Pairwise Sequence Alignments

Lecture 4
Bioinformatics level 3
Biotechnology program

Dr. Omnia Badr
Department of Genetics

Pevsner, J., 2015. *Bioinformatics and functional genomics*. John Wiley & Sons.

Pairwise sequence alignment is the most fundamental operation of bioinformatics

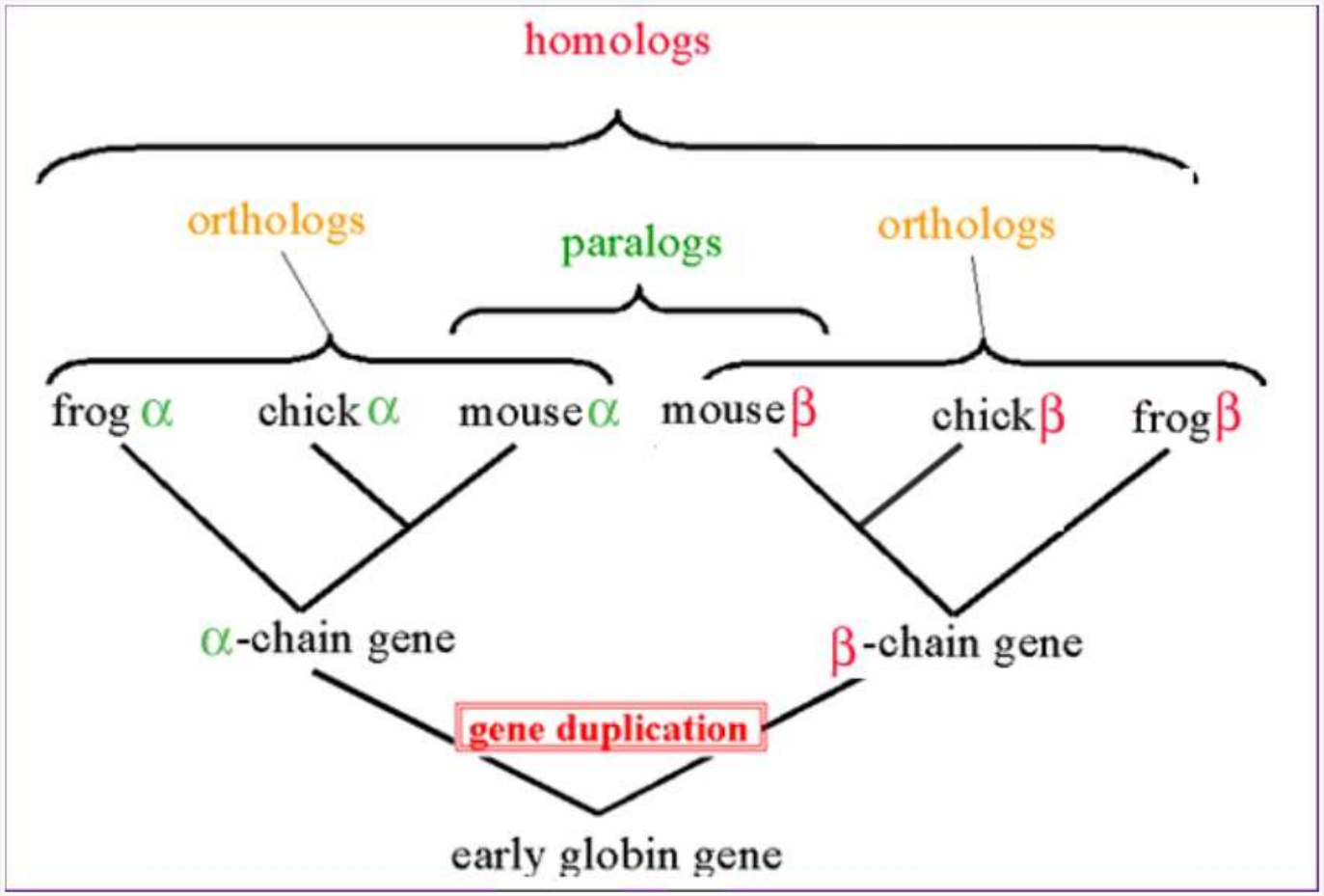
- It is used to decide if two proteins (or genes) are related structurally or functionally
- It is used to identify domains or motifs that are shared between proteins
- It is the basis of BLAST searching
- It is used in the analysis of genomes

Pairwise Sequence Alignment

- What is an alignment, and why might it be significant?
  - An alignment is *a mapping from one sequence to another, identifying elements that are likely to have arisen from a common ancestor*
  - A good alignment is an indication of homology

Similarity vs. Homology

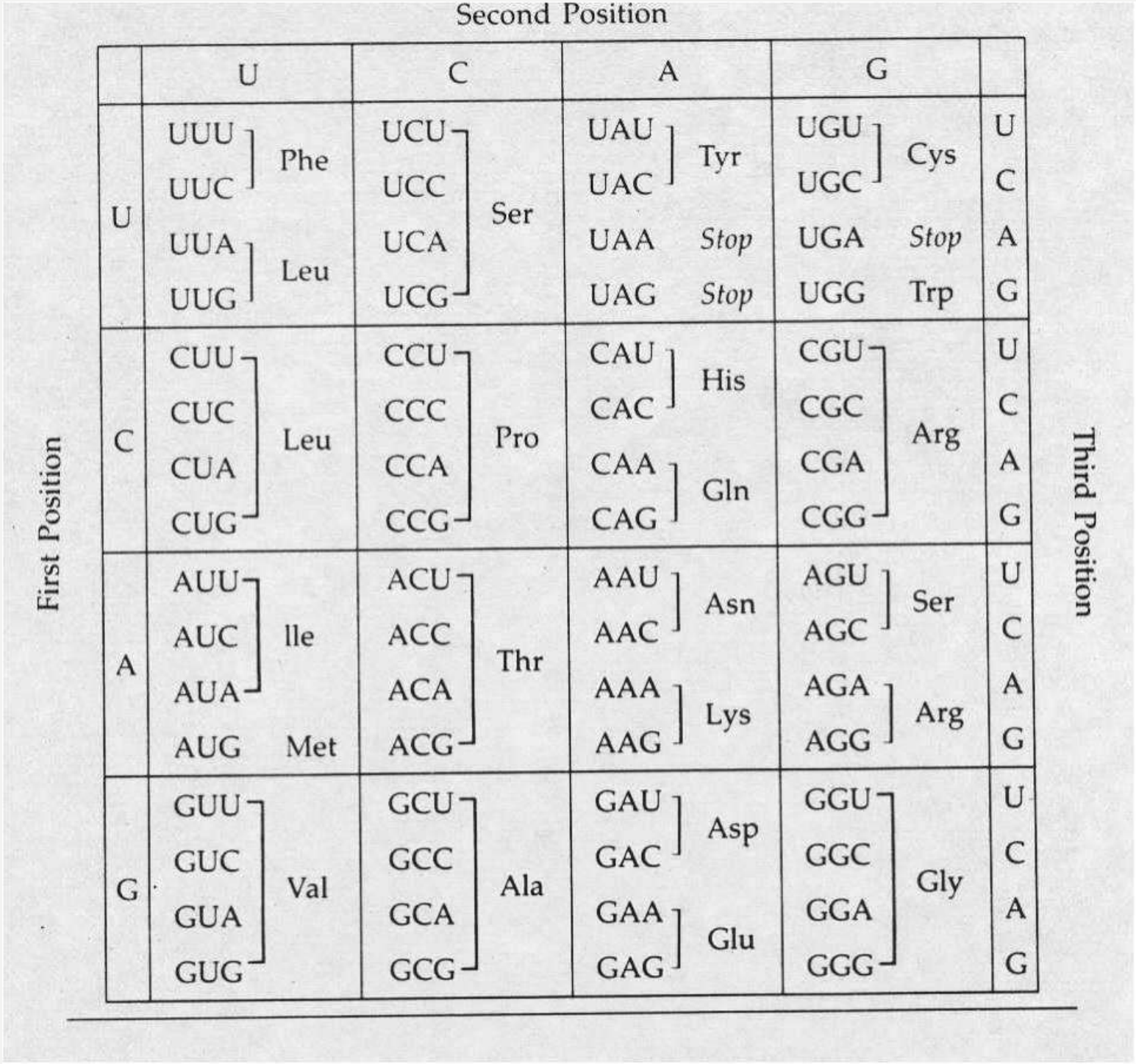
- *Homology* is an evolutionary relationship that either exists or does not. It cannot be partial.
- *Similarity* is a measure of the quality of alignment between two sequences. High similarity is evidence for homology.



Homologous sequences. Orthologs and Paralogs are two types of homologous sequences. Orthology describes genes in different species that derive from a common ancestor. Orthologous genes may or may not have the same function. Paralogy describes homologous genes within a single species that diverged by gene duplication.

Pairwise alignment: protein sequences can be more informative than DNA

- protein is more informative (20 vs 4 characters); many amino acids share related biophysical properties
- codons are degenerate: changes in the third position often do not alter the amino acid that is specified
- protein sequences offer a longer “look-back” time (relatedness over millions or billions of years)
- DNA sequences can be translated into protein, and then used in pairwise alignments


Second Position
First Position
Third Position
U
C
A
G
UUU Phe
UUC
UUA Leu
UUG
UCU Ser
UCC
UCA
UCG
UAU Tyr
UAC
UAA Stop
UAG Stop
UGU Cys
UGC
UGA Stop
UGG Trp
CUU Leu
CUC
CUA
CUG
CCU Pro
CCC
CCA
CCG
CAU His
CAC
CAA Gln
CAG
CGU Arg
CGC
CGA
CGG
AUU Ile
AUC
AUA
AUG Met
ACU Thr
ACC
ACA
ACG
AAU Asn
AAC
AAA Lys
AAG
AGU Ser
AGC
AGA Arg
AGG
GUU Val
GUC
GUA
GUG
GCU Ala
GCC
GCA
GCG
GAU Asp
GAC
GAA Glu
GAG
GGU Gly
GGC
GGA
GGG

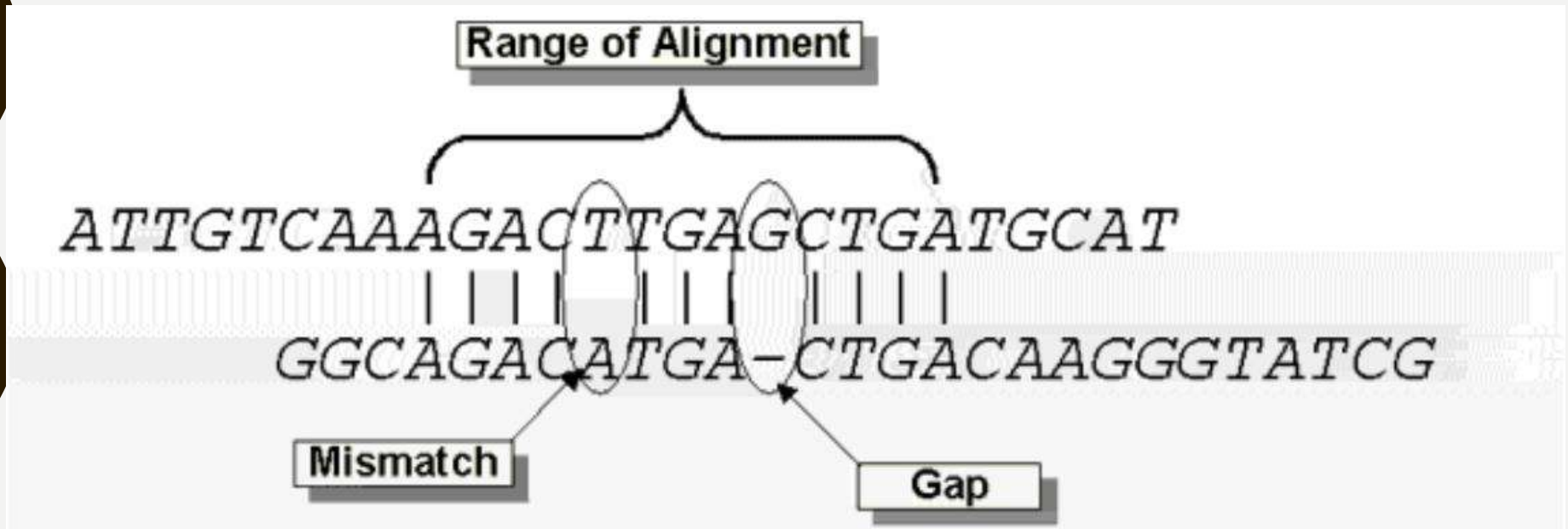
How do we compute similarity?

- Similarity can be defined by counting positions that are identical between two sequences

```
Query: 181 catcaactacaactccaaagacacccttacacccactaggatatcaacaaacctacccac 240
           ||||||||  ||||  ||||||  |||||  | ||||||||||||||||||||||||||||
Sbjct: 189 catcaactgcaaccccaaagccacccct-cacccactaggatatcaacaaacctacccac 247
```

- Gaps (insertions/deletions) can be important

An alignment scoring system is required to evaluate how good an alignment is



$$S = \sum(\text{identities, mismatches}) - \sum(\text{gap penalties})$$

$$\text{Score} = \text{Max}(S)$$

- use of a substitution matrix

Many possible alignments to consider

- Without gaps, there are NxM possible alignments between sequences of length N and M
- Once we start allowing gaps, there are many possible arrangements to consider:

```
abcbcd      abcbcd      abcbcd
abc---d     a--bcd      ab--cd
```

- This becomes a very large number when we allow mismatches, since we then need to look at every possible pairing between elements: there are roughly N_M possible alignments.

Scoring schemes

- For nucleotide sequences a simple scheme would be to assign +1 for a match, -1 for a mismatch and maybe assign a gap penalty of -2 for gaps

Scoring schemes for proteins

- Can be simple (+1 for match, -1 for mismatch) based on physiochemical properties for instance
- Or better yet take into account evolution and the rate of mutations over time
- Substitution matrices are exactly that

Global alignment versus local alignment

Global alignment (Needleman-Wunsch) extends from one end of each sequence to the other

Local alignment finds optimally matching regions within two sequences (“subsequences”)

Local alignment is almost always used for database searches such as BLAST. It is useful to find domains (or limited regions of homology) within sequences