Modeling Molecular evolution

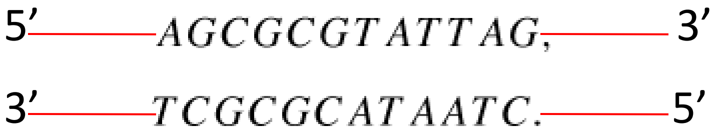
Phylogenetic distances

Preview

- Mathematical of mutation in DNA is covered in the next few lectures.
- Mutations that occur in the DNA sequences are random.
- To capture the random mutation knowledge of basic probability is required.
- Application of Probabilistic models leads naturally to linear models.
- Sequence similarity in terms of phylogenetic distances will emerge from these linear models.
- So evolutionary relationship among different DNA sequences will be covered in these lectures.

Background on DNA

- The DNA molecule forms a double helix, a twisted ladder-like structure.
- DNA is made of nucleotides or bases -- adenine, guanine, cytosine, and thymine -- are denoted by the letters A, G, C , and T .
- Because of chemical similarity, adenine and guanine are called purines, while cytosine and thymine are called pyrimidines.
- A is paired with T and G is paired with C.

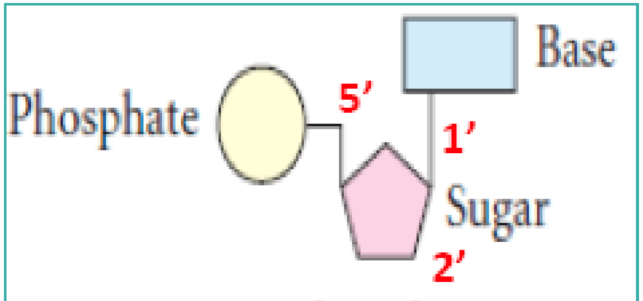


5’ AGCGCGTATTAG, 3’

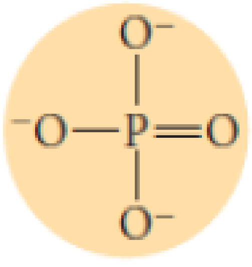
3’ TCGCGCATAATC. 5’

Complementary strands

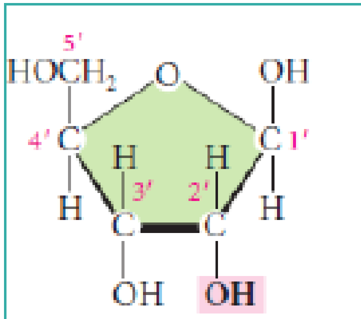
Structure of nucleotide, nucleosides , ribose and DNA



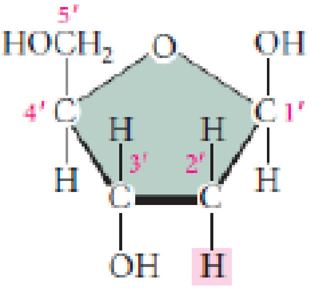
Nucleotide



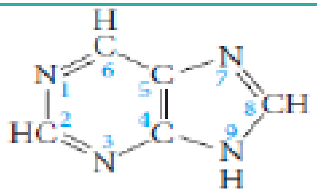
Phosphate



Ribose

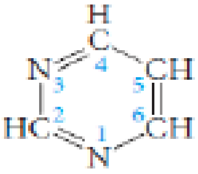


Deoxyribose



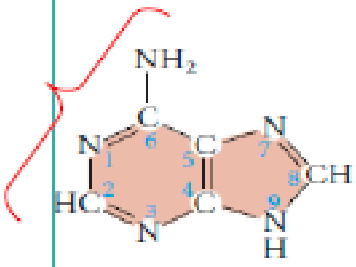
(6,5 rings)

Purine
(basic structure)



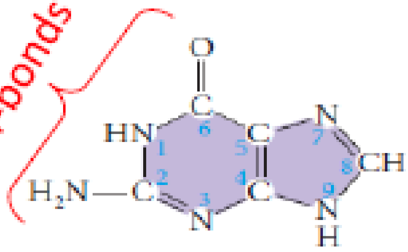
(6 membered ring)

Pyrimidine
(basic structure)

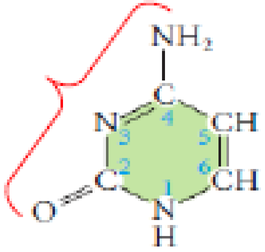


Adenine (A)

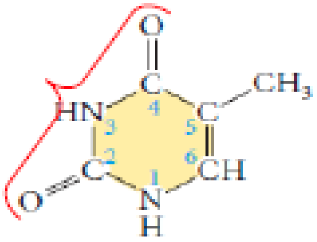
H-bonds



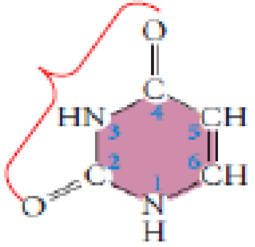
Guanine (G)



Cytosine (C)

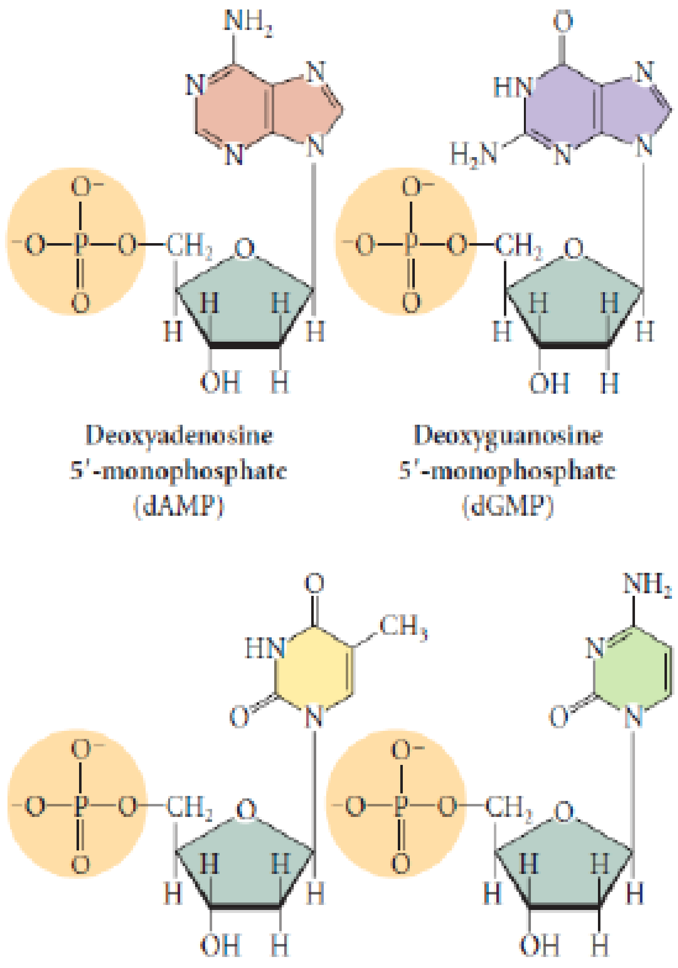


Thymine (T)
(present in DNA)



Uracil (U)
(present in RNA)

Four types of DNA nucleotides



Deoxyadenosine 5′-monophosphate (dAMP)

Deoxyguanosine 5′-monophosphate (dGMP)

Deoxythymidine 5′-monophosphate (dTMP)

Deoxycytidine 5′-monophosphate (dCMP)

CONCEPT CHECK

Q: How do the sugars of RNA and DNA differ?

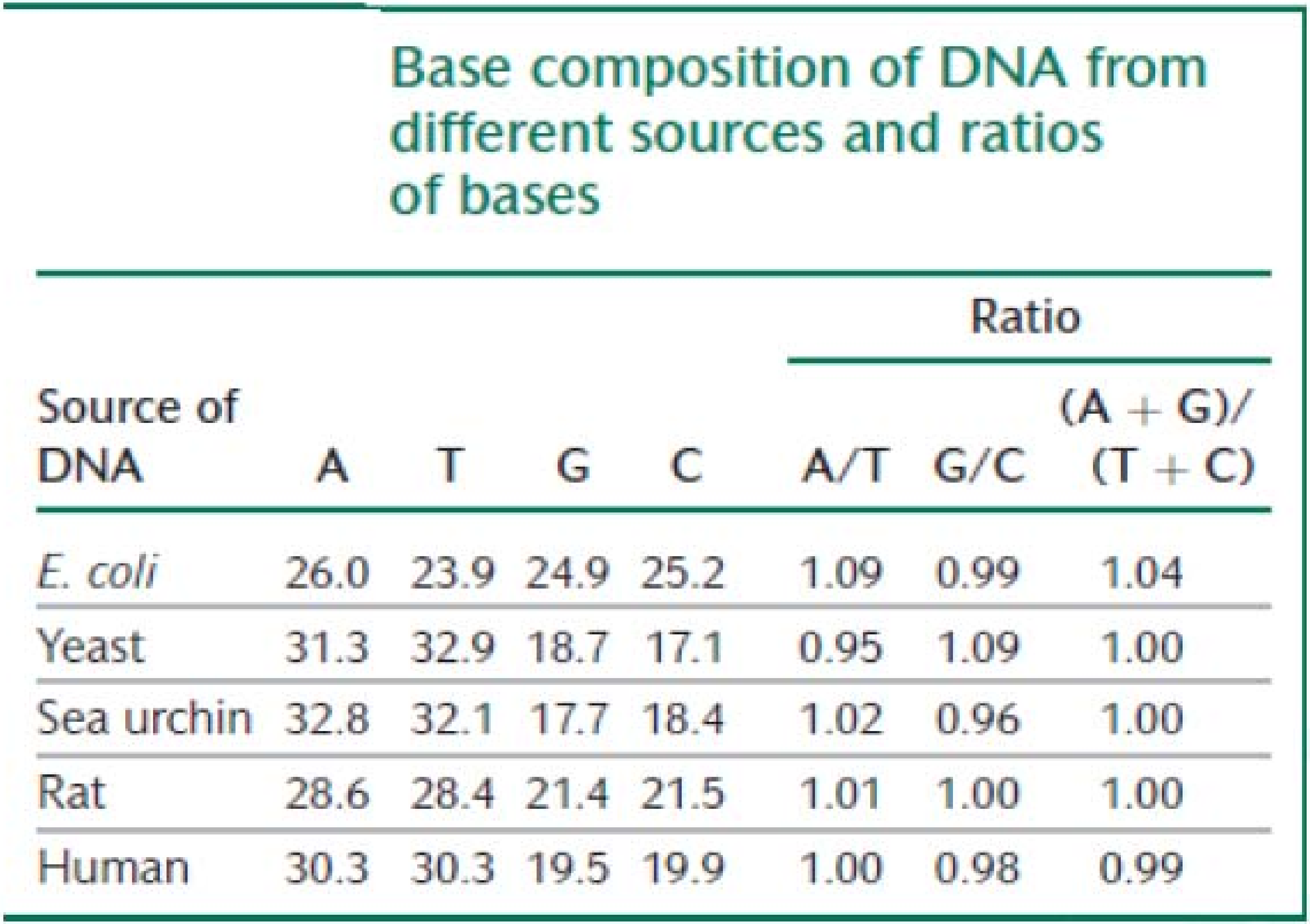
a. RNA has a six-carbon sugar; DNA has a five-carbon sugar.

b. The sugar of RNA has a hydroxyl group that is not found in the sugar of DNA.

c. RNA contains uracil; DNA contains thymine.

d. DNA's sugar has a phosphorus atom; RNA's sugar does not.

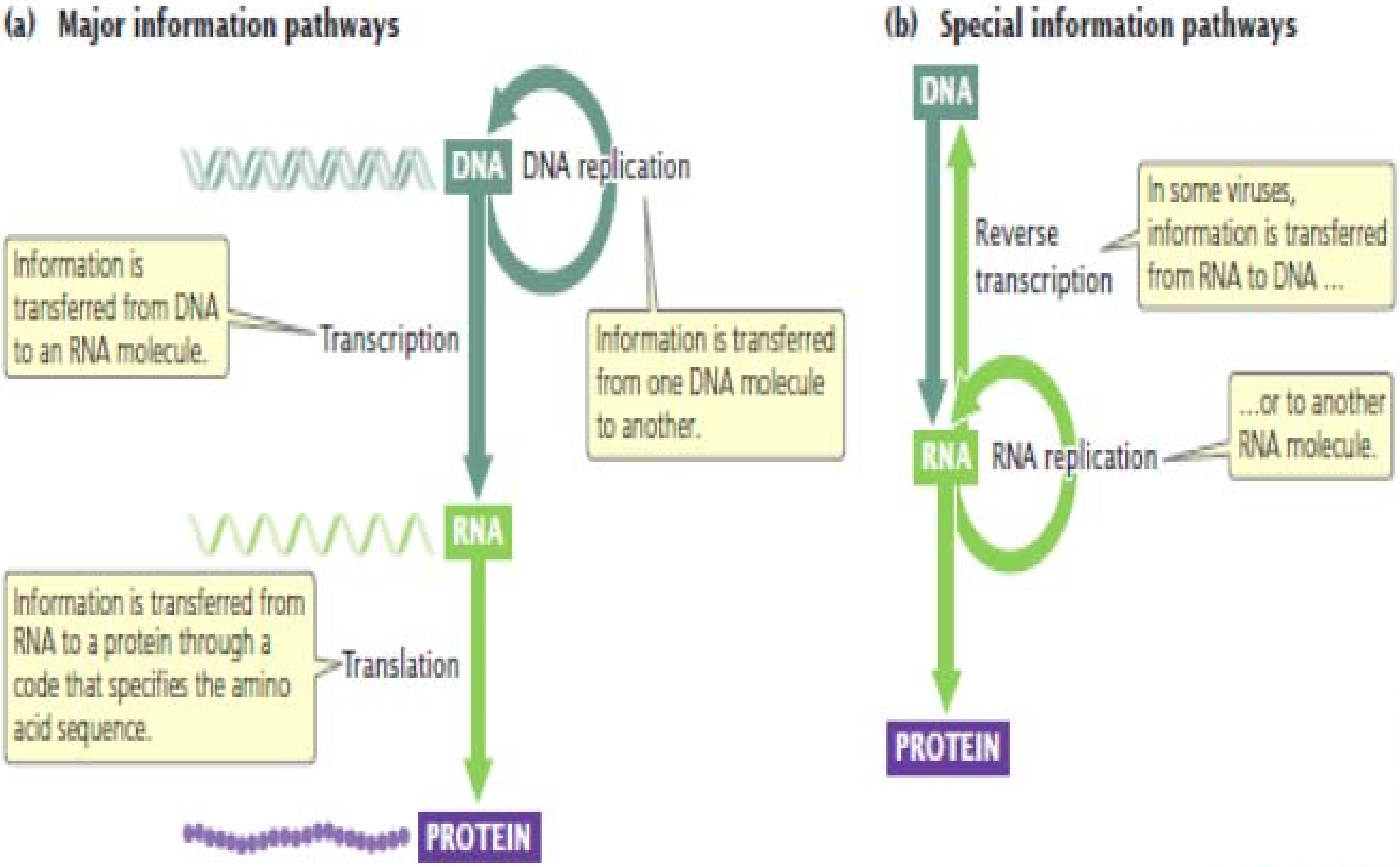
Chargaff’s rule- A/T = G/C



Base composition of DNA from different sources and ratios of bases

					Ratio		
Source of							(A + G)/
DNA	A	T	G	C	A/T	G/C	(T + C)
E. coli	26.0	23.9	24.9	25.2	1.09	0.99	1.04
Yeast	31.3	32.9	18.7	17.1	0.95	1.09	1.00
Sea urchin	32.8	32.1	17.7	18.4	1.02	0.96	1.00
Rat	28.6	28.4	21.4	21.5	1.01	1.00	1.00
Human	30.3	30.3	19.5	19.9	1.00	0.98	0.99

Pathways of information transfer


(a) Major information pathways
DNA
DNA replication
Information is transferred from DNA to an RNA molecule.
Transcription
Information is transferred from one DNA molecule to another.
RNA
Information is transferred from RNA to a protein through a code that specifies the amino acid sequence.
Translation
PROTEIN
(b) Special information pathways
DNA
Reverse transcription
In some viruses, information is transferred from RNA to DNA ...
RNA
RNA replication
...or to another RNA molecule.
PROTEIN

Genes, codons, and genetic code

- Some sections of DNA form genes that encode instructions for the manufacturing of proteins (though the production of the protein is accomplished through the intermediate production of messenger RNA).

- In these genes, triplets of consecutive bases form codons, with each codon specifying a particular amino acid to be placed in the protein chain according to the genetic code.

- For example, the codon TGC always means that the amino acid cysteine will occur at that location in the protein the end of the protein.

- Certain codons specify stop signal.

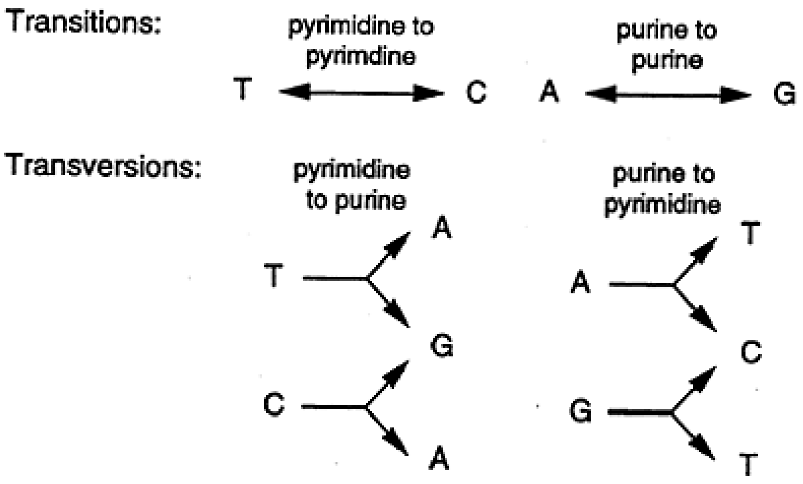
Mutations in DNA

- The most common mutation that is introduced in the copying of sequences of DNA is a **base substitution**, a replacement of one base for another at a certain site in the sequence.

- For instance, if the sequence AAT**C**GC in an ancestor becomes AAT**G**GC in a descendent, then a base substitution C → G has occurred at the fourth site.

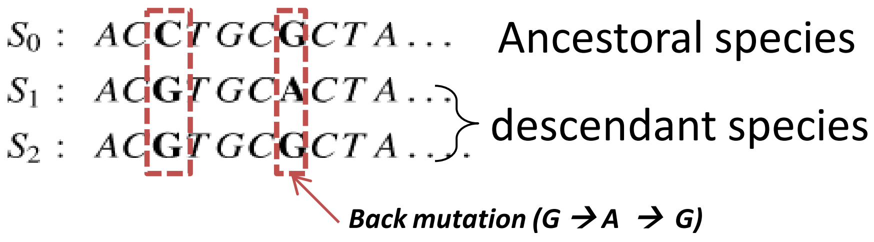
- A base substitution that replaces a purine with a purine, or a pyrimidine with a pyrimidine, is called a **transition**, whereas an interchange of these classes is called a **transversion.**

Summary



How to determine the amount of mutation that occurred over the time?

Consider below the three aligned DNA sequences of the same species out time. Overall there are 10 sites (nucleotides) are there.



S_0 : ACCTGCGCTA... Ancestoral species

S_1 : ACGTGCACTA...

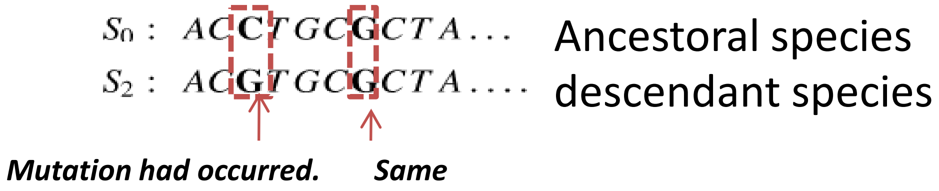
S_2 : ACGTGCGCTA...

descendant species

Back mutation (G → A → G)

3/10 of mutations occurred per site.

Now, if **we only saw the sequences for S_0 and S_2** only one base substitution among the first 10 sites, the one appearing in the third site.



S_0 : ACCTGCGCTA... Ancestoral species

S_2 : ACGTGCGCTA.... descendant species

Mutation had occurred. *Same*

1/10 of mutations occurred per site.

So we need a mathematical model **to reconstruct the number of mutations** that are likely to have occurred from those we see in **comparing only the initial and final DNA sequences**.

What to know in probability?

- **Example.** To apply this language to a DNA sequence, suppose a 40-base sequence reads as follows:

AGCTTCCGATCCGCTATAATCGTTAGTTGTTACACCTCTG

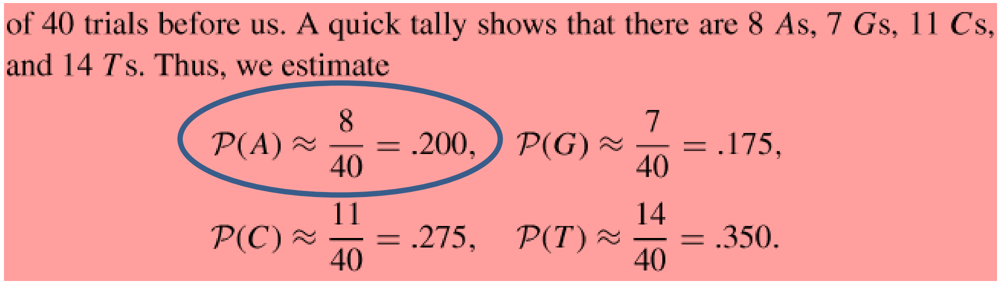
- What is the probability that the next base, in site 41, should be an A?

What to know in probability?

- **Example.** To apply this language to a DNA sequence, suppose a 40-base sequence reads as follows:

AGCTTCCGATCCGCTATAATCGTTAGTTGTTACACCTCTG

- What is the probability that the next base, in site 41, should be an A?



of 40 trials before us. A quick tally shows that there are 8 *A*s, 7 *G*s, 11 *C*s, and 14 *T*s. Thus, we estimate

$$\boxed{\mathcal{P}(A) \approx \frac{8}{40} = .200,} \quad \mathcal{P}(G) \approx \frac{7}{40} = .175,$$

$$\mathcal{P}(C) \approx \frac{11}{40} = .275, \quad \mathcal{P}(T) \approx \frac{14}{40} = .350.$$

Thus, we estimate the probability of an A in site 41 as .2.

What to know in probability?

- What is a trial and an Event?
- Addition rule
- Mutually exclusive events and sums of probabilities.
- Independent events and products of probabilities.
- Multiplication Rule
- Conditional Probabilities
- Definition of Independence

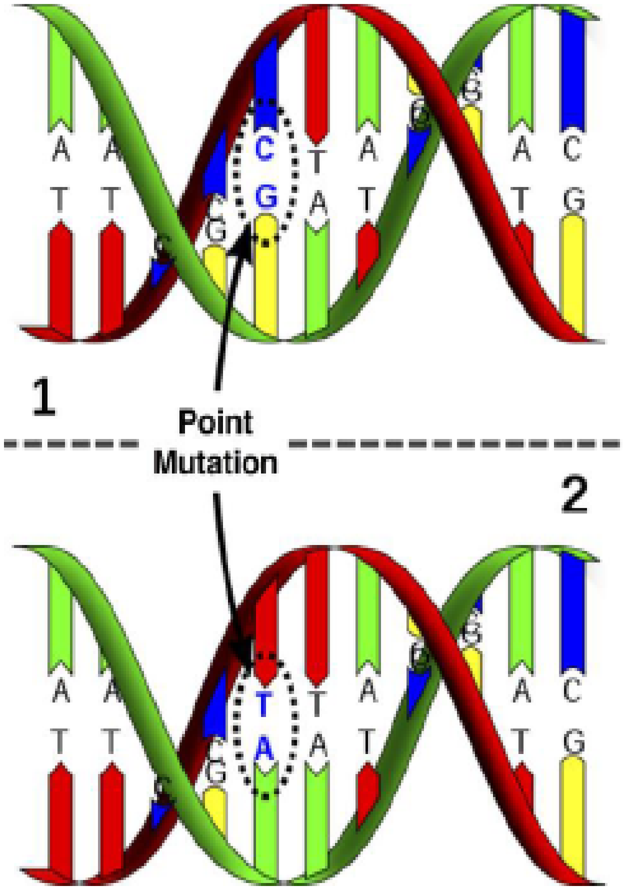
Variation in DNA sequences

- **Mutations** –mistakes in DNA replication.

- Mutations are rare.

  On average, one mistake per 200 million to 1 billion nucleotides

- Consequently –most mutations in DNA are *inherited from previous generations.*



http://rosalind.info/media/point_mutation.png

Types of mutations

- Mutations originate in single individuals.

- Mutations can become *fixed in a population.*
  –every individual has that mutation.

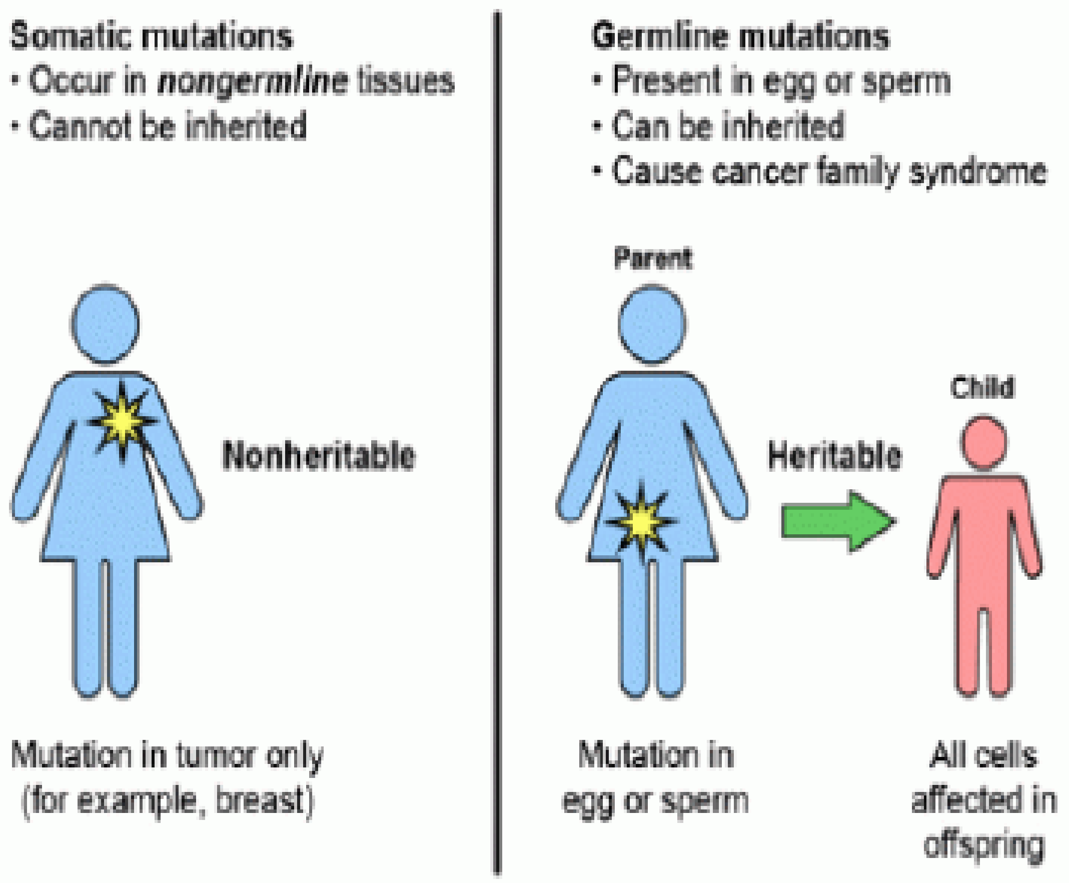
- Broadly categorized into three different ways:

- Neutral mutations: do not affect the organisms functions or ability to generate offspring

- Deleterious mutations: disrupt some functions
  –Under negative selection

- Advantageous mutations: enhance some function
  –Under positive selection

Germline mutations



Somatic mutations
- Occur in *nongermline* tissues
- Cannot be inherited

Germline mutations
- Present in egg or sperm
- Can be inherited
- Cause cancer family syndrome

Nonheritable
Mutation in tumor only (for example, breast)
Parent
Child
Heritable
Mutation in egg or sperm
All cells affected in offspring

https://autismsciencefoundation.wordpress.com/2015/12/02/brain-tissue-reveals-more-genetic-clues-to-autism/

Models of mutation

- Models are built to show how to determine ultimate probabilities of specific nucleotides appearing, given mutation rate(s) in a given site of the DNA sequences.

- Three approaches are used:

(I) Analytic approach: By calculus to understand the underlying theory

(II) Numeric Approach: By Markov Chains to process simple numerical examples.

(III) Linear Algebraic approach: By transition matrix eigen structure to handle complex models and complicated matrices.

Basic Molecular Biology

- A (purine) binds (2 H bonds) to T (pyrimidine)
- C (pyrimidine) binds (3 H bonds) to G (purine)

Possible mutations:

(i) Transitions

- Purine → Purine; i.e., A →G, G → A (transition)
- Pyrimidine → Pyrimidine; i.e., C →T , T→C (transition)

2 transitions are seen

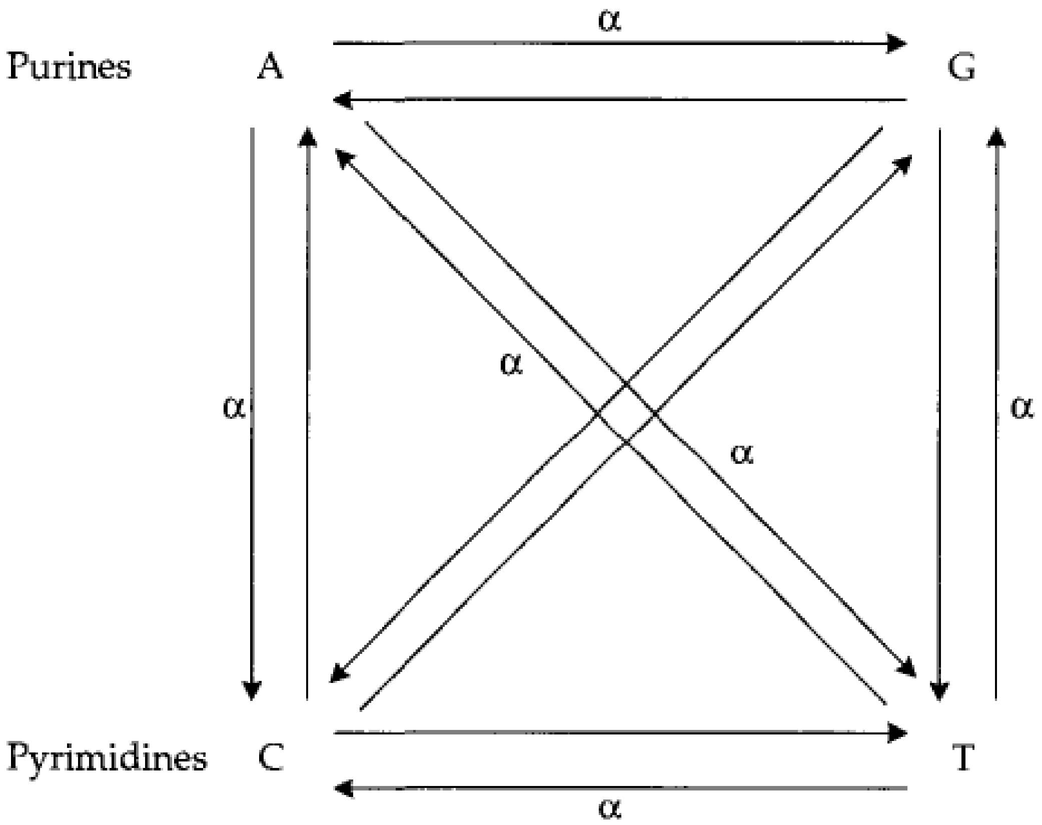
(ii) Transversions

- Purine → Pyrimidine; i.e., A→T, A→C, G→T, G→C (transversion)
- Pyrimidine → Purine; i.e., T→A, C→A, T→G, C→G (transversion)

4 are seen

Note: There are twice as many transversions as transitions possible.

In Graphical form



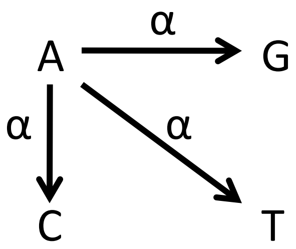
Transitions are more common

–In humans, transitions are at least 2 times more likely than transversions

Nucleotide Substitution, A Simple Model: Jukes and Cantor

- This simple model assumes that substitutions occur with equal probability among the four nucleotide types.

- For example, if the nucleotide under consideration is A, it will change to T, C, or G with equal probability.



- In this model, the **rate α** of substitution for each nucleotide is 3α per unit time

- Because the model involves a single parameter, α, it is called the one-parameter model.

Jukes-Cantor model: At time t = 0 and t =1

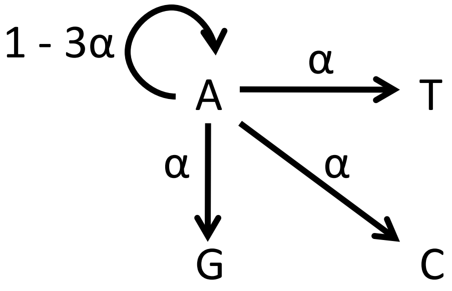
- Since we start with A, the probability that this **site is occupied** by A at time t=0 is $P_{A(0)} = 1$.

- At time t=1, the probability of still having A at this site is given by

$$P_{A(1)} = 1 - 3\alpha$$

3α = the probability of A changing to T, C, or G.

$1 - 3\alpha$ = the probability that A has remained unchanged.



Jukes-Cantor model: Explanation for time t =2

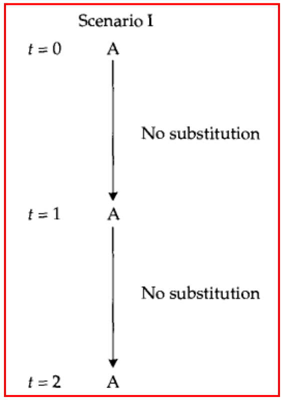
- **Two possible scenarios**:

(1) the nucleotide has remained unchanged from time 0 to time 2, and

(2) the nucleotide has changed to T, C, or G at time 1, but has subsequently reverted to A at time 2.

- In **scenario-1**
- The probability of the nucleotide being A at time t=1 is $P_{A(1)}$, and the probability that it has remained A at time t=2 is 1 - 3α.
- The **product of these two independent** variables gives us the probability

$$P_{A(2)} = (1-3\alpha)\ P_{A(1)}$$



Jukes-Cantor model: At time t =2

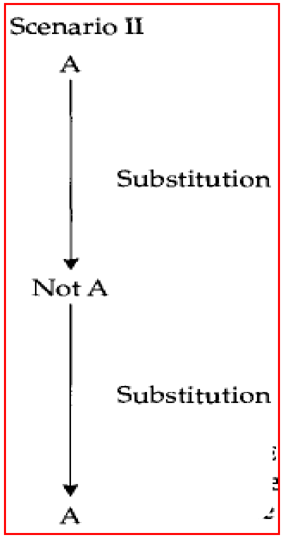
- The second scenario:

- The probability of the nucleotide not being A at time 1 is 1 - $P_{A(1)}$, and

- the probability of changing back to A at time 2 is α.

- The product of these two variables gives us the probability

$$P_{A(2)} = \alpha\ (1-P_{A(1)})$$

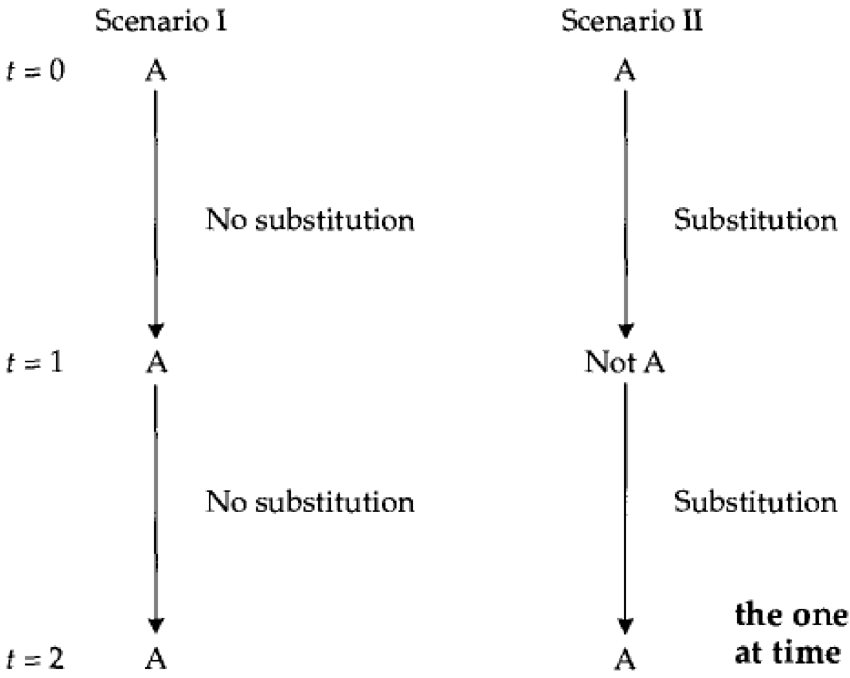


- The **overall probability** is

$$P_{A(2)}=(1-3\alpha)\ P_{A(1)} + \alpha\ [1-P_{A(1)}]$$

Summary Jukes-Cantor model: At time t =2

- To determine the probability of having A at time t=2, **two scenarios** are considered.



Two possible scenarios according to the one-parameter model for having A at a site at time $t = 2$, given that the site had A at time 0.

$$P_{A(2)} = \overbrace{(1\text{-}3\alpha)\ P_{A(1)}}^{\text{Scenario-1}} + \overbrace{\alpha\ [1\text{-}P_{A(1)}]}^{\text{Scenario-2}}$$

Jukes-Cantor model generalization: At time t =t

- Using the above formulation, we can show that the following recurrence equation

$$P_{\mathrm{A}(t+1)} = (1-3\alpha)P_{\mathrm{A}(t)} + \alpha\left[1 - P_{\mathrm{A}(t)}\right]$$

- We can rewrite the above equation in terms of the amount of change in PA(t) per unit time as

$$\Delta P_{\mathrm{A}(t)} = P_{\mathrm{A}(t+1)} - P_{\mathrm{A}(t)} = -3\alpha P_{\mathrm{A}(t)} + \alpha\left[1 - P_{\mathrm{A}(t)}\right] = -4\alpha P_{\mathrm{A}(t)} + \alpha$$

- This is the discrete equation. The **continuous equation** will be

$$\frac{dP_{\mathrm{A}(t)}}{dt} = -4\alpha P_{\mathrm{A}(t)} + \alpha$$

Solving the first order differential equation

$$\frac{dP_{A(t)}}{dt} = -4\alpha P_{A(t)} + \alpha$$

The solution of the equation is given as **(HW: check this !)**

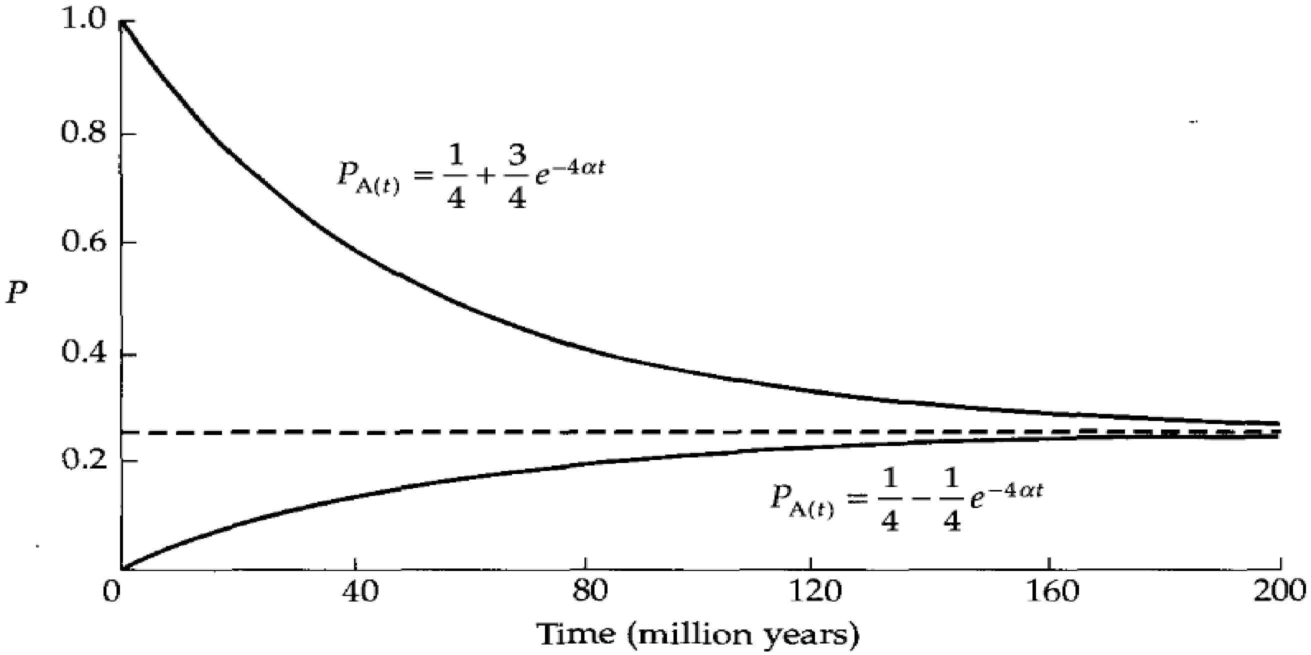
$$P_{A(t)} = \frac{1}{4} + \left(P_{A(0)} - \frac{1}{4}\right)e^{-4\alpha t}$$

Since we started with A, the probability that the site has A at time 0 is 1. Thus, $P_{A(o)}$ = 1 and consequently,

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

For example, if the initial nucleotide is not A, then $P_{A(0)}$ = 0, and the probability of having A at this position at time t is

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



Temporal changes in the probability, *P*, of having a certain nucleotide at a position starting with either the same nucleotide (upper line) or with a different nucleotide (lower line). The dashed line denotes the equilibrium frequency (*P* = 0.25). $\alpha = 5 \times 10^{-9}$ substitutions per site per year.

Under the Jukes-Cantor model, the probability of each of the four nucleotides at equilibrium is 1/4.

After reaching equilibrium, there will be no further changes in probabilities, i.e., $P_{A(t)} = P_{T(t)} = P_{C(t)} = P_{G(t)} = 1/4$ all subsequent times.

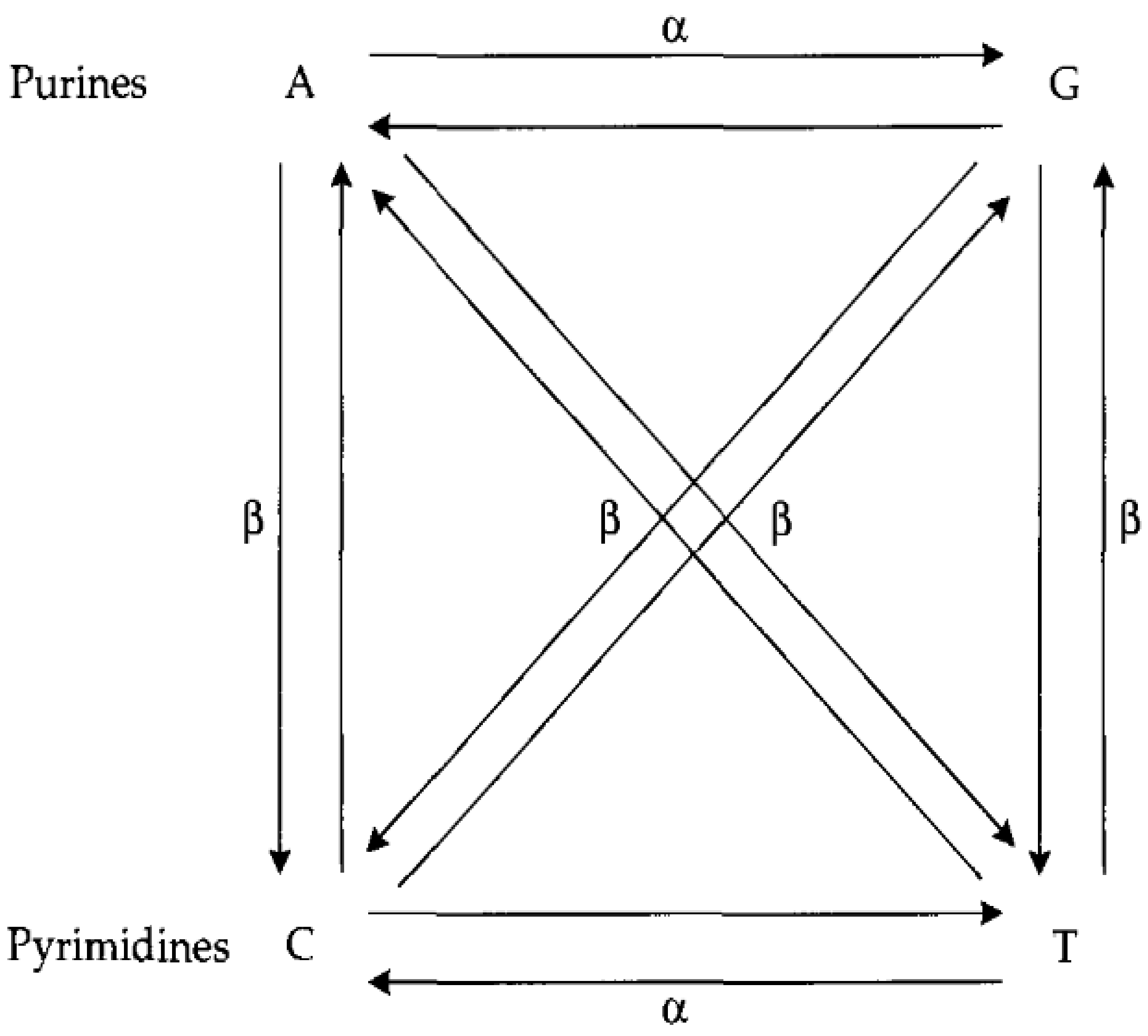
Kimura's two-parameter model

- The assumption that all **nucleotide substitutions occur with equal probability**, as in Jukes and Cantor's model, is unrealistic in most cases.

- For example, transitions (i.e., changes between A and G or between C and T) are generally more frequent than transversions.

- To take this fact into account, Kimura(1980) proposed a two-parameter model.

- In this scheme, the rate of transitional substitution at each nucleotide site is a per unit time, whereas the rate of each type of transversional substitution is β (3 per unit time).

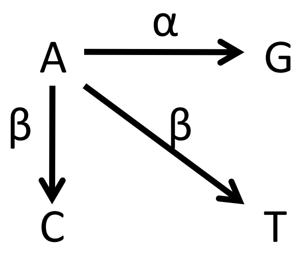


Two-parameter model of nucleotide substitution. The rate of transition (α) may not be equal to the rate of transversion (β).

Kimura’s Model

- Let us first consider the probability that a site that has A at time t = 0 will have A at time t.
- After one time unit t=1, the probability of A changing into G is α, and
- the probability of A changing into either C or T is 2β
- Thus, the probability of **A remaining unchanged** after one time unit is

$$P_{AA(1)} = 1 - \alpha - 2\beta$$



Four scenarios at time t = 2

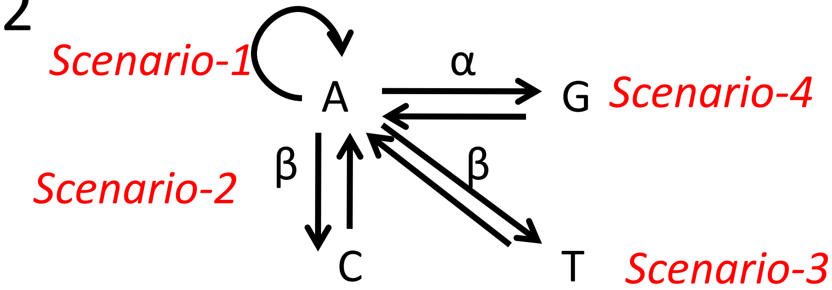
- At time t=2, the probability of having A at this site is given by the sum of the probabilities of four different scenarios:

- (1) A remained unchanged at t = 1 and t = 2;

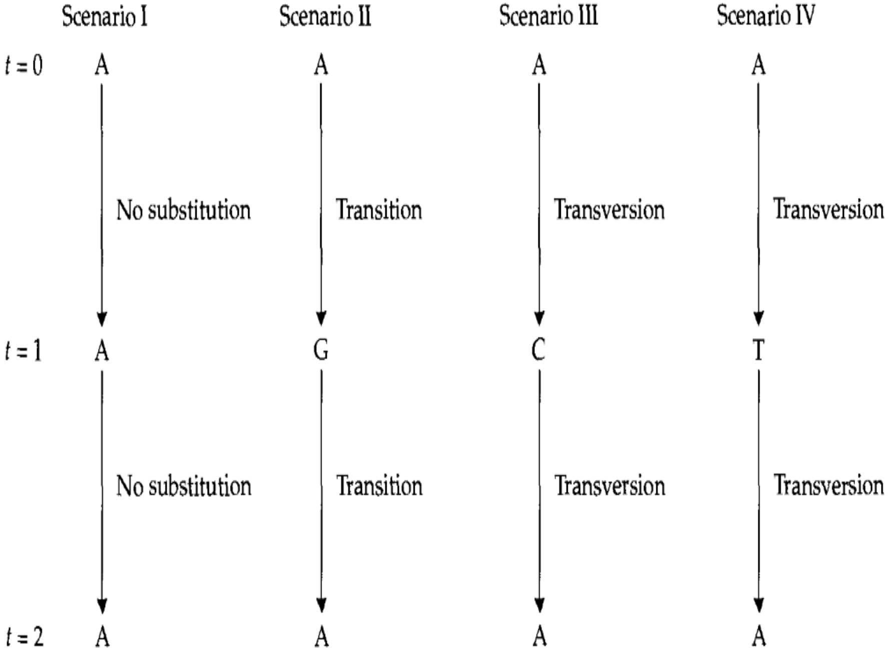
- (2) A changed into G at t = 1 and reverted by a transition to A at t = 2;

- (3) A changed into C at t = 1 and reverted by a transversion to A at t = 2;and

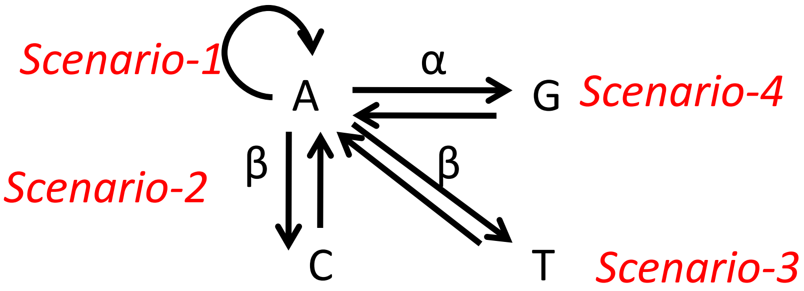
- (4) A changed into T at t = 1 and reverted by a transversion to A at t = 2



Possible mutation scenario’s in Kimura’s model



Four possible scenarios, according to Kimura's (1980) two-parameter model, for having A at a site at time $t = 2$, given that the site had A at time 0.



$$P_{\mathrm{AA}(2)} = (1-\alpha-2\beta)P_{\mathrm{AA}(1)} + \beta P_{\mathrm{TA}(1)} + \beta P_{\mathrm{CA}(1)} + \alpha P_{\mathrm{GA}(1)}$$

By extension we obtain the following recurrence equation for the general case:

$$P_{\mathrm{AA}(t+1)} = (1-\alpha-2\beta)P_{\mathrm{AA}(t)} + \beta P_{\mathrm{TA}(t)} + \beta P_{\mathrm{CA}(t)} + \alpha P_{\mathrm{GA}(t)}$$

Approximating the discrete-time model by the continuous-time model, we get

$$\frac{dP_{\mathrm{AA}(t)}}{dt} = -(\alpha+2\beta)P_{\mathrm{AA}(t)} + \beta P_{\mathrm{TA}(t)} + \beta P_{\mathrm{CA}(t)} + \alpha P_{\mathrm{GA}(t)}$$

We arrive at the following solution:

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

Thus, as in the case of Jukes and Cantor's model, the equilibrium frequencies of the four nucleotides are 1/4.

Substitution rate and mutation rate

- Two rates to consider:
- Mutation rate: rate at which new mutations arise
- Substitution rate: rate at which new mutations become fixed in a species depends on mutation rate and election.
- Mutation rates and substitution rates are related
- Substitutions can happen only after mutations occur
- But they refer to different processes.

Matrix models of base substitution

- We begin by modeling the ancestral sequence probabilistically.

- Each site in the sequence is one of the four bases A, G, C ,or T , chosen randomly according to some probabilities P_A , P_G , P_C , and P_T .

- These four probabilities must satisfy

$$\mathcal{P}_A + \mathcal{P}_G + \mathcal{P}_C + \mathcal{P}_T = 1$$

- For convenience, we will always use **the order A, G, C, T** for the bases (so the purines come first and then the pyrimidines) and put these four probabilities into a vector as

$$P_0 = (P_A, P_G, P_C, P_T).$$

- This vector describes the ancestral base distribution, with its entries giving the fraction of sites we would expect to be occupied by each of the four bases.

Construction of Matrix Model

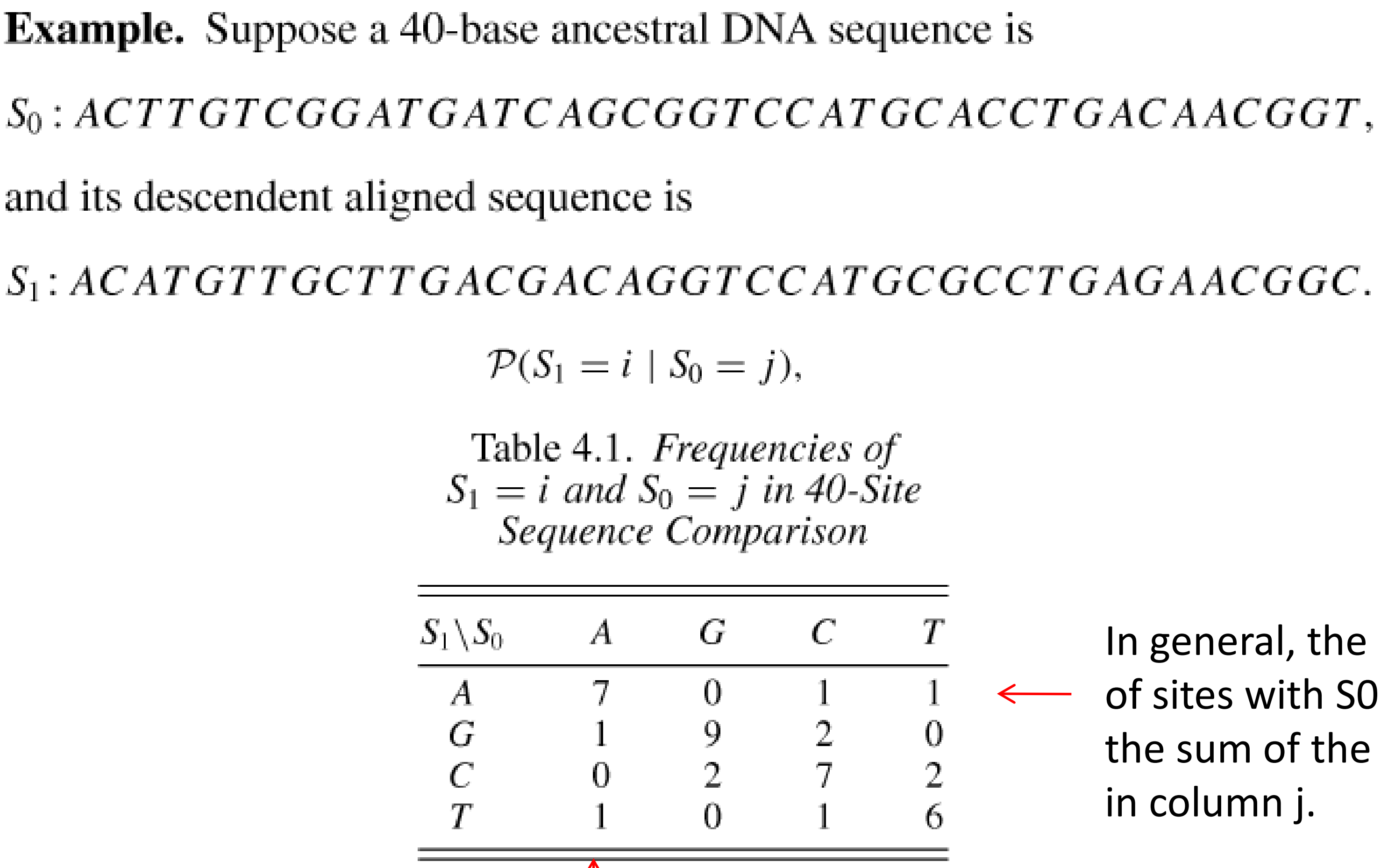
- We model the mutation process **over one time step**, assuming that only base substitutions can occur -- no deletions, insertions, or inversions are considered.
- We specify the 16 conditional probabilities of observing a base substitution,
- $P(S_1=i|S_0=j)$,for i, j = A, G, C, and T.
- It will be convenient to put these numbers into a 4 × 4 matrix, using the ordering A, G, C, and T .
- In column of the matrix are entries referring to the same ancestral base, and in each row are entries referring to the same descendent base.

Ancesteral base

$$M = \overset{\xrightarrow{\qquad\qquad\qquad\qquad\qquad}}{\begin{pmatrix} \mathcal{P}_{A|A} & \mathcal{P}_{A|G} & \mathcal{P}_{A|C} & \mathcal{P}_{A|T} \\ \mathcal{P}_{G|A} & \mathcal{P}_{G|G} & \mathcal{P}_{G|C} & \mathcal{P}_{G|T} \\ \mathcal{P}_{C|A} & \mathcal{P}_{C|G} & \mathcal{P}_{C|G} & \mathcal{P}_{C|T} \\ \mathcal{P}_{T|A} & \mathcal{P}_{T|G} & \mathcal{P}_{T|C} & \mathcal{P}_{T|T} \end{pmatrix}} \Bigg\downarrow$$

Descendant base

Example



Example. Suppose a 40-base ancestral DNA sequence is

$S_0 : ACTTGTCGGATGATCAGCGGTCCATGCACCTGACAACGGT,$

and its descendent aligned sequence is

$S_1 : ACATGTTGCTTGACGACAGGTCCATGCGCCTGAGAACGGC.$

$$\mathcal{P}(S_1 = i \mid S_0 = j),$$

Table 4.1. *Frequencies of $S_1 = i$ and $S_0 = j$ in 40-Site Sequence Comparison*

$S_1 \backslash S_0$	A	G	C	T
A	7	0	1	1
G	1	9	2	0
C	0	2	7	2
T	1	0	1	6

In general, the number of sites with S0 = j is the sum of the entries in column j.

Total number of sites with a particular base in S0. S0 = A is 7+1+0+1=9.

Conditional probability

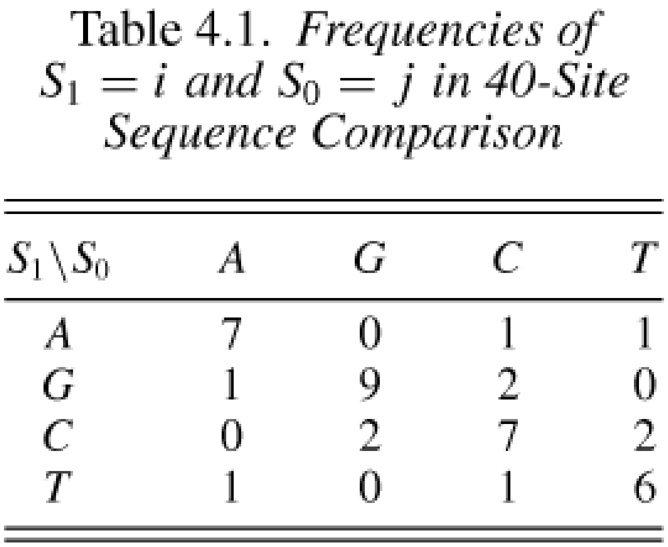


Table 4.1. *Frequencies of $S_1 = i$ and $S_0 = j$ in 40-Site Sequence Comparison*

$S_1 \backslash S_0$	A	G	C	T
A	7	0	1	1
G	1	9	2	0
C	0	2	7	2
T	1	0	1	6

$$\begin{array}{c} \xrightarrow{9/40} \\ \xrightarrow{12/40} \\ \xrightarrow{11/40} \\ \xrightarrow{8/40} \end{array} \begin{pmatrix} \mathcal{P}_A \\ \mathcal{P}_G \\ \mathcal{P}_C \\ \mathcal{P}_T \end{pmatrix}$$

approximate values;

sum adds to 1.

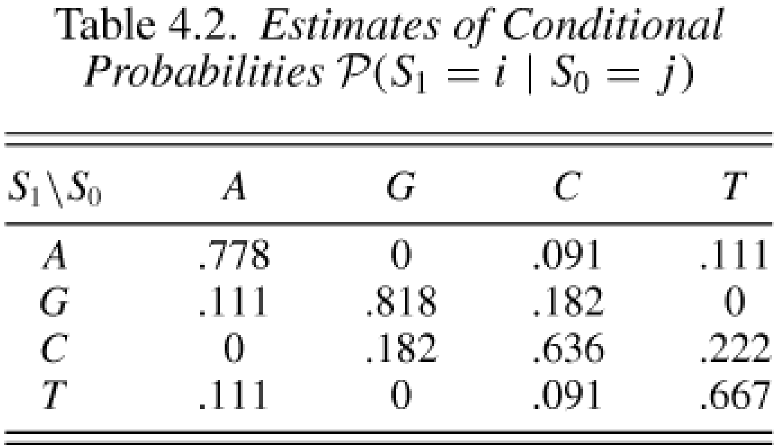


Table 4.2. *Estimates of Conditional Probabilities $\mathcal{P}(S_1 = i \mid S_0 = j)$*

$S_1 \backslash S_0$	A	G	C	T
A	.778	0	.091	.111
G	.111	.818	.182	0
C	0	.182	.636	.222
T	.111	0	.091	.667

Transition matrix and initial vector

$$\begin{pmatrix} \mathcal{P}_A \\ \mathcal{P}_G \\ \mathcal{P}_C \\ \mathcal{P}_T \end{pmatrix}$$

$$M = \begin{pmatrix} \mathcal{P}_{A|A} & \mathcal{P}_{A|G} & \mathcal{P}_{A|C} & \mathcal{P}_{A|T} \\ \mathcal{P}_{G|A} & \mathcal{P}_{G|G} & \mathcal{P}_{G|C} & \mathcal{P}_{G|T} \\ \mathcal{P}_{C|A} & \mathcal{P}_{C|G} & \mathcal{P}_{C|G} & \mathcal{P}_{C|T} \\ \mathcal{P}_{T|A} & \mathcal{P}_{T|G} & \mathcal{P}_{T|C} & \mathcal{P}_{T|T} \end{pmatrix}$$

Initial vector P_o

Transition matrix

probabilities of each base in the ancestral sequence S0 are transformed into the probabilities of each base in the descendent sequence S1 one time step later.

For a 40 site sequence comparison,

$$\mathbf{p}_0 \approx (.225, .275, .275, .225) \quad \text{and} \quad M \approx \begin{pmatrix} .778 & 0 & .091 & .111 \\ .111 & .818 & .182 & 0 \\ 0 & .182 & .636 & .222 \\ .111 & 0 & .091 & .667 \end{pmatrix}.$$

Interpretation

$$M\mathbf{p}_0 = \begin{pmatrix} \mathcal{P}_{A|A} & \mathcal{P}_{A|G} & \mathcal{P}_{A|C} & \mathcal{P}_{A|T} \\ \mathcal{P}_{G|A} & \mathcal{P}_{G|G} & \mathcal{P}_{G|C} & \mathcal{P}_{G|T} \\ \mathcal{P}_{C|A} & \mathcal{P}_{C|G} & \mathcal{P}_{C|C} & \mathcal{P}_{C|T} \\ \mathcal{P}_{T|A} & \mathcal{P}_{T|G} & \mathcal{P}_{T|C} & \mathcal{P}_{T|T} \end{pmatrix} \begin{pmatrix} \mathcal{P}_A \\ \mathcal{P}_G \\ \mathcal{P}_C \\ \mathcal{P}_T \end{pmatrix}$$

$$= \begin{pmatrix} \mathcal{P}_{A|A}\mathcal{P}_A + \mathcal{P}_{A|G}\mathcal{P}_G + \mathcal{P}_{A|C}\mathcal{P}_C + \mathcal{P}_{A|T}\mathcal{P}_T \\ \mathcal{P}_{C|A}\mathcal{P}_A + \mathcal{P}_{C|G}\mathcal{P}_G + \mathcal{P}_{C|C}\mathcal{P}_C + \mathcal{P}_{C|T}\mathcal{P}_T \\ \mathcal{P}_{G|A}\mathcal{P}_A + \mathcal{P}_{G|G}\mathcal{P}_G + \mathcal{P}_{G|C}\mathcal{P}_C + \mathcal{P}_{G|T}\mathcal{P}_T \\ \boxed{\mathcal{P}_{T|A}\mathcal{P}_A + \mathcal{P}_{T|G}\mathcal{P}_G + \mathcal{P}_{T|C}\mathcal{P}_C + \mathcal{P}_{T|T}\mathcal{P}_T} \end{pmatrix}.$$

$$\downarrow$$

$$\mathcal{P}_{T|A}\mathcal{P}_A + \mathcal{P}_{T|G}\mathcal{P}_G + \mathcal{P}_{T|C}\mathcal{P}_C + \mathcal{P}_{T|T}\mathcal{P}_T.$$

Informally, this gives the total probability that a site in S1 has base T.

$$\underbrace{\mathcal{P}_{T|A}\mathcal{P}_A} + \mathcal{P}_{T|G}\mathcal{P}_G + \mathcal{P}_{T|C}\mathcal{P}_C + \mathcal{P}_{T|T}\mathcal{P}_T.$$

$$\mathcal{P}_{T|A}\mathcal{P}_A = \mathcal{P}(S_1 = T \mid S_0 = A)\mathcal{P}(S_0 = A).$$

Markov models

How it evolves over next time step t=1 & t=2 from t = 0?

$$\mathbf{p}_1 = M\mathbf{p}_0 = \begin{pmatrix} .225 \\ .275 \\ .300 \\ .200 \end{pmatrix}, \quad \mathbf{p}_2 = M\mathbf{p}_1 = \begin{pmatrix} .222 \\ .274 \\ .320 \\ .183 \end{pmatrix}.$$

To **build Markov models** the following are required.

(a) States have to be defined (n)

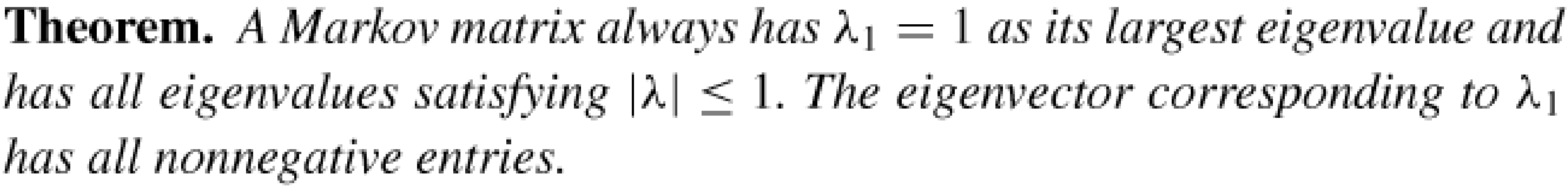
(b) Transition matrix ($M_{n \times n}$) has to provided and its also called Markov matrix.

(c) Initial population vector has to be provided for evolution of the system ($P_{n \times 1}$)

(d) The entries of M must all be ≥0 (because they are probabilities), and each column must add to one.

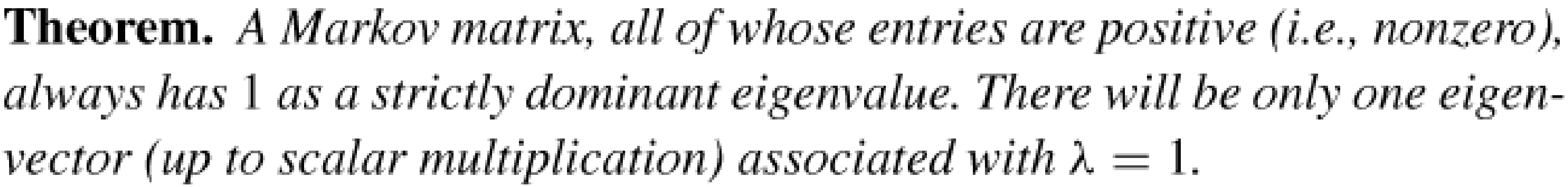
Assumption: No memory. The present is dependent only on the immediate earlier time step.

Two important theorems



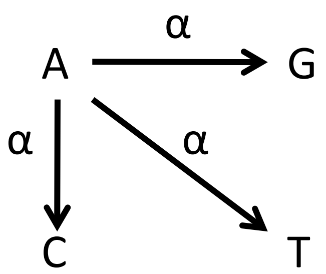
Theorem. *A Markov matrix always has* $\lambda_1 = 1$ *as its largest eigenvalue and has all eigenvalues satisfying* $|\lambda| \leq 1$. *The eigenvector corresponding to* λ_1 *has all nonnegative entries.*

This does not rule out -1 as an eigenvalue or having several different eigenvectors with eigenvalue 1.



Theorem. *A Markov matrix, all of whose entries are positive (i.e., nonzero), always has* 1 *as a strictly dominant eigenvalue. There will be only one eigenvector (up to scalar multiplication) associated with* $\lambda = 1$.

Jukes Cantor Model



$$M = \begin{pmatrix} 1-\alpha & \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & 1-\alpha & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & 1-\alpha & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} & 1-\alpha \end{pmatrix} \qquad \mathbf{p}_0 = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$$

Question: What proportion of the sites will each base appear after one time step?

$$\mathbf{p}_1 = M\mathbf{p}_0 = \begin{pmatrix} 1-\alpha & \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & 1-\alpha & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & 1-\alpha & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} & 1-\alpha \end{pmatrix} \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix} = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix}$$

Is it a Markov matrix?
Check!!!

Is it equilibrium base distribution?

The base composition of the sequence does not change under the Jukes-Cantor model.

In the language of linear algebra, we would say that the vector (1/4 , 1/4 , 1/4 , 1/4) is an eigenvector of M with eigenvalue 1. (Av = λv; Here v = P_o, and λ = 1).

Example-1

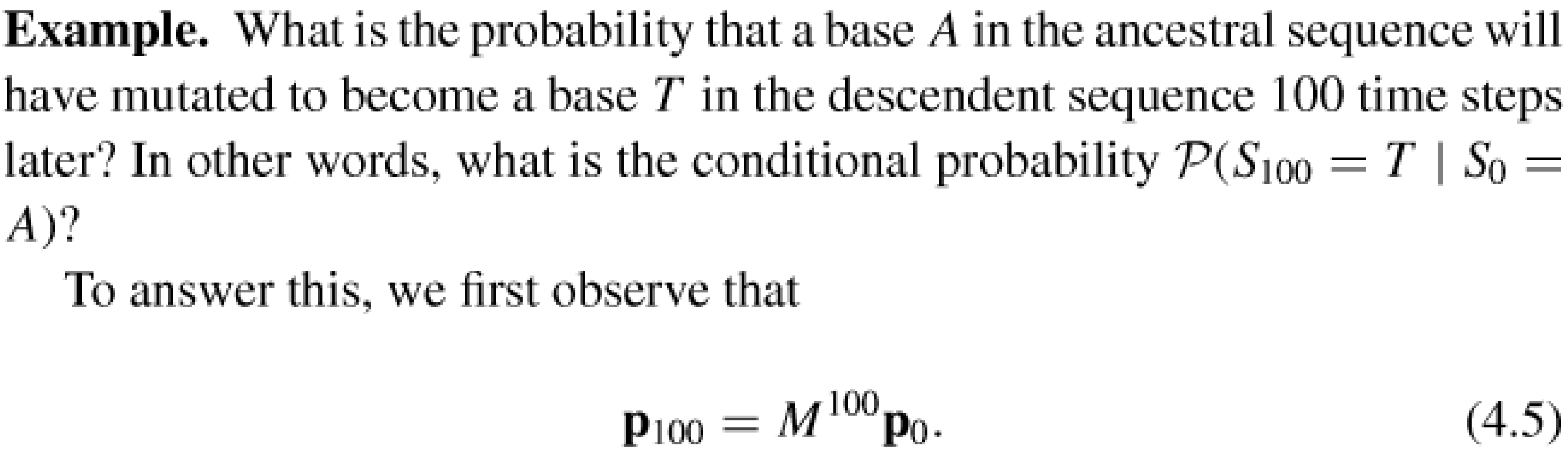
Example. What proportion of the sites will have a base A in the ancestral sequence and a T in the descendent one time step later? In other words, what is $p(S_0 = A \text{ and } S_1 = T)$?

$$M = \begin{pmatrix} 1-\alpha & \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & 1-\alpha & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & 1-\alpha & \frac{\alpha}{3} \\ \boxed{\frac{\alpha}{3}} & \frac{\alpha}{3} & \frac{\alpha}{3} & 1-\alpha \end{pmatrix} \qquad \mathbf{p}_0 = \left(\boxed{\frac{1}{4}}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$$

$$\mathcal{P}(S_0 = A \text{ and } S_1 = T) = \mathcal{P}(S_1 = T \mid S_0 = A)\mathcal{P}(S_0 = A).$$

Now the conditional probability $\mathcal{P}(S_1 = T \mid S_0 = A) = \frac{\alpha}{3}$ can be found as the (4,1) entry in M, while $\mathcal{P}(S_0 = A) = \frac{1}{4}$ is an entry in $\mathbf{p}_0$. Thus, $\mathcal{P}(S_0 = A \text{ and } S_1 = T) = \frac{\alpha}{12}$.

Example-2



Example. What is the probability that a base A in the ancestral sequence will have mutated to become a base T in the descendent sequence 100 time steps later? In other words, what is the conditional probability $\mathcal{P}(S_{100} = T \mid S_0 = A)$?

To answer this, we first observe that

$$\mathbf{p}_{100} = M^{100}\mathbf{p}_0. \tag{4.5}$$

Fall entries of M^t for all t will give the conditional probabilities of base substitutions over various numbers of time steps.

Calculate of M^t on the insight from eigenvectors provide the best approach to understanding how powers of matrices behave.

Also try for Kimura’s model.

(Try it as HW!)

Eigenvalues and Eigenvectors in JC model to generalize the model for any time

$$M = \begin{pmatrix} 1-\alpha & \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & 1-\alpha & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & 1-\alpha & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} & 1-\alpha \end{pmatrix} \qquad \mathbf{p}_0 = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$$

We require M^t to calculate the occurrence of base substitution for large time.

Step-1

Eigenvectors	Eigenvalues
$\mathbf{v}_1 = (1, 1, 1, 1)$	$\lambda_1 = 1$
$\mathbf{v}_2 = (1, 1, -1, -1)$	$\lambda_2 = 1 - \frac{4}{3}\alpha$
$\mathbf{v}_3 = (1, -1, 1, -1)$	$\lambda_3 = 1 - \frac{4}{3}\alpha$
$\mathbf{v}_4 = (1, -1, -1, 1)$	$\lambda_4 = 1 - \frac{4}{3}\alpha$

To find the entries of M^t, we begin by focusing on the first column of M^t The first column can be isolated by taking the product

Step-2

$$M^t \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \text{first column of } M^t.$$

$$M^t \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \text{first column of } M^t$$

Step-3 Now we can express $(1, 0, 0, 0)$ in terms of the eigenvectors as

$$(1, 0, 0, 0) = \frac{1}{4}\mathbf{v}_1 + \frac{1}{4}\mathbf{v}_2 + \frac{1}{4}\mathbf{v}_3 + \frac{1}{4}\mathbf{v}_4 .$$

Step-4 Express it in terms of eigenvalue equation

$$M^t \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \frac{1}{4}\underbrace{M^t \mathbf{v}_1} + \frac{1}{4}\underbrace{M^t \mathbf{v}_2} + \frac{1}{4}M^t \mathbf{v}_3 + \frac{1}{4}M^t \mathbf{v}_4$$

Eigenvalue equation: AV = λV

$$= \frac{1}{4}1^t \mathbf{v}_1 + \frac{1}{4}\left(1 - \frac{4}{3}\alpha\right)^t \mathbf{v}_2 + \frac{1}{4}\left(1 - \frac{4}{3}\alpha\right)^t \mathbf{v}_3 + \frac{1}{4}\left(1 - \frac{4}{3}\alpha\right)^t \mathbf{v}_4 .$$

Step-5 Substituting in the vectors $\mathbf{v}_i$, we find

$$M^t \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{4} + \frac{3}{4}\left(1 - \frac{4}{3}\alpha\right)^t \\ \frac{1}{4} - \frac{1}{4}\left(1 - \frac{4}{3}\alpha\right)^t \\ \frac{1}{4} - \frac{1}{4}\left(1 - \frac{4}{3}\alpha\right)^t \\ \frac{1}{4} - \frac{1}{4}\left(1 - \frac{4}{3}\alpha\right)^t \end{pmatrix} .$$

The other columns of M^t are found similarly, giving

$$M^t =$$

$$\begin{pmatrix} \frac{1}{4}+\frac{3}{4}\left(1-\frac{4}{3}\alpha\right)^t & \frac{1}{4}-\frac{1}{4}\left(1-\frac{4}{3}\alpha\right)^t & \frac{1}{4}-\frac{1}{4}\left(1-\frac{4}{3}\alpha\right)^t & \frac{1}{4}-\frac{1}{4}\left(1-\frac{4}{3}\alpha\right)^t \\ \frac{1}{4}-\frac{1}{4}\left(1-\frac{4}{3}\alpha\right)^t & \frac{1}{4}+\frac{3}{4}\left(1-\frac{4}{3}\alpha\right)^t & \frac{1}{4}-\frac{1}{4}\left(1-\frac{4}{3}\alpha\right)^t & \frac{1}{4}-\frac{1}{4}\left(1-\frac{4}{3}\alpha\right)^t \\ \frac{1}{4}-\frac{1}{4}\left(1-\frac{4}{3}\alpha\right)^t & \frac{1}{4}-\frac{1}{4}\left(1-\frac{4}{3}\alpha\right)^t & \frac{1}{4}+\frac{3}{4}\left(1-\frac{4}{3}\alpha\right)^t & \frac{1}{4}-\frac{1}{4}\left(1-\frac{4}{3}\alpha\right)^t \\ \frac{1}{4}-\frac{1}{4}\left(1-\frac{4}{3}\alpha\right)^t & \frac{1}{4}-\frac{1}{4}\left(1-\frac{4}{3}\alpha\right)^t & \frac{1}{4}-\frac{1}{4}\left(1-\frac{4}{3}\alpha\right)^t & \frac{1}{4}+\frac{3}{4}\left(1-\frac{4}{3}\alpha\right)^t \end{pmatrix}$$

Question: How do to find the other columns ?

To find the entries of M^t, we begin by focusing on the ~~first~~ 2nd column of M^t. The ~~first~~ 2 nd column can be isolated by taking the product

Step-2

$$M^t \cancel{\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}} \begin{matrix} 0 \\ 1 \\ 0 \\ 0 \end{matrix} = \text{first column of } M^t.$$

$$M^t \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \text{~~first column~~ 2 nd of } M^t$$

Express (0 1 0 0) in terms of eigenvectors
(0 1 0 0) = c1 v1 + c2 v2 + c3 v3 + c4 v4.
(find the weights c1, c2, c3, and c4)

Rest of the steps do it as HW using computers

Problems

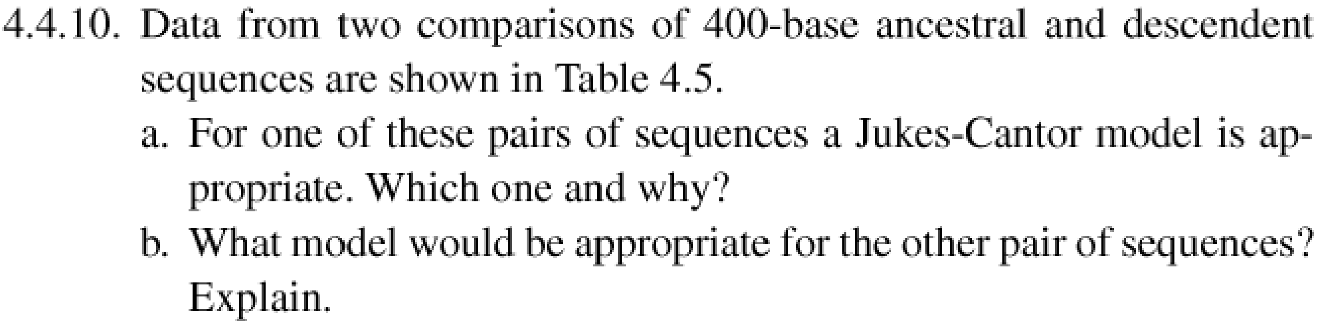
$$M = \begin{pmatrix} 1-\alpha & \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & 1-\alpha & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & 1-\alpha & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} & 1-\alpha \end{pmatrix}$$

JC

$$M = \begin{pmatrix} * & \beta & \gamma & \gamma \\ \beta & * & \gamma & \gamma \\ \gamma & \gamma & * & \beta \\ \gamma & \gamma & \beta & * \end{pmatrix}$$

Kimura

What are β's and γ's?
What should be the diagonal (*) values?



4.4.10. Data from two comparisons of 400-base ancestral and descendent sequences are shown in Table 4.5.

a. For one of these pairs of sequences a Jukes-Cantor model is appropriate. Which one and why?

b. What model would be appropriate for the other pair of sequences? Explain.

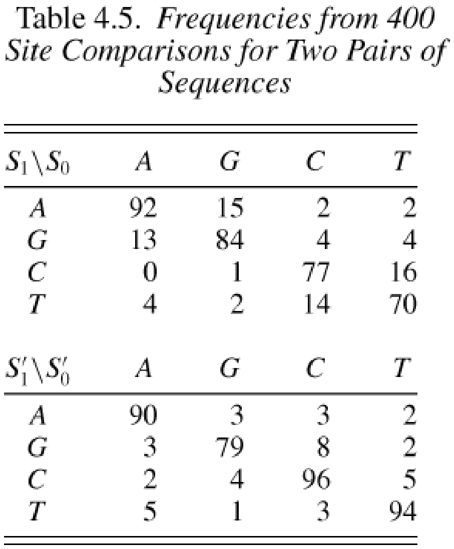


Table 4.5. *Frequencies from 400 Site Comparisons for Two Pairs of Sequences*

$S_1 \backslash S_0$	A	G	C	T
A	92	15	2	2
G	13	84	4	4
C	0	1	77	16
T	4	2	14	70
$S_1' \backslash S_0'$	A	G	C	T
A	90	3	3	2
G	3	79	8	2
C	2	4	96	5
T	5	1	3	94

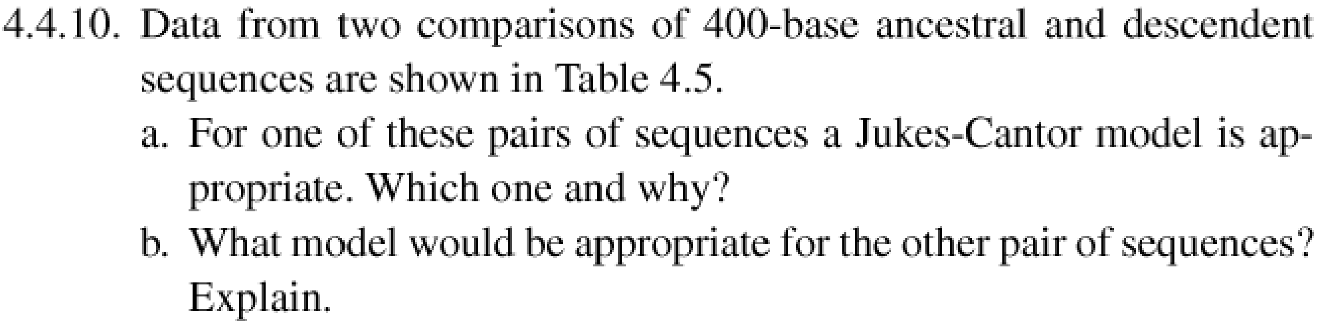
Solutions

$$M = \begin{pmatrix} 1-\alpha & \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & 1-\alpha & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & 1-\alpha & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} & 1-\alpha \end{pmatrix} \qquad M = \begin{pmatrix} * & \beta & \gamma & \gamma \\ \beta & * & \gamma & \gamma \\ \gamma & \gamma & * & \beta \\ \gamma & \gamma & \beta & * \end{pmatrix}$$

JC **Kimura**

What are β's and γ's? mutation rates β for transitions and γ for transversions

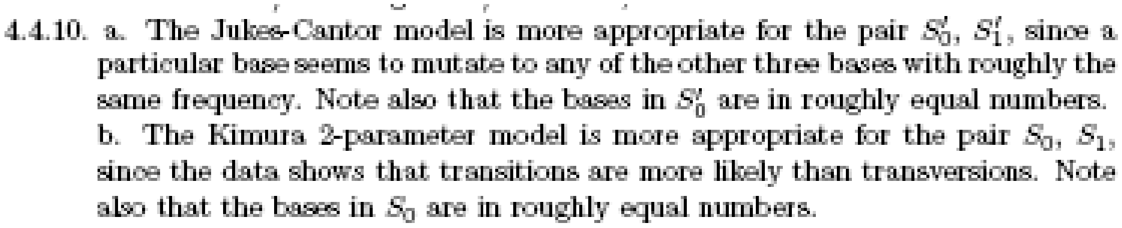
What should be the diagonal (*) values? **1 - β - 2γ .**



4.4.10. Data from two comparisons of 400-base ancestral and descendent sequences are shown in Table 4.5.

a. For one of these pairs of sequences a Jukes-Cantor model is appropriate. Which one and why?

b. What model would be appropriate for the other pair of sequences? Explain.



4.4.10. a. The Jukes-Cantor model is more appropriate for the pair S_0', S_1', since a particular base seems to mutate to any of the other three bases with roughly the same frequency. Note also that the bases in S_0' are in roughly equal numbers.
b. The Kimura 2-parameter model is more appropriate for the pair S_0, S_1, since the data shows that transitions are more likely than transversions. Note also that the bases in S_0 are in roughly equal numbers.

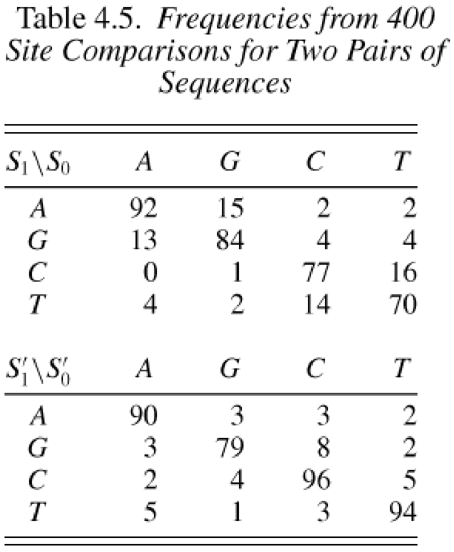
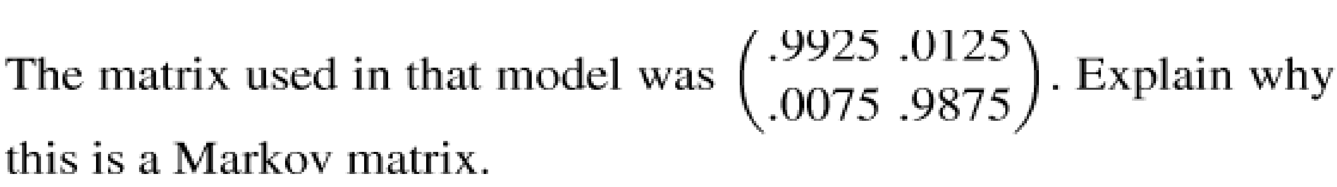


Table 4.5. *Frequencies from 400 Site Comparisons for Two Pairs of Sequences*

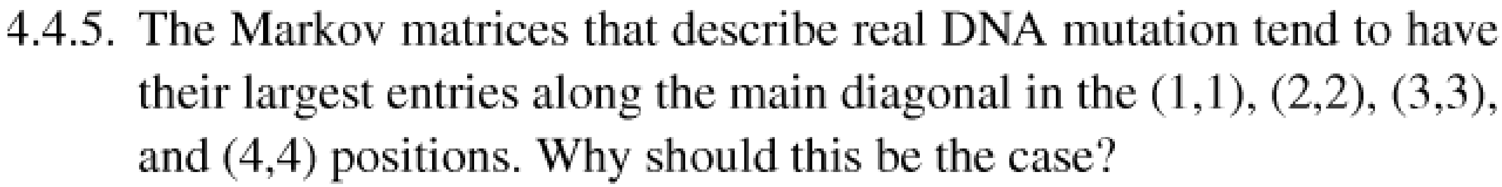
$S_1 \setminus S_0$	A	G	C	T
A	92	15	2	2
G	13	84	4	4
C	0	1	77	16
T	4	2	14	70

$S_1' \setminus S_0'$	A	G	C	T
A	90	3	3	2
G	3	79	8	2
C	2	4	96	5
T	5	1	3	94

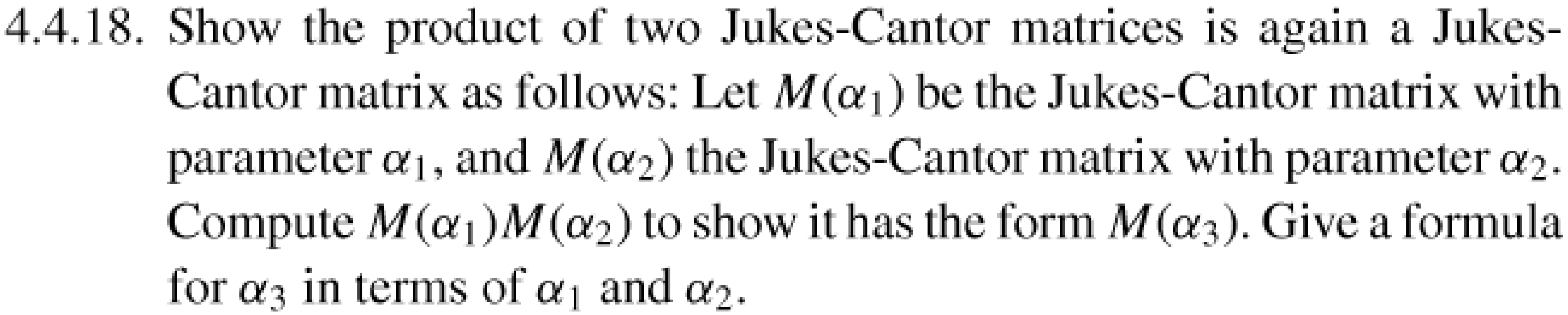
Problems



The matrix used in that model was $\begin{pmatrix} .9925 & .0125 \\ .0075 & .9875 \end{pmatrix}$. Explain why this is a Markov matrix.

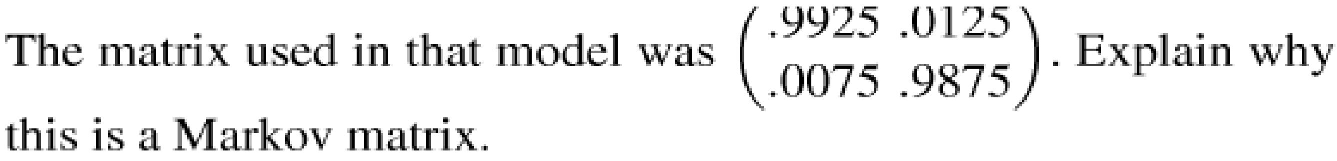


4.4.5. The Markov matrices that describe real DNA mutation tend to have their largest entries along the main diagonal in the (1,1), (2,2), (3,3), and (4,4) positions. Why should this be the case?

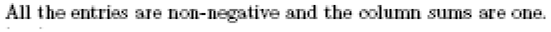


4.4.18. Show the product of two Jukes-Cantor matrices is again a Jukes-Cantor matrix as follows: Let $M(\alpha_1)$ be the Jukes-Cantor matrix with parameter α_1, and $M(\alpha_2)$ the Jukes-Cantor matrix with parameter α_2. Compute $M(\alpha_1)M(\alpha_2)$ to show it has the form $M(\alpha_3)$. Give a formula for α_3 in terms of α_1 and α_2.

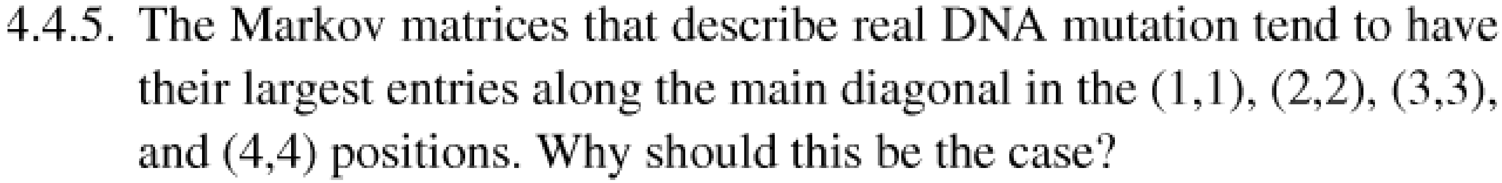
Solutions



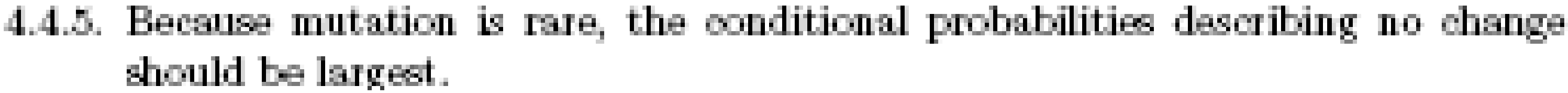
The matrix used in that model was $\begin{pmatrix} .9925 & .0125 \\ .0075 & .9875 \end{pmatrix}$. Explain why this is a Markov matrix.



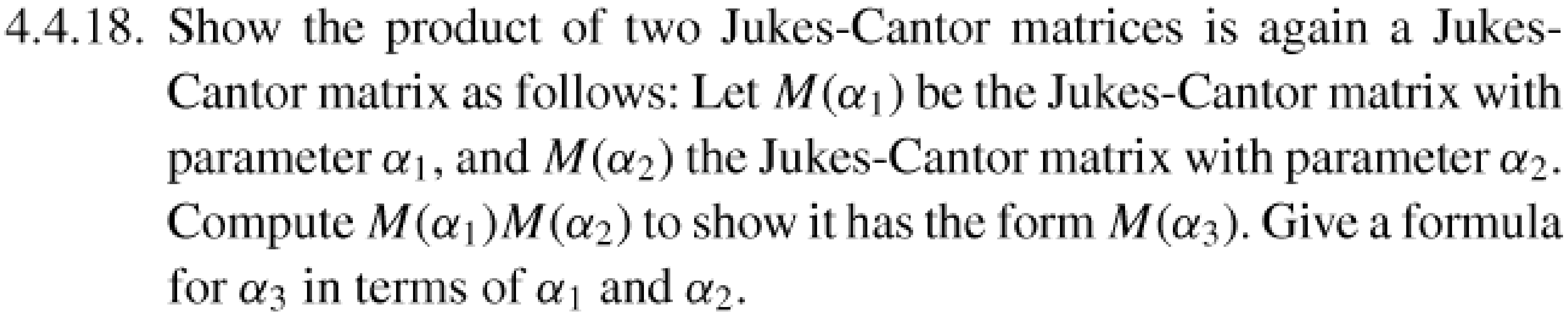
All the entries are non-negative and the column sums are one.



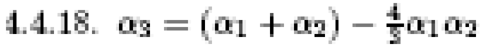
4.4.5. The Markov matrices that describe real DNA mutation tend to have their largest entries along the main diagonal in the (1,1), (2,2), (3,3), and (4,4) positions. Why should this be the case?



4.4.5. Because mutation is rare, the conditional probabilities describing no change should be largest.

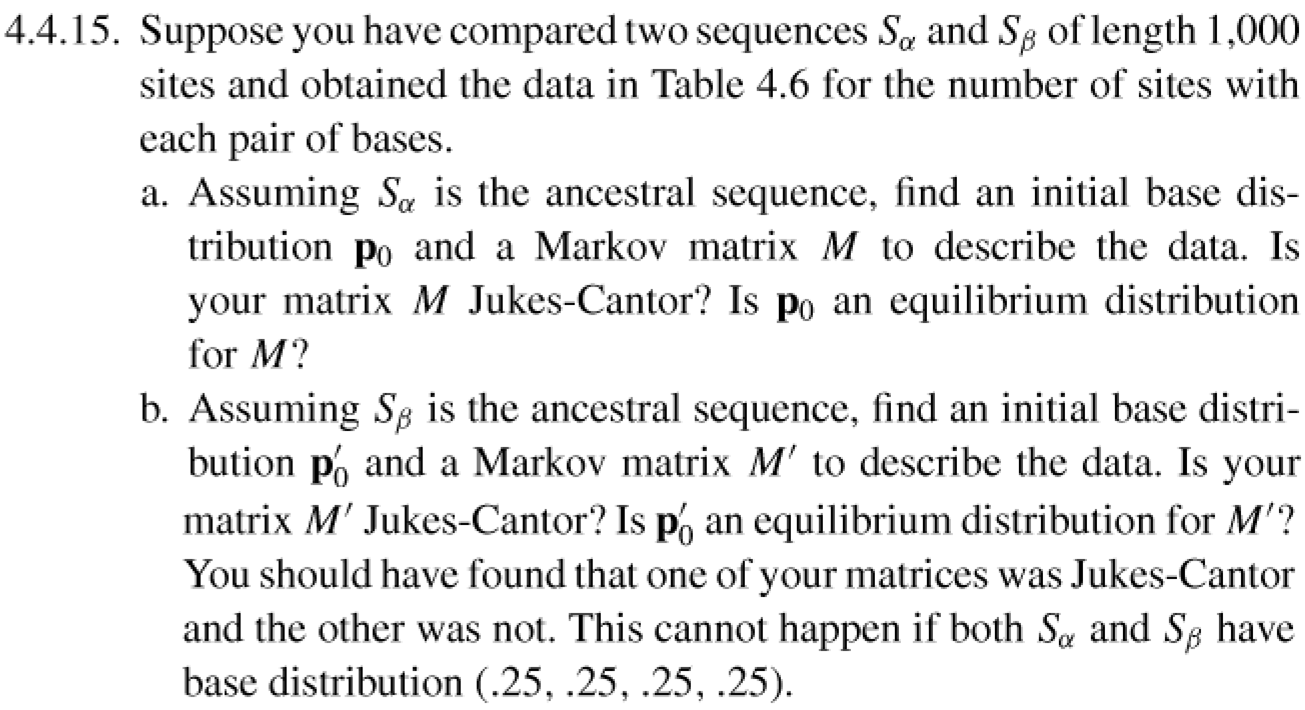


4.4.18. Show the product of two Jukes-Cantor matrices is again a Jukes-Cantor matrix as follows: Let $M(\alpha_1)$ be the Jukes-Cantor matrix with parameter α_1, and $M(\alpha_2)$ the Jukes-Cantor matrix with parameter α_2. Compute $M(\alpha_1)M(\alpha_2)$ to show it has the form $M(\alpha_3)$. Give a formula for α_3 in terms of α_1 and α_2.



4.4.18. $\alpha_3 = (\alpha_1 + \alpha_2) - \frac{4}{3}\alpha_1\alpha_2$

Problems



4.4.15. Suppose you have compared two sequences S_α and S_β of length 1,000 sites and obtained the data in Table 4.6 for the number of sites with each pair of bases.

a. Assuming S_α is the ancestral sequence, find an initial base distribution $\mathbf{p}_0$ and a Markov matrix M to describe the data. Is your matrix M Jukes-Cantor? Is $\mathbf{p}_0$ an equilibrium distribution for M?

b. Assuming S_β is the ancestral sequence, find an initial base distribution $\mathbf{p}_0'$ and a Markov matrix M' to describe the data. Is your matrix M' Jukes-Cantor? Is $\mathbf{p}_0'$ an equilibrium distribution for M'? You should have found that one of your matrices was Jukes-Cantor and the other was not. This cannot happen if both S_α and S_β have base distribution (.25, .25, .25, .25).

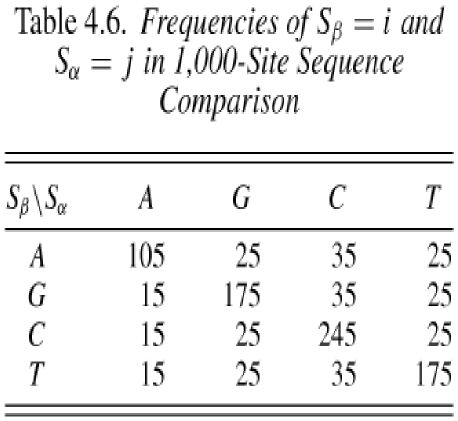
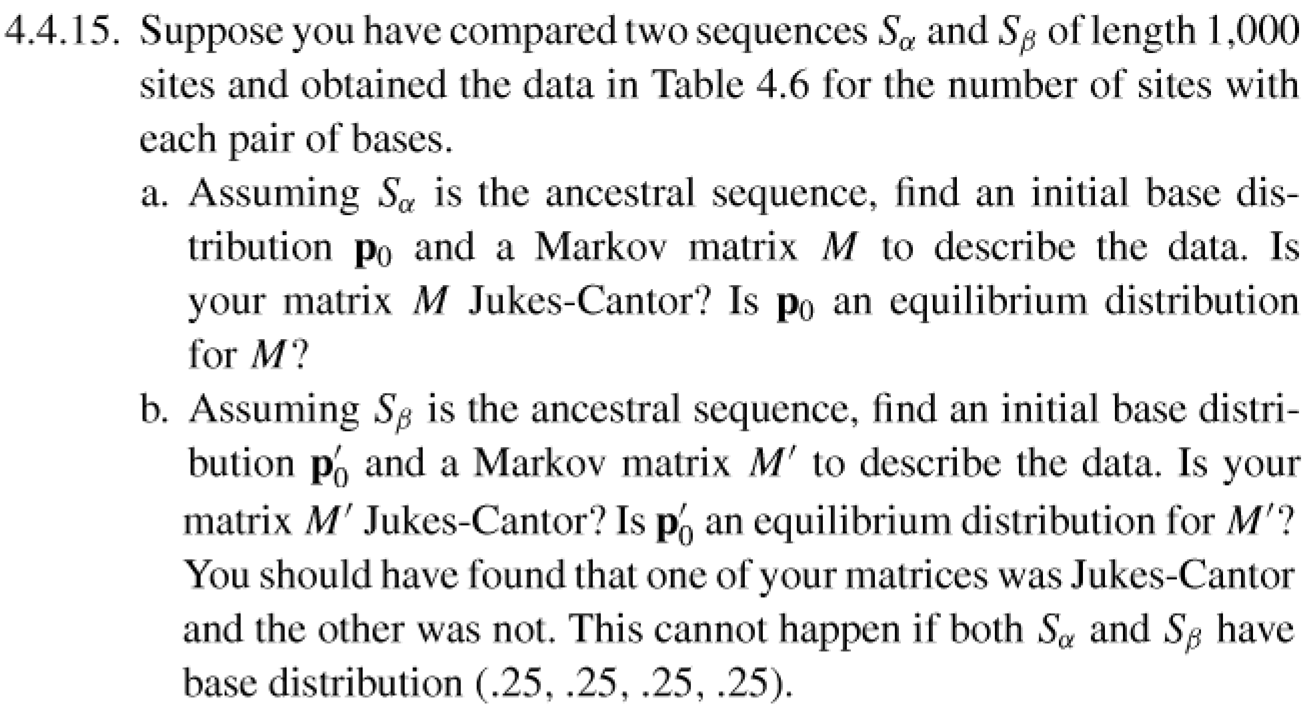


Table 4.6. *Frequencies of $S_\beta = i$ and $S_\alpha = j$ in 1,000-Site Sequence Comparison*

$S_\beta \backslash S_\alpha$	*A*	*G*	*C*	*T*
A	105	25	35	25
G	15	175	35	25
C	15	25	245	25
T	15	25	35	175

Solutions



4.4.15. Suppose you have compared two sequences S_α and S_β of length 1,000 sites and obtained the data in Table 4.6 for the number of sites with each pair of bases.

a. Assuming S_α is the ancestral sequence, find an initial base distribution $\mathbf{p}_0$ and a Markov matrix M to describe the data. Is your matrix M Jukes-Cantor? Is $\mathbf{p}_0$ an equilibrium distribution for M?

b. Assuming S_β is the ancestral sequence, find an initial base distribution $\mathbf{p}_0'$ and a Markov matrix M' to describe the data. Is your matrix M' Jukes-Cantor? Is $\mathbf{p}_0'$ an equilibrium distribution for M'? You should have found that one of your matrices was Jukes-Cantor and the other was not. This cannot happen if both S_α and S_β have base distribution (.25, .25, .25, .25).

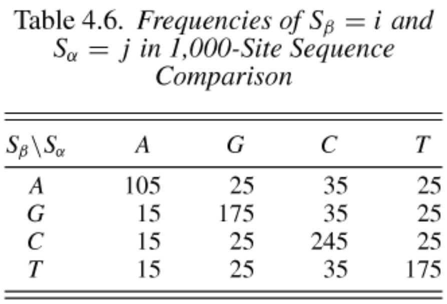
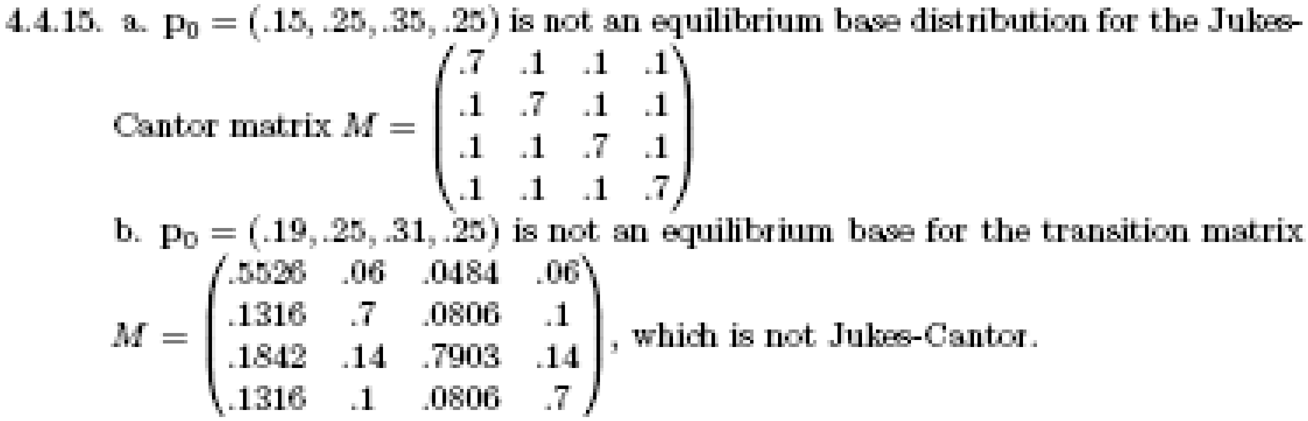


Table 4.6. *Frequencies of $S_\beta = i$ and $S_\alpha = j$ in 1,000-Site Sequence Comparison*

$S_\beta \backslash S_\alpha$	*A*	*G*	*C*	*T*
A	105	25	35	25
G	15	175	35	25
C	15	25	245	25
T	15	25	35	175



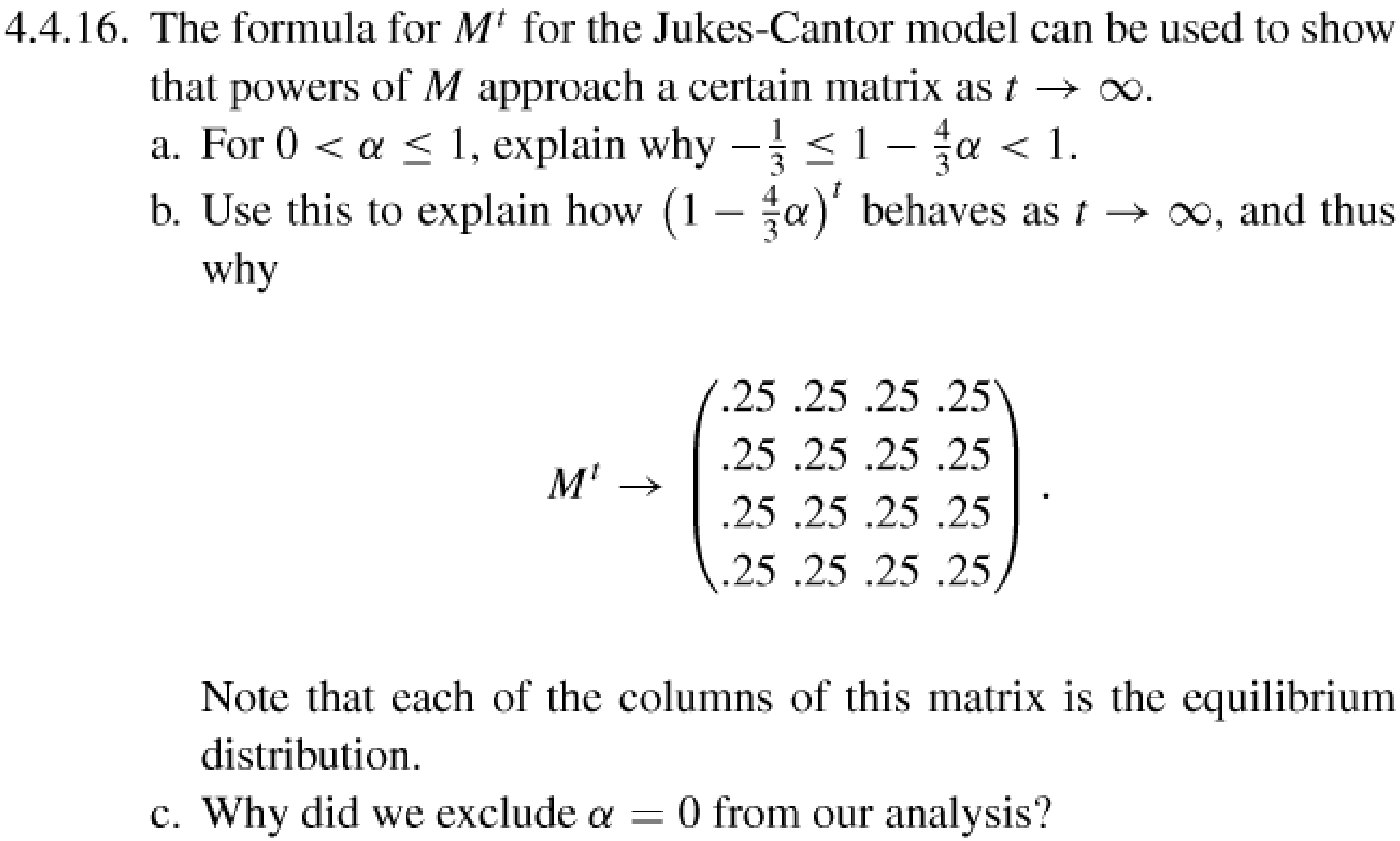
4.4.15. a. $\mathbf{p}_0 = (.15, .25, .35, .25)$ is not an equilibrium base distribution for the Jukes-Cantor matrix $M = \begin{pmatrix} .7 & .1 & .1 & .1 \\ .1 & .7 & .1 & .1 \\ .1 & .1 & .7 & .1 \\ .1 & .1 & .1 & .7 \end{pmatrix}$

b. $\mathbf{p}_0 = (.19, .25, .31, .25)$ is not an equilibrium base for the transition matrix

$$M = \begin{pmatrix} .5526 & .06 & .0484 & .06 \\ .1316 & .7 & .0806 & .1 \\ .1842 & .14 & .7903 & .14 \\ .1316 & .1 & .0806 & .7 \end{pmatrix},$$ which is not Jukes-Cantor.

(Check the P_0's for both)

Problems



4.4.16. The formula for M^t for the Jukes-Cantor model can be used to show that powers of M approach a certain matrix as $t \to \infty$.

a. For $0 < \alpha \leq 1$, explain why $-\frac{1}{3} \leq 1 - \frac{4}{3}\alpha < 1$.

b. Use this to explain how $\left(1 - \frac{4}{3}\alpha\right)^t$ behaves as $t \to \infty$, and thus why

$$M^t \to \begin{pmatrix} .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \end{pmatrix}.$$

Note that each of the columns of this matrix is the equilibrium distribution.

c. Why did we exclude $\alpha = 0$ from our analysis?

Problems

4.4.22. The Jukes-Cantor model can be presented in a different form as a 2×2 Markov model. Let q_t represent the fraction of sites that agree between the ancestral sequence and the descendent sequence at time t, and p_t the fraction that differ, so $q_0 = 1$ and $p_0 = 0$. Assume that over each time step, the probability that a base substitution occurs is α, and that each of the three possible base substitutions is equally likely. Then

$$\begin{pmatrix} q_{t+1} \\ p_{t+1} \end{pmatrix} = \begin{pmatrix} 1-\alpha & \frac{\alpha}{3} \\ \alpha & 1-\frac{\alpha}{3} \end{pmatrix} \begin{pmatrix} q_t \\ p_t \end{pmatrix}, \quad \begin{pmatrix} q_0 \\ p_0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

a. Explain why each entry in the matrix has the value it does. (Observe that $1 - \frac{\alpha}{3} = (1-\alpha) + \frac{2\alpha}{3}$.)

b. Compute the steady state of the model by finding the eigenvector with eigenvalue 1.

c. Find the other eigenvalue and eigenvector for the matrix.

d. Use parts (b) and (c), together with the initial conditions $(q_0, p_0) = (1, 0)$, to give a formula for q_t and p_t as functions of time.

Solutions

4.4.22. The Jukes-Cantor model can be presented in a different form as a 2×2 Markov model. Let q_t represent the fraction of sites that agree between the ancestral sequence and the descendent sequence at time t, and p_t the fraction that differ, so $q_0 = 1$ and $p_0 = 0$. Assume that over each time step, the probability that a base substitution occurs is α, and that each of the three possible base substitutions is equally likely. Then

$$\begin{pmatrix} q_{t+1} \\ p_{t+1} \end{pmatrix} = \begin{pmatrix} 1-\alpha & \frac{\alpha}{3} \\ \alpha & 1-\frac{\alpha}{3} \end{pmatrix} \begin{pmatrix} q_t \\ p_t \end{pmatrix}, \quad \begin{pmatrix} q_0 \\ p_0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

a. Explain why each entry in the matrix has the value it does. (Observe that $1 - \frac{\alpha}{3} = (1-\alpha) + \frac{2\alpha}{3}$.)

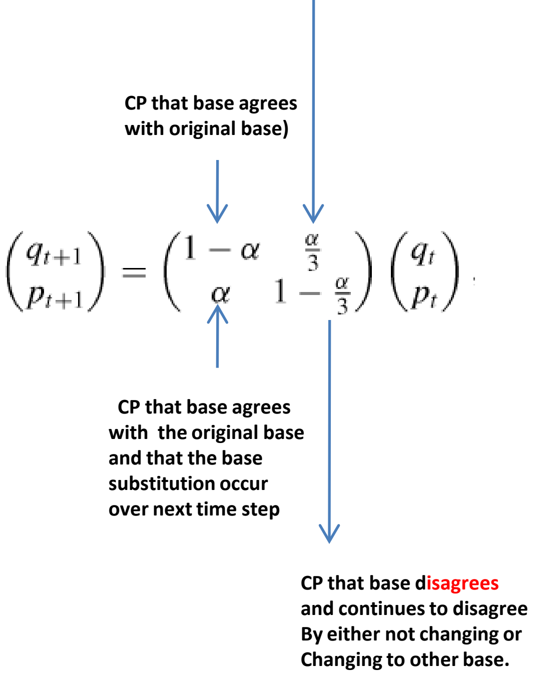
b. Compute the steady state of the model by finding the eigenvector with eigenvalue 1. (1/4, 3/4)

c. Find the other eigenvalue and eigenvector for the matrix. c. $\lambda = 1 - 4\alpha/3$ with eigenvector $(1, -1)$.

d. Use parts (b) and (c), together with the initial conditions $(q_0, p_0) = (1, 0)$, to give a formula for q_t and p_t as functions of time.

c. $\lambda = 1 - 4\alpha/3$ with eigenvector $(1, -1)$.

d. Since $\begin{pmatrix} q_t \\ p_t \end{pmatrix} = M^t \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, and $\begin{pmatrix} 1 \\ 0 \end{pmatrix} = 1 \begin{pmatrix} 1/4 \\ 3/4 \end{pmatrix} + \frac{3}{4} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$, then $q_t = \frac{1}{4} + \frac{3}{4}(1 - \frac{4\alpha}{3})^t$ and $p_t = \frac{3}{4} - \frac{3}{4}(1 - \frac{4\alpha}{3})^t$.



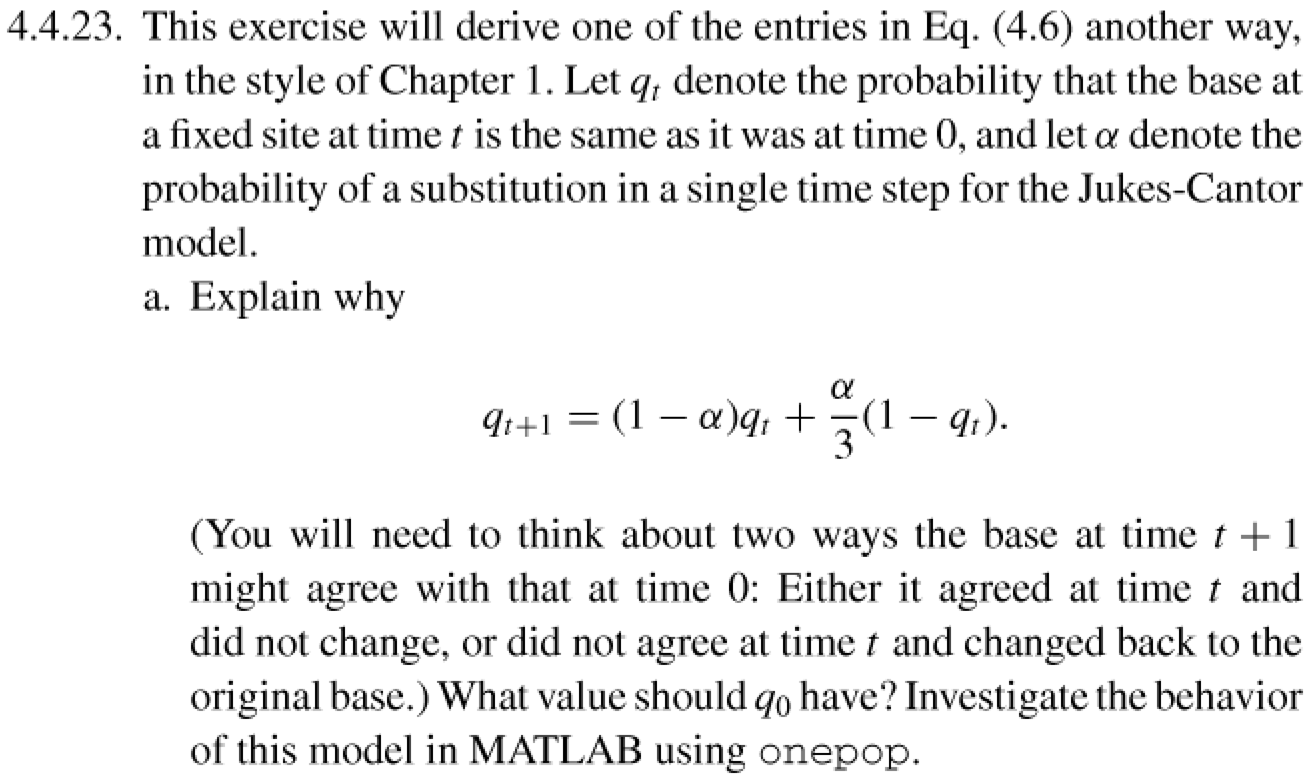
$$\begin{pmatrix} q_{t+1} \\ p_{t+1} \end{pmatrix} = \begin{pmatrix} 1-\alpha & \frac{\alpha}{3} \\ \alpha & 1-\frac{\alpha}{3} \end{pmatrix} \begin{pmatrix} q_t \\ p_t \end{pmatrix}.$$

CP that base agrees with original base)

CP that base agrees with the original base and that the base substitution occur over next time step

CP that base disagrees and continues to disagree By either not changing or Changing to other base.

Problems



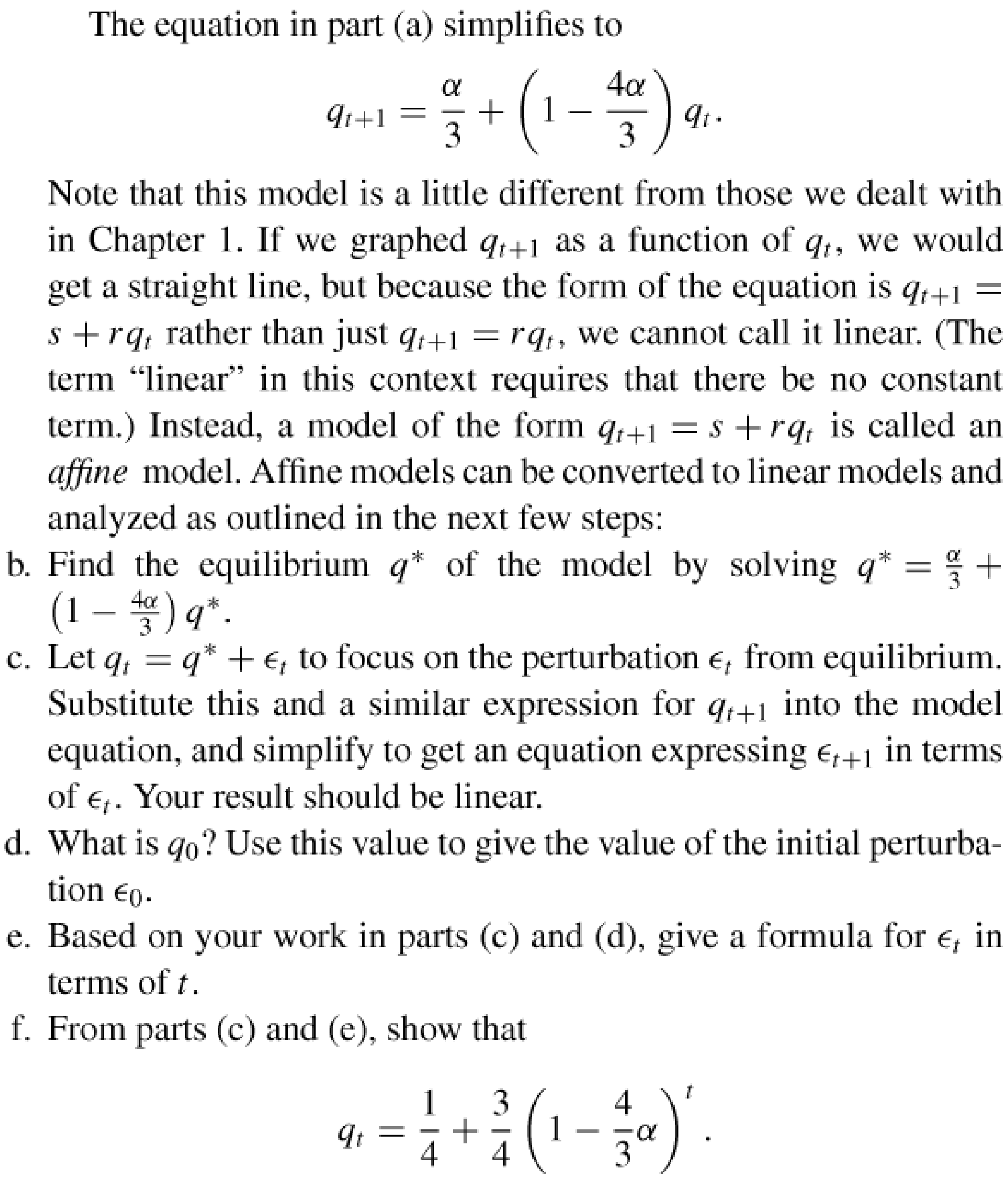
4.4.23. This exercise will derive one of the entries in Eq. (4.6) another way, in the style of Chapter 1. Let q_t denote the probability that the base at a fixed site at time t is the same as it was at time 0, and let α denote the probability of a substitution in a single time step for the Jukes-Cantor model.

a. Explain why

$$q_{t+1} = (1-\alpha)q_t + \frac{\alpha}{3}(1-q_t).$$

(You will need to think about two ways the base at time $t+1$ might agree with that at time 0: Either it agreed at time t and did not change, or did not agree at time t and changed back to the original base.) What value should q_0 have? Investigate the behavior of this model in MATLAB using `onepop`.

Continued from earlier slide.



The equation in part (a) simplifies to

$$q_{t+1} = \frac{\alpha}{3} + \left(1 - \frac{4\alpha}{3}\right) q_t .$$

Note that this model is a little different from those we dealt with in Chapter 1. If we graphed q_{t+1} as a function of q_t, we would get a straight line, but because the form of the equation is $q_{t+1} = s + rq_t$ rather than just $q_{t+1} = rq_t$, we cannot call it linear. (The term "linear" in this context requires that there be no constant term.) Instead, a model of the form $q_{t+1} = s + rq_t$ is called an *affine* model. Affine models can be converted to linear models and analyzed as outlined in the next few steps:

b. Find the equilibrium q^* of the model by solving $q^* = \frac{\alpha}{3} + \left(1 - \frac{4\alpha}{3}\right) q^*$.

c. Let $q_t = q^* + \epsilon_t$ to focus on the perturbation ϵ_t from equilibrium. Substitute this and a similar expression for q_{t+1} into the model equation, and simplify to get an equation expressing ϵ_{t+1} in terms of ϵ_t. Your result should be linear.

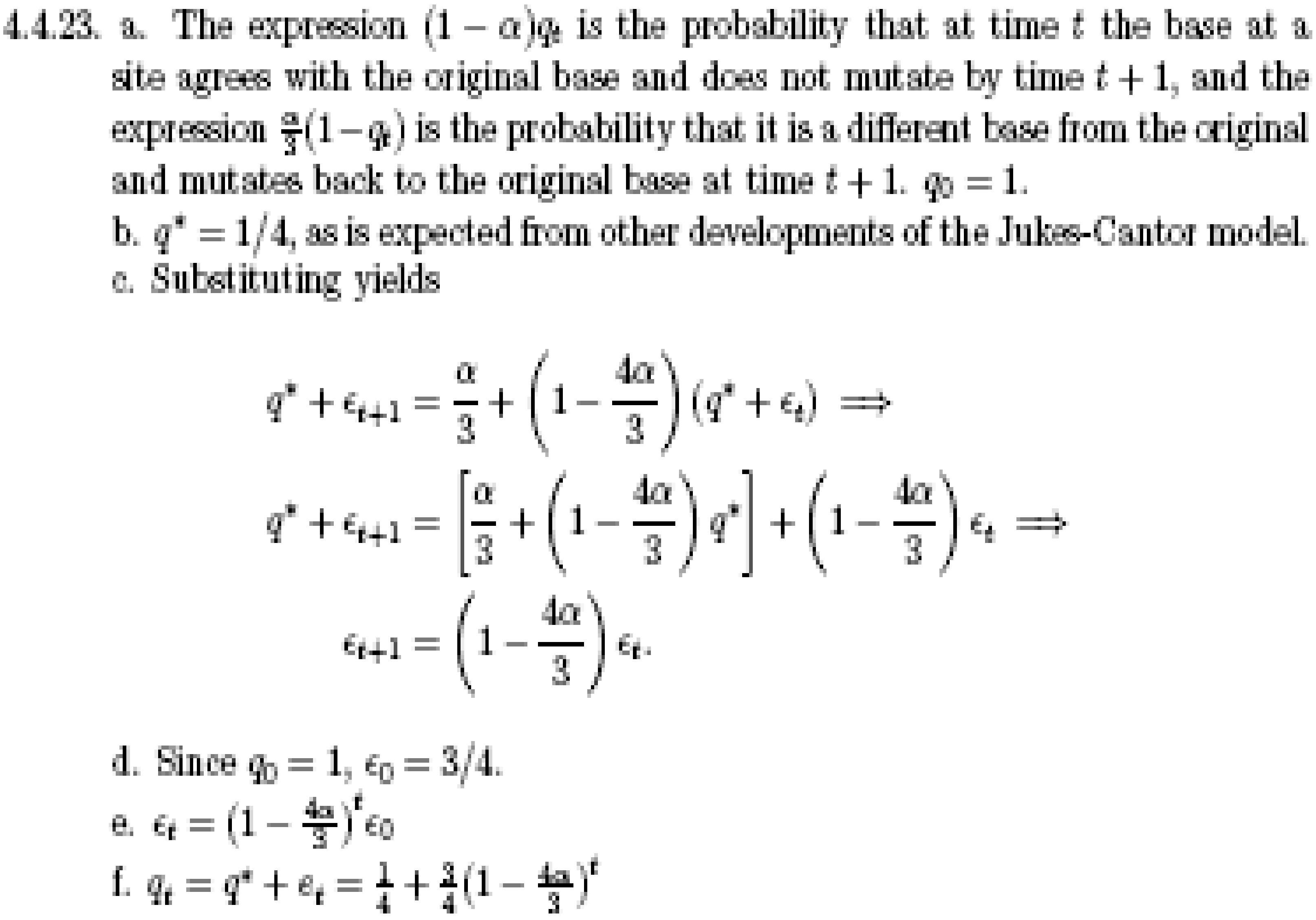
d. What is q_0? Use this value to give the value of the initial perturbation ϵ_0.

e. Based on your work in parts (c) and (d), give a formula for ϵ_t in terms of t.

f. From parts (c) and (e), show that

$$q_t = \frac{1}{4} + \frac{3}{4}\left(1 - \frac{4}{3}\alpha\right)^t .$$

Solutions



4.4.23. a. The expression $(1-\alpha)q_t$ is the probability that at time t the base at a site agrees with the original base and does not mutate by time $t+1$, and the expression $\frac{\alpha}{3}(1-q_t)$ is the probability that it is a different base from the original and mutates back to the original base at time $t+1$. $q_0 = 1$.

b. $q^* = 1/4$, as is expected from other developments of the Jukes-Cantor model.

c. Substituting yields

$$q^* + \epsilon_{t+1} = \frac{\alpha}{3} + \left(1 - \frac{4\alpha}{3}\right)(q^* + \epsilon_t) \implies$$

$$q^* + \epsilon_{t+1} = \left[\frac{\alpha}{3} + \left(1 - \frac{4\alpha}{3}\right)q^*\right] + \left(1 - \frac{4\alpha}{3}\right)\epsilon_t \implies$$

$$\epsilon_{t+1} = \left(1 - \frac{4\alpha}{3}\right)\epsilon_t.$$

d. Since $q_0 = 1$, $\epsilon_0 = 3/4$.

e. $\epsilon_t = (1 - \frac{4\alpha}{3})^t \epsilon_0$

f. $q_t = q^* + \epsilon_t = \frac{1}{4} + \frac{3}{4}(1 - \frac{4\alpha}{3})^t$

Phylogenetic distances

- "Distance" here is an abstract notion of how different the sequences are because of mutations.

- Two models with variations are proposed for the molecular evolution of mutation.

- Kimura’s is a two-parameter (K2P) and three parameter (K3P) model.

K2P

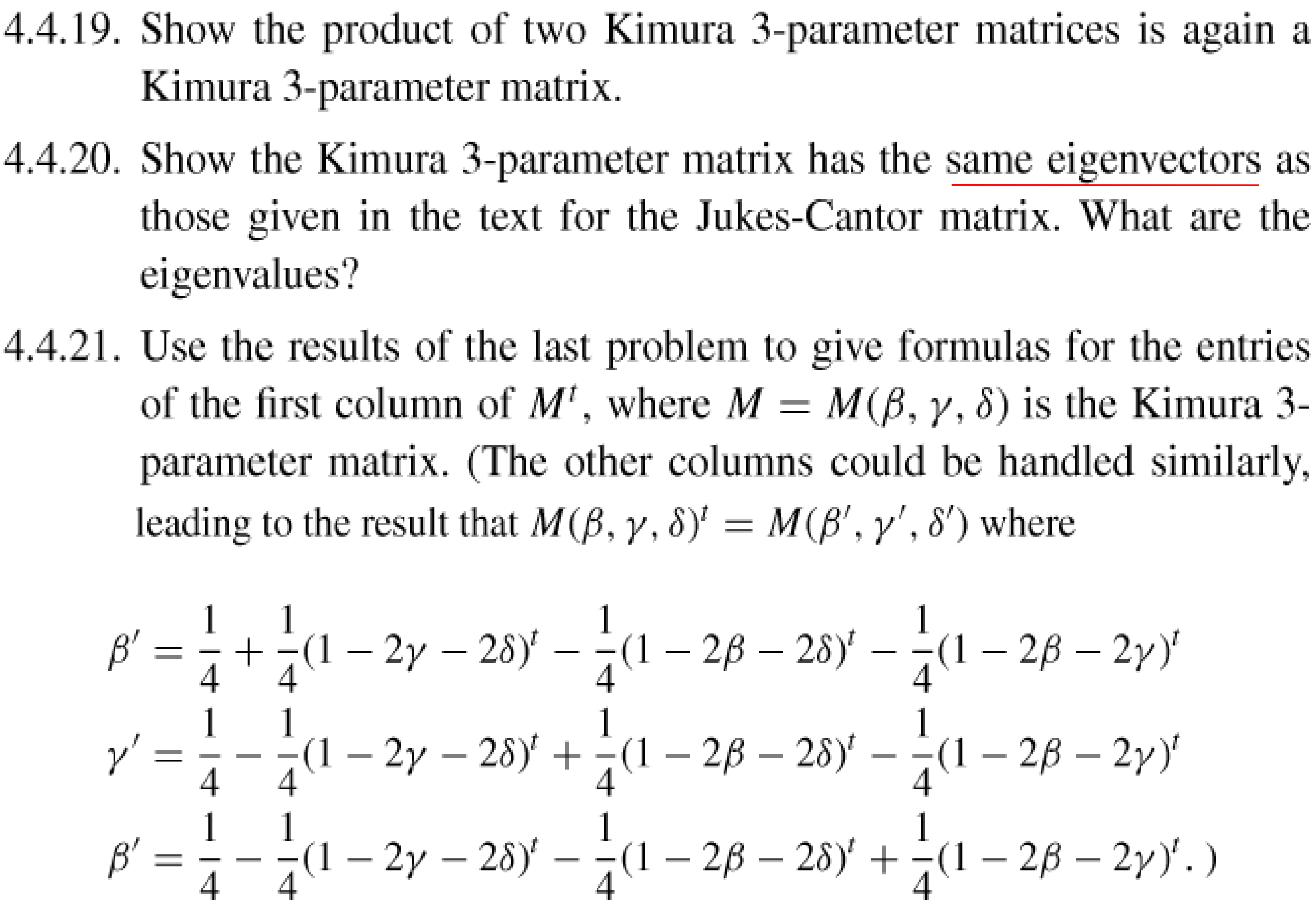
$$M = \begin{pmatrix} * & \beta & \gamma & \gamma \\ \beta & * & \gamma & \gamma \\ \gamma & \gamma & * & \beta \\ \gamma & \gamma & \beta & * \end{pmatrix}$$

K3P

$$M = \begin{pmatrix} * & \beta & \gamma & \delta \\ \beta & * & \delta & \gamma \\ \gamma & \delta & * & \beta \\ \delta & \gamma & \beta & * \end{pmatrix}.$$

What is the arrangement of bases in the matrix?

Home-work



4.4.19. Show the product of two Kimura 3-parameter matrices is again a Kimura 3-parameter matrix.

4.4.20. Show the Kimura 3-parameter matrix has the same eigenvectors as those given in the text for the Jukes-Cantor matrix. What are the eigenvalues?

4.4.21. Use the results of the last problem to give formulas for the entries of the first column of M^t, where $M = M(\beta, \gamma, \delta)$ is the Kimura 3-parameter matrix. (The other columns could be handled similarly, leading to the result that $M(\beta, \gamma, \delta)^t = M(\beta', \gamma', \delta')$ where

$$\beta' = \frac{1}{4} + \frac{1}{4}(1 - 2\gamma - 2\delta)^t - \frac{1}{4}(1 - 2\beta - 2\delta)^t - \frac{1}{4}(1 - 2\beta - 2\gamma)^t$$

$$\gamma' = \frac{1}{4} - \frac{1}{4}(1 - 2\gamma - 2\delta)^t + \frac{1}{4}(1 - 2\beta - 2\delta)^t - \frac{1}{4}(1 - 2\beta - 2\gamma)^t$$

$$\beta' = \frac{1}{4} - \frac{1}{4}(1 - 2\gamma - 2\delta)^t - \frac{1}{4}(1 - 2\beta - 2\delta)^t + \frac{1}{4}(1 - 2\beta - 2\gamma)^t.\,)$$

How to identify the amount of hidden mutation over time in JC model?

$$M = M(\alpha) = \begin{pmatrix} 1-\alpha & \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & 1-\alpha & \frac{\alpha}{3} & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & 1-\alpha & \frac{\alpha}{3} \\ \frac{\alpha}{3} & \frac{\alpha}{3} & \frac{\alpha}{3} & 1-\alpha \end{pmatrix}$$

At time t=t, the transition matrix M takes the form (look eqn (4.6) from the book)

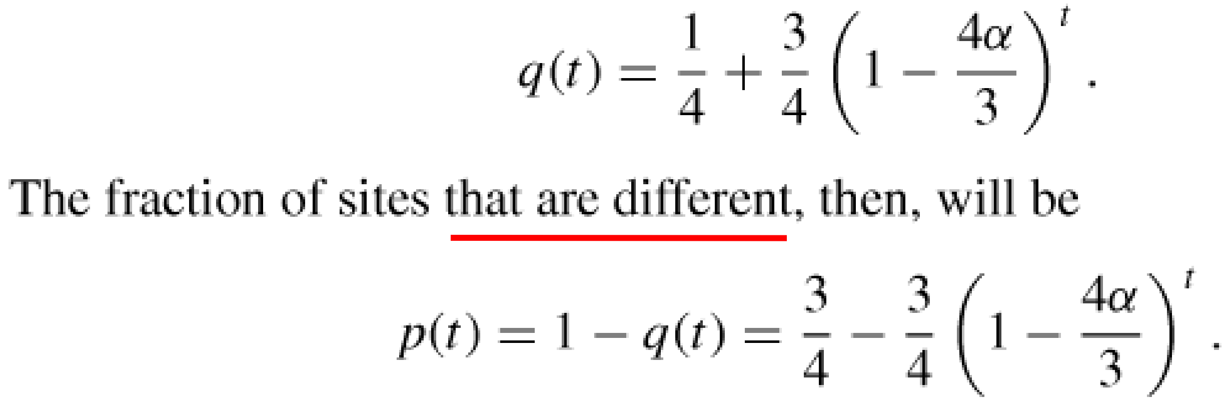
$$M^t = \begin{pmatrix} \frac{1}{4}+\frac{3}{4}\left(1-\frac{4}{3}\alpha\right)^t & \frac{1}{4}-\frac{1}{4}\left(1-\frac{4}{3}\alpha\right)^t & \frac{1}{4}-\frac{1}{4}\left(1-\frac{4}{3}\alpha\right)^t & \frac{1}{4}-\frac{1}{4}\left(1-\frac{4}{3}\alpha\right)^t \\ \frac{1}{4}-\frac{1}{4}\left(1-\frac{4}{3}\alpha\right)^t & \frac{1}{4}+\frac{3}{4}\left(1-\frac{4}{3}\alpha\right)^t & \frac{1}{4}-\frac{1}{4}\left(1-\frac{4}{3}\alpha\right)^t & \frac{1}{4}-\frac{1}{4}\left(1-\frac{4}{3}\alpha\right)^t \\ \frac{1}{4}-\frac{1}{4}\left(1-\frac{4}{3}\alpha\right)^t & \frac{1}{4}-\frac{1}{4}\left(1-\frac{4}{3}\alpha\right)^t & \frac{1}{4}+\frac{3}{4}\left(1-\frac{4}{3}\alpha\right)^t & \frac{1}{4}-\frac{1}{4}\left(1-\frac{4}{3}\alpha\right)^t \\ \frac{1}{4}-\frac{1}{4}\left(1-\frac{4}{3}\alpha\right)^t & \frac{1}{4}-\frac{1}{4}\left(1-\frac{4}{3}\alpha\right)^t & \frac{1}{4}-\frac{1}{4}\left(1-\frac{4}{3}\alpha\right)^t & \frac{1}{4}+\frac{3}{4}\left(1-\frac{4}{3}\alpha\right)^t \end{pmatrix}$$

The diagonal entries are

$$\frac{1}{4} + \frac{3}{4}\left(1 - \frac{4}{3}\alpha\right)^t$$

Diagonal entries of M^t give conditional probabilities that the base at time t is the same as the base at time 0.

It indicates that the fraction of initial base sequence at t=0 agreed at time t=t.



$$q(t) = \frac{1}{4} + \frac{3}{4}\left(1 - \frac{4\alpha}{3}\right)^t.$$

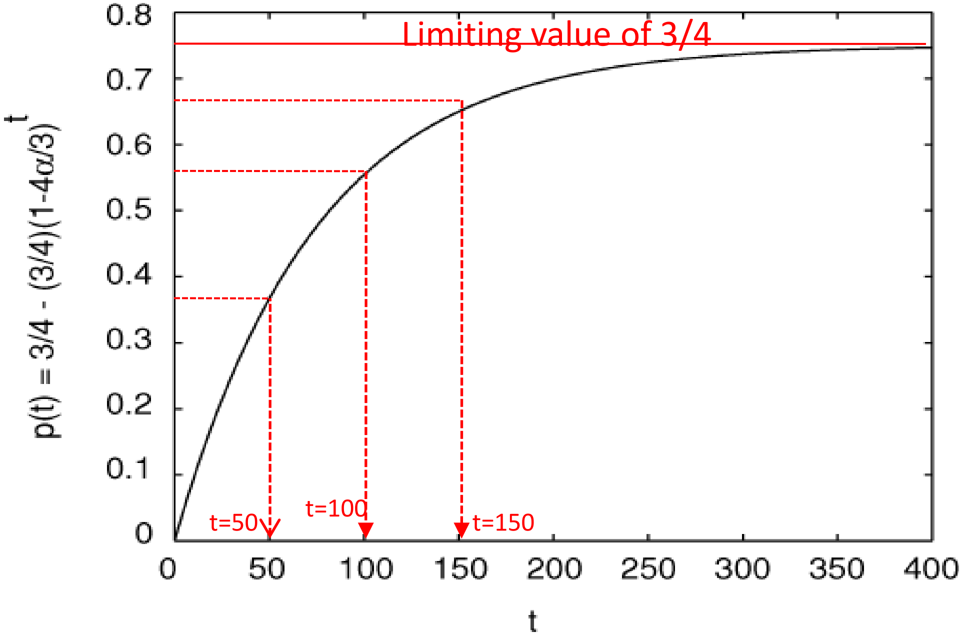
The fraction of sites that are different, then, will be

$$p(t) = 1 - q(t) = \frac{3}{4} - \frac{3}{4}\left(1 - \frac{4\alpha}{3}\right)^t.$$

Plot the graph and see what information can be got

Graph of P(t): Two important information

This part is got easily from the data



This part is **not** got easily from the data

Fraction of sites that differ from their original base gradually increases with t , approaching the value 3/4 , and P(t) never exceeds 3/4.

For each time t , p(t) has a different value. This means that given any value $0 \leq p \leq 3/4$, we should be able to find a t with p(t) = p. i.e., from the proportion of sites that differ between two sequences, time elapsed is extracted.

The Jukes-Cantor distance

How to determine mutation rate α or the number of elapsed time steps t?

With p = p(t) estimated, we get t; i.e.,

$$p = \frac{3}{4} - \frac{3}{4}\left(1 - \frac{4}{3}\alpha\right)^t$$

Subjective →

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

However, choice of a step size for time in the model affects both the value of the mutation rate α, and the number of elapsed time steps between ancestor and descendent.

So the product of elapsed time and mutation rate is taken.

d: Substitution/site during elapsed time

$$\begin{aligned} d &= t\alpha \\ &= \text{(no. of time steps)(mutation rate)} \\ &= \text{(no. of time steps)(no. of substitutions per site/time step)} \\ &= \text{(expected no. of substitutions per site during the elapsed time).} \end{aligned}$$

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

$$\ln\left(1 - \frac{4}{3}\alpha\right) \approx -\frac{4}{3}\alpha.$$

mutation rate per time step, α, is very small.

$$\begin{aligned} t &\approx \frac{\ln\left(1 - \frac{4}{3}p\right)}{-\frac{4}{3}\alpha} \\ &\approx -\frac{3}{4\alpha}\ln\left(1 - \frac{4}{3}p\right), \end{aligned}$$

$$d = t\alpha \approx -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right).$$

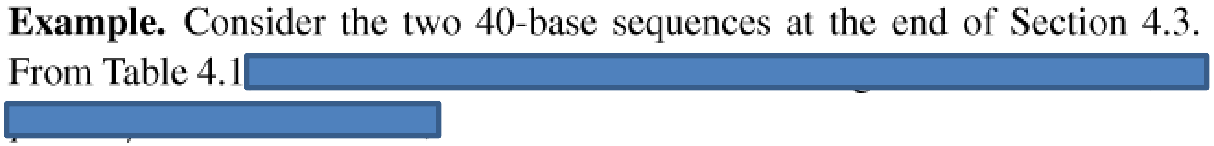
(OR)

$$\boxed{d_{JC}(S_0, S_1) = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right)}$$

Jukes-Cantor distance between DNA sequences S0 and S1

p is the fraction of sites that disagree in comparing S0 with S1.

Example



Example. Consider the two 40-base sequences at the end of Section 4.3. From Table 4.1 [redacted]

Determine JC distance and the hidden mutations.

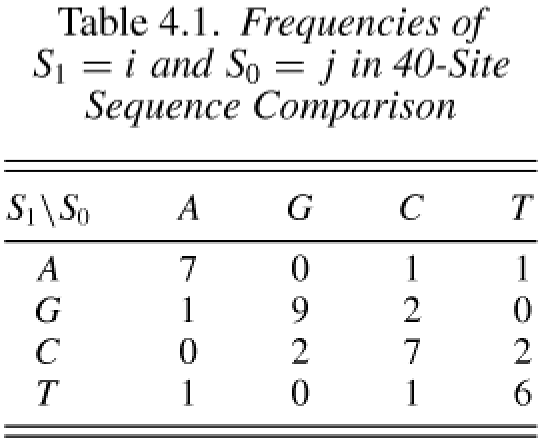


Table 4.1. *Frequencies of $S_1 = i$ and $S_0 = j$ in 40-Site Sequence Comparison*

$S_1 \backslash S_0$	A	G	C	T
A	7	0	1	1
G	1	9	2	0
C	0	2	7	2
T	1	0	1	6

Example

Example. Consider the two 40-base sequences at the end of Section 4.3. From Table 4.1, we find that 11 of the sites have undergone a substitution, so $p = 11/40 = \underline{.2750}$. Thus,

$$d_{JC}(S_0, S_1) = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right)$$

$$d_{JC}(S_0, S_1) = -\frac{3}{4}\ln\left(1 - \frac{4}{3}\frac{11}{40}\right) \approx \underline{.3426}.$$

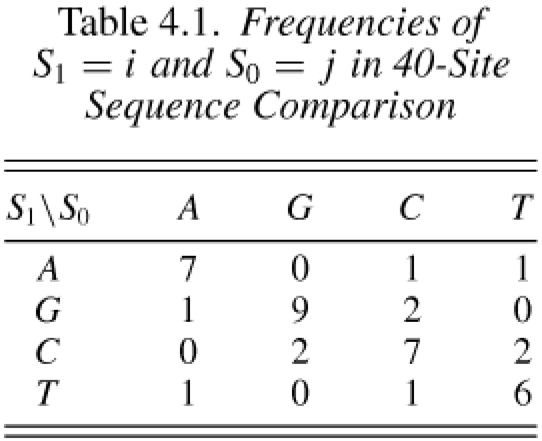


Table 4.1. *Frequencies of $S_1 = i$ and $S_0 = j$ in 40-Site Sequence Comparison*

$S_1 \backslash S_0$	A	G	C	T
A	7	0	1	1
G	1	9	2	0
C	0	2	7	2
T	1	0	1	6

Observed = 0.2750 substitutions per site on average,
Estimate = 0 .3426 substitutions per site occurred over evolution.

The difference is due to **Hidden mutations**.

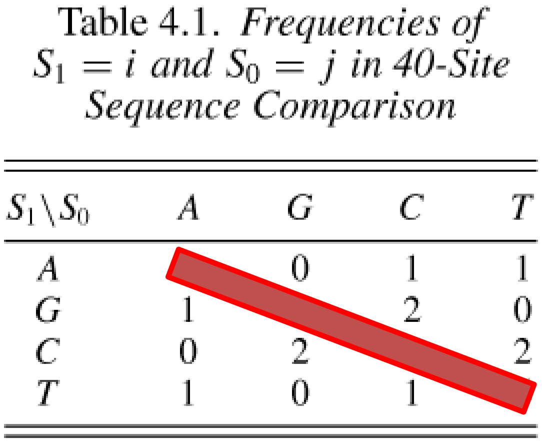


Table 4.1. *Frequencies of $S_1 = i$ and $S_0 = j$ in 40-Site Sequence Comparison*

$S_1 \backslash S_0$	A	G	C	T
A		0	1	1
G	1		2	0
C	0	2		2
T	1	0	1	

Do not take diagonal element for this problem.

Derive Kimura distances for 2 (K2P)and 3 (K3P) parameter models

K2P

$$M = \begin{pmatrix} * & \beta & \gamma & \gamma \\ \beta & * & \gamma & \gamma \\ \gamma & \gamma & * & \beta \\ \gamma & \gamma & \beta & * \end{pmatrix}$$

$$\boxed{d_{K2} = -\frac{1}{2}\ln(1 - 2p_1 - p_2) - \frac{1}{4}\ln(1 - 2p_2).}$$

γ = δ

Goes from 3 to 2 parameter model

K3P

β is the probability of a transition

&

γ + δ = 2γ is the probability of a transversion.

$$M = \begin{pmatrix} * & \beta & \gamma & \delta \\ \beta & * & \delta & \gamma \\ \gamma & \delta & * & \beta \\ \delta & \gamma & \beta & * \end{pmatrix}.$$

$$\boxed{d_{K3} = -\frac{1}{4}(\ln(1 - 2\beta - 2\gamma) + \ln(1 - 2\beta - 2\delta) + \ln(1 - 2\gamma - 2\delta))}$$

Additive and symmetric distances: Log-det. A general distance formula for Markov base substitution model

- The distance formulas assume that data are consistent with either the JC, K2P or K3P models.

- However, it may not be consistent as models are not elaborate or do not describe sequence properties carefully.

- So, distance formula for the general Markov model is necessary.

Multiplication of JC matrices

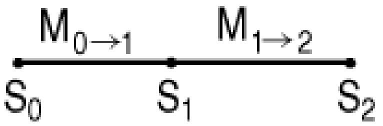


Figure 4.2. Three sequences in evolutionary order.

Let $M_{0\rightarrow1}$ = M(α1) and $M_{1\rightarrow2}$ = M(α2) be two Jukes-Cantor matrices describing the two mutation.

Then, we can calculate a mutation matrix $M_{0\rightarrow2}$ for the full passage from S_0 to S_2 as the product

$$M_{0\to2} = M_{1\to2} M_{0\to1}.$$

$$M_{0\to2} = M(\alpha_3), \text{ with } \alpha_3 = \alpha_1 + \alpha_2 - \frac{4}{3}\alpha_1\alpha_2.$$ We solved it in the earlier class.

Multiplication of JC matrices → Adding distances

$$M(\alpha_1): \quad -\frac{3}{4}\ln\left(1-\frac{4}{3}\alpha_1\right)$$

$$M(\alpha_2): \quad -\frac{3}{4}\ln\left(1-\frac{4}{3}\alpha_2\right)$$

$$M(\alpha_1)M(\alpha_2)=M(\alpha_3): \quad -\frac{3}{4}\ln\left(1-\frac{4}{3}(\alpha_1+\alpha_2-\frac{4}{3}\alpha_1\alpha_2)\right).$$

But a little algebra shows

Note: $\boxed{d_{JC}(S_0,S_1)=-\frac{3}{4}\ln\left(1-\frac{4}{3}p\right)}$

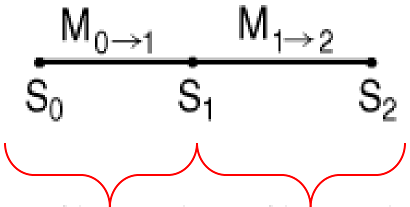
$$-\frac{3}{4}\ln\left(1-\frac{4(\alpha_1+\alpha_2-\frac{4}{3}\alpha_1\alpha_2)}{3}\right)=\left(-\frac{3}{4}\ln\left(1-\frac{4\alpha_1}{3}\right)\right)$$

$$+\left(-\frac{3}{4}\ln\left(1-\frac{4\alpha_2}{3}\right)\right)$$

Multiplying two Jukes-Cantor matrices corresponds to adding the associated distances.

Additive-symmetry properties

For the general Markov model, distance between sequences are defined as the additive property



$$d(S_0, S_2) = d(S_0, S_1) + d(S_1, S_2)$$

We want something like this !!

So we define log-det distance (also called the paralinear distance in this form) between S0 and S1, and it is defined by

$$d_{LD}(S_0, S_1) = -\frac{1}{4}\left(\ln\left(\det(F)\right) - \frac{1}{2}\ln(g_0 g_1)\right)$$

F is the 4 × 4 frequency array obtained by comparing sites in sequences S0 and S1.

g_i is the product of the 4 entries in f_i

"det" denotes the determinant of a matrix.

$$d(S_0, S_1) = d(S_1, S_0).$$

(Symmetry property)