Modeling Molecular Evolution

What are the basic processes of molecular evolution?

What is one looking for while comparing sequences?

When comparing two sequences one is looking for evidence that they have diverged from a common ancestor by a process of mutation or natural selection

Basic mutational processes are:

substitution – which change residues in a sequence

insertions and deletions – which add or remove residues, together referred as gaps or indels

Further apart the two sequences are from each other, the more frequent these changes are expected to occur.

Mutations potentially affect the function of the gene, which can either be **beneficial**, or lead to **reduction** in functionality & adaptability of the protein

Natural selection comes into play – allowing mutations that are either evolutionarily advantageous or, occur in non-functional regions of the sequence

Natural selection has the effect of screening the mutations – some changes are seen **more often** than others

Natural selection - fundamental mechanism through which evolution occurs, but for selection to be possible there must be some underlying variability in the genetic makeup within a species.

Since selection usually acts to reduce variability, there must also be a source of new genetic variation.

This is introduced at the molecular level in the DNA of individuals, through random changes as the molecules are copied into new generations.

Let’s try to develop mathematical models of DNA mutation processes, i.e., of molecular evolution using the language of probability to describe random mutations

- the concept of phylogenetic distance as a measure of sequence similarity will emerge from these probabilistic models.

When base substitutions occur in the evolution, the probability of a particular base appearing at a site in the descendent sequence might depend on the ancestral base.

e.g., if the ancestral base is T, then

- probability of seeing a T in the descendant sequence is higher – why?
- probability of seeing a C in the descendant sequence is lower than T – why?
- probability of seeing an A or G in the descendant sequence is lowest – why?

To formalize this we shall use the concept of conditional probability

When base substitutions occur in the evolution, the probability of a particular base appearing at a site in the descendent sequence **might** depend on the ancestral base.

e.g., if the ancestral base is T, then

- probability of seeing a T in the descendant sequence is higher – why?

- probability of seeing a C in the descendant sequence is lower than T – why?

- probability of seeing an A or G in the descendant sequence is lowest – why?

To formalize this we shall use the concept of **conditional probability**

Conditional Probability

If E and F are two events, then conditional probability of F given E is defined by

$$p(F \mid E) = \frac{p(F \cap E)}{p(E)}$$

The concept of conditional probability also clarifies the notion of independence of events.

Events E and F are independent if knowledge that one has occurred gives no information as to whether the other occurred, i.e.,

p(F|E) = p(F) and P(E|F) = p(E)

Conditional Probability

If E and F are two events, then conditional probability of F given E is defined by

$$p(F \mid E) = \frac{p(F \cap E)}{p(E)}$$

The concept of conditional probability also clarifies the notion of independence of events.

Events E and F are independent if knowledge that one has occurred gives no information as to whether the other occurred, i.e.,

p(F|E) = p(F) and P(E|F) = p(E)

Most common mutation introduced in copying sequences of DNA is a base substitution

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.

Which of these two substitutions is more common?

Conditional Probability

Ex: Taking into account the likelihood of transitions and transversions, which of the following is likely to be the smallest? Which is likely to be the largest?

(i) $p(S_1 = C| S_0 = C)$ “no change”,
(ii) $p(S_1 = T| S_0 = C)$, “transition”,
(iii) $p(S_1 = A| S_0 = C)$ “transversion”, and
(iv) $p(S_1 = G| S_0 = C)$ “transversion”.

What is the sum of the above four probabilities?

Other mutations observed include:

- **Deletion of a base or consecutive bases,**
- **Insertion of a base or consecutive bases,**
- **Inversion (reversal) of a section of the sequence**

- these mutations seen more rarely in natural populations.

- not surprising, since these mutations have a dramatic effect on the protein.

Ignore such possibilities to make our modeling task both clearer & mathematically tractable.

Focusing solely on base substitutions, a basic problem is how to deduce the amount of mutation during evolution:

S0: ACCTGCGCTA
S1: ACGTGCACTA
S2: ACGTGCGCTA

If G → A → C at the 7th position?

Comparing S0 & S2 - 1/10 mutations per site

Comparing S0, S1 & S2 - 3/10 mutations per site from S0 to S2

– a simple ratio of mutations per site obtained from comparing 1st & 3rd sequences gives a lower estimate of the mutation that actually occurred.

Assuming that mutations are rare,

- ignore the probability of a hidden mutations having occurred (G → A → G)

we can reconstruct a mathematical model for the no. of mutations that are likely to have occurred from those observed in comparing only the initial and final DNA sequences.

Matrix Models
of
Base Substitution

Matrix Models of Base Substitution

Model ancestral sequence probabilistically:

Assume each site in the sequence is one of the 4 bases chosen randomly with probabilities p_A, p_G, p_C, p_T,

- these probabilities describe the ancestral base distribution in a vector as

$$\mathbf{p}_0 = (p_A, p_G, p_C, p_T), \qquad p_A + p_G + p_C + p_T = ?$$

Is the assumption that all bases in the sequence chosen “at random” reasonable?

Would it matter if the DNA sequence was coding or noncoding?

Matrix Models of Base Substitution

Model the mutation process over one-time step, assuming that only base <u>substitutions</u> can occur.

⇒ 16 conditional probabilities of observing a base substitution, $p(S_1 = i|S_0 = j)$ for i, j = A, G, C, and T:

Descendent base

$$M = \begin{pmatrix} p_{A|A} & p_{A|G} & p_{A|C} & p_{A|T} \\ p_{G|A} & p_{G|G} & p_{G|C} & p_{G|T} \\ p_{C|A} & p_{C|G} & p_{C|C} & p_{C|T} \\ p_{T|A} & p_{T|G} & p_{T|C} & p_{T|T} \end{pmatrix}$$

Ancestral base

Assuming only base substitutions – is it reasonable for coding regions of DNA?

Column sum ?

Row sum?

Example

For the 40-base ancestral sequence, S0:

ACTTGTCGGATGATCAGCGGTCCATGCACCTGACAACGGT

and its descendent sequence, S1:

ACATGTTGCTTGACGACAGGTCCATGCGCCTGAGAACGGC

$$p_0 = (p_A, p_G, p_C, p_T) = (.225, .275, .275, .225)$$

$$M = \begin{pmatrix} .778 & 0 & .091 & .111 \\ .111 & .818 & .182 & 0 \\ 0 & .182 & .636 & .222 \\ .111 & 0 & .091 & .667 \end{pmatrix}$$

No. of A’s in S_0 = 9, probability of A in S_0 = 9/40 = 0.225, ...

First entry in M is $p_{A|A}$ = 7/9 = 0.778, …

Matrix Models of Base Substitution

Multiplying:

$$M\mathbf{p}_0 = \begin{pmatrix} p_{A|A} & p_{A|G} & p_{A|C} & p_{A|T} \\ p_{G|A} & p_{G|G} & p_{G|C} & p_{G|T} \\ p_{C|A} & p_{C|G} & p_{C|C} & p_{C|T} \\ p_{T|A} & p_{T|G} & p_{T|C} & p_{T|T} \end{pmatrix} \begin{pmatrix} p_A \\ p_G \\ p_C \\ p_T \end{pmatrix}$$

$$= \begin{pmatrix} p_{A|A}\, p_A + p_{A|G}\, p_G + p_{A|C}\, p_C + p_{A|T}\, p_T \\ p_{G|A}\, p_A + p_{G|G}\, p_G + p_{G|C}\, p_C + p_{G|T}\, p_T \\ p_{C|A}\, p_A + p_{C|G}\, p_G + p_{C|C}\, p_C + p_{C|T}\, p_T \\ p_{T|A}\, p_A + p_{T|G}\, p_G + p_{T|C}\, p_C + p_{T|T}\, p_T \end{pmatrix} = \begin{pmatrix} p(S_1 = A) \\ p(S_1 = G) \\ p(S_1 = C) \\ p(S_1 = T) \end{pmatrix}$$

we find that $M\,\mathbf{p}_0 = \mathbf{p}_1$

$\mathbf{p}_1$ - the vector of base probabilities in sequence S_1

Matrix Models of Base Substitution

M - is a **transition matrix**, gives how probabilities of each base in ancestral seq. S_0 are transformed into probabilities of each base in the descendent seq. S_1 one-time step later.

What would be the meaning of *M*p1?

$$p_1 = Mp_0 = \begin{pmatrix} .225 \\ .275 \\ .300 \\ .200 \end{pmatrix}, \quad p_2 = Mp_1 = \begin{pmatrix} .222 \\ .274 \\ .320 \\ .183 \end{pmatrix}$$

What is the sum of the entries in p1? In p2?

Why must this be the case? Why use same *M*?

Matrix Models of Base Substitution

To make sense biologically, we must assume that the probabilistic mutation process over the first time step is identical to that over the next time step.

Using the same transition matrix M of conditional probabilities means each type of base substitution has the same likelihood of occurring as it did before.

- reasonable assumption for small time intervals.

Matrix Models of Base Substitution

Furthermore, what happens during the second step depends only on:

- what the base was at time t = 1 (the information in p1), and
- the conditional probabilities (the information in M)

Whether that site experienced a substitution during the previous time step is irrelevant.

This is an example of a **Markov model**.

Markov Models

- **In a <u>Markov</u> model, a system is described in one of *n* different states, and may switch from one state to another with time.**
- **In our DNA substitution model, the system is a site in a DNA sequence, which is initially in one of the 4 states (A, G, C, or T) according to the base that occupies it.**
- **Initial probabilities that the system is in one of the states is given by a vector, p_0.**
- **Conditional probabilities of the switch from every state to every other state over one-time step is given by a 4 x 4 transition matrix, *M*.**

Markov Models

An important assumption is made in any Markov model:

What happens to the system over a given time step depends only on the state the system is in at the start of that step and the transition probabilities.

- there is “no memory” of what changes might have occurred during earlier time steps, i.e., conditional probabilities are independent of the past history.

Can a Markov model be used to identify spatial patterns in a DNA sequence, e.g. CpG islands, protein-coding regions?

Markov Models

Q. For a DNA substitution model, is it reasonable to assume this independence?

In our DNA model we also assumed that each site in the sequence behaves identically and independently of every other site to find various probabilities from sequence data, by considering each site as an independent trial of the same probabilistic process.

Q. How reasonable is this assumption?

Markov Models

This assumption may not be a very reasonable one for gene sequences:

- Genetic code allows for many changes in the third site of each codon to have no effect on the gene product, as a consequence, substitutions in the third sites might be more likely than in the first two, violating the assumption that each site behaves identically
- Since genes lead to the production of proteins, the likelihood of change at one site may well be tied to changes at another, violating the assumption of independence.

Markov Models

Can we find ways to go around these assumptions?

- allowing for different conditional probabilities for various sites.
- be careful to take assumptions into account when using the tools we develop on real data

- for instance, we might ignore the third base of each codon in estimating information from our data, so that it is more reasonable to treat sites as independent and following identical processes.

Markov Models

Markov matrix has all entries ≥ 0 and its columns sum to 1.

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

There will be only one eigenvector associated with $\lambda_1 = 1$.

Theorem-2: A Markov matrix, all of whose entries are positive (i.e., nonzero), always has 1 as a strictly dominant eigenvalue.

Markov Models

We will now discuss a few special Markov models of base substitutions:

- Jukes-Cantor Model
- Kimura Models

Jukes-Cantor model

- the simplest Markov model of base substitution.

Additional assumptions to the basic Markov model:

First, all bases occur with equal probability in the ancestral sequence, i.e.,

$$p_0 = (\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4})$$

Second, all the 16 conditional probabilities of base substitutions are same, i.e., all possible substitutions are equally likely:

A ↔ G, C ↔ T, A ↔ C,
A ↔ T, C ↔ G, T ↔ G

i.e., it assumes transitions & transversions occur at the same rate.

Jukes-Cantor model

If we define α/3 as the conditional probability of a base substitution of any type:

$$p(S_1 = i|S_0 = j) = \alpha/3, \quad \text{for all } i, j$$

i.e., the 12 off-diagonal entries of the matrix *M* will be α/3.

Since the entries in any column of *M* add to 1, what would be the diagonal entries?

Jukes-Cantor model

Transition matrix for Jukes-Cantor model:

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

Value of α depends on the time step we use and features of the particular DNA sequence being modeled.

Jukes-Cantor model

Although α is a probability, we can interpret it as a rate:

- it is the rate at which observable base substitutions occur over one time step and is measured in units of

$$\alpha = \text{(substitutions per site) / time step}$$

Mutational rates α for DNA in real organisms is not easily found.

Jukes-Cantor model

Estimates of α:

- 1.1 x 10^{-9} mutations per site per year for certain sections of chloroplast DNA of maize & barley
- 10^{-8} mutations per site per year for mtDNA in mammals
- 0.01 mutations per site per year for influenza A virus
- Rate of mutation is generally a bit lower in coding regions than in noncoding DNA

After 1 million years, compute the amount of mutation in the descendant sequence if $\alpha = 10^{-8}$ & length of sequence is 1000 bases.

Jukes-Cantor model

In the development of our model, we shall treat α as an unknown constant.

In reality, the mutation rate may not be constant; it may change with time, or with location within the DNA.

For shorter periods of time and for DNA serving a fixed purpose, the assumption of a constant mutational rate is reasonable.

When mutation rates are constant, there is said to be a molecular clock operating.

Jukes-Cantor model

Ex-1: For the Jukes-Cantor model, in what proportion of the sites will each base appear after one time-step?

Ex-2: What proportion of the sites will have a base A in the ancestral sequence and a T in the descendent one time-step later:

$$p(S_0 = A \text{ and } S_1 = T)?$$

Ex-3: 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: $p(S_{100} = T \mid S_0 = A)$?

Jukes-Cantor model

Ex-1: For the Jukes-Cantor model, in what proportion of the sites will each base appear after one time-step?

$$p_1 = M\,p_0 = \begin{pmatrix} 1-\alpha & \alpha/3 & \alpha/3 & \alpha/3 \\ \alpha/3 & 1-\alpha & \alpha/3 & \alpha/3 \\ \alpha/3 & \alpha/3 & 1-\alpha & \alpha/3 \\ \alpha/3 & \alpha/3 & \alpha/3 & 1-\alpha \end{pmatrix} \begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix} = \begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix}$$

Base composition of the sequence does not change under Jukes-Cantor model

(¼, ¼, ¼, ¼) is an equilibrium base distribution for sequences under the Jukes-Cantor model

Jukes-Cantor model

Ex-2: What proportion of sites will have a base A in ancestral sequence and a T in descendent one time-step later, $p(S_0 = A \text{ and } S_1 = T)$?

$$p(S_0 = A \text{ and } S_1 = T)$$

$$= p(S_1 = T \mid S_0 = A)\, p(S_0 = A)$$

$$= (\alpha/3)(1/4) = \alpha\,/12$$

Jukes-Cantor model

Ex-3: What is the probability that a base A in the ancestral sequence will have mutated to become a base T in descendent sequence 100 time-steps later, i.e.,

compute the probability $p(S_{100} = T \mid S_0 = A)$?

$$\boldsymbol{p}_{100} = \boldsymbol{M}^{100}\, \boldsymbol{p}_0$$

What is the (4,1) entry of $\boldsymbol{M}^{100}$?

Jukes-Cantor model

Generalizing to any *t*, let's find all entries of M^t – using the eigenvectors approach.

Why do we need to compute eigenvalues and eigenvectors?

How do we compute the eigenvectors & eigenvalues of a matrix?

Jukes-Cantor model

Theorem: If A is an n x n matrix, v a non-zero vector, and λ a scalar such that $Av = \lambda v$, then v is an eigenvector of A with eigenvalue λ.

Equilibrium base distribution is one eigenvector with eigenvalue λ = 1, there are 3 more that can be found by trial and error or a long computation

$$p_1 = M\,p_0 = \begin{pmatrix} 1-\alpha & \alpha/3 & \alpha/3 & \alpha/3 \\ \alpha/3 & 1-\alpha & \alpha/3 & \alpha/3 \\ \alpha/3 & \alpha/3 & 1-\alpha & \alpha/3 \\ \alpha/3 & \alpha/3 & \alpha/3 & 1-\alpha \end{pmatrix} \begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix} = \begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix}$$

Jukes-Cantor model

Eigenvectors & eigenvalues of Jukes-Cantor matrix are:

$\mathbf{v}_1 = (1,1,1,1)$ $\lambda_1 = 1$

$\mathbf{v}_2 = (1,1,-1,-1)$ $\lambda_2 = 1 - 4/3\ \alpha$

$\mathbf{v}_3 = (1,-1,1,-1)$ $\lambda_3 = 1 - 4/3\ \alpha$

$\mathbf{v}_4 = (1,-1,-1,1)$ $\lambda_4 = 1 - 4/3\ \alpha$

Check by multiplying Mv_i for each i.

Jukes-Cantor model

Theorem: If v is an eigenvector of A with eigenvalue λ, then for any scalar c, cv is also an eigenvector of A with the same eigenvalue λ, i.e.,

If $Av = \lambda v$, then

$$A(cv) = cAv = c \lambda v = \lambda(cv)$$

Theorem: Let A be a *n* x *n* matrix with *n* eigenvectors v_1, v_2, ..., v_n, whose corresponding eigenvalues are λ_1, λ_2, ..., λ_n.

Let these *n* eigenvectors form columns of the matrix S.

Jukes-Cantor model

If S has an inverse, then any vector can be written as a sum of column eigenvectors. Expressing initial eigenvector as

$$x_0 = c_1v_1 + c_2v_2 + \ldots + c_nv_n$$

Then, $x_1 = A\,x_0 = A(c_1v_1 + c_2v_2 + \ldots + c_nv_n)$

$$= c_1Av_1 + c_2Av_2 + \ldots + c_nAv_n$$

$$= c_1\lambda_1v_1 + c_2\lambda_2v_2 + \ldots + c_n\lambda_nv_n$$

$$x_2 = Ax_1 = A(c_1\lambda_1v_1 + c_2\lambda_2v_2 + \ldots + c_n\lambda_nv_n)$$

$$= c_1\lambda_1Av_1 + c_2\,\lambda_2Av_2 + \ldots + c_n\lambda_nAv_n$$

$$= c_1(\lambda_1)^2v_1 + c_2(\lambda_2)^2v_2 + \ldots + c_n(\lambda_n)^2v_n$$

And so on, we obtain

$$x_t = c_1(\lambda_1)^tv_1 + c_2(\lambda_2)^tv_2 + \ldots + c_n(\lambda_n)^tv_n$$

Jukes-Cantor model

To find all entries of M^t: Let's first focus on the first column of M^t, which can be isolated by taking the product

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

Expressing (1,0,0,0) in terms of the eigenvectors:

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

Jukes-Cantor model

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

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

Substituting in the vectors **v**$_i$,

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

v1 = (1,1,1,1)	λ1 = 1
v2 = (1,1,-1,-1)	λ2 = 1 – 4/3 α
v3 = (1,-1,1,-1)	λ3 = 1 – 4/3 α
v4 = (1,-1,-1,1)	λ4 = 1 – 4/3 α

Jukes-Cantor model

Other columns of M^t are found similarly:

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

Note: all diagonal entries (prob. of a base remaining unchanged) are identical, also all non-diagonal entries (prob. of a base undergoing substitution) are identical.

Jukes-Cantor model

This formula for M^t is of the Jukes-Cantor form itself, with the Jukes-Cantor parameter being

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

Ex-3: Can we now answer the question of the probability that a base A in ancestral sequence will have mutated to become a base T in the descendent sequence 100 time-steps later?

This is the (4,1) entry of M^{100} which is

$$\frac{1}{4}-\frac{1}{4}\left(1-\frac{4}{3}\alpha\right)^{100}$$

Jukes-Cantor Model

Jukes-Cantor model is a one-parameter model of mutation,

- it depends on the single parameter α to specify the mutation.

Other models use several different parameters to specify mutation rates for several different types of mutations, e.g., Kimura 2-parameter and Kimura 3-parameter models

The Kimura Models

Kimura 2-parameter model allows for different rates for transitions (β) and transversions (γ).

If we assume these rates are independent of initial base, then off-diagonal entries of the transition matrix are given by:

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

Since the columns sum to 1, this means all the diagonal entries must be 1 - β - 2γ.

The Kimura Models

Kimura 3-parameter model assumes a transition matrix of the form

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

The equilibrium base distribution vector

$$\mathbf{p_0} = (\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4})$$

is an eigenvector with eigenvalue 1 for both Kimura 2- and 3- parameter models, i.e., sequences evolving according to these models have uniform base distribution at all times.

Phylogenetic Distances

Phylogenetic Distances

With a model of DNA mutations, we can better understand how to relate the amount of mutation that we observe in comparing an ancestral sequence and descendent sequence to the amount of mutation that must have actually occurred.

That is, we will be able to uncover the amount of hidden mutation that was obscured by subsequent mutations at the same site.

Phylogenetic Distances

Considering Jukes-Cantor model of sequence mutation:

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

Compute the entries of $\mathbf{M}^t$ for *t* = 0, 1, 2, 3, …

Diagonal entries of $\mathbf{M}^t$ – probability of observing no change at a site are

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

Phylogenetic Distances

Fraction of sites that agreed with their initial base are given by

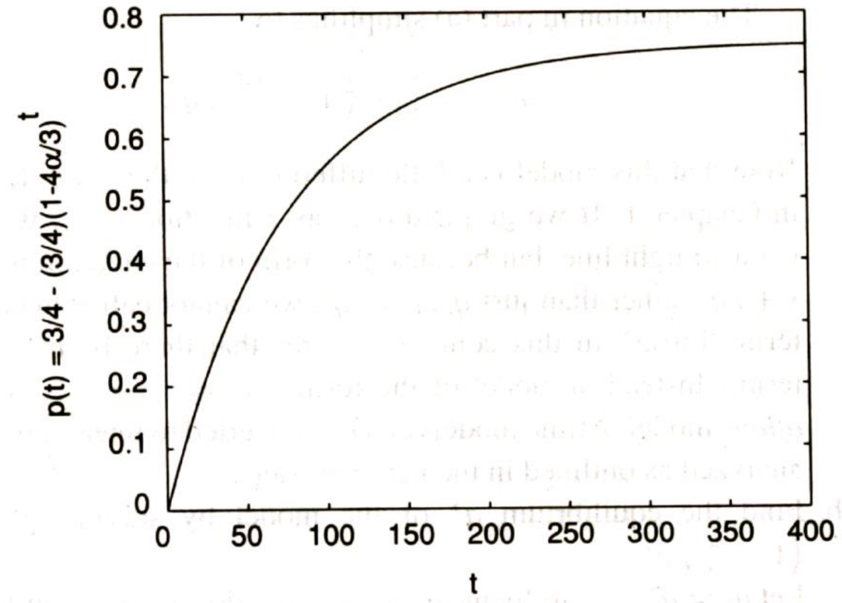
$$q(t) = \frac{1}{4} + \frac{3}{4}\left(1 - \frac{4}{3}\alpha\right)^{t}$$

Fraction of sites that are different will be

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

Phylogenetic Distances

Fraction of sites that differ from original base gradually increases with *t*, approaching the value ¾, and never exceeds ¾. Why?



Jukes Cantor model with alpha = 0.01

Q. Even if so much mutation has occurred that the two sequences appear to be completely unrelated, you would expect to find agreement at 1/4 of the sites. Why?

Jukes-Cantor Distance

For each time *t*, *p*(*t*) has a different value, i.e., given any value $0 \leq p \leq 3/4$, we can find a *t* with *p*(*t*) = *p*, corresponding to the proportion of sites that differ between two sequences

$\Rightarrow$ We should be able to recover the number of elapsed time steps (assuming we know α)

For real sequence data, *p* is easily estimated, although the elapsed time *t* and the mutation rate α usually are not known.

Recovering them from data is our goal.

Jukes-Cantor Distance

Suppose we have records of an original DNA sequence and a mutated version of it at a later time, but do not know either the mutation rate α nor the number of elapsed time steps *t*.

- we can estimate p = p(t) by comparing the two sequences and using the proportion of sites that disagree in the two sequences as an estimate.

- if the original & mutated sequences are ATTGAC and ATGGCC, our estimate is

$$p(t) = 2/6 = 0.333$$

Jukes-Cantor Distance

With p = p(t) estimated, how do we recover information on the mutation rate α and the amount of elapsed time t?

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

Solving for *t*, $$t = \frac{\ln(1 - 4/3\,p)}{\ln(1 - 4/3\,\alpha)}$$

Note: Choice of a step size for time in formulating our model affects both the value of mutation rate α, and the elapsed time between ancestor and descendent.

We cannot really expect to recover both of these.

Jukes-Cantor Distance

Product of the two does have a meaning which is more intrinsic to what we are modeling:

d = tα **= (no. of time steps)(mutation rate)**

= (no. of time steps)(no. of substitutions per site/time step)

= (expected no. of substitutions per site during the elapsed time)

$$\ln(1-\frac{4}{3}\alpha) \approx -\frac{4}{3}\alpha \qquad t \approx \frac{\ln(1-4/3p)}{-4/3\alpha} \approx -\frac{3}{4\alpha}\ln\left(1-\frac{4}{3}p\right)$$

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

Using the approximation ln(1+x) ~ x, for small x

Jukes-Cantor Distance

Jukes-Cantor distance between DNA sequences S_0 & S_1 is defined as

$$d_{\mathrm{JC}}(\mathrm{S}_0, \mathrm{S}_1) = t\alpha \approx -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right)$$

p - fraction of sites that disagree in comparing S_0 & S_1

Provided that Jukes-Cantor model accurately describes the evolution of one sequence into another, it is an estimate of the total number of substitutions per site that occurred during the evolution

"Distance" here is an abstract notion of how different two sequences are because of mutations

Jukes-Cantor Distance

If molecular clock hypothesis holds, distance computed is proportional to the amount of elapsed time; the constant of proportionality being the mutation rate

$\Rightarrow$ the distance can be thought of as a measure of how much time was required for one sequence to mutate into the other.

If molecular clock hypothesis does not hold, it is still a reconstruction of the average number of substitutions that occurred at any one site.

- the larger it is, greater the evolutionary change

Jukes-Cantor Distance

If there is some other data (such as geological record) suggesting the time evolved, then the mutation rate can be found from d_{JC}.

- this is one way that real DNA mutation rates are estimated.

For e.g., if t = 10Myrs by some geological records, then

$d = t\alpha = 10^7 \times \alpha = 0.33$ (from d_{JC})

$\Rightarrow \qquad \alpha = 0.33 \times 10^{-7}$

Jukes-Cantor Distance

Ex: If between two 40-base sequence 11 sites have undergone a substitution, then p = 11/40 = 0.275

$$d_{\mathrm{JC}}(\mathrm{S}_0,\mathrm{S}_1) = -\frac{3}{4}\ln\left(1-\frac{4}{3}\frac{11}{40}\right) \approx .3426$$

while we observe .275 substitutions per site, we estimate that in the course of evolution 0.3426 substitutions per site occurred

- Hidden mutations account for the difference.

The Kimura distances

For Kimura 3-parameter model,

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

If $\gamma = \delta$, this expression gives the distance for the Kimura 2-parameter model, with β being the probability of transition and $\gamma + \delta = 2\gamma$, the probability of transversion.

If from sequence data we estimate probability of transition as p_1 and transversion as p_2

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

References

- **Mathematical Models in Biology: An Introduction, E.S. Allman and J.A. Rhodes**
- **Bioinformatics Sequence & Genome Analysis, David W. Mount**
- **Biological Sequence Analysis, Probabilistic Models of Proteins and Nucleic Acids, R. Durbin, S.R. Eddy, A. Keoghs and G. Mitchison**