BI694
Bioinformatics & Phylogenetics

Winter Semester 2017

WEEK 10

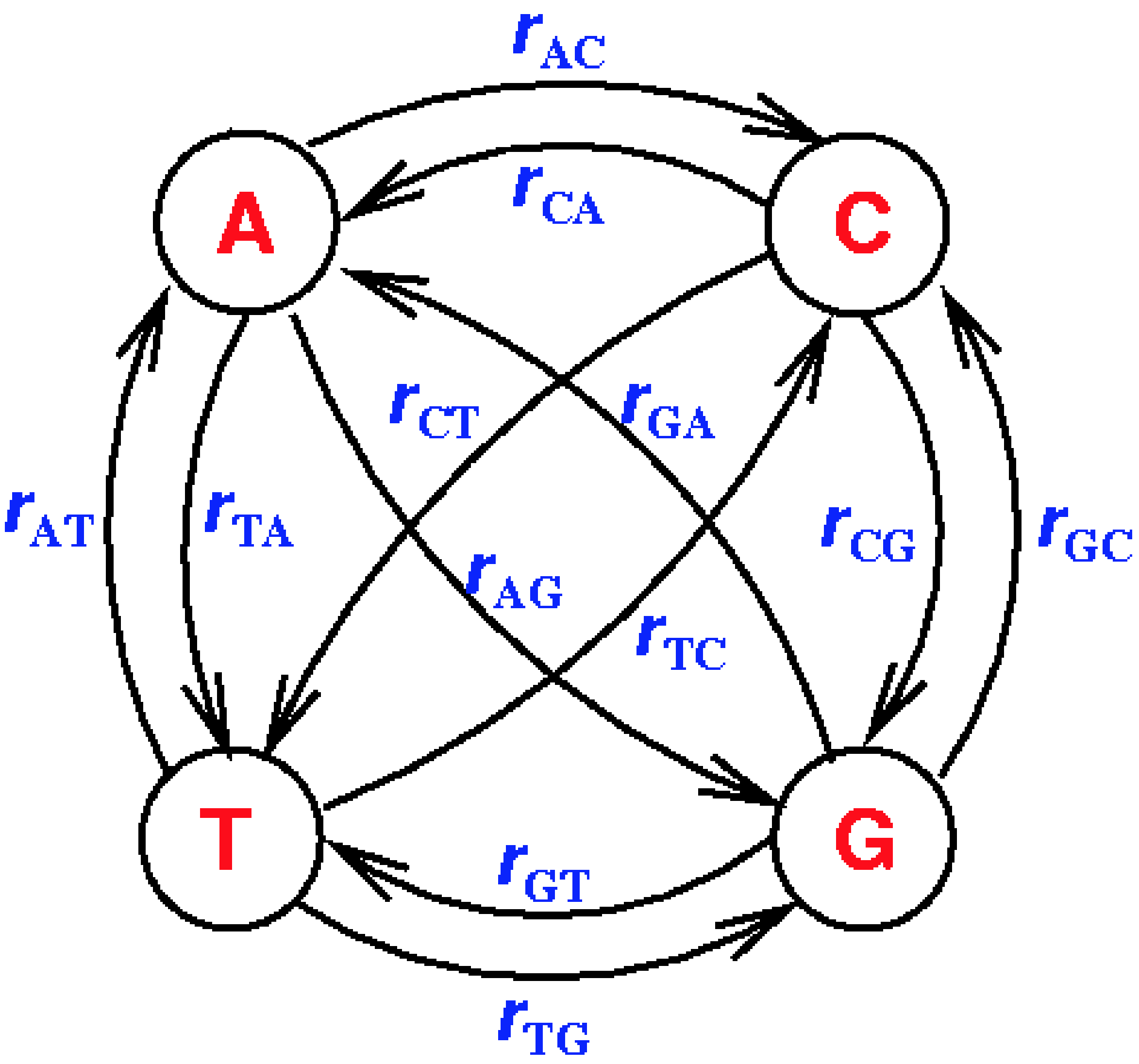
Intro to Phylogenetics and Parsimony

Models of Molecular Evolution

Substitution model specifies way in which characters evolve

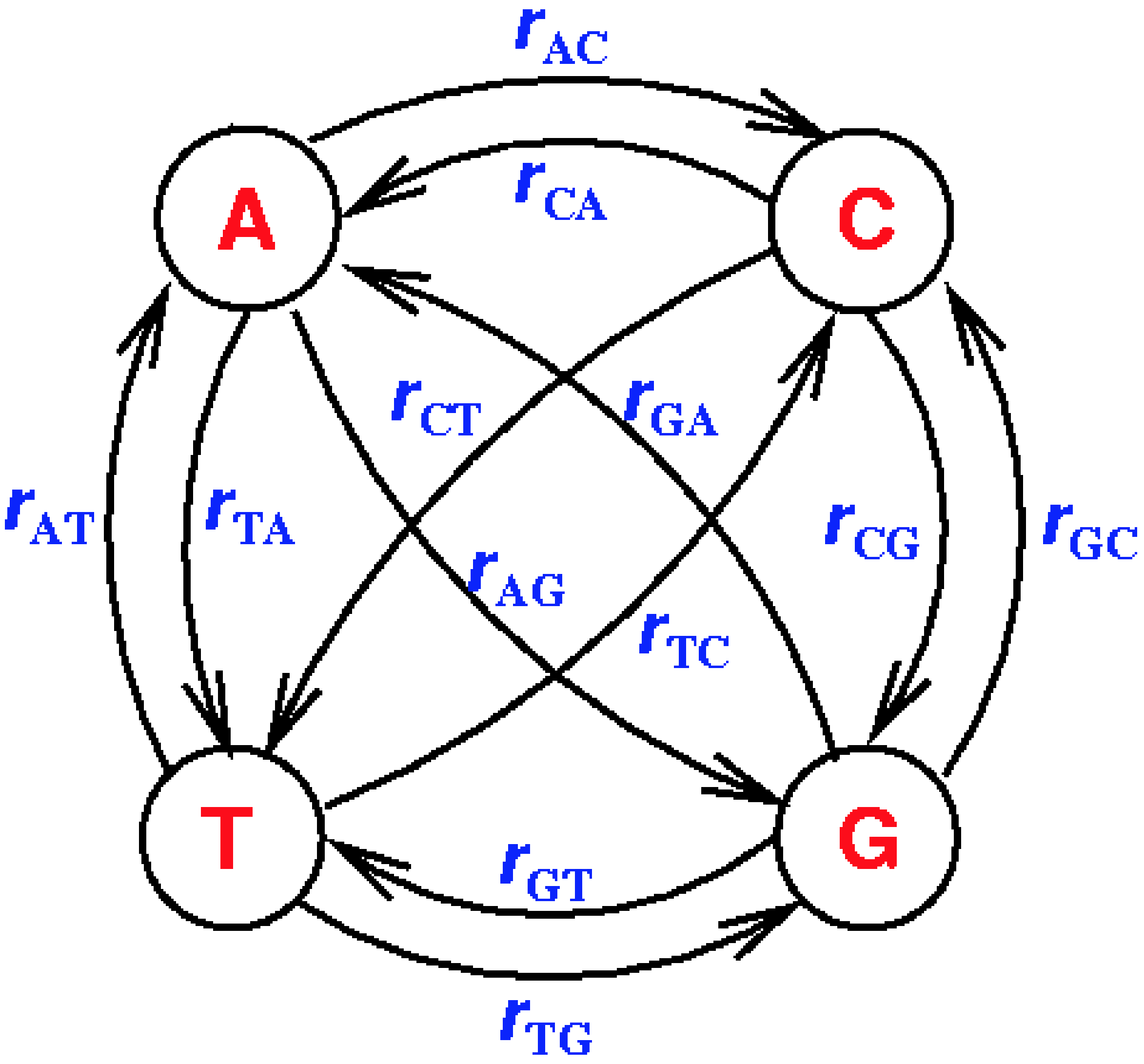
Models of Molecular Evolution

Substitution model specifies way in which characters evolve



Models of Molecular Evolution

Substitution model specifies way in which characters evolve

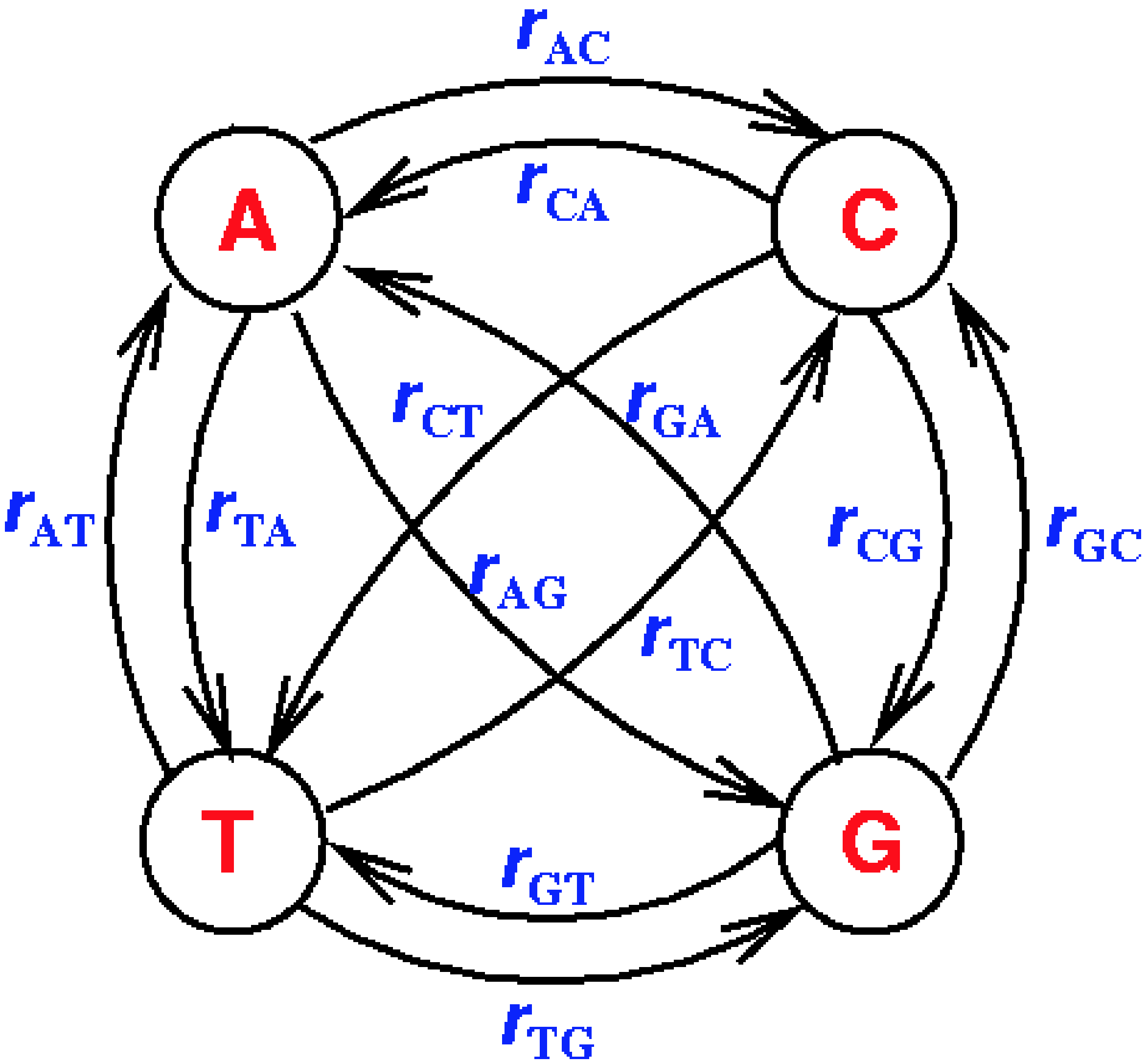


Markov models:

Process in which probability of an event happening in some time window is dependent only on the state at that time and independent on how it came to be in that state (eg, coin toss).

Models of Molecular Evolution

Substitution model specifies way in which characters evolve



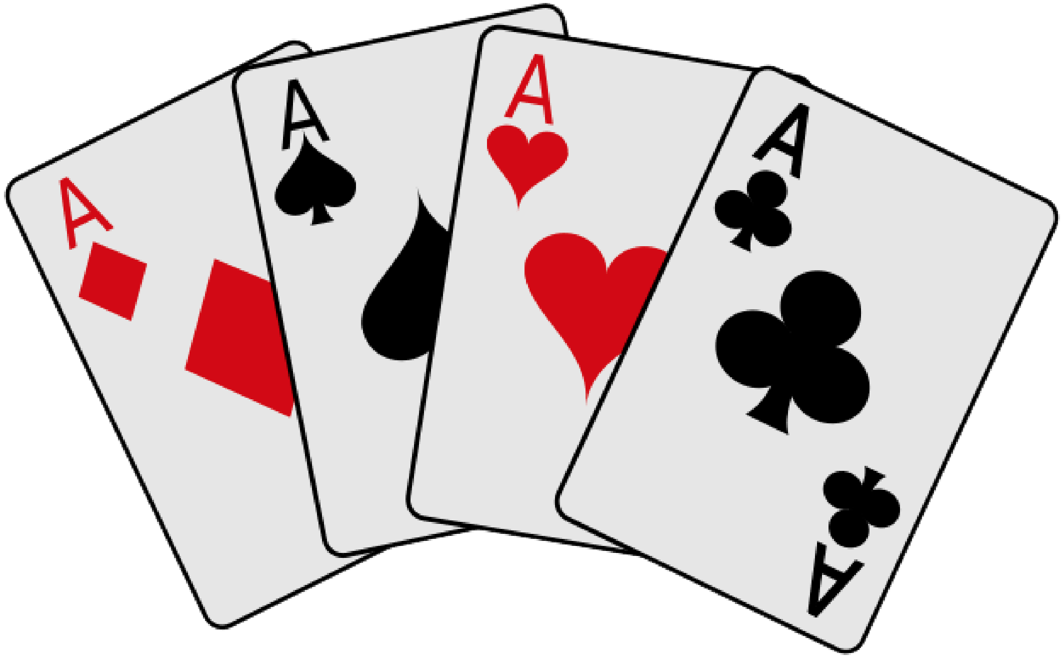
DNA:

A, G, C, T (gaps treated as missing data)

Time reversible models:

A → T equals T → A

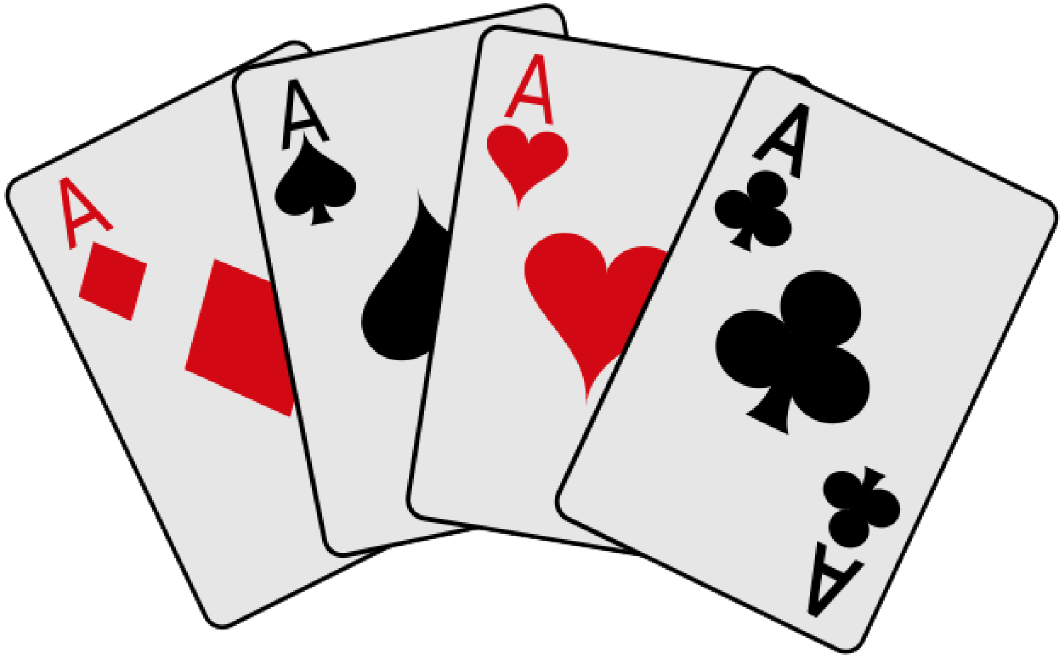
Models of Molecular Evolution



Imagine we have a card on the table. Every so often the card is placed in the deck and replaced by drawing a card from the deck. If the new card if the same suit there would be no visible change. Replacements by cards of the same suit would be invisible or “hidden” to our eye.

The rate or frequency with which card changes occur is denoted by **μ**

Models of Molecular Evolution



Let's say the card changes with a rate of 0.6 substitutions per minute...

We expect to see 36 (0.6 x 60 minutes) changes per hour and average wait times Between changes of 1 minute and 40 seconds (1/0.6 minutes).

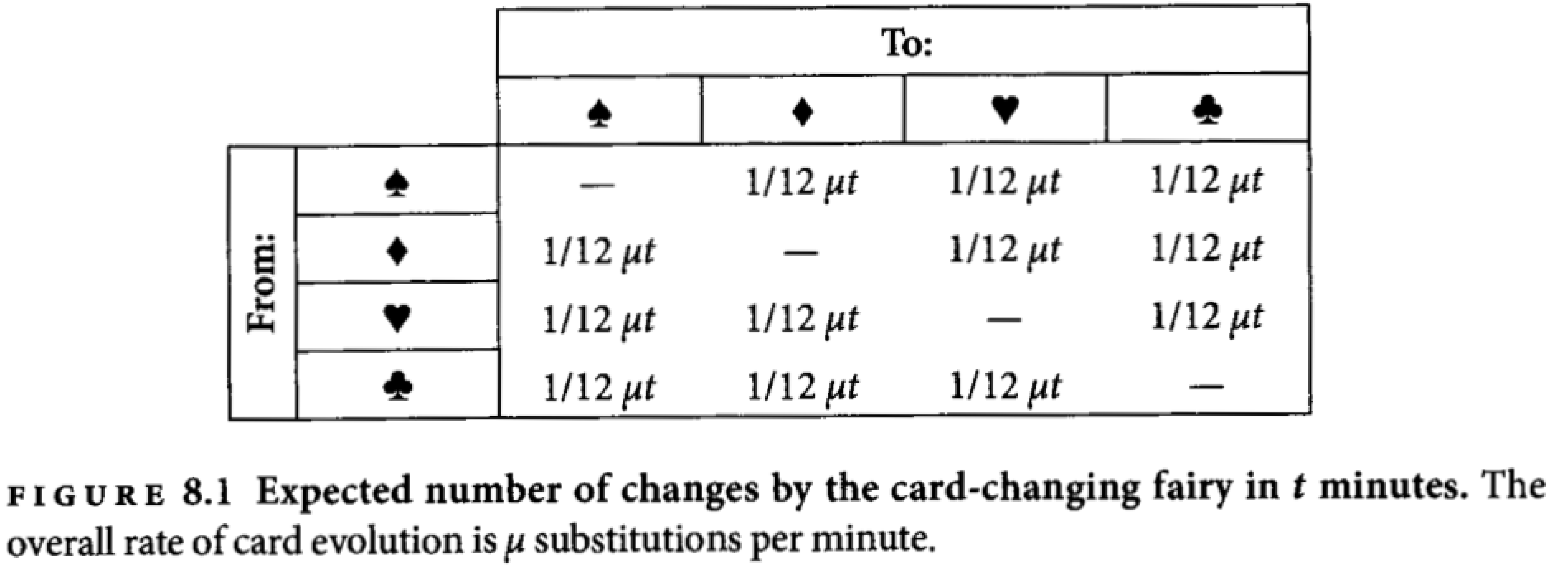
Watching for a long time you will see the card cycling through all possible suits.

Models of Molecular Evolution

Let's say the card changes with a rate of 0.6 substitutions per minute...

We expect to see 36 (0.6 x 60 minutes) changes per hour and average wait times
Between changes of 1 minute and 40 seconds (1/0.6 minutes).

Watching for a long time you will see the card cycling through all possible suits.



		To:			
		♠	♦	♥	♣
From:	♠	—	1/12 μt	1/12 μt	1/12 μt
	♦	1/12 μt	—	1/12 μt	1/12 μt
	♥	1/12 μt	1/12 μt	—	1/12 μt
	♣	1/12 μt	1/12 μt	1/12 μt	—

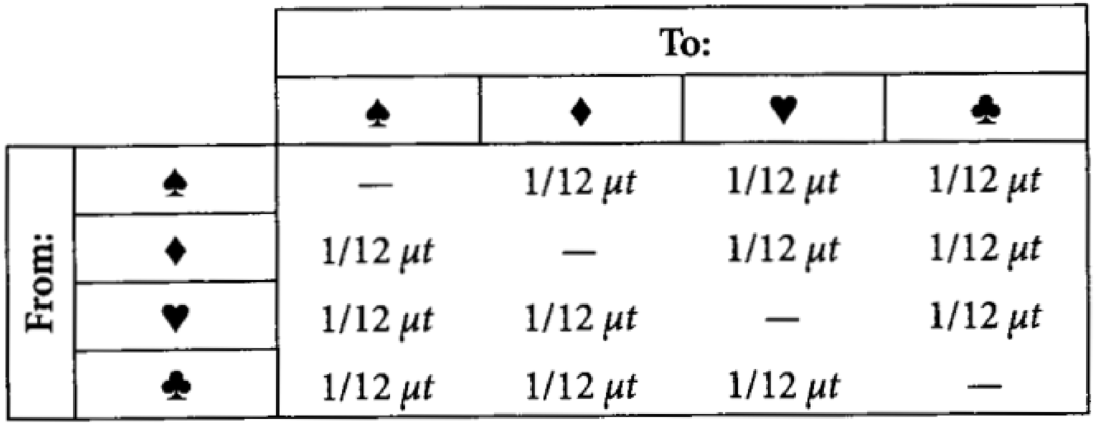
FIGURE 8.1 Expected number of changes by the card-changing fairy in t minutes. The overall rate of card evolution is μ substitutions per minute.

Models of Molecular Evolution

Let's say the card changes with a rate of 0.6 substitutions per minute...

We expect to see 36 (0.6 x 60 minutes) changes per hour and average wait times Between changes of 1 minute and 40 seconds (1/0.6 minutes).

Overall rate of visible change is μ and thus each of the possible changes of suits occurs at a rate of μ/12.

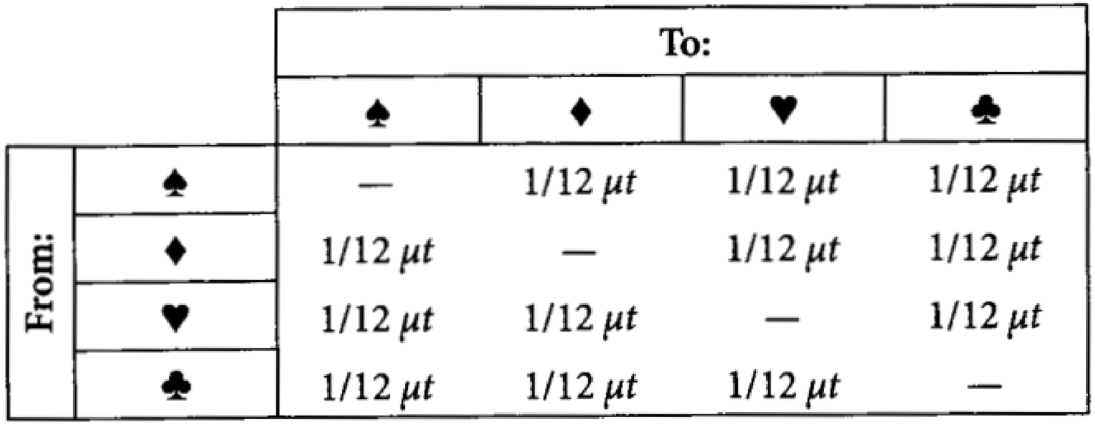


		To:			
		♠	♦	♥	♣
From:	♠	—	1/12 μt	1/12 μt	1/12 μt
	♦	1/12 μt	—	1/12 μt	1/12 μt
	♥	1/12 μt	1/12 μt	—	1/12 μt
	♣	1/12 μt	1/12 μt	1/12 μt	—

FIGURE 8.1 Expected number of changes by the card-changing fairy in t minutes. The overall rate of card evolution is μ substitutions per minute.

Models of Molecular Evolution

At 0.6 changes per minute overall we expect 0.05 (0.6/12) specific changes per minute.

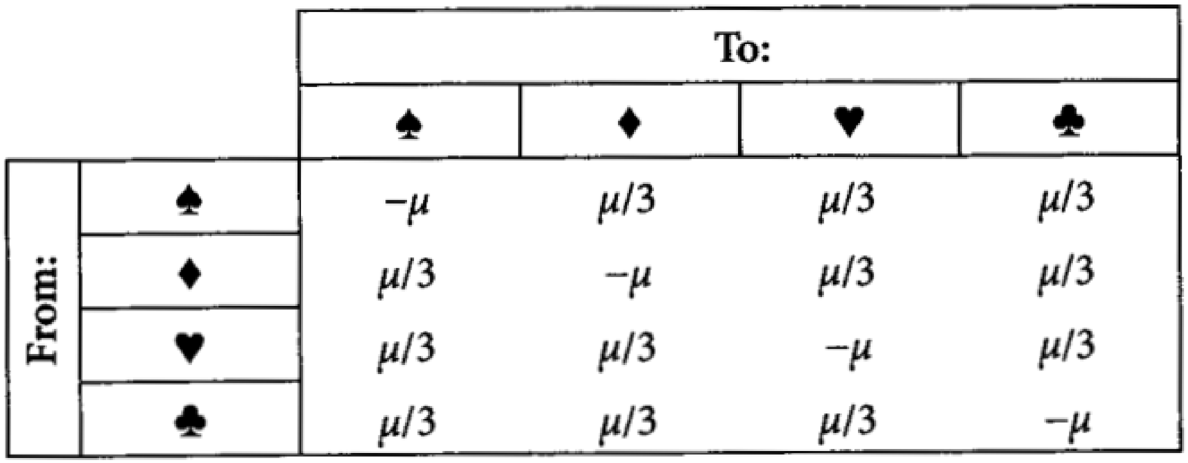


		To:			
		♠	♦	♥	♣
From:	♠	—	1/12 μt	1/12 μt	1/12 μt
	♦	1/12 μt	—	1/12 μt	1/12 μt
	♥	1/12 μt	1/12 μt	—	1/12 μt
	♣	1/12 μt	1/12 μt	1/12 μt	—

FIGURE 8.1 **Expected number of changes by the card-changing fairy in t minutes.** The overall rate of card evolution is μ substitutions per minute.

Models of Molecular Evolution

Instantaneous Rate Matrix:
Rate at which the suit to the left will change to another suit. Since the rate of change is μ, and there are 3 alternative suits of equal frequency in the deck, the rate of change is μ/3. Staying in the same suit has a rate of -μ so that each row sums to zero (the net rate of leaving the row ought to be zero).

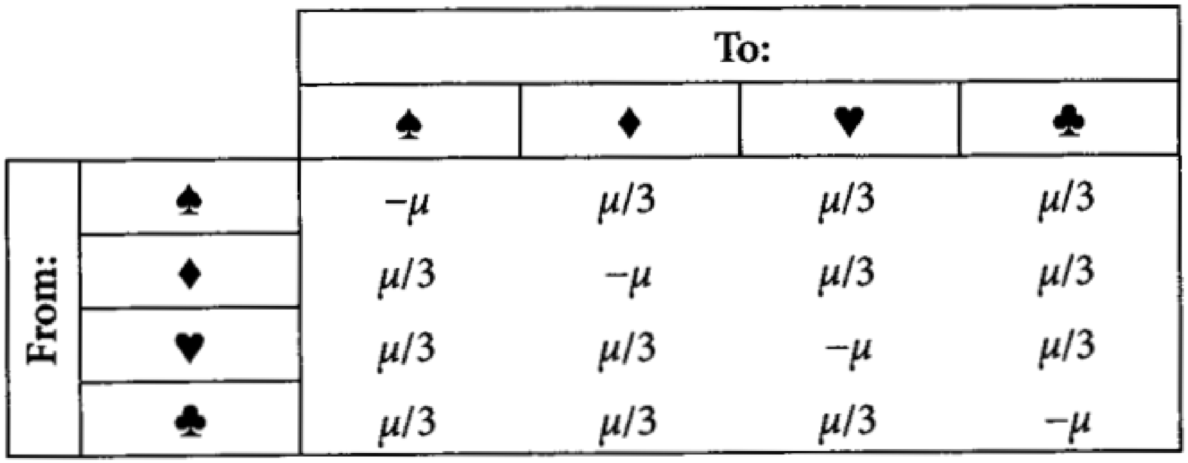


		To:			
		♠	♦	♥	♣
From:	♠	$-\mu$	$\mu/3$	$\mu/3$	$\mu/3$
	♦	$\mu/3$	$-\mu$	$\mu/3$	$\mu/3$
	♥	$\mu/3$	$\mu/3$	$-\mu$	$\mu/3$
	♣	$\mu/3$	$\mu/3$	$\mu/3$	$-\mu$

FIGURE 8.2 Instantaneous rates of substitution by the card-changing fairy. The overall rate of card evolution is μ substitutions per minute.

Models of Molecular Evolution

For DNA sequences we want to determine how long ago two sequences shared a common ancestor. We focus on how far apart two taxa are – ie, we calculate their evolutionary distance.



		To:			
		♠	♦	♥	♣
From:	♠	$-\mu$	$\mu/3$	$\mu/3$	$\mu/3$
	♦	$\mu/3$	$-\mu$	$\mu/3$	$\mu/3$
	♥	$\mu/3$	$\mu/3$	$-\mu$	$\mu/3$
	♣	$\mu/3$	$\mu/3$	$\mu/3$	$-\mu$

FIGURE 8.2 **Instantaneous rates of substitution by the card-changing fairy.** The overall rate of card evolution is μ substitutions per minute.

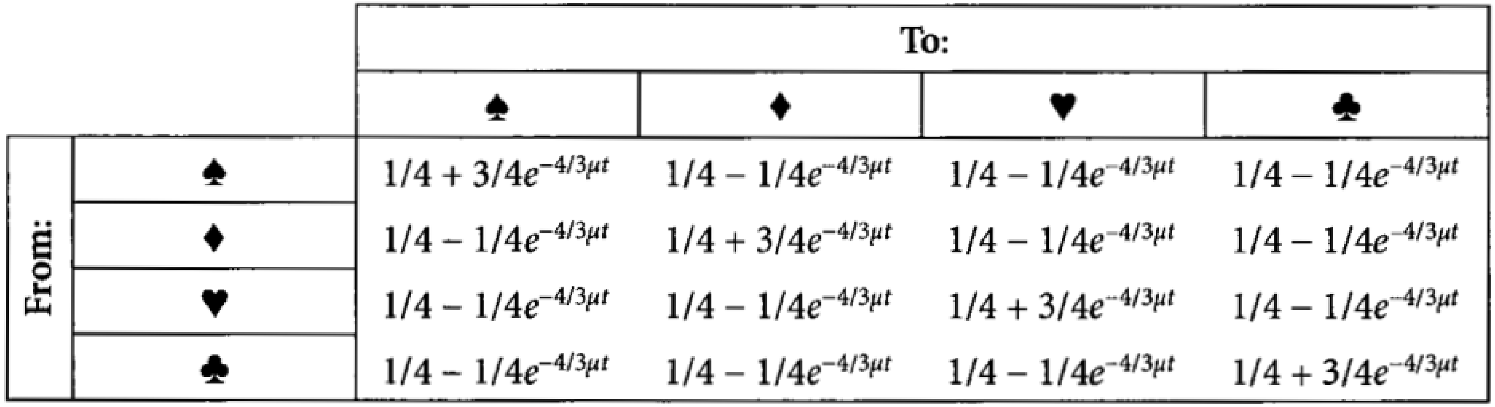
Models of Molecular Evolution

Suppose we leave the room for 10 minutes... Based on the suit of the card observed when we return, can we say anything about the number of changes that occurred while we were gone?

Models of Molecular Evolution

Suppose we leave the room for 10 minutes... Based on the suit of the card observed when we return, can we say anything about the number of changes that occurred while we were gone?

We need to know how the probability of observing each of the four suits changes as a function of the substitution rate, μ and the time, t.

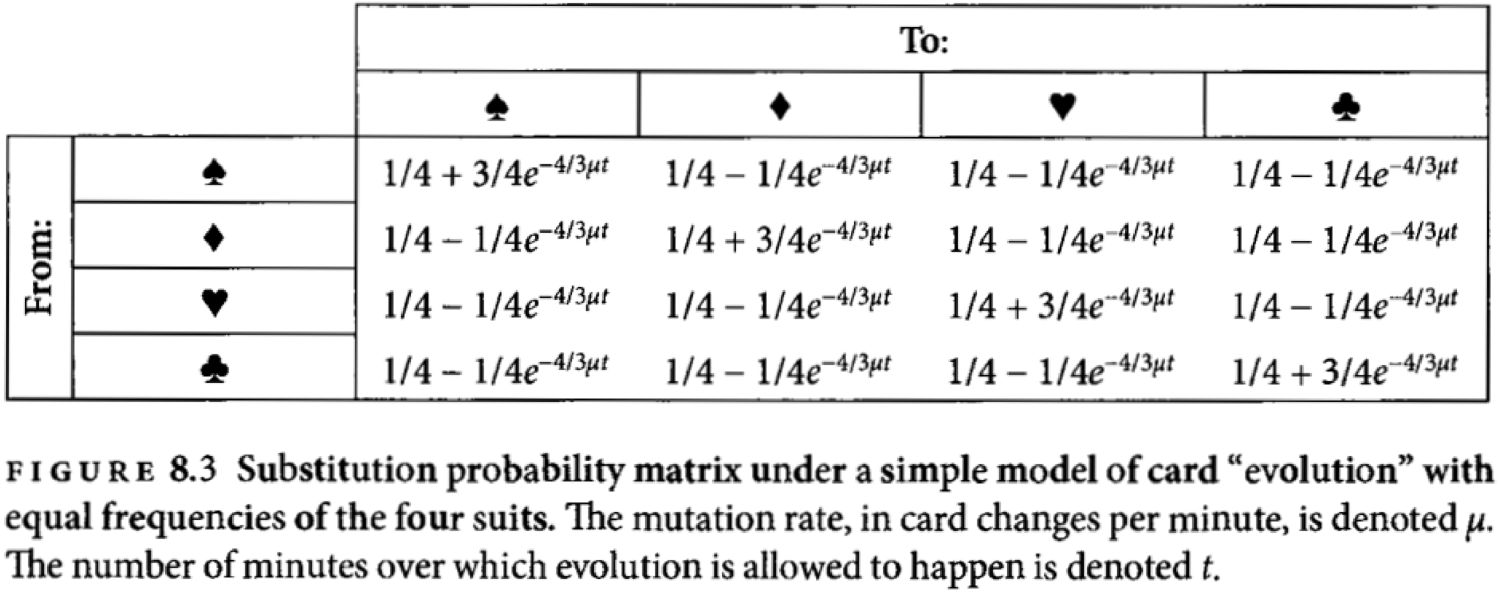


		To:			
		♠	♦	♥	♣
From:	♠	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	♦	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	♥	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	♣	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$

FIGURE 8.3 **Substitution probability matrix under a simple model of card "evolution" with equal frequencies of the four suits.** The mutation rate, in card changes per minute, is denoted μ. The number of minutes over which evolution is allowed to happen is denoted t.

Models of Molecular Evolution

What is the probability of starting as a spade and ending as a spade a **short** time later?



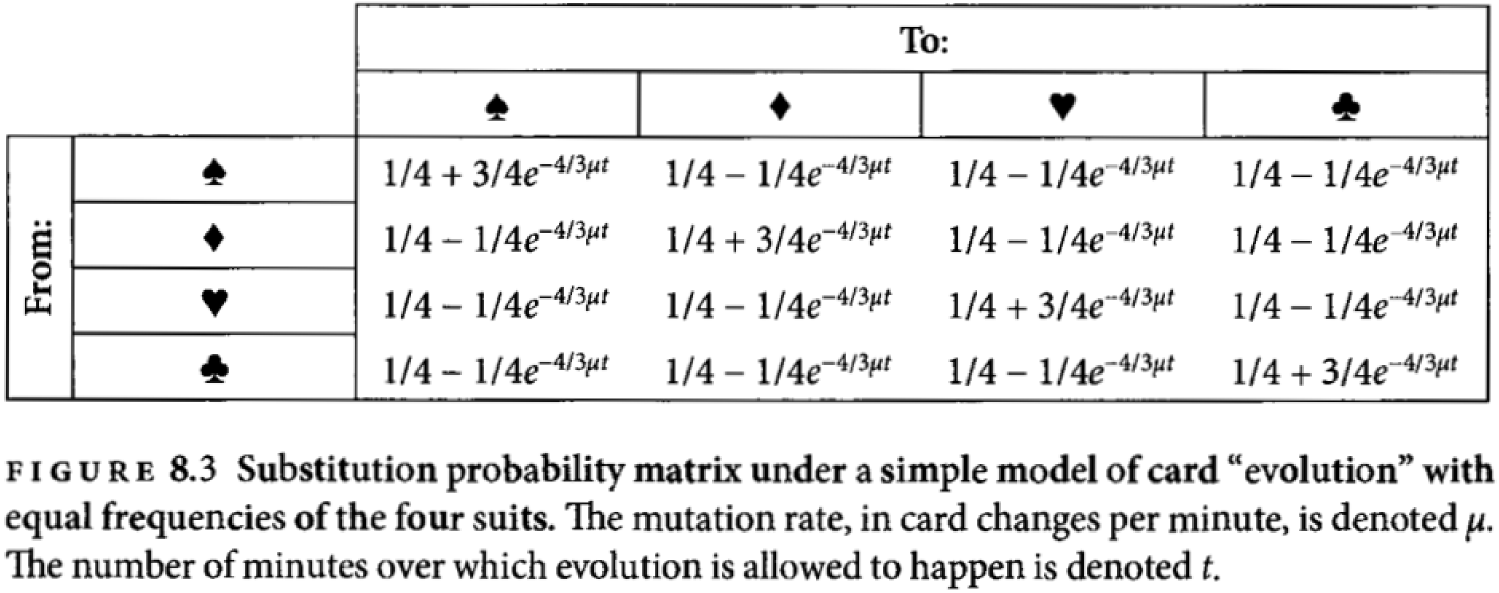
		To:			
		♠	♦	♥	♣
From:	♠	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	♦	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	♥	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	♣	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$

FIGURE 8.3 Substitution probability matrix under a simple model of card "evolution" with equal frequencies of the four suits. The mutation rate, in card changes per minute, is denoted μ. The number of minutes over which evolution is allowed to happen is denoted t.

Models of Molecular Evolution

What is the probability of starting as a spade and ending as a spade a **short** time later?

If µt is small then e to the power of -µt is ~1.0 (a number to the power of 0 = 1). In that case probability of spade being a spade is ~1 (¼ + ¾).



		To:			
		♠	♦	♥	♣
From:	♠	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	♦	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	♥	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	♣	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$

FIGURE 8.3 **Substitution probability matrix under a simple model of card "evolution" with equal frequencies of the four suits.** The mutation rate, in card changes per minute, is denoted μ. The number of minutes over which evolution is allowed to happen is denoted t.

Models of Molecular Evolution

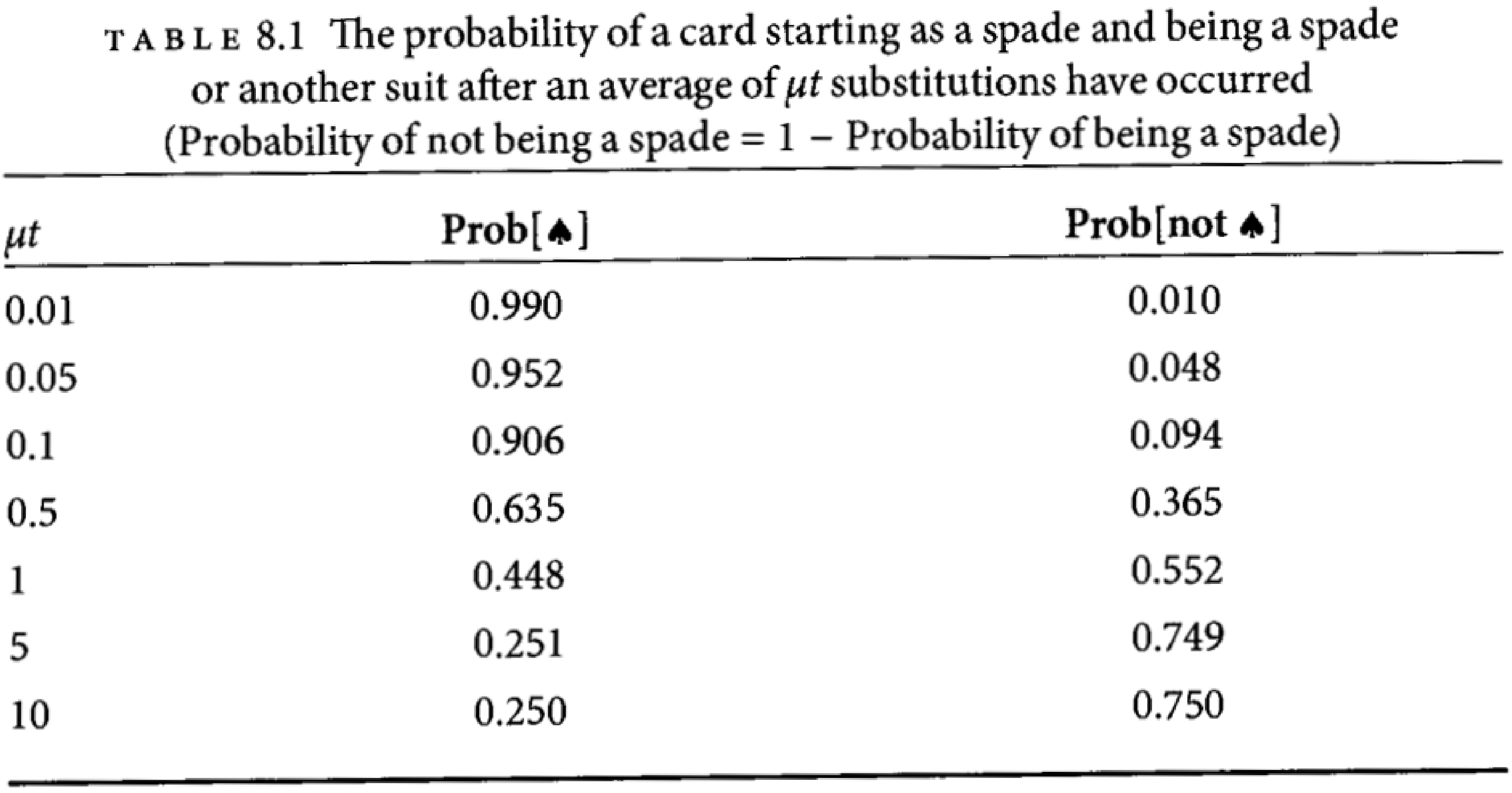
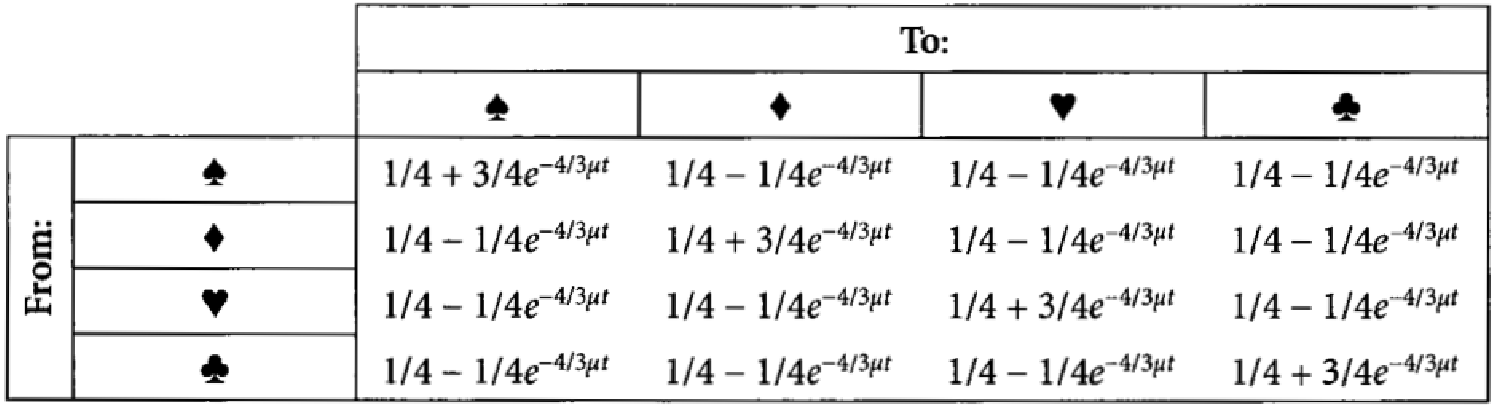


TABLE 8.1 The probability of a card starting as a spade and being a spade or another suit after an average of μt substitutions have occurred (Probability of not being a spade = 1 − Probability of being a spade)

μt	**Prob[♠]**	**Prob[not ♠]**
0.01	0.990	0.010
0.05	0.952	0.048
0.1	0.906	0.094
0.5	0.635	0.365
1	0.448	0.552
5	0.251	0.749
10	0.250	0.750

Models of Molecular Evolution

Without knowing μ we cannot determine the the evolutionary distance μt – there could have been few or many changes. μ cannot be estimated from a single card.



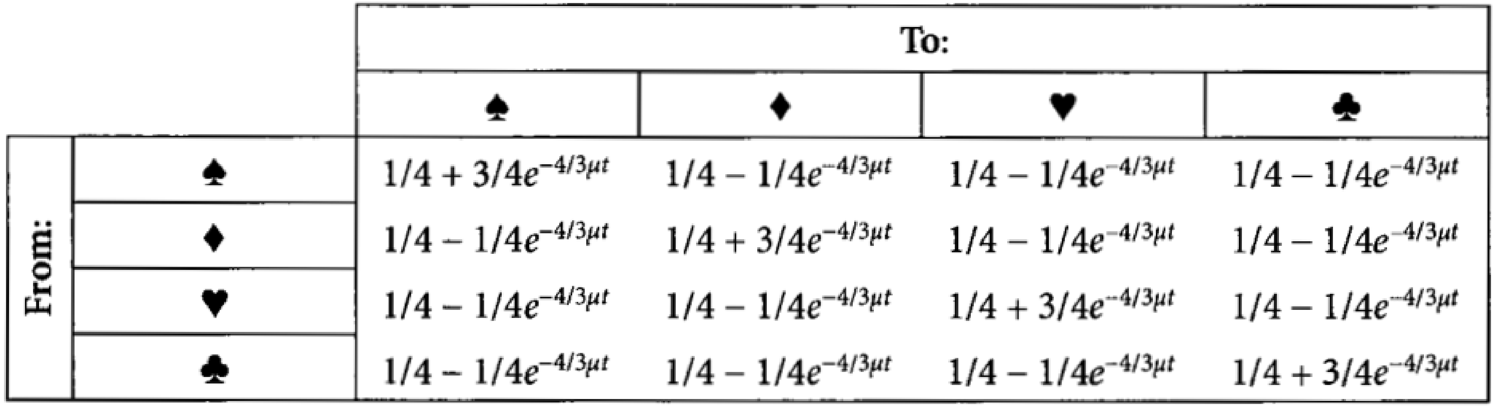
		To:			
		♠	♦	♥	♣
From:	♠	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	♦	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	♥	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	♣	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$

FIGURE 8.3 **Substitution probability matrix under a simple model of card "evolution" with equal frequencies of the four suits.** The mutation rate, in card changes per minute, is denoted μ. The number of minutes over which evolution is allowed to happen is denoted t.

Models of Molecular Evolution

Without knowing μ we cannot determine the the evolutionary distance μt – there could have been few or many changes. μ cannot be estimated from a single card.

Solution: we leave several cards behind. From several cards we can estimate the proportion of cards that changed suit.

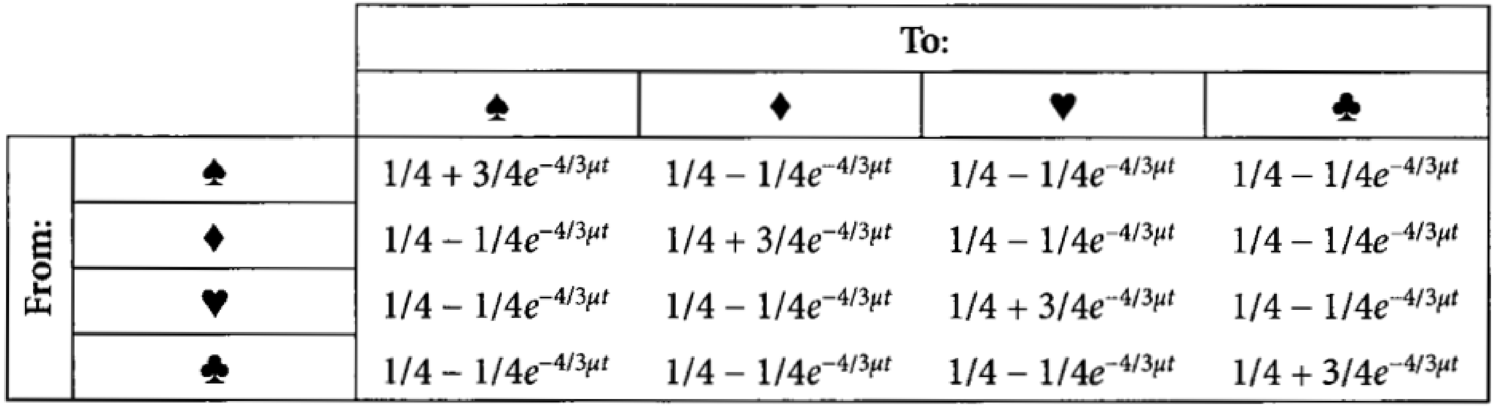


		To:			
		♠	♦	♥	♣
From:	♠	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	♦	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	♥	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	♣	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$

FIGURE 8.3 **Substitution probability matrix under a simple model of card "evolution" with equal frequencies of the four suits.** The mutation rate, in card changes per minute, is denoted μ. The number of minutes over which evolution is allowed to happen is denoted t.

Models of Molecular Evolution

Suppose we leave 100 cards and upon return we find 40 cards that changed suit.



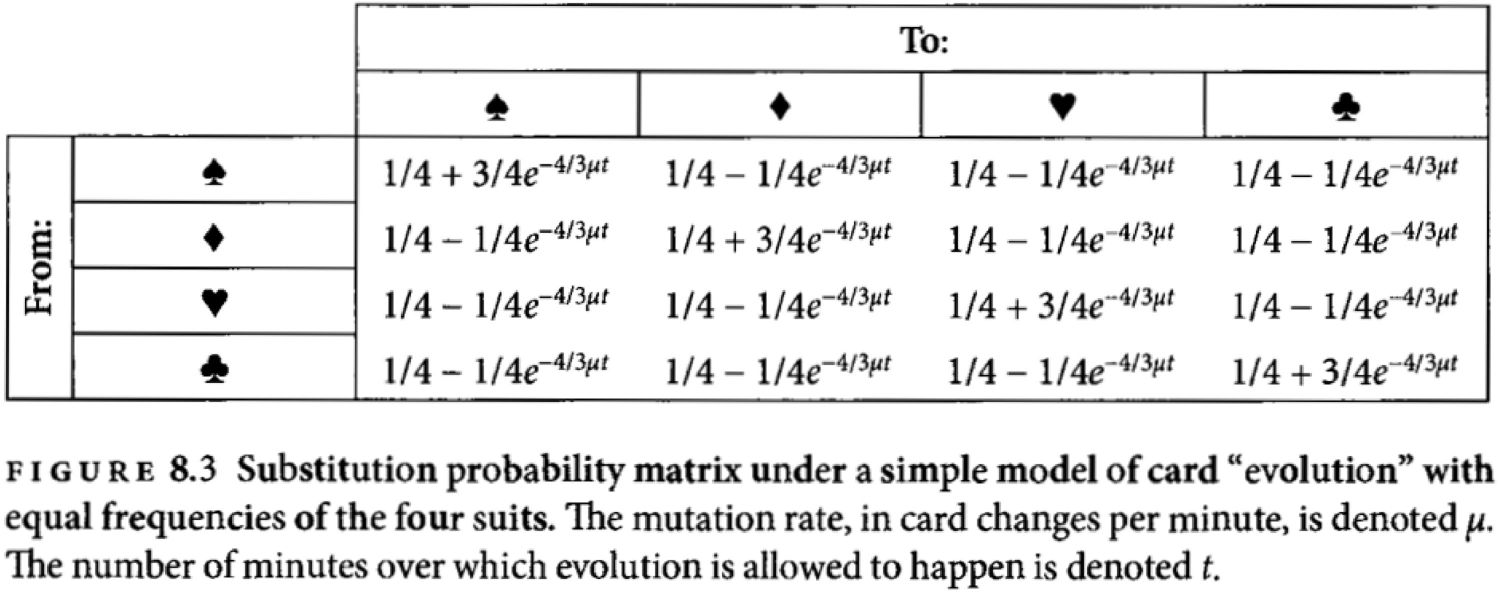
		To:			
		♠	♦	♥	♣
From:	♠	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	♦	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	♥	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	♣	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$

FIGURE 8.3 **Substitution probability matrix under a simple model of card "evolution" with equal frequencies of the four suits.** The mutation rate, in card changes per minute, is denoted μ. The number of minutes over which evolution is allowed to happen is denoted t.

Models of Molecular Evolution

Suppose we leave 100 cards and upon return we find 40 cards that changed suit.

This means that the ratio of unchanged cards is 0.6 (probability of no change is 0.6). Solving ¼ + ¾ $e^{-4/3\mu t}$ = 0.6 yields 0.572 μt (evolutionary distance). Using the substitution matrix we take hidden changes into account...



		To:			
		♠	♦	♥	♣
From:	♠	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	♦	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	♥	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	♣	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$

FIGURE 8.3 **Substitution probability matrix under a simple model of card "evolution" with equal frequencies of the four suits.** The mutation rate, in card changes per minute, is denoted μ. The number of minutes over which evolution is allowed to happen is denoted t.

Models of Molecular Evolution

The figure below shows that a frequency of 0.4 observed changes equals an evolutionary distance of 0.57. So what we did in the previous slide was estimate the number of hidden And observed changes from the observed changes alone.

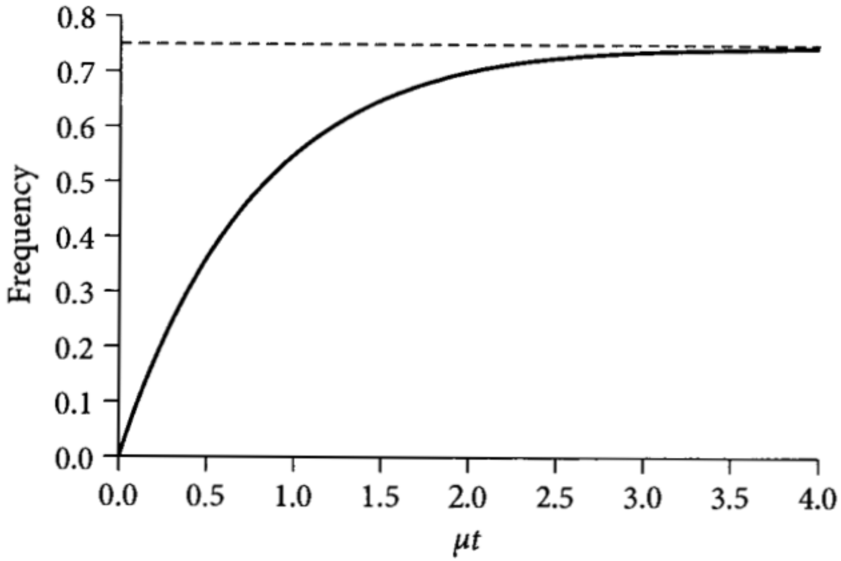
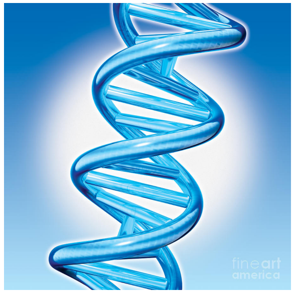


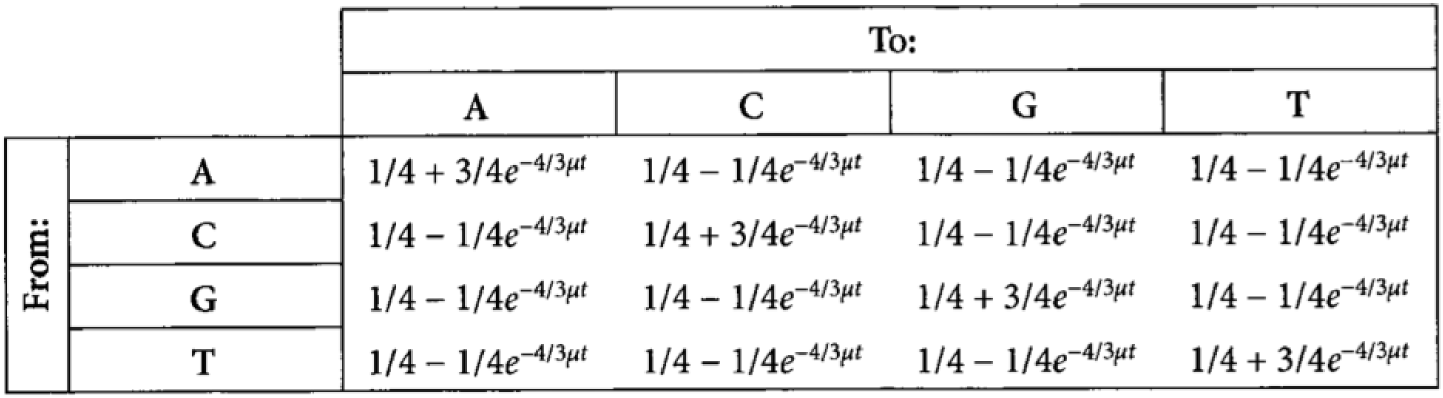
FIGURE 8.4 Expected frequency of cards that have a different suit from the ancestor as a function of μt.

Models of Molecular Evolution



Replace cards with nucleotides and you have Jukes Cantor (JC) model.

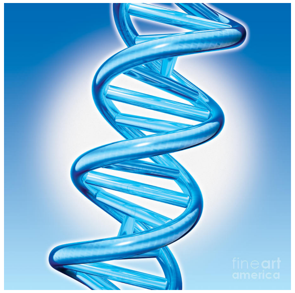
Core assumptions:
1) equal base frequencies
2) equal rates of substitution
3) rates of substitution are the same across the sequence



		To:			
		A	C	G	T
From:	A	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	C	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	G	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	T	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$

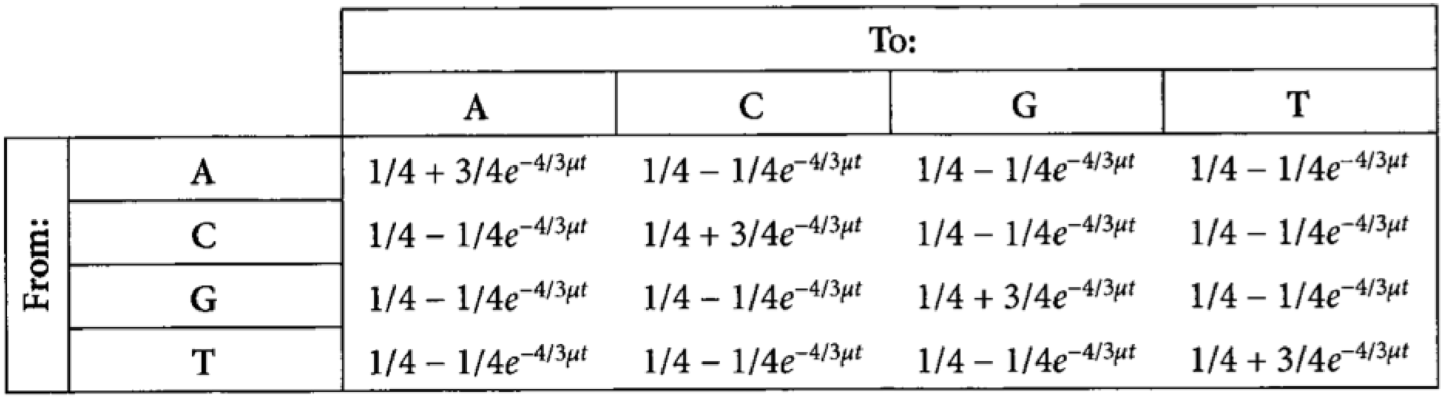
FIGURE 8.5 Substitution probability matrix under the JC model of DNA sequence evolution. The mutation rate, in substitutions per unit time, is denoted μ. The time interval over which evolution is allowed to happen is denoted t.

Models of Molecular Evolution



Replace cards with nucleotides and you have Jukes Cantor (JC) model.

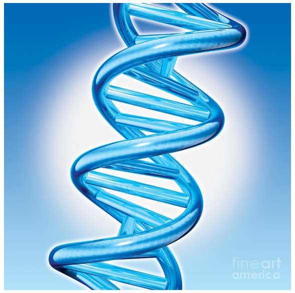
Good starting point but all of its core assumptions are violated by real DNA data



		To:			
		A	C	G	T
From:	A	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	C	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	G	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$
	T	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 - 1/4e^{-4/3\mu t}$	$1/4 + 3/4e^{-4/3\mu t}$

FIGURE 8.5 Substitution probability matrix under the JC model of DNA sequence evolution. The mutation rate, in substitutions per unit time, is denoted μ. The time interval over which evolution is allowed to happen is denoted t.

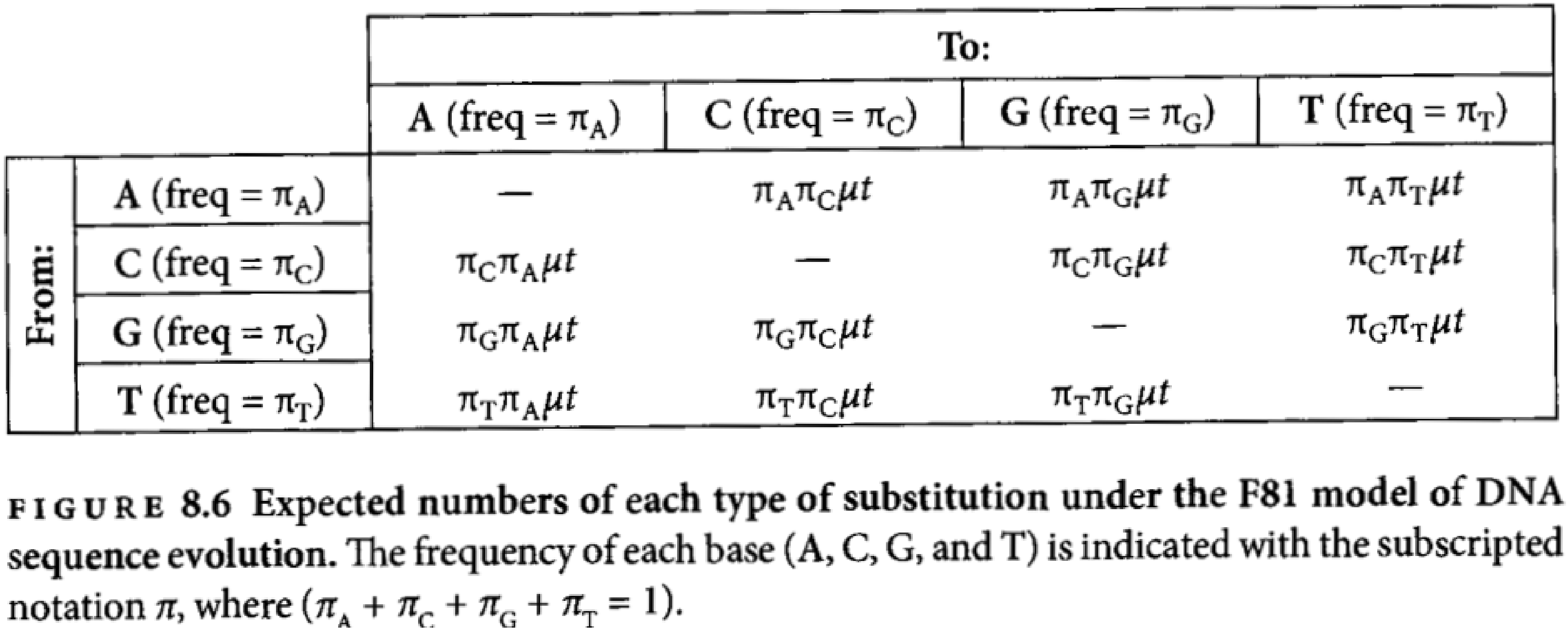
Models of Molecular Evolution



F81 (Felsenstein 1981) model:

Allows bases to occur at different frequencies in the "deck"

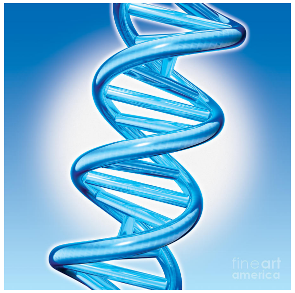
Eg, rRNA may be biased towards G and C



		To:			
		A (freq = π_A)	C (freq = π_C)	G (freq = π_G)	T (freq = π_T)
From:	A (freq = π_A)	—	$\pi_A\pi_C\mu t$	$\pi_A\pi_G\mu t$	$\pi_A\pi_T\mu t$
	C (freq = π_C)	$\pi_C\pi_A\mu t$	—	$\pi_C\pi_G\mu t$	$\pi_C\pi_T\mu t$
	G (freq = π_G)	$\pi_G\pi_A\mu t$	$\pi_G\pi_C\mu t$	—	$\pi_G\pi_T\mu t$
	T (freq = π_T)	$\pi_T\pi_A\mu t$	$\pi_T\pi_C\mu t$	$\pi_T\pi_G\mu t$	—

FIGURE 8.6 Expected numbers of each type of substitution under the F81 model of DNA sequence evolution. The frequency of each base (A, C, G, and T) is indicated with the subscripted notation π, where ($\pi_A + \pi_C + \pi_G + \pi_T = 1$).

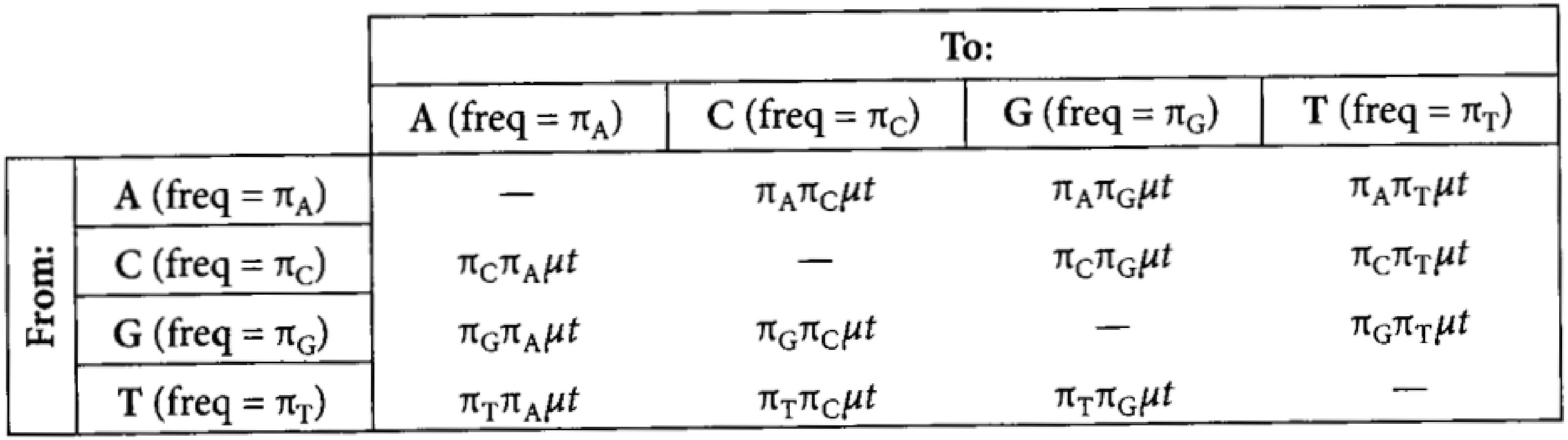
Models of Molecular Evolution



F81 model:

Allows bases to occur at different frequencies in the “deck”

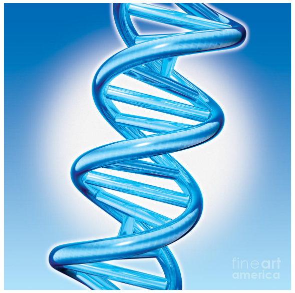
Eg, rRNA may be biased towards G and C



		To:			
		A (freq = π_A)	C (freq = π_C)	G (freq = π_G)	T (freq = π_T)
From:	A (freq = π_A)	—	$\pi_A \pi_C \mu t$	$\pi_A \pi_G \mu t$	$\pi_A \pi_T \mu t$
	C (freq = π_C)	$\pi_C \pi_A \mu t$	—	$\pi_C \pi_G \mu t$	$\pi_C \pi_T \mu t$
	G (freq = π_G)	$\pi_G \pi_A \mu t$	$\pi_G \pi_C \mu t$	—	$\pi_G \pi_T \mu t$
	T (freq = π_T)	$\pi_T \pi_A \mu t$	$\pi_T \pi_C \mu t$	$\pi_T \pi_G \mu t$	—

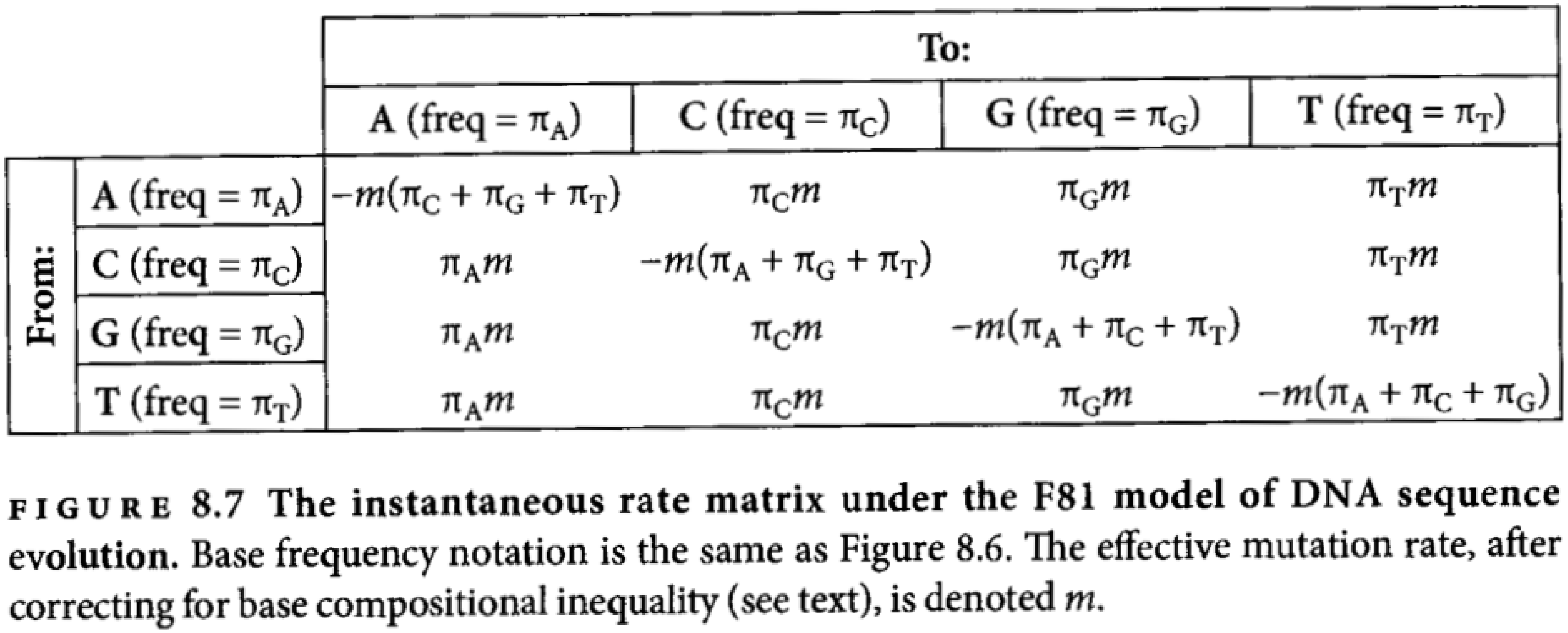
FIGURE 8.6 **Expected numbers of each type of substitution under the F81 model of DNA sequence evolution.** The frequency of each base (A, C, G, and T) is indicated with the subscripted notation π, where ($\pi_A + \pi_C + \pi_G + \pi_T = 1$).

Models of Molecular Evolution



F81 model:

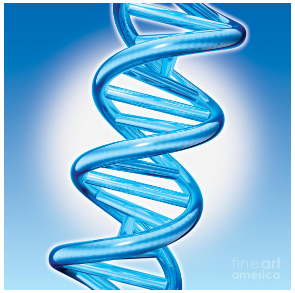
m is a modified version of μt that takes the effect of unequal base frequencies into account. Rare bases will persist for less time than common bases that will often be substituted by themselves.



		To:			
		A (freq = π_A)	C (freq = π_C)	G (freq = π_G)	T (freq = π_T)
From:	A (freq = π_A)	$-m(\pi_C + \pi_G + \pi_T)$	$\pi_C m$	$\pi_G m$	$\pi_T m$
	C (freq = π_C)	$\pi_A m$	$-m(\pi_A + \pi_G + \pi_T)$	$\pi_G m$	$\pi_T m$
	G (freq = π_G)	$\pi_A m$	$\pi_C m$	$-m(\pi_A + \pi_C + \pi_T)$	$\pi_T m$
	T (freq = π_T)	$\pi_A m$	$\pi_C m$	$\pi_G m$	$-m(\pi_A + \pi_C + \pi_G)$

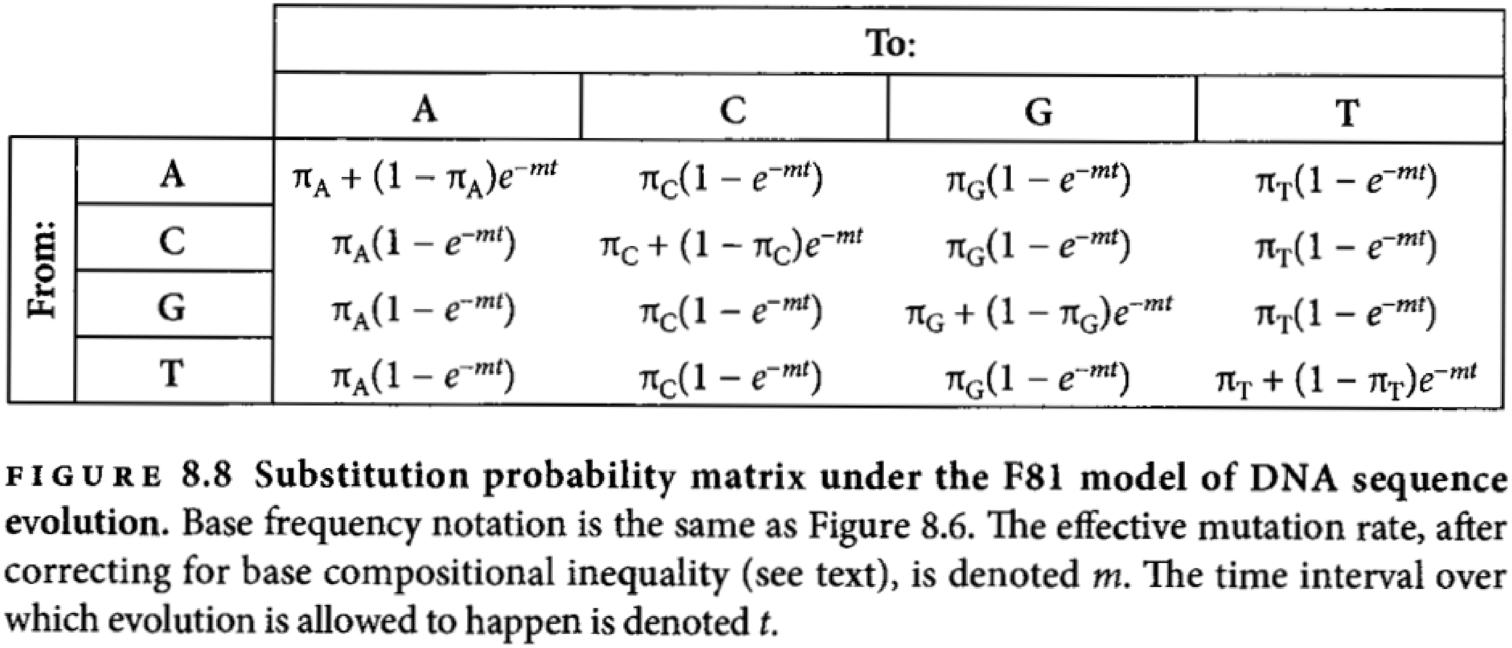
FIGURE 8.7 The instantaneous rate matrix under the F81 model of DNA sequence evolution. Base frequency notation is the same as Figure 8.6. The effective mutation rate, after correcting for base compositional inequality (see text), is denoted *m*.

Models of Molecular Evolution



F81 model:

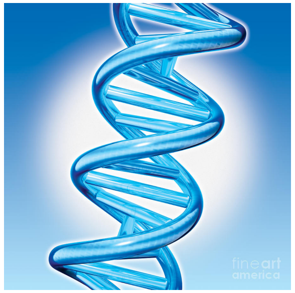
m is a modified version of µt that takes the effect of unequal base frequencies into account. Rare bases will persist for less time than common bases that will often be substituted by themselves.



		To:			
		A	C	G	T
From:	A	$\pi_A + (1 - \pi_A)e^{-mt}$	$\pi_C(1 - e^{-mt})$	$\pi_G(1 - e^{-mt})$	$\pi_T(1 - e^{-mt})$
	C	$\pi_A(1 - e^{-mt})$	$\pi_C + (1 - \pi_C)e^{-mt}$	$\pi_G(1 - e^{-mt})$	$\pi_T(1 - e^{-mt})$
	G	$\pi_A(1 - e^{-mt})$	$\pi_C(1 - e^{-mt})$	$\pi_G + (1 - \pi_G)e^{-mt}$	$\pi_T(1 - e^{-mt})$
	T	$\pi_A(1 - e^{-mt})$	$\pi_C(1 - e^{-mt})$	$\pi_G(1 - e^{-mt})$	$\pi_T + (1 - \pi_T)e^{-mt}$

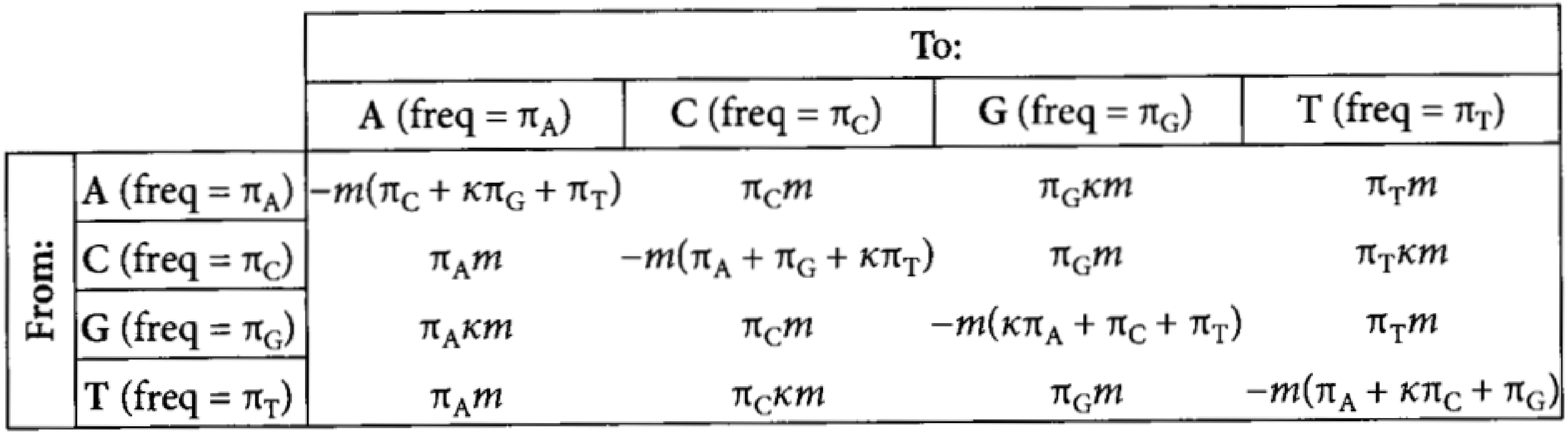
FIGURE 8.8 **Substitution probability matrix under the F81 model of DNA sequence evolution.** Base frequency notation is the same as Figure 8.6. The effective mutation rate, after correcting for base compositional inequality (see text), is denoted *m*. The time interval over which evolution is allowed to happen is denoted *t*.

Models of Molecular Evolution



HKY (Hasegawa, Kishino, Yano 1985) model:

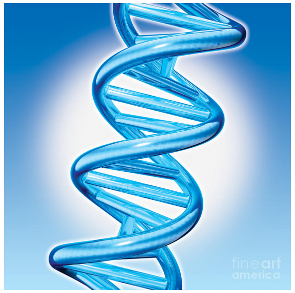
Transitions and transversions occur at different rates...



		To:			
		A (freq = π_A)	C (freq = π_C)	G (freq = π_G)	T (freq = π_T)
From:	A (freq = π_A)	$-m(\pi_C + \kappa\pi_G + \pi_T)$	$\pi_C m$	$\pi_G \kappa m$	$\pi_T m$
	C (freq = π_C)	$\pi_A m$	$-m(\pi_A + \pi_G + \kappa\pi_T)$	$\pi_G m$	$\pi_T \kappa m$
	G (freq = π_G)	$\pi_A \kappa m$	$\pi_C m$	$-m(\kappa\pi_A + \pi_C + \pi_T)$	$\pi_T m$
	T (freq = π_T)	$\pi_A m$	$\pi_C \kappa m$	$\pi_G m$	$-m(\pi_A + \kappa\pi_C + \pi_G)$

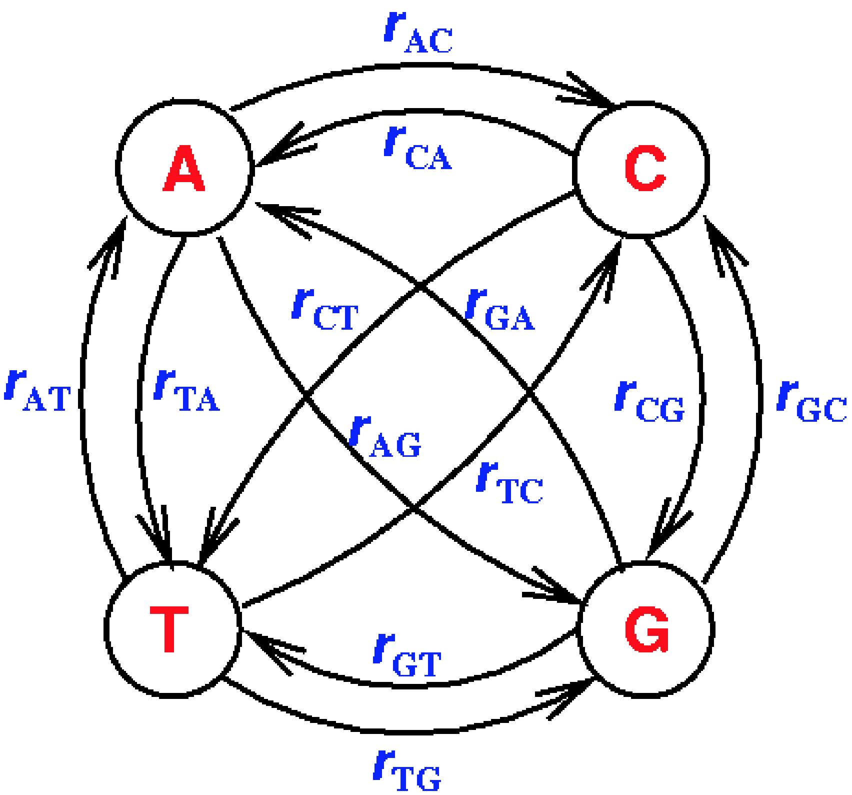
FIGURE 8.9 The instantaneous rate matrix under the HKY model of DNA sequence evolution. Notation is the same as Figure 8.7 except for the addition of a rate multiplier, κ, which indicates how many times faster transitions occur than transversions.

Models of Molecular Evolution

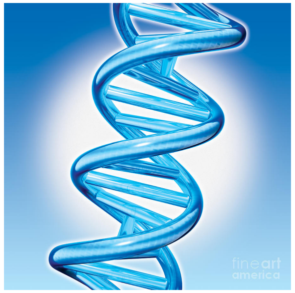


GTR (general, time reversible) model:

The different types of transitions and transversions can have different rates (ie, A → T may differ from T → A)

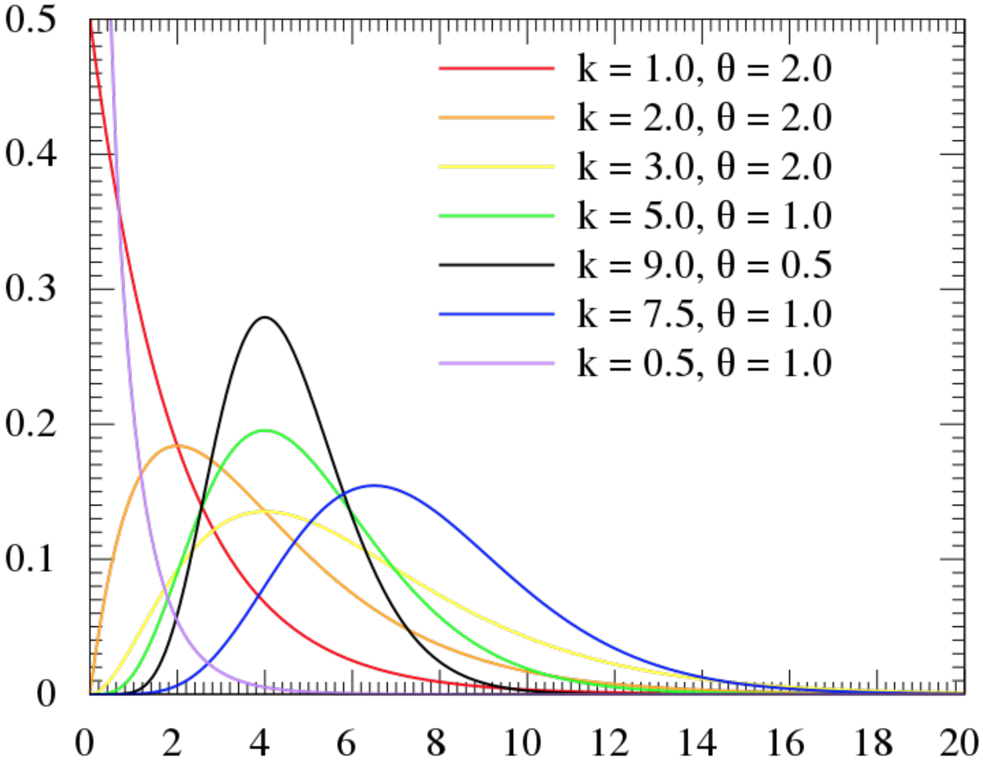


Models of Molecular Evolution

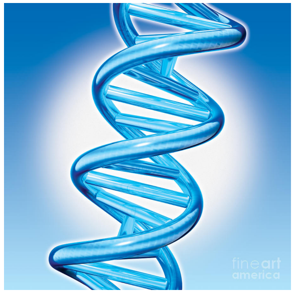


Rate Heterogeneity:

Not all sites in a DNA molecule evolve at the same rate. We could partition the data using Auxiliary info on rates or draw rates from a distribution.



Models of Molecular Evolution



Models of molecular evolution are nested...

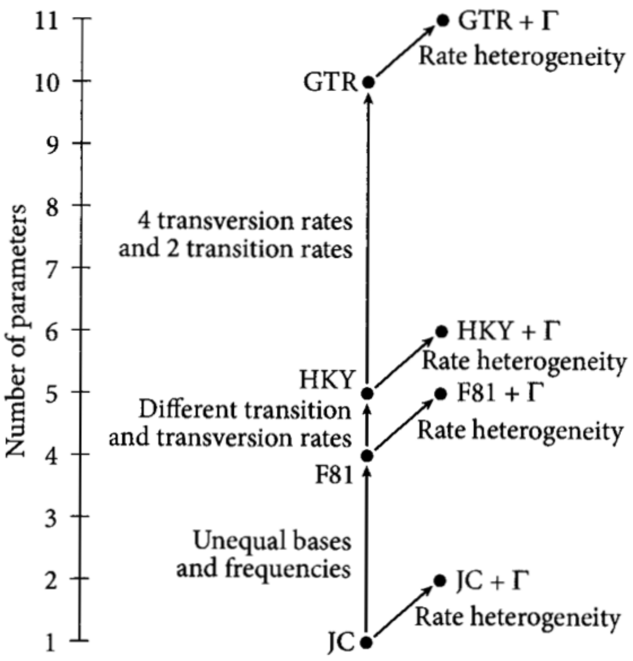


FIGURE 8.10 Depiction of the relationship between some commonly used models of evolution. Any two models that are connected by arrows that proceed in the same direction are nested: the simpler model (closer to the bottom of the chart) contains a subset of the free parameters of the more complex model. The axis on the left shows the number of free rate parameters in the model. The figure assumes that site-to-site rate heterogeneity is modeled using a discrete approximation to a gamma (Γ) distribution, which adds one free parameter to the model.

Distance Trees

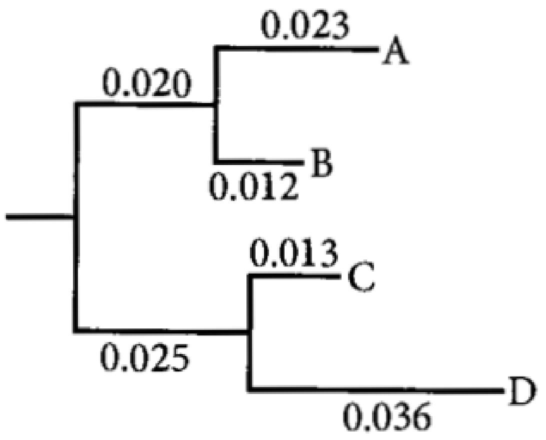


FIGURE 8.11 Phylogram showing evolutionary distances (in substitutions/site).

TABLE 8.2 Evolutionary distances between the taxa in Figure 8.11

	A	B	C	D
A	0			
B	0.035	0		
C	0.081	0.070	0	
D	0.104	0.093	0.049	0

Distance Trees

Pairwise distances can be translated into evolutionary distances by using the expected relationship between the two under a model of sequence evolution.

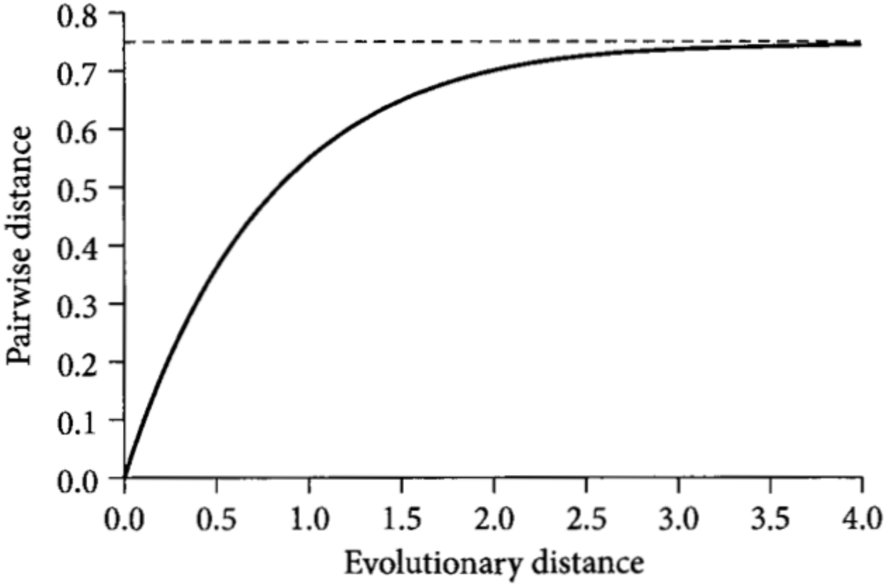


FIGURE 8.12 Relationship of evolutionary and pairwise distances under the JC model. The dashed line indicates the maximum expected pairwise distance, 0.75.

Distance Trees

Minimum evolution, for example, adjusts branch lengths to minimize the sum of squared deviations between observed and expected distances.

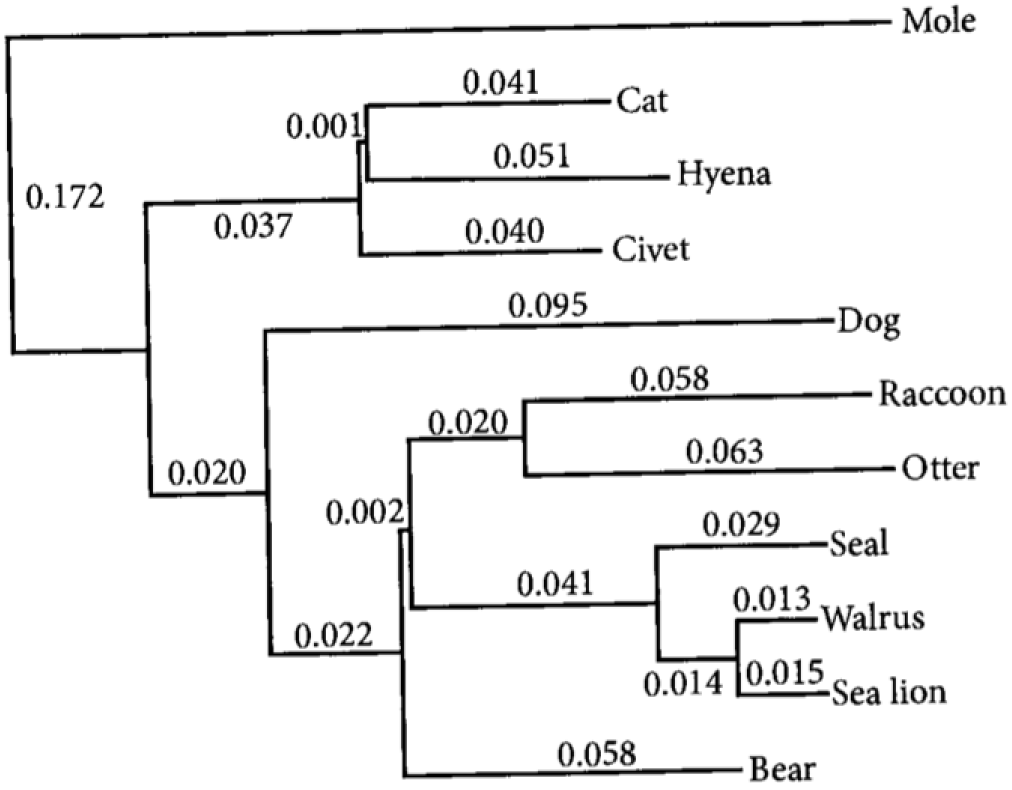


FIGURE 8.14 Minimum evolution tree based on the JC distances obtained from the carnivoran molecular data. Branch lengths are given in average number of character state changes per character.

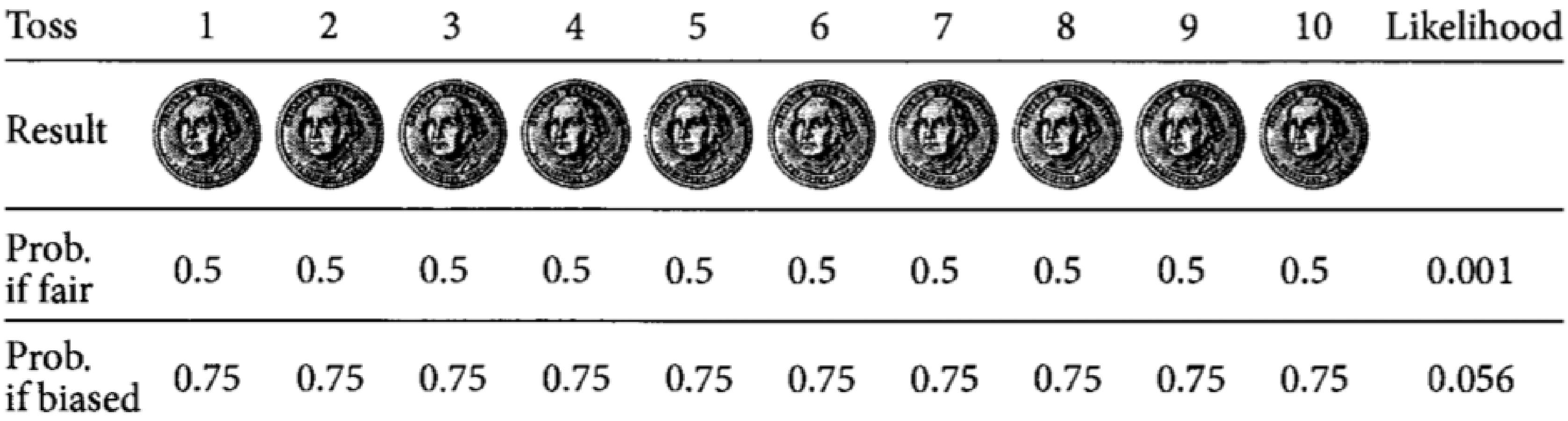
Maximum Likelihood

Find the tree that has the highest probability of giving rise to the observed data

Maximum Likelihood

A coin toss example...

How probable are the observed data under the hypothesis that the coin is biased?



Toss	1	2	3	4	5	6	7	8	9	10	Likelihood
Result											
Prob. if fair	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.001
Prob. if biased	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.056

FIGURE 8.15 Likelihood of ten sequential heads for a fair or a biased coin. The likelihood is the product of the probabilities of each individual toss, e.g. 0.5^{10}, under the fair coin hypothesis.

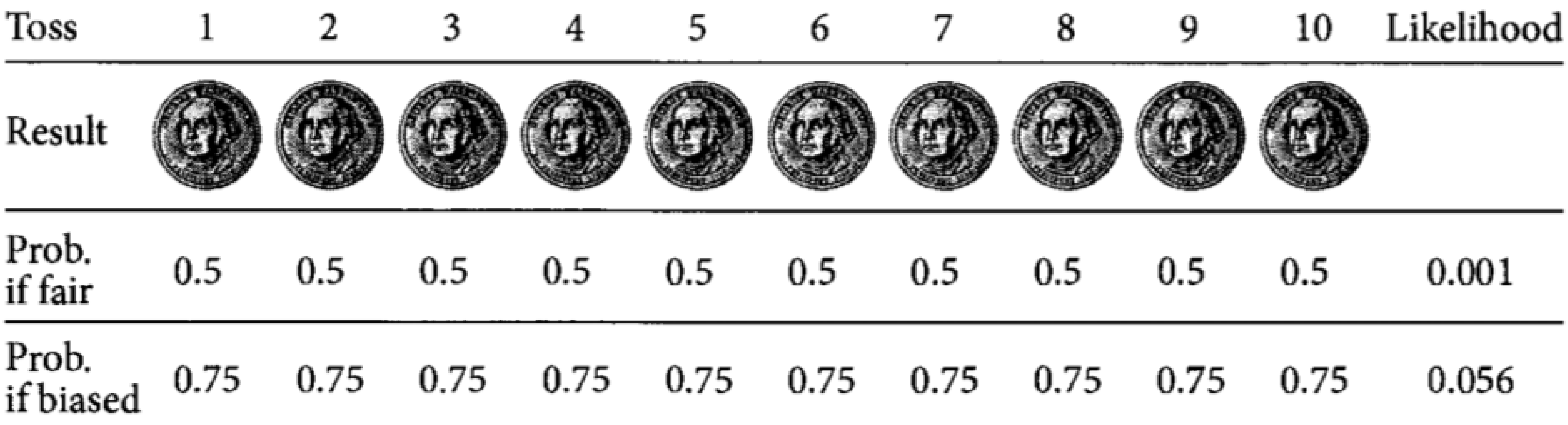
Maximum Likelihood

A coin toss example...

How probable are the observed data under the hypothesis that the coin is biased?

Not very probably but much more likely that the alternate hypothesis:

56 times more likely 0.056/0.001 (likelihood ratio). Log-likelihood ratio is
Ln(56) = 4.02. (log-likelihood ratio of 2 ~ equals P < 0.05)



Toss	1	2	3	4	5	6	7	8	9	10	Likelihood
Result											
Prob. if fair	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.001
Prob. if biased	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.056

FIGURE 8.15 Likelihood of ten sequential heads for a fair or a biased coin. The likelihood is the product of the probabilities of each individual toss, e.g. 0.5^{10}, under the fair coin hypothesis.

Maximum Likelihood

Only unknown is this tree is branch length.

How long is it under JC model?

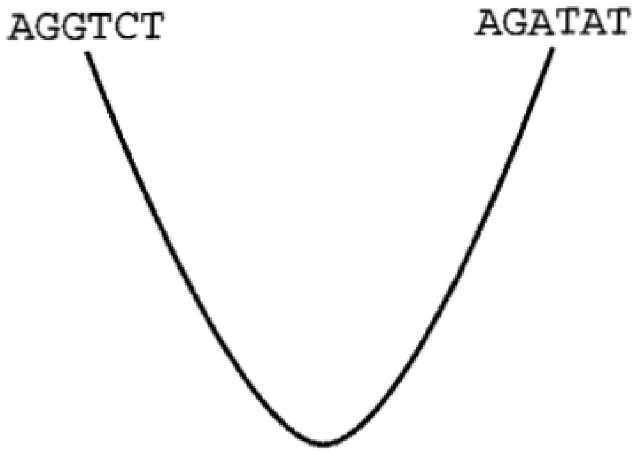


FIGURE 8.16 A two-taxon tree with six base-pair DNA sequences marked at the tips. The only element of uncertainty in this tree is the length of the single branch.

Maximum Likelihood

Only unknown is this tree is branch length.

How long is it under JC model?

Cannot be infinitely long since 4 matching bases (¾ of bases should be mismatches after infinite time)

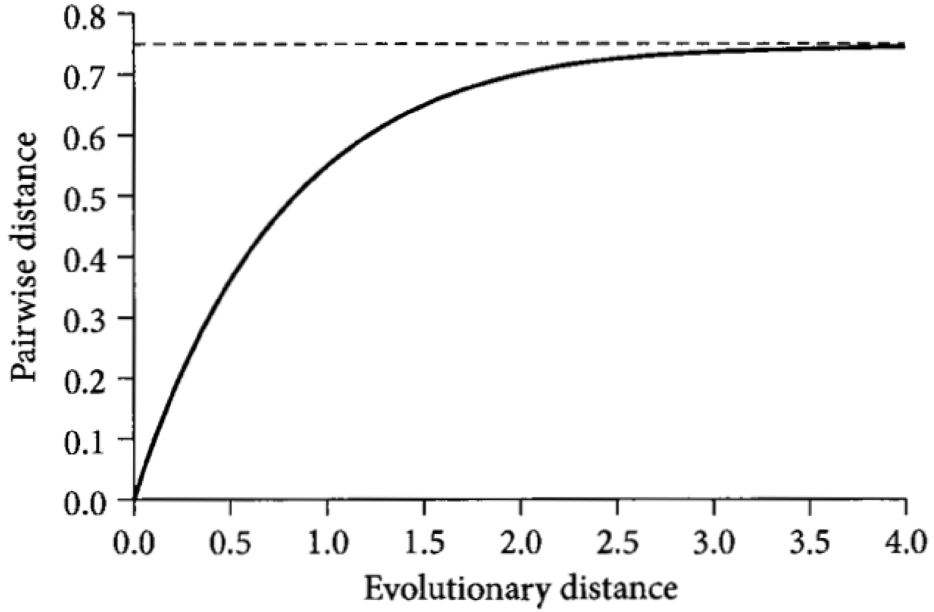


FIGURE 8.12 Relationship of evolutionary and pairwise distances under the JC model. The dashed line indicates the maximum expected pairwise distance, 0.75.

Maximum Likelihood

Only unknown is this tree is branch length.

How long is it under JC model?

Since two bases are mismatches is also cannot be zero length

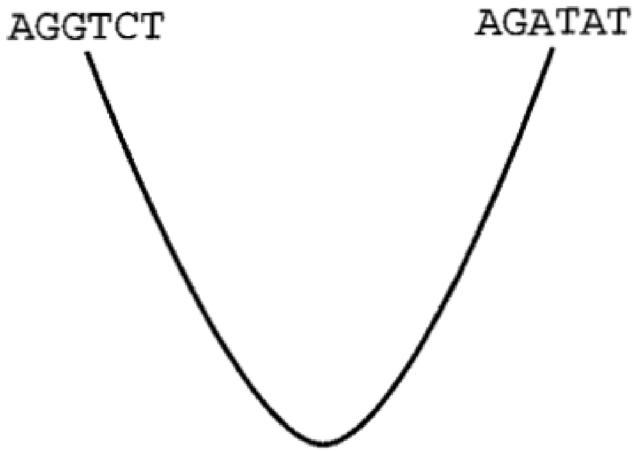


FIGURE 8.16 **A two-taxon tree with six base-pair DNA sequences marked at the tips.** The only element of uncertainty in this tree is the length of the single branch.

Maximum Likelihood

The rate of evolution (μ) and time (t) will determine the branches length. μ and t are difficult to disentangle but all we need to worry about is their product μt.

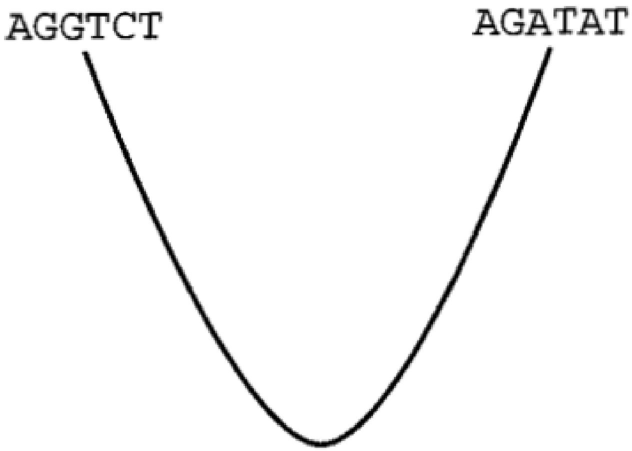


FIGURE 8.16 A two-taxon tree with six base-pair DNA sequences marked at the tips. The only element of uncertainty in this tree is the length of the single branch.

Maximum Likelihood

The rate of evolution (μ) and time (t) will determine the branches length. μ and t are difficult to disentangle but all we need to worry about is their product μt.

The probability of the first position in one taxon to be A is ¼ (equal base frequencies) Given that, the probability of observing a in both taxa is:
$\frac{1}{4} (\frac{1}{4} + \frac{3}{4} e^{-4/3\mu t})$

We can substitute any value for μt into the equation to obtain the probability that this site evolved under that branch length. This is the **site likelihood**.

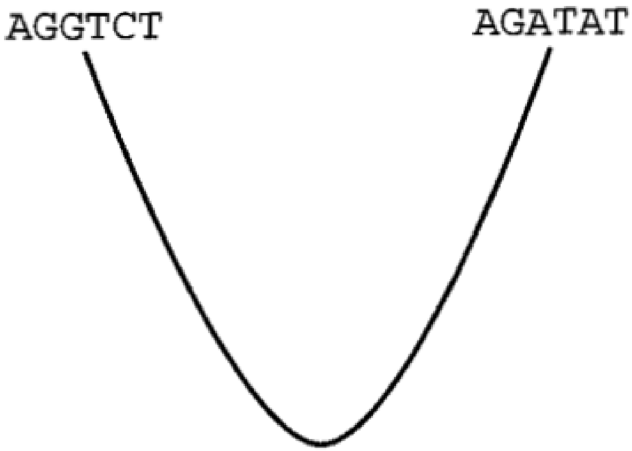


FIGURE 8.16 **A two-taxon tree with six base-pair DNA sequences marked at the tips.** The only element of uncertainty in this tree is the length of the single branch.

Maximum Likelihood

The probability (site likelihood) of a mismatch is
$\frac{1}{4} (\frac{1}{4} - \frac{1}{4} e^{-4/3\mu t})$

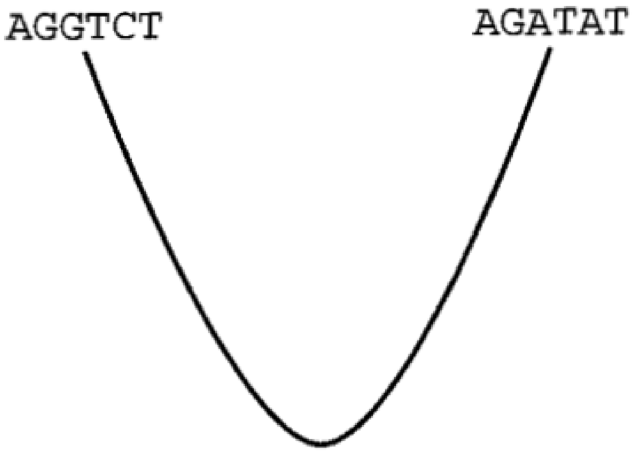


FIGURE 8.16 A two-taxon tree with six base-pair DNA sequences marked at the tips. The only element of uncertainty in this tree is the length of the single branch.

Maximum Likelihood

The likelihood of a tree knowing the entire characters matrix is the product of all site likelihoods.

$[\frac{1}{4} (\frac{1}{4} + \frac{3}{4} e^{-4/3\mu t})]^2 * [\frac{1}{4} (\frac{1}{4} + \frac{3}{4} e^{-4/3\mu t})]^4$

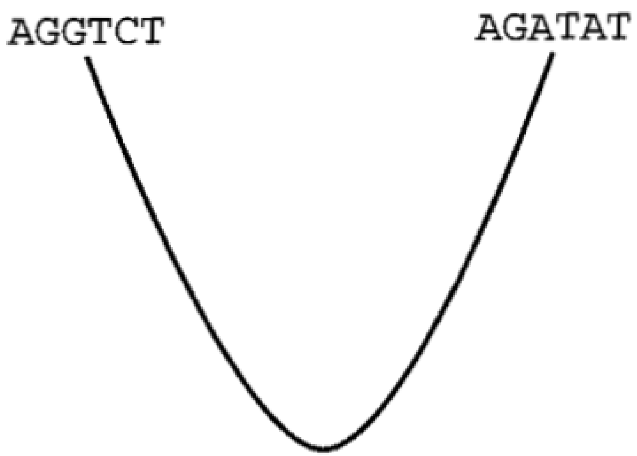


FIGURE 8.16 A two-taxon tree with six base-pair DNA sequences marked at the tips. The only element of uncertainty in this tree is the length of the single branch.

Maximum Likelihood

The likelihood of a tree knowing the entire characters matrix is the product of all site likelihoods.

Branch length of 0.44 has highest likelihood (0.595 x 10^{-6})

Log likelihood is -14.33 (avoids computer issues with small numbers. Highest likelihood corresponds to least negative log likelihood.

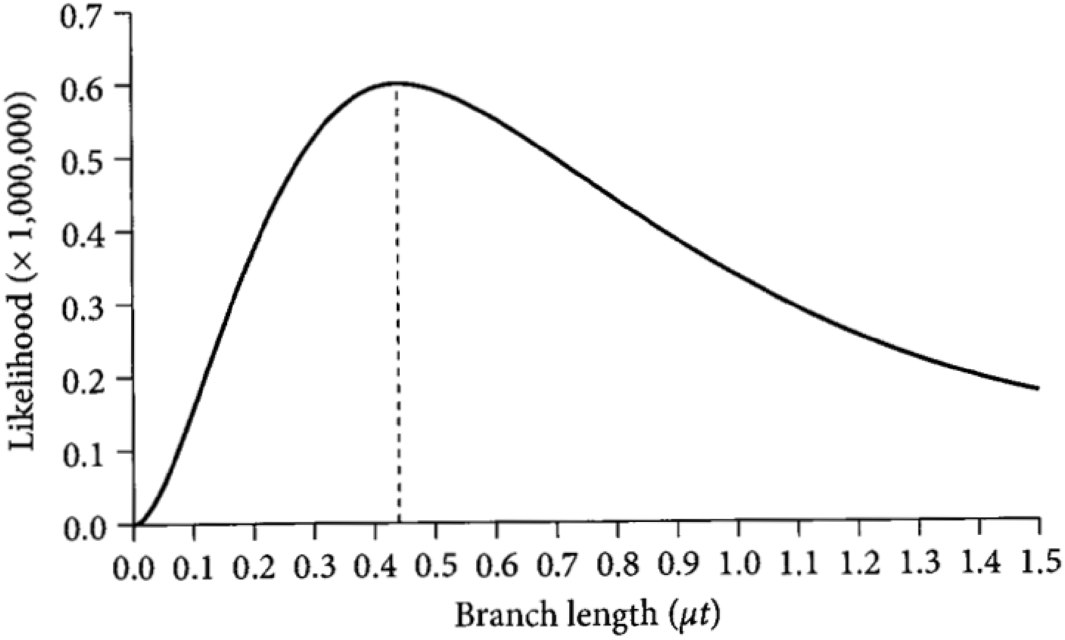


FIGURE 8.17 Likelihood values for different branch lengths given the data shown in Figure 8.16.

Maximum Likelihood

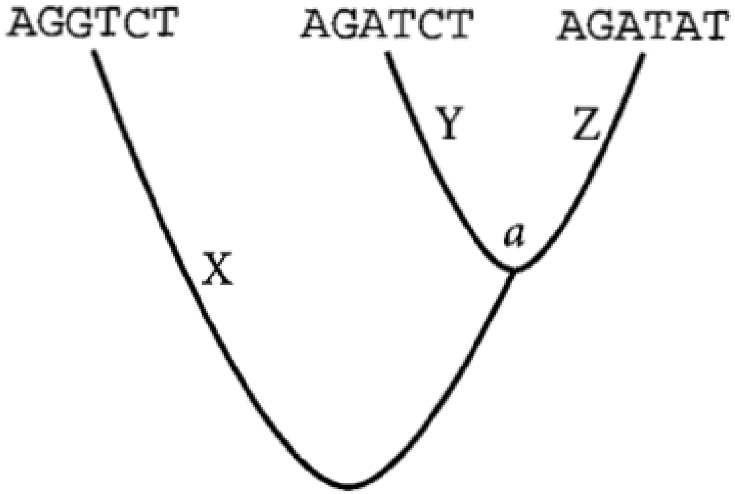


FIGURE 8.18 **Three-taxon tree with a six base-pair sequence at each tip.** The only items of uncertainty are the three branch lengths (X, Y, and Z) and the sequence at internal node *a*. We use the maximum likelihood criterion to estimate the value of the branch lengths, while summing over all possible sequences at node *a*.

Testing Model Fit

Pick a more complex model only when the gain in likelihood is higher than expected if the simpler model were true.

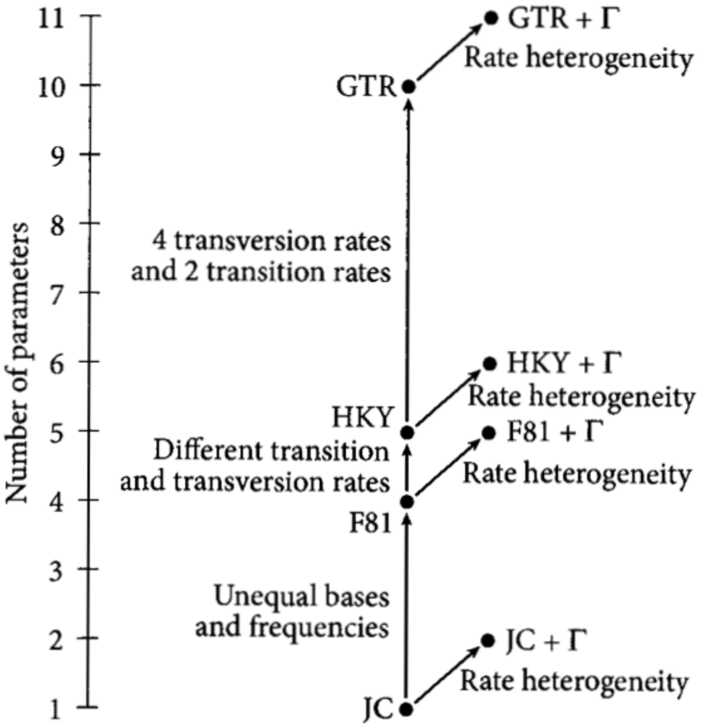


FIGURE 8.10 **Depiction of the relationship between some commonly used models of evolution.** Any two models that are connected by arrows that proceed in the same direction are nested: the simpler model (closer to the bottom of the chart) contains a subset of the free parameters of the more complex model. The axis on the left shows the number of free rate parameters in the model. The figure assumes that site-to-site rate heterogeneity is modeled using a discrete approximation to a gamma (Γ) distribution, which adds one free parameter to the model.