Evolutionary models

Ruriko Yoshida
Dept. of Statistics University of Kentucky

Discrete time Markov chain

There are your favorite bars, $S = \{A, C, G, T\}$, in Lexington. You and your friends, Arne, Connie, Dave, decide to go to bar hopping one night. Since you are already drunk, you and your friends decide where you and your friends are going next by rollong a four faced die (tetrahedron).

Each of you and your friends has a four faced die. You will roll your die if you and your friends are currently at Bar A, Arne rolls his die if you and your friends are currently at Bar C, Connie rolls her die if you and your friends are currently at Bar G, and Dave rolls his die if you and your friends are currently at Bar T.

Each die has different weights on each face.

Some example...

The probability of obtaining each letter differs depending on which die you are rolling:

	A	C	G	T
Your die	P_{AA}	P_{AC}	P_{AG}	P_{AT}
Arne's die	P_{CA}	P_{CC}	P_{CG}	P_{CT}
Connie's die	P_{GA}	P_{GC}	P_{GG}	P_{GT}
Dave's die	P_{TA}	P_{TC}	P_{TG}	P_{TT}

Where P_{xy} for any x, y in S, $P_{AA} + P_{AC} + P_{AG} + P_{AT} = 1$, $P_{CA} + P_{CC} + P_{CG} + P_{CT} = 1$, $P_{GA} + P_{GC} + P_{GG} + P_{GT} = 1$, and $P_{TA} + P_{TC} + P_{TG} + P_{TT} = 1$.

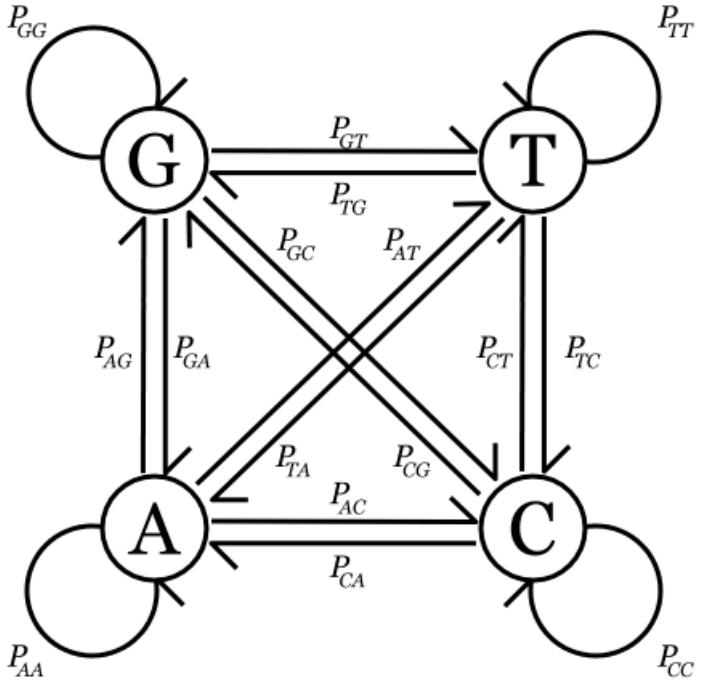
Some example...

Here is a specific example:

	A	C	G	T
Your die	$1/4$	$1/4$	$1/4$	$1/4$
Arne's die	$1/5$	$1/5$	$2/5$	$1/5$
Connie's die	$1/3$	$1/3$	$1/6$	$1/6$
Dave's die T	$1/6$	$1/3$	$1/3$	$1/6$

We can describe this process by drawing a picture...

Some example...



This is an example of **Discrete Time Markov process**.

Some definitions on MC

Definition A **discrete time stochastic process** is a collection of random variables $\{X_0, X_1, X_2, \ldots\}$ defined on a common sample space and state space S which depends on time $n = 0, 1, 2, \cdots$.

Definition A **discrete time Markov process** is a discrete time stochastic process $\{X_n\}_{n=0}^{\infty}$ which satisfies the **Markov property**, that is, for all $n \in \{0, 1, \ldots\}$ and any states $x_0, x_1, \cdots x_n, y \in S$,

$$P(X_{n+1} = y | X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = P(X_{n+1} = y | X_n = x_n).$$

Definition A **time homogeneous Markov process** $\{X_n\}_{n=0}^{\infty}$ is a stochastic process such that for all $n \in \{0, 1, \ldots\}$ and any states $x, y \in S$,

$$P(X_{n+1} = y | X_n = x) = P(X_1 = y | X_0 = x).$$

Finite State MC

Definition Given a Markov chain with finite state space S, a **transition matrix** is a matrix whose entry in the ith row and jth column is

$$P(X_{t+1} = j | X_t = i).$$

Definition Given a Markov chain with finite state space S, the **transition graph** has vertex set S, and has directed edges (i, j) with weight

$$P(X_{t+1} = j | X_t = i)$$

whenever this weight is positive.

Definition We call a Markov process **Markov chain** if we can describe the process as the transition graph.

Go back to our example...

Since the probability where you are going next depends only on the place where you are currently at, this process satisfies Markov property.

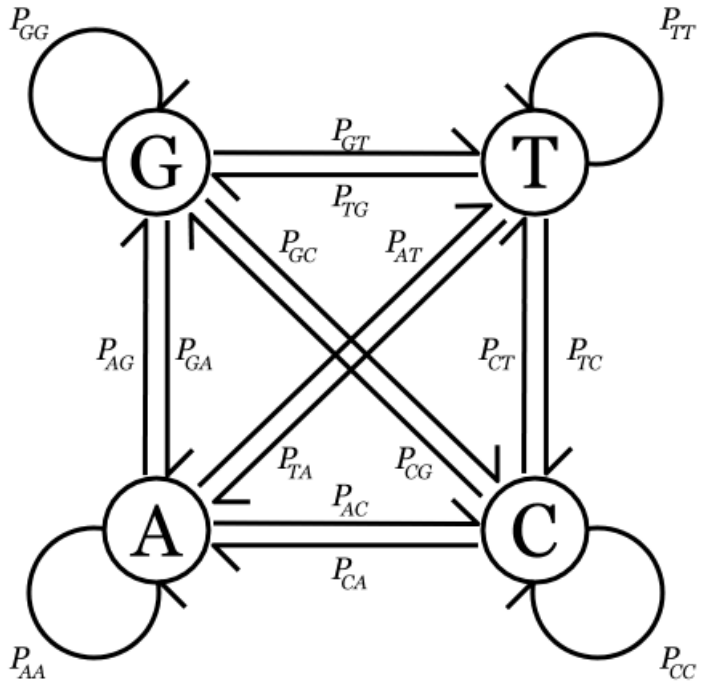
Since the probability where you are going next is not affected by the time this process is time homogeneous.

The transition martix of our Markov chain is:

	A	C	G	T
Your die	$1/4$	$1/4$	$1/4$	$1/4$
Arne's die	$1/5$	$1/5$	$2/5$	$1/5$
Connie's die	$1/3$	$1/3$	$1/6$	$1/6$
Dave's die T	$1/6$	$1/3$	$1/3$	$1/6$

Some example...

The transition graph is:



This is an example of **Discrete Time Markov chain**.

Question

Question You and your friends decide the initial bar by rolling a fair die (X_0). Then You and your friends move twice ($n = 2$). What is the probability that you and your friends have been at Bar C at $n = 0$, at Bar T at $n = 1$, and at Bar A at $n = 2$?

Review Suppose A_1, A_2 are events such that $P(A_2) \neq 0$. Then the **conditional probability** of event A_1 given A_2 is:

$$P(A_1|A_2) = \frac{P(A_1A_2)}{P(A_2)}.$$

Thus we have

$$P(A_1|A_2)P(A_2) = P(A_1A_2).$$

Computing probability

By this way we can compute the probability of a path, $x_0x_1x_2\cdots x_nx_{n+1}$, in a graph by:

$$\begin{aligned} & P(X_{n+1}=x_{n+1},X_n=x_n,\cdots,X_1=x_1,X_0=x_0) \\ = \quad & P(X_{n+1}=x_{n+1}|X_n=x_n,\cdots,X_1=x_1,X_0=x_0) \\ & \times P(X_n=x_n|X_{n-1}=x_{n-1},\cdots,X_1=x_1,X_0=x_0) \\ & \times\cdots\times P(X_1=x_1|X_0=x_0)\times P(X_0). \end{aligned}$$

Note that this is a discrete time Markov chain so it satisfies Markov property, thus we have:

$$\begin{aligned} & P(X_{n+1}=x_{n+1},X_n=x_n,\cdots,X_1=x_1,X_0=x_0) \\ = \quad & P(X_{n+1}=x_{n+1}|X_n=x_n)\times P(X_n=x_n|X_{n-1}=x_{n-1}) \\ & \times\cdots\times P(X_1=x_1|X_0=x_0)\times P(X_0). \end{aligned}$$

Go back to the example...

We want to compute the probability that you and your friends have been at Bar C at $n = 0$, at Bar T at $n = 1$, and at Bar A at $n = 2$. So we want to compute

$$P(X_2 = A, X_1 = T, X_0 = C).$$

By using conditional probability and Markov property we have

$$\begin{aligned} & P(X_2 = A, X_1 = T, X_0 = C) \\ = \quad & P(X_2 = A | X_1 = T) P(X_1 = T | X_0 = C) P(X_0 = C) \end{aligned}$$

Now use the transition martix of our Markov chain:

	A	C	G	T
Your die A	$1/4$	$1/4$	$1/4$	$1/4$
Arne's die C	$1/5$	$1/5$	$2/5$	$1/5$
Connie's die G	$1/3$	$1/3$	$1/6$	$1/6$
Dave's die T	$1/6$	$1/3$	$1/3$	$1/6$

Then we have:

$$\begin{aligned} & P(X_2 = A, X_1 = T, X_0 = C) \\ = \; & P(X_2 = A | X_1 = T) P(X_1 = T | X_0 = C) P(X_0 = C) \\ = \; & P_{TA} \times P_{CT} \times P(X_0) \\ = \; & 1/6 \times 1/5 \times 1/4 \end{aligned}$$

How about...

Question You and your friends decide the initial bar by rolling a fair die (X_0). Then You and your friends move five times ($n = 5$). What is the probability that you and your friends have been at Bar A at $n = 5$?

Fact $P(X_n = y|X_0 = x)$ for states x, y is the (x, y)th element of the matrix P^n where P is the transition matrix.

P^5 is:

0.24352	0.27621	0.28410	0.19617
0.24346	0.27611	0.28421	0.19621
0.24357	0.27630	0.28400	0.19613
0.24352	0.27622	0.28410	0.19616

Thus the answer is

$$
\begin{aligned}
& P(X_5 = A) \\
= \quad & \textstyle\sum_{x \in S} P(X_5 = A | X_0 = x) P(X_0 = x) \\
= \quad & 0.24352 * 1/4 + 0.24346 * 1/4 + 0.24357 * 1/4 + 0.24352 * 1/4 \\
= \quad & 0.24352
\end{aligned}
$$

Definition π is a **stationary** or **invariant** distribution for a Markov chain if $X_t \sim \pi$ implies that $X_{t+1} \sim \pi$ (i.e. $\sum_{y \in S} \pi_y p(y, x) = \pi_x$).

Definition For countable state Markov chains, if

$$\sum_{x \in S} \pi_x P(X_{t+1} = y | X_t = x) = \pi_y$$

then π is a stationary distribution. These are called the **balance equations.**

Definition π is a **limiting** distribution for a countable state Markov chain if

$$\lim_{t \to \infty} P(X_t = i | X_0 = j) = \pi_i,$$

for all states i and j.

Note. The limiting distribution is a stationary distribution but not always true that a stationary distribution is the limiting distribution.

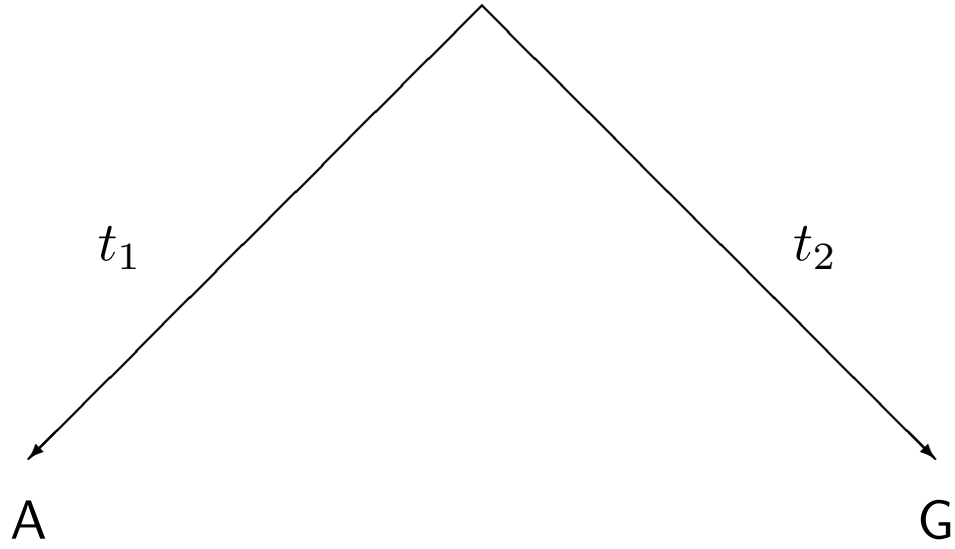
I will show a proof for a fact that the limiting distribution is a stationary distribution.

Example. $S = \{1, 2, 3\}$. If the transition matrix $P = I_3$, then any distribution is stationary but it does not have to be the limiting distribution.

Evolutionary Model

Pairwise sequences

Suppose we have a pair of sequences at a single site such that:



Assuming time reversibility.... (let $t = t_1 + t_2$)

Continuous time Markov chain (CTMC)

Suppose X_t, $t > 0$ is a stochastic process, **Markov property** states:

$$P[X_{t+h} = y | X_s = x_s, s \leq t] = P[X_{t+h} = y | X_t = x_t], \forall h > 0.$$

A stochastic process with the Markov property is usually called a **Markov process**. Markov processes are called (time-) homogeneous if

$$P[X_{t+h} = y | X_t = x_t] = P[X_h = y | X_0 = x_0], \forall t, h > 0.$$

A Markov process is called **time continuous** if

$$P[X_{t+h} = j | X_t = i] = q_{ij}h + o(h)$$

where q_{ij} is the rate of the transition probability and $o(h)$ is some constant in h such that $o(h) \to 0$ as $h \to 0$.

DTMC to CTMC

How can we obtain the transition probability for a CTMC?

We discrete the time interval $[0, T]$ into n pieces and send $n \to \infty$.

If we do then we have $h = T/n$.

From the definition above we have

$$P[X_{t+h} = j | X_t = i] = q_{ij}h + o(h)$$

Example: Poisson process

A **Poisson process** is a process $\{X_t\}$ satisfying:

- $X_0 = 0$.
- The number of events during one time interval does not affect the number of events during a different time interval.
- The average rate at which events occur remains constant.
- Events occur once at a time.

Example: Poisson process

From the definition we have the state space $\Sigma = \mathbb{Z}_+ = \{0, 1, 2, \cdots\}$.

Let $x \in \Sigma$ and $\lambda > 0$. With the time interval $(t, t+h)$ the probability to change from x to $x+1$ is $\lambda h + o(h)$ and the probability to stay at x is $1 - \lambda h + o(h)$.

We discretize the time interval $[0, t]$ into n pieces. Then we use Binomial distribution:

$$P(X_t = x | X_0 = 0) = \binom{n}{x} (\lambda \frac{t}{n})^x (1 - \lambda \frac{t}{n})^{n-x}.$$

If we send $n \to \infty$ then we get the distribution:

$$P(X_t = x | X_0 = 0) = \frac{(\lambda t)^x e^{-\lambda t}}{x!}$$

CTMC with a finite state space

How about a CTMC with a finite state space?

The idea to obtain the transition probability is the same. As soon as we have the rates for transitions we can compute the transition probability.

Definition: A **rate matrix** (or **infinitesimal matrix**) is a square matrix $Q = (q_{ij})$, with rows and columns indexed by the state space Σ.

Rate matrices must satisfy the following requirements:

$$q_{ij} \geq 0 \quad \text{for} \quad i \neq j,$$

$$\sum_{j \in \Sigma} q_{ij} = 0 \quad \text{for all} \quad i \in \Sigma,$$

$$q_{ii} < 0 \quad \text{for all} \quad i \in \Sigma.$$

Example

Consider the state space $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$.

The Jukes-Cantor matrix is the matrix

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

where $\alpha \geq 0$ is a parameter.

Transition probability

If we have a finite state space we can compute the transition probability matrix by

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

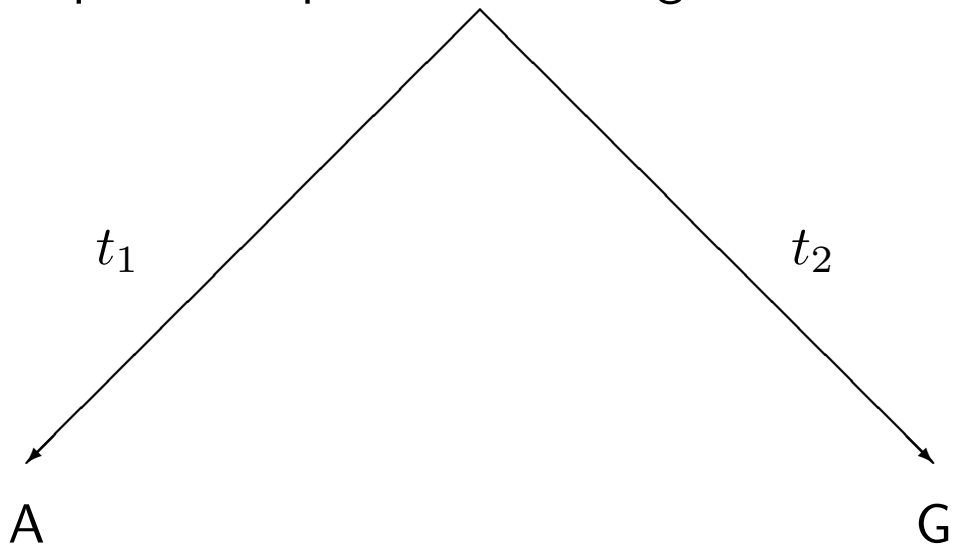
Example: For J-C model we have

$$P(t) = \frac{1}{4}\begin{pmatrix} 1+3e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} \\ 1-e^{-4\alpha t} & 1+3e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} \\ 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1+3e^{-4\alpha t} & 1-e^{-4\alpha t} \\ 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1+3e^{-4\alpha t} \end{pmatrix}.$$

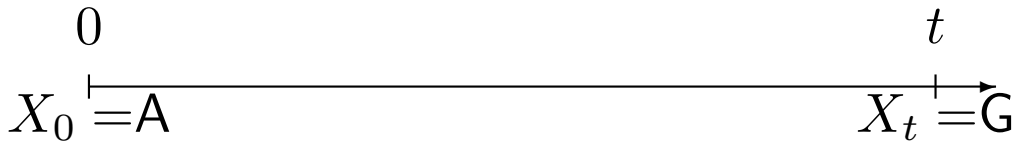
Note that if we send $t \to \infty$ then we get the limiting distribution π. Under the JC model we have $\pi = (1/4, 1/4, 1/4, 1/4)$.

Example

Suppose we have a pair of sequences at a single site under the JC model:



Assuming time reversibility.... (let $t = t_1 + t_2$)



Example

We estimate the initial distribution by the limiting distribution (because this is time reversible). Let P_{AG} be the probability to observe $X_0 = A$ and $X_t = G$.

$$\begin{aligned} & P_{AG} \\ = \; & P(X_0 = A,\, X_t = G) \\ = \; & P(X_0 = A) P(X_t = G | X_0 = A) \\ = \; & \tfrac{1}{4} \times \tfrac{1}{4} \left(1 - e^{-4\alpha t}\right) \end{aligned}$$

By this way we can compute the probability P_{ab} where for any pairs $a, b \in \Sigma = \{A, C, G, T\}$, namely:

$$\begin{aligned} P_{ab} = \; & \tfrac{1}{4} \times \tfrac{1}{4} \left(1 - e^{-4\alpha t}\right) \quad \text{if } a \neq b \\ P_{ab} = \; & \tfrac{1}{4} \times \tfrac{1}{4} \left(1 + 3e^{-4\alpha t}\right) \quad \text{if } a = b \end{aligned}$$

Example: Triples

How about we have the number of leaves $n = 3$?

To make it simple we consider the two state model: $\Sigma = \{0, 1\}$.

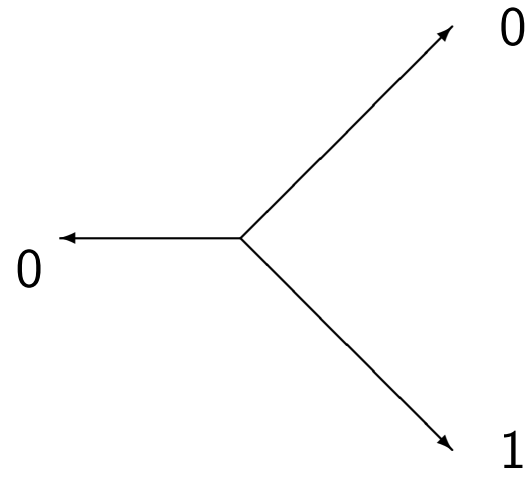


Figure 1: Tree with three leaves.

Example: Triples

Consider the JC model on $\Sigma = \{0, 1\}$. Then we have the rate matrix:

$$\begin{pmatrix} -\alpha & \alpha \\ \alpha & -\alpha \end{pmatrix}$$

Then we have the transition probability matrix:

$$\begin{pmatrix} 1-p(t) & p(t) \\ p(t) & 1-p(t) \end{pmatrix} \text{ where } p(t) = \frac{1}{2}(1 - e^{-4\alpha t}).$$

There is two cases

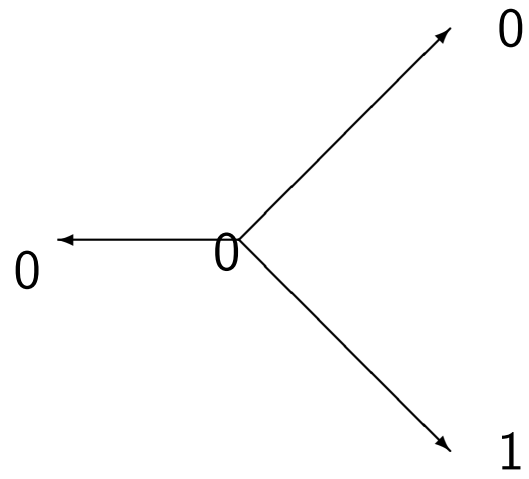


Figure 2: The interior node with a state 0.

There is two cases

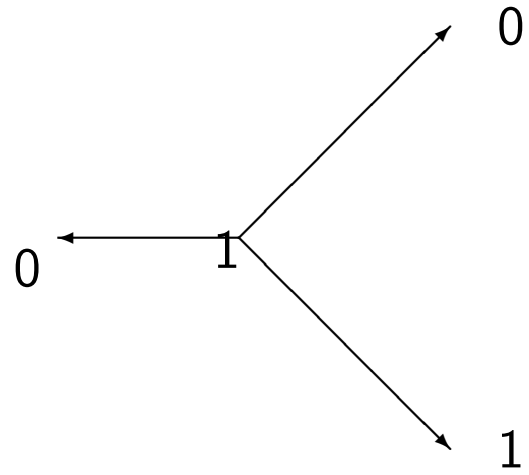


Figure 3: The interior node with a state 1.

Triples

The probability of this will be given by:

$$\begin{aligned}
& P(X_{t_1}=0, X_{t_2}=0, X_{t_3}=1) \\
= \; & P(\{X_{t_1}=0, X_{t_2}=0, X_{t_3}=1\} \cap \{X_0=0\}) \\
& + P(\{X_{t_1}=0, X_{t_2}=0, X_{t_3}=1\} \cap \{X_0=1\}) \\
= \; & P(Figure2) + P(Figure3) \\
= \; & P(X_{t_1}=0, X_{t_2}=0, X_{t_3}=1 | X_0=0) P(X_0=0) \\
& + P(X_{t_1}=0, X_{t_2}=0, X_{t_3}=1 | X_0=1) P(X_0=1) \\
= \; & P(X_{t_1}|X_0=0) P(X_{t_2}|X_0=0) P(X_{t_3}|X_0=0) P(X_0=0) \\
& + P(X_{t_1}|X_0=1) P(X_{t_2}|X_0=1) P(X_{t_3}|X_0=1) P(X_0=1).
\end{aligned}$$

Under this model, since this is time reversable, we use the stationary distribution $(1/2, 1/2)$ for $P(X_0=0)$ and $P(X_0=1)$, Thus, $P(X_0=0) = P(X_0=1) = 1/2$.

Generalizing to multiple sites....

Suppose we have multiple site and assume that each site is independently mutated.

Example:

$$\begin{aligned} X^1 &= GATTACA \\ X^2 &= GCCATAC \end{aligned}$$

Since each site mutated independently we have

$$\begin{aligned} & P(X^1, X^2) \\ = \; & P_{GG} P_{AC} P_{TC} P_{TA} P_{AT} P_{CA} P_{AC} \\ = \; & P^1_{GG} P^2_{AC} P^1_{TC} P^1_{TA} P^1_{AT} P^1_{CA} \end{aligned}$$

Thus all we have to do is to count frequencies of pairs (a, b) where $a, b \in \Sigma$ to compute the probability.

The GTR model

Consider the general time reversible (GTR) model.

Let $\pi_a,\ a \in \Sigma,\ \sum_a \pi_a = 1$, be the stationary distribution of the Markov chain.

The GTR model has substitution rate matrix:

$$Q_\theta = \begin{bmatrix} \cdot & \theta_{AG}\pi_G & \theta_{AC}\pi_C & \theta_{AT}\pi_T \\ \theta_{AG}\pi_A & \cdot & \theta_{GC}\pi_C & \theta_{GT}\pi_T \\ \theta_{AC}\pi_A & \theta_{GC}\pi_G & \cdot & \theta_{CT}\pi_T \\ \theta_{AT}\pi_A & \theta_{GT}\pi_G & \theta_{CT}\pi_C & \cdot \end{bmatrix}$$

where the diagonal elements are such that each row sums to zero.

The 6 unknown parameters are $\theta = (\theta_{AG}, \theta_{AC}, \theta_{AT}, \theta_{GC}, \theta_{GT}, \theta_{CT})$.

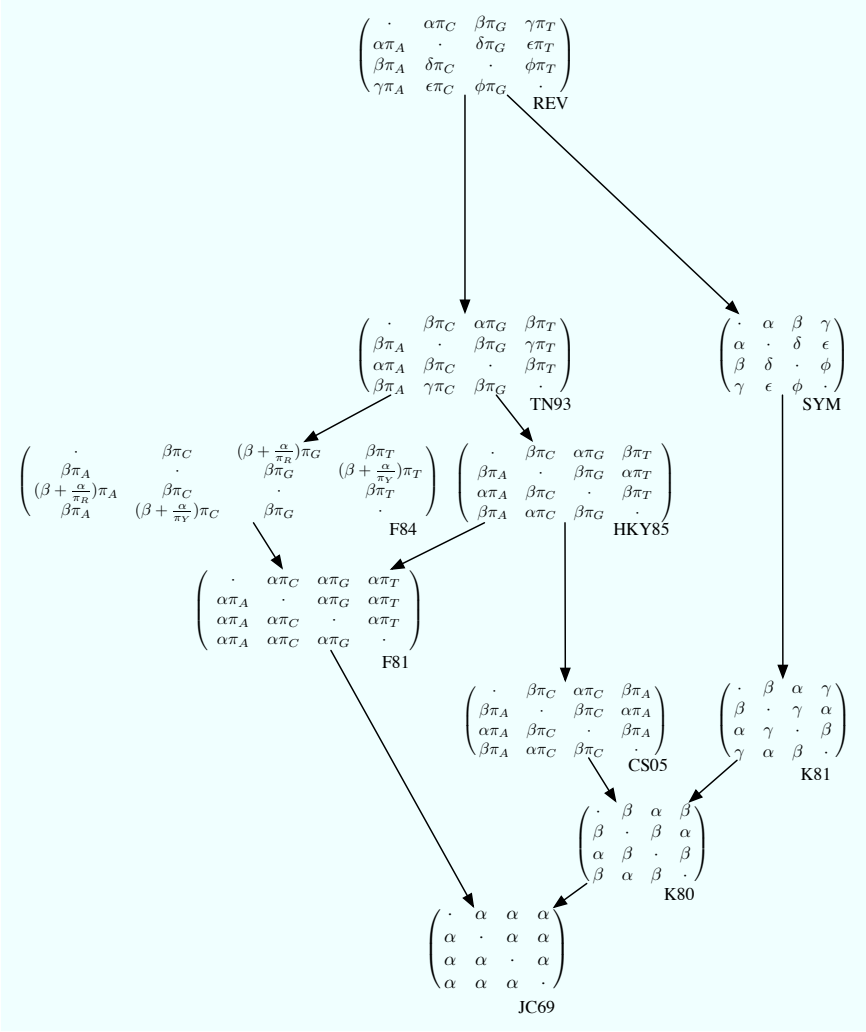


Figure 4: The Felsenstein hierarchy of evolutionary models