

Molecular Evolution

Molecules as Documents of Evolutionary History, Molecular Phylogenetics, and Rates of Molecular Evolution

Molecular evolution is the study of how DNA, RNA and protein sequences change over time, and how those changes can be used to reconstruct the history of life. Because every heritable change in a lineage is, in principle, recorded in its genome, biological macromolecules act as historical "documents" - and comparing these documents across species is the basis of molecular phylogenetics and of estimating when lineages diverged.

1. Molecules as Documents of Evolutionary History

The idea that biological macromolecules record evolutionary history was formalised by **Emile Zuckerkandl and Linus Pauling**. Comparing haemoglobin sequences across species in the early 1960s, they noticed that the degree of dissimilarity between protein sequences corresponded approximately to the phylogenetic distance between the source species. This led to their influential 1965 paper "*Molecules as Documents of Evolutionary History*", which argued that informational macromolecules - DNA, RNA and proteins - carry a historical record of the lineage that can, in principle, be read much like a document, independent of (and complementary to) the fossil record.

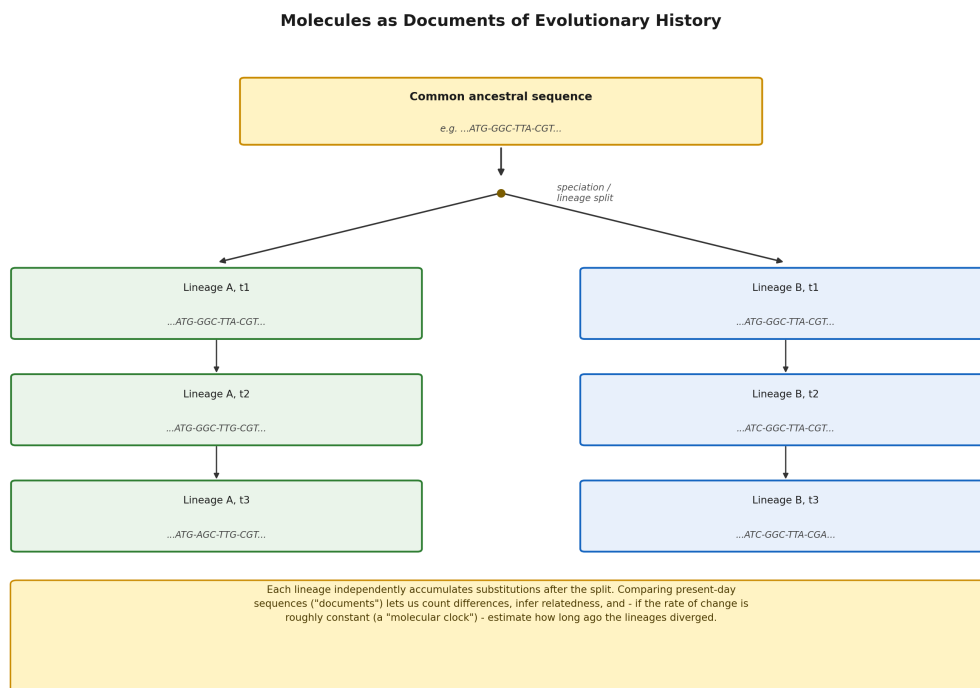


Fig. 1 - After a lineage splits, each descendant independently accumulates substitutions; comparing present-day sequences reveals this shared and divergent history.

1.1 Homology: the basis of comparison

Any comparison of molecules across species is only meaningful between **homologous** sequences - those descended from a common ancestral sequence. Homologous genes are further classified by how they came to exist in the genomes being compared:

- **Orthologs** - homologous genes in different species that arose from a single ancestral gene through a **speciation** event. Orthologs typically retain the same function and are the most reliable markers for reconstructing species relationships.
- **Paralogs** - homologous genes within the same genome that arose through a **gene duplication** event. Paralogs are free to diverge in function after duplication (one copy can retain the ancestral role while the other evolves a new one), which is a major route to new gene functions.

Orthologs and Paralogs: Gene Trees Inside a Species Tree

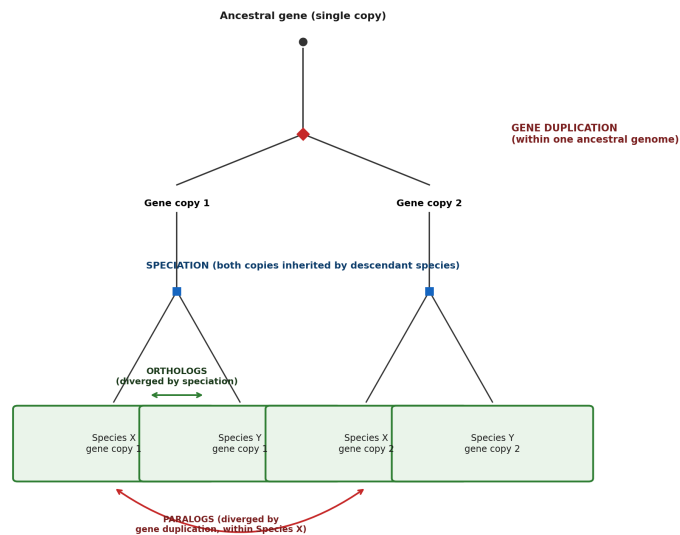


Fig. 2 - A gene duplication followed by speciation produces two pairs of orthologs and two pairs of paralogs from a single ancestral gene.

Key points

- Using paralogs instead of orthologs to infer a species tree is a classic pitfall - the resulting "gene tree" can genuinely differ from the true "species tree" because it is tracking gene duplication history, not speciation history.
- The molecular clock concept - that molecular differences accumulate roughly in proportion to time since divergence - grew directly out of the "documents" idea and underpins molecular dating of evolutionary events (Section 3).

2. Molecular Phylogenetics

Molecular phylogenetics uses sequence data to reconstruct a **phylogenetic tree** - a branching diagram representing hypothesised evolutionary relationships. The workflow begins with a **multiple sequence alignment (MSA)**, which arranges homologous sequences so that positions (sites) sharing a common ancestral origin line up in the same column; all downstream tree-building methods operate on this aligned matrix.

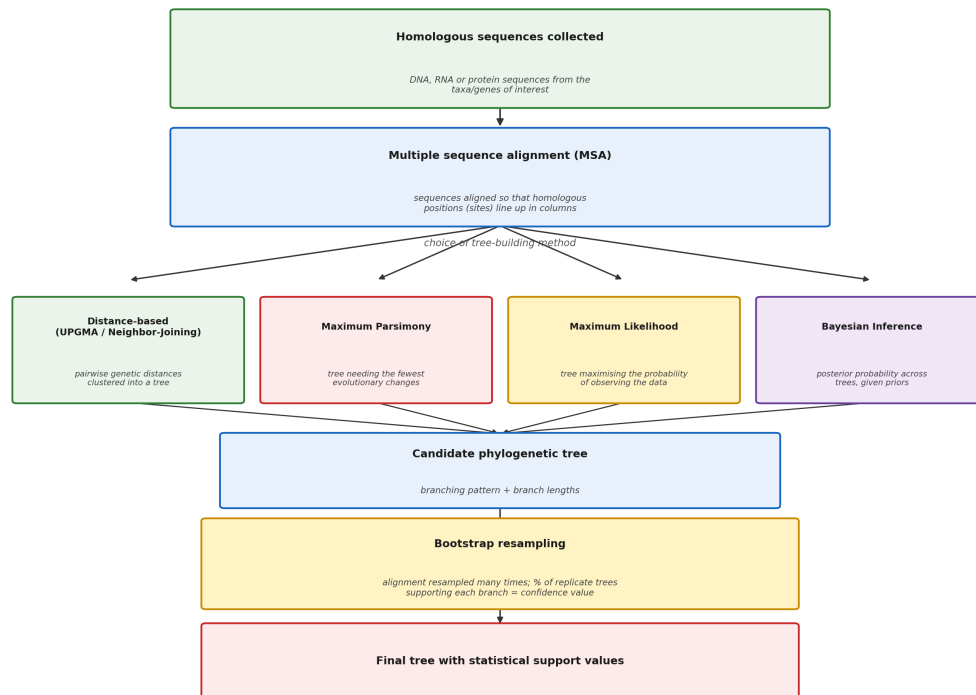


Fig. 3 - From raw homologous sequences to a phylogenetic tree with statistical support: alignment, choice of tree-building method, and bootstrap evaluation.

2.1 Tree-building methods

Method	Principle	Strength / limitation
Distance-based (UPGMA)	Pairwise distances clustered assuming a constant rate of change (a molecular clock)	Fast and simple; produces a rooted tree, but inaccurate if rates vary across lineages
Distance-based (Neighbor-Joining)	Pairwise distances clustered without assuming equal rates across lineages	Fast and widely used; produces an unrooted tree; less accurate than character-based methods for complex data

Method	Principle	Strength / limitation
Maximum Parsimony (MP)	Selects the tree topology requiring the fewest evolutionary changes overall	Conceptually simple; can be statistically inconsistent ("long-branch attraction") for fast-evolving lineages
Maximum Likelihood (ML)	Selects the topology that maximises the probability of observing the actual data, under an explicit model of sequence evolution	Statistically robust and widely trusted; computationally intensive
Bayesian Inference	Combines the likelihood with prior probabilities to give a posterior probability distribution over trees	Directly quantifies uncertainty in the tree; computationally heavy, needs careful priors

2.2 Assessing confidence: bootstrapping

Because a tree is only an estimate, its reliability must be assessed. In the standard **bootstrap** procedure, columns of the alignment are resampled with replacement many times (typically 100-1000 replicates) to generate pseudo-replicate data sets, and the tree-building method is re-run on each. The percentage of bootstrap replicate trees that recover a given branch (a **bootstrap support value**) is reported alongside that branch - values above roughly 70-80% are generally considered well supported.

Worth noting

- No single method is universally "best": UPGMA/Neighbor-Joining are useful for a quick first look or very large data sets, Maximum Parsimony is intuitive but can be misled by convergent/parallel changes, and Maximum Likelihood/Bayesian methods are generally preferred for rigorous analysis because they use an explicit, testable model of how sequences evolve.

3. Rates of Molecular Evolution

3.1 The molecular clock hypothesis

Zuckerkandl and Pauling's observation that protein differences scaled roughly with time since divergence led to the formal **molecular clock hypothesis**: for a given molecule, substitutions accumulate at an approximately constant rate over time, so the number of differences between two sequences can be used to estimate how long ago their lineages diverged (after calibrating the clock against at least one known divergence date, usually from the fossil record). In practice the clock is rarely perfectly constant - rates vary between genes and between lineages - so modern studies typically use **relaxed molecular clock** models that allow rate variation across branches of a tree.

3.2 The neutral theory of molecular evolution

Proposed by **Motoo Kimura**, the **neutral theory** holds that the great majority of substitutions observed at the molecular level are selectively **neutral or nearly neutral** - neither beneficial nor harmful - and become fixed in a population by random **genetic drift** rather than by positive natural selection. This does not contradict Darwinian selection acting on morphology and phenotype; it simply proposes that, at the sequence level, most fixed differences are "noise" rather than adaptation. A key consequence is that under strict neutrality, the rate of substitution is expected to equal the underlying mutation rate and to be roughly constant over time - providing a theoretical justification for the molecular clock.

3.3 Synonymous vs non-synonymous substitutions

Within protein-coding DNA, substitutions are classified by their effect on the encoded protein. A **synonymous (silent) substitution** changes a codon without changing the amino acid it specifies (exploiting redundancy in the genetic code), while a **non-synonymous (replacement) substitution** changes the amino acid. Because synonymous changes are largely invisible to natural selection acting on protein structure/function, they accumulate close to the neutral (mutation) rate, whereas non-synonymous changes are much more exposed to selection - most are removed by purifying selection, so they typically accumulate more slowly.

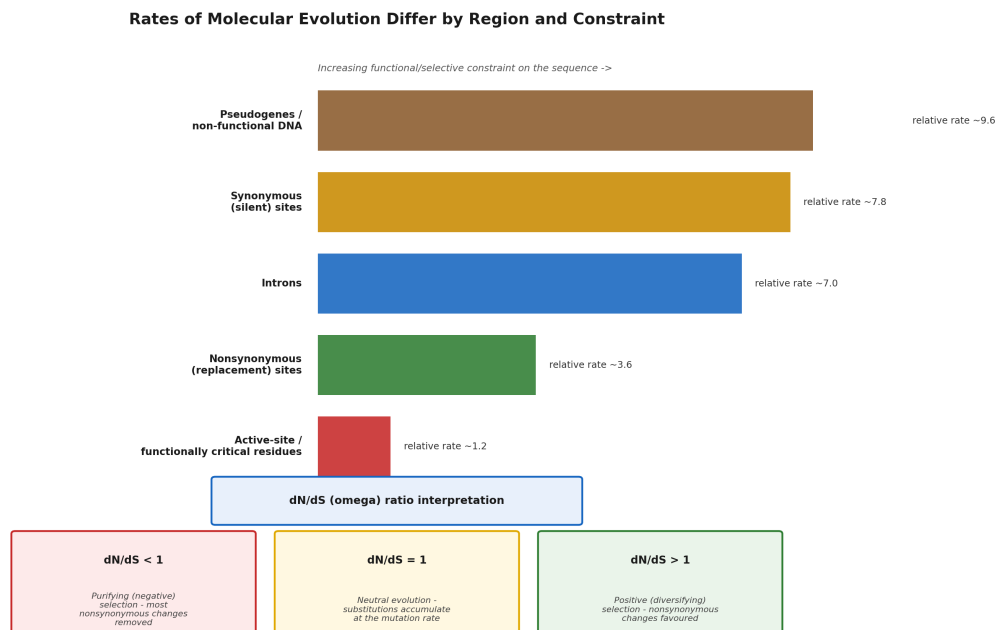


Fig. 4 - Substitution rates rise with decreasing functional constraint (top) and the standard dN/dS interpretation of selection pressure on protein-coding sequence (bottom).

The ratio of non-synonymous to synonymous substitution rates, **dN/dS (also written omega)**, is therefore a widely used measure of the type and strength of selection acting on a gene: **dN/dS < 1** indicates purifying (negative) selection removing most amino-acid-changing mutations, **dN/dS = 1** is consistent with neutral evolution, and **dN/dS > 1** indicates positive (diversifying) selection actively favouring amino acid change.

3.4 Factors causing rate variation

- **Functional constraint:** regions critical to a molecule's structure or function (e.g. enzyme active sites, histone proteins) evolve slowly because most changes are deleterious and removed by purifying selection; non-functional DNA (e.g. pseudogenes) evolves fastest because it is essentially unconstrained.
- **Generation-time effect:** since most mutations arise from errors during DNA replication, species with shorter generation times (more germline replications per year) tend to accumulate mutations, and therefore substitutions, faster than species with long generation times.
- **Population size and genetic drift:** in smaller populations, drift fixes mildly deleterious or nearly neutral variants more readily, which can alter the effective substitution rate relative to large populations where selection is more efficient.
- **DNA repair efficiency and metabolic rate:** lineages differ in baseline mutation rate due to differences in replication fidelity, repair mechanisms, and metabolic (oxidative damage) rate.

Key points

- "Rate of molecular evolution" is not one number - it varies by gene, by region within a gene (synonymous vs non-synonymous, structural vs flexible regions), and by lineage, which is exactly why a strict molecular clock often needs to be "relaxed" in real analyses.
- dN/dS is a per-gene or per-site average; a gene under overall purifying selection can still contain a handful of individual positively selected sites (detectable with site-specific dN/dS models).

4. Quick Revision Summary

Concept	Essence
Molecules as documents	Zuckerlandl & Pauling (1965): sequences record evolutionary history, readable like a document
Orthologs	Homologs arising by speciation; best markers for species relationships
Paralogs	Homologs arising by gene duplication within a genome; can mislead species-tree inference
Multiple sequence alignment	Aligns homologous sites into columns before any tree- building method is applied
UPGMA / Neighbor-Joining	Distance-based tree methods; fast but less robust than character-based methods
Maximum Parsimony	Tree needing fewest evolutionary changes; can be misled by long-branch attraction
Maximum Likelihood / Bayesian	Model-based methods; statistically robust, computationally intensive
Bootstrap value	% of resampled replicate trees supporting a branch - a confidence measure
Molecular clock hypothesis	Substitutions accumulate roughly constantly over time, enabling divergence-time estimation
Neutral theory (Kimura)	Most molecular substitutions are neutral, fixed by genetic drift, not selection
dN/dS (omega)	<1 purifying selection, =1 neutral, >1 positive selection
Rate-affecting factors	Functional constraint, generation time, population size, DNA repair/metabolic rate

Notes prepared for study purposes. Consult current primary literature for details required in formal citation or research use.