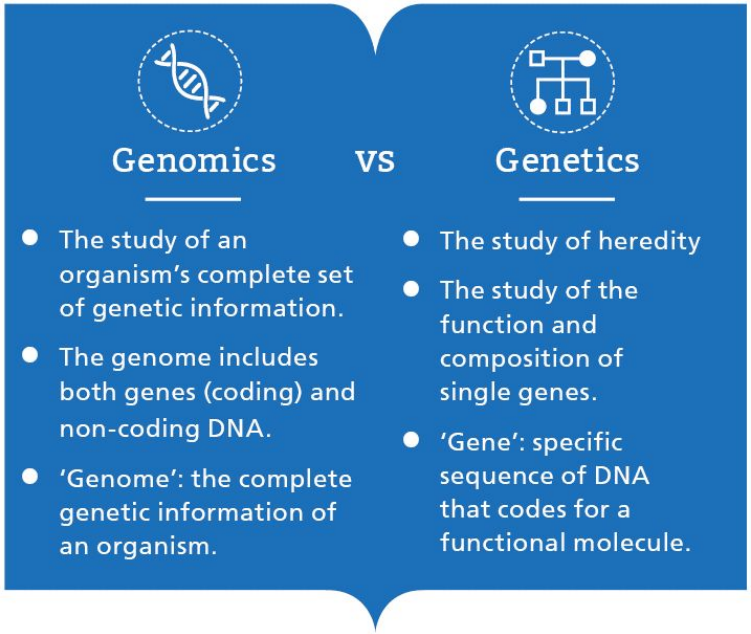
Lab 2: From Genomics to Sequence Alignment

Intro to Genomics and Sequence Alignment Techniques

As. Univ. Drd. Ing. Bozdog Alexandru

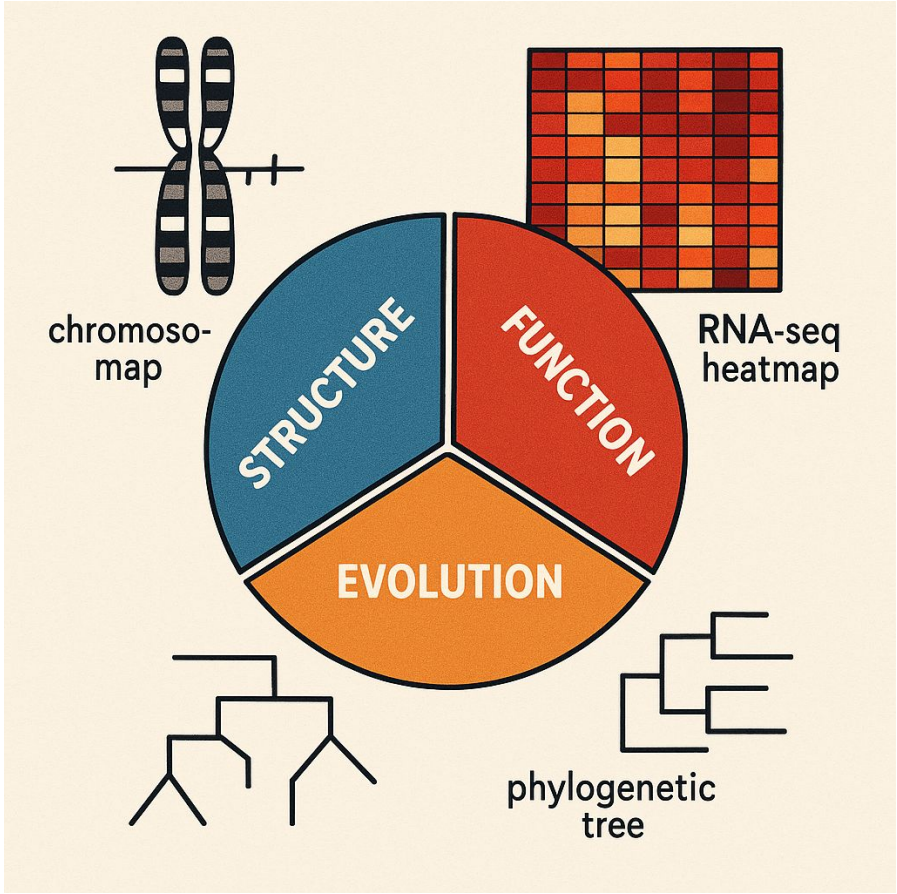
Learning Objectives

- ✓ Define genus, genome, sequences
- ✓ ADN → ARN → protein
- ✓ Recognize biological file formats
- ✓ Understanding sequence alignment
- ✓ Practical application with BLAST, Clustal and Biopython



What is Genomics?

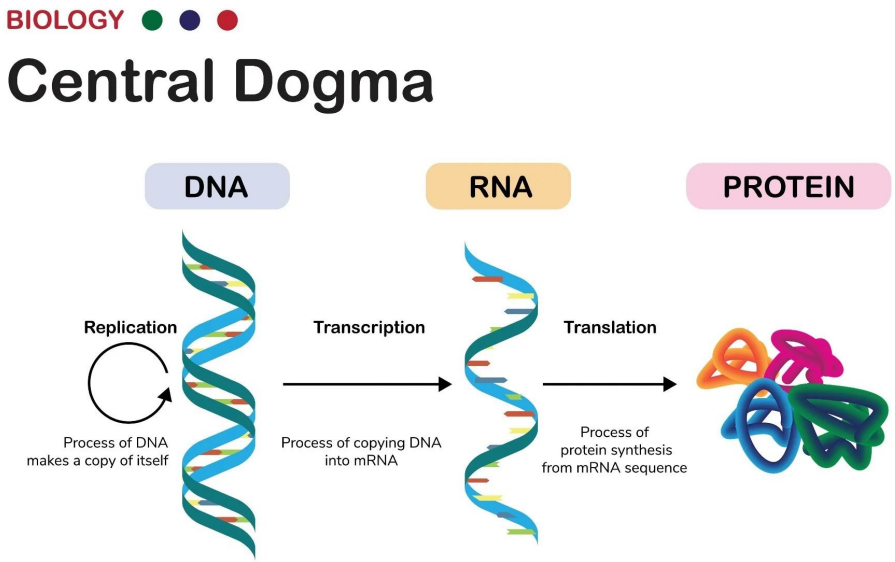
- Study of the whole DNA of an organism
- Fields of interest: Structural, Functional, Comparative, Translational
- Data obtained from high-throughput sequencing
- Relevance → medicine, evolution, biotechnology



Central Dogma

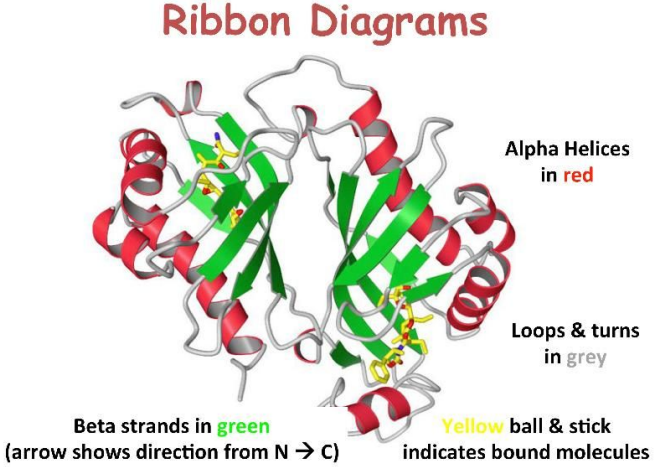
DNA → RNA → Protein

- Transcription and Translation
- Each level - different alphabet
- Flow of information = flow of data



Sequences and Biological formats

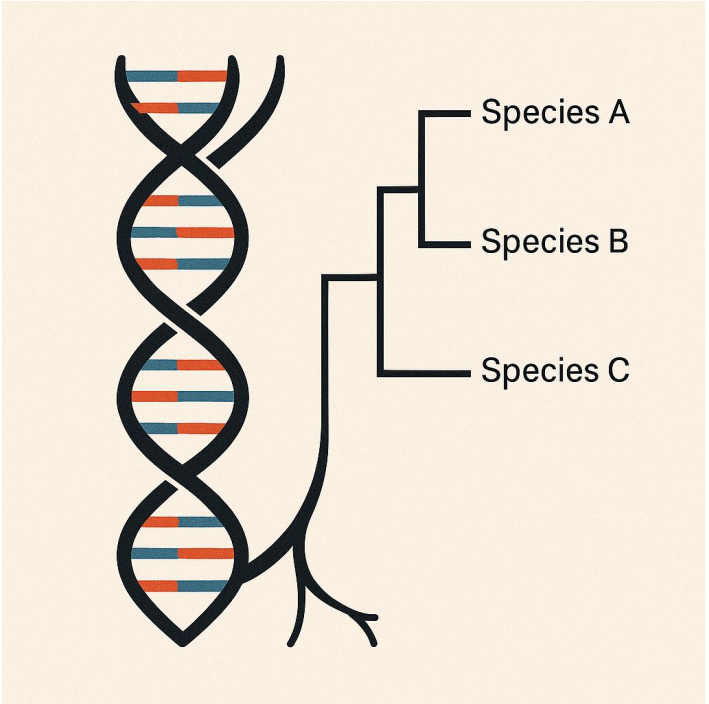
Type	Alphabet	Format	Example
DNA	A T C G	FASTA / FASTQ	`>TP53 ATGCGTAAC`
RNA	A U C G	FASTA	`AUGCGAU`
Protein	20 aa	FASTA / PDB	`MKTAYIAKQ`




```
>NM_011640.3 Mus musculus transformation related protein 53 (Trp53), transcript variant 1, mRNA
TTTCCCCTCCCACGTGCTCACCCTGGCTAAAGTTCTGTAGCTTCAGTTCATTGGGACCATCCTGGCTGTA
GGTAGCGACTACAGTTAGGGGGCACCTAGCATTCAGGCCCTCATCCTCCTCCTTCCCAGCAGGGTGTCAC
GCTTCTCCGAAGACTGGATGACTGCCATGGAGGAGTCACAGTCGGATATCAGCCTCGAGCTCCCTCTGAG
CCAGGAGACATTTTCAGGCTTATGGAAACTACTTCCTCCAGAAGATATCCTGCCATCACCTCACTGCATG
GACGATCTGTTGCTGCCCCAGGATGTTGAGGAGTTTTTTGAAGGCCCAAGTGAAGCCCTCCGAGTGTCAG
GAGCTCCTGCAGCACAGGACCCTGTCACCGAGACCCCTGGGCCAGTGGCCCCTGCCCCAGCCACTCCATG
GCCCCTGTCATCTTTTGTCCCTTCTCAAAAAACTTACCAGGGCAACTATGGCTTCCACCTGGGCTTCCTG
CAGTCTGGGACAGCCAAGTCTGTTATGTGCACGTACTCTCCTCCCCTCAATAAGCTATTCTGCCAGCTGG
CGAAGACGTGCCCTGTGCAGTTGTGGGTCAGCGCCACACCTCCAGCTGGGAGCCGTGTCCGCGCCATGGC
CATCTACAAGAAGTCACAGCACATGACGGAGGTCGTGAGACGCTGCCCCCACCATGAGCGCTGCTCCGAT
GGTGATGGCCTGGCTCCTCCCCAGCATCTTATCCGGGTGGAAGGAAATTTGTATCCCGAGTATCTGGAAG
ACAGGCAGACTTTTCGCCACAGCGTGGTGGTACCTTATGAGCCACCCGAGGCCGGCTCTGAGTATACCAC
CATCCACTACAAGTACATGTGTAATAGCTCCTGCATGGGGGGCATGAACCGCCGACCTATCCTTACCATC
ATCACACTGGAAGACTCCAGTGGGAACCTTCTGGGACGGGACAGCTTTGAGGTTCGTGTTTGTGCCTGCC
CTGGGAGAGACCGCCGTACAGAAGAAGAAAATTTCCGCAAAAAGGAAGTCCTTTGCCCTGAACTGCCCCC
AGGGAGCGCAAAGAGAGCGCTGCCCACCTGCACAAGCGCCTCTCCCCCGCAAAAGAAAAAACCACTTGAT
GGAGAGTATTTCACCCTCAAGATCCGCGGGCGTAAACGCTTCGAGATGTTCCGGGAGCTGAATGAGGCCT
TAGAGTTAAAGGATGCCCATGCTACAGAGGAGTCTGGAGACAGCAGGGCTCACTCCAGCTACCTGAAGAC
CAAGAAGGGCCAGTCTACTTCCCGCCATAAAAAAACAATGGTCAAGAAAGTGGGGCCTGACTCAGACTGA
CTGCCTCTGCATCCCGTCCCCATCACCAGCCTCCCCCTCTCCTTGCTGTCTTATGACTTCAGGGCTGAGA
CACAATCCTCCCGGTCCCTTCTGCTGCCTTTTTTACCTTGTAGCTAGGGCTCAGCCCCCTCTCTGAGTAG
```

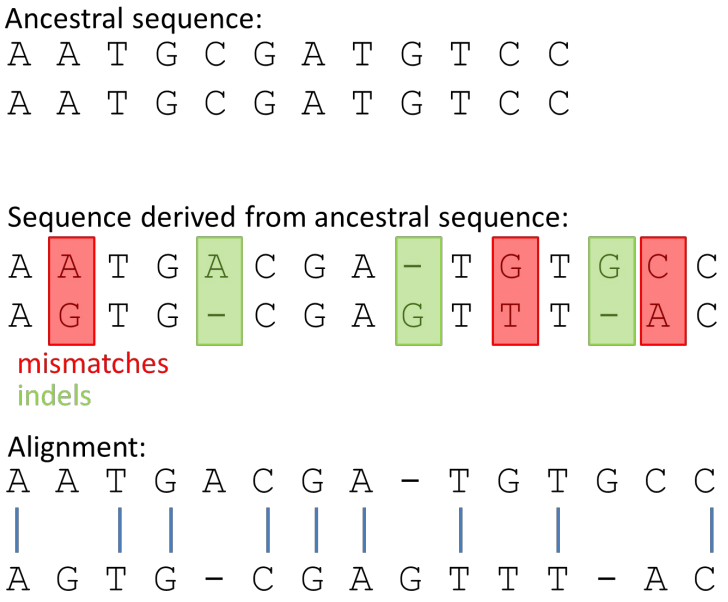
Why sequence alignment?

- Detect homology (common origin)
- Identify mutations, insertions, deletions
- Deduce function and structure
- Build phylogenetic trees



What is Sequence Alignment

- Rearranging sequences to maximize similarity
- Introducing gaps (–) for insertions/deletions
- Calculating an optimal score



Types of Alignment

Type	Purpose	Algorithm	Use case
Global	Full length	Needleman–Wunsch	Similar genes
Local	Best fit region	Smith–Waterman	Motifs, domains
Semi-global	No edge loss	Hybrid	NGS Reads

Scores and Substitution Matrices

$$S = \sum_{i=1}^{L} s(x_i, y_i) - \sum \text{penalizări pentru gap-uri}$$

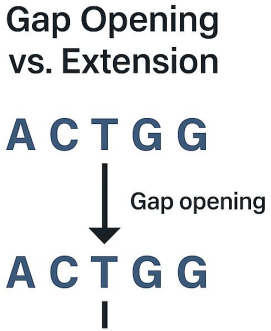
unde:

- $s(x_i, y_i)$ este scorul din matrice (ex. BLOSUM);
- penalizările provin din evenimentele de inserție/deleție.

- DNA: +1 match / -1 missmatch
- Protein: Matrices (Blossum / PAL)

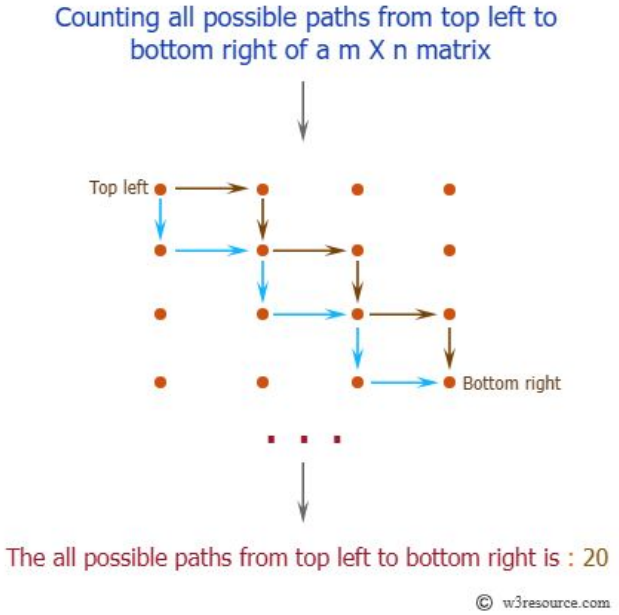
- Gap penalties: linear, afine
- Goal: Maximize score

A	C	G	T
−1	−2	−2	1
−2	−2	−2	−1
−2	−2	−2	0



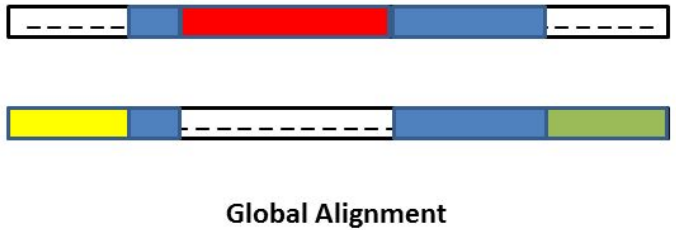
Dynamic programming

1. Initiate matrix
2. Recurently : `S(i,j) = max(diag+match, stânga+gap, sus+gap)`
3. Backtracking → Optimal Alignment
4. O(n×m) Complexity

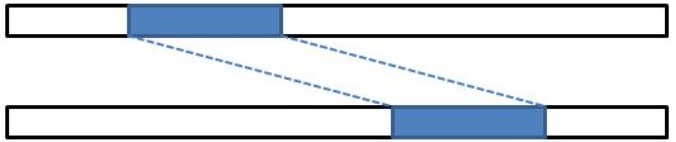


Local vs Global Alignment

- **Global:** Covers complete sequence
- **Local:** covers only maximum score region
- **BLAST =** quick Smith-Waterman



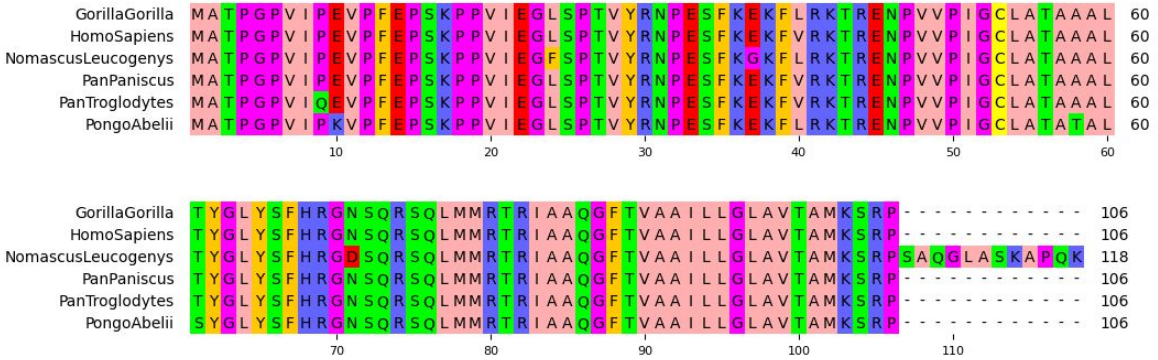
Global Alignment



Local Alignment

Multiple Sequence Alignment

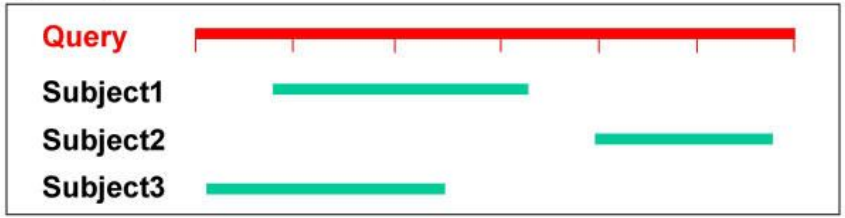
- 2 or more sequences aligned simultaneously
- Heuristics: progressive(Clustal Ω), consistency(T-Coffee)
- Identify conserved motifs, domains, evolutionary relationships



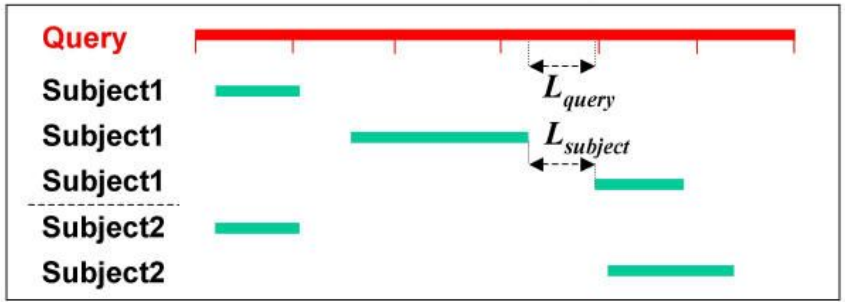
Tools and Libraries

```
from Bio import pairwise2
```

Type1 alignment: continuous match



Type2 alignment: discontinuous matches in the same subject



CLI: `needle`, `water`, `blastn`, `clustalo`

Python (Biopython): `pairwise2`, `AlignIO`, `Bio.Blast.NCBIWWW`

Results :Alignment, Score, E-value

Alignment Evaluation

Metric	Semnification	Interpretation
Score	Sum of matches - penalties	Algorithm specific
% Identity	Exact matches - length	Similarity Degree
E-value	Probability Score appears by chance	Smaller = More significant
Consensus	Dominant character per column	MSA/ conservation

Use Cases & Conclusions

- Identify mutations and SNP-s
- Construct phylogenetic trees
- Predict proteic domains
- Drug repositioning
- Metagenomic analysis

✅ Genomics = data source

✅ Sequence alignment = First step of analysis

✅ Instruments: BLAST, Clustal, Biopython

➡️ Next lab : Aligning NGS reads and variant analysis