Exercises

1) Calculate the nucleotide substitution rate using JC69 model and this alignment:

CGATCGATCGA
CAAGCCA-CTA

$$K = -\frac{3}{4}\ln\left(1-\frac{4}{3}p\right)$$

p = 4/10 = 0.40
K= 0.57

2) Why is p < K in the JC69 Model? correct for multiple hits

3) Under the JC69 model, how often would you expect to see an A in one sequence? How often would you expect to see an A aligned to A between two sequences as divergence time goes to infinity? P(A) = 0.25 (equal base frequency), P_{ii}(infinite) = 0.25, P_{ii}(infinite)*P(A) = 0.0625

4) What is the probability of three identities given a substitution rate (αt) of 0.5? $P_{ii} = 1/4+3/4 * e^{-4*0.5} = 0.124$

$P_{ij} = 1/4-1/4*e^{-4*0.5} = 0.216$ $P_{ij}*P_{ii}*P_{ii} = 0.0267$

$$P_{ii}(t) = \frac{1}{4}+\frac{3}{4}e^{-4\alpha t}$$

5) What are the assumptions of JC model?

1) Equal base frequencies
2) Equal mutation rates between the bases
3) Constant mutation rate (time & sequence)
4) No selection
5) Sites evolve independent of one another

Today's objectives

- Introduction to comparative genomics
- Substitution rates vs differences
- Jukes-Cantor model
- Markov Models
- Markov substitution models
- Maximum likelihood

Markov Models

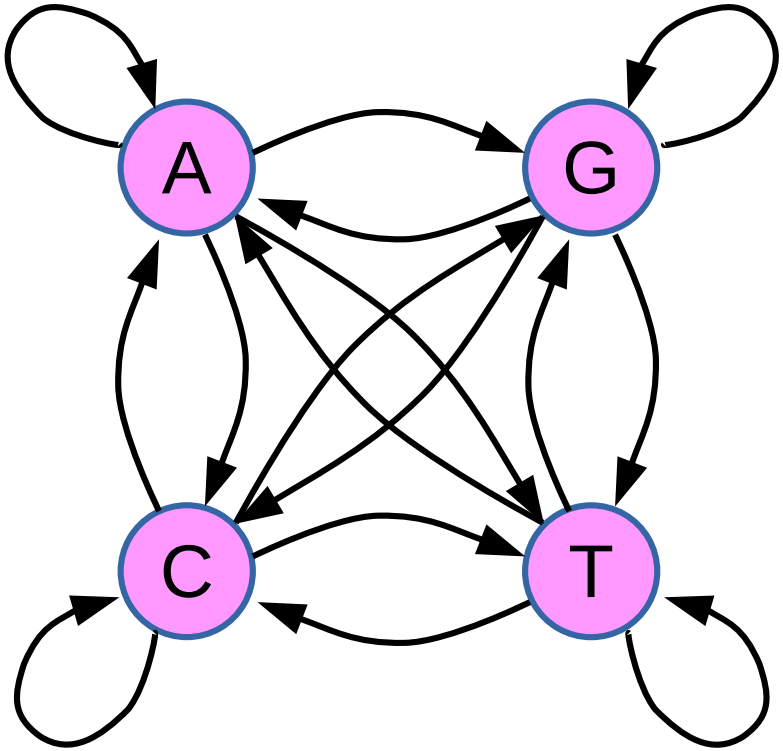
A Markov chain is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event.
A Markov chain is the sequences of transitions through time of a Markov process.

Markov process:
(a) The number of possible outcomes or states is finite.
(b) The outcome at any stage depends only on the outcome of the previous stage.
(c) The probabilities are constant over time.
Transitions determined by transition rate matrix

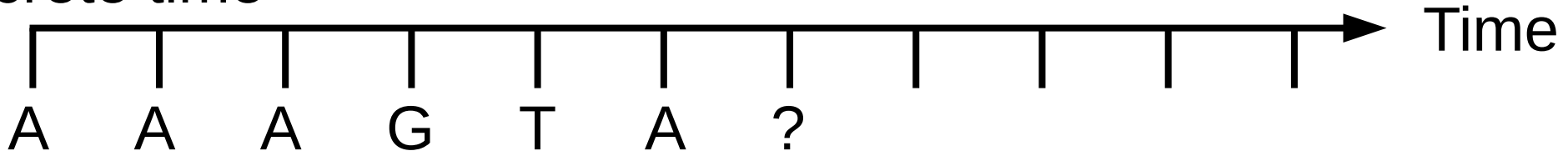
Markov Chains:
Discrete or Continuous
We will also cover: HMM and MCMC

Markov Models

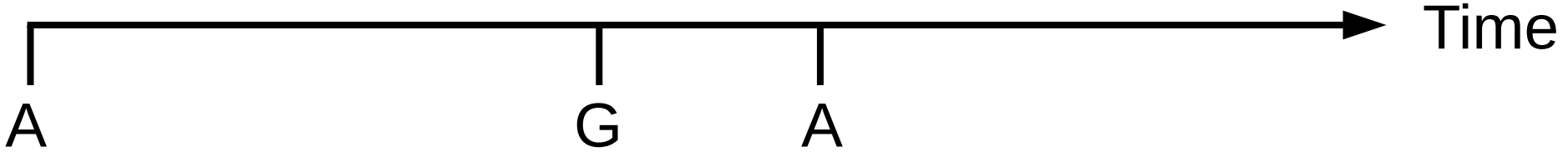
- States {A,G,C,T}
- Transitions between states are governed by transition matrix
- Time between transitions are exponential random variables (continuous)



Discrete time



Continuous time

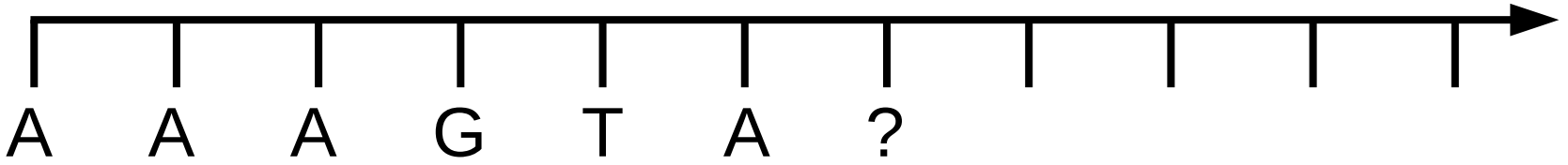


Markov Chains in Biology

- Birth - death process (ecology)
- Substitution rate (molecular evolution)
- DNA sequence features (genes, hidden)

Discrete Time Markov Chain

A Markov Chain:



$$P_{i \to j} = P\left(X_{n+1} = j \,|\, X_n = i\right)$$

Depends on current state

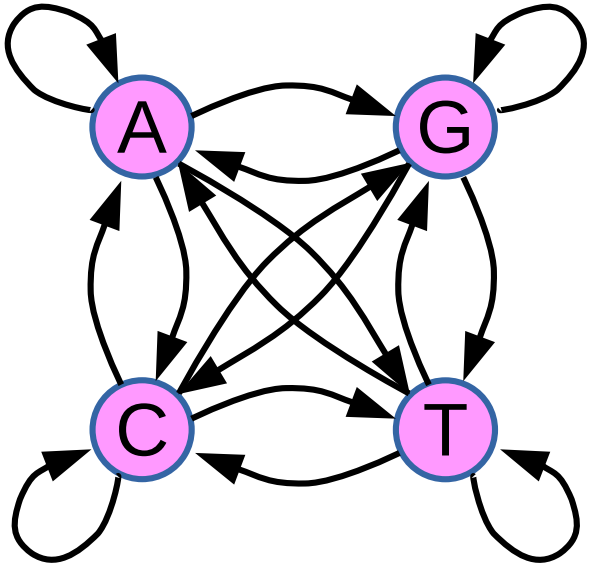
$$P_{i \to j} = \begin{pmatrix} p_{AA} & p_{AG} & p_{AC} & p_{AT} \\ p_{GA} & p_{GG} & p_{GC} & p_{GT} \\ p_{CA} & p_{CG} & p_{CC} & p_{CT} \\ p_{TA} & p_{TG} & p_{TC} & p_{TT} \end{pmatrix} \quad P_{ij} = \begin{pmatrix} 1-3\alpha & \alpha & \alpha & \alpha \\ \alpha & 1-3\alpha & \alpha & \alpha \\ \alpha & \alpha & 1-3\alpha & \alpha \\ \alpha & \alpha & \alpha & 1-3\alpha \end{pmatrix}$$

Jukes-Cantor Model
No Memory

http://setosa.io/blog/2014/07/26/markov-chains/

Graphical and matrix representation

Graphical



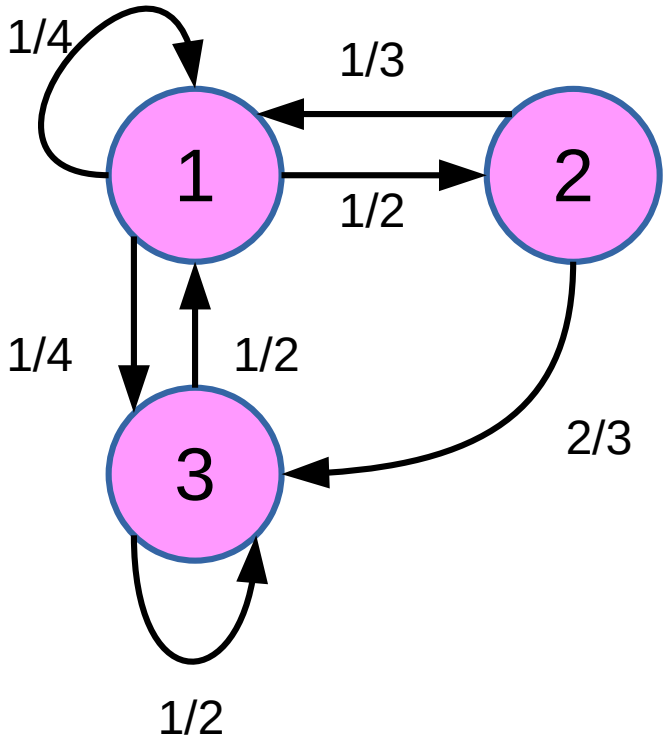
Each arrow is a transition
Each circle is a state

Matrix

$$P_{i \to j} = \begin{pmatrix} p_{AA} & p_{AG} & p_{AC} & p_{AT} \\ p_{GA} & p_{GG} & p_{GC} & p_{GT} \\ p_{CA} & p_{CG} & p_{CC} & p_{CT} \\ p_{TA} & p_{TG} & p_{TC} & p_{TT} \end{pmatrix}$$

Each cell is a transition probability
Each row/column is a state

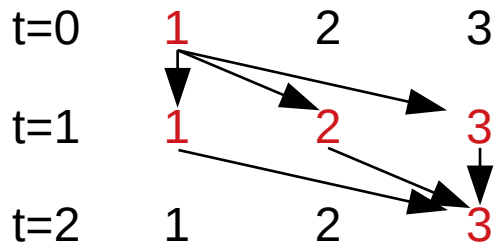
Calculating transition probabilities



$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix} \qquad P = \begin{bmatrix} 1/4 & 1/2 & 1/4 \\ 1/3 & 0 & 2/3 \\ 1/2 & 0 & 1/2 \end{bmatrix}$$

1. What is the probability of $X_0 = 1$, $X_1 = 2$, $X_2 = 3$?
$P(X_0 = 1) * P(X_1 = 2 \mid X_0 = 1) * P(X_2 = 3 \mid X_1 = 2)$

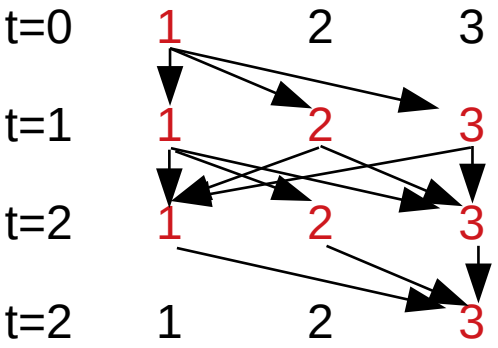
2. What if $P(X_0 = 1) = 1$ (given)
$1*p_{12}*p_{23} = 1*1/2 * 2/3$



123
133
113

3. What is the probability of $X_2 = 3$ given $X_0 = 1$?
$P(123) + P(133) + P(113) =$
$P(X_1 = 2 \mid X_0 = 1) * P(X_2 = 3 \mid X_2 = 2) +$
$P(X_1 = 3 \mid X_0 = 1) * P(X_2 = 3 \mid X_2 = 3) +$
$P(X_1 = 1 \mid X_0 = 1) * P(X_2 = 3 \mid X_2 = 1) =$
$1/2 * 2/3 + 1/4 * 1/2 + 1/4 * 1/4 = 0.52083$

As time increases, so do potential paths



What is the probability of $X_3 = 3$ given $X_0 = 1$?

P(1113) + P(1123) + P(1133) +
P(1213) + ~~P(1223)~~ + P(1233) +
P(1313) + ~~P(1323)~~ + P(1333)

$$P = \begin{bmatrix} 1/4 & 1/2 & 1/4 \\ 1/3 & 0 & 2/3 \\ 1/2 & 0 & 1/2 \end{bmatrix}$$

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}$$

Problem:

- Paths/Probabilities ~ $O(n^t)$ with n states
- We need to evaluate any t, n=4 for DNA but can be larger, e.g. amino acids

Solution: Chapman-Kolmogorov equation & matrix algebra

Chapman–Kolmogorov equation

$P_{ij}^{1} = P_{ij}$ P^1_{ij} = probability of i to j in one generation

$$P_{ij}^{n+m} = \sum_{k=0}^{\infty} P_{ik}^{n} P_{kj}^{m}$$ for all n, m > 0, all i, j

Squared

$$P^{(2)} = P^{(1+1)} = P \cdot P = P^2$$

$$P^{(n)} = P^{(n-1+1)} = P^{n-1} \cdot P = P^n$$

Power n

Thus, we can calculate any P_{ij} if we can multiple matrices of transition probabilities

Matrix multiplication
i.e. matrix algebra

How do we calculate P^2 and P^n?
Where P is a matrix

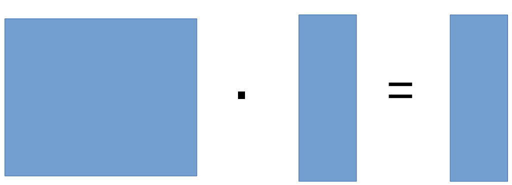
Vector multiplication

Let: $\mathbf{a} = (a_1, a_2, a_3, \ldots a_n)$

& $\mathbf{b} = (b_1, b_2, b_3, \ldots b_n)$

$$\mathbf{a} \cdot \mathbf{b} = a_1 b_1 + a_2 b_2 + a_3 b_3 + \ldots + a_n b_n$$

Scalar product or “dot product”



Square Matrix by vector

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \quad b = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \quad Ab = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 \end{pmatrix}$$

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 \end{pmatrix}$$

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 \end{pmatrix}$$

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 \end{pmatrix}$$

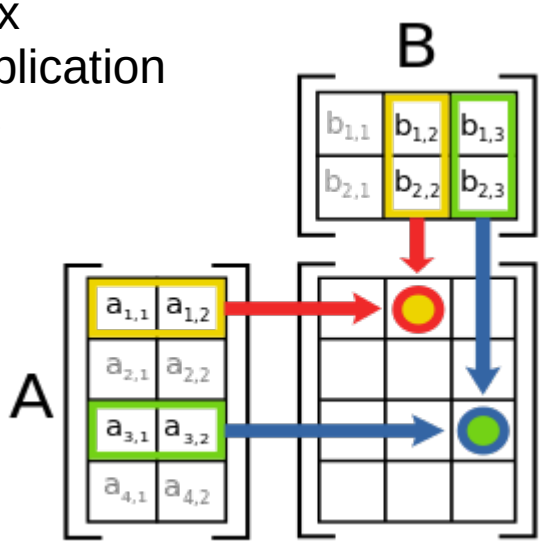
http://xaktly.com/MatrixOperations.html

Generalized multiplication

Given Matrix A (m x n) and Matrix B (n x p):

- A x B is defined to be a (m x p) matrix
- Undefined if columns in A != rows in B

Matrix multiplication A x B



$$\underset{4\times2 \text{ matrix}}{\begin{bmatrix} a_{11} & a_{12} \\ \cdot & \cdot \\ a_{31} & a_{32} \\ \cdot & \cdot \end{bmatrix}} \underset{2\times3 \text{ matrix}}{\begin{bmatrix} \cdot & b_{12} & b_{13} \\ \cdot & b_{22} & b_{23} \end{bmatrix}} = \underset{4\times3 \text{ matrix}}{\begin{bmatrix} \cdot & x_{12} & x_{13} \\ \cdot & \cdot & \cdot \\ \cdot & x_{32} & x_{33} \\ \cdot & \cdot & \cdot \end{bmatrix}}$$

$$x_{12} = a_{11}b_{12} + a_{12}b_{22}$$
$$x_{33} = a_{31}b_{13} + a_{32}b_{23}$$

The dot product of two vectors a = [a1, a2, …, an] and b = [b1, b2, …, bn] is defined as:

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

Thus, each element in 4x3 matrix is dot product A-rows and B-columns

Matrix algebra

$$A=\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{12} & a_{22} & a_{23} & a_{24} \\ a_{13} & a_{32} & a_{33} & a_{34} \\ a_{14} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

$$B=\begin{bmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{12} & b_{22} & b_{23} & b_{24} \\ b_{13} & b_{32} & b_{33} & b_{34} \\ b_{14} & b_{42} & b_{43} & b_{44} \end{bmatrix}$$

$$AB=\begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{12} & x_{22} & x_{23} & x_{24} \\ x_{13} & x_{32} & x_{33} & x_{34} \\ x_{14} & x_{42} & x_{43} & x_{44} \end{bmatrix}$$

$$x_{11}=a_{11}b_{11}+a_{12}b_{21}+a_{13}b_{31}+a_{14}b_{41}$$

$$x_{32}=a_{31}b_{12}+a_{32}b_{22}+a_{33}b_{32}+a_{34}b_{42}$$

Matrix algebra

$$A=\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{12} & a_{22} & a_{23} & a_{24} \\ a_{13} & a_{32} & a_{33} & a_{34} \\ a_{14} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

$$B=\begin{bmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{12} & b_{22} & b_{23} & b_{24} \\ b_{13} & b_{32} & b_{33} & b_{34} \\ b_{14} & b_{42} & b_{43} & b_{44} \end{bmatrix}$$

$$AB=\begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{12} & x_{22} & x_{23} & x_{24} \\ x_{13} & x_{32} & x_{33} & x_{34} \\ x_{14} & x_{42} & x_{43} & x_{44} \end{bmatrix}$$

$$x_{11}=a_{11}b_{11}+a_{12}b_{21}+a_{13}b_{31}+a_{14}b_{41}$$

$$x_{32}=a_{31}b_{12}+a_{32}b_{22}+a_{33}b_{32}+a_{34}b_{42}$$

Matrix algebra

$$A=\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

$$B=\begin{bmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \\ b_{31} & b_{32} & b_{33} & b_{34} \\ b_{41} & b_{42} & b_{43} & b_{44} \end{bmatrix}$$

$$AB=\begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ x_{41} & x_{42} & x_{43} & x_{44} \end{bmatrix}$$

$$x_{11}=a_{11}b_{11}+a_{12}b_{21}+a_{13}b_{31}+a_{14}b_{41}$$

$$x_{32}=a_{31}b_{12}+a_{32}b_{22}+a_{33}b_{32}+a_{34}b_{42}$$

Whats the probability of going from state 3 to state 2 in 2 moves?

x_{32} = 312, 322, 332, 342

If A is for move one and B is for move two, then x_{32}

Matrix algebra

$X_0 = [1,0,0]$ states

$X_1 = X_0 * P$

$X_2 = X_1 * P$

$X_2 = X_0 * P^2$

$$X_{n+1} = X_n P$$

$$X_{n+3} = X_n P^3$$

$$\lim_{n \to \infty} P^n = \pi$$

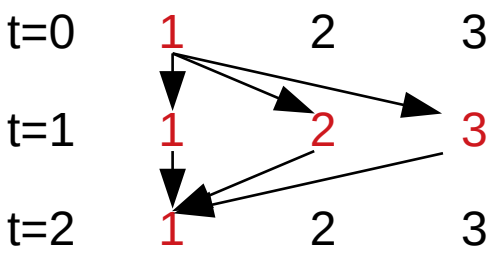
```
In [1]: import numpy as np
        x0 = np.array( [1,0,0 ])
        P = np.array( [[1/4, 1/2, 1/4], [1/3, 0, 2/3],[1/2, 0, 1/2]])

In [2]: print(x0)
        print(P)

        [1 0 0]
        [[ 0.25       0.5        0.25      ]
         [ 0.33333333 0.         0.66666667]
         [ 0.5        0.         0.5       ]]

In [3]: x1 = x0.dot(P)
        x1 = np.dot(x0,P) # both lines are equivalent
        print(x1)

        [ 0.25  0.5   0.25]

In [4]: x2 = x1.dot(P)
        print(x2)

        [ 0.35416667  0.125       0.52083333]
```

stationary distribution (π) is the left eigenvector of the transition matrix

Dot product calculation

$$P=\begin{bmatrix} 1/4 & 1/2 & 1/4 \\ 1/3 & 0 & 2/3 \\ 1/2 & 0 & 1/2 \end{bmatrix}$$

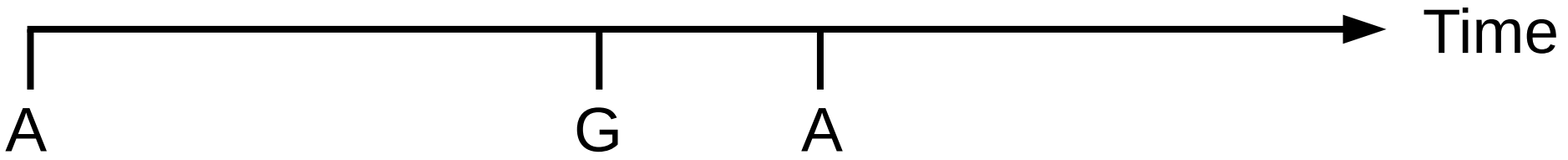
$$P=\begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}$$




```
In [1]: import numpy as np
        x0 = np.array( [1,0,0 ])
        P = np.array( [[1/4, 1/2, 1/4], [1/3, 0, 2/3],[1/2, 0, 1/2]])

In [2]: print(x0)
        print(P)

        [1 0 0]
        [[ 0.25        0.5         0.25      ]
         [ 0.33333333  0.          0.66666667]
         [ 0.5         0.          0.5       ]]

In [3]: x1 = x0.dot(P)
        x1 = np.dot(x0,P) # both lines are equivalent
        print(x1)

        [ 0.25  0.5   0.25]

In [4]: x2 = x1.dot(P)
        print(x2)

        [ 0.35416667  0.125       0.52083333]
```

1/4*1/4 + 1/2*1/3 + 1/4*1/2 = 0.354

Continuous time Markov Chain



Definition of derivative

$$\frac{dP(t)}{dt} = \lim_{\delta t \to 0} \frac{P(t+\delta t) - P(t)}{\delta t}$$

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

$$P(t) = \exp(Qt)$$

$$P(t) = \sum_{n=0}^{\infty} \frac{Q^n t^n}{n!}$$

Compute: O(m^3) for m by m matrix using NumPy numerical approximation

$$\lim_{n \to \infty} P^n = \pi$$

Q is the (instantaneous) transition rate matrix

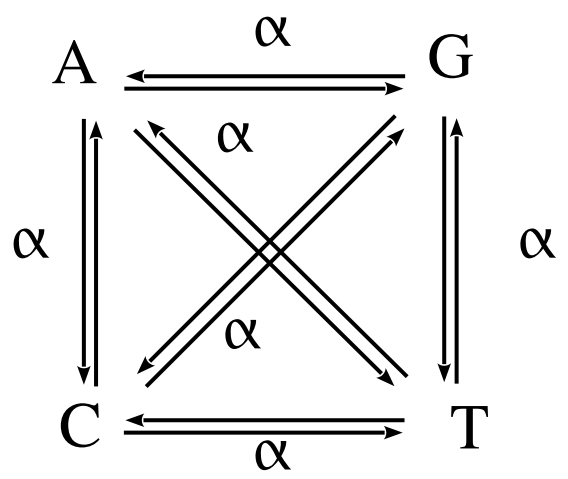
P is the transition probability matrix

πP = π

π = equilibrium states

$\pi_i Q_{ij} = \pi_j Q_{ji}$ (reversible)

Transition Rate Matrix



$$Q = \begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix} \end{array}$$

$$\sum_{j=0}^{n} Q_{ij} = 0$$

Sum of the rows = 0

Therefore $$Q_{ii} = -\sum_{j\,;\,j \neq i}^{n} Q_{ij}$$

Transition probabilities

$$P(t) = \exp(Qt) \qquad Q = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}$$

If alpha = u/3, then:

$$P(t) = \exp(Qt)$$

$$Q = \begin{pmatrix} -u & \frac{u}{3} & \frac{u}{3} & \frac{u}{3} \\ \frac{u}{3} & -u & \frac{u}{3} & \frac{u}{3} \\ \frac{u}{3} & \frac{u}{3} & -u & \frac{u}{3} \\ \frac{u}{3} & \frac{u}{3} & \frac{u}{3} & -u \end{pmatrix}$$

$$P(t) = \exp(Qut)$$

$$Q = \begin{pmatrix} -1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & -1 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & -1 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & -1 \end{pmatrix}$$

Transition probabilities

$$P(t) = \exp(Qt)$$

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

If alpha = u/3, then:

$$P(t) = \exp(Qut)$$

$$Q = \begin{pmatrix} -1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & -1 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & -1 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & -1 \end{pmatrix}$$

Substitution rate = mutation (u) * time (t)

$$P(1) = \begin{array}{c} \begin{matrix} \text{A} & \text{G} & \text{C} & \text{T} \end{matrix} \\ \begin{pmatrix} 0.4477 & 0.1841 & 0.1841 & 0.1841 \\ 0.1841 & 0.4477 & 0.1841 & 0.1841 \\ 0.1841 & 0.1841 & 0.4477 & 0.1841 \\ 0.1841 & 0.1841 & 0.1841 & 0.4477 \end{pmatrix} \end{array}$$

Scaling: Because u is the mutation rate per generation and t is time in generations, we would like the substitution rate, ut, to be measured in units of substitutions per site. This can be accomplished by scaling the rate matrix such that when the substitution rate ut = 1 there is one substitution per site. A simple way to achieve this scaling is to scale u, the total number of substitutions per unit time to 1. This can be done by setting the flux to 1.

Transition probabilities

$$P(t) = \exp(Qt)$$

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

$$Q = \begin{pmatrix} -1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & -1 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & -1 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & -1 \end{pmatrix}$$

CGAC
CTTC
What is the probability of the data under JC69 model with t=1?
$P_{CC} * P_{GT} * P_{AT} * P_{CC}$

$$P(1) = \begin{array}{c} \begin{matrix} A & G & C & T \end{matrix} \\ \begin{pmatrix} 0.4477 & 0.1841 & 0.1841 & 0.1841 \\ 0.1841 & 0.4477 & 0.1841 & 0.1841 \\ 0.1841 & 0.1841 & 0.4477 & 0.1841 \\ 0.1841 & 0.1841 & 0.1841 & 0.4477 \end{pmatrix} \end{array}$$

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t}$$ 1/4+3/4*exp(-4/3) =0.447

$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t}$$ 1/4-1/4*exp(-4/3) =0.184

JC69, K80 and HKY85 Models

JC69 $$Q=\begin{pmatrix} * & \alpha & \alpha & \alpha \\ \alpha & * & \alpha & \alpha \\ \alpha & \alpha & * & \alpha \\ \alpha & \alpha & \alpha & * \end{pmatrix} \text{ or } \begin{pmatrix} * & 1 & 1 & 1 \\ 1 & * & 1 & 1 \\ 1 & 1 & * & 1 \\ 1 & 1 & 1 & * \end{pmatrix}$$

K80 $$Q=\begin{pmatrix} * & \kappa & 1 & 1 \\ \kappa & * & 1 & 1 \\ 1 & 1 & * & \kappa \\ 1 & 1 & \kappa & * \end{pmatrix}$$

HKY85 $$Q=\begin{pmatrix} * & \pi_G \kappa & \pi_C & \pi_T \\ \pi_A \kappa & * & \pi_C & \pi_T \\ \pi_A & \pi_G & * & \pi_T \kappa \\ \pi_A & \pi_G & \pi_C \kappa & * \end{pmatrix}$$

κ = transitions/transversion
AG and CT changes
π = equilibrium base frequencies

K80, Kimura 1980 model, distinguishes between transitions (AG), i.e. from purine to purine, or (CT), i.e. from pyrimidine to pyrimidine) and transversions (from purine to pyrimidine or vice versa).

HKY85, the Hasegawa, Kishino and Yano 1985 model allows unequal base frequencies.

$$P_{i \to j}=\begin{pmatrix} p_{AA} & p_{AG} & p_{AC} & p_{AT} \\ p_{GA} & p_{GG} & p_{GC} & p_{GT} \\ p_{CA} & p_{CG} & p_{CC} & p_{CT} \\ p_{TA} & p_{TG} & p_{TC} & p_{TT} \end{pmatrix}$$

Scaling with JC69 model

$$P(t) = \exp(Qt)$$

$$Q = \begin{pmatrix} * & \alpha & \alpha & \alpha \\ \alpha & * & \alpha & \alpha \\ \alpha & \alpha & * & \alpha \\ \alpha & \alpha & \alpha & * \end{pmatrix}$$

$$Q_{ii} = -\sum_{i \neq j}^{n} Q_{ij}$$

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

$$P(t) = \exp(Qut)$$

$$Q = \begin{pmatrix} * & 1 & 1 & 1 \\ 1 & * & 1 & 1 \\ 1 & 1 & * & 1 \\ 1 & 1 & 1 & * \end{pmatrix}$$

$$Q = \begin{pmatrix} -1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & -1 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & -1 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & -1 \end{pmatrix}$$

Divide Q by the flux, so that the total flux is set to one.

$$-\sum_{i=1}^{n} \pi_i Q_{ii}$$

= - 4* (0.25 * -3α) = 3α

Flux

What is the substitution rate?

Jukes Cantor

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t}$$

$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t}$$

$$K = -\frac{3}{4} \ln\left(1 - \frac{4}{3} p\right)$$

HKY85

$$P(t) = \exp(Qt)$$

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

What is the substitution rate?

Jukes Cantor

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t}$$

$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t}$$

$$K = -\frac{3}{4} \ln\left(1 - \frac{4}{3} p\right)$$

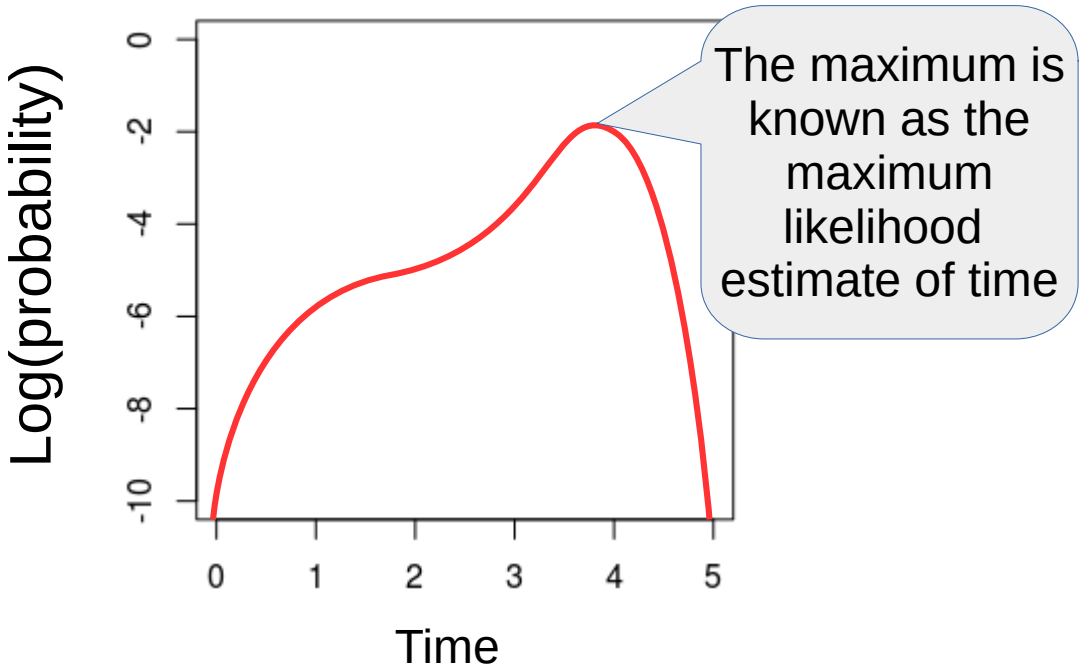
HKY85

$$P(t) = \exp(Qt)$$

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

Lets say we estimate kappa and equilibrium frequencies, then all we need to know is time (t)

We can then calculate the probability of the data for any value of t



Maximum Likelihood

- Suppose we have a coin that lands heads with probability θ and tails with probability $1-\theta$.
- Assuming sample $x_1, x_2, ..., x_n$ is from a parametric distribution $f(x|\theta)$, find the maximum likelihood estimate of θ.
- The probability mass function (density) is Bernoulli: $f(x|\theta) = \theta^k(1-\theta)^{1-k}$ for $k \in \{0,1\}$

Several random variables and probability distributions may be derived from the Bernoulli process:

- The number of successes in the first n trials, ~ binomial distribution B(n, p)
- The number of trials needed to get r successes, ~ negative binomial distribution NB(r, p)
- The number of trials needed to get one success, ~ geometric distribution NB(1, p), a special case of the negative binomial distribution

Probability and Likelihood

Sum = 1

$P(x \mid \theta)$: Probability of event x given model θ

- Viewed as a function of x (fixed θ), it's a probability
- Viewed as a function of θ (fixed x), it's a likelihood

Sum != 1

$P(x_1..x_n \mid \theta) = \theta^k(1-\theta)^{n-k}$ where k is number of heads

$P(HHTHH \mid .6) = (0.6)^4(1-0.6)^1 = 0.05184$

$P(HHTHH \mid .5) = (0.5)^4(1-0.5)^1 = 0.03125$

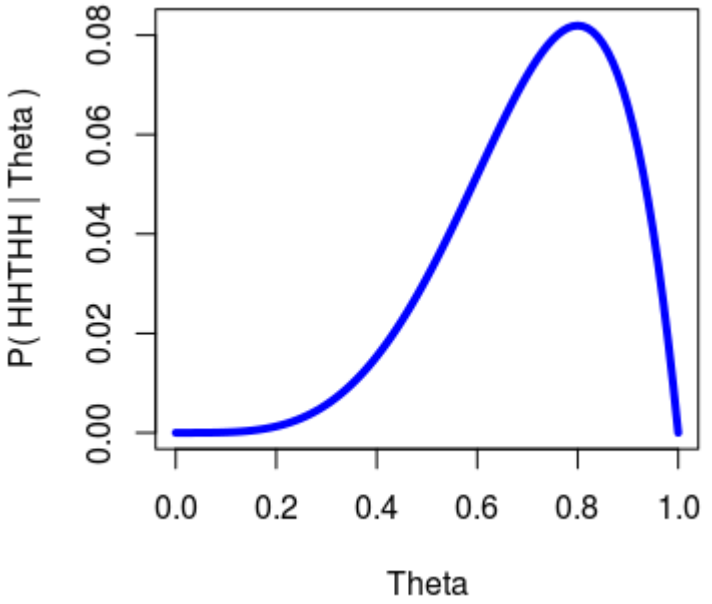
$P(HHTHH \mid .6) > P(HHTHH \mid .5)$

θ = probability of heads

HHTHH is more likely when $\theta = .6$ than $\theta = .5$

Maximum Likelihood

θ	$\theta^4(1-\theta)$
0.2	0.0013
0.5	0.0313
0.8	0.0819
0.95	0.0407



To find the maximum likelihood, find when the derivative is zero.

Maximum Likelihood Estimate

$$L(\theta|x_1, x_2, \ldots x_n) = f(x_1|\theta) \ldots f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

$$L(\theta|x_1, x_2, \ldots x_n) = \theta^k (1-\theta)^{n-k}$$

$$\log\left(L(\theta|x_1, x_2, \ldots x_n)\right) = k\log(\theta) + (n-k)\log(1-\theta)$$

$$\frac{\partial}{\partial\theta}\log\left(L(\theta|x_1, x_2, \ldots x_n)\right) = \frac{k}{(\theta)} + \frac{(k-n)}{(1-\theta)}$$

Set derivative to zero and solve for θ $\hat{\theta} = \frac{k}{n}$

The maximum likelihood estimate of theta is k/n or the number of heads divided by the number of trials

Bayes' Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Proof using definition of conditional probability:
P(B)*P(A|B) = P(A and B)

P(A|B) = P(A and B) / P(B)
P(B|A) = P(A and B) / P(A)

P(A|B)*P(B) = P(B|A)*P(A)

P(A|B) = P(B|A)*P(A)/P(B)

Bayes' Theorem

L(A|B) = P(B|A)

Likelihood A

Prior

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

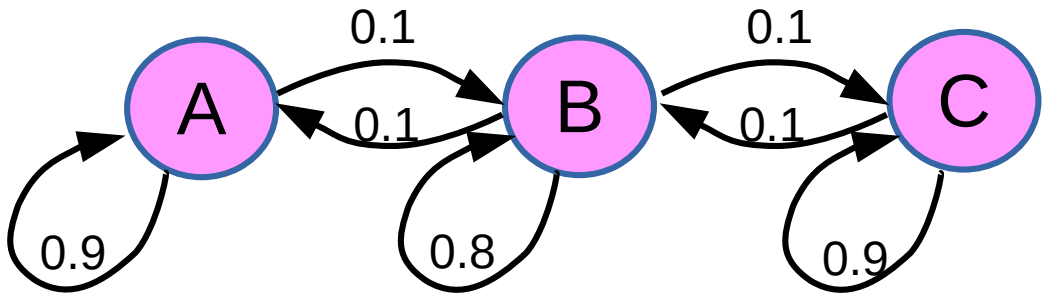
Posterior A

Marginal probability

$$P(\theta|x_1, x_2, \ldots x_n) = \frac{f(x_1, x_2, \ldots x_n|\theta)P(\theta)}{P(x_1, x_2, \ldots x_n)}$$

Exercises

1) What is the transition rate matrix for the following discrete time Markov Chain?



	A	B	C
A			
B			
C			

2) What is the probability of X_2=A, after two steps in a discrete time Markov Chain illustrated above, given that X_0=A? What is P(X_3=A|X_0=A)?

3) Why do we use a model with memory for nucleotide substitution?