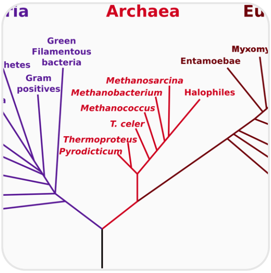
ChatGPT

Measuring Evolutionary Closeness with Genetic Data

Scientists infer “closeness” by comparing DNA or protein sequences from different species. Historically, this was done one gene at a time (e.g. a single ribosomal RNA gene), but modern studies often use many genes or whole genomes nature.com nature.com . For example, multi-locus approaches (like MLST) compare a handful of conserved genes across species, while **phylogenomics** uses hundreds of single-copy genes or genome-wide SNPs to build trees nature.com nature.com . In practice:

- **Whole genome vs markers:** When full genomes are available, researchers may use genome-wide SNPs or thousands of orthologous genes to compute distances nature.com . But for speed or limited data, they still often use a few standard “marker” genes (see below) nature.com nature.com .

Fig: A ribosomal RNA-based phylogenetic tree of life (Carl Woese 1990) showing Bacteria, Archaea and Eukarya. Ribosomal RNA genes (16S/18S) are highly conserved markers that reveal deep relationships *en.wikipedia.org* *en.wikipedia.org* .



Common Genetic Markers

Most phylogenetic analyses focus on genes that are universally present and evolve at appropriate rates. **Highly conserved genes** give reliable signals across distant groups, while faster-evolving regions help with close relatives. Common choices include:

- **Ribosomal RNA genes:** The 16S rRNA gene (in bacteria/archaea) and 18S (in eukaryotes) are classic universal markers en.wikipedia.org en.wikipedia.org . They change slowly and are easy to align across broad taxonomic groups. (The figure above shows a tree built from rRNA data en.wikipedia.org .)
- **Mitochondrial DNA:** In animals, mitochondrial genes (like **COI**, cytochrome *b*, 12S/16S rRNA) are popular because they mutate relatively quickly and are abundant in cells. (Typical animal mtDNA carries 16S and 12S rRNAs plus protein genes such as **COI** pmc.ncbi.nlm.nih.gov .) Plant phylogenies often use chloroplast genes (e.g. *rbcL*, *matK*).
- **Conserved protein-coding genes:** Housekeeping genes (like RNA polymerase subunits, translation factors, tubulins, etc.) are also used. For deep analyses, researchers may select dozens or hundreds of single-copy orthologs that exist in all species of interest pmc.ncbi.nlm.nih.gov .
- **Nuclear ribosomal spacers:** In fungi and plants, the internal transcribed spacer (ITS) regions between rRNA genes evolve rapidly and serve as species-level barcodes (similar to COI in animals).

By combining markers (multi-gene or genome-wide alignments), one can improve accuracy. For example, a "supermatrix" of many genes or concatenated alignments is common in modern studies nature.com nature.com .

Phylogenetic Tree Construction Methods

Once sequences are chosen and aligned, computational methods infer the tree. There are two broad approaches pmc.ncbi.nlm.nih.gov pmc.ncbi.nlm.nih.gov : distance-based and character-based. Representative methods include:

- **Distance-based (e.g. Neighbor-Joining, UPGMA):** Compute pairwise genetic distances (e.g. percent differences or model-corrected distances) and cluster them. Neighbor-Joining (NJ) is very popular – it builds an unrooted tree by iteratively joining nearest neighbors until one tree is formed pmc.ncbi.nlm.nih.gov pmc.ncbi.nlm.nih.gov . (UPGMA is simpler but assumes a strict molecular clock.)

- **Maximum Parsimony (MP):** Searches for the tree topology that requires the fewest evolutionary changes overall. It simply counts how many substitutions explain the data, without assuming a specific model of sequence evolution pmc.ncbi.nlm.nih.gov pmc.ncbi.nlm.nih.gov . MP is less used now for DNA data but is conceptually simple.
- **Maximum Likelihood (ML):** Uses explicit statistical models (see below) to find the tree that maximizes the probability (likelihood) of observing the given sequences pmc.ncbi.nlm.nih.gov pmc.ncbi.nlm.nih.gov . ML methods consider branch lengths and allow different rates at sites or branches, yielding more accurate trees especially for more divergent data. Software like RAxML, IQ-TREE or PhyML implements ML tree search.
- **Bayesian Inference:** Similar to ML but uses Bayesian statistics and Markov Chain Monte Carlo (MCMC) to sample tree space under a given model. It produces a set of trees from the posterior distribution, from which a consensus can be made pmc.ncbi.nlm.nih.gov pmc.ncbi.nlm.nih.gov . Tools like MrBayes or BEAST use this approach, which also incorporates prior information and gives credibility values on clades.
- **Other methods:** (Not requested but often mentioned: “neighbor-net” or “split networks” for non-tree structures, and coalescent-based “species tree” methods that account for gene tree discordance.)

Each method has trade-offs: distance methods are fast and handle large data well pmc.ncbi.nlm.nih.gov , while ML/Bayesian are more computationally intensive but statistically powerful pmc.ncbi.nlm.nih.gov . The table in pmc.ncbi.nlm.nih.gov summarizes these approaches.

Genetic Distance and Substitution Models

A key step is computing *genetic distances* between sequences. The simplest distance (p-distance) is just the fraction of differing sites, but this underestimates true changes when sequences diverge because multiple substitutions can hit the same site. **Substitution models** correct for this by modeling how bases change over time. Examples:

- **Jukes–Cantor (JC69) model:** Assumes equal probability for all substitutions (any base → any other) and equal base frequencies megasoftware.net . It provides a simple correction formula: if p is the observed difference fraction, the corrected distance is $-\frac{3}{4}\ln(1-\tfrac{4}{3}p)$. This adjusts for unseen multiple hits under the "all rates equal" assumption megasoftware.net .
- **Kimura two-parameter (K80) model:** Distinguishes transitions (purine↔purine A↔G or pyrimidine↔pyrimidine C↔T) from transversions (purine↔pyrimidine) en.wikipedia.org . Transitions often occur more frequently, so K80 uses one rate for transitions and another for transversions. The formula (shown in en.wikipedia.org) uses two proportions p (transitions) and q (transversions) to get a distance.
- **More complex models:** Other models allow unequal base frequencies (HKY85), separate transversion rates (Tamura–Nei), or the General Time Reversible (GTR) model which has a rate for each possible substitution en.wikipedia.org . In practice, software selects an appropriate model (e.g. by likelihood tests).

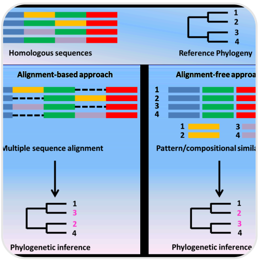
In sum, distances are computed by converting sequence differences into expected numbers of substitutions under a model. These corrected distances can then be used directly (in distance methods) or plugged into likelihood calculations (for ML/Bayesian).

Cutting-Edge Approaches

With rapidly dropping sequencing costs, evolutionary genomics is embracing new methods:

- **Whole-genome phylogeny:** Instead of a few genes, researchers now use *phylogenomics* – comparing entire genomes or large SNP datasets. For example, tens of thousands of SNPs across bacterial genomes can be used to build high-resolution trees nature.com . Genome-wide data capture more variation (neutral SNPs accumulate uniformly) and can resolve very close or very deep relationships nature.com . (Tools like PhyloPhlAn or large-species trees often concatenate hundreds of genes.)

- **Alignment-free methods:** Traditional phylogenetics relies on sequence alignment, which can be slow or ambiguous for very large genomes. Alignment-free approaches instead compare sequences by k-mer (substring) frequencies or other compositional features. For example, methods compute distances from patterns of k-mers in each genome and then build NJ trees

  . These methods can quickly handle thousands of genomes or highly divergent sequences where alignment is hard. (The figure above contrasts alignment-based vs alignment-free phylogeny.)

- **Machine learning (ML) and AI:** Emerging studies apply ML to phylogenetics. For instance, deep learning models (like convolutional neural networks) have been proposed to classify sequences or predict evolutionary relationships using large training data bmcbioinformatics.biomedcentral.com . One example (Ph-CNN) embeds phylogenetic distances in a CNN for microbiome classification bmcbioinformatics.biomedcentral.com . Other work uses neural nets to choose the best substitution model or even directly infer small trees. While not yet mainstream, ML holds promise for handling huge genomic datasets and complex models.

- **Other technologies:** Real-time sequencing and metagenomics allow phylogenetic placement of unknown samples on a reference tree (e.g. using tools like pplacer or EPA). Single-cell genomics and ancient DNA are also expanding the range of species we can analyze.

In summary, measuring evolutionary closeness is done by aligning homologous sequences (whole genomes or selected genes), computing genetic distances using substitution models, and then building trees with methods like neighbor-joining, maximum likelihood or Bayesian inference pmc.ncbi.nlm.nih.gov pmc.ncbi.nlm.nih.gov . Modern genomics adds more data and even AI techniques to refine these relationships, giving ever-more detailed pictures of how species are related.

Sources: Authoritative reviews and textbooks on phylogenetics pmc.ncbi.nlm.nih.gov pmc.ncbi.nlm.nih.gov nature.com nature.com ; classic works on substitution models megasoftware.net en.wikipedia.org ; and example studies on phylogenomics and ML applications en.wikipedia.org en.wikipedia.org bmcbioinformatics.biomedcentral.com .

Citations

Standardized phylogenetic and molecular evolutionary analysis applied to sp...
https://www.nature.com/articles/s41598-020-58356-1

Standardized phylogenetic and molecular evolutionary analysis applied to sp...
https://www.nature.com/articles/s41598-020-58356-1

16S ribosomal RNA - Wikipedia
https://en.wikipedia.org/wiki/16S_ribosomal_RNA

18S ribosomal RNA - Wikipedia
https://en.wikipedia.org/wiki/18S_ribosomal_RNA

General properties and phylogenetic utilities of nuclear ribosomal DNA and ...
https://pmc.ncbi.nlm.nih.gov/articles/PMC2733198/

Phylogenetic Resolution of Deep Eukaryotic and Fungal Relationships Using ...
https://pmc.ncbi.nlm.nih.gov/articles/PMC5631032/

Common Methods for Phylogenetic Tree Construction and Their Implementa...
https://pmc.ncbi.nlm.nih.gov/articles/PMC11117635/

Common Methods for Phylogenetic Tree Construction and Their Implementa...
https://pmc.ncbi.nlm.nih.gov/articles/PMC11117635/

Common Methods for Phylogenetic Tree Construction and Their Implementa...
https://pmc.ncbi.nlm.nih.gov/articles/PMC11117635/

Common Methods for Phylogenetic Tree Construction and Their Implementa...

https://pmc.ncbi.nlm.nih.gov/articles/PMC11117635/

Common Methods for Phylogenetic Tree Construction and Their Implementa...
https://pmc.ncbi.nlm.nih.gov/articles/PMC11117635/

Common Methods for Phylogenetic Tree Construction and Their Implementa...
https://pmc.ncbi.nlm.nih.gov/articles/PMC11117635/

Common Methods for Phylogenetic Tree Construction and Their Implementa...
https://pmc.ncbi.nlm.nih.gov/articles/PMC11117635/

Common Methods for Phylogenetic Tree Construction and Their Implementa...
https://pmc.ncbi.nlm.nih.gov/articles/PMC11117635/

Jukes-Cantor distance
https://www.megasoftware.net/web_help_7/hc_jukes_cantor_distance.htm

W **Models of DNA evolution - Wikipedia**
https://en.wikipedia.org/wiki/Models_of_DNA_evolution

W **Models of DNA evolution - Wikipedia**
https://en.wikipedia.org/wiki/Models_of_DNA_evolution

Phylogenetic convolutional neural networks in metagenomics | BMC Bioinfor...
https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2033-5

All Sources

nature W en.wikipedia pmc.ncbi.nlm.nih megasoftware bmcbioin...edcentral