1 Nei & Gojobori's unweighted method

1.1 Synonymous/Nonsynonymous site

For example, codons UUA UUG CUU CUC CUA CUG represent the same amino acid

We first calculate the synonymous (the amino acid type is unchanged after substitution) fraction of UUA

U $f_1 = \frac{1}{3}$		U $f_2 = 0$		A $f_3 = \frac{1}{3}$	
$U \to C$	✓	$U \to C$	×	$A \to U$	×
$U \to A$	×	$U \to A$	×	$A \to G$	✓
$U \to G$	×	$U \to G$	×	$A \to C$	×

$s = f_1 + f_2 + f_3 = \frac{2}{3}$ (synonymous site of a codon)

$n = 3 - s$ (nonsynonymous site of a codon)

$S = \sum_i s_i$ (synonymous sites of the nucleotide sequence)

$N = \sum_i 3 - s_i$ (nonsynonymous sites of the nucleotide sequence)

When we are comparing two sequences, calculate the S and N for both sequences for the length of the shorter sequence and average the values from the two sequences.

1.2 Synonymous/Nonsynonymous nucleotide difference

The difference of two sequences is counted codon by codon. When comparing two condons, there are three types of possibilities: one, two and three nucleotide pairs are mismatched. Each case is taken care of separately.

Case 1 *1 mismatch nucleotide*

If the two condons represent the same amino acid, $S_d = S_d + 1$.

If the two condons represent different amino acid, $N_d = N_d + 1$.

Case 2 *2 mismatch nucleotide pairs*

There are two different paths of the same probability for such substitution. The reason that there are two paths is due to the different order of substitution of the mismatched pair. If there are two mismatch pairs, given the assumption that only one nucleotide is substituted at a time, either the front one is substituted first or the rear one is substituted first, hence there are two different paths. For example, $CCC \to CAA$.

1. $CCC \to CCA \to CAA$
2. $CCC \to CAC \to CAA$

There are 4 substitutions in total, 1 is synonymous ($CCC \to CCA$), 3 are nonsynonymous. Therefore $S_d = S_d + \frac{1}{4}, N_d = N_d + \frac{3}{4}$. There are always two paths for 2 mismatch pairs, but if one of the path involves a stop codon, that path is ignored. For example, $AAA \to TAT$.

1. $AAA \to TAA \to TAT$
2. $AAA \to AAT \to TAT$

In path 1, TAA is a stop codon, therefore path 1 is ignored and we only use the 2 substitutions in path 2. Both substitutions in path 2 are nonsynonymous. Therefore $S_d = S_d + \frac{0}{2}, N_d = N_d + \frac{2}{2}$.

Case 3 *3 mismatch nucleotide pairs*

There are six different paths of the same probability for such substitution. The reason that there are six paths is the same as the previous case, only this time we have 3 substitutions hence we have $3! = 6$ paths. For example, $CTT \rightarrow AGG$.

1. $CTT \rightarrow ATT \rightarrow AGT \rightarrow AGG$
2. $CTT \rightarrow ATT \rightarrow ATG \rightarrow AGG$
3. $CTT \rightarrow CTG \rightarrow AGT \rightarrow AGG$
4. $CTT \rightarrow CTG \rightarrow ATG \rightarrow AGG$
5. $CTT \rightarrow CGT \rightarrow CGG \rightarrow AGG$
6. $CTT \rightarrow CTG \rightarrow CGG \rightarrow AGG$

There are 18 substitutions in total, 6 are synonymous, 12 are nonsynonymous. Therefore $S_d = S_d + \frac{6}{18}, N_d = N_d + \frac{12}{18}$. Similarly, if there is a stop codon in any path, that path is ignored and treated as non-existant.

2 Jukes and Cantor's model for the multiple nucleotide substitution correction

$K = -\frac{3}{4}\ln(1 - \frac{4}{3}p)$

For synonymous substitution: $p_s = \frac{S_d}{S}, ds = K_s = -\frac{3}{4}\ln(1 - \frac{4}{3}p_s)$

For nonsynonymous substitution: $p_n = \frac{N_d}{N}, dn = K_n = -\frac{3}{4}\ln(1 - \frac{4}{3}p_n)$

3 Estimate divergence time

$T = \frac{p_n}{2dn}$ (in Million Years)