How do we reconstruct molecular phylogenetic relationships?

- The first step
  - Sequence selection and alignment,
  - Determine site by site homologies
  - Detect DNA or amino acid differences.
- The second step
  - Build a mathematical model describing the evolution in time of the sequences.
- A model can be built empirically
  - Using properties calculated through comparisons of observed sequences
  - Parametrically, using chemical or biological properties.
  - Standard models will consider substitutions at the DNA, amino acid, or codon level.
• Zuckerkandl and Pauling (1965) the theory of molecular clock.

• Rate of molecular evolution is approximately constant over time for all the proteins in all lineages.

• Any time of divergence between genes, proteins, or lineages can be dated simply by measuring the number of changes between sequences.

• Aris-Brosou and Yang 2002. Relax the clock assumption.

• Any molecular clock seems have different rates for different DNA positions.
  
  – Rates of transitions/ transversions (Brown and Simpson 1982).
• Mutation rates seem to vary both among and within genomes, being affected by factors as:

• Chromosomal position (Sharp et al. 1989)

• Nearest neighbor bases (Blake et al. 1992)

• Different efficiency of the repair systems between DNA strands during replication and transcription (Veauite and Fuchs 1993)

• For general models it is difficult to implement all different mutation rules and patterns that we detect in the genetic material belonging to different species.

• Models have incorporated only the simplest rules or have been based on empirical observation with little understanding of the underlying biology.

• For all models there is an assumption they all share, the Markov property.
Markov property

• Consider a stochastic model for DNA or amino acid sequence evolution.

• Assumption independence of evolution at different sequence sites and thus can consider sites one by one.

• Values 1,2,3,4 to represent the nucleotides A, T, C, G for DNA sequences and 1,...,20 for amino acid sequences.

• A Markov process can have 3 important properties:

  • Homogeneity means that the rate matrix is independent of time.

  • Stationarity means that the process is at that equilibrium, that nucleotide frequencies have remained more or less the same during the course of evolution.

  • Reversibility means that the process of sequence evolution is theoretically indistinguishable from the same process watched in reverse.
How much evolutionary change has occurred between two sequences?

- Simplest
  Count the number of nucleotide sites as which the two sequences differ.

```
ATGCTCG
```

```
ATCCGCA
```

This is a poor measure of the actual number of evolutionary changes.

Change of DNA sequence of a species, is actually mutation, then substitution in a population
Substitutions exchange a purine for purine (C and T) or pyrimidine for other pyrimidine (A and G) transitions.

Exchange purine for pyrimidine or pyrimidine for purine transversions.

Protein coding genes and their consequences for the protein the gene encodes.

Substitution that does not change the amino acid is a synonymous.

Substitution that does change the amino acid is non-synonymous.
### Matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td>No mutation</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>No mutation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>No mutation</td>
<td>No mutation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>No mutation</td>
<td>No mutation</td>
<td>No mutation</td>
<td></td>
</tr>
</tbody>
</table>

### Models of evolution

**Variation in nucleotide frequencies.**

Different kinds of substitution to occur with different probabilities.

**Variation in the rate of substitution between sites.**
Models of evolution

- Jukes and Cantor (1969) proposed a stochastic model for DNA substitution.

- All nucleotide substitutions occur at an equal rate, and when a nucleotide is substituted, any one of the other nucleotides is equally likely to be its replacement.
The Jukes-Cantor model (1969)  
the simplest symmetrical model of DNA evolution

Transition probabilities under Jukes Cantor model

• All sites change independently.

• Each base in the sequence has an equal chance of changing (when 1 change, it changes to 1 of the 3 other bases with equal probability \( \frac{u}{3} \)).

• Assumption: There is not difference of rates of substitution between transitions and transversions.

• Models calculate the probability of transition from one state C to another A.
Kimura's (1980) K2P model of DNA change, which allows for different rates of transitions and transversions

I. At rate $a$, if the site has a purine (A or G), choose one of the two purines at random and change to it. If the site has a pyrimidine (C or T), choose one of the pyrimidines at random and change to it.

II. At rate $b$, choose one of the four bases at random and change to it.

For any nucleotide, there is one chance at rate $a$, that causes a transition, and two, a rate $b$, that cause transversions. The ratio of transitions to tranversions $R$, will be $a/(2b)$.

The total rate of change will be $a + 2b$. 
• Felsenstein (1981) proposed a model in which the rate of substitution to a nucleotide depends only on the equilibrium frequency of that nucleotide. Assumption: Non equal expected frequencies. Program PHYLIP (Kishino and Hasegawa 1989).

• Gojobori et al. (1982), Tamura and Nei (1993) 6 parameters.

The models imply reversibility:
The probability of starting at A and ending at T in evolution is the same as the probability of starting at T and ending at A.
• Lanave et al. (1984), Tavaré (1986), Yang (1994) and Rodríguez et al. (1990) considering the a model that implies no-reversibility.

• \[ \text{G} \rightarrow \text{A} \]
  \[ \text{G} \rightarrow \text{C} \]
  \[ \text{C} \rightarrow \text{T} \]
  \[ \text{T} \rightarrow \text{A} \]

• Ancestor – descendent relationships.

Page and Holmes 1998
ModelTest (Posada and Crandall 1998)
Models of protein evolution

• Amino acid models (probabilities, Markov process)

• Models depending on secondary structure (Hidden Markov Model –HMM-)

• Codon based models

First molecular sequences available were protein sequences.

Eck and Dayhoff (1966) First molecular phylogeny with protein sequences.
• Dayhoff (1972) and Schwartz and Dayhoff (1978) developed a model of protein evolution that resulted in the development of replacement matrices.

• Replacement rates are derived from alignments of protein sequences that are at least 85% identical.

• Probability of accepted mutation (PAM) matrices, from the limited amount of protein sequence data available at the time.

• Matrices 20 x 20 (amino acids)

• In Dayhoff Markovian model, the dynamics of amino acid substitution resembles a continuous time.

• (Crooks and Brenner 2004) suggested that in short time, the dynamics of amino acid substitution are not Markovian, stationary, nor homogeneous.
Models (Michael 2007) including selective pressure.

Gene duplication, deletion and fold addition

Most proteins encoded in a genome have a defined 3D structure that can be classified into distinct protein folds.

The folds are defined by the topology of the peptide chain, it is possible to determine whether two proteins adopt the same fold by sequence comparison.

Different categories of structural environment, for example, α-helix, β-sheet, turn, and loop.

• Grantham (1974). Codon mutation model calculates the distance between the amino acid coded by the different codons on the basis of the physicochemical properties of the amino acids.

• Goldman and Yang (1994) described a codon mutation model.

• In coding regions, natural selection determines the fixation of, amino acid replacements, insertions, and deletions.
• Database search programs to compare protein sequences:
  • FASTA (Pearson and Lipman 1988)
  • BLAST (Atschul et al. 1990, 1997)
  • Standard matrices are often used to compare sequences that have different amino acid compositions due to functional constraints.

• Specialized substitution matrices have been constructed for certain classes of proteins (Ng et al. 2000 and Muller 2001) and adjustments for those matrices by Kuo and Altschul 2005.

• Multiple sequence alignment of proteins:
  – Identify highly conserved amino acids,
  – patterns of conserved sequences,
  – generating clues to functional and evolutionary relationships among proteins (Rodi et al. 2004).

• Amino acid diversity, measure of how far the amino acid abundances of a particular position differ from that of a uniform random distribution.

• DIVAA software (http://relic.bio.anl.gov)
Other models

- 4 different chemical subunits analogous to DNA.
- RNA folds into a variety of complex tertiary structures, analogous to structured proteins.
Rzhetsky (1995) introduced a model to estimate base substitution in ribosomal RNA genes and to infer phylogenetic relationships.

Phylogenetic analyses of ribosomal RNA (rRNA) sequences have given important results about ancient events because of their high levels of conservation over extremely long evolutionary times.

Takes into account rRNA secondary structure elements, namely stem and loop regions.

It is very difficult to infer rRNA tertiary structure.
• Many RNA molecules are subject to functional constraints, resulting in highly conserved secondary structure over long evolutionary times.

• PHASE http://www.bioinf.man.ac.uk/resources/

In conclusion, the modeling of processes of sequence evolution is a prospering field of research.

All evolutionary models started simply with some assumptions; development of software introduced more complexity in the models.

There is not a tight relationship between models of evolution of macromolecules and reconstruction of evolutionary history of genes and organisms.