Phylogenetics: Trees and Splits, Basic Types of Models

Jiayue Qi [1]

Combinatorial and Algebraic Statistics (Spring 2021), KTH, Stockholm

2021.04.30.

[1] The author is supported by the Austrian Science Fund (FWF): W1214-N15, project DK9.

Phylogenetics

- Phylogenetics is a subject in mathematical biology.
- The goal is to construct the evolutionary tree from data about extant species.
- Extant species correspond to leaves.
- Ancestral species correspond to internal (non-leaf) vertices.
- Phylogeny of these species will be described by a rooted tree.

For the convenience of further discussion, let us come to some basic concepts of the type of trees that we will focus on.

X-tree

- X: set of labels.
- $\mathcal{T} = (V, E)$: a tree.
- $\phi : X \to V$ a map from label set to vertices of tree T.
- The pair $T = (\mathcal{T}, \phi)$ is called an *X-tree* if each vertex of degree 1 or 2 is in the image of ϕ.

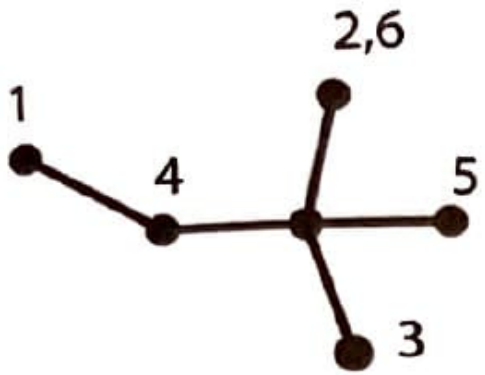


Figure: A [6]-tree T, where $[6] := \{1, 2, 3, 4, 5, 6\}$. [1, Figure 15.1.1.]

binary phylogenetic X-tree

- X: set of labels.
- $\mathcal{T} = (V, E)$: a tree.
- $\phi : X \to V$ a map from label set to vertices of tree T.
- The pair $T = (\mathcal{T}, \phi)$ is called a *phylogenetic X-tree* if each leaf has exactly one label and no label are assigned to non-leaves, i.e., the image of ϕ is the leaf set and ϕ is injective.
- My understanding: a phylogenetic X-tree is not necessarily an X-tree, since there can be inner vertices of degree 2 in the tree.
- A **binary phylogenetic X-tree** is a phylogenetic X-tree where each nonleaf vertex has degree 3.
- Note that in a rooted tree, the root is assigned with a special label ρ. And we always draw the edges as directed ones away from the root.

rooted binary phylogenetic X-tree

- Rooted binary phylogenetic X-tree?
  - Each leaf has one label, the root has a special label.
  - There is no label elsewhere assigned.
  - All vertices except for the root has degree 3.
  - The root has degree 2.
- The phylogenetic models that will be discussed today are all focusing on constructing a rooted binary phylogenetic X-tree (but maybe with extra labels attached to inner vertices).

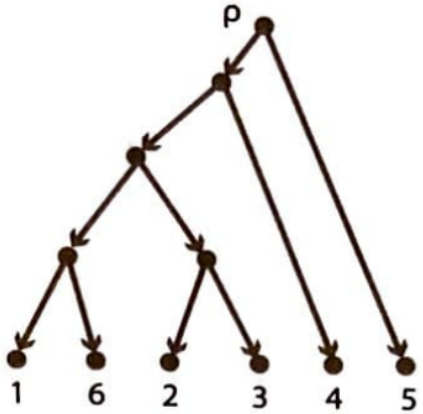


Figure: A rooted binary phylogenetic [6]-tree. [1, Figure 15.1.1.]

"splits representation"

- A *split* $A \mid B$ of X is a bi-partition of X into two disjoint non-empty sets.
- A split is *valid* for the X-tree T if it can be obtained by removing an edge of T and collecting labels in the two connected components respectively, forming the sets A and B.
- We usually denote the set of all valid splits of T as $\Sigma(T)$.

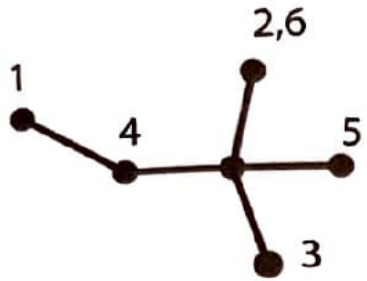


Figure: A $[6]$-tree T, where $[6] := \{1, 2, 3, 4, 5, 6\}$. [1, Figure 15.1.1.]

"splits representation"

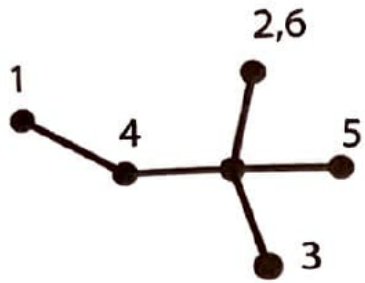


Figure: A [6]-tree T, where $[6] := \{1, 2, 3, 4, 5, 6\}$. [1, Figure 15.1.1.]

- $\Sigma(T) = \{1 \mid 23456, 14 \mid 2356, 26 \mid 1345, 3 \mid 12456, 5 \mid 12346\}$
- Every set of splits can be obtained from some X-tree?

"splits representation"

- Every set of splits can be obtained from some X-tree?
- No! Only those where splits are **pairwise compatible**.
- A pair of splits $A_1 \mid B_1$, $A_2 \mid B_2$ are *pairwise compatible* if at least one of the sets $A_1 \cap A_2$, $A_1 \cap B_2$, $B_1 \cap A_2$, $B_1 \cap B_2$ is empty.
- There is a one-to-one correspondence between the set of X-trees and the set of pairwise compatible splits of X.
- If T is an X-tree, then $\Sigma(T)$ is a pairwise compatible set of splits. Proof: on the next slide.
- The converse is also true, namely there exists a unique X-tree T such that $\Sigma(T) = \Sigma$ for any pairwise compatible slits set Σ.

"splits representation"

Proposition

If T is an X-tree, then $\Sigma(T)$ is a pairwise compatible set of splits.

Proof.

Proof: Let us refer to [2, Proposition 22]. Let $A_1 \mid B_1$, $A_2 \mid B_2$ be two splits in $\Sigma(T)$. They correspond to two edges, say e_1 and e_2. When we remove these two edges, we obtain three connected components. Each of the four sets $A_1 \cap A_2$, $A_1 \cap B_2$, $B_1 \cap A_2$, $B_1 \cap B_2$ equals to one of the leaf sets of these three compoents, if not empty. Also, it is not hard to check that these four sets are pairwise disjoint. Therefore, at least one of them must be empty

"splits representation"

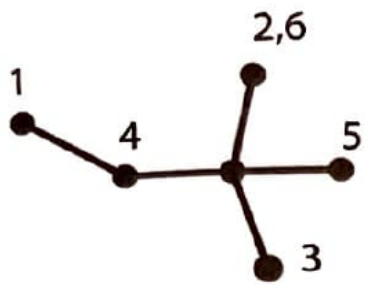
- The converse is also true, namely there exists a unique X-tree T such that $\Sigma(T) = \Sigma$ for any pairwise compatible slits set Σ.
- The book offers a "tree-growing" construction, let me present another one that is given in [2, Definition 2.4], on our running example.
- $\Sigma = \{1 \mid 23456, 14 \mid 2356, 26 \mid 1345, 3 \mid 12456, 5 \mid 12346\}$.
- First we pick any split, you decide?

"splits representation"

- $\Sigma = \{1 \mid 23456, 14 \mid 2356, 26 \mid 1345, 3 \mid 12456, 5 \mid 12346\}$.
- $14 \mid 2356$ was picked! First step is to create two vertices having $\{1, 4\}$ and $\{2, 3, 5, 6\}$ respectively as their labeling sets.
- Then collect all subsets of splits in Σ in one set, and pick from this set those that is either a subset of $\{1, 4\}$, or a subset of $\{2, 3, 5, 6\}$.
- Consider the chosen subsets, together with $\{1, 4\}$ and $\{2, 3, 5, 6\}$. We build a Hasse diagram from these sets, according to the set containment order. Each set is attached to a vertex (of the diagram) as the labeling set.
- Then we add an extra edge between vertices of $\{1, 4\}$ and $\{2, 3, 5, 6\}$.
- Finally we start from leaves of the current graph, delete from the labeling sets of their ancestral vertices those labels of the leaves.

"splits representation"

- Then we see that we indeed get the same [6]-tree as given in our running example!
- The usage of this equivalent representation of X-trees will not show up in the remaining discussion of today's lecture, but may be used in some other places, say next week's lecture.
- With this, we conclude the first part of today's lecture.



phylogenetics: basic idea

- To describe the evolutionary history of sequence data over time.
- Associate a sequence S_i to each vertex i of a rooted tree $\mathcal{T}$.
- The sequences S_i are on some alphabet, for instance the DNA bases $\{A, C, G, T\}$, in which case the sequences are fragments of DNA and the vertex i corresponds to a species that has that fragment in the corresponding gene.
- Given some sequences for the extant species, we aim to build such a tree, where the root carries the most ancestral sequence, and all other sequences in the tree have descended from it.
- We usually do not have information on the sequences of ancestral species, so the sequences for the internal vertices are viewed as hidden variables.

two assumptions for phylogenetic models

- **Only point mutations:** at a random point in time, a change to a single position in a sequence occurs. Also, the point mutations happen independently at each position in the sequence.
- A diagram of DNA sequences as shown in the figure below is called an *alignment*. In our consideration, all sequences have the same length, because of the first assumption.
- **Site independence:** each column of the alignment has the same underlying probability distribution.
- **Goal:** figure out which rooted tree produce the given alignment, best explaining the observed data.

Human:	ACCGTGCAACGTGAACGA
Chimp:	ACCTTGGAAGGTAAACGA
Gorilla:	ACCGTGCAACGTAAACTA

Figure: [1, Page 340]

phylogenetic methods

- Clustering algorithms: create a phylogeny with no need to test all possible trees.
  - Distance methods
- Optimality Criteria: must test all possible trees using search algorithm and give each a score (criterion).
  - Maximum Parsimony
  - Minimum Distance
  - Maximum Likelihood
  - Bayesian Probability

phylogenetic methods

- Clustering algorithms: create a phylogeny with no need to test all possible trees.
  - Distance methods
- Optimality Criteria: must test all possible trees using search algorithm and give each a score (criterion).
  - Maximum Parsimony
  - Minimum Distance
  - Maximum Likelihood
  - Bayesian Probability

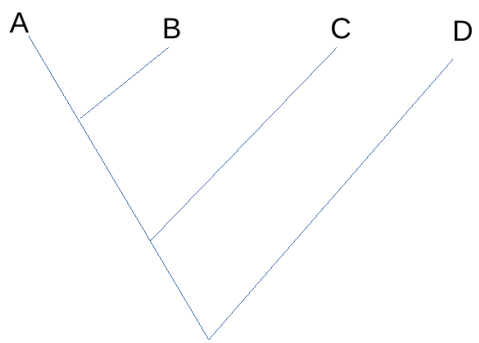
maximum parsimony

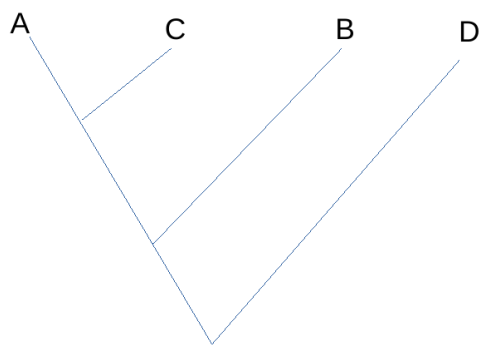
- The principle of **parsimony**: a theory should provide the simplest possible explanation for a phenomenon.
- Given the alignment for sequences at the leaves.
- Need to determine on the topology of the tree, the sequences at the inner vertices, so that it requires minimum number of mutations to explain the given sequences at the leaves.
- Consider a given alignment. Actually we should go through all possible tree structures with the sequences marked at the leaves, and try out all possible sequences for the inner vertices, then find a construction such that least number of mutations is needed.
- NP-hard! No known polynomial time algorithms yet.

maximum parsimony: an example

- For simplicity, in the upcoming example[2], we fix one species as the outgroup, and only consider two tree topologies, to have an intuitive idea of the method.
- Given an alignment as below. Let species D be the outgroup, consider the following two tree topologies.
- Illustrate the process on the whiteboard.

Nucleotide Position / Taxon	1	2	3
A	G	G	G
B	G	T	G
C	T	G	T
D	T	T	T



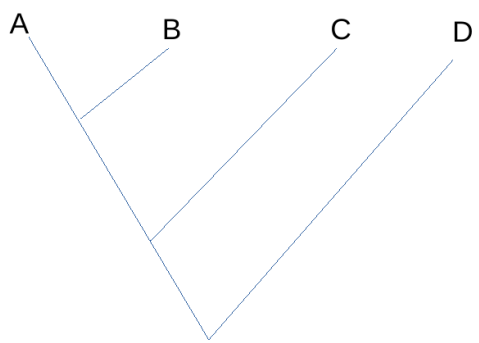


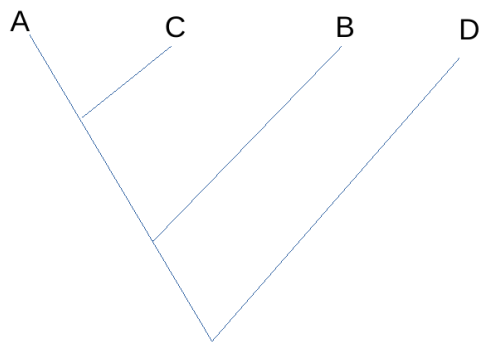
[2]This example is taken from the video https://www.youtube.com/watch?v=gXb_WuLCD8g&t=365s

maximum parsimony: an example

- For the first and third nucleotide position, the minimum number of mutations on the left tree are both 1, while those on the right tree are both 2.
- For the second nuceotide position, the minimum number of mutations on the left tree is 2, while that on the right tree is 1.
- Consider the number of mutations of all three positions, it is $1 + 2 + 1 = 4$ for the left tree, and $2 + 1 + 2 = 5$ for the second tree. Therefore, we choose the first tree as the most parsimonious one.

Nucleotide Position / Taxon	1	2	3
A	G	G	G
B	G	T	G
C	T	G	T
D	T	T	T

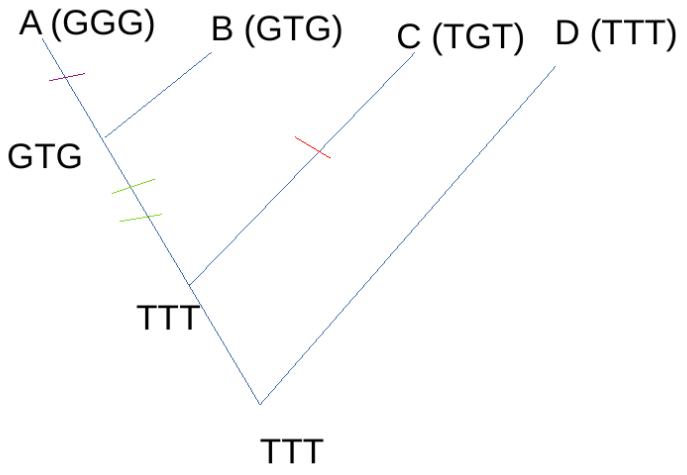




maximum parsimony: an example

- The final tree, with evolutionary history illustrated.

Taxon \ Nucleotide Position	1	2	3
A	G	G	G
B	G	T	G
C	T	G	T
D	T	T	T



- Mximum parsimony method assumes that the probability of mutation is way smaller than the probability of staying in the same nucleobase. (“mutations are rare”)
- With this method, we figured out a tree where the probability that we obtain the observed sequences is the highest, i.e., a tree that best explains the observed data.

Three-leaf tree

- Consider the tree depicted below: The labels of the internal vertices Y_1, Y_2 are hidden variables, while those of leaves X_1, X_2, X_3 are observed random variables.
- Recall the DAG model associated to this tree.

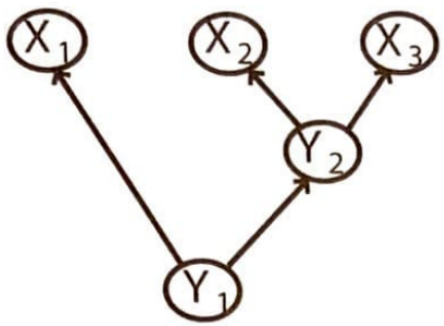
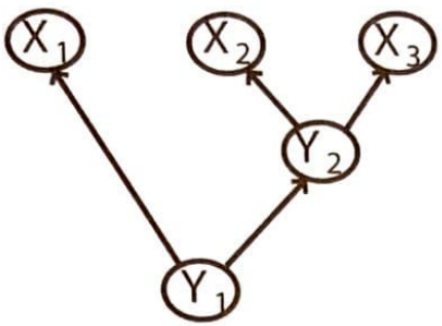


Figure: [1, Figure 15.2.1]

Three-leaf tree

- The probability of observing $X_1 = x_1$, $X_2 = x_2$, $X_3 = x_3$ can be computed by the formula in the figure below, where $P_{Y_2|Y_1}(y_2 \mid y_1) := P(Y_2 = y_2 \mid Y_1 = y_1)$. ($[\kappa]$: state space)
- Markov property! The sequence at node v only depends on the sequence at its parent vertex u.



$$P(x_1, x_2, x_3) = \sum_{y_1 \in [\kappa]} \sum_{y_2 \in [\kappa]} P_{Y_1}(y_1) P_{Y_2|Y_1}(y_2 \mid y_1) \times P_{X_1|Y_1}(x_1 \mid y_1) P_{X_2|Y_2}(x_2 \mid y_2) P_{X_3|Y_2}(x_3 \mid y_2),$$

Figure: [1, Example 15.2.1]

continuous time Markov chain

- The conditional probability distributions that appear in the previous directed graphical model is a square matrix of size $\kappa \times \kappa$, which can be viewed as the transition matrix of Markov chains. (note that the state space $[\kappa] := \{1, \ldots, \kappa\}$)
- The usual approach in phylogenetics: consider the model arising from a *continuous time Markov process*.
- We specify such a model by giving the *rate matrix*, which is usually denoted by Q.
- $Q \in \mathbb{R}^{\kappa \times \kappa}$ describes a continuous time Markov chain if it fulfills the following condition: $q_{ij} \geq 0$ for all $i \neq j$, and $\sum_{j=1}^{\kappa} q_{ij} = 0$.

continuous time Markov chain

- An intuitive explanation of the rate matrix:
  - If the process is in state i, we need to wait an exponentially distributed amount of time with parameter $-q_{ii}$ (usually denoted by q_i) until the next substitution.
  - "How long you already waited doesn't mean anything..."
  - When the state change happens, the change from i to j has probability $\frac{q_{ij}}{q_i}$.
- We can compute the probability matrix (transition matrix) from rate matrix by

$$P(t) = exp(Qt) = I + Qt + \frac{Q^2t^2}{2!} + \frac{Q^3t^3}{3!} + \cdots$$

- Each edge e has a parameter t_e called *branch length*, $P(t_e)$ is the transition matrix of edge e.
- Intuitively, the branch length indicates the time duration between two major speciation events (denoted by vertices).

continuous time Markov chain

- Note that the transition matrix (probability matrix) is a function in t, while the probability $\frac{q_{ij}}{q_i}$ of changing from state i to j is a different concept.
- The latter refers to that probability when the mutation/substitution really happens, excluding the consideration of staying at state i.

continuous time Markov chain (CTMC)

- My understanding on the function of transition matrix in phylogenetics:
  - Mximum likelihood method: once we have the transition matrix, given some data at the leaves, we can then try all possible tree topologies, all possible branch lengths, all possible data for the ancestral species, then find a setting that maximize the probability of our observed data.
  - Do simulations and construct a phylogeny tree: when all parameters (mentioned in the last subitem) are fixed, we can do simulations, constructing an evolutionary phylogeny.
- To specify a continuous time Markov chain, it suffices to give the rate matrix.

phylogenetic models: CFN, JC69

- Cavender-Farris-Neyman model, the the idea is to group the nucleobases into two groups — purine $\{A, G\}$ and pyrimidine $\{C, T\}$ — and just consider the transitions between the two groups.
- The simplest model for DNA bases — Jukes-Cantor model (JC69): once a mutation happens, it has equal probability to change to the any of the other three bases.

phylogenetic models: CFN, JC69

Figures below are taken from [1, Page 343]: the form of rate matrices of CFN model, the form of the transition matrices of CFN model, the form of rate matrices of JC69 model. ($\alpha \in \mathbb{R}^+$)

$$Q^{CFN} = \begin{pmatrix} -\alpha & \alpha \\ \alpha & -\alpha \end{pmatrix} \qquad P(t) = \exp(Q^{CFN}t) = \begin{pmatrix} \frac{1+\exp(-2\alpha t)}{2} & \frac{1-\exp(-2\alpha t)}{2} \\ \frac{1-\exp(-2\alpha t)}{2} & \frac{1+\exp(-2\alpha t)}{2} \end{pmatrix}$$

$$Q^{JC} = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}$$

phylogenetic models: K2P, K3P

- However biologically, transitions within the purine group or within the pyrimidine group is way more than those from one group to another. Kimura tried to improve the model based on this fact, introducing Kimura 2-parameter and Kimura 3-parameter models.
- Figures below are taken from [1, Page 343 344]: the forms of rate matrices of K2P model and those of K3P model.

$$Q^{K2P} = \begin{pmatrix} -\alpha-2\beta & \beta & \alpha & \beta \\ \beta & -\alpha-2\beta & \beta & \alpha \\ \alpha & \beta & -\alpha-2\beta & \beta \\ \beta & \alpha & \beta & -\alpha-2\beta \end{pmatrix} \qquad Q^{K3P} = \begin{pmatrix} -\alpha-\beta-\gamma & \beta & \alpha & \gamma \\ \beta & -\alpha-\beta-\gamma & \gamma & \alpha \\ \alpha & \gamma & -\alpha-\beta-\gamma & \beta \\ \gamma & \alpha & \beta & -\alpha-\beta-\gamma \end{pmatrix}$$

CTMC: stationary distribution

- Recall one step of the maximum likelihood method mentioned earlier.
- Once we have the transition matrix, given some data at the leaves, we can take a random tree topology, random branch lengths for edges, go through all possible ancestral sequences at the inner vertices, then compute the probability of obtaining the observed data.
- What is missing?
- Probability for the sequence at the root vertex is missing!
- So this would be the stationary distribution of CTMC.

CTMC: stationary distribution

- An intuitive idea: start from any distribution μ, when t goes to infinity, $\mu \cdot P(t)$ converges to π, where π is the stationary distribution.
- We imagine a really long branch stops at the root vertex, hence the probability for the sequence at root is given by the stationary distribution.

CTMC: stationary distribution

- For example, consider the JC69 model rate matrix (the left figure), if we do simulations with the time parameter rather long.
- Starting say from A, out of 10000 samples, we may get 2500 samples respectively with the state A, C, G or T.
- So the the matrix for $P(t)$ when $t \to \infty$ would be as given in the right figure.
- But this intuitive idea might not hold for all models, it holds for *irreducible, aperiodic* Markov chains!

-1	1/3	1/3	1/3
1/3	-1	1/3	1/3
1/3	1/3	-1	1/3
1/3	1/3	1/3	-1

To / From	A	C	G	T
A	1/4	1/4	1/4	1/4
C	1/4	1/4	1/4	1/4
G	1/4	1/4	1/4	1/4
T	1/4	1/4	1/4	1/4

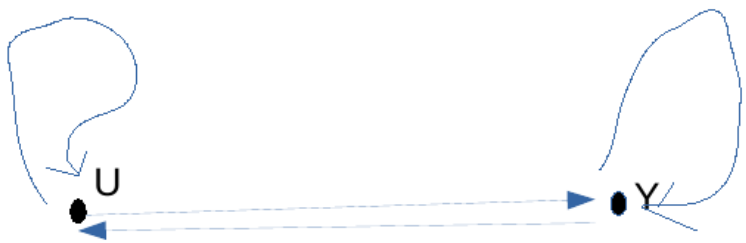
Perron-Frobenius theorem

- Let each state be a vertex, draw an edge $i \to j$ if $p_{ij} > 0$ in the transition matrix.
- The Markov chain is *irreducible* if there is a directed path between any ordered pair (i, j).
- The Markov chain is *aperiodic* if the greatest common divisor of cycle lengths is 1.
- Let Q be a rate matrix s.t. $P(t) = exp(Qt)$ describes an irreducible, aperiodic Markov chain. Then there is a unique probability distribution $\pi \in \Delta_{\kappa-1}$ such that $\pi \cdot P(t) = \pi$, i.e., $\pi \cdot Q = 0$. ($\Delta_{\kappa-1}$: the probability simplex)
- This distribution is called the *stationary distribution* of the considered Markov chain.

Perron-Frobenius theorem

- See the left figure below for the corresponding graph of the CFN model, where U denotes the purine group and Y denotes the pyrimidine group.
- Based on the probability matrix (see the left figure) — recall that $\alpha \in \mathbb{R}^+$ — this graph is not hard to compute.
- CFN model: irreducible, aperiodic Markov chain.
- By Perron-Frobenius theorem, it has a unique stationary distribution.

$$P(t) = \exp(Q^{CFN}t) = \begin{pmatrix} \frac{1+\exp(-2\alpha t)}{2} & \frac{1-\exp(-2\alpha t)}{2} \\ \frac{1-\exp(-2\alpha t)}{2} & \frac{1+\exp(-2\alpha t)}{2} \end{pmatrix}$$



Perron-Frobenius theorem

- An easy computation tells us that $\pi = (\frac{1}{2}, \frac{1}{2})$ satisfies $\pi \cdot Q = 0$, hence it is the unique stationary distribution of the considered Markov chain.
- With the similar method, we should be able to get that JC69,K2P,K3P all have $\pi = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ as their stationary distribution.

non-uniform stationary distribution models

- Biologically, Cs and Gs are generally rarer than As and Ts. Therefore, it makes more sense to study models where the stationary distribution is non-uniform.
- Felsenstein model (F81), Hasegawa-Kishino-Yano model (HKY85).
- $*$ is to make sure that row sum equals zero, α in HKY model is to allow adjusting the probability of exchanging within purine versus pyrimidine.
- Figures below are taken from [1, Page 345]: the forms of rate matrices of F81 model and those of HKY model.

$$Q^{F81} = \begin{pmatrix} * & \pi_C & \pi_G & \pi_T \\ \pi_A & * & \pi_G & \pi_T \\ \pi_A & \pi_C & * & \pi_T \\ \pi_A & \pi_C & \pi_G & * \end{pmatrix}$$

$$Q^{HKY} = \begin{pmatrix} * & \pi_C & \alpha\pi_G & \pi_T \\ \pi_A & * & \pi_G & \alpha\pi_T \\ \alpha\pi_A & \pi_C & * & \pi_T \\ \pi_A & \alpha\pi_C & \pi_G & * \end{pmatrix}$$

time reversibility

- Time reversibility: the amount of change from state x to y is equal to the amount of change from y to x.
- Mathematically: $\pi_i \cdot Q_{ij} = \pi_j \cdot Q_{ji}$.
- The general time-reversible model (GTR) depicts a general form of the rate matrix whenever the model is time reversible.
- The figure below is taken from [1, Page 345]: the form of rate matrices of GTR model.

$$Q^{GTR} = \begin{pmatrix} * & \pi_C\alpha & \pi_G\beta & \pi_T\gamma \\ \pi_A\alpha & * & \pi_G\delta & \pi_T\epsilon \\ \pi_A\beta & \pi_C\delta & * & \pi_T\zeta \\ \pi_A\gamma & \pi_C\epsilon & \pi_G\zeta & * \end{pmatrix}$$

some Maths considerations?

- Usually the rate matrix is fixed for all edges of the evolutionary tree, but it is probably unreasonable across large evolutionary distances.
- Let us consider the situation where different edges have different rate matrices.
- Consider a degree-two vertex in the tree, $exp(Q_1 t_1)$ and $exp(Q_2 t_2)$ are the two transition matrices corresponding to its two incident edges.
- Naturally we want some consistency when this vertex gets suppressed, i.e., we would like $exp(Q_1 t_1) \cdot exp(Q_2 t_2)$ to be able to be expressed as $exp(Q_3(t_1 + t_2))$.
- So which set of rate matrices can satisfy the above mentioned property? We need some Lie algebra.

Lie Markov models

- A set $L \subset \mathbb{K}^{\kappa \times \kappa}$ is a *matrix Lie algebra* if L is a $\mathbb{K}$-vector space and for all $A, B \in L$, $[A, B] \in AB - BA \in L$.
- A set of rate matrices $\mathcal{L}$ is called a *Lie Markov model* if it is a Lie algebra.
- Theorem 15.2.5 of [1]: let $\mathcal{L} \in \mathbb{R}^{\kappa \times \kappa}$ be a collection of rate matrices. If $\mathcal{L}$ is a Lie Markov model, then for any $Q_1, Q_2 \in \mathcal{L}$ and t_1, t_2, there is a $Q_3 \in \mathcal{L}$ such that

$$exp(Q_1 t_1) \cdot exp(Q_2 t_2) = exp(Q_3(t_1 + t_2)).$$

- CFN, JC69, K2P, K3P, F81 are all Lie Markov models.
- HKY, GTR are not Lie Markov models.

References


Sullivant Seth.
Algebraic statistics. Vol. 194. American Mathematical Soc., 2018.


Qi Jiayue and Josef Schicho.
Five Equivalent Ways to Describe a Phylogenetic Tree. arXiv preprint arXiv:2011.11774 (2020).

Thank You