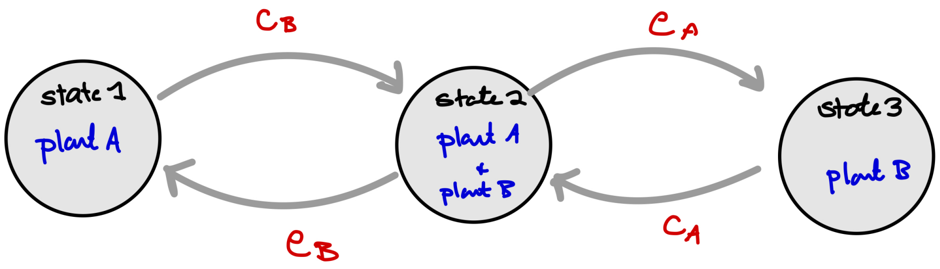
What is P_{12} and what does this probability represent.

$P_{12} = 0.3$ is the probability that a Plant B is introduced to an ecosystem and coexists with Plant A that is already present until the next year/time step.

DTDS stochastic processes work well for systems in which are naturally described in discrete time (e.g., annual life cycles) or are only observed by researchers on discrete time intervals (e.g., biannual surveys), but many other systems are better described by events occurring continuously and not synchronously. For example, consider the ecological dynamics described above but in a tropical environments where plants may exhibit reduced synchrony.

An example CTDS process

Let's reimagine the coexistence problem but this time assuming that the plants colonize and are excluded in continuous time. Let c_X be the colonization rate of species X and e_X be the rate at which species X is excluded from an environment containing the other species. We could logically model the system them as:



I have represented these **rates** with grey arrows.

Note there is no longer a arrow of going from State X to State X as this is the rate at which noting happens which is the default.

Transition Rate Matrices

As we did for the DTDS stochastic process we can describe a CTDS process with a matrix known as the **transition rate matrix**, $\mathbf{Q}$. In this matrix element $q_{ij} \;\; i \neq j$ represents the rate of going from state i to state j whereas element $q_{ii} = -\sum_{j \neq i} q_{ij}$,

Properties of the Transition-Rate Matrix

- The off-diagonal elements of $\mathbf{Q}$ are positive
- The diagonal elements of $\mathbf{Q}$ are negative
- The rows of $\mathbf{Q}$ sum to 0 (Not 1 like the transition probability matrix!)

Example: 5.2 Species coexistence in continuous time

1. Give the CT coexistence model described above, what is the transition rate matrix for this system?

$$\mathbf{Q} = \begin{bmatrix} -c_B & c_B & 0 \\ e_B & -(e_B + e_A) & e_A \\ 0 & c_A & -c_A \end{bmatrix}$$

Deriving the Transition Probability Matrix

Given that the system is in state i at time t, what is the probability that it is in state j at time $t + \Delta t$? In other words, what is:

$$\Pr(X_{t+\Delta t} = j | X_t = i)$$

Note that when $\Delta t = 1$, this is equivalent to calculating the corresponding transition probability.

Recall that *transition probabilities* satisfy the Chapman-Kolmogorob equation:

$$p_{ij}^{t+h} = \sum_k p_{ik}^t p_{kj}^h$$

Subtracting p_{ij}^t from both sides we have:

$$\begin{aligned} p_{ij}^{t+h} - p_{ij}^t &= \left(\sum_k p_{ik}^h p_{kj}^t \right) - p_{ij}^t \\ &= \left(\sum_{k \neq i} p_{ik}^h p_{kj}^t \right) + p_{ii}^h p_{ij}^t - p_{ij}^t \\ &= \left(\sum_{k \neq i} p_{ik}^h p_{kj}^t \right) + (p_{ii}^h - 1) p_{ij}^t \end{aligned}$$

Dividing through by h and taking the limit we have:

$$\lim_{h \to 0} \frac{p_{ij}^{t+h} - p_{ij}^t}{h} = \left(\sum_{k \neq i} p_{kj}^t \lim_{h \to 0} \frac{p_{ik}^h}{h} \right) + p_{ij}^t \lim_{h \to 0} \frac{(p_{ii}^h - 1)}{h}$$

We then note that:

$$\begin{aligned} &\lim_{h \to 0} \frac{p_{ij}^h}{h} = q_{ij} \\ &\lim_{h \to 0} \frac{p_{ii}^h - 1}{h} = q_{ii} \\ &\lim_{h \to 0} \frac{p_{ij}^{t+h} - p_{ij}^t}{h} = \frac{dp_{ij}}{dt} \end{aligned}$$

So:

$$\frac{dp_{ij}}{dt} = \sum_k p_{kj}^t q_{ik}$$

This means wethe transition probabilities are given by the system of ODEs represented by the matrix equation:

$$P'(t) = P(t)Q$$

If we ignore for a second that P and Q are matrices, we can recognize that this is a simple linear ODE that we know the solution of (analogous to exponential population growth) and hence we have:

$$P(t) = P(0)e^{Qt}$$

and hence:

$$\Pr(X_{t+\Delta t} = j | X_t = i) = P(t)e^{Q\Delta t}$$

But what does it mean to have e to a matrix? Recall that $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ so we can express e^{Qt} as:

$$e^{Q\Delta t} = \sum_{n=0}^{\infty} \frac{(Q\Delta t^n)}{n!} \approx \underbrace{\mathbf{I}}_{\text{nothing happens}} + \underbrace{Q\Delta t}_{\text{1 event}} + \underbrace{\frac{Q^2 \Delta t^2}{2}}_{\text{2 events}} + \mathcal{O}(\Delta t^3)$$

where the approximation describes the outcome over a short amount of time Δt. This can be on one way of thinking about why the exponential function shows up all the time in biology as it captures and weights the probability of different numbers of events ocurring.

Example: 5.3 Species coexistence in continuous time cont

Using the transition rate matrix above:

$$\mathbf{Q} = \begin{bmatrix} -c_B & c_B & 0 \\ e_B & -(e_B + e_A) & e_A \\ 0 & c_A & -c_A \end{bmatrix}$$

and letting

$$c_B = 0.3\frac{1}{\text{year}} \quad c_A = 0.1\frac{1}{\text{year}} \quad e_B = 0.25\frac{1}{\text{year}} \quad e_A = 0.01\frac{1}{\text{year}}$$

What is the probability that an ecosystem starting with both species present (state 2) has only species A present (state 1) after 1 month ($\Delta t = \frac{1}{12}$)?

1. Approximate this value assuming that only one event occurs in time Δt

Here we want the $(2, 1)$ element of the matirx:

$$\mathbf{I} + \mathbf{Q}\Delta t = \begin{bmatrix} 1 - c_B\Delta t & c_B\Delta t & 0 \\ e_B\Delta t & 1 - (e_B + e_A)\Delta t & e_A\Delta t \\ 0 & c_A\Delta t & 1 - c_A\Delta t \end{bmatrix}$$

which is: $e_B\Delta t = 0.25 * 0.083 = 0.0208$

2. Approximate this value assuming that up to two events occurs in time Δt

Here we want the $(2, 1)$ element of the matirx:

$$\mathbf{I} + \mathbf{Q}\Delta t + \frac{(\mathbf{Q}\Delta t)^2}{2}$$

This gives:

$$\begin{bmatrix} 0.97615 & 0.02383 & 0.00002 \\ 0.01986 & 0.97933 & 0.00081 \\ 0.00017 & 0.00808 & 0.99174 \end{bmatrix}$$

And a resulting probability of $P_{2,1}(1/12) = 0.01986$

Discussion:How do these answers compare and what does that tell us about the validtiy of the approximation?

The Taylor Series expression above is great as a conceptual understanding and the purposes of simulating a CTDS process incrementally. But what if we want to calculate the transition probability over long time spans? For this we use Eigen-decomposition.

Eigendecomposition

Recall that a square (non-singular matrix) can be decomposed into the product:

$$\mathbf{Q} = \mathbf{A}.\mathbf{D}.\mathbf{A}^{-1}$$

where $\mathbf{A}$ is a matrix who's columns are the (left) eigenvectors of $\mathbf{Q}$, $\mathbf{D}$ is a diagonal matrix of eigenvalues, and $\mathbf{A}^{-1}$ is the inverse of $\mathbf{A}$ which results in a matrix who's rows are the right eighevalues of $\mathbf{Q}$.

$$\mathbf{A}=\begin{bmatrix}\vec{u}_1 & \vec{u}_1 & \dots & \vec{u}_n\end{bmatrix}\quad \mathbf{D}=\begin{bmatrix}\lambda_1 & 0 & \dots & 0\\ 0 & \lambda_2 & \ddots & 0\\ \dots & & \ddots & 0\\ 0 & 0 & \dots & \lambda_n\end{bmatrix}\quad \mathbf{A}^{-1}=\begin{bmatrix}\vec{v}_1\\ \vec{v}_2\\ \vdots\\ \vec{v}_n\end{bmatrix}$$

So (without a formal derivation) we have:

$$e^{\mathbf{Q}t}=e^{\mathbf{A}.\mathbf{D}.\mathbf{A}^{-1}t}=\mathbf{A}.e^{\mathbf{D}t}.\mathbf{A}^{-1}$$

where:

$$e^{\mathbf{D}t}=\begin{bmatrix}e^{\lambda_1 t} & 0 & \dots & 0\\ 0 & e^{\lambda_2 t} & \ddots & 0\\ \dots & & \ddots & 0\\ 0 & 0 & \dots & e^{\lambda_n t}\end{bmatrix}$$

This equality follows can be derived from the fact that e^x can be defined as a series of powers. The derivation itself is cool but not really important for us here.

We can re-order the eigenvalues/eigenvectors in any order, so lets assume that $\lambda_1 > \lambda_2 > \dots$. The eigenvalues of a transition rate matrix have a few convenient properties:

1. There is at least one eigenvalue of $\lambda = 0$ if Q is "strongly connected" (it is possible to reach every state from every other state eventually, this is necessary for there to be a unique stationary distribution).
2. All the other eigenvalues are less then 0, $0 > \lambda > 2\min q_{i,i}$

Then recall that our goal here is to understand changes over large amounts of time, in which case:

$$e^{\lambda_1 t}=1 \gg e^{\lambda_2 t} \gg \dots$$

So we can approximate:

$$e^{\mathbf{Q}t}\approx\begin{bmatrix}\vec{u}_1 & \vec{u}_1 & \dots & \vec{u}_n\end{bmatrix}.\begin{bmatrix}1 & 0 & \dots & 0\\ 0 & 0 & \ddots & 0\\ \dots & & \ddots & 0\\ 0 & 0 & \dots & 0\end{bmatrix}.\begin{bmatrix}\vec{v}_1\\ \vec{v}_2\\ \vdots\\ \vec{v}_n\end{bmatrix}=\underbrace{\vec{u}_1.\vec{v}_1}_{\approx\mathbf{P}}$$

Note that $\vec{u}_1$ is a column vector and $\vec{v}_1$ is a row vector so their product is a matrix. The net result is approximately the transition probability matrix over long periods of time $\mathbf{P}$.Note that the result here doesn't depend on t and hence is the transition matrix in the realm of time over which the **stationary distribution** applies.

What are $\vec{u}_1$ and $\vec{v}_1$? $\vec{u}_1$ is known as the left eigenvector and $\vec{v}_1$ as the right eigenvector. By definition of an eigenvector we have:

$$\mathbf{Q}.\vec{u}_1=\lambda\vec{u}_1=\vec{0}$$
$$\vec{v}_1.\mathbf{Q}=\lambda\vec{v}_1=\vec{0}$$

Recall that the leading eigenvalue is 0. This gives two systems of equations:

$$\sum_k u_{1,k}Q_{j,k}=0\ \forall j \quad \sum_j Q_{j,k}v_{1,j}=0\ \forall k$$

Given that the rows of $\mathbf{Q}$ sum to 0 by definition $\sum_k Q_{j,k}$ we know that $u_{1,1}=u_{1,2}=\dots=u_{1,n}$ and the row vector $\vec{v}_1$ (the leading right eigenvector) is the stationary distribution!

Example: 5.4 Species coexistence in continuous time cont

Python: Lecture5_1.ipynb

1. What is the probability of going from state 2 to state 1 in the long term?

Discussion: How does this answer compare to the results over the short term?

2. What is the long-term probability of ending up in state 1?

Lecture 5.2 Models of Molecular Evolution

One important application of CTDS models in biology is in describing how genomes evolve. Specifically, in describing how mutations occur at different bases in the genome. Such models are known as **models of moleduclar evolution**. There are two general types of such models: "nucleotide substitution" models and "amino acid substitution" models. Nucleotide models describe the rate at which each of the four nucleotides (A,C,G,T) mutates to become each of the other nucleotides. Amino-acid models describe how each of the 20 essential Amino acids evolve to become each of the other 20.

The motivation and design of AA models is complicated having to do with when and how mutations occur and the relative chemical properties of different AAs. The standard models for AA substitution are called the **BLOSUM** matrices (https://en.wikipedia.org/wiki/BLOSUM). This is all I am going to mention about this.

There are a array of Nucleotide substitution models, each with different properties and assumptions. In this class we are going to talk about three different options, each of which is known by an acronym: JC69, HKY84, GTR. Each of these is by definition a 4 state process.

JC69: Jukes Cantor 1969

This is the simplist model of molueclar evolution. It assumes that each nucleotide is equally likely to become each other nucleotides. Specifically the rate of going from nucleotide x to nucleotide y is denoted as μ.

Example: 5.5 JC69

1. Draw a transition diagram for the JC69 model.

2. What is the transition rate matrix for JC69?

$$\mathbf{Q}^{JC69} = \begin{bmatrix} -3\mu & \mu & \mu & \mu \\ \mu & -3\mu & \mu & \mu \\ \mu & \mu & -3\mu & \mu \\ \mu & \mu & \mu & -3\mu \end{bmatrix}$$

3. Suppose that $\mu = 10^{-3} \frac{\text{mut}}{\text{site*year}}$ (a value appropriate for the evolution of bases in a virus genome). Suppose it takes approximately 2 weeks ($\Delta t = \frac{1}{24}$) for patient A to infect patient B. What is the probability that a mutation occurred at a focal site during the course of patient A's infection? What about two nmutations?

The per-base pair mutation rate is μ meaning that that the total rate at which a mutations occur is 3μ (see the off diagonal element of the rate matrix). We need to ask, what is the probability that a mutation occurs at or before $t = \frac{1}{24}$?

The CDF of the exponential distribution is:

$$\Pr(T \leq t^* = \frac{1}{24}) = 1 - e^{-\lambda t^*} = 1.24 \times 10^{-4}$$

Where $\lambda = 3\mu$
The probability that two mutations occurred can be calculated from the CDF of the Erlang distribution with $k = 2$ (this CDF is a function called a Gamma Regularlized Function, so we will just evaluate it numerically)

$$\Pr(2 \text{ mutations}) = 7.81 * 10^{-9}$$

To complement this let's calculate the 1-step and 2-step approximations to the transition probability matrix.

$$P_1(\Delta t = \frac{1}{24}) = \mathbf{I} - \mathbf{Q}\Delta t = \begin{bmatrix} 1-3\mu\Delta t & \mu\Delta t & \mu\Delta t & \mu\Delta t \\ \mu\Delta t & 1-3\mu\Delta t & \mu\Delta t & \mu\Delta t \\ \mu\Delta t & \mu\Delta t & 1-3\mu\Delta t & \mu\Delta t \\ \mu\Delta t & \mu\Delta t & \mu\Delta t & 1-3\mu\Delta t \end{bmatrix}$$

$$= \begin{bmatrix} 0.99987 & 0.000042 & 0.000042 & 0.000042 \\ 0.000042 & 0.99987 & 0.000042 & 0.000042 \\ 0.000042 & 0.00004 & 0.99987 & 0.00004 \\ 0.000042 & 0.000042 & 0.000042 & 0.99987 \end{bmatrix}$$

$$P_2(\Delta t) = \begin{bmatrix} 6\Delta \mathrm{t}^2\mu^2 - 3\Delta \mathrm{t}\mu + 1 & \Delta \mathrm{t}\mu(1-2\Delta \mathrm{t}\mu) & \Delta \mathrm{t}\mu(1-2\Delta \mathrm{t}\mu) & \Delta \mathrm{t}\mu(1-2\Delta \mathrm{t}\mu) \\ \Delta \mathrm{t}\mu(1-2\Delta \mathrm{t}\mu) & 6\Delta \mathrm{t}^2\mu^2 - 3\Delta \mathrm{t}\mu + 1 & \Delta \mathrm{t}\mu(1-2\Delta \mathrm{t}\mu) & \Delta \mathrm{t}\mu(1-2\Delta \mathrm{t}\mu) \\ \Delta \mathrm{t}\mu(1-2\Delta \mathrm{t}\mu) & \Delta \mathrm{t}\mu(1-2\Delta \mathrm{t}\mu) & 6\Delta \mathrm{t}^2\mu^2 - 3\Delta \mathrm{t}\mu + 1 & \Delta \mathrm{t}\mu(1-2\Delta \mathrm{t}\mu) \\ \Delta \mathrm{t}\mu(1-2\Delta \mathrm{t}\mu) & \Delta \mathrm{t}\mu(1-2\Delta \mathrm{t}\mu) & \Delta \mathrm{t}\mu(1-2\Delta \mathrm{t}\mu) & 6\Delta \mathrm{t}^2\mu^2 - 3\Delta \mathrm{t}\mu + 1 \end{bmatrix}$$

4. Of course a virus genome has more than 1 site. SARS-CoV-2 has ≈ 1000 bases. What is the probability that at least one mutation has occurred at any one site during the course of patient A's infection.

Now $\lambda = 3 * 1000 * 10^{-3} = 3$.

The probability that there is *at least* one mutation then is: $1 - e^{-\frac{3}{24}} = 0.117$

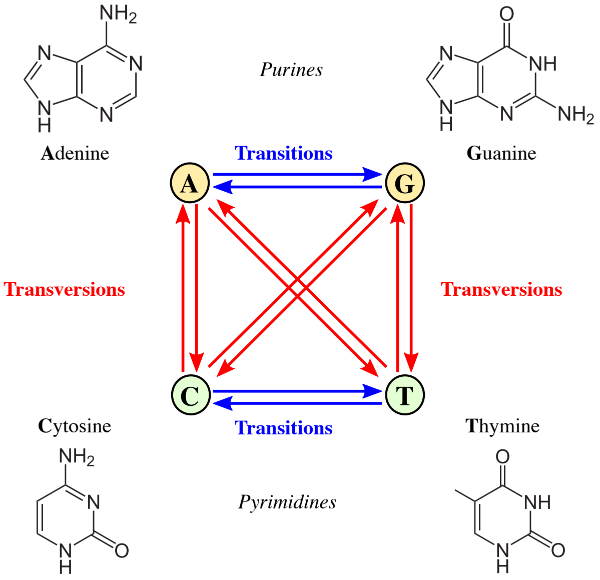
The probability that two mutations occurred is: 0.007.

So the probability that exactly one mutation occurring is: $\approx 0.117 - 0.007 = 0.11$ assuming that three or more mutations are very unlikely to occur.

Discussion: JC69 is an extreme approximation of genome evolution, what would you add/change abou tthis model for realism? Note that the goal is also to minimize parameters that have to be estimated.

K80: Kimura 1980 & HKY84: Hasegawa-Kishino-Yano 1984

JC69 assumes that all mutations are equally likely to occur, this is certainly not the case. We can model some of this complexity by



There are two general types of base pairs "purines" which are large and "pyrimidines" which are small. In order to maintain the stability of the genome mutations are more likely to occur within these groups (purine->purine or pyrimidine->pyrimidine), mutations known as transitions, then between them, transversions. Hence we purpose the following model known as K80:

$$\mathbf{Q}^{K80} = \begin{bmatrix} -(\alpha+2\beta) & \alpha & \beta & \beta \\ \alpha & -(\alpha+2\beta) & \beta & \beta \\ \beta & \beta & -(\alpha+2\beta) & \alpha \\ \beta & \beta & \alpha & -(\alpha+2\beta) \end{bmatrix}$$

Where the states are ordered (1: T, 2: C, 3:A, 4:G) and transitions occur at a rate α and transversions occur at a rate β.

Example: 5.6 K80

1. What is the stationary distribution of the K80 model above?

Recall that the leading eigenvalue (of a strongly connected transition rate matrix like this one) is $\lambda = 0$, so we just want to solve for the corresponding right eigenvectors.

$$\mathbf{Q}.\vec{v}_1 = \lambda \vec{v}_1 = 0$$

$$\begin{aligned}
\begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} v_1 & v_2 & v_3 & v_4 \end{bmatrix}. &\begin{bmatrix} -(\alpha+2\beta) & \alpha & \beta & \beta \\ \alpha & -(\alpha+2\beta) & \beta & \beta \\ \beta & \beta & -(\alpha+2\beta) & \alpha \\ \beta & \beta & \alpha & -(\alpha+2\beta) \end{bmatrix} \\
= &\begin{bmatrix} v_1(-\alpha-2\beta)+\alpha v_2+\beta v_3+\beta v_4 \\ v_2(-\alpha-2\beta)+\alpha v_1+\beta v_3+\beta v_4 \\ v_3(-\alpha-2\beta)+\alpha v_4+\beta v_1+\beta v_2 \\ v_4(-\alpha-2\beta)+\alpha v_3+\beta v_1+\beta v_2 \end{bmatrix}^T
\end{aligned}$$

and hence $v_1 = v_2 = v_3 = v_4$. Recall that eigevectors have aribitrary length but we want to normalize so that the eigenvector sums to 1, so...

$$\vec{\pi} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$$

Discussion: What does this stationary distribution mean for the frequency of A, C,G,T in the genome in the long term?

2. Consider a sequences with 50 bases, simulate molecular evolution for $t = 20$ time steps assuming $\alpha = 0.1$ and $\beta = 0.04$.

Phython: Lecture5_2.ipynb

3. Do the observed frequencies of each nucleotide at the end of your simulation match your expectations from the stationary distribution?

Phython: Lecture5_2.ipynb

This model predicts that the frequencies of A,C,G, and T's observed in the genome in the long term should be equal. Given that evolution has been occurring for a long time, this means that all genomes should be made up of approximately equal amounts of each nucleotide. This is not true though in the real world, and in fact the proportion of each nucleotide varies significantly across the tree of life. Hence, the HKY84 model takes the backbone of K80 and adjusts it so that the stationary distribution reflects an the "empirical distribution".

Specifically, let $\vec{\pi}$ be a vector representing the OBSERVED proportion of each nucleotide in the genome. Then the HKY84 model is represented by the transition rate matrix:

$$\mathbf{Q}^{HKY} = \begin{bmatrix} -(\alpha\pi_C+\beta(\pi_A+\pi_G)) & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha\pi_T & -(\alpha\pi_T+\beta(\pi_A+\pi_G)) & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & -(\alpha\pi_G+\beta(\pi_T+\pi_C)) & \alpha\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha\pi_A & -(\alpha\pi_A+\beta(\pi_C+\pi_T)) \end{bmatrix}$$

The stationary distribution of this model is:

$$\vec{\pi} = [\pi_T, \pi_C, \pi_A, \pi_G]$$

Or the empirical stationary distribution.

GTR: General Time Reversible

(Model originally developed by Simon Taveré in 1986)

By convention in this model the nucleotides are listed in alphebetical order so the states are in order $\{A, C, G, T\}$

$$\mathbf{Q}^{GTR} = \begin{bmatrix} -(a\pi_C + b\pi_G + c\pi_T) & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & -(a\pi_A + d\pi_G + e\pi_T) & d\pi_G & e\pi_T \\ \pi_A b & \pi_C d & -(\pi_A b + \pi_C d + f\pi_T) & f\pi_T \\ \pi_A c & \pi_C e & f\pi_G & -(\pi_A c + \pi_C e + f\pi_G) \end{bmatrix}$$

A CTDS stocahstic process is known as **Time Reversable** if it satisfies:

$$\pi_i \mathbf{Q}_{ij} = \pi_j \mathbf{Q}_{ji} \quad \forall i, j$$

This implies:

$$\begin{aligned} \pi_i \mathbf{P}_{ij}(t) &= \pi_j \mathbf{P}_{ji}(t) \\ \mathbf{P}_{ij}(t) &= \frac{\pi_j}{\pi_i} \mathbf{P}_{ji}(t) \end{aligned}$$

More intuitively a stochastic process is time reversible if:

$$\begin{aligned} &\Pr(X_t = x_0, X_{t+\delta_1} = x_1, X_{t+\delta_1+\delta_2} = x_2, \ldots X_{t+\sum_j^n \delta_j} = x_n) = \\ &\Pr(X_t = x_n, X_{t+\delta_n} = x_{n-1}, X_{t+\delta_n+\delta_{n-1}} = x_{n-2}, \ldots X_{t+\sum_j^n \delta_j} = x_0) \end{aligned}$$

Example: 5.7

1. Show that the GTR model is time reversible

Let's check a few elements: (1:T, 2:C, 3:A, 4:G)

Element 1,2: $Q_{1,2} = a\pi_T$ and $Q_{2,1} = a\pi_C \quad a\pi_C \times \frac{\pi_T}{\pi_C} = Q_{1,2}$

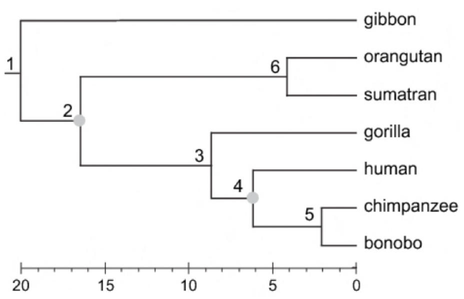
Element 2,4: $Q_{2,4} = e\pi_C$ and $Q_{4,2} = e\pi_G \quad e\pi_G \times \frac{\pi_C}{\pi_G} = Q_{2,4}$

Lecture 5.3 Tree Likelihoods

Phylogenetic Trees

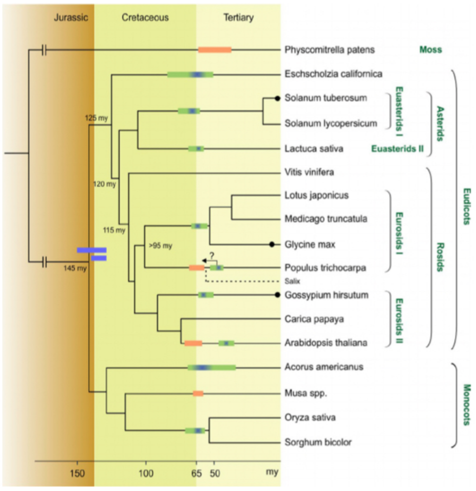
A phylogenetic tree is a diagrammatic representation of the evolutionary relationships among a group **taxa** (distantly related organisms, for example different species, different viruses as viruses do not reproduce sexually and evolve rapidly, or individuals from distinct populations). It depicts the common ancestry and divergence of these taxa over time.

Apes



Yang + Rannala 2005

Flowering Plants



Fawcett et al 2015 PNAS

Discussion: Given this tree, what is the closest relative of a human?

There are multiple different uses of phylogenetic trees:

Classification and Taxonomy: Phylogenetic trees provide a framework for classifying and organizing biodiversity. By identifying evolutionary relationships, biologists can create taxonomies that reflect the true genetic or evolutionary relatedness among species.

Comparative Biology: Phylogenies provide a framework for comparative biology, allowing researchers to study the diversity of traits, behaviours, and physiological processes across related species while accounting for their evolutionary relationships.

Evolutionary Processes: Phylogenies help researchers investigate the mechanisms of evolutionary change, such as the relative strengths of natural selection and genetic drift, or the drivers of speciation and extinction.

Phylodynamics: Inference of the rate at which lineages speciate/go extinct. In the context of viral phylogenetic trees this can inform the transmission rate and recovery rate of an infectious disease.

Here are some key components and concepts related to phylogenetic trees:

1. **Nodes:** Represent points where lineages split, indicating common ancestors. Nodes are often labeled with the estimated time of divergence or other relevant information (such as the statistical certainty of that node).
2. **Branches:** Connect nodes and represent the evolutionary pathways from common ancestors to descendant species or taxa.
3. **Tips or Leaves:** Represent the terminal endpoints of the tree, indicating currently existing taxa.
4. **Root:** The point on the tree that represents the common ancestor of all the taxa included in the tree.
5. **Branch Lengths:** The lengths of branches can represent evolutionary distances, such as genetic divergence or time. Longer branches indicate more evolutionary change (e.g., mutations). In this class we will use branch lengths to indicate time.

Discussion: How many tips and how many nodes are in the ape tree above? Which two species are most cloesly related?

Phylogenetic trees can be constructed based on various types of data, including molecular sequences (DNA, RNA, or protein), morphological characteristics, or a combination of both. Tree-building methods use algorithms to analyze the data and infer the most likely evolutionary relationships among the taxa. There are three general classes of tree building methods: Parsimony, Likelihood, and Bayesian. Here we will only talk about the second and only in the case of DNA (i.e. molecular) data.

When constructing a phylogenetic tree, we often consider a *single genome per taxa* known as the **consensus sequence**. Note that species have alot of diversity within them as we explored with the coalescent. But for a phylogeny we ignore this and go with a single "most representative" sequence.

Calculating the tree likelihood

We will talk much more about this in Topic 7, but here we want to calculate the Likelihood of a model, Θ, given some observed data $\mathcal{D}$:

$$\mathcal{L}(\Theta|\mathcal{D}) \overset{\text{def}}{=} \Pr(\mathcal{D}|\Theta)$$

In our case the data is the observed genome sequences and the model is the parameters of a molecular evolution transition rate matrix.

Example: 5.8: Molecular models

1. Assuming a JC69 model what is Θ?

$\Theta^{JC69} = \{\mu\}$

2. Assuming a HKY84 model what is Θ?

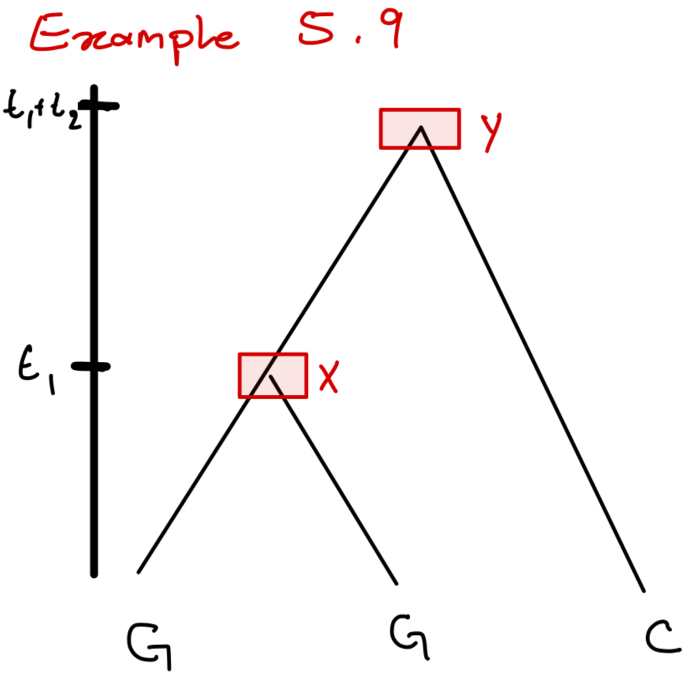
$\Theta^{HKY84} = \{\alpha, \beta\}$

Note that $\vec{\pi}$ is obtained directly from the data and is not a flexible parameter.

Example: 5.9: Tree Likelihood

Consider the following phylogenetic tree.

Newick Format
$((G:t_1, G:t_1):t_2, C)$



Parameters:

$$\mu = 1.2$$
$$t_1 = 0.3$$
$$t_2 = 0.5$$

1.What is the likelihood of this tree given the observed sequences under the JC69 model?

Let $\mathcal{D}$ be the "Data" which is the states at the tips of the tree.

Let Θ^{JC69} be the parameters of the JC69 model
Let $\mathcal{T}$ the the tree topology (branching pattern) and branch lengths.

The likelihood of the tree given the data is the probability of the data given the tree:

$$\mathcal{L}(\mathcal{T}, \Theta|\mathcal{D}) = \Pr(\mathcal{D}|\mathcal{T}, \Theta)$$

To calculate the probability of the tree we:

- Calculate the probability of the changes on each branch (edges in the tree graph).
- When there is an internal node that we don't know the state of, we simply sum over all the possible values it could take on.
- We weight the root node states by the stationary distribution $\Pr(Y) = \pi_Y$

$$\mathcal{L}(\mathcal{T}, \Theta|\mathcal{D}) = \Pr(\mathcal{D}|\mathcal{T}, \Theta) = \sum_{Y \in \{A,C,G,T\}} \Pr(Y) P_{Y,C}(t_1 + t_2) \sum_{X \in \{A,C,G,T\}} P_{X,Y}(t_2) P_{G,X}(t_1) P_{G,X}(t_1)$$

So how do we calculate $P_{i,j}(t)$? Recall that the transition probability matrix of a CTDS process is given by:

$$\mathbf{P}(t) = e^{\mathbf{Q}t}$$

We can solve for this numerically in python.
Phython: Lecture5_3.ipynb

2.What is the probability that root node Y is a G?

To calculate this we use a modified version of Bayes Theorem that allows for additional conditioning of all terms:

$$\Pr(A|B, C) = \frac{\Pr(B|A, C) \Pr(A|C)}{\Pr(B|C)}$$

Here we consider a specific tree $\mathcal{T}$ (see the tree topology above) and molecular model Θ (JC69 with $\mu = 1.2$) so:

$$\Pr(Y = G|\mathcal{D}, \mathcal{T}, \Theta) = \frac{\Pr(\mathcal{D}|Y = G, \mathcal{T}, \Theta) \Pr(Y = G|\mathcal{T}, \Theta)}{\Pr(\mathcal{D}|\mathcal{T}, \Theta)}$$

where $\mathcal{D}$ is the 'data' at the tips.

Here $\Pr(Y = G|\mathcal{T}, \Theta) = \pi_G$ is the "prior" probability that the root is a G, assuming that evolution has occured for a long time prior to this tree our best guess for the state is the stationary distribution of the JC69 model or $\pi_G = \frac{1}{4}$

The denominator $\Pr(\mathcal{D}|\mathcal{T}, \Theta)$ is the likelihood we calculated in the last part.

Finally, $\Pr(\mathcal{D}|Y = G, \mathcal{T}, \Theta)$ is very similar to the likelihood above but fixing the root state at a G

$$\Pr(\mathcal{D}|Y = G, \mathcal{T}, \Theta) = P_{G,C}(t_1 + t_2) \sum_{X \in \{A,C,G,T\}} P_{X,G}(t_2) P_{G,X}(t_1) P_{G,X}(t_1)$$

When there is more than one site/base being considered, we assume that evolution is independent at each base. There are more complex models that allows the mutation rate to vary from base to base, but lets assume it is equal for each base for now.

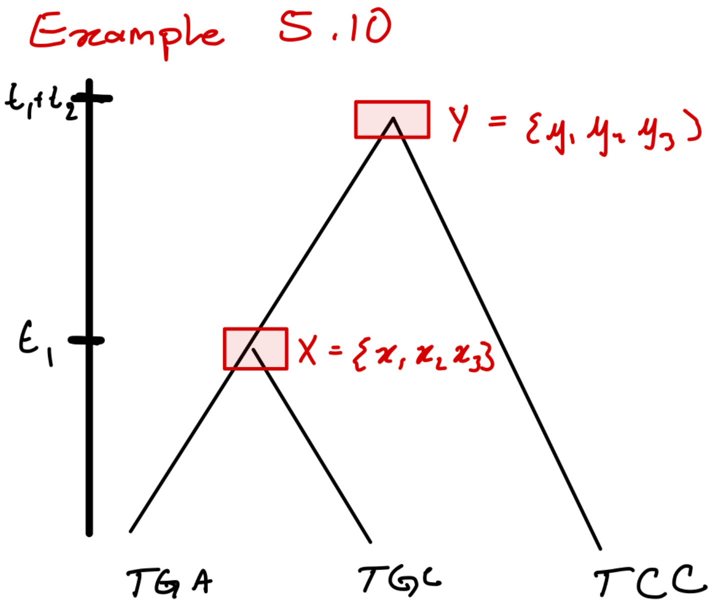
Example: 5.10: Tree Likelihood 2

Phython: Lecture5_3.ipynb

Consider the following phylogenetic tree. What is the likelihood of observing the following phylogenetic tree?

Newick Format

$((TGA:t_1, TGC:t_1):t_2, TCC)$



When considering multiple sites/bases, we assume that each of them evolves independently such that:

$$\begin{aligned}\Pr(((TGA:t_1, TGC:t_1):t_2, TCC)) = &\Pr(((T:t_1, T:t_1):t_2, T)) \\ &\times \Pr(((G:t_1, G:t_1):t_2, C)) \\ &\times \Pr(((A:t_1, C:t_1):t_2, C))\end{aligned}$$

Lecture 5.4 Simulations and Stochastic Mapping

Finish This

Simulating Evolution on a Tree

Here we discussed how to simulate genome sequences that are consistent with a given **tree topology** (the branching pattern and branch lengths of a tree). But, as in the coalescent, we often know the genome sequences at the present day but want to understand who the ancestors are and particular trajectories of mutations that may have given rise to the OBSERVED sequences.

Stochastic Mapping

- Method from Nielsen 2002

Stochastic mapping is a method for 1) simulating molecular evolution on a tree and 2) inferring the ancestral sequence at internal nodes in the tree.

Example: 5.11: Ancestral Reconstruction

1. Consider the following tree, propose a likely ancestor at node X and node Y? How much does each tip in the tree inform the likely sequence of each ancestor?

Conclusion: Each tip in the tree contains information about each node in the tree proportional to the inverse of the distance between them.

2. Where could mutations have happened in the tree to be consistent with the state of the tips?

Stochastic mapping provides a method for answering the above questions in a manner that is consistent with a given model of molecular evolution. This procedure has four steps:

Step 1: Post-order traversal calculation of node probabilities

- When looking at probabilities and stochastic processes on trees we consider two different types of calculations: **post-order traversal** and **pre-order traversal**.

Pre-order traversal: calculations that proceed from the root through the descants to the tips.
Post-order traversal:calculations that proceed from the tips through the ancestors to the root.

Let k be an index over all the internal nodes in a tree.
Let D_k be the data (sequences) of all the descendant tips from node k
Let Y_k be the (unobserved) state at internal node k

First e calculate $\Pr(D_k|Y_k = i) = f_{k,i}$

Example: 5.11: Ancestral Reconstruction

3. Calculate $f_{X,i}$ and f_{Y_i} What information is being used to do these calculations?.

Step 2: Root Probability

At the root node r calculate the probability

$$\Pr(Y_r = i|D_r = D) = \frac{f_{r,i}\pi_i}{\sum_j f_{r,j}\pi_j}$$

where π_j is the stationary distribution of the molecular evolution model or, if applicable, also equal to the empirical distribution of states.

Note that at the root D_r is all the data so we can also just call this D.

Example: 5.11: Ancestral Reconstruction

4. What is the root node in this case?

5. Calculate $\Pr(Y_r = i|D_r = D)$ for the root node. What information is being used here?

Step 3: Pre-order traversal calculation of node probabilities

Now we travel down the tree calculating the probability that each node is in a given state given the ancestor.

Notationally, for a node $k-1$ let node k be its direct ancestor.
Let t_{k-1} be the length of the branch between nodes $k-1$ and k

$$\Pr(Y_{k-1} = j|Y_k = i, D) = \frac{f_{k-1,j}P_{i,j}(t_{r-1})}{\sum_k f_{k-1,k}P_{i,k}(t_{r-1})}$$

Example: 5.11: Ancestral Reconstruction

6. What is/are the internal nodes in this case and who are their direct ancestors.

7. Calculate $\Pr(Y_{k-1} = j|Y_k = i, D) \quad \forall \text{ internal nodes}$

Step 4: Simulation of ancestral states and mutations

Step 4A: Given $\Pr(Y_r = i|D_r = D)$ and $\Pr(Y_{k-1} = j|Y_k = i, D)$, simulate the ancestor at each internal node.

Example: 5.11: Ancestral Reconstruction

8. **Simulate the state of the internal nodes in the tree.**

Step 4B: Given the states at the nodes and the tips, simulate mutational trajectories along the branches that are consistent with the tip states.