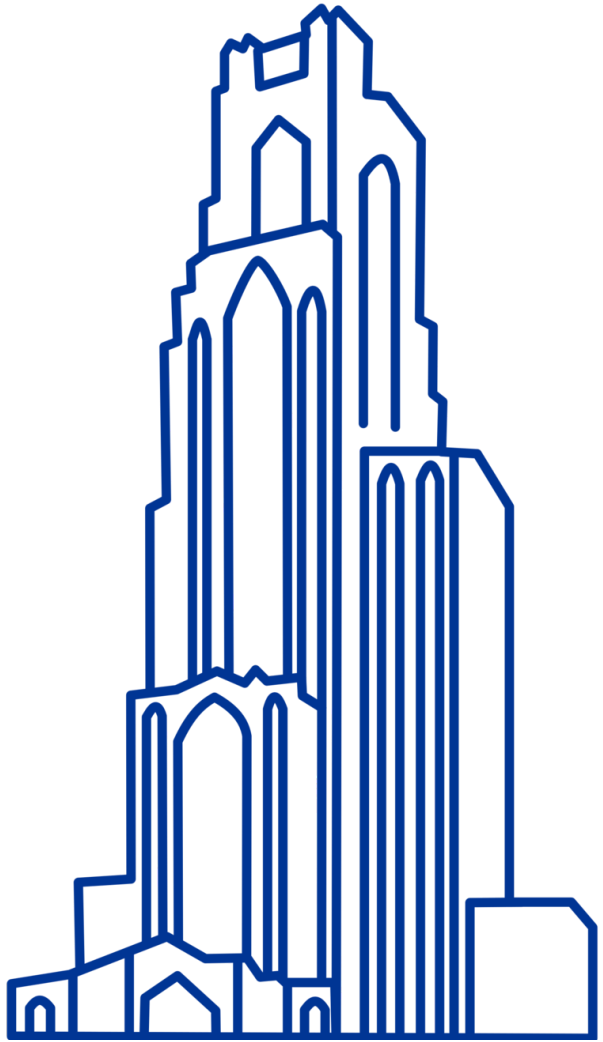


Computational Biology

(BIOSC 1540)

Lecture 05A

Sequence Alignment

Foundations

Feb 4, 2025

University of Pittsburgh

Announcements

Assignments

- Assignment P01D is due Monday (Feb 10)
- Assignment P01E will be released on Saturday (Feb 8)

Quizzes

- Quiz 02 is on Feb 18 and will cover lectures 04A to 06B

CBytes

- CByte 02 expires on Feb 7
- CByte 03 expires on Feb 15
- CByte 04 releases on Feb 8

Next reward: Checkpoint Submission Feedback

ATP until the next reward: 1,653

After today, you should have a better understanding of

Why sequence alignment matters

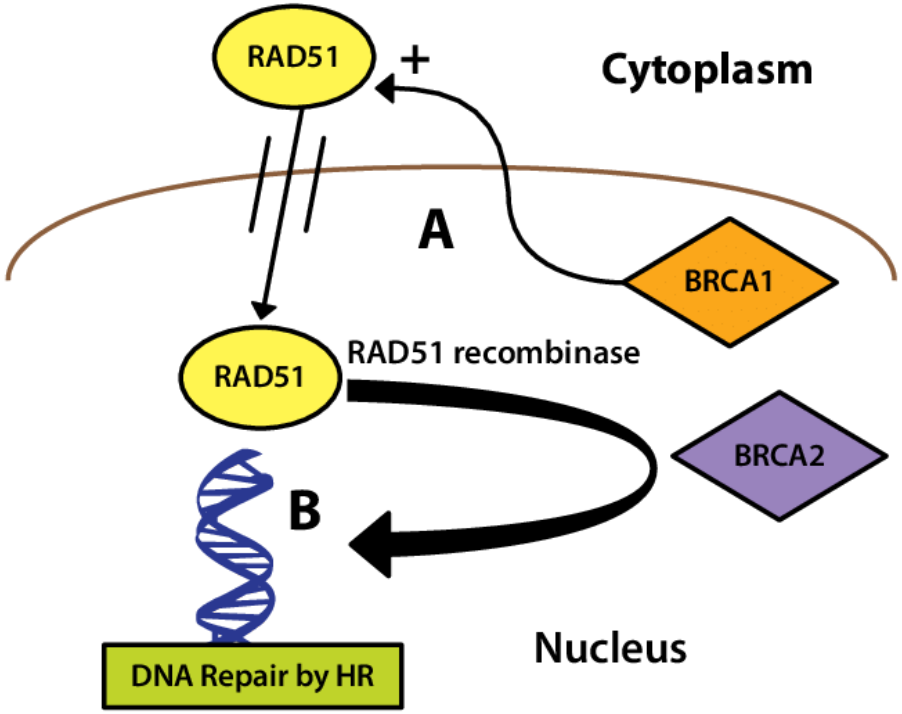
Homology

Homology describes the evolutionary relationship between sequences and is key to understanding biological function and evolution

Homologous sequences share a common ancestor, even if they have diverged over time.

Homology helps transfer knowledge from well-studied genes to newly discovered ones.

Example: The identification of BRCA1 as a breast cancer gene was based on identifying its association with RAD51, which function was known due to its high homology with yeast DNA repair

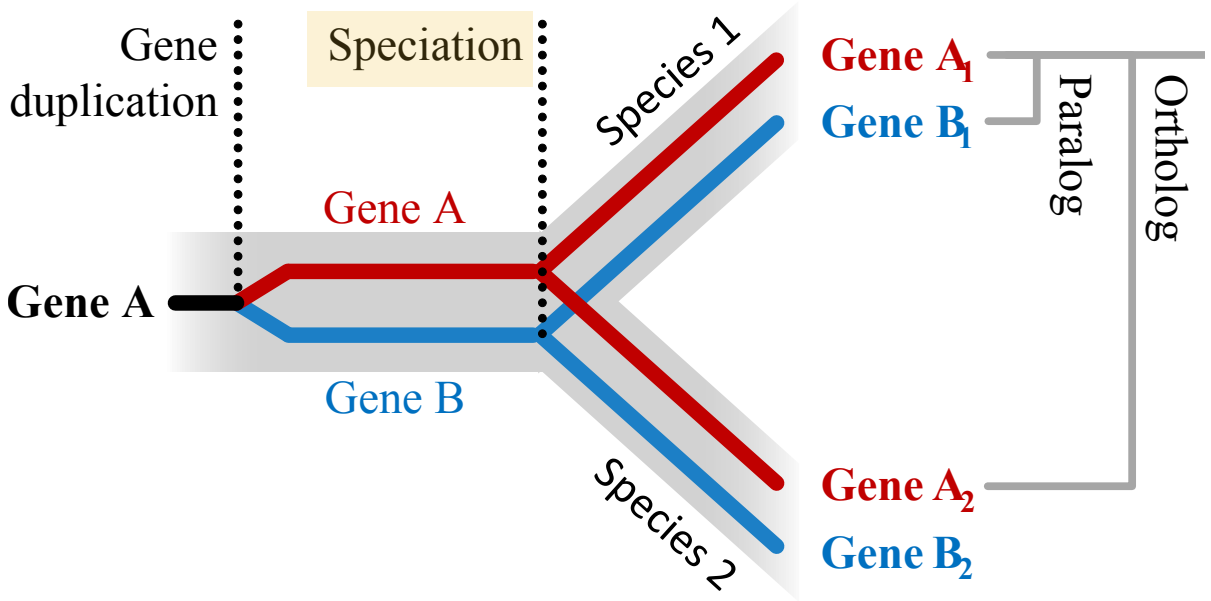


DOI: 10.1016/S0092-8674(00)81847-4

Orthologs are genes in different species that originated from a common ancestor and usually retain the same function

Orthologs arise from speciation events, meaning a single ancestral gene diverges into different species.

They typically perform the same function across species but may accumulate minor adaptations.

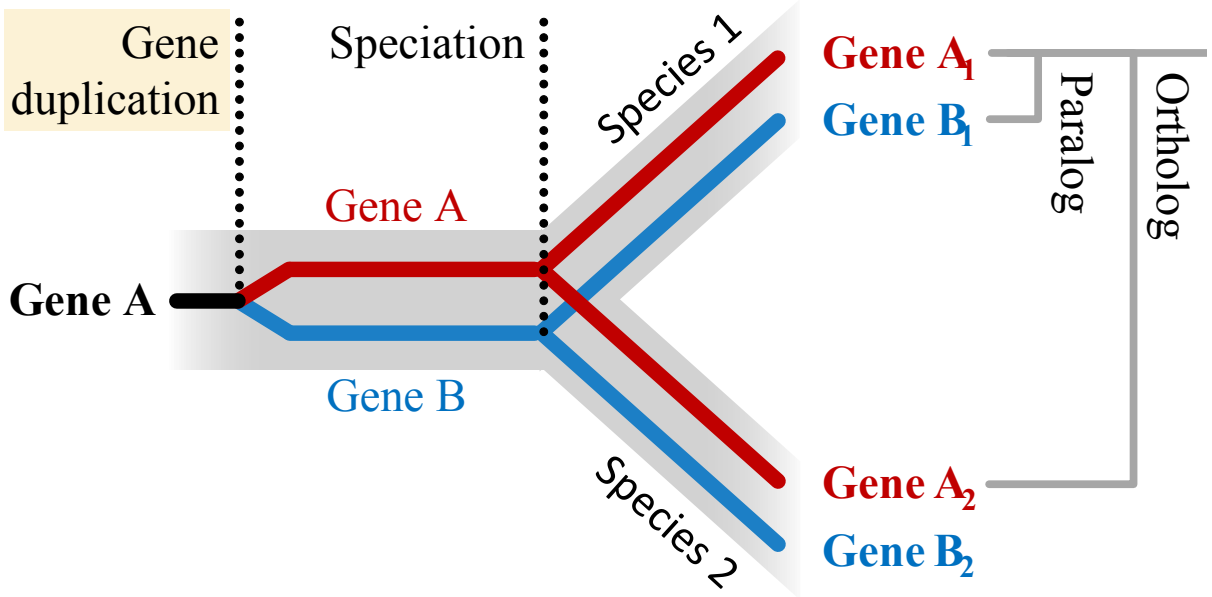


Example: The hemoglobin gene in humans and mice is orthologous, both encoding oxygen-carrying proteins in red blood cells.

Paralogs are genes that arise from duplication within the same genome and may evolve new functions

Paralogs originate from duplication events, which allows one copy to retain the original function while the other copy can

- Gain a new function (neofunctionalization).
- Specialize in a subset of the original function (subfunctionalization).
- Become a nonfunctional pseudogene.

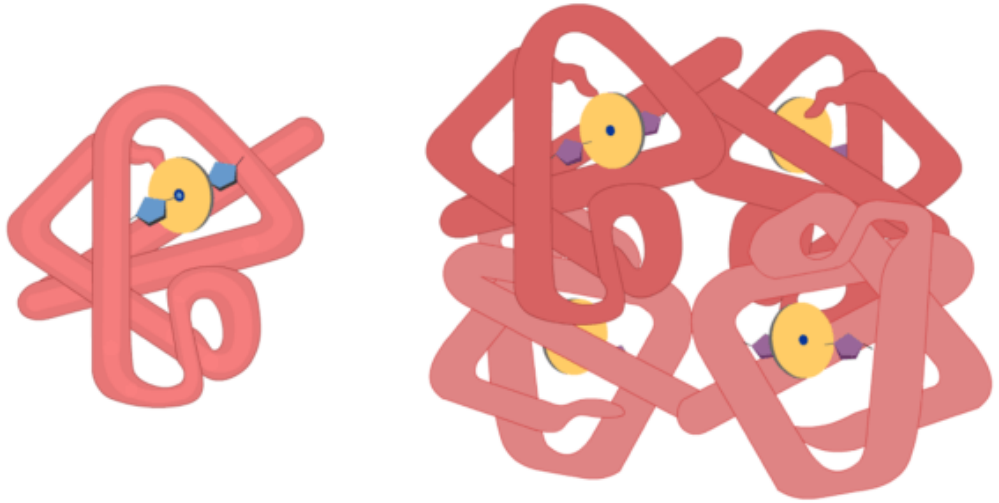


Paralogs drive gene family expansions, leading to specialized and diverse biological functions.

Gene duplication allows new biological functions to emerge while preserving essential roles

Examples of paralog-driven functional diversification:

- **Globin family**: Myoglobin (muscle oxygen storage) and hemoglobin (blood oxygen transport) evolved from a common ancestor.
- **HOX genes**: Regulate body plan development, with duplicates specializing in different body regions.
- **Opsin genes**: Responsible for color vision in vertebrates, arising from multiple duplications.



Myoglobin Hemoglobin

After today, you should have a better understanding of

Why sequence alignment matters

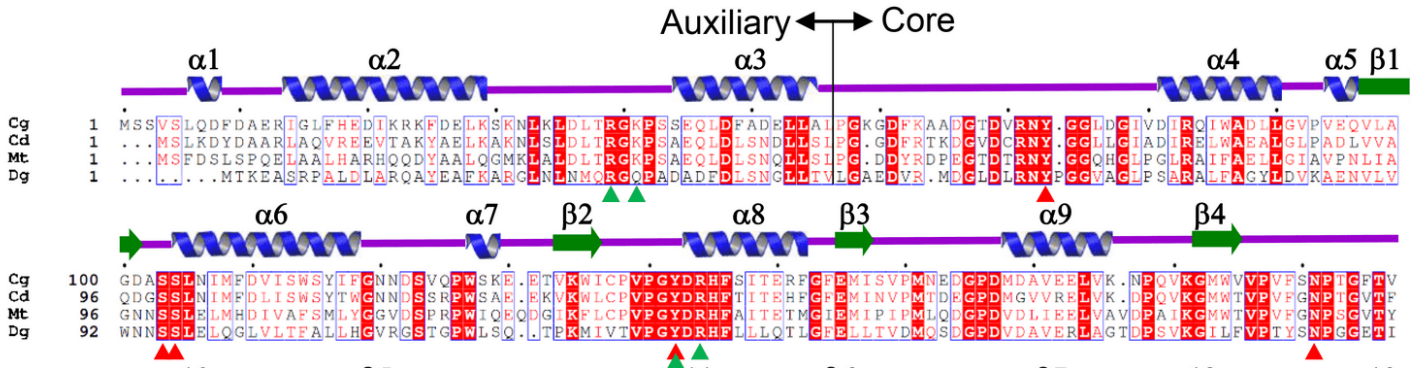
Homology applications

Functional annotation of genes and proteins relies on identifying homologous sequences

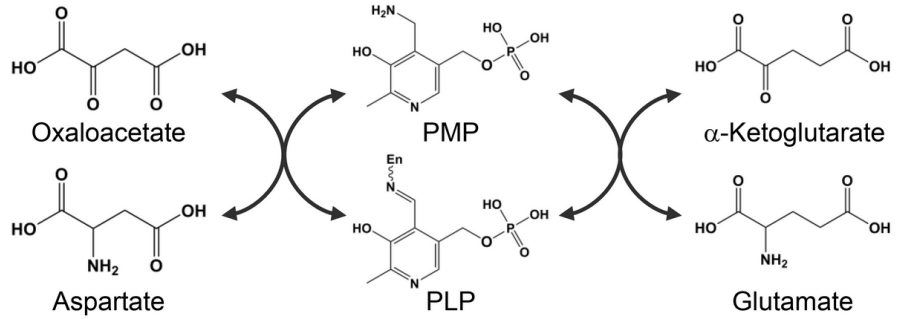
Conserved motifs suggests similar functional roles in different organisms.

For example, if a newly **discovered protein aligns with a known enzyme**, it likely shares the same biochemical function.

Homology-based searches (e.g., BLAST) rapidly annotate unknown sequences by comparing them to well-characterized databases.



Conserved residues that bind to PLP cofactor are shown with triangles



Aspartate aminotransferase (AspAT)

Source

Protein sequence homology predicts 3D structure and function

Highly conserved residues often indicate key structural or catalytic sites.

Structural homology models unknown proteins based on alignment with known structures.

Protein threading techniques align sequences to structural databases like the Protein Data Bank.

AlphaFold uses sequence alignments as inputs to its deep learning model

We will cover this in Lectures 11A and 11B

Homologous genes allow functional studies using model organisms to understand human biology

Importance in research:

- Used to **infer gene function across species** (e.g., using model organisms to study human genes).
- Enable **comparative genomics** and evolutionary studies.
- Help in **drug discovery**, as conserved drug targets can be tested in different organisms.
- **Knockout and mutation studies** in animals help determine gene function in humans.
- **Evolutionarily conserved pathways** (e.g., DNA repair, metabolism) can be studied using orthologs.

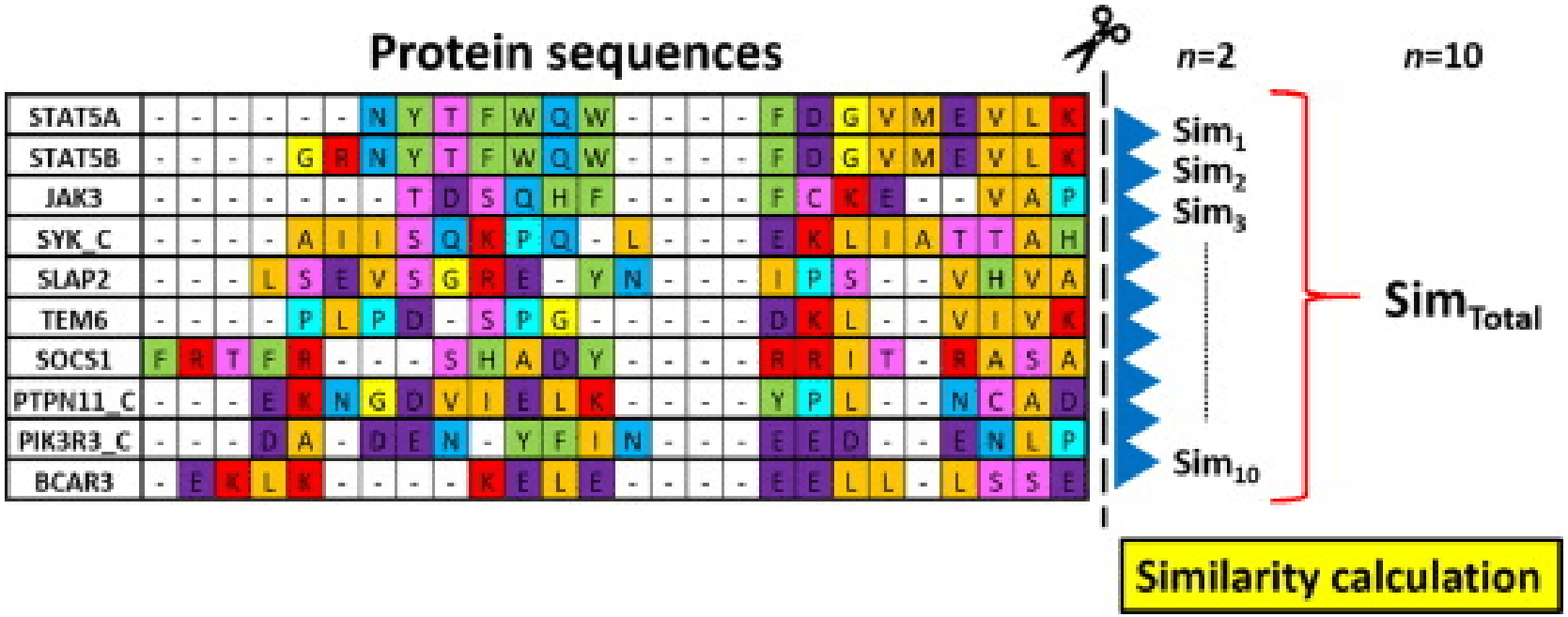
After today, you should have a better understanding of

Conceptual interpretation of alignment results

Match and mismatches

Alignment results provide insight into sequence similarity, evolutionary relationships, and functional conservation

Alignment patterns reflect evolutionary events, including mutations, conservation, and sequence divergence



Alignment algorithms compare sequences to identify conservation, mutations, and functional domains.

Identical residues at aligned positions suggest evolutionary conservation and functional stability

```
CGACGATTCTATAGTCTAACATGCGAGCGTGACGAATAAAAGATCTCGCG
||||||||||||||||||||||||||||||||||||||||||||||||||
CGACGATTCTATAGTCTAACATGCGAGCGTGACGAATAAAAGATCTCGCG
```

Matches (|) indicate strong evolutionary constraints, meaning the sequence is critical for function.

Highly **conserved sequences** often correspond to:

- **Protein active sites** (e.g., catalytic residues in enzymes).
- **DNA regulatory elements** (e.g., promoters, enhancers).
- **RNA structural motifs** (e.g., ribosomal RNA stems and loops).

Sequence mismatches ( ) highlight mutations that contribute to genetic variation and adaptation

```
CGCGATTCTATAGTCTAACATGCGAGCGTGGAAAAAAGATCTCGCG
||    ||||||||   |   | | | |  ||| |||      |||
CGACGATCTATAGTAACATGCGAGCGTGACGAATAAAAGATCTGCG
```

Mismatches occur when **different residues are aligned**

Point mutations can be neutral, beneficial, or deleterious.

- **Synonymous mutations** that do not alter the protein sequence.
- **Nonsynonymous mutations** that may change protein function.

After today, you should have a better understanding of

Conceptual interpretation of alignment results

Insertions and deletions

Insertions (–) introduce new genetic material, impacting protein structure and genome evolution

```
CGACGATTCTATAGTC-------------TGACGAATAAAAGATCTCGCG
|||||||||||||||              ||||||||||||||||||||||
CGACGATTCTATAGTCTAACATGCGAGCGTGACGAATAAAAGATCTCGCG
```

Gaps are used to indicate insertions

Causes of insertions:

- Gene duplications leading to new protein functions.
- Insertion of transposable elements modifying gene regulation.
- Microindels affecting protein structure and function.

Functional and evolutionary impact:

- Short insertions in proteins modify binding sites or enzyme activity.
- Insertions in DNA regulatory regions affect gene expression patterns.

Deletions (–) remove genetic material, leading to functional changes or species divergence

```
CGACGATTCTATAGTCTAACATGCGAGCGTGACGAATAAAAGATCTCGCG
||||||||||||||||           ||||||||||||||||||||||
CGACGATTCTATAGTC-----------TGACGAATAAAAGATCTCGCG
```

Gaps are used to indicate deletions

Causes of deletions:

- Loss of nonessential genes in parasitic or symbiotic organisms.
- Regulatory deletions affecting developmental pathways.
- Frameshift deletions that drastically alter protein coding.

Functional and evolutionary consequences:

- Can disable genes, leading to loss of function.
- Can optimize metabolic efficiency, as seen in endosymbiotic bacteria with streamlined genomes.

Small insertions and deletions (indels) are a major cause of genetic disorders

Indels play a major role in speciation by modifying gene structure and expression.

Frameshift mutations caused by small indels result in completely altered protein sequences.

Indels can disrupt coding sequences or regulatory elements, leading to disease.

- Cancer-related genes (e.g., TP53, BRCA1).
- Neurological disorders (e.g., Huntington's disease, caused by repeat expansions).
- Metabolic disorders (e.g., cystic fibrosis, due to a 3-base deletion in CFTR).

After today, you should have a better understanding of

Conceptual interpretation of alignment results

Alignment scores

Alignment scores measure sequence similarity by rewarding matches and penalizing mismatches and gaps

Alignment algorithms assign numerical scores to quantify how well two sequences align.

Matches receive **positive scores** (e.g., +1 or +2)

- Higher values are assigned to matches in functionally critical regions.

Mismatches receive **negative scores** (e.g., -1 or -2)

- Lower penalties for conservative substitutions (e.g., leucine to isoleucine).
- Higher penalties for radical substitutions (e.g., leucine to arginine, which changes charge and structure).

Gaps are **heavily penalized** (e.g., -2, or -3).

- Ensures meaningful evolutionary comparisons.
- This reflects that large insertions/deletions are less common than point mutations.

Substitution matrices model evolutionary relationships by assigning biologically meaningful scores to amino acid replacements

Not all mutations are equally likely—some substitutions occur more frequently due to biochemical properties.

Substitution matrices **assign different scores to different amino acid replacements** based on their evolutionary likelihood.

Impact on alignment quality:

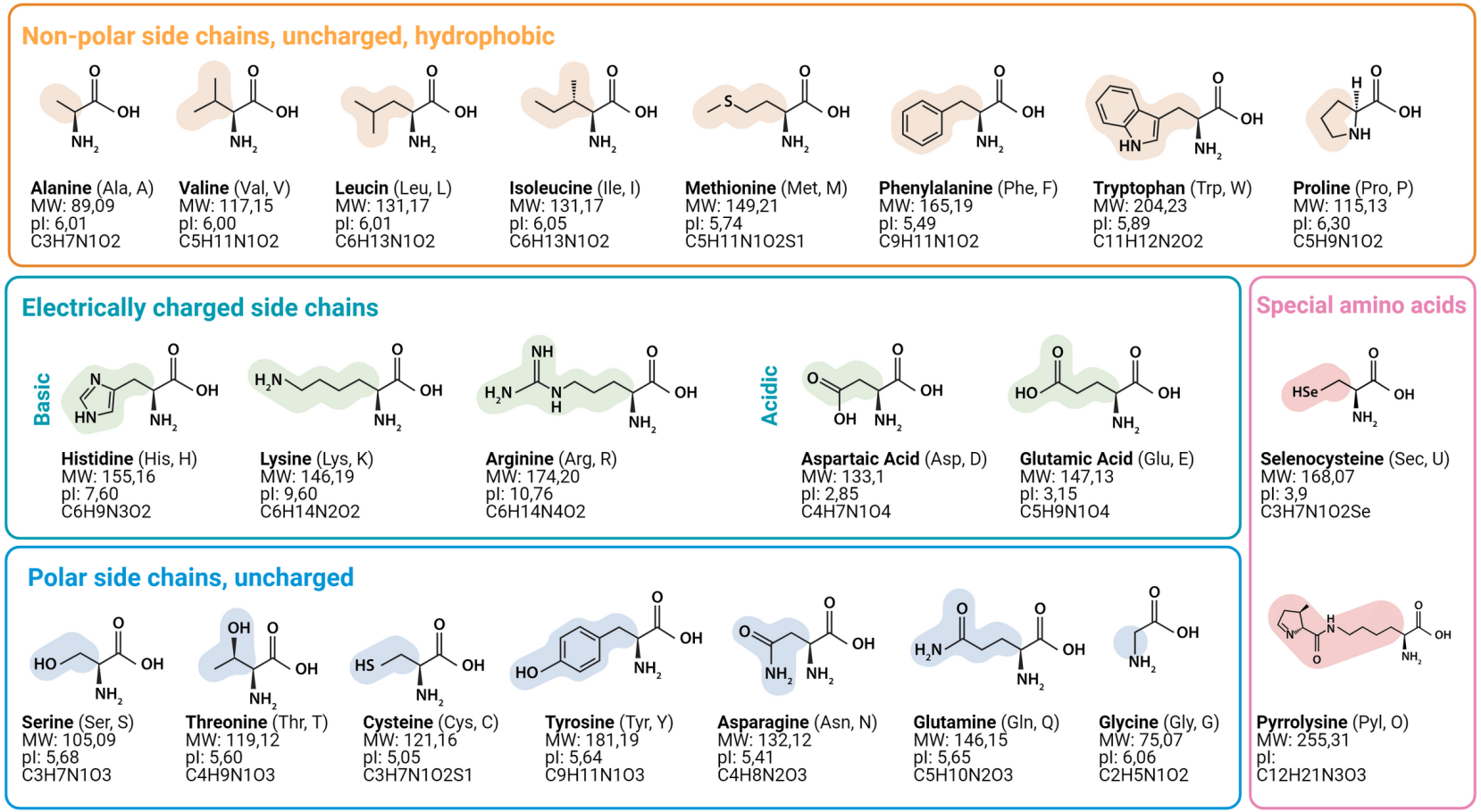
- Helps distinguish true homology from random similarity.
- Improves evolutionary modeling.
- Adjusts mismatch penalties based on real-world observations.

Two widely used matrices are **PAM (Point Accepted Mutation) matrices** and **BLOSUM (Blocks Substitution Matrix)**

More frequent substitutions have lower penalties, while rare substitutions are penalized more heavily

Physicochemical properties influence substitution likelihood:

- Hydrophobic residues often replace other hydrophobic residues.
- Charged residues tend to substitute with others of similar charge.



After today, you should have a better understanding of

Conceptual interpretation of alignment results

E-values

E-values measure the likelihood that an alignment occurs by chance, helping assess biological relevance

E-value (Expectation Value): Number of expected random matches in a database search.

Lower E-value = Higher significance (e.g., E = 0.001 means only 1 in 1,000 alignments is due to chance).

Database size affects E-value: Larger databases increase the probability of chance alignments.

After today, you should have a better understanding of

Pairwise versus multiple sequence alignment

Pairwise sequence alignment is the fundamental method for comparing two biological sequences

Pairwise alignment finds the optimal arrangement of two sequences to maximize similarity and minimize differences.

Methods like global and local alignment provide different perspectives on sequence similarity.

```
Query  1        ATGACTTTATCCATTCTAGTTGCACATGACTTGCAACGAGTAATTGGTTTTGAAAATCAA  60
Sbjct  2555705  .....A.....A..AA.T..C..T..C..TAAA...A....C.....G.ACC........  2555646

Query  61       TTACCTTGGCATCTACCAAATGATTTGAAGCATGTTAAAAAATTATCAACTGGTCATACT  120
Sbjct  2555645  ..........CT............A......A.....C..C.GA.C.....GA....A  2555586

Query  121      TTAGTAATGGGTCGTAAGACATTTGAATCGATTGGTAAACCACTACCGAATCGTCGAAAT  180
Sbjct  2555585  C.T.......CA..G..A..T...A.T..T..A..G..G...T.G..A...A.A..T..C  2555526

Query  181      GTTGTACTTACTTC---AGATACAAGTTTCAACGTAGAGGGCGTTGATGTAATTCATTCT  237
Sbjct  2555525  ..C.....C...AACCA..C.T..--.....C.A.GA....-..A.....T..AA.C...  2555469

Query  238      ATTGAAGATATTTATCAACTACCGGGCCATGTTTTTATATTTGGAGGGCAAACATTATTT  297
Sbjct  2555468  C....T..A...A.AG.GT..T.T..T...................A.....G....AC  2555409

Query  298      GAAGAAATGATTGATAAAGTGGACGACATGTATATTACTGTTATTGAAGGTAAATTTCGT  357
Sbjct  2555408  ....C.........CC.G..A..T..T........C..A..A..A..T..A..G....AA  2555349

Query  358      GGTGATACGTTCTTTCCACCTTATACATTTGAAGACTGGGAAGTTGCCTCTTCAGTTGAA  417
Sbjct  2555348  ..A..C..A..........A..C.....C...A..........C.AA........A...  2555289

Query  418      GGTAAACTAGATGAGAAAAATACAATTCCACATACCTTTCTACATTTAATTCGTAAAAAA  477
Sbjct  2555288  ...C.........A........T..A..G.....A..CT........G.G....G....  2555229
```

While pairwise alignment is effective for comparing two sequences, it has limitations for analyzing multiple sequences

Strengths:

- Computationally efficient for two sequences.
- Provides a direct, detailed comparison.
- Useful for identifying single mutations or evolutionary changes.

Limitations:

- Cannot reveal conserved regions across multiple species.
- Cannot model evolutionary relationships between many sequences.
- Performance and accuracy decline when extended to multiple sequences

Example: A pairwise comparison of hemoglobin genes between humans and chimpanzees provides insight into species divergence but does not reveal broader evolutionary trends across mammals.

MSA extends pairwise alignment to multiple sequences, enabling more powerful biological interpretations

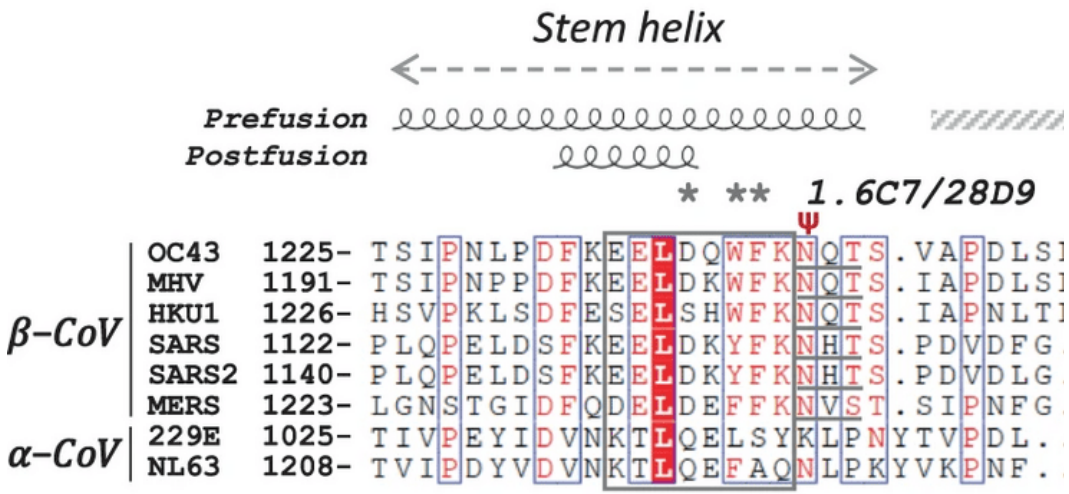
MSA aligns three or more sequences to reveal **conserved motifs, functional domains, and evolutionary relationships**.

Unlike pairwise alignment, MSA considers multiple substitutions, insertions, and deletions across species.

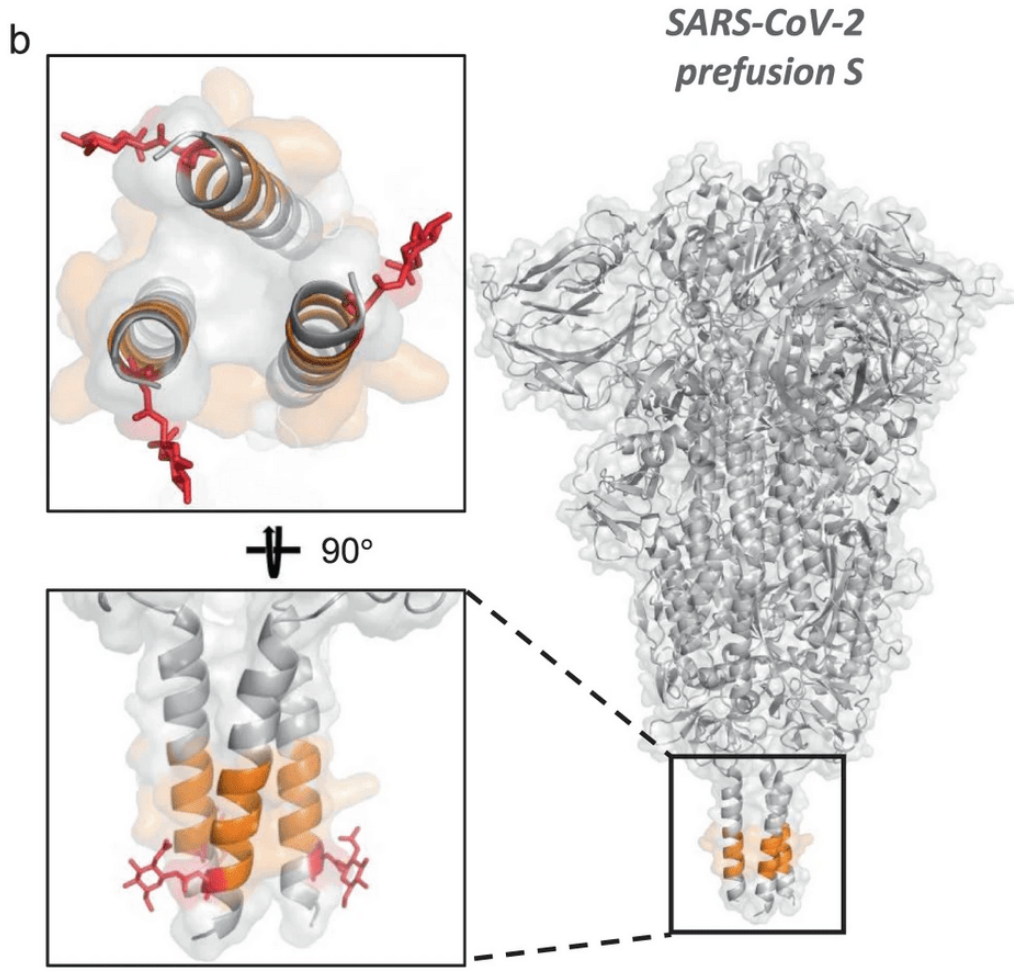
Example: **ClustalW and MUSCLE** generate MSAs to compare entire protein families

Example: MSA of SARS-CoV-2 spike proteins identifies conserved regions for vaccine development

Sequence alignment identified key conserved residues that are often epitopes



Residues in orange are stem helix epitope region



DOI: 10.1038/s41467-021-21968-w

Before the next class, you should

Lecture 05A:

Sequence alignment - Foundations

Lecture 05B:

Sequence alignment - Methodology

Today

Thursday

- Start P01D (due Feb 10)