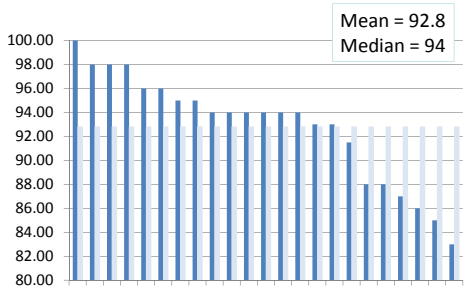
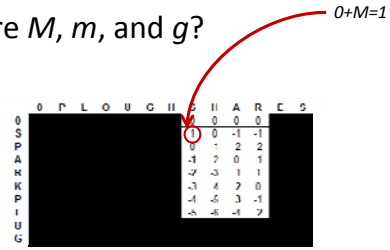
Exam 1 scores



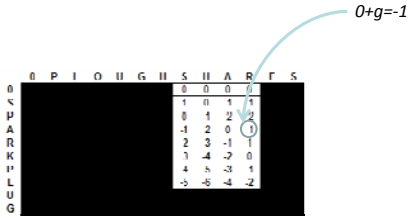
1(b) What are *M*, *m*, and *g*?

M = 1



1(b)

M = 1
g = -1



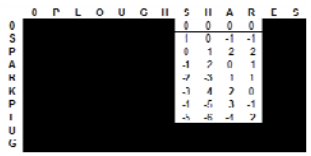
1(b)

M = 1
g = -1
m = -2

Three possibilities

- ~~*m > -2*~~ Not possible because *1+m* would replace *-1* with a better score.
- *m = -2* *1+m = 1+2g = -1*
- ~~*m < -2*~~ Violates *m > 2g*

1(b) Which algorithm?



Semi-global alignment
- Not global: top row has 0's
- Not local: the matrix has some negative values

1(c) What are *M*, *m*, and *g*?

1(c)

d(x,x) = 0

A 0-valued match cell in the first column ->
- Distance function
- *d(x,x) = 0*
- First column must be 0's

1(c)

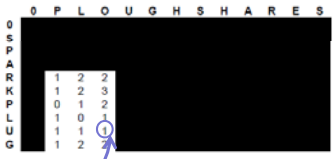
d(x,x) = 0
d(x,_) = 1

min { *0+g*, *1+m*, *1+g* }

0+g: -> *d(x,_) = g = 1*
~~*1+m:*~~ Require *d(x,y)>0*
~~*1+g:*~~ 1+g > 0+g

1(c)

d(x,x) = 0
d(x,_) = 1
d(x,y)= 1



min { *1+g*, *0+m*, *1+g* }

~~*1+g*~~: ~~-> 1+g = 2~~
0+m: -> m= *d(x,y)= 1*
~~*1+g*~~: ~~-> 1+g = 2~~

1(c) Which algorithm?

Semi-global alignment
- Not global: First column must have 0s, if the second column has 0s and 1s
- Not local: Local alignment requires a similarity function

3) Consider *k* sequences s_1, s_2, … s_k and let *M = 1* and *m = g = -1* be the scoring fn.
Let
- *X* be the score of the optimal global alignment of s_1 and s_2.
- *Y* be the score of the optimal local alignment of s_1 and s_2.
- *Z* be the score of the pairwise alignment of s_1 and s_2 induced by the optimal global alignment of s_1, s_2, … s_k .

Give an inequality that expresses the relationship between *X, Y* and *Z*.
Explain your answer.

$$Y \geq X$$

Why? *Y* must be positive. *X* can be negative.

3) Consider *k* sequences s_1, s_2, … s_k and let *M = 1* and *m = g = -1* be the scoring fn.
Let
- *X* be the score of the optimal global alignment of s_1 and s_2.
- *Y* be the score of the optimal local alignment of s_1 and s_2.
- *Z* be the score of the pairwise alignment of s_1 and s_2 induced by the optimal global alignment of s_1, s_2, … s_k .

Give an inequality that expresses the relationship between *X, Y* and *Z*.
Explain your answer.

~~*Y > X*~~

	0	S	A	N	K	A
0	0	0	0	0	0	0
S	0	1	0	0	0	**0**
A	0	0	2	1	0	1
N	0	0	1	3	2	1
T	0	0	0	2	2	1
A	0	0	1	1	1	3

Y = local alignment

	0	S	A	N	K	A
0	0	1	2	3	4	5
S	1	1	1	2	3	4
A	2	1	2	1	2	4
N	3	2	1	3	2	3
T	4	3	2	2	2	2
A	5	4	4	3	2	3

X = global alignment

3) Consider k sequences $s_1, s_2, \ldots s_k$ and let $M = 1$ and $m = g = -1$ be the scoring fn.
Let

- X be the score of the optimal global alignment of s_1 and s_2.
- Y be the score of the optimal local alignment of s_1 and s_2.
- Z be the score of the pairwise alignment of s_1 and s_2 induced by the optimal global alignment of $s_1, s_2, \ldots s_k$.

Give an inequality that expresses the relationship between X, Y and Z.
Explain your answer.

$$X \geq Z$$

Why? The induced alignment reflects relationships in the entire family. The optimal alignment of s_1 and s_2 with respect to $s_3, \ldots s_k$ may result in a sub-optimal alignment between s_1 and s_2.

```
A_CT
AG_T
ACGT
```

3) Consider k sequences $s_1, s_2, \ldots s_k$ and let $M = 1$ and $m = g = -1$ be the scoring fn.
Let

- X be the score of the optimal global alignment of s_1 and s_2.
- Y be the score of the optimal local alignment of s_1 and s_2.
- Z be the score of the pairwise alignment of s_1 and s_2 induced by the optimal global alignment of $s_1, s_2, \ldots s_k$.

Give an inequality that expresses the relationship between X, Y and Z.
Explain your answer.

$$Y \geq X \geq Z$$

5(a)

Exact dynamic programming requires time t_1 to obtain an MSA of k sequences of length n. Let t_2 be the time required to obtain an MSA of k sequences of length $3n$ using the same method. What is t_2/t_1?

5(a)

$$t_1 = O(n^k 2^k k^2)$$

$$t_2 = O((3n)^k 2^k k^2)$$

$$\frac{t_2}{t_1} = \frac{O((3n)^k 2^k k^2)}{O(n^k 2^k k^2)}$$

$$\frac{t_2}{t_1} = 3^k$$

5(b)

Exact dynamic programming requires time t_1 to obtain an MSA of k sequences of length n. Let t_3 be the time required to obtain an MSA of $3k$ sequences of length n using the same method. What is t_3/t_1?

5(b)

$$t_1 = O(n^k 2^k k^2)$$

$$t_3 = O(n^{3k} 2^{3k} (3k)^2)$$

$$\frac{t_3}{t_1} = \frac{O(n^{3k} 2^{3k} (3k)^2)}{O(n^k 2^k k^2)}$$

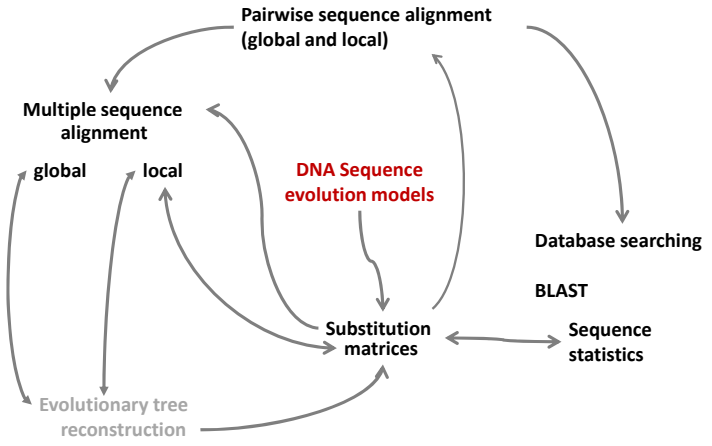
$$\frac{t_3}{t_1} = O(n^{2k} 2^{2k})$$

5(c)

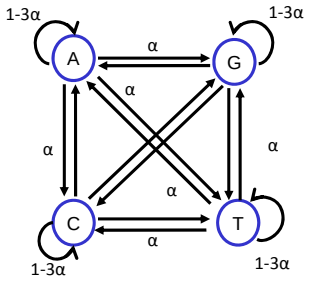
$$O(3^k) \text{ versus } O(n^{2k} 4^k)$$

Increasing k has more impact because

- "increasing k is exponentially increasing the time, while increasing n is adding a constant multiplier"
- "This makes sense because by increasing n, you're simply adding extra boxes in the same dimensions, but by increasing k, you're adding extra dimensions"



Jukes-Cantor model: sequence substitution at a single site



Rate of substitutions

$P(xy) = 3\alpha$

Probability nucleotide x remains unchanged

$P(xx) = 1-3\alpha$

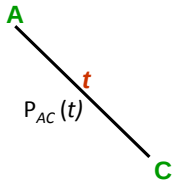
In a discrete time framework, α is the probability of a given substitution occurring in a single time step.

$\pi^* = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$

From this, we can derive several quantities of interest…

Given ancestral nucleotide *z*, the probability of observing nucleotide *x* after time *t* is given by

$P_{xx}(t) = ¼ + ¾e^{-4\alpha t}$,
$P_{zx}(t) = ¼ - ¼e^{-4\alpha t}$, $y \neq x$



Note that:

- At *t=0,* $P_{xx}(0) = 1$ and $P_{xy}(0) = 0$. This makes sense because if no time has elapsed, then no substitution can have occurred (yet).
- As $t \rightarrow \infty$, $P_{xx}(t) = P_{xy}(t) = 0.25$. This says that the steady state distribution of nucleotide frequencies is uniform under the Jukes Cantor model

2. Given an alignment of *n* nucleotides that differs at *m* positions, the expected number of substitutions since the divergence of the two sequences is given by

$$E[\text{sub}] = -\frac{3}{4}\ln\left(1 - \frac{4}{3}\frac{m}{n}\right)n.$$

For example, if we observe 200 mismatches in an alignment of 1000 nucleotides, then the number of actual substitutions is

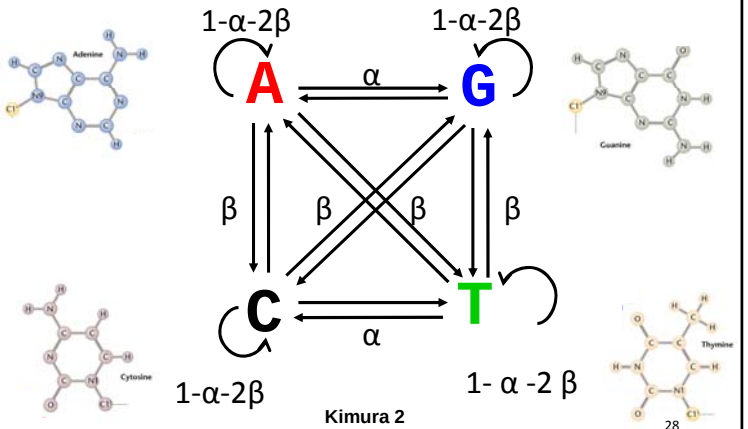
$-\frac{3}{4}\ln(1 - \frac{800}{3000}) = 233$ substitutions

…CACATACGAAGATACGAACGAGC…
…CAGATAGGAAGAGACGATCTAGC…
n nucleotides with *m* mismatches

Note that:

- If *m = 0,* then *E[sub] = 0* and the distance between the sequences is zero.
- *m/n* ≤ *.75* in sequences governed by the Jukes Cantor model.
- As *m/n* → 0.75, E[sub] → ∞. This is because once we reach the steady state distribu- tion of nucleotide frequencies, *m/n* provides no information about how long the sequences have been diverging.

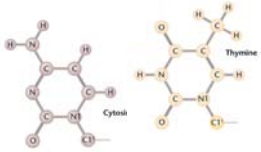
A more complex model
different probabilities for transitions and transversions

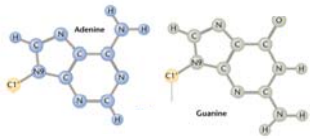


Kimura 2 parameter

Transitions and Transversions

Pyrimidines have one ring



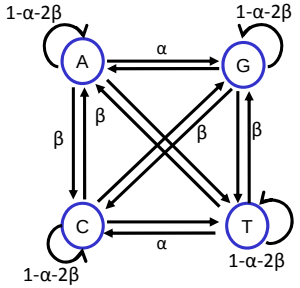


Purines have two rings

Transitions: substitutions within the same class of nucleotide
(purine – purine or pyrimidine-pyrimidine)

Transversions: substitutions between classes
(purine – pyrimidine or pyrimidine-purine)

Kimura 2 Parameter model



Rate of substitutions

$P(xy) = \alpha+2\beta$

Probability nucleotide x remains unchanged

$P(xx) = 1-\alpha-2\beta$

in a single time step

$\pi = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$

From this, we can derive several quantities of interest…

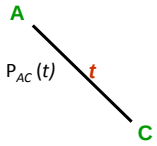
Kimura 2 Parameter model

Given ancestral nucleotide *z*, the probability of observing nucleotide *x* after time *t* is given by

$P_{xx}(t) = \frac{1}{4}(1 + e^{-4\beta t} + 2e^{-2(\alpha+\beta)t})$,

$P_{zx}(t) = \frac{1}{4}(1 + e^{-4\beta t} - 2e^{-2(\alpha+\beta)t})$, *z→x is a transition*

$P_{zx}(t) = \frac{1}{4}(1 - e^{-4\beta t})$, *z→x is a transversion*



Given an alignment of *n* nucleotides that differs at $m = m_s + m_v$ positions, where

m_s = number of transitions ,

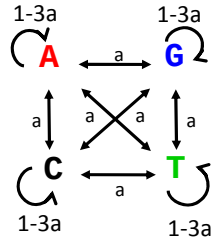
m_v = number of transversions,

the expected number of substitutions is given by

$$E[\text{sub}] = \left[-\frac{1}{2}\ln\left(1 - \frac{m_s}{n} - \frac{m_v}{n}\right) - \frac{3}{4}\ln\left(1 - \frac{4}{3}\frac{m_v}{n}\right)\right] n$$

Jukes-Cantor model (1969)

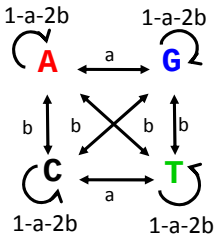
p(A)=0.25
p(G)=0.25
p(C)=0.25
p(T)=0.25



- All substitutions have equal probability
- Base frequencies are equal

Kimura 2 parameter model (K2P) (1980)

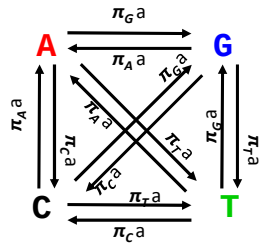
*p(**A**)=0.25*
*p(**G**)=0.25*
*p(**C**)=0.25*
*p(**T**)=0.25*



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

Felsenstein (1981)

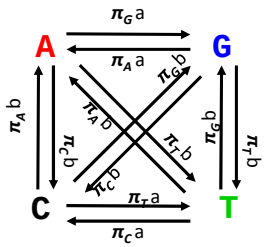
*p(**A**)=π_A*
*p(**G**)=π_G*
*p(**C**)=π_C*
*p(**T**)=π_T*



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

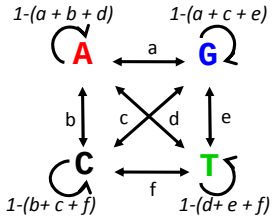
Hasegawa, Kishino & Yano (HKY) (1985)

*p(**A**)=π_A*
*p(**G**)=π_G*
*p(**C**)=π_C*
*p(**T**)=π_T*



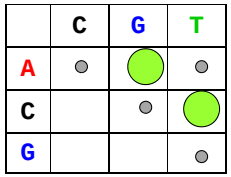
- Transitions and transversions have different probabilities
- Unequal base frequencies $p(A) \neq p(G) \neq p(C) \neq p(T)$

General Time Reversible model



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

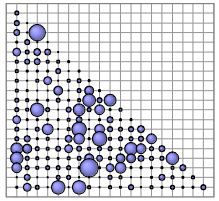
DNA substitution models



DNA substitution model
(e.g., JC, K2P, GTr)

- Four states (A, C, G, T)
- Model specifies the probability of substitution for all possible pairs of nucleotides

Amino acid substitution models



Amino acid substitution matrix
(e.g., PAM, WAG, JTT, MtREV etc)

- Twenty states (A, C, … Y)
- Model specifies the probability of substitution for all possible pairs of amino acids

43