

Announcements

Problem Set 2 due Friday at midnight.

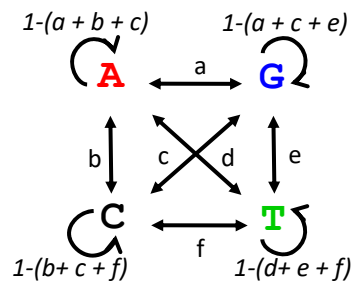
First exam:

- Tuesday, Sept 30, 7:30pm – 9:30pm
- WH 5403
 - This exam is closed book. You may bring two pages (or one page, front and back) of your own notes.
 - The exam covers material covered in Lectures 1 through 9.
 - You will not need a calculator

Models of sequence evolution

- | | |
|---------------------------------------|--------------|
| • Nucleotide substitution models | Last Tuesday |
| • Applications of substitution models | Today |

Properties of DNA substitution models

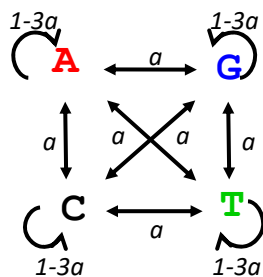


- State space: $\{E_1 = A, E_2 = G, E_3 = C, E_4 = T\}$
- States are fully connected
- Transition probabilities:
substitution frequencies $(a, b, c, d \dots)$
- Implicitly also specifies stationary base frequencies: $\varphi^* = (p_A, p_G, p_C, p_T)$

GACTAGCTAGACATAGCTAGACAGATACGAAGATACGAAGCTAGCTAGACATATTACATATAC

3

Jukes-Cantor model (1969)



	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	$1-3a$

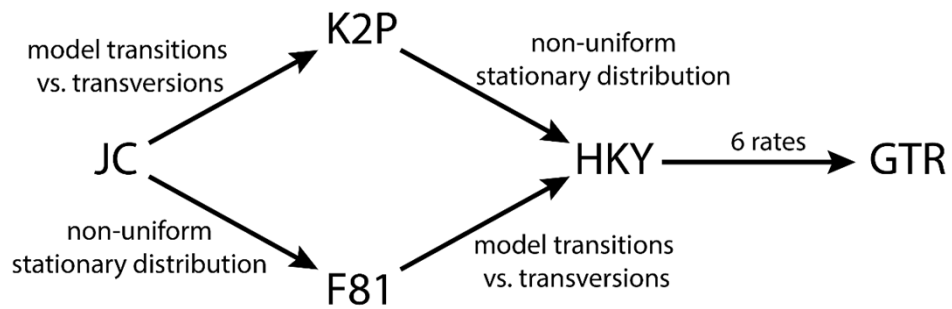
Assumptions:

- All substitutions have equal probability
- Base frequencies are equal

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

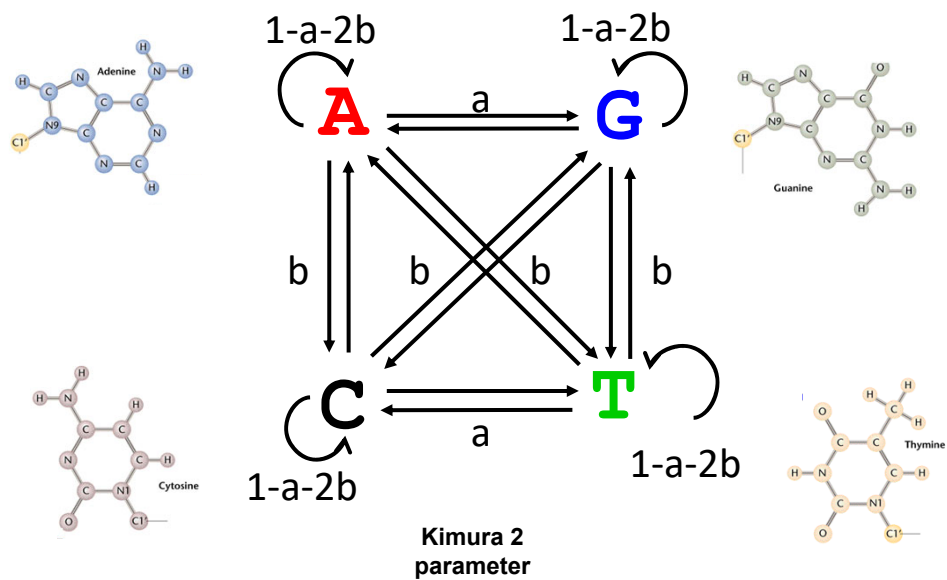
4

Models of DNA sequence evolution



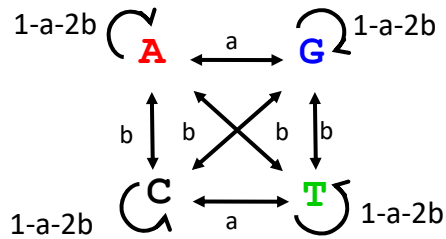
5

Kimura 2 parameter model



6

Kimura 2 parameter model (K2P) (1980)



	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

Assumptions

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

$$p(\text{A})=0.25$$

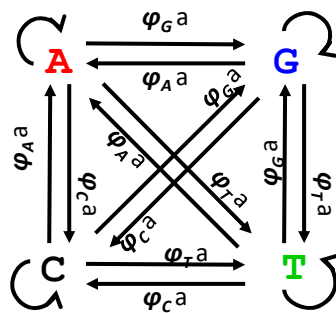
$$p(\text{G})=0.25$$

$$p(\text{C})=0.25$$

$$p(\text{T})=0.25$$

7

Felsenstein (1981)



$$p(\text{A})=\varphi_{\text{A}}$$

$$p(\text{G})=\varphi_{\text{G}}$$

$$p(\text{C})=\varphi_{\text{C}}$$

$$p(\text{T})=\varphi_{\text{T}}$$

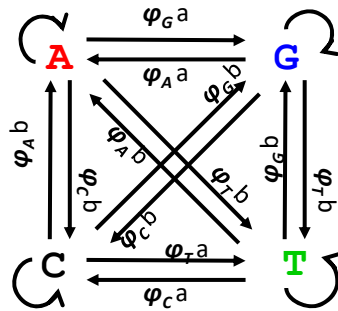
- All substitutions have the same base rate

- Unequal base frequencies

$$p(\text{A}) \neq p(\text{G}) \neq p(\text{C}) \neq p(\text{T})$$

8

Hasegawa, Kishino & Yano (HKY) (1985)

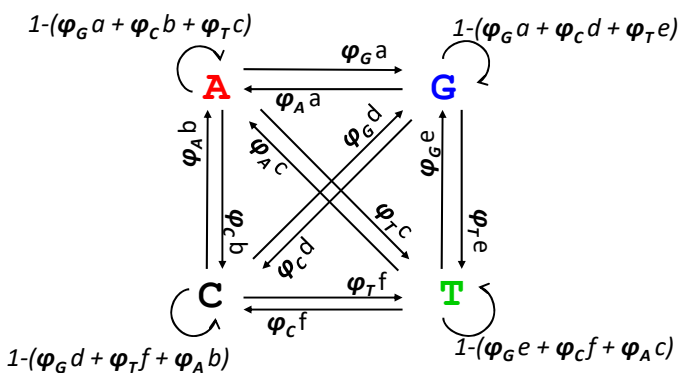


$$\begin{aligned}
 p(\text{A}) &= \varphi_{\text{A}} \\
 p(\text{G}) &= \varphi_{\text{G}} \\
 p(\text{C}) &= \varphi_{\text{C}} \\
 p(\text{T}) &= \varphi_{\text{T}}
 \end{aligned}$$

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

9

General Time Reversible model



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

10

Models of sequence evolution

- Nucleotide substitution models Last Tuesday
- Applications of substitution models
 - Simulation
 - Estimating the probability of observing x aligned with y Today
 - Estimating rates of evolution
 - Phylogenetic inference
 - Correcting for multiple substitutions at the same site Next Tuesday
 - Deriving scoring matrices, $S[x,y]$ October

Questions to ask:

- Given a site evolving according to Jukes Cantor with parameter α , what is the probability of observing x aligned with y ?

```

...ATGCGAGGACTCXCAGTGA...
...ATGTGAGGTCTCYCAATGA...
  
```

Given an alignment of s_1 and s_2 with m observed mismatches, how many substitutions occurred since the divergence of s_1 and s_2 ?

```

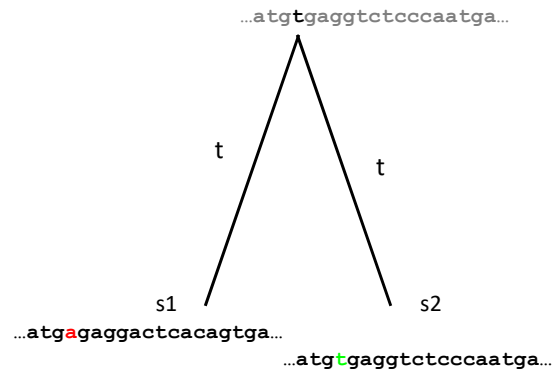
...CACATACGAAGATACGAACGAGC...
...CAGATAGGAAGAGACGATCTAGC...
<----->
  n nucleotides with m mismatches
  
```

We apply these models to the following scenario

Sequences s_1 and s_2 are DNA sequences of length n

s_1 and s_2 have been diverging from a common ancestor for t million years (MY)
according to a given model with parameters $a, b, c, \dots$

- This framework does not model indels
- Assumes site independence
- Parameters are estimated from data



We apply these models to the following scenario

Sequences s_1 and s_2 are DNA sequences of length n

s_1 and s_2 have been diverging from a common ancestor for t million years (MY)
according to the **Jukes Cantor (JC) model with parameter α**

- This framework does not model indels
- Assumes site independence
- **All substitutions are equally probable**
- **Parameter α is estimated from data**

