Sequence comparison theory

A sequence by itself does not provide any information

information can be retrieved by

- Sequence analysis
- <u>Similarity searching by means of comparison between sequences</u>

4 GOALS FOR SIMILARITY SEARCHING

1. Localization of a new sequence through the similarity with a previously localized sequence.

2. Hypothesis on function of a new sequence, if it is identical or similar to a known sequence

3. Classification of a protein in a specific family

4. Assessment of evolutionary relations between sequences

- Identification of conserved sequences

sequences where any change in a specific position (amino acid or DNA) does not change the molecule chemico-physical traits

You can always evaluate similarity of sequences but sometimes you cannot establish the mechanisms which caused similarity.

Biological similarity may occur for

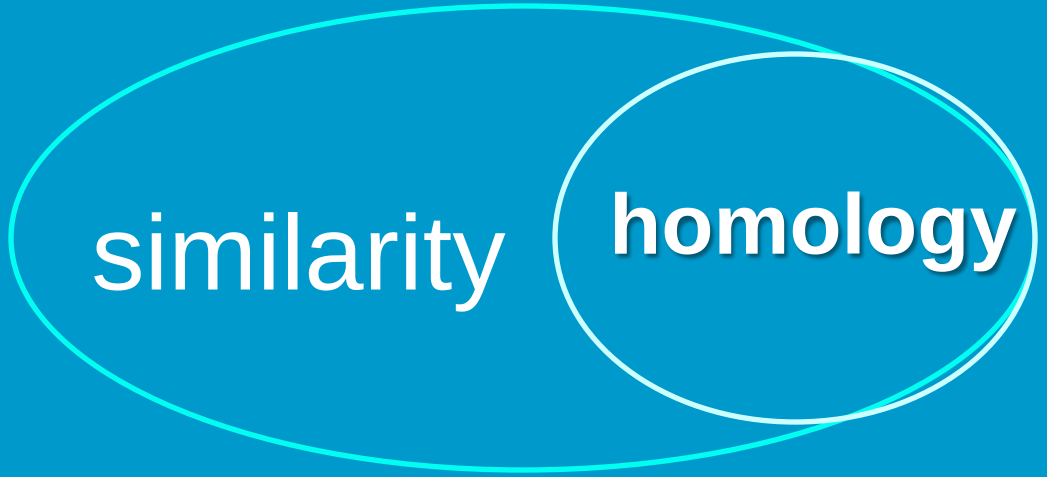
- Random events
- Convergent adaptation
- Homology

Example: wings of birds and bats are not homologous since are derived from independent evolutionary steps

Sequence comparison

similarity ≠ **homology**

Apart from the causes — Common ancestor



Through the sequence comparison you can explore the relations between sequences

- <u>Similarity is a quantitative value</u>
  - concerns identity between sequences
- <u>Homology is a qualitative trait</u>
  - refers to an evolutionary relation between sequences

Some words… concerning relation between sequences

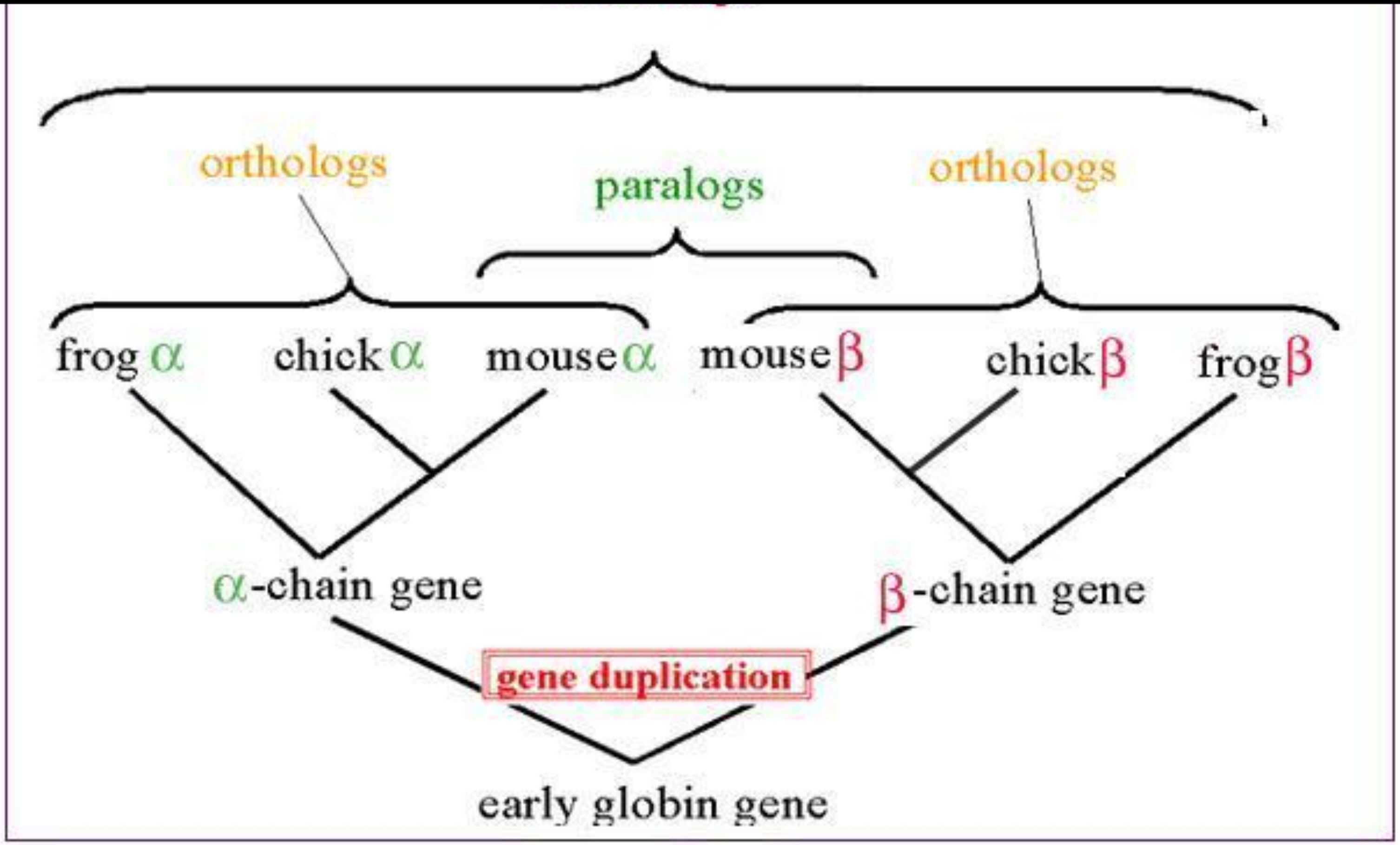
Similar: sequences which are characterized by an identity level

Homologous: sequences which derive from the same evolutionary pathway

Hortologous: homologous sequences which derive from a common ancestor but evolved independently. Function is or is not maintained.

Paralogous: homologous sequences which evolved through gene duplication in the same species

Homologous sequences



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.

A high similarity likely indicates homology

but

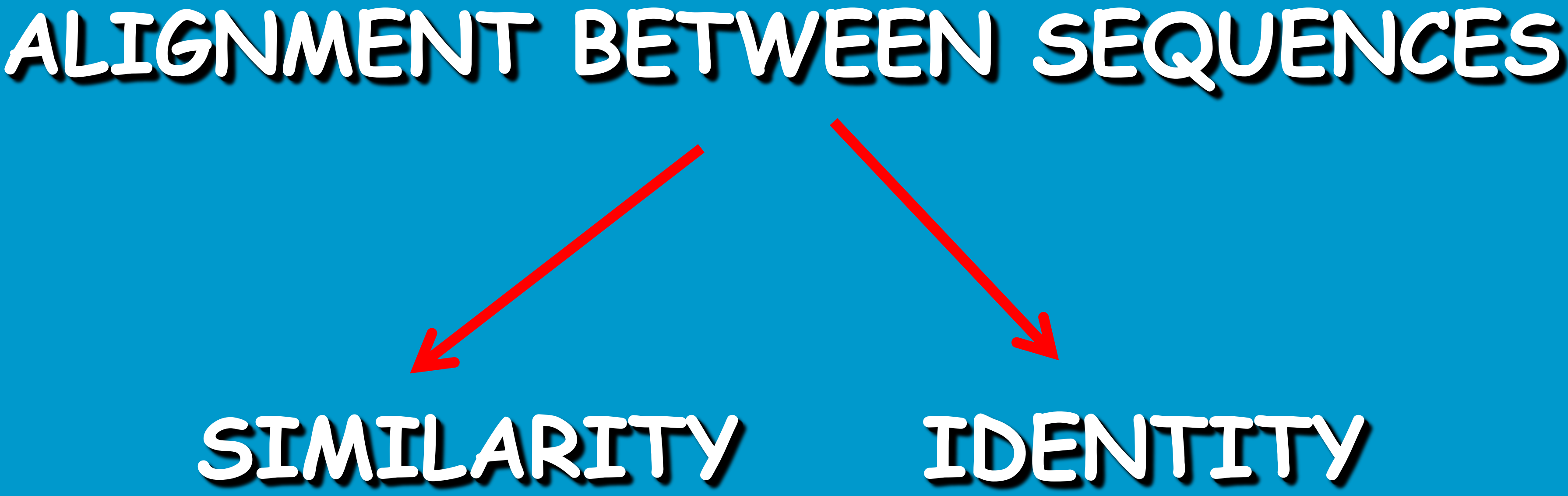
homology only sometimes corresponds

to high similarity

Homology, Similarity, and Identity

- <u>Identity</u> is a *measure* made on an alignment
  - Sequence A can be “32 % identical to” Sequence B
- <u>Similarity</u> is a measure of how close are
  - Two sequences
  - Two amino acids (isoleucine and leucine)
- <u>Homology is a *property* that exists or does not exist</u>
  - Sequence A ***IS*** or ***IS NOT*** homologous to Sequence B
  - Sequence A cannot be “40% homologous to” B

Homology is established on the basis of measured similarity or identity



ALIGNMENT BETWEEN SEQUENCES

SIMILARITY

IDENTITY

SIMILARITY SEARCHING IS BASED ON ALIGNMENT BETWEEN SEQUENCES.

- Sequence identity
  - Percent of matches (nucleotides, amino acids)
- Sequence similarity
  - measure of how close are two sequences

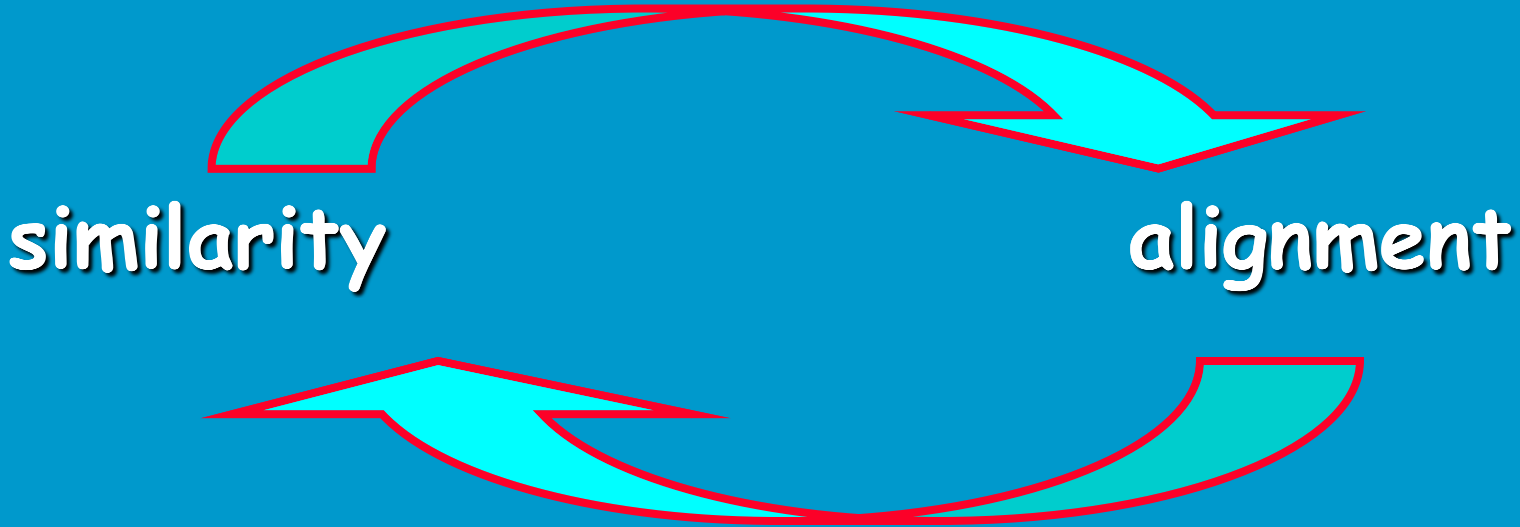
to evaluate identity or similarity between sequences comparison can be perfoprmed

- Between two sequences (pairwise alignment)
  - With a subject sequence
  - In public database

- Between more sequences (multiple alignment)
  - Multiple comparison
  - In public database

Pairwise comparison between two sequences

Similarity between two sequences



1 – definition of similarity criteria
2 – sequence alignment
3- similarity evaluation

Alignment of the two strings of characters.

All the possible alignments are tested

Similarity between two sequences

1- alignment of the string of characters
All the possible alignments are tested

For example: Align sequence 2 on sequence 1

Query sequence → AAKQW Sequence 1

Subject sequence → AAKKQW Sequence 2

Similarity between two sequences

AAKKQW	6 characters
AAKQW	5 caharacters

We tested 10 (5+5) possible alignments

and compared 30 (6x5) characters

Similarity between two sequences

1- Line up the sequences against each other

```
AAKKQW
     AAKQW
```

```
AAKKQW
    AAKQW
```

```
AAKKQW
   AAKQW
```

```
AAKKQW
  AAKQW
```

```
AAKKQW
 AAKQW
```

```
AAKKQW
AAKQW
```

```
 AAKKQW
AAKQW
```

```
  AAKKQW
AAKQW
```

```
   AAKKQW
AAKQW
```

```
    AAKKQW
AAKQW
```

Similarity between two sequences

1- alignment of the two strings of characters. All the possible alignments are tested

- without gaps

2- evaluation of similarity by the sum of the characters which align perfectly.

Similarity between two sequences

2- score all the possible alignments through the sum of the characters which align perfectly (matches).

```
AAKKQW
     AAKQW    0

AAKKQW
    AAKQW     0

AAKKQW
   AAKQW      0

AAKKQW
  AAKQW       0

AAKKQW
 AAKQW        4
```

```
AAKKQW
AAKQW         3

 AAKKQW
AAKQW         1

  AAKKQW
AAKQW         0

   AAKKQW
AAKQW         0

    AAKKQW
AAKQW         0
```

Simple pairwise alignment

Two sequences are written one on the top of the other.

```
CGCTTCGGACGAAATCGCATCAGCATACGATCGCATGCCGGCGGATAAC

CGAAATCGCATCAGCATACGATCGCATGC
```

the query sequence is moved on the subject sequence.

```
CGCTTCGGACGAAATCGCATCAGCATACGATCGCATGCCGGCGGATAAC
||     |    |
CGAAATCGCATCAGCATACGATCGCATGC
```

```
CGCTTCGGACGAAATCGCATCAGCATACGATCGCATGCCGGCGGATAAC
         |   | |  |  | |  | |
 CGAAATCGCATCAGCATACGATCGCATGC
```

```
CGCTTCGGACGAAATCGCATCAGCATACGATCGCATGCCGGCGGATAAC
  |        |        |
  CGAAATCGCATCAGCATACGATCGCATGC
```

```
CGCTTCGGACGAAATCGCATCAGCATACGATCGCATGCCGGCGGATAAC
         || |   ||||  |    |   |
   CGAAATCGCATCAGCATACGATCGCATGC
```

```
CGCTTCGGACGAAATCGCATCAGCATACGATCGCATGCCGGCGGATAAC
        |    |||     |  ||   ||
    CGAAATCGCATCAGCATACGATCGCATGC
```

```
CGCTTCGGACGAAATCGCATCAGCATACGATCGCATGCCGGGCGGATAAC
     || |              |   ||   ||
     CGAAATCGCATCAGCATACGATCGCATGC
```

```
CGCTTCGGACGAAATCGCATCAGCATACGATCGCATGCCGGGCGGATAAC
       ||        || ||     |
      CGAAATCGCATCAGCATACGATCGCATGC
```

```
CGCTTCGGACGAAATCGCATCAGCATACGATCGCATGCCGGGCGGATAAC
           |   |        |      |
       CGAAATCGCATCAGCATACGATCGCATGC
```

```
CGCTTCGGACGAAATCGCATCAGCATACGATCGCATGCCGGGCGGATAAC
           ||
        CGAAATCGCATCAGCATACGATCGCATGC
```

```
CGCTTCGGACGAAATCGCATCAGCATACGATCGCATGCCGGGCGGATAAC
         |||||||||||||||||||||||||||||
         CGAAATCGCATCAGCATACGATCGCATGC
```

The highest score corresponds to the best alignment

Global or local alignment ?

Global (full-length) alignment:

```
LTGARDWEDIPLWTDWDIEQESDFKTRAFGTANCHK
 ||   | |  |   |      |  ||  || | ||
 TGIPLWTDWDLEQESDNSCNTDHYTREWGTMNAHK
```

local (sub-sequences) alignment :

```
TGARDWEDIPLWTDWDIEQESDFKTRAFGTANCHK
        |||||||| |||||
      TGIPLWTDWDLEQESDNSCNTDHYTREWGTMNAHK
```

Sub-sequences

When a sequence A is identical to any portion of a sequence B, sequence A is said to be a sub-sequence of B.

```
A     .............ESDFGHKLPV......
                   ||||||||||
B     CHKIPLMTRWDQQESDFGHKLPVIYTREW
```

GLOBAL OR LOCAL ALIGNMENT?

id.

global alignment :

```
LTGARDWEDIPLWTDWDIEQESDFKTRAFGTANCHK      13
 ||   | |  |          |  ||  || | ||
 TGIPLWTDWDLEQESDNSCNTDHYTREWGTMNAHKAG
```

local alignment :

```
TGARDWEDIPLWTDWDIEQESDFKTRAFGTANCHK          13
        |||||||| |||||
      TGIPLWTDWDLEQESDNSCNTDHYTREWGTMNAHK
```

Which is the best ?

GLOBAL OR LOCAL ALIGNMENT?

1) Select the best alignment by a biological point of view

2) At computational level the best alignment measure should reflect the best biological alignment

Local alignment often better reflect a common biological function

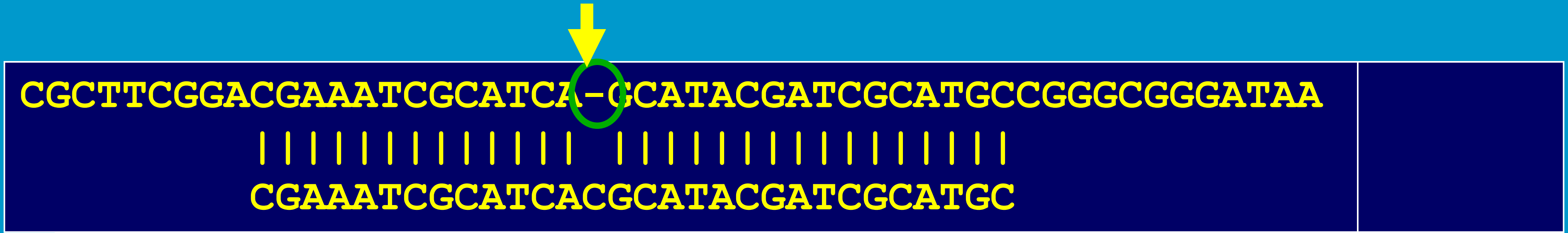
GLOBAL OR LOCAL ALIGNMENT?

- In theory, global alignment evaluates similarity of the overall sequence. It is best for describing relations between sequences.

- In practice, local alignment is of more general use.
  - In proteins only parts are homologous (share conserved domains)

Gaps

Simple alignment often does not work



```
CGCTTCGGACGAAATCGCATCA-GCATACGATCGCATGCCGGGCGGATAA
         ||||||||||||| ||||||||||||||||
         CGAAATCGCATCACGCATACGATCGCATGC
```

It can be necessary to include gaps (nucleotide insertion/deletion) and evaluate alignment with “gaps”

Gap

A SPACE INTRODUCED INTO AN ALIGNMENT TO COMPENSATE FOR INSERTIONS AND DELETIONS IN ONE SEQUENCE RELATIVE TO ANOTHER.

Similarity between two sequences with gaps

Insertion and/or deletion (INDEL) of characters (*gaps*)

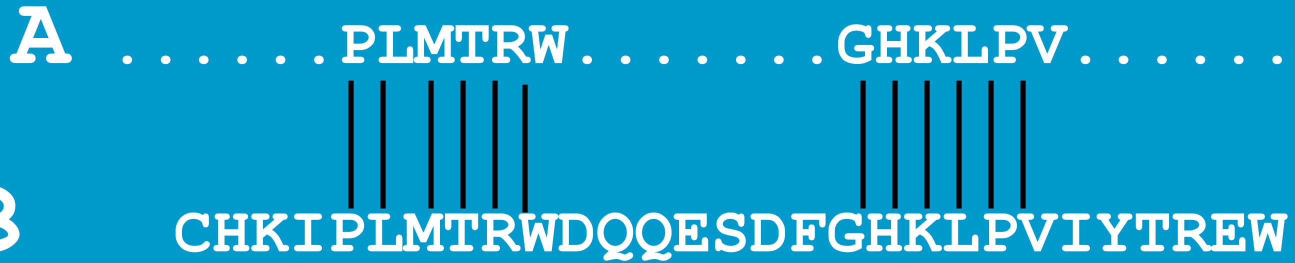
```
  IPLMTRWDQEQESDFGHKLPIYTREWCTRG
            ||||||||||                  10
CHKIPLMTRWDQQESDFGHKLPVIYTREW

                       ↓
   IPLMTRWDQEQESDFGHKLP-IYTREWCTRG
   ||||||||| |||||||||| ||||||          25
CHKIPLMTRWDQ-QESDFGHKLPVIYTREW
            ↑
```

Simple alignment does not work between A and B

A . . . . . .PLMTRWGHKLPV. . . . . .

B …. . . . . .CHKIPLMTRWDQQESDFGHKLPVIYTREW. . . . . .…

A is identical to different parts of B



```
A  ......PLMTRW.......GHKLPV......

B      CHKIPLMTRWDQQESDFGHKLPVIYTREW
```

More than one block of gaps must be inserted

GAP PENALTY

Is based on two notions:

- Deletion or insertion (gap) is much less likely to occur than the most radical amino acid substitution. It should be heavily penalized.
- Once a deletion or insertion (gap) has occurred in a given position deletion or insertion of additional residues (gap extension) becomes much more likely.

GAP PENALTY

In the scoring of an alignment introduction of a gap and extension of the gap causes the deduction of a fixed amount (the gap score, *G*).

$$G=a+bx\ ,\quad a>>b$$

a is the gap opening penalty, *b* is the gap extension penalty, *x* is the extension of the gap after the opening.

The choice of gap costs is empirical, but it is customary to choose a high value for gap existence (10-15) and a low value for gap extension (1-2).

Similarity between two sequences with gaps

Similarity evaluation - with gap penalty

```
   IPLMTRWDQEQESDFGHKLP-IYTREWCTRG
   ||||||||| |||||||||| ||||||      Score = 25 - 2
CHKIPLMTRWDQ-QESDFGHKLPVIYTREW
```

gap creation penalty (es.: -1 for each *gap*)

```
   IPLMTRWDQEQESDFGHKLP----IYTREWCTRG
   ||||||||| ||||||||||    ||||||
CHKIPLMTRWDQ-QESDFGHKLPVGSSIYTREW
```

gap extension penalty

Score = 25 - 1-(1+(0.1*3))

(es.: -0.1 for each ins/del following the first)

QUANTITATIVE EVALUATION OF SIMILARITY (SCORE) OF AN ALIGNMENT

Comprehensive alignment score

Must account for the identity of all the characters in both sequences and the gap penalties.

The procedure for the score evaluation should maximise the number of identical matches between sequences by inserting gaps

RAW SCORE

THE SCORE OF AN ALIGNMENT, S,
IS CALCULATED AS THE SUM OF

1- SUBSTITUTION SCORES
2- GAP SCORES.