

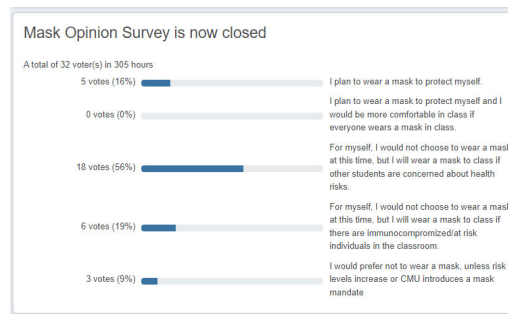
## Today

- **Announcements**
- Markov chains
  - Review
  - Stationary distributions
- Models of sequence evolution
  - Nucleotide substitution models
  - (Amino acids in about 2 weeks)

## Announcements

- Office hours with Dr. Stolzer will be held via zoom
  - Tuesdays 1-2pm **But not today**
  - Fridays 4-5pm
  - By appointment (If you cannot make either time)

## COVID-19 Update



## Today

- Announcements
- **Markov chains**
  - **Review**
  - Stationary distributions
- Models of sequence evolution
  - Nucleotide substitution models
  - (Amino acids in about 2 weeks)

**Box 2: Summary of Markov chain notation**

A Markov chain has states  $E_1, E_2, \dots, E_n$  corresponding to the range of the associated random variable.

$\varphi_j(t)$  is the probability that the chain is in state  $E_j$  at time  $t$ . The vector  $\varphi(t) = (\varphi_1(t), \dots, \varphi_n(t))$  is the *state probability distribution* at time  $t$ .

$\pi = \varphi(0)$  is the *initial state probability distribution*.

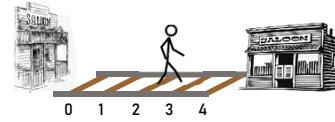
$P$  is the *transition probability matrix*.  $P_{jk}$  gives the probability of making a transition to state  $E_k$  at time  $t+1$ , given that the chain was in state  $E_j$  at time  $t$ . The rows of this matrix sum to one:  $\sum_k P_{jk} = 1$ .

The state probability distribution at time  $t+1$  is given by  $\varphi(t+1) = \varphi(t) \cdot P$ . The probability of being in state  $E_k$  at  $t+1$  is

$$\varphi_k(t+1) = \sum_j \varphi_j(t) P_{jk}$$

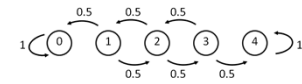
The *Markov property* states that Markov chains are memoryless. The probability that the chain is in state  $E_k$  at time  $t+1$ , depends only on  $\varphi(t)$  and is independent of  $\varphi(t-1), \varphi(t-2), \varphi(t-3), \dots$

**Random walk with absorbing boundaries**



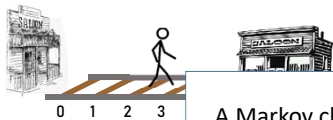
$$P = \begin{bmatrix} E_0 & E_1 & E_2 & E_3 & E_4 \\ E_0 & 1 & 0 & 0 & 0 & 0 \\ E_1 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ E_2 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ E_3 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ E_4 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

- Every time the clock ticks, the drunk takes one step
- to the left with prob 0.5
  - to the right with prob 0.5



If the drunk reaches State  $E_1$  or  $E_4$  the drunk enters the bar and stays there

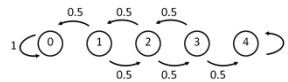
**Random walk with absorbing boundaries**



$$P = \begin{bmatrix} E_0 & E_1 & E_2 & E_3 & E_4 \\ E_0 & 1 & 0 & 0 & 0 & 0 \\ E_1 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ E_2 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ E_3 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ E_4 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

A Markov chain with absorbing boundaries is not Irreducible

- Every time the clock ticks the drunk takes one step
- to the left with prob 0.5
  - to the right with prob 0.5



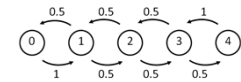
If the drunk reaches State  $E_1$  or  $E_4$  the drunk enters the bar and stays there

**Random walk with reflecting boundaries**



$$P = \begin{bmatrix} E_0 & E_1 & E_2 & E_3 & E_4 \\ E_0 & 0 & 1 & 0 & 0 & 0 \\ E_1 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ E_2 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ E_3 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ E_4 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

- Every time the clock ticks, the drunk takes one step
- to the left with prob 0.5
  - to the right with prob 0.5
- }  $E_1, E_2, E_3$



If the drunk enters State  $E_0$  ( $E_4$ ) the drunk reverses direction and enters  $E_1$  ( $E_3$ ) at the next tick

Random walk with *reflecting* boundaries

$$P = \begin{bmatrix} E_0 & E_1 & E_2 & E_3 & E_4 \\ E_0 & 0 & 1 & 0 & 0 & 0 \\ E_1 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ E_2 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ E_3 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ E_4 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Every time the clock ticks, the drunk takes one step

- to the left with prob 0.5
- to the right with prob 0.5

If the drunk enters State  $E_0$  ( $E_4$ ) the drunk reverses direction and enters  $E_1$  ( $E_3$ ) at the next tick

## Today

- Announcements
- **Markov chains**
  - Review
  - **Stationary distributions**
- Models of sequence evolution
  - Nucleotide substitution models
  - (Amino acids in about 2 weeks)

Another random walk

$$P = \begin{bmatrix} E_0 & E_1 & E_2 & E_3 & E_4 \\ E_0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ E_1 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ E_2 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ E_3 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ E_4 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

Every time the clock ticks, the drunk takes one step

- to the left with prob 0.5
- to the right with prob 0.5

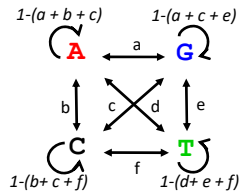
If the drunk enters State  $E_0$  ( $E_4$ ), the drunk

- rests with prob 0.5
- reverses direction and enters  $E_1$  ( $E_3$ ) with prob 0.5

## Today

- Announcements
- Markov chains
  - Review
  - Stationary distributions
- **Models of sequence evolution**
  - **Nucleotide substitution models**
  - (Amino acids in about 2 weeks)

### Properties of DNA substitution models

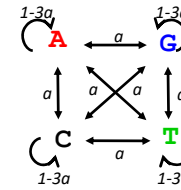


- State space:  $\{E_1 = A, E_2 = G, E_3 = C, E_4 = T\}$
- States are fully connected
- Transition probabilities: substitution frequencies  $(a, b, c, d, \dots)$
- Implicitly also specifies stationary base frequencies:  $\varphi^* = (p_A, p_G, p_C, p_T)$

GACTAGCTAGACATAGCTAGACAGATACGAAGATACGAACTAGCTAGACATATTACATATAC

14

### Jukes-Cantor model (1969)



$$p(A)=0.25$$

$$p(G)=0.25$$

$$p(C)=0.25$$

$$p(T)=0.25$$

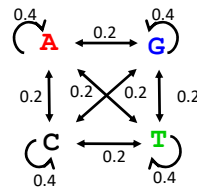
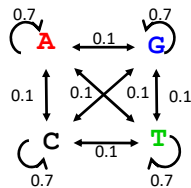
Assumptions:

- All substitutions have equal probability
- Base frequencies are equal

15

### Two Jukes Cantor models with different rates

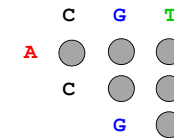
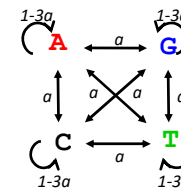
All Jukes Cantor models have a single rate parameter,  $a$ ,  $0 \leq a \leq 1/3$   
 Different instances of the JC model can have different rates. Rates are typically learned from data



The model on the right changes twice as fast as the model on the left.  
 In both models, all substitutions are equally probable

16

### Three representations of the Jukes Cantor model



	A	C	G	T
A	$1-3a$	$a$	$a$	$a$
C	$a$	$1-3a$	$a$	$a$
G	$a$	$a$	$1-3a$	$a$
T	$a$	$a$	$a$	$1-3a$

17

### More nucleotide substitution models

Jukes Cantor

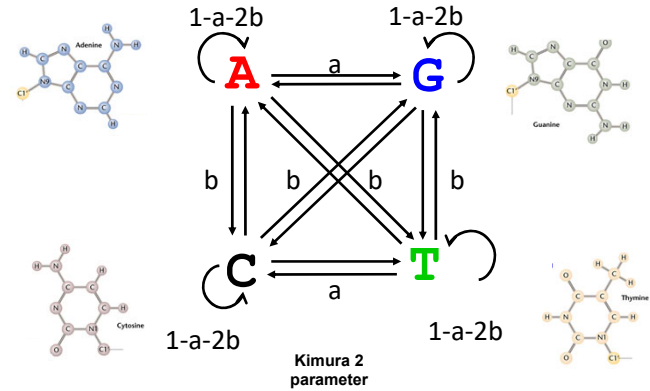
- Uniform substitution probabilities
- Uniform base frequencies

Substitution models can be extended by allowing

- different substitution probabilities for different base pairs
- non-uniform base frequencies

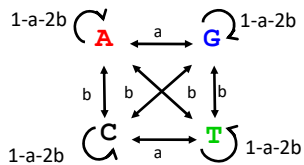
or both

### A more complex model different probabilities for transitions and transversions



Kimura 2 parameter

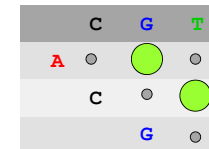
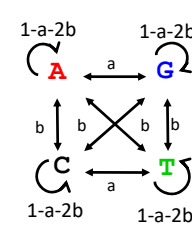
### Kimura 2 parameter model (K2P) (1980)



$$\begin{aligned}
 p(A) &= 0.25 \\
 p(G) &= 0.25 \\
 p(C) &= 0.25 \\
 p(T) &= 0.25
 \end{aligned}$$

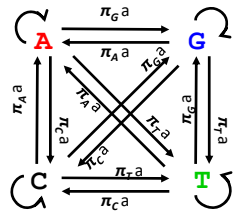
- Transitions and transversions have different probabilities
- Base frequencies are equal

### Three representations of the Kimura 2-parameter model



	A	C	G	T
A	1-a-2b	b	a	b
C	b	1-a-2b	b	a
G	a	b	1-a-2b	b
T	b	a	b	1-a-2b

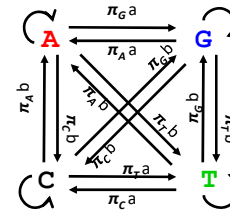
Felsenstein (1981)



$$\begin{aligned}
 p(A) &= \pi_A \\
 p(G) &= \pi_G \\
 p(C) &= \pi_C \\
 p(T) &= \pi_T
 \end{aligned}$$

- All substitutions have equal probability
- Unequal base frequencies  $p(A) \neq p(G) \neq p(C) \neq p(T)$

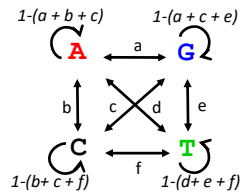
Hasegawa, Kishino & Yano (HKY) (1985)



$$\begin{aligned}
 p(A) &= \pi_A \\
 p(G) &= \pi_G \\
 p(C) &= \pi_C \\
 p(T) &= \pi_T
 \end{aligned}$$

- Transitions and transversions have different probabilities
- Unequal base frequencies  $p(A) \neq p(G) \neq p(C) \neq p(T)$

General Time Reversible model



- All six pairs have different substitution frequencies
- Unequal base frequencies  $p(A) \neq p(G) \neq p(C) \neq p(T)$