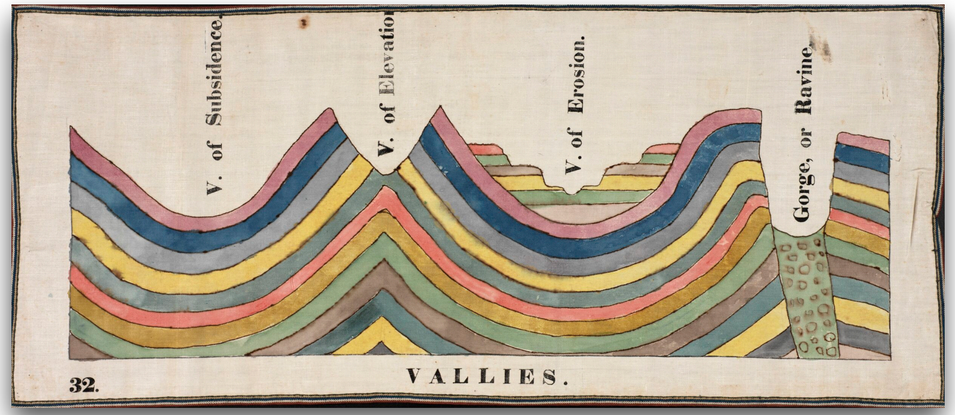

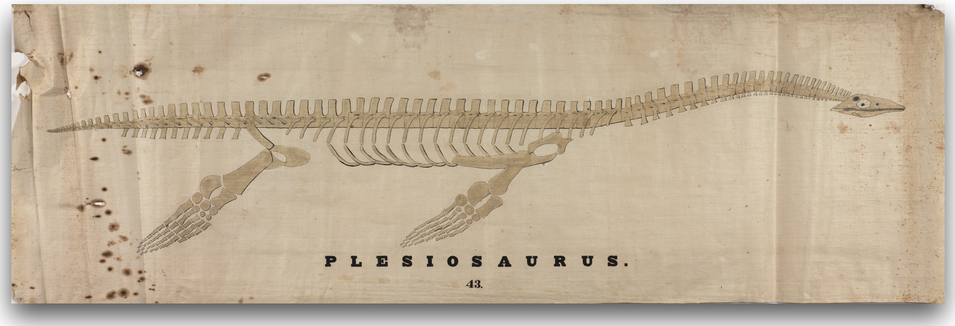

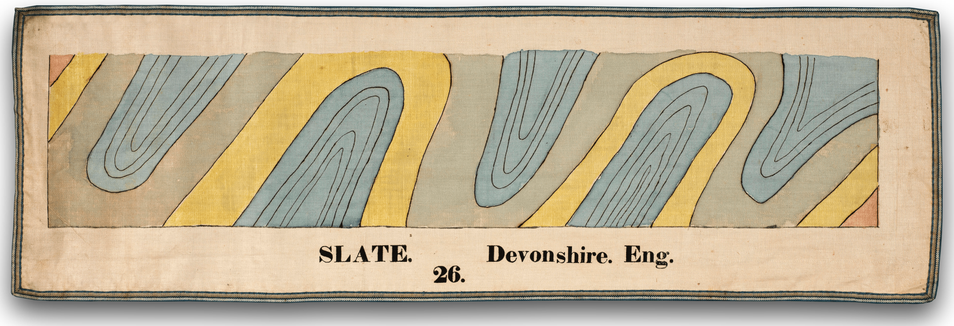

Inferring selection

Phylogenetics aims to reconstruct patterns of **shared ancestry** among different organisms …

… while phylogenetic trees are themselves the result of these inferences, they can also be powerful **tools for studying evolutionary processes**.

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**.

Given that in the universal genetic code **there are 61 codons for a specific amino acid**, and three are stop codons, **but only 20 amino acids**.

Most amino acids are encoded by more than one codon.

The degeneracy of the genetic code allows two types of substitutions:

Nonsynonymous substitutions:

- Do **change** the amino acid sequence of a protein.
- Their rate (*dN*) is expected to reflect **selective pressure** on protein.

Synonymous substitutions:

- Do **not change** the amino acid sequence of a protein.
- Their rate (*dS*) is assumed to be **neutral** with respect to selection.

These assumptions - especially about *dS* neutrality - are **widely debated**!

Nonetheless, **codon models** provide insight into selective forces.

Muse and Gaut (MG94)

In 1994, Muse and Gaut introduced a codon substitution model that has **different rates for synonymous versus nonsynonymous substitutions**.

In codon models, two parameters are used:

- ratio of the number of syn. substitutions per syn. site (**dS**).
- ratio of the number of nonsyn. substitutions per nonsyn. site (**dN**).

This allows for the estimation of the absolute rates of each type, reflecting selective constraints on protein-coding genes.

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** - it includes a transition/transversion bias κ - and **codon frequency parameters**- using empirical 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.

More complex codon models to detect selection followed shortly after:

- **site-specific models** let dNdS varies across codons
- **branch-specific models** let dNdS vary across lineages
- **branch-site models** let dNdS vary across lineages and codons

- **site-specific models** let dNdS vary across codons … dNdS values are averaged throughout all branches of our phylogeny … but we know that selection can differ between different lineages!

- **branch-specific models** let dNdS vary across lineages … dNdS values are averaged across all codons of our alignment … but we know that selection can differ widely between different gene!

- **branch-site models** relax at the same time the previous assumptions to detect positive selection on specific sites and lineages …

…some **jargon** in codon models ,,,

Foreground: different class(es) of branches of the phylogeny that are allowed to have a different dN/dS - whether higher or lower - than the rest.

Background: all other branches not under explicit testing and that provide the baseline for comparison with the foreground.

Purifying Selection (ω < 1) also known as negative selection, acts to eliminate harmful mutations that alter protein function. It preserves the integrity of important genes by discouraging changes at the amino acid level. This is the most common form of selection in coding regions, especially in genes essential for basic cellular processes.

Neutral Evolution (ω = 1) occurs when mutations neither benefit nor harm the organism—they simply drift through populations over time. In these cases, synonymous and nonsynonymous changes accumulate at similar rates. This process is expected at sites where function is not tightly constrained, such as in pseudogenes or certain non-coding regions.

Positive Selection (ω > 1) promotes mutations that offer an advantage to the organism. These changes are often fixed rapidly in populations and can lead to new or enhanced functions. It's a hallmark of evolutionary innovation, commonly observed in genes involved in immune defense, sensory perception, or reproductive function. An elevated dN/dS ratio is a strong indicator that positive selection may be at work.

Relaxed Selection (Δω > 0 between lineages) isn't defined by an absolute ω value, but by a comparison between two evolutionary contexts. If a gene or lineage previously under strong purifying selection (low ω) experiences reduced functional constraint, its ω may increase — not necessarily above 1, but higher than before. It reflects weaker purifying selection, not active adaptation and os detected by comparing ω across branches and gens.

A **dNdS > 1** is considered **unrealistic** when inferred for whole gene.

Such a value would suggest that all codons are accumulating **twice as many nonsynonymous mutations** as synonymous ones.

This signal is rarely linked to genuine positive selection, even when associated with specific branches.

These extreme values point to issues such as **pseudogenization** or **errors** in alignment and orthology, which can impact **downstream analyses**.

This aspect can be leveraged for **phylogenetic subsampling** and dNdS can be useful for finding problematic regions, sequences or whole genes.

codon-based alignments

A type of multiple sequence alignment where **nucleotide sequences are aligned in sets of three nucleotides** (codons).

Insertions or deletions - **indels** - must involve entire codons - **multiples of three nucleotides** - to avoid disrupting the reading frame.

Ensures that the reading frame is preserved across all sequences, **preventing frame-shift mutations** that are not biologically plausible.

… fundamental for codon models!

FINISH