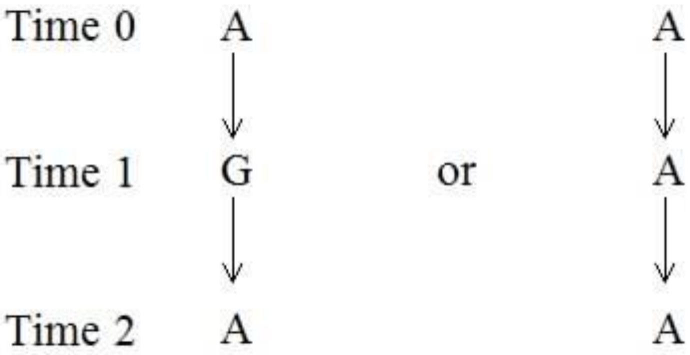
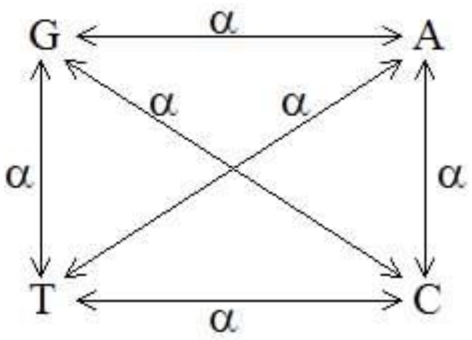
Nucleotides substitution model & Evolution

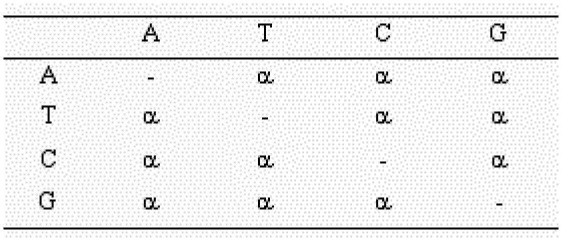
Jukes-Cantor distance

In the Jukes and Cantor (1969) model, the rate of nucleotide substitution is the same for all pairs of the four nucleotides A, T, C, and G. As is shown below, the multiple hit correction equation for this model produces a maximum likelihood estimate of the number of nucleotide substitutions between two sequences. It assumes an equality of substitution rates among sites (see the related gamma distance), equal nucleotide frequencies, and it does not correct for higher rate of transitional substitutions as compared to transversional substitutions. Jukes - Cantor Model (1969) - simple model assuming each nucleotide equally likely to change into any other nucleotide with a rate of alpha.





The Jukes-Cantor model



	A	T	C	G
A	-	α	α	α
T	α	-	α	α
C	α	α	-	α
G	α	α	α	-

If a site within a gene at time 0 is a C, the probability (P) that the site would be a C at time 0 is $P_{C(0)} = 1$.

If a site within a gene at time 0 is a C, the probability (P) that the site would still be a C at time 1 is $P_{C(1)} = 1 - 3a$.

What is the probability that the site would be a C at time 2?

$$P_{C(2)} = (1 - 3a)P_{C(1)} + a[1-P_{A(1)}]$$

At any time (t) in the future, the probability that the site contains a C is defined by

$$P_{C(t)} = ¼ +(3/4)e^{-4at}$$

Through manipulations of this equation, we can derive the following equation for determining K:

$$K = -3/4 \ln[1 - (4/3)p],$$

Where p is the fraction of nucleotides that a simple count reveals to be different between two sequences

The Jukes-Cantor model of a global uniformity in substitution patterns is a clear oversimplification, but it is adequate in many cases where multiple substitutions are possible.

Kimura's Two-Parameter Model

Nucleotide bases fall into two categories depending on the ring structure of the base.

- Purines: A or G (these are two ring bases)
- Pyrimadines: C or T (these are single ring bases)

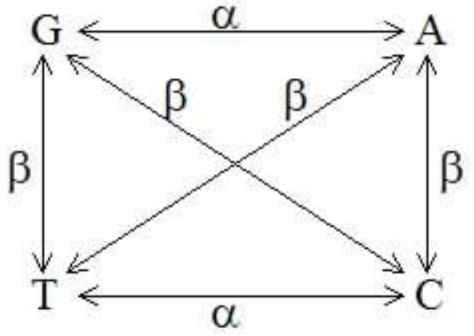
Mutations in DNA are changes in which one base is replaced by another.

A mutation that conserves the ring number is called a transition (e.g., A -> G, G -> A, T -> C, or C -> T).

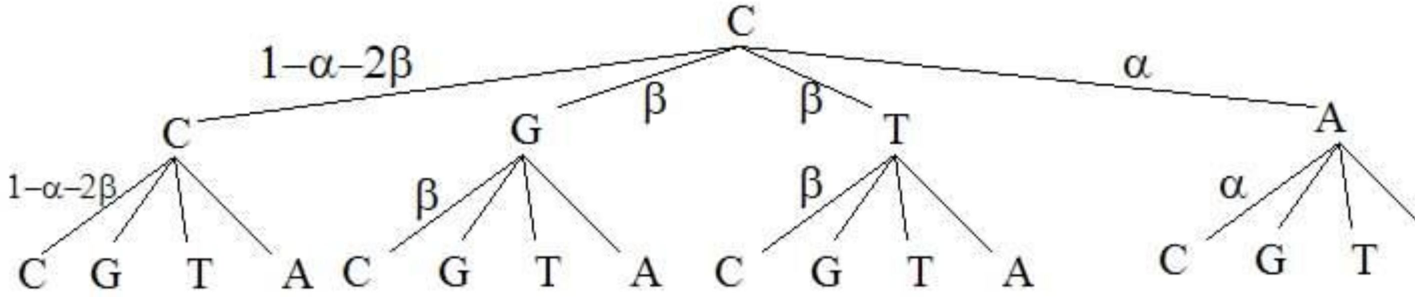
A mutation that changes the ring number is called transversion. (e.g. A -> C, A -> T, C -> G, etc.).

The number of transitions observed to occur in nature (i.e., when comparing related DNA sequences) is at least 3 times as frequent as the number of transversions.

Kimura's Two-Parameter model incorporates these different rates for transitions and transversions.



Transitions are assumed to occur at a uniform rate a and transversions at a different uniform rate b



$P_{CC(2)} = (1 - a - 2b)P_{CC(1)} + bP_{GC(1)} + bP_{TC(1)} + aP_{AC(1)}$

At any time (t) in the future, the probability that the site contains a C is defined by

$P_{CC(t)} = 1/4+(1/4)e^{-4bt} +(1/2)e^{-2(a+b)t}$

Through manipulations of this equation, we can derive the following equation for determining *K*:

$K = 1/2 \ln[1/(1 - 2P - Q)] + 1/4 \ln[1/(1 - 2Q)]$, where

P is the fraction of nucleotides that a simple count reveals to be transitions, and

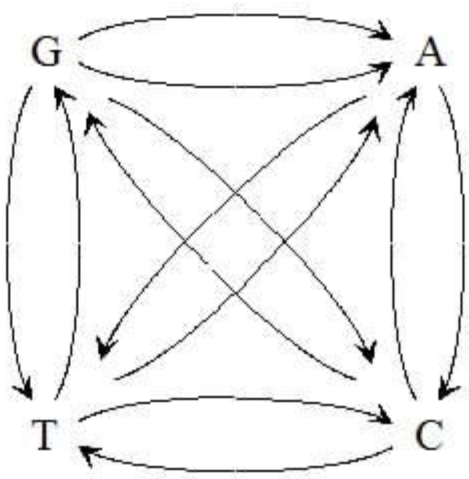
Q is the fraction of nucleotides that a simple count reveals to be transversions

If no distinction between transitions and transversions are made, this equation reduces to the simple Jukes-Cantor equation.

Models with Even More Parameters

As more data became available, it was revealed that Kimura's assumption of two differ uniform rates was also an oversimplification.

By analysis, we can determine relative frequencies for each of the 12 possible types of substitutions



12-parameter models can be generated that attach probabilities to each of these possible types of substitutions. (See table 3.4)

A 13th parameter can be added to compensate for differences in the relative frequencies caused by substitution biases associated with regional genomic GC context.

Due to the assumptions about the original data, the complex models often perform less reliably than the simple Jukes-Cantor or Kimura models.

Substitutions between Protein Sequences

The proportion *p* of different amino acids between two protein sequences can be calculated as:

$p = n / L$, where

n is the number of amino acids that differ between the two sequences, and

L is the number of positions at which at which differences could be observed in the aligned sequences.

Estimating the number of substitutions between two protein sequences is more difficult than the same task for DNA sequences due to:

- back mutations - some substitutions occur more frequently
- substitutional path from one amino acid to another is not always the same length

One solution is to weight each amino acid substitution differently by using empirical data from a variety of different protein comparisions to generate a matrix such as the PAM matrix.