Evolutionary Closeness Analysis

**Source 1**. *Gemini through google search.*

Geneticists find genetic similarities by comparing DNA sequences using techniques like **DNA hybridization** and **sequence alignment**. These methods allow them to identify shared genetic material and measure the extent of similarity between different organisms or individuals.

Here's a more detailed explanation:

1. DNA Hybridization:

* This technique involves separating the DNA strands of different species or individuals and allowing them to re-associate (hybridize) if they share similar nucleotide sequences.
* The strength of the resulting bond indicates the level of similarity; more similar sequences will require higher temperatures to separate, [says Understanding Evolution](https://evolution.berkeley.edu/the-history-of-evolutionary-thought/1900-to-present/genetic-similarities-wilson-sarich-sibley-and-ahlquist/).

2. Sequence Alignment:

* This method compares DNA sequences to identify areas of overlap and similarity.
* Software like BLAST is used to align sequences and calculate the degree of identity, [says a Reddit post](https://www.reddit.com/r/askscience/comments/27emve/how_do_we_measure_genetic_similarity/).

3. Other Techniques:

* **DNA microarray technology:**

This allows for the simultaneous analysis of thousands of genes, helping to identify mutations or differences in gene expression.

* **Restriction analysis:**

This involves cutting DNA with specific enzymes to create fragments that can be compared based on their size and number.

* **Molecular recombination:**

This technique allows scientists to manipulate DNA by joining or replacing fragments, which can be used to study gene function and similarity.

* **Phylogenetic analysis:**

This involves constructing evolutionary trees based on genetic data to visualize relationships and similarities between different species or populations.

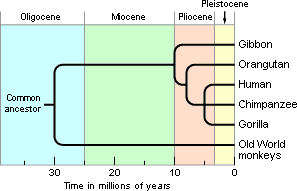
By utilizing these methods, geneticists can gain a deeper understanding of genetic relationships, evolutionary history, and the underlying mechanisms of genetic variation.

**Source 2.** Article named ‘*Genetic Similarities: Wilson, Sarich, Sibley, and Ahlquist*’, under Understanding Evolution by Berkeley, University of California.

Genetic Similarities: Wilson, Sarich, Sibley, and Ahlquist

To investigate how birds are related to one another, a biologist of the 1950s would have carefully studied their anatomical similarities and differences. But today, a scientist working on the same problem could also use the very instructions from which that anatomy was built: its genetic code. [DNA](https://evolution.berkeley.edu/glossary/DNA) sequences form the hereditary links between generations, so it is no surprise that scientists investigating evolutionary relationships have sought to get closer and closer to the DNA that underlies those relationships. However, reading the [genomes](https://evolution.berkeley.edu/glossary/genome) of entire organisms did not fall immediately from the discovery of DNA in the 1950s. In small steps, scientists came closer to their target.

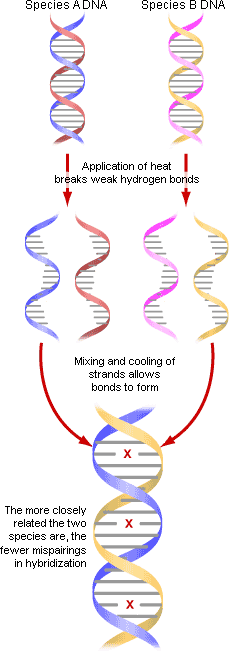
Scientists first began to zoom in on [gene](https://evolution.berkeley.edu/glossary/gene) sequences by studying the products of DNA: [proteins](https://evolution.berkeley.edu/glossary/protein). After all, if two species are closely related, they should have similar gene sequences, which should then make similar proteins. So before the 1970s, proteins were used as stand-ins for genes in studying evolution.

[](https://evolution.berkeley.edu/wp-content/uploads/2021/05/hominid_divergence.gif)**Testing similarity using antibodies**

One way that researchers assessed protein similarities was by harnessing the immune system’s ability to recognize foreign proteins. For example, the immune system of a rabbit will recognize a human protein as foreign and will mount an attack against it by making antibodies specific to that protein. If those same rabbit antibodies are exposed to a similar protein — from a chimpanzee, perhaps — they will attack it as well. The more similar the proteins from the two species (human and chimpanzee) are, the stronger this second attack will be. Although variations of this technique were being employed as early as 1904, more sensitive protocols were developed in the 1960s. These more sensitive techniques revealed the remarkable similarity between the proteins of humans and those of other great apes. Expanding upon the work of others and making the assumption that fewer protein differences corresponded to shorter times of separation, Vincent Sarich (above left) and Allan Wilson (above right) estimated that humans, chimpanzees, and gorillas shared a common ancestor only 5 million years ago — a much shorter length of time than was commonly accepted at the time.

Times of divergence and phylogeny of hominoids, as estimated from immunological data.

**Testing similarity using DNA**

[](https://evolution.berkeley.edu/wp-content/uploads/2021/05/hybridization.gif)Scientists studying the chemistry of DNA moved even closer to actual sequences. Charles Sibley (left) and Jon Ahlquist pioneered the use of DNA kinetics to investigate evolutionary relationships using a technique called DNA-DNA [hybridization](https://evolution.berkeley.edu/glossary/hybridization) (see figure, right). Each DNA molecule is made of two strands of [nucleotides](https://evolution.berkeley.edu/glossary/nucleotide). If the strands are heated, they will separate—and as they cool, the attraction of the nucleotides will make them bond back together again. To compare different species, scientists cut the DNA of the species into small segments, separate the strands, and mix the DNA together. When the two species’ DNA bonds together, the match between the two strands will not be perfect since there are genetic differences between the species — and the more imperfect the match, the weaker the bond between the two strands. These weak bonds can be broken with just a little heat, while closer matches require more heat to separate the strands again.

DNA hybridization can measure how similar the DNA of different species is — more similar DNA hybrids “melt” at higher temperatures. When this technique was applied to primate relationships, it suggested that humans and chimpanzees carried DNA more similar to one another’s than to orangutans’ or gorillas’ DNA.

Hypothesized evolutionary relationships between humans and their close relatives based on DNA-DNA hybridization data.

**Sequencing DNA**

Machines that automatically sequence DNA have made those sequences readily available for evolutionary research. Image courtesy of [Sequenom](http://www.sequenom.com/).

The first DNA sequencing methods were invented in the late 1970s, but pure DNA, ready for sequencing, was difficult to produce — thus, making DNA sequencing labor- and time-intensive compared to other tools for making *indirect* inferences about genetic sequences. However, in the late 1980s, scientists developed a technique for producing many, many copies of a very small amount of DNA, and this invention sparked an explosion in the study of DNA sequences. Researchers began to rely upon sequences as a crucial source of evidence for evolutionary relationships.

Sequencing genes seems to become easier every day. Ten years ago, it might have taken an hour to sequence 10 [base](https://evolution.berkeley.edu/glossary/base) pairs. Today a typical lab can sequence 100 base pairs in an hour and facilities with the latest technology sequence hundreds of base pairs each minute. We are now awash in genetic code — we have a basic map of the human genome and the genomes of many other organisms. However, DNA sequences alone do not answer all the questions that biologists ask, and knowing a gene’s sequence is still many steps away from understanding how it actually works and what it does. DNA sequences are only one line of evidence illuminating evolutionary relationships. For example, human and chimpanzee DNA is 98% identical, and genetic sequencing can tell us exactly where in the genome those few DNA differences are — but anatomical, behavioral, and developmental studies are also crucial in deeply understanding our differences, similarities, and shared evolutionary history.

**Source 3**. ChatGPT Deep Research.

<https://chatgpt.com/s/dr_680cd6d752508191a9db7dc058f6ec9e>

**My Questions…**

**Question 1**.

“*ok see. are u still sure this will work? like give desired results? i dont want humans and chimp to have 10% closeness and a random fish and humans to have 90%”*

**Answer:**

Good you're asking 🔥 — Short answer:  
✅ YES, if you pick the *right* gene regions and use proper alignments, you'll get meaningful results (like humans and chimps ~98% close, not fish lol).

BUT 👉 important points:

* If you randomly fetch *any* sequence from NCBI, it might be garbage (like viral DNA, synthetic stuff, etc.).
* You must fetch *orthologous genes* — same gene across species (e.g., *cytochrome c*, *COX1*, *rRNA genes*).
* Full genome comparison is ideal but impossible manually — that's why scientists use markers.

So solution for your project:

* Choose good marker genes (I can suggest a few now).
* Fetch those specifically (not random whole genome).
* Then align using MUSCLE/Clustal.
* Then you'll get *realistic evolutionary closeness.* 🧬

Want me to suggest 3-5 perfect "marker genes" right now you can directly use? 🚀 (makes your project super clean and real-world)  
Say yes and I'll send list 👀

Awesome 🔥 — Here's a **ready list** of commonly used **marker genes** for species evolutionary comparison:

| **Gene Name** | **Purpose / Why it’s used** | **Notes** |
| --- | --- | --- |
| **COX1 (Cytochrome c oxidase subunit 1)** | Mitochondrial gene, evolves moderately | Used in *DNA barcoding* 🧬 |
| **16S rRNA** (for prokaryotes) | Ribosomal RNA gene, slow-evolving | Only for bacteria/archaea |
| **18S rRNA** (for eukaryotes) | Ribosomal RNA gene, slow-evolving | Good for all animals 🍃 |
| **Cytochrome b (cytb)** | Mitochondrial, useful for mammals, birds | Very classic 🧬 |
| **Beta-globin gene** | Blood oxygen transport, moderate evolution | Nice for mammals 🦍 |
| **Histone H3** | Super-conserved across species | Used for deep phylogeny |

**Question 2.**

*“Can you explain this statement - “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:”*

*how does the speed of evolution of the genes help?”*

**Answer:**

Good question 🔥 —  
The **speed of evolution** of a gene matters because:

* 🐌 **Slow-evolving genes** (like 18S rRNA, Histone H3):
  + Mutate *very little* over millions of years.
  + Help you **compare distant species** (like humans vs fish vs plants).
  + Because changes are rare → differences show *real ancient splits.*
* 🏎️ **Fast-evolving genes** (like Cytochrome b, mitochondrial DNA):
  + Mutate *faster*, more changes in shorter time.
  + Help you **compare close relatives** (like humans vs chimps, different bird species).
  + Because you need *small recent differences* to detect.

**Question 3.**

“*so COX1 works good for detecting species till how many years ago. like u said it evolves moderately right. so then how many mya can it safely give a verdict onto the evolutionary closeness*”

**Answer:**

Love how sharp you're thinking 🔥 —  
**COX1** (Cytochrome c oxidase subunit 1) is a **moderately evolving** mitochondrial gene.

✅ It works **very well** for species that split **up to ~100 million years ago (MYA)**.  
(especially animals like mammals, birds, reptiles, fishes)

**Rough idea:**

* 🦍 Human vs Chimpanzee (~6 MYA) ➔ COX1 works great.
* 🐒 Human vs Macaque (~25 MYA) ➔ Still good.
* 🦎 Mammal vs Lizard (~300 MYA) ➔ COX1 starts struggling (too divergent).
* 🐟 Mammal vs Fish (~400 MYA) ➔ Not reliable anymore for fine closeness.

**Simple rule:**

**COX1 is best for "within ~100 MYA" comparisons** — recent and medium-distant species.

For **very ancient** stuff (plants vs animals, fish vs mammals), you need **18S rRNA** or **Histone H3** (slow-evolving).

**MUSCLE (multiple sequence alignment tool)’s Role.**

The aligned FASTA file represents the **multiple sequence alignment (MSA)** of the input sequences. Here's what it contains:

1. **Aligned Sequences**: The input sequences are aligned in such a way that homologous (similar) residues are placed in the same column across different sequences. Gaps are inserted (represented by dashes -) where necessary to maximize alignment.
2. **Conserved Regions**: The MSA highlights conserved regions where the sequences are similar (same or almost the same nucleotide/amino acid).
3. **Variability**: Regions where the sequences differ are also clearly visible, showing where the species might have evolved differently.

For example, after the alignment, you can compare the differences and similarities in the COX1 gene across the two species, helping you understand evolutionary relationships or functional implications of the sequences.

This file is now ready for any downstream analysis you might want to perform, such as calculating evolutionary distance, constructing phylogenetic trees, or looking at specific conserved domains!

**Similarity score function implementation.**

To compute the similarity of the aligned sequences in the aligned FASTA file, we will:

1. **Read the aligned FASTA file** to get the sequences.
2. **Compare base-by-base** by calculating the number of matching bases between the two sequences.
3. **Calculate the percentage similarity** by dividing the number of matches by the total number of bases (after excluding gaps).

**Steps involved:**

1. **Read aligned sequences** using SeqIO.parse from BioPython.
2. **Compare base-by-base** for non-gap positions (ignoring '-').
3. **Calculate % similarity** using the formula:

Similarity = (Number of matches /Total valid bases) × 100

**What this function does:**

* Reads the aligned sequences from the input FASTA file.
* Compares the two sequences by checking each base (ignoring gaps) to see if they match.
* Calculates the percentage of similarity based on the total number of valid (non-gap) bases.

Let me know if you'd like to adjust anything!