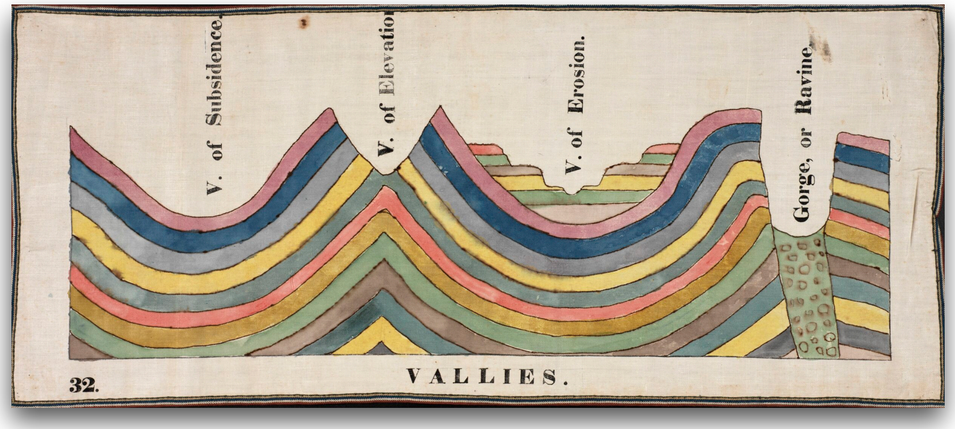
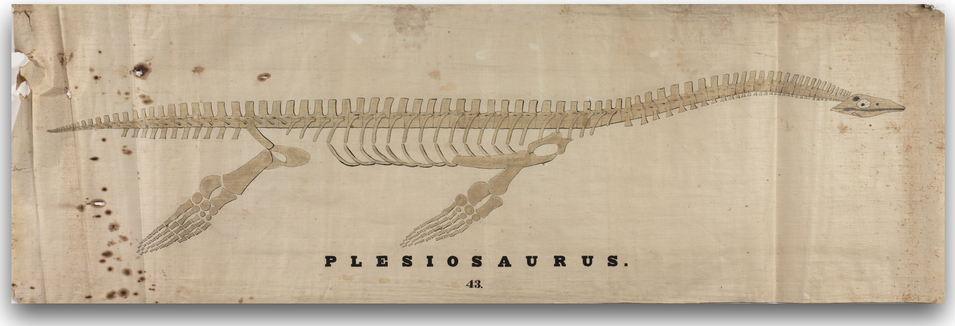
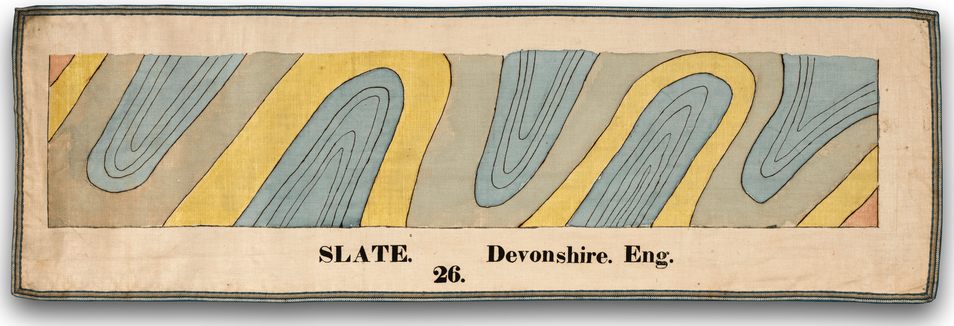






Models of sequence evolution

Orthology inferences are inferred **pairwise**.

When considering **multiple species**,
we should move to the concept of … **orthogroup**.

An orthogroup is defined by a **reference speciation event**.

An orthogroup is defined as a **set of homologous genes** that descend **from a single gene in the last common ancestor** of species considered.

An **orthogroup** contains:

Orthologs – genes of different species that **evolved from a single ancestral gene** through **speciation events**.

In-paralogs – gene duplicates that **occurred after the last common ancestor of the species considered**.

BEWARE:

Orthogroups **do not separate orthologs and in-paralogs** but group all genes descended from a single one in the LCA of the studied species.

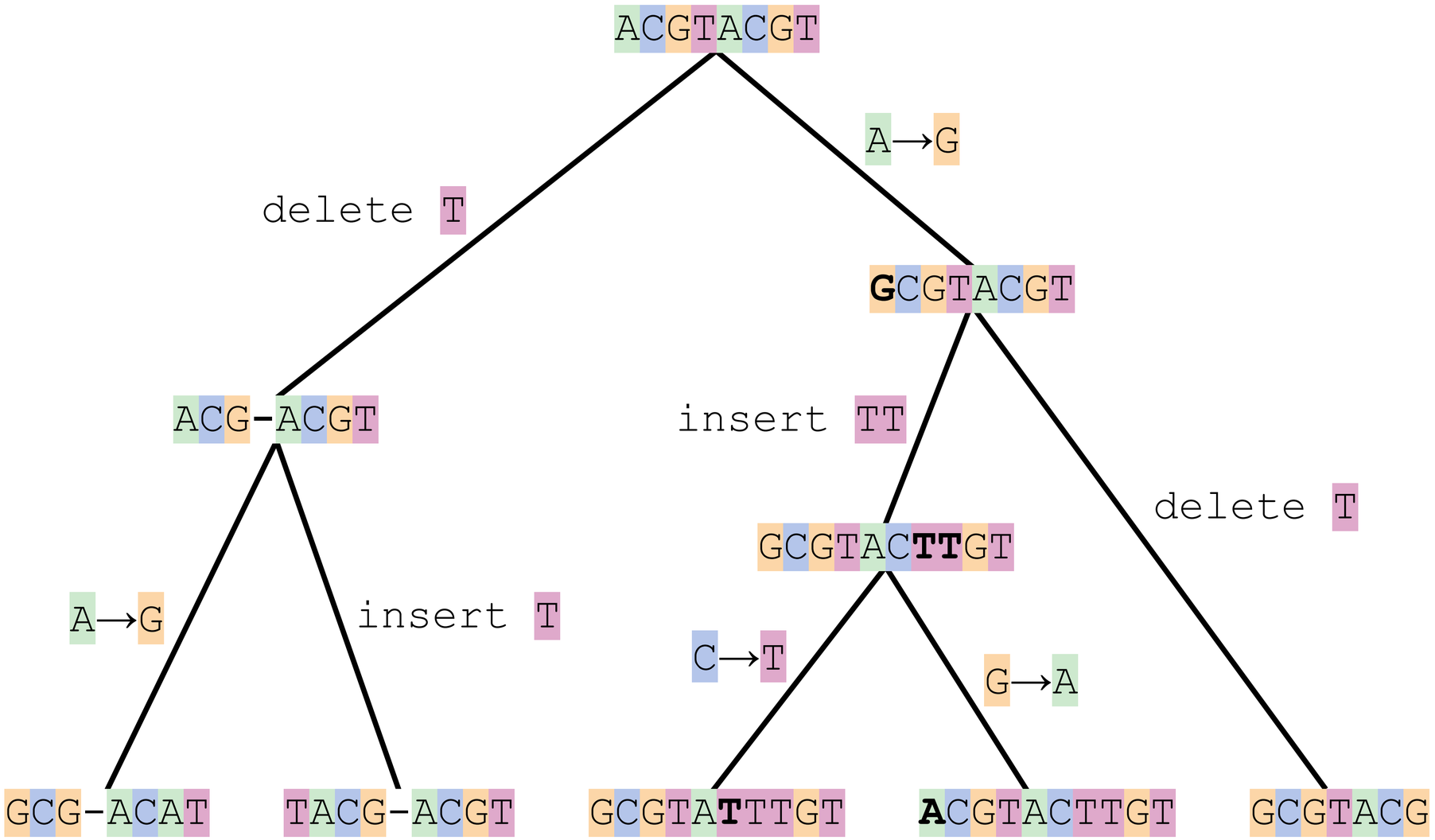
Out-paralogs - gene duplications that occurred before the LCA - are **not included in an orthogroup**, as they belong to a different orthogroup.

orthologous genes, homologous sites!

```
species 1   AGGATCTGCAATTGCTCTTCTAATCTGTCTGATCAGGAT
species 2   AGG------AATTGCTCTTCTAATCTGTCT---CAGGAT
species 3   AGGATCTGCAATTGC---TCTAATCTGTCTGATCAGGAT
species 4   AGAATCTGCAATTGCTCTTCTGATCTGTCTGATCACGAT
species 5   AGGATCTGC---TGCTCTTCTGATCTGTCTGATCAGGAT
```

The goal of alignement is to identify which **positions** are **homologous**.

That is, their **evolutionary history** reflect the **species relationships**!


ACGTACGT
delete T
A→G
ACG-ACGT
GCGTACGT
insert TT
delete T
A→G
insert T
GCGTACTTGT
C→T
G→A
GCG-ACAT
TACG-ACGT
GCGTATTTGT
ACGTACTTGT
GCGTACG

Distance-based methods

Distance-based phylogenetic trees are based on the **total number of evolutionary changes** between pairs of sequences.

Starting from the alignment, these methods look at all possible pairs of the aligned sequences and count how many characters are different at each position. These pairwise differences are represented in a **distance matrix**.

These methods are now best used for **exploratory analysis** of large datasets before conducting more intensive tree building using character-base methods.

Character-based methods

Character-based methods compare all sequences by **considering one character** (nucleotides or amino acids) in the alignment **at a time**.

As character-based methods incorporate **evolutionary models** making these methods more accurate than distance-based methods.

Character-based phylogenetic inference is really about tree-scoring, not tree-finding .. ⛺

Nearest Neighbour Interchange (NNI) is a local search method that makes small, localized changes to the tree. It is the simplest and fastest tree rearrangement algorithm!

Subtree Pruning and Regrafting (SPR) is a more extensive tree rearrangement method than NNI. It moves entire subtrees rather than just swapping immediate neighbors.

Tree Bisection and Reconnection (TBR) is the most aggressive of these tree rearrangement methods and explores a much larger space of possible tree topologies. It has the lesser chance of local optima!

"*All models are wrong, but some are useful*"

Box, 1979

In biology, a **substitution model**, also called **models of sequence evolution**, are Markov models that describe changes over evolutionary time.

A substitution model is a mathematical framework that describes the probabilities of **changes** (substitutions) occurring in **sequence evolution over time**.

It accounts for different types of substitutions (e.g., nucleotide, amino acid, or codon changes) and is used to estimate evolutionary distances between sequences and to

by correcting for multiple substitutions that may have occurred at the same site.

In a **Markov model** is a type of stochastic model used to represent systems that transition between a finite or countable number of states in a chain-like manner.

The defining characteristic of a Markov model is the **Markov property**, which stipulates that the probability of transitioning to the next state depends solely on the current state and not on the sequence of events that preceded it.

Key components:

- **States:** The distinct configurations or conditions that the system can occupy.

- **Transition Probabilities:** The probabilities associated with moving from one state to another.

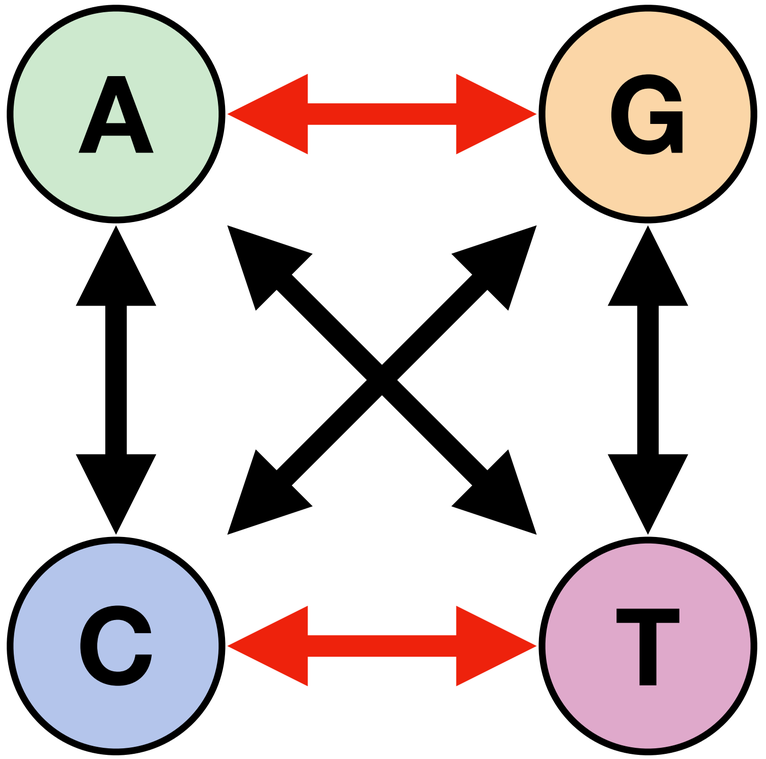
Mechanistic models: consider the biological process involved: mutational biases, natural selection… they have more interpretative power and particularly useful for studying the evolutionary forces and mechanisms.

Empirical models: describe the relative rates of substitution do not consider explicitly factors that influence the evolutionary process large quantities of sequence data.

Models of sequence substitution can be built for different evolutionary units:

- **nucleotides**: 4 states (A, C , G , T)
- **codons**: 61 states (AAA, AAC , AAT , …)
- **amino acids**: 20 states (Phe, Leu, Ile, …)

Nucleotide Substitution Models



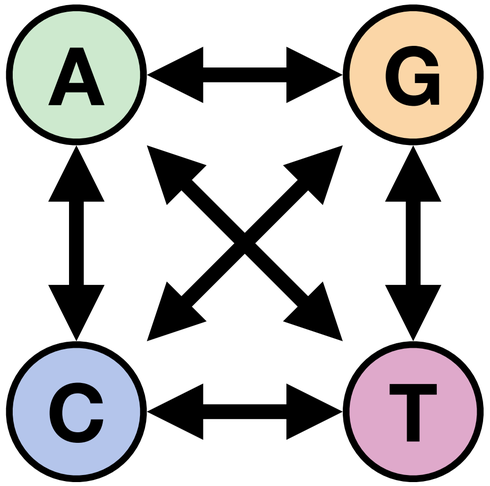
Transitions (Ts): Substitutions within the same nucleotide class A ↔ G (purines), C ↔ T (pyrimidines). More frequent than transversions due to structural similarity.

Transversions (Tv): Substitutions between different nucleotide classes (purine ↔ pyrimidine). Examples: A ↔ C, A ↔ T, G ↔ C, G ↔ T. Less common due to greater structural change.

Jukes–Cantor (JC69)

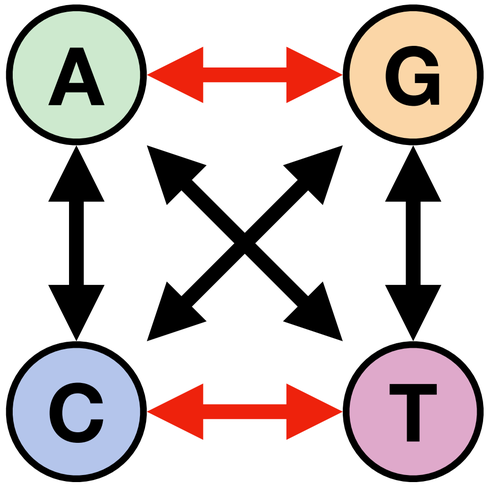
The simplest nucleotide model, JC69 assumes all nucleotides are equally frequent and all substitution rates are equal.

It has a single parameter (overall substitution rate) and corrects for multiple substitutions analytically. This model is often used as a baseline, though it may oversimplify real sequence evolution.



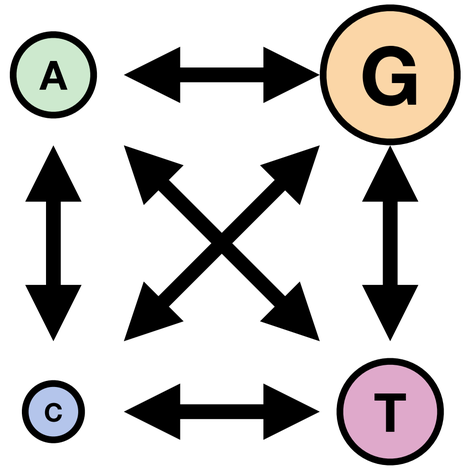
Kimura 2-Parameter (K80)

Kimura (1980) introduced a two-parameter DNA model that distinguishes between transitions and transversions. K80 still assumes equal base frequencies (25% each) but better fits data with the commonly observed transition bias. It uses a separate rate (alfa) for transitions (interchanges of A↔G or C↔T) and another (beta) for transversions.



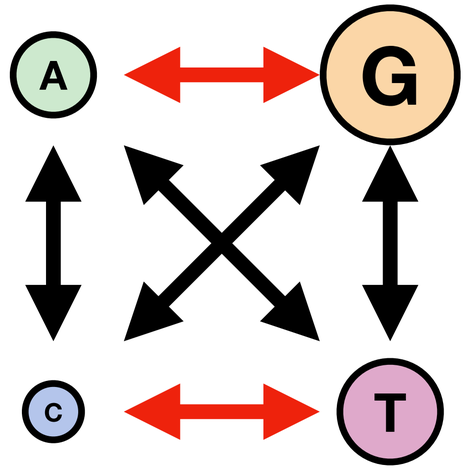
Felsenstein (F81)

Felsenstein’s 1981 model extends JC69 by allowing unequal base frequencies while keeping all substitution rates equal . In F81, each nucleotide has its own equilibrium frequency … π_A, π_C, π_G, π_T … This accounts for compositional bias in sequences but assumes no transition/transversion rate difference.



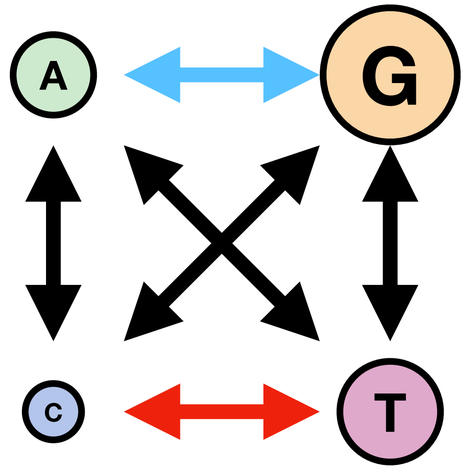
Felsenstein (F81)

Felsenstein’s 1981 model extends JC69 by allowing unequal base frequencies while keeping all substitution rates equal . In F81, each nucleotide has its own equilibrium frequency … π_A, π_C, π_G, π_T … This accounts for compositional bias in sequences but assumes no transition/transversion rate difference.



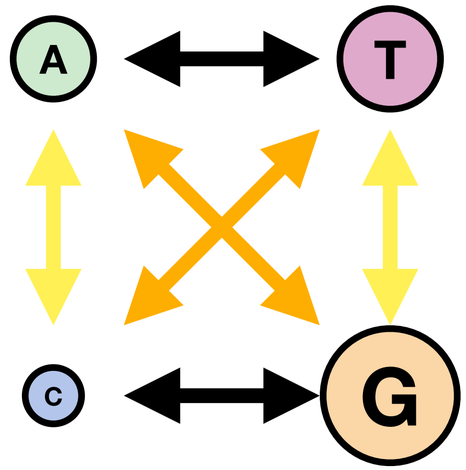
Tamura–Nei (TN93)

Tamura and Nei (1993) further refine transition rate parameters. TN93 allows two distinct transition rates while keeping a single rate for all transversions. It also permits unequal base frequencies.



Kimura 3-parameter (K3P)

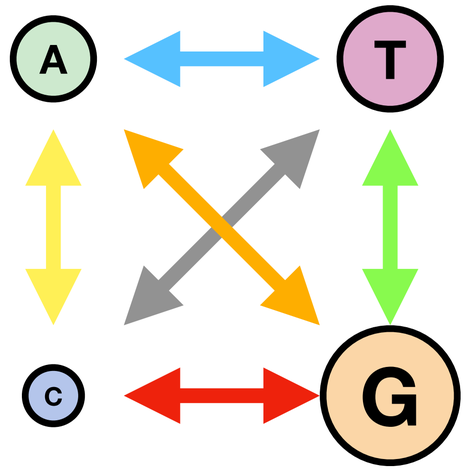
Kimura (1981) introduced the Kimura 3-parameter (K3P) model, which extends the Kimura 2-parameter (K2P) model by incorporating variable base frequencies. K3P assumes equal transition rates but allows two distinct transversion rates, making it more flexible than K2P while maintaining computational efficiency.



General Time Reversible (GTR):

GTR (Tavaré 1986) is the most general reversible model for nucleotides. It includes a full set of six substitution rate parameters (one for each pair of nucleotides) and allows any set of equilibrium base frequencies. GTR usually provides the best fit to data, at the cost of more parameters.

All simpler models (JC, HKY, TN93, etc.) are special cases of GTR with certain rates constrained equal.



Models are defined by the **Q-matrix**, specifying the instantaneous rates of substitution. It incorporates:

Equilibrium frequencies of nucleotides:

- π_A - equilibrium frequencies of A
- π_C - equilibrium frequencies of C
- π_G - equilibrium frequencies of G
- π_T - equilibrium frequencies of T

Substitution rate parameters between nucleotide pairs:

- **α** - substitution rate A to G
- **β** - substitution rate A to C
- **γ** - substitution rate A to T
- **δ** - substitution rate G to C
- **ϵ** - substitution rate G to T
- **η** - substitution rate C to T

Assuming equal frequencies, the off-diagonal elements of the Q matrix can be calculated as:
A→C: $q_{ac} = r_{ac} \times \pi_c = 1.6385 \times 0.25 = 0.4096$
A→G: $q_{ag} = r_{ag} \times \pi_g = 3.0090 \times 0.25 = 0.7523$
A→T: $q_{at} = r_{at} \times \pi_t = 1.6385 \times 0.25 = 0.4096$

The diagonal element q_{aa} is then calculated to ensure the row sums to zero: $q_{aa} = -(q_{ac} + q_{ag} + q_{at}) = -(0.4096 + 0.7523 + 0.4096) = -1.5715$. This process is repeated for all nucleotides!

General Time Reversible (GTR):

$$Q = \begin{pmatrix} -(\alpha\pi_G + \beta\pi_C + \gamma\pi_T) & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A & -(\alpha\pi_A + \delta\pi_C + \epsilon\pi_T) & \delta\pi_C & \epsilon\pi_T \\ \beta\pi_A & \delta\pi_G & -(\beta\pi_A + \delta\pi_G + \eta\pi_T) & \eta\pi_T \\ \gamma\pi_A & \epsilon\pi_G & \eta\pi_C & -(\gamma\pi_A + \epsilon\pi_G + \eta\pi_C) \end{pmatrix}$$

Key Points:

The diagonal elements are set such that each row sums to zero, ensuring that the total probability remains conserved over time.

The model is time-reversible, meaning the process of substitution looks the same forward and backward in time when the system is at equilibrium.

Models ranked by complexity

Jukes-Cantor (JC): equal base frequencies, all substitutions equally likely

Felsenstein 1981 (F81): variable base frequencies, all substitutions equally likely

Kimura 2-parameter (K80): equal base frequencies, one transition rate and one transversion rate

Hasegawa-Kishino-Yano (HKY): variable base freqs, one transition and one transversion rate

Tamura-Nei (TrN): variable base frequencies, equal transversion rates, variable transition rates

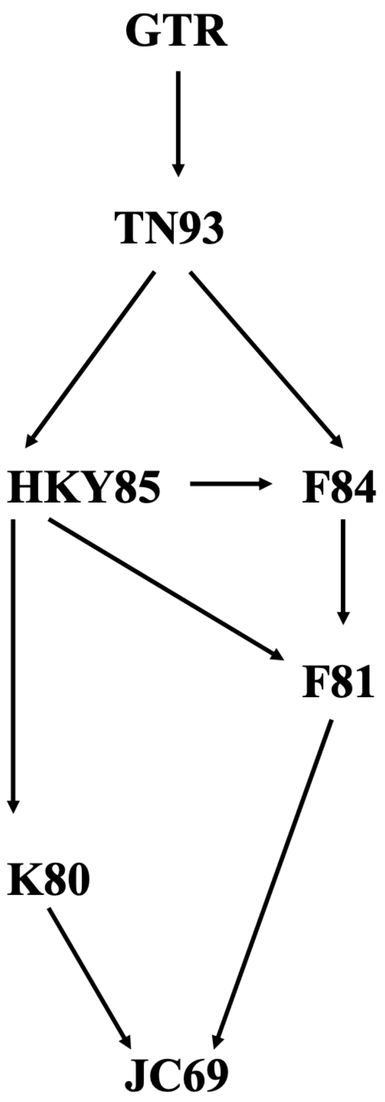
Kimura 3-parameter (K3P): variable base freqs, equal transition rates, two transversion rates

transition model (TIM): variable base freqs, variable transition rates, two transversion rates

transversion model (TVM): variable base freq, variable transversion rates, transition rates equal

symmetrical model (SYM): equal base freqs, symmetrical substitution matrix (A to T = T to A)

general time reversible (GTR): variable base frequencies, symmetrical substitution matrix



8

a, *b*, *c*, *d*, *e*, π_A, π_C, π_G,
(*a*+*b*+*c*+*d*+*e*+*f* =1 π_A,+ π_C,+ π_G,+ π_T,=1)

5

a=*c*=*d*=*f*, *b*=*e*, π_A, π_C, π_G,
(*Ti/Tv ratio, Y/R ratio and and three frequencies since* π_A,+ π_C,+ π_G,+ π_T,=1)

4

a=*b*=*c*=*d*=1, *b*=*e*=k, π_A, π_C, π_G,
(*Ti/Tv ratio and only thre e frequencies since* π_A,+ π_C,+ π_G,+ π_T,=1)

3

a=*c*=*d*=*f*=*b*=*e*=*f*, π_A, π_C, π_G,
(*only three nt frequencies since* π_A,+ π_C,+ π_G,+ π_T,=

1

a=*b*=*c*=*d*=1, *b*=*e*=k, π_A=π_C=π_G =π_T
(*Ti/Tv ratio*)

0

a=*b*=*c*=*d*=*e*=*f*, π_A=π_C=π_G =π_T

Codon Substitutions Models

Translation involves the sequential recognition of triplets of adjacent nucleotides called **codons**.

Since there are four nucleotides, there are 4*4*4 = **64 possible codons**.

In the universal genetic code, 61 code for a specific amino acid, while the other three are **stop codons**. Given that **there are 61 codons but only 20 amino acids**, most amino acids are encoded by more than one codon.

The degeneracy of the genetic code allows **substitutions** to occur in DNA that do **not result in a change in the amino acid sequence**.

These substitutions are known as **synonymous substitutions** (i.e., those substitutions that do not result in an amino acid change), and are generally thought to be under relaxed selective pressure and more common than **non-synonymous substitutions** (i.e., those substitutions that do result in an amino acid change).

Muse and Gaut (MG94)

In 1994, Muse and Gaut introduced a codon substitution model that assigns different rates to synonymous vs. nonsynonymous substitutions.

In MG94, two parameters can be used:

• one for the rate of synonymous changes (**dS**)

• one for nonsynonymous changes (**dN**).

This allows for the estimation of the absolute rates of each type, reflecting selective constraints on protein-coding genes. Natural selection operates mainly at the protein level.

Goldman–Yang 1994 (GY94)

Goldman and Yang's codon model, also published in 1994, similarly distinguishes codon changes by their effect on the encoded amino acid. Rather than separate dN and dS, it is often parameterized in terms of their ratio **ω = dN/dS**.

ω < 1 - purifying selection
ω = 1 - neutral evolution
ω > 1 - positive selection

The GY94 framework also typically incorporates a nucleotide mutation model (for the underlying base changes, e.g. a transition/transversion bias κ) and codon frequency parameters (e.g. using empirical codon frequencies or an F3×4 model of base frequencies at codon positions)

GY94 and MG94 laid the groundwork for most modern codon phylogenetic analyses by explicitly modeling selective pressure on proteins.

Building on GY94, Yang and Nielsen and others developed more complex codon models to detect selection.

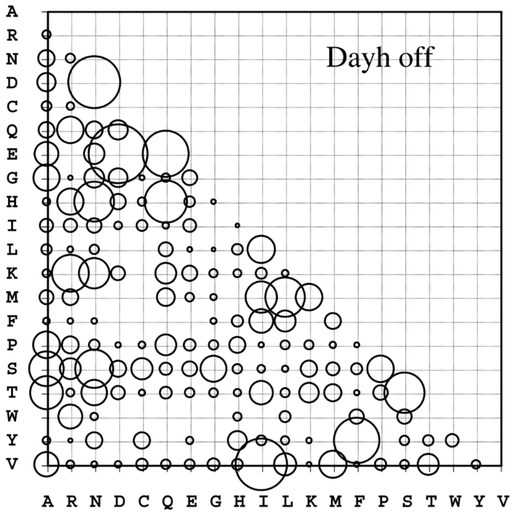
- **Site-specific models** in which the ω ratio varies across codon sites
- **branch-specific models** (Yang 1998) let ω differ among lineages to find episodes of adaptive evolution on particular branches.
- **Branch-site models** combine these ideas to detect positive selection affecting certain sites along specific lineages.

We will learn more on how to infer selection across sites and branches in **lesson 18**!

Amino Acid Substitution Models

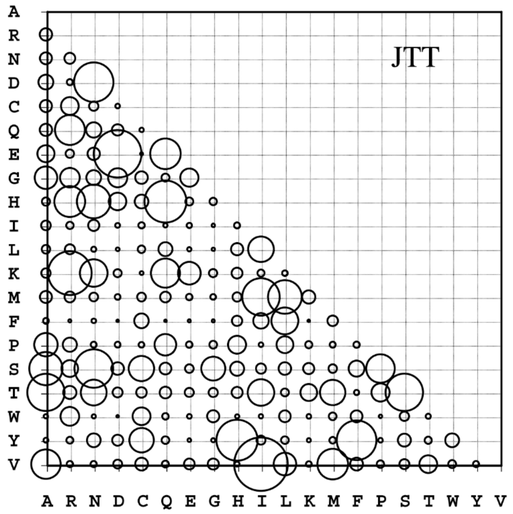
Dayhoff's PAM Model

Dayhoff et al. (1978) pioneered empirical protein evolution models! They compiled alignments of closely related proteins (>85% identity) to estimate amino acid replacement frequencies, resulting in the PAM (Percent Accepted Mutation) matrices.
For example, the PAM250 matrix corresponds to 250 accepted changes per 100 amino acids. The Dayhoff model (PAM) was foundational, though its limited dataset caused some rate biases (e.g. favoring globular protein substitutions).



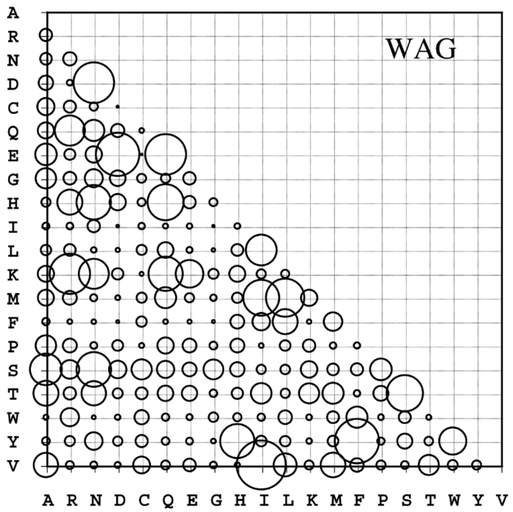
JTT (Jones–Taylor–Thornton):

JTT is an updated empirical model published in 1992 that built on Dayhoff's approach with a larger, more diverse protein database. Jones et al. derived a new 20×20 substitution matrix that better represents typical protein evolution, and it has been widely adopted for maximum-likelihood phylogenetic analyses.
They also produced a separate matrix for membrane proteins, noting that Dayhoff's matrix was biased toward soluble proteins!



WAG (Whelan and Goldman):

WAG is a 2001 amino acid model estimated using a maximum-likelihood approach on a broad set of protein families.
Unlike Dayhoff and JTT (which used observed counts and log-odds methods), WAG optimized the replacement rates to maximize likelihood, given a large training alignment dataset. This resulted in a more accurate general-purpose matrix that often fits protein data better than older matrices.



LG (Le and Gascuel): LG (2008) is a an improved empirical model derived from a vast alignment database using ML methods. Generally outperforming WAG/JTT, LG is a common choice for phylogenetics.

MtREV (Mitochondrial REV): To address the unique unique constraints of mitochondrial genomes, Adachi and Hasegawa (1996) developed the MtREV model from 20 vertebrate mitochondrial protein sequences. Similarly, specialized matrices like MtMam for mammalian mitochondria and HIVb for HIV proteins have been created.

BLOSUM (BLOcks SUbstitution Matrix): (Henikoff & Henikoff 1992) is a series of matrices (e.g. BLOSUM62) derived from conserved ungapped blocks of protein alignments. BLOSUM matrices are widely used for sequence alignment and database search, though they are not derived from an explicit evolutionary model.

+ possible to estimate a protein substitution rate matrix using **QMaker**!

1. Empirical and Mechanistic models:

Protein Substitution Models: Derived from extensive datasets, these models have pre-estimated parameters that remain fixed during analysis. While others, such as amino acid frequencies, can be estimated from the specific dataset.

Nucleotide Substitution Models: Mechanistic Models: Parameters, including substitution rates and base frequencies, are typically estimated directly from the dataset under analysis.

2. Distinguishing Model Parameters and Optimized Values:

Model Parameters: Refer to the structural components of the model, such as the number of substitution types or the inclusion of rate heterogeneity.

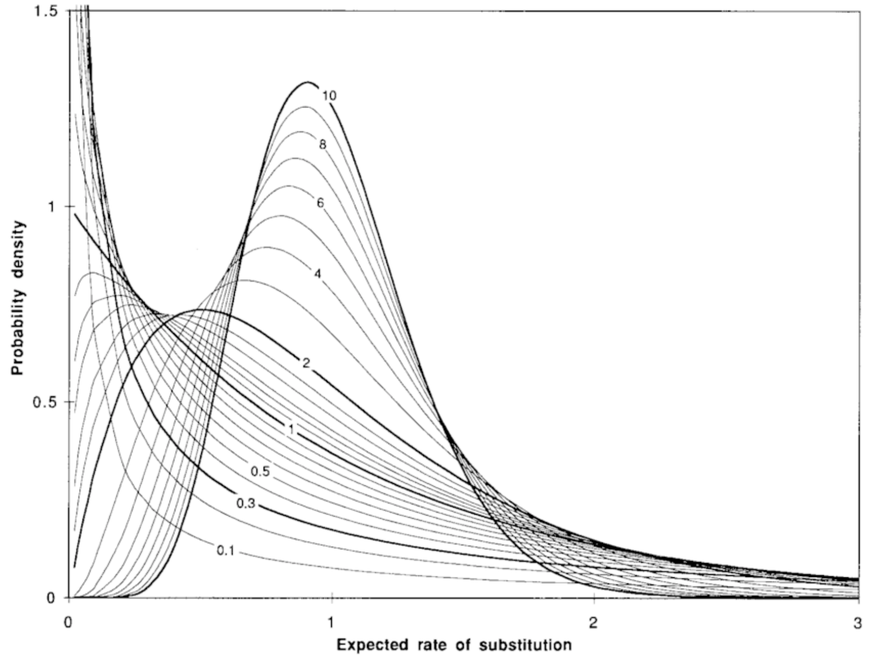
Optimized Numerical Values: These are the specific estimates of the parameters obtained by fitting the model to the data, reflecting the unique evolutionary characteristics of the dataset.

Key Assumptions:

The probability of a nucleotide changing from state ***i*** to state ***j*** depends solely on the current state ***i***, without influence from prior states. This reflects the Markov property, where the process lacks memory of past events.

The rates at which substitutions occur are constant over time. This implies that the evolutionary process is uniform throughout the timeline, without periods of accelerated or decelerated mutation rates.

The nucleotide frequencies remain constant over time, indicating that the system has reached a stationary distribution. At equilibrium, the proportion of each nucleotide does not change, reflecting a balance in the substitution process.



Rate Variation: Real sequence data rarely evolve at a uniform rate across all sites.

- **Gamma-distributed rate heterogeneity (+Γ)** assumes sites have variable rates drawn from a gamma distribution (with shape parameter α), allowing sites to evolve faster or slower than average

- **Invariant sites (+I)** models allow a proportion of sites to be evolutionarily invariable (zero rate)

These modifiers can be added to any base model (e.g. HKY+Γ, GTR+Γ+I) to better capture among-site rate variation, significantly improving likelihood fit in many cases!

What about distances?

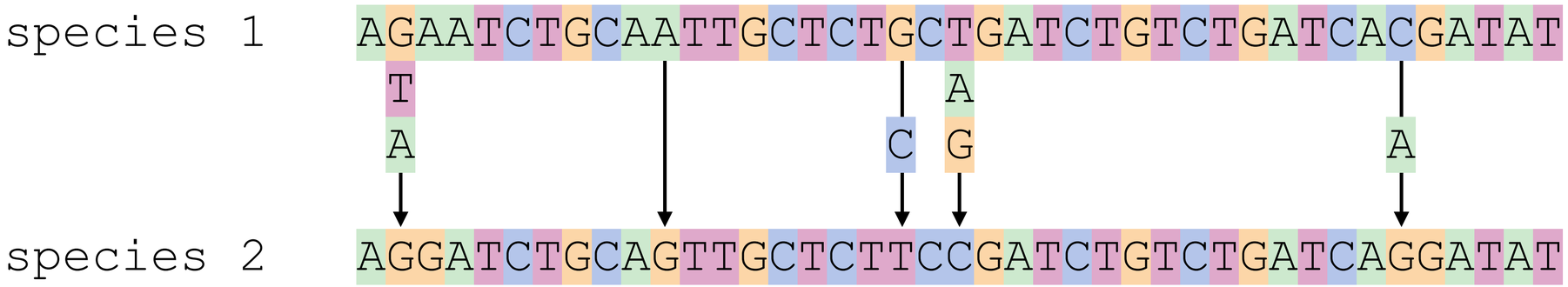
How to define the genetic distance between two taxa?

Observed distance or p-distance

the proportion of nucleotide sites at which the two sequences differ.

Expected distance or corrected distance

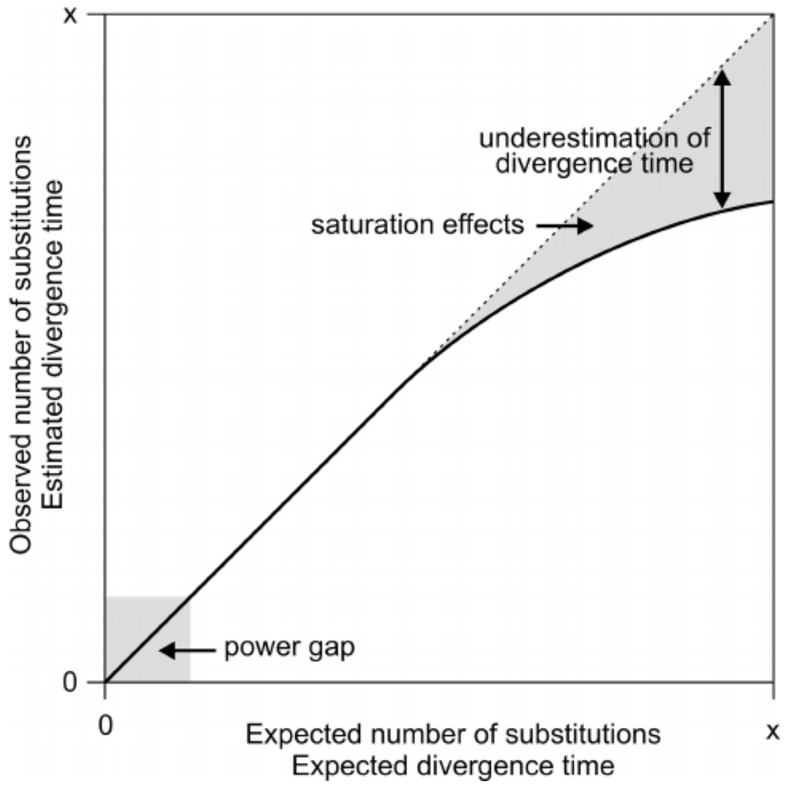
an estimate of the actual number of substitutions that have occurred, using substitution models to correct observed distances by considering the likelihood of multiple substitutions over time.

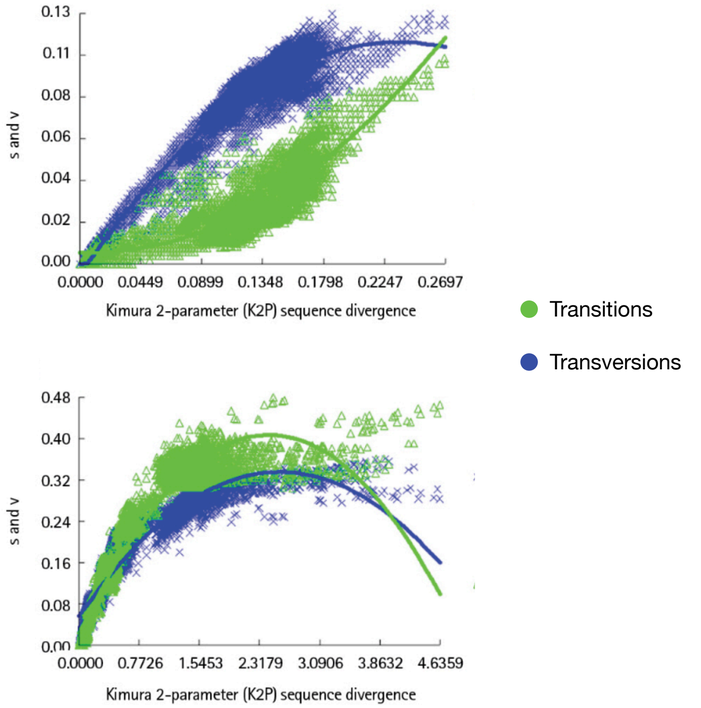


Substitution Saturation: when multiple nucleotide or amino acid substitutions have happened at the same site over time, leading to an underestimation of the true evolutionary divergence between sequences. Leads to:

Loss of Phylogenetic Signal: Saturation diminishes the informative value of sequences.

Long Branch Attraction: Highly divergent lineages may appear erroneously similar due to saturation.





FINISH