SEQUENCE COMPARISON

SEQUENCE COMPARISON

- BASED ON ALIGNMENT
  - IDENTITY
  - SIMILARITY
    - GAP
- BIOLOGICAL MEANING

GOALS IN SEARCHING FOR SIMILARITY

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

A SEQUENCE BY ITSELF DOES NOT PROVIDE ANY INFORMATION

information can be retrieved by

- Sequence analysis
- Similarity searching by means of comparison between sequences

COMPARISON IS BASED ON ALIGNMENT OF THE STRINGS OF CHARACTERS (SEQUENCES)

AAKKQW 6 characters

AAKQW 5 characters

STEPS IN COMPARING TWO OR MORE SEQUENCES

- To determine the element by element correspondence
- To measure their similarity
- To observe conserved or variable patterns

Some concepts concerning relation between sequences

Similar: sequences which are characterized by an identity level

Homologous: sequences which derive from the same evolutionary pathway

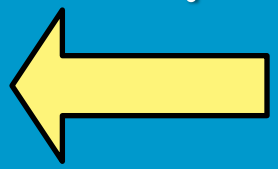
Hortologous: homologous sequences which derive from a common ancestor but evolved independently. Function is or is not maintained.

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

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



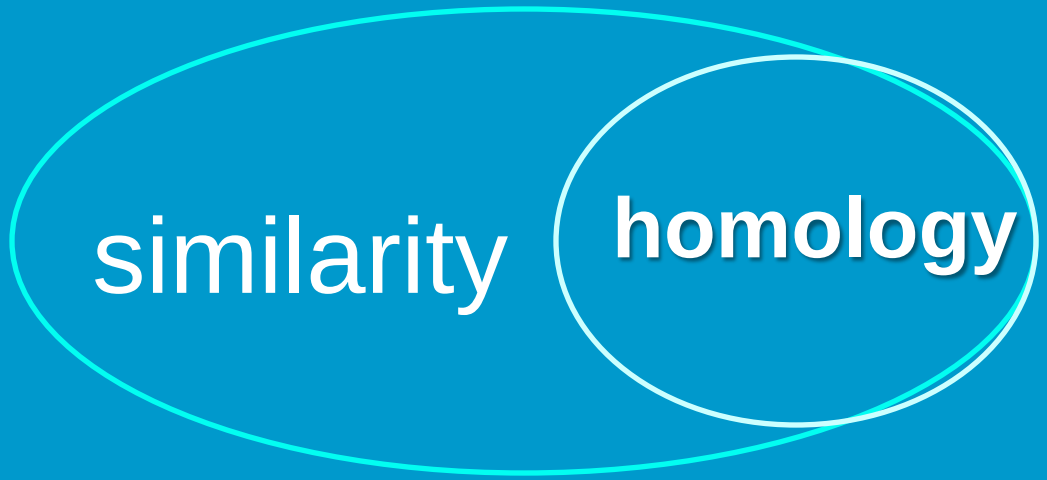
Sequence comparison

similarity

Apart from the causes

homology

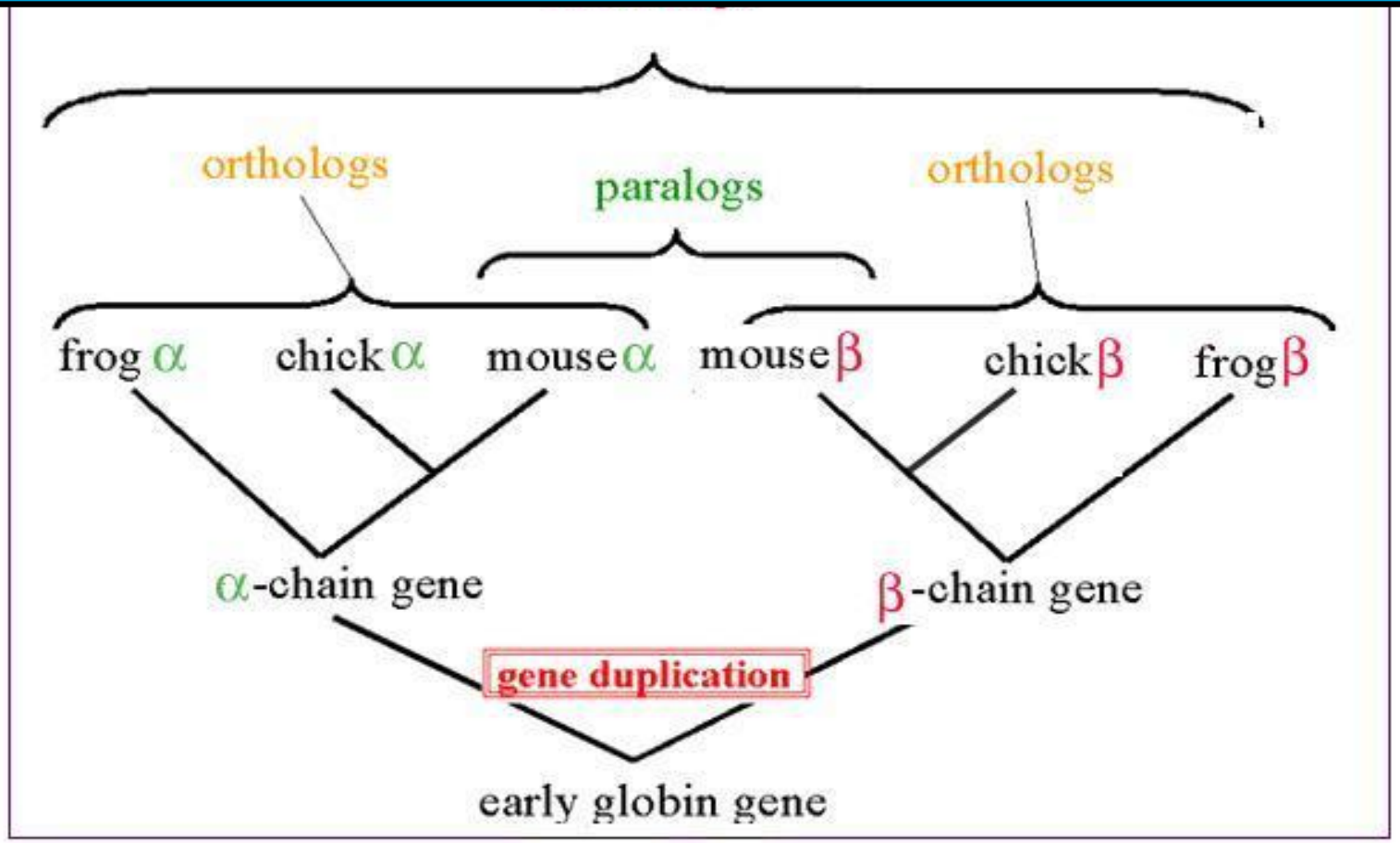
Common ancestor



- Similarity is a quantitative value
  - Identity

- Homology is a qualitative trait
  - refers to an evolutionary relation between sequences

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

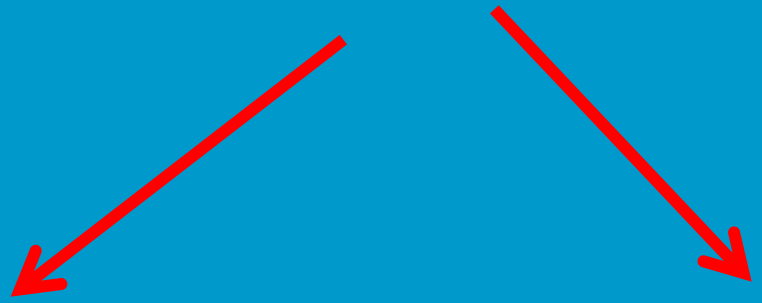
to high similarity

Identity, Similarity, Homology

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

Homology is established on the basis of measured identity or similarity

ALIGNMENT BETWEEN SEQUENCES



SIMILARITY

IDENTITY

AAKKQW

6 characters

AAKQW

5 characters

AAKKQW
AA KQW

6 characters
5 characters

AAKKQW

6 characters

AA KQW

5 characters

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

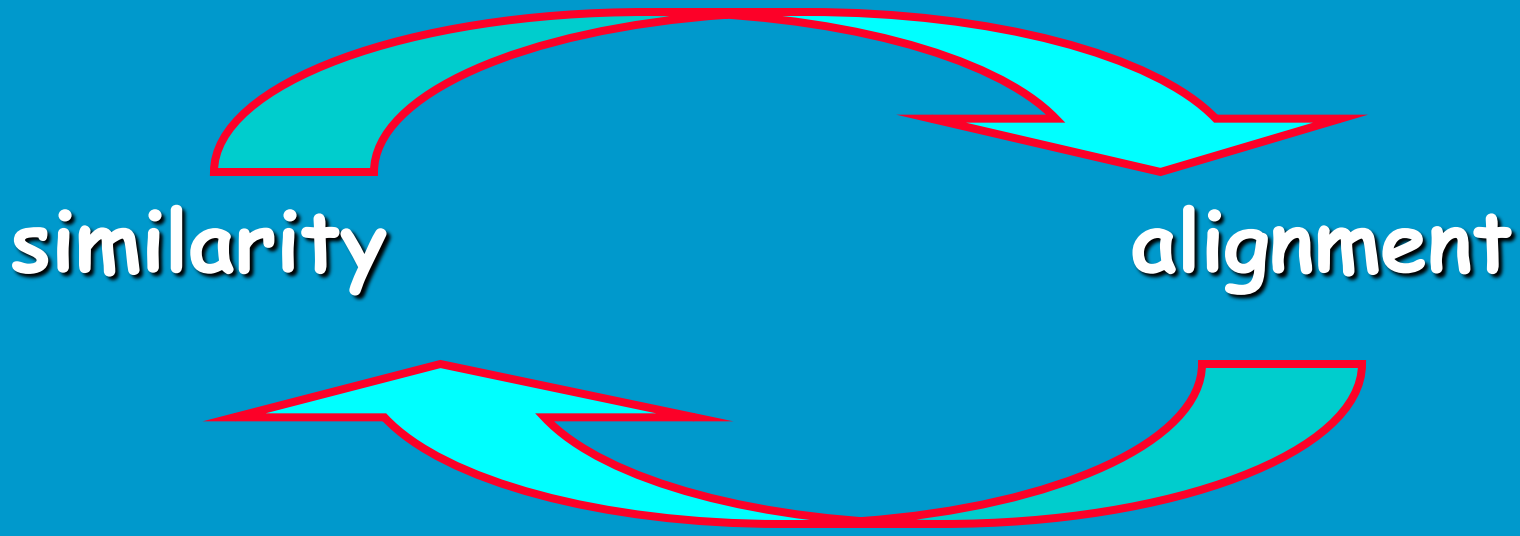
SIMILARITY SEARCHING IS BASED ON ALIGNMENT BETWEEN SEQUENCES.

- 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 MATCHES (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.

```
CGCTTCGGACGAAATCGCATCAGCATACGATCGCATGCCGGGCGGGATAAC

CGAAATCGCATCAGCATACGATCGCATGC
```

the query sequence is moved on the subject sequence.

```
CGCTTCGGACGAAATCGCATCAGCATACGATCGCATGCCGGGCGGGATAAC
||      |     |
CGAAATCGCATCAGCATACGATCGCATGC
```

```
CGCTTCGGACGAAATCGCATCAGCATACGATCGCATGCCGGGCGGGATAAC
          |    | |   |   | |   | |
 CGAAATCGCATCAGCATACGATCGCATGC
```

```
CGCTTCGGACGAAATCGCATCAGCATACGATCGCATGCCGGGCGGGATAAC
  |          |          |
  CGAAATCGCATCAGCATACGATCGCATGC
```

```
CGCTTCGGACGAAATCGCATCAGCATACGATCGCATGCCGGGCGGGATAAC
          ||  |    ||||   |      |    |
    CGAAATCGCATCAGCATACGATCGCATGC
```

```
CGCTTCGGACGAAATCGCATCAGCATACGATCGCATGCCGGGCGGGATAAC
         |     |||       |   ||    ||
     CGAAATCGCATCAGCATACGATCGCATGC
```

```
CGCTTCGGACGAAATCGCATCAGCATACGATCGCATGCCGGGCGGGATAAC
      ||  |                   |    ||    ||
      CGAAATCGCATCAGCATACGATCGCATGC
```

```
CGCTTCGGACGAAATCGCATCAGCATACGATCGCATGCCGGGCGGGATAAC
        ||          ||  ||     |
       CGAAATCGCATCAGCATACGATCGCATGC
```

```
CGCTTCGGACGAAATCGCATCAGCATACGATCGCATGCCGGGCGGGATAAC
            |    |          |       |
        CGAAATCGCATCAGCATACGATCGCATGC
```

```
CGCTTCGGACGAAATCGCATCAGCATACGATCGCATGCCGGGCGGGATAAC
            ||
         CGAAATCGCATCAGCATACGATCGCATGC
```

```
CGCTTCGGACGAAATCGCATCAGCATACGATCGCATGCCGGGCGGGATAAC
          ||||||||||||||||||||||||||||
         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.

GLOBAL OR LOCAL ALIGNMENT?

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

AAKKQW

6 characters

AAKQW

5 characters

AAKKQW
AA KQW

6 characters
5 characters

AAKKQW

6 characters

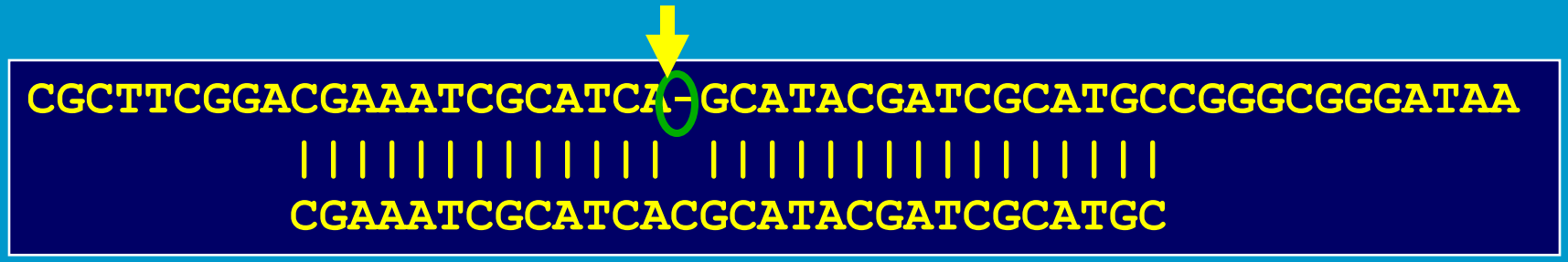
AA KQW

5 characters

Simple alignment often does not work

AAKKQW — 6 characters

AA KQW — 5 characters



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

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).

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

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
   |||||||||| |||||||||| |||||| 
CHKIPLMTRWDQ-QESDFGHKLPVIYTREW
```

Score = 25 - 2

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

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

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

Score = 25 – 1-(1+(0.1*3)

QUANTITATIVE EVALUATION (SCORE)

OF SIMILARITY

OF AN ALIGNMENT

Comprehensive alignment score

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

PERCENT SEQUENCE IDENTITY (PID)

THERE IS NO UNIVERSAL DEFINITION

- "PID1":100 * (identical positions) / (aligned positions + internal gap positions)
- "PID2":100 * (identical positions) / (aligned positions)
- "PID3":100 * (identical positions) / (length shorter sequence)
- "PID4":100 * (identical positions) / (average length of the two sequences)

PERCENTAGE IDENTITY (PID).

PID is also strongly length dependent, so, the shorter a pair of sequences is, the higher the PID you might expect by chance.