Multiple Sequence alignment (MSA)

PROTEIN MULTIPLE ALIGNMENTS

Give information about regions of conserved sequence

Useful for:

1. Function prediction
2. Structure prediction
3. Identification of new members in family proteins
4. Test and function modification in specific proteins.

Useless in two extreme cases:

- Very similar sequences, which have had no time for divergence
- Sequences which have diverged a lot and have no similar regions

MSA APPLICATIONS

Application	*Procedure*
Extrapolation	A good multiple alignment can help convincing you that an uncharacterized sequence is really a member of a protein family.
Phylogenetic analysis	If you carefully chose the sequences to include in your multiple alignment, you can reconstruct the history of these proteins.
Pattern Identification	By discovering very conserved positions you can identify a region that is characteristic of a function (in proteins or in nucleic acid sequences).
Domain identification	It is possible to turn a multiple sequence alignment into a profile that describes a protein family or a protein domain. You can use this profile to scan databases for new members of the family.
DNA regulatory elements	You can turn a DNA multiple alignment of a binding site into a weight matrix and scan other DNA sequences for potential similar binding sites.
Structure prediction	A good multiple alignment can give you an almost perfect prediction of your protein secondary structure for proteins or RNA. Sometimes it can also help building a 3-D model.
PCR analysis	A good multiple alignment can help you identifying the less degenerated portions of a protein family
nsSNP	Identify the nsSNP that are the most likely to alter the function

MULTIPLE ALIGNMENTS ADVANTAGES

Advantages: Can reveal information which has not been found in simple sequence analysis.

It is better to go from more simple to more complex:

SIMPLE ANALYSIS ➔ MULTIPLE ANALYSIS

Groups of related sequences

Proteins can be related by means of homology or convergence, being multiple analysis the most adequate in both cases.

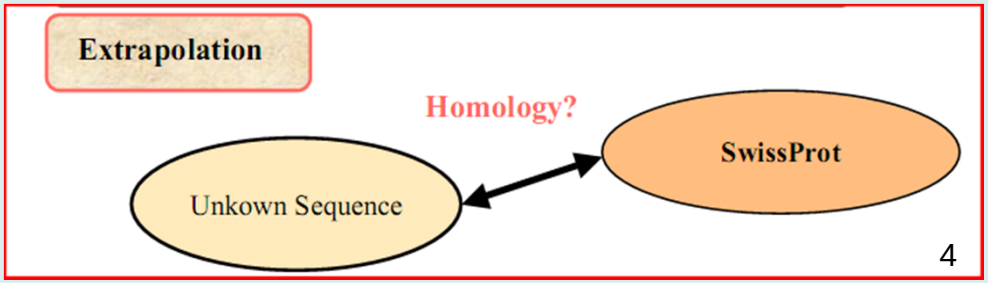
Homologue: common ancestor and function (normally)

Convergent: evolve independently

to have a common sequence

which typically has a common

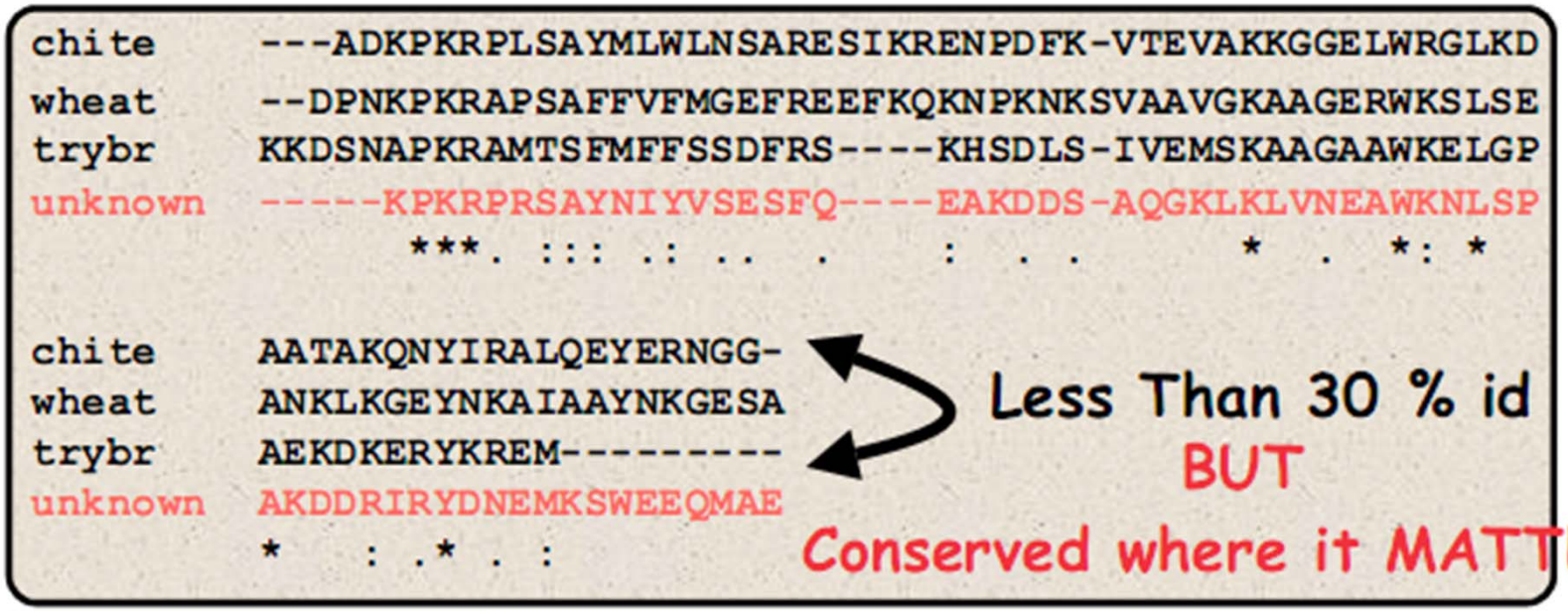
function or common structure.



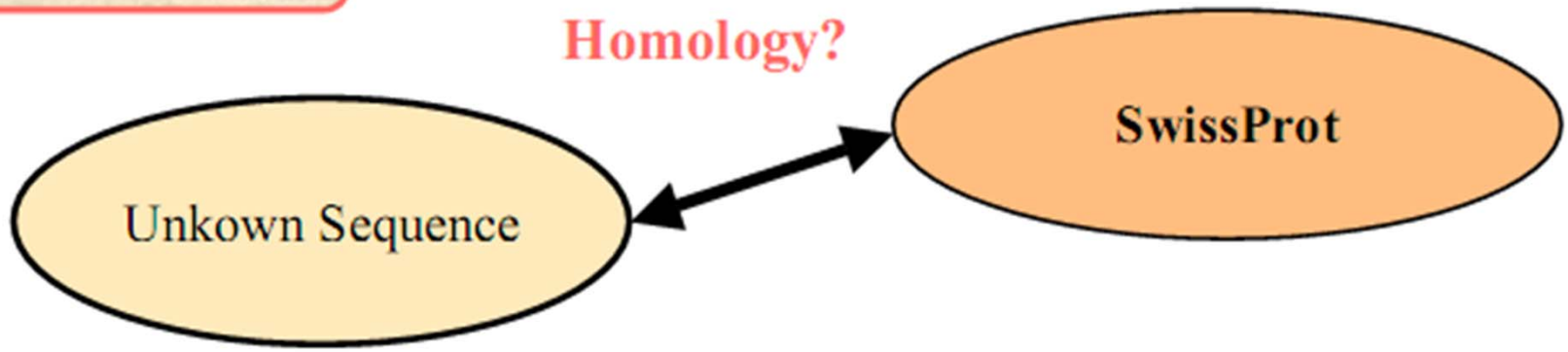
MAIN CRITERIA FOR BUILDING MSA

Criterion	*Meaning*
Structure similarity	**Amino acids that play the same role in each structure are in the same column.** Structure superposition programs are the only ones that use this criterion.
Evolutionary similarity	**Amino acids or nucleotides related to the same amino acid (or nucleotide) in the common ancestor of all the sequences are put in the same column.** No automatic program explicitly uses this criterion, but they all try to deliver an alignment that respects it.
Functional similarity	**Amino acids or nucleotides with the same function are in the same column.** No automatic program explicitly uses this criterion, but if the information is available, you can force some programs to respect it or you can edit your alignment manually.
Sequence similarity	**Amino acids in the same column are those that yield an alignment with maximum similarity.** Most programs use sequence similarity because it is the easiest criterion. When the sequences are closely related, structure, evolutionary and functional similarities are equivalent to sequence similarity.

MSA UTILITY: EXTRAPOLATION

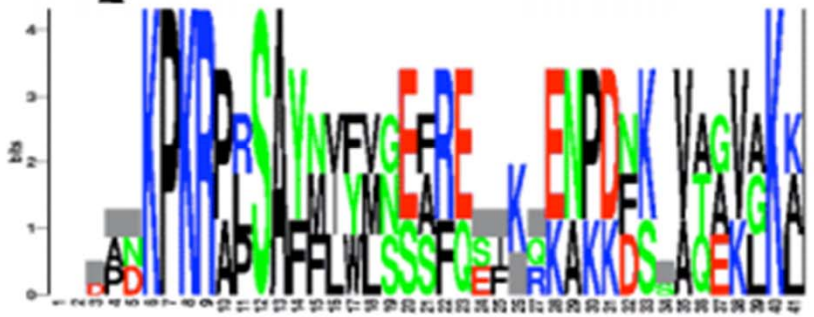


Extrapolation



MSA UTILITY: PATTERNS

```
chite   ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat   --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr   KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse   -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
              ***. ::: .: ..  .    :  . .      *  .  *: *

chite   AATAKQNYIRALQEYERNGG-
wheat   ANKLKGEYNKAIAAYNKGESA
trybr   AEKDKERYKREM---------
mouse   AKDDRIRYDNEMKSWEEQMAE
        *   : .* . :
```



Prosite Patterns

P-K-R-[PA]-x(1)-[ST]...

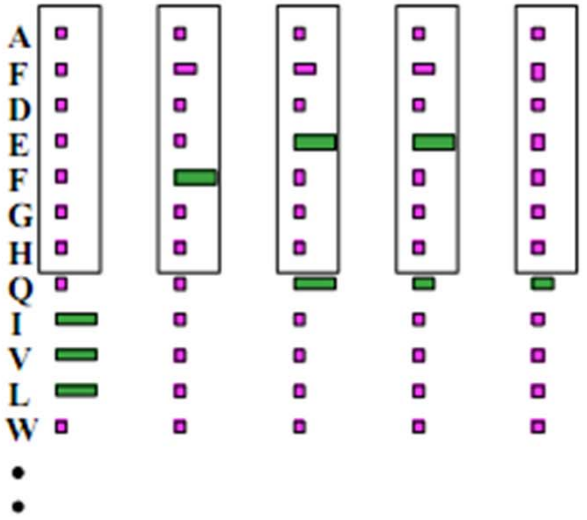
MSA UTILITY: PROFILES

```
chite   ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat   --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr   KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse   -----KPKRPRSAYNIYVSESFQ----EAKDDS-IQGKLKLVNEAWKNLSP
              ***. ::: .: ..   .    :  . .      *  .  *  *

chite   AATAKQNYIRALQEYERNGG-
wheat   ANKLKGEYNKAIAAYNKGESA
trybr   AEKDKERYKREM---------
mouse   AKDDRIRYDNEMKSWEEQMAE
        *   : .* . :
```

L?

K>R

Prosite Profiles



• • •

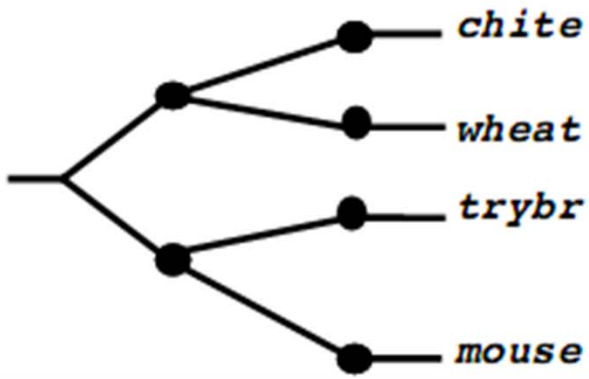
-More Sensitive

-More Specific

MSA UTILITY: PHYLOGENY

```
chite   ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat   --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr   KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse   -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
              ***. ::: .: ..  .    :  . .      *  .  *: *

chite   AATAKQNYIRALQEYERNGG-
wheat   ANKLKGEYNKAIAAYNKGESA
trybr   AEKDKERYKREM---------
mouse   AKDDRIRYDNEMKSWEEQMAE
        *   : .* . :
```

Phylogeny



-Evolution

-Paralogy/Orthology

MSA UTILITY: STRUCTURE PREDICTION

```
chite   ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat   --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr   KKDSNAPKRAMSSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse   -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
              ***. .. .: ..  .    :  . .      *  .  *: *

chite   AATAKQNYIRALQEYERNGG-
wheat   ANKLKGEYNKAIAAYNKGESA
trybr   AEKDKERYKREM---------
mouse   AKDDRIRYDNEMKSWEEQMAE
        *   : .* . :
```

Struc. Prediction

Column Constraint

⇔

Evolution Constraint

⇔

Structure Constraint

HOW TO DISTINGUISH A GOOD MSA?

The problem: Same as the pairwise alignment problem:

1. We do NOT know how sequences evolve
2. We do NOT understand the relation between structures and sequences.

So….

a) How can I choose my sequences?

b) What is the best substitution matrix?

c) What about Insertions and Deletions?

d) What is the best method?

e) How can I use my alignment?

CHOOSING SEQUENCES

How to find sequences of related proteins

Usually, there is only one query sequence

Search in databases: BLAST

Selection using statistical parameters (E-value) and experience.

Use **experimental data**, if available, to construct the alignments (i.e. positions in a catalytic centre should be forced)

How many sequences are needed?

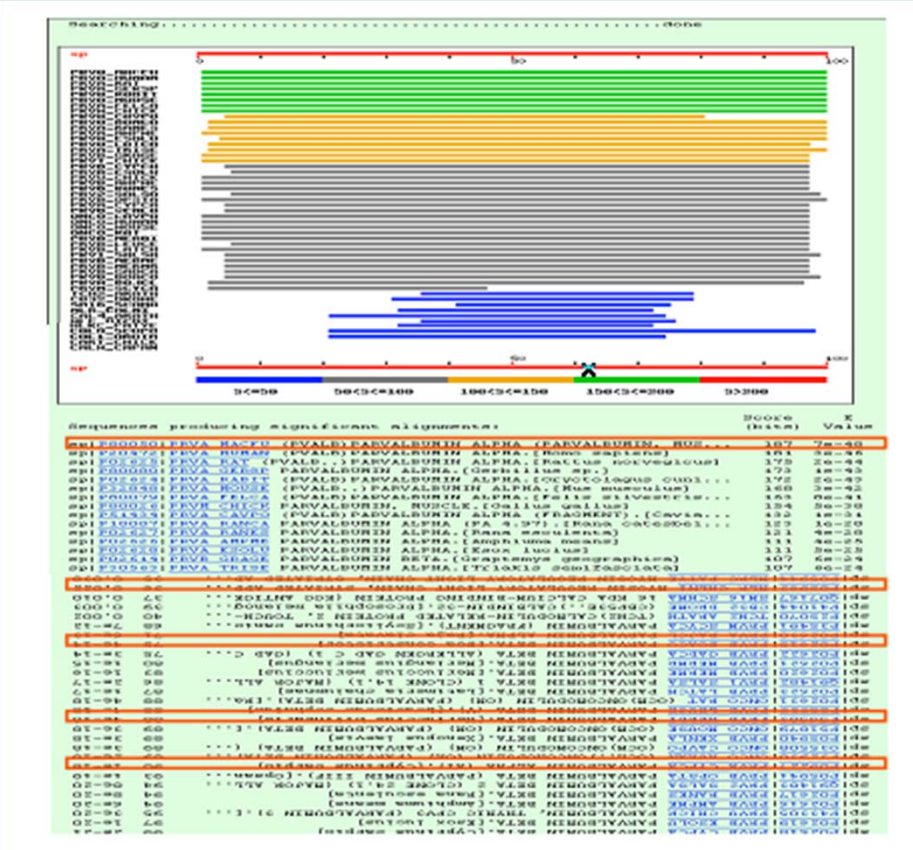
As many as there could be. Alignments of two or three sequences have a limited success.

Observe if sequence subgroups are formed and analyze separately.

Eliminate redundancy: highly similar sequences do not contribute to improve information

Divide and win: MOSAIC proteins

DO NOT CHOOSE IDENTICAL SEQUENCES!



Identical sequences brings no information
Multiple sequence alignments thrive on diversity

CHOOSING THE BEST SUBSTITUTION MATRIX

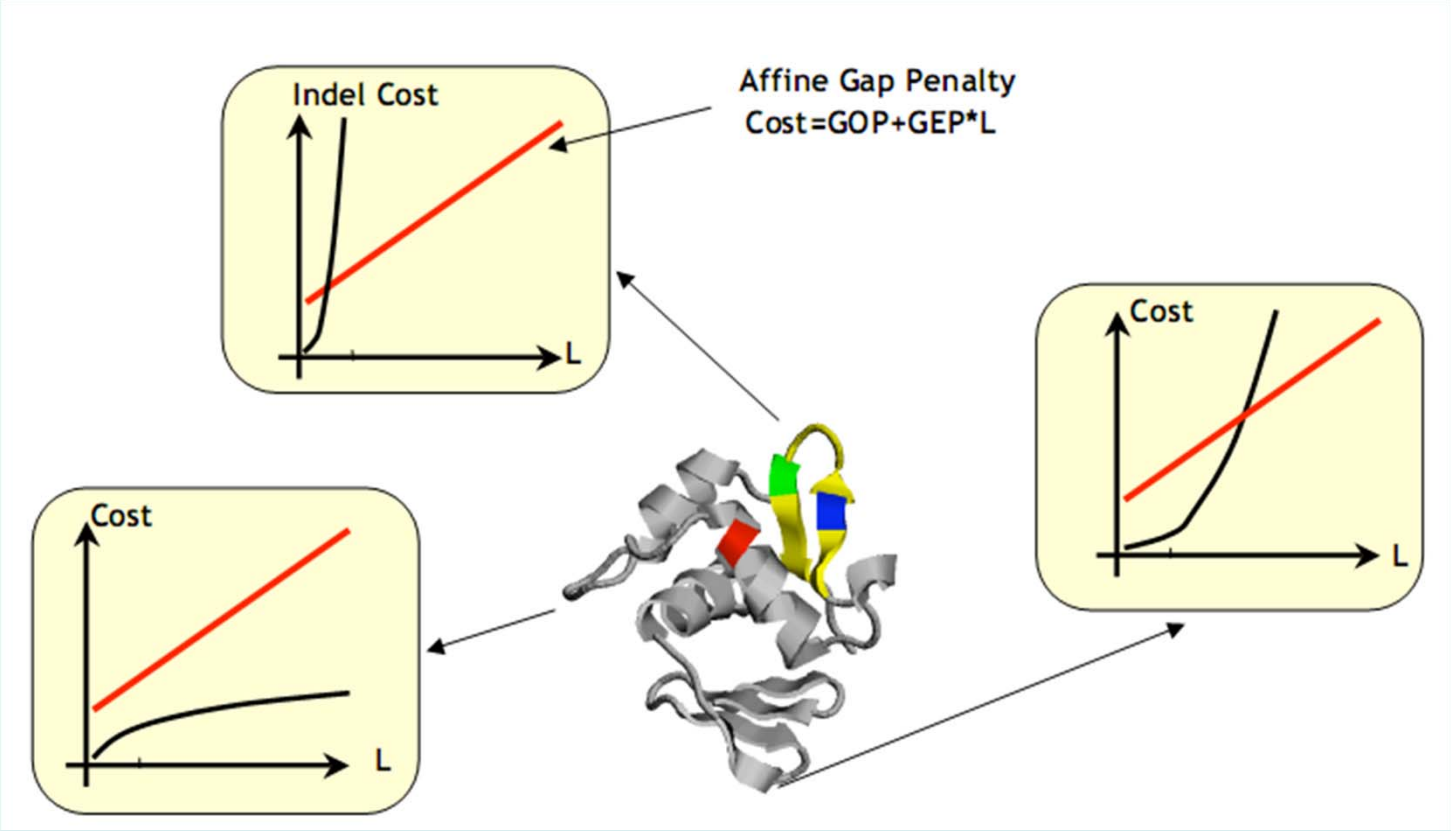
- Mutation rates depend on families

Family	S	N
Histone3	6.4	0
Insulin	4.0	0.1
Interleukin I	4.6	1.4
α–Globin	5.1	0.6
Apolipoprot. AI	4.5	1.6
Interferon G	8.6	2.8

Rates in Substitutions/site/Billion Years as measured on Mouse Vs Human (0.08 Billion years)

- Choosing the right matrix may be tricky
  - Gonnet250 > BLOSUM62 > PAM250
  - Depends on the family, the program used and its tuning

INSERTIONS AND DELETIONS (INDELS)



Multiple sequence alignments insertions and deletions
A big problem since the cost of gap open penalties (GOP) and extension (GEP) may be different in different parts of the protein

METHODOLOGY FOR MSA

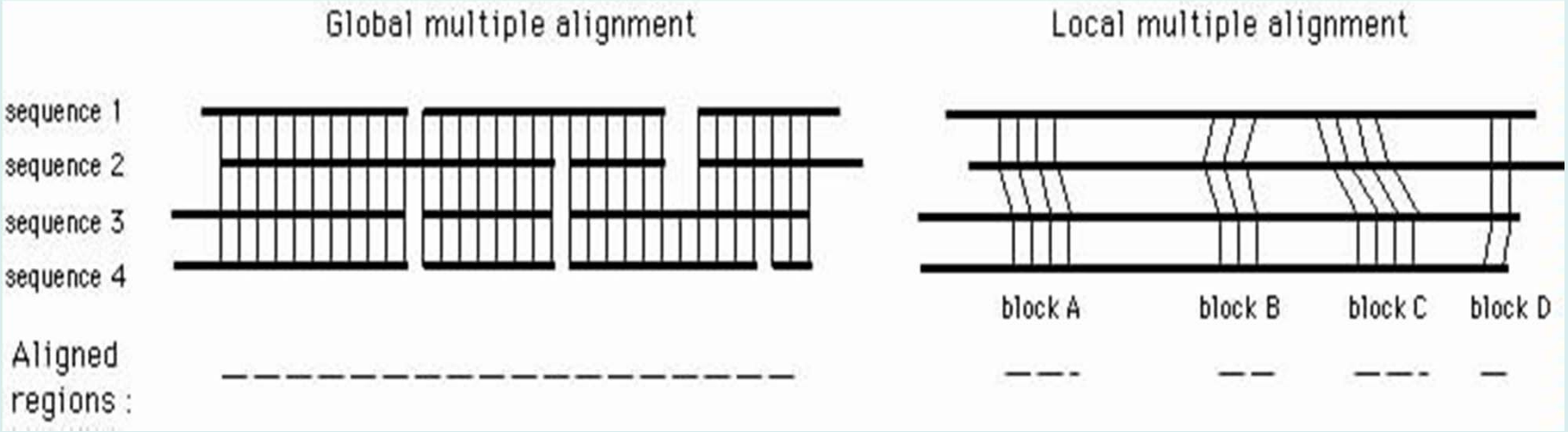
There are two mains ways of aligning sequences:

GLOBAL: whole length

LOCAL: obly certain regions

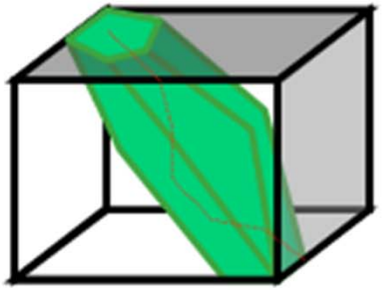
(valid for pair and multiple alignments)

Global alignment need the use of gaps



ALIGNMENT METHODS

1-Carillo and Lipman:

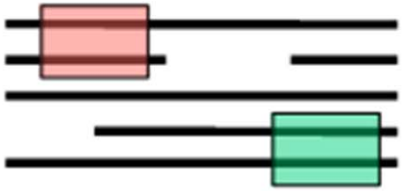


-MSA, DCA.

-Few Small Closely Related Sequence.

-Do Well When They Can Run.

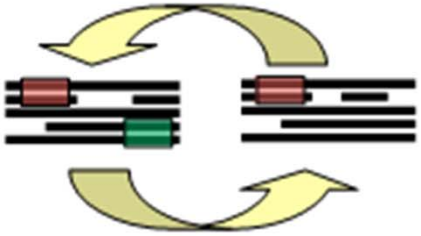
2-Segment Based:



-DIALIGN, MACAW.

-May Align Too Few Residues

3-Iterative:



-HMMs, HMMER, SAM.

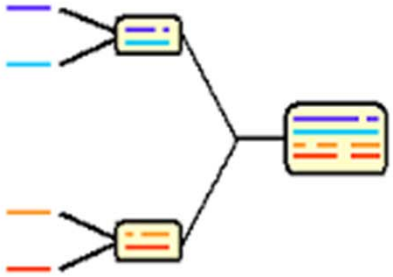
-Slow, Sometimes Inacurate

-Good Profile Generators

4-Progressive:

-ClustalW, Pileup, Multalign...

- Fast and Sensitive

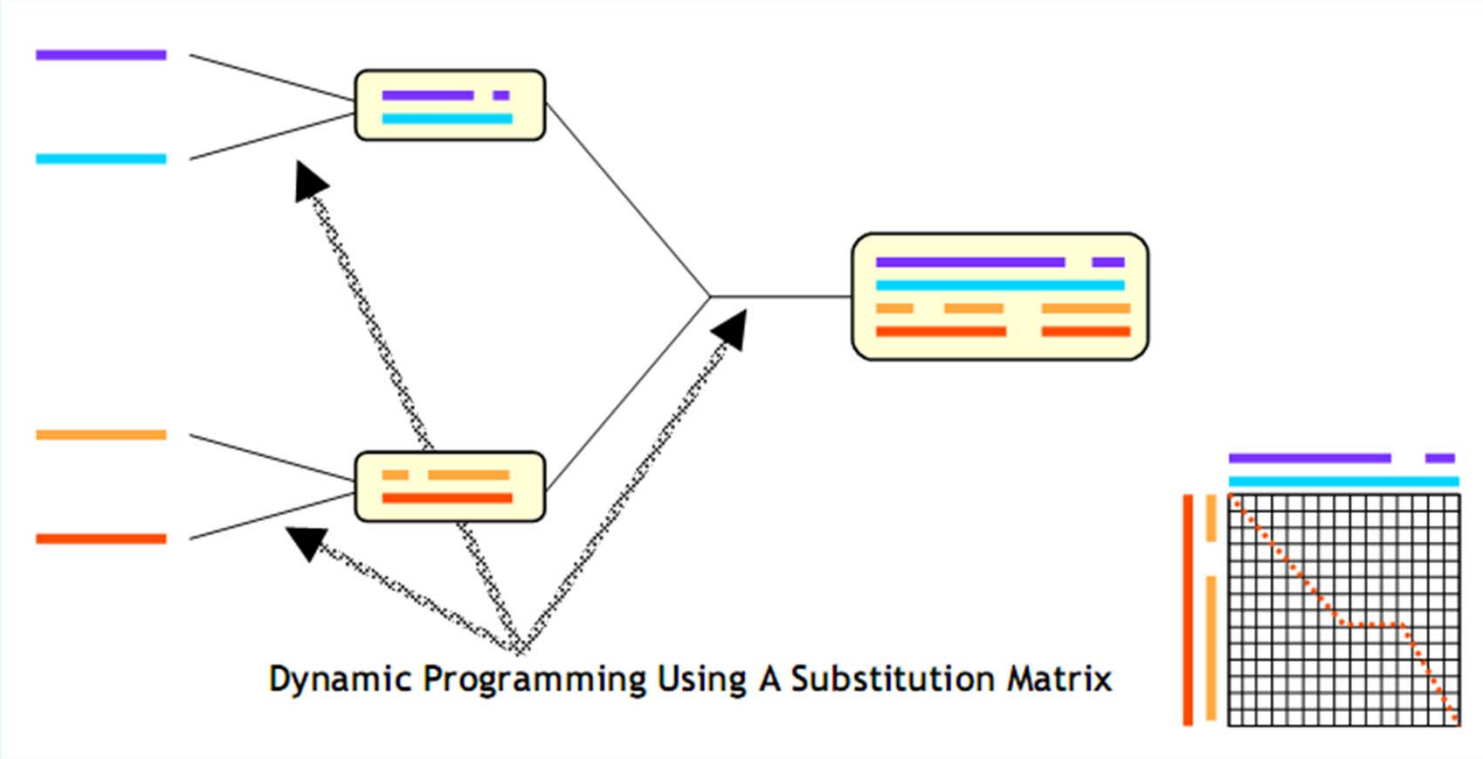


5-Mixtures:

-T-Coffee, MAFFT, MUSCLE, ProbCons, Psi-Praline,

- Very sensitive

PROGRESSIVE ALIGNMENT



Depends on the CHOICE of the sequences
Depends on the ORDER of the sequences (tree)
Depends on the parameters:

- Substitution matrix
- Penalties
- Sequence weight
- Tree making algorithm

ALIGNMENT TOOLS

Global

CLUSTALW – Automatic. Few adjustable parameters. Possibility of phylogenetic tree calculations

Local

BLOCKMAKER – Multiple alignments without GAPS. Excludes gappd sequences and only admit proteins with the same blocks in the same order.

MEME – Expected motifs should be specified. Motifs found in proteins do not have to show necessarily the same order.

MACAW - Semi-manual local multiple alignment detecting motifs in proteins.

Alignment Database Searches

BLIMPS	Searches in P and N databases using motifs and viceversa
MAST	Searches in databases using motifs
LAMA	Searches in motif databases using motifs

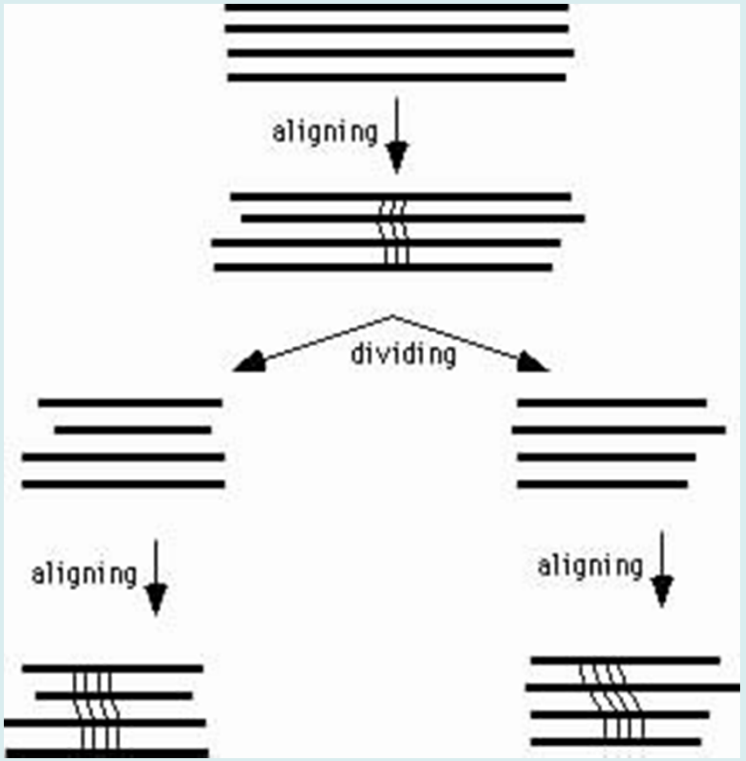
WHAT MAKES A GOOD ALIGNMENT…

The more divergent the sequences, the better
The fewer INDELS, the better
Nice ungapped blocks separated with INDELS
Different classes of residues within a block:

- Completely conserved (*)
- Size and hydropathy conserved (:)
- Size or hydropathy conserved (.)

If the same alignment is found in other databases, it would probably be correct.

If nothing works: **Divide and win**
Separate blocks and search for
good multiple alignments separately



Going further: Remote homologues
PSI-BLAST

The ultimate evaluation is a matter of personal judgment and knowledge

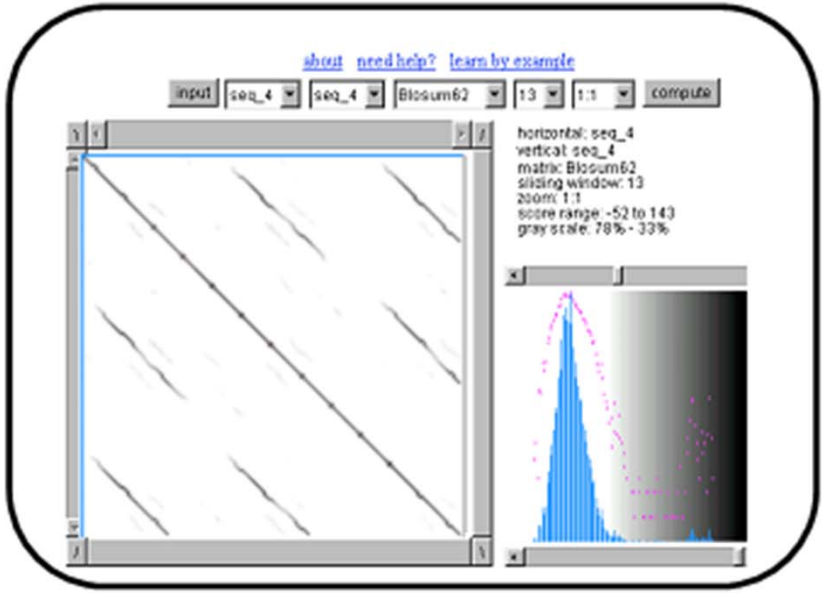
DO NOT USE TOO MANY SEQUENCES

* **It is difficult to *compute* big alignments**. Public servers do not have infinite resources. Your jobs may take a very long time to run (if it runs). For you, this makes it difficult to tune parameters and check alternatives.
* **It is difficult to *build* big alignments**. Multiple alignment programs are not very good at handling very large sets of sequences.
* **It is difficult to *display* big alignments**: you cannot print them and they clog your computer when you want to visualize them. If columns are longer than one page interpretation becomes impossible.
* **It is difficult to *use* big alignments**: tree building and structure prediction programs cannot handle them easily.
* **It is difficult to make *accurate* big alignments**. Multiple sequence alignment programs make mistakes. The curse is that **these mistakes do not add up, they multiply!** This is why it is easy to ruin an entire alignment with a tiny number of bad sequences. Of course the more sequences you have the more likely this is to happen.

BEWARE OF REPEATS

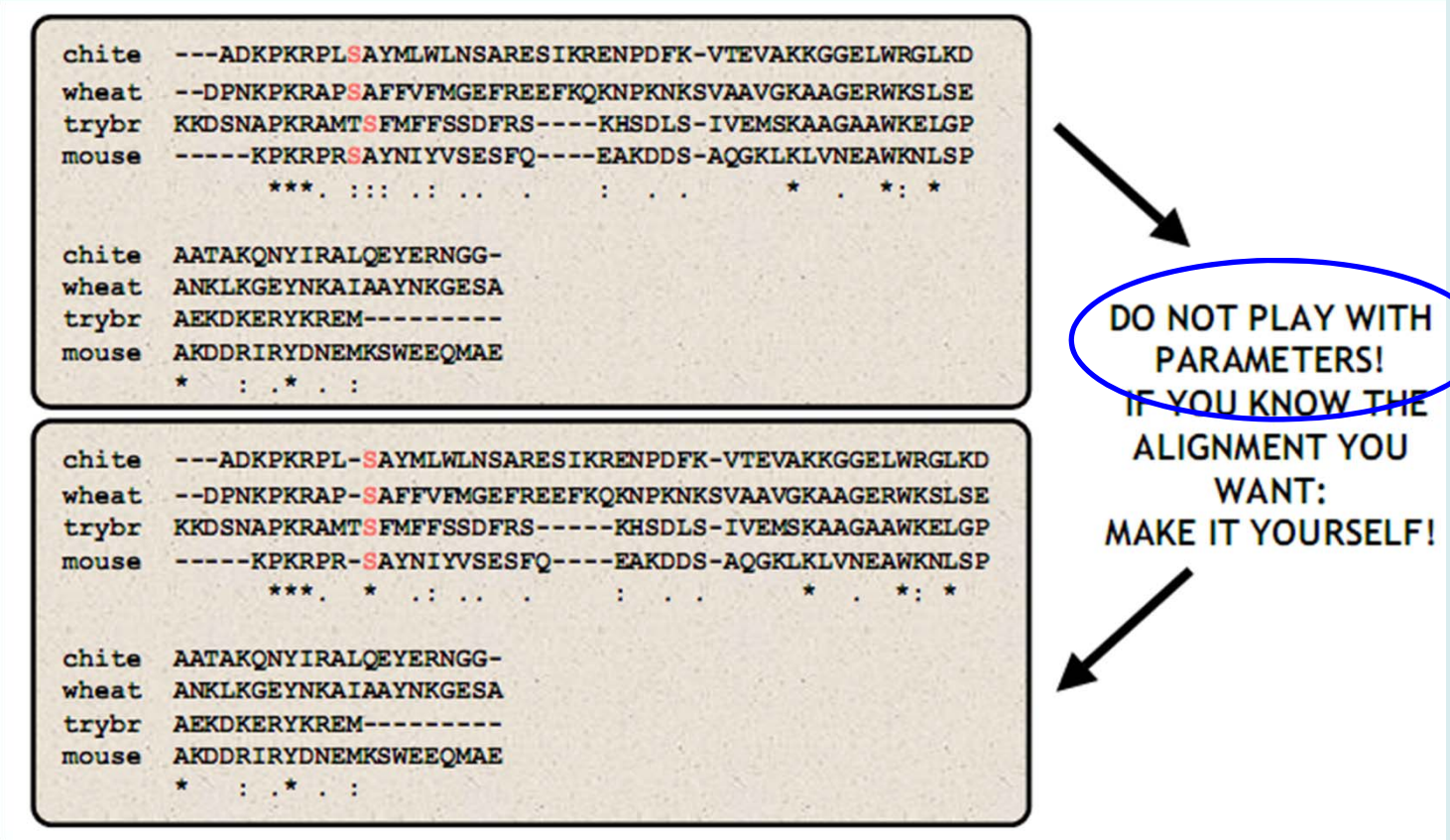
- There is a problem when two sequences do not contain the same number of repeats

- It is then better to manually extract the repeats and to align them separately. Individual repeats can be recognized using Dotlet or Dotter.



KEEP A BIOLOGICAL PERSPECTIVE

DO NOT OVERPLAY WITH PARAMETERS



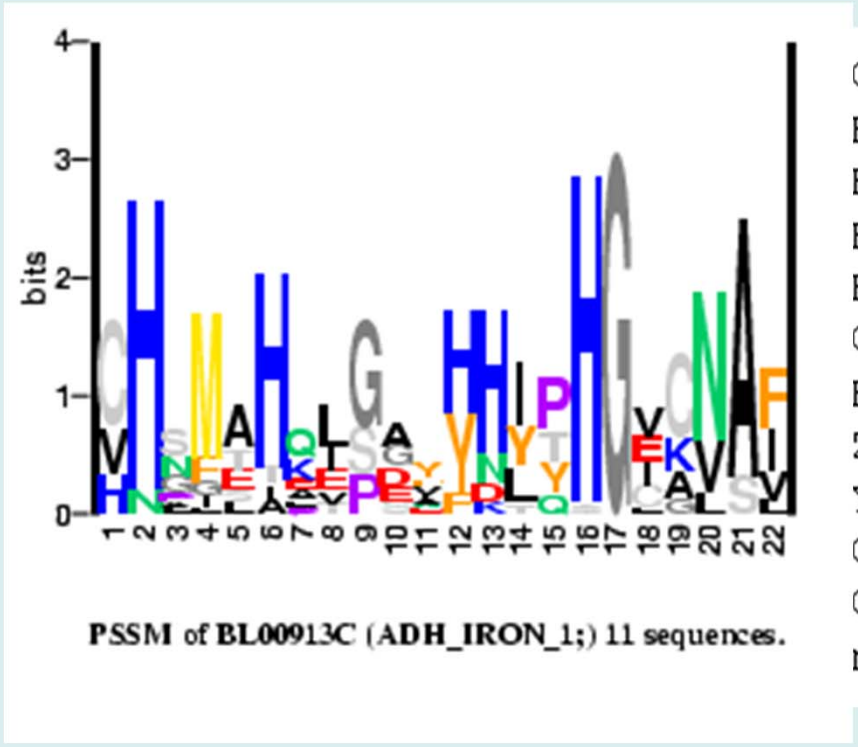
MULTIPLE ALIGNMENT VISUALIZATION

They are difficult to visualize

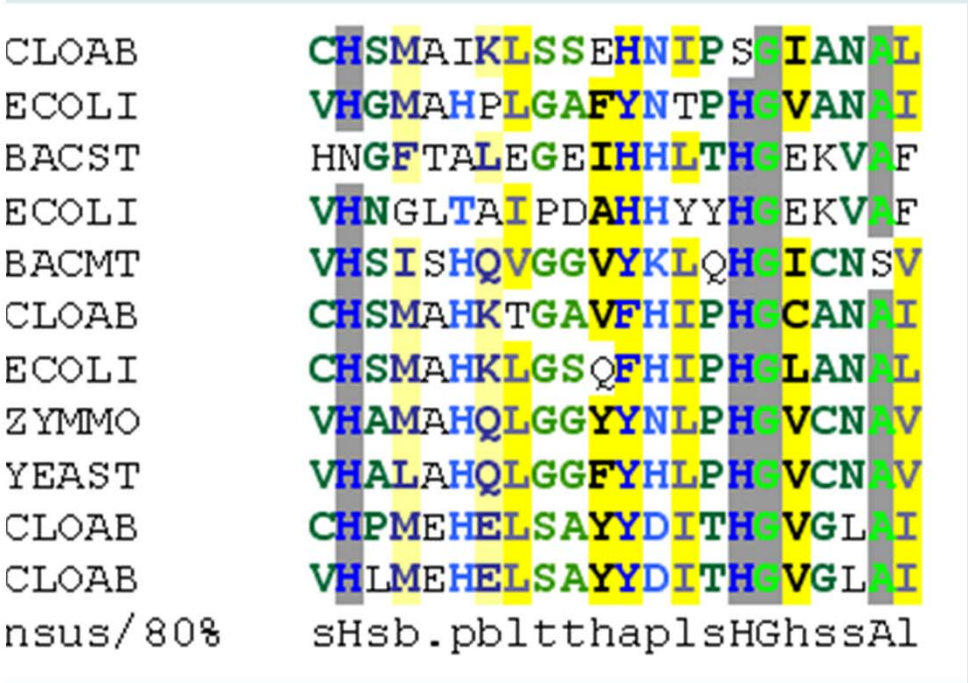
```
ADHE_CLOAB ( 720) CHSMAIKLSSEHNIPSGIANAL  66

FUCO_ECOLI ( 262) VHGMAHPLGAFYNTPHGVANAI  44

GLDA_BACST ( 259) HNGFTALEGEIHHLTHGEKVAF 100

GLDA_ECOLI ( 269) VHNGLTAIPDAHHYYHGEKVAF 100

MEDH_BACMT ( 259) VHSISHQVGGVYKLQHGICNSV  78

ADH1_CLOAB ( 258) CHSMAHKTGAVFHIPHGCANAI  47
ADHE_ECOLI ( 721) CHSMAHKLGSQFHIPHGLANAL  47

ADH2_ZYMMO ( 261) VHAMAHQLGGYYNLPHGVCNAV  36
ADH4_YEAST ( 263) VHALAHQLGGFYHLPHGVCNAV  41

ADHA_CLOAB ( 266) CHPMEHELSAYYDITHGVGLAI  50
ADHB_CLOAB ( 266) VHLMEHELSAYYDITHGVGLAI  49
//
```

GRAPHS AND COLOR

Graphic strategies are used: sequence logos or color



PSSM of BL00913C (ADH_IRON_1;) 11 sequences.



MSA CONCLUSIONS

- The best alignment method:
  - Your brain
  - The right data
- The best evaluation method:
  - Your eyes
  - Experimental information (Swiss-Prot)

- Choosing the sequences well is important
- Beware of repeated elements

- What can I conclude?
  - Homology => information extrapolation
- How can I go further?
  - Patterns
  - Profiles
  - HMMs
  - o…

Phylogenetic trees

PHYLOGENETIC ANALYSIS. MOLECULAR CLOCK

Punctual mutations are continuously accumulated inDNA and some of them lead to changes in aminoacids.

Time 0	1	2
Species000atgctagcta	Species001atgttagcta	Species001atgttagcta
	Species002atgctagcta	Species001atgataggta
Variacion		
0/10	1/10	2/10

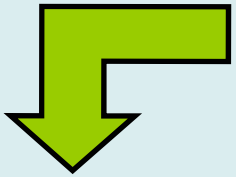
The period of time of two sequences/species separations is calculated

Time= Variation x velocity (year/mutation)

ex. 5 % variation and 50 Mill years/% variation = 250 Mill. years.

Constraints. Velocity differs with genes and organisms (at least one magnitude order) and could not be constant in time. Moreover, corrected mutations (backtracking) could subestimate the period of divergence.

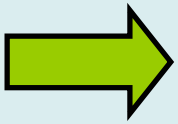
DISTANCE-BASED TREE BUILDING

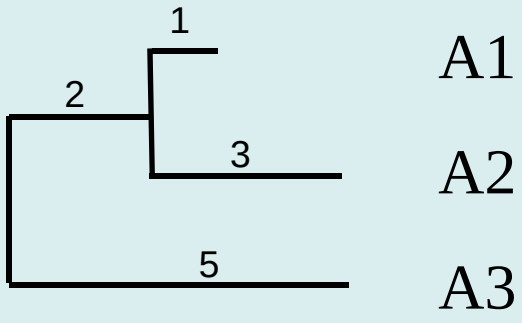


A1	MKFYSLPNFPEN
A2	MKYYKLPDLPDE
A3	MRFYTACENPRS

Distance matrix

	A2	A3
A1	4	8
A2		10

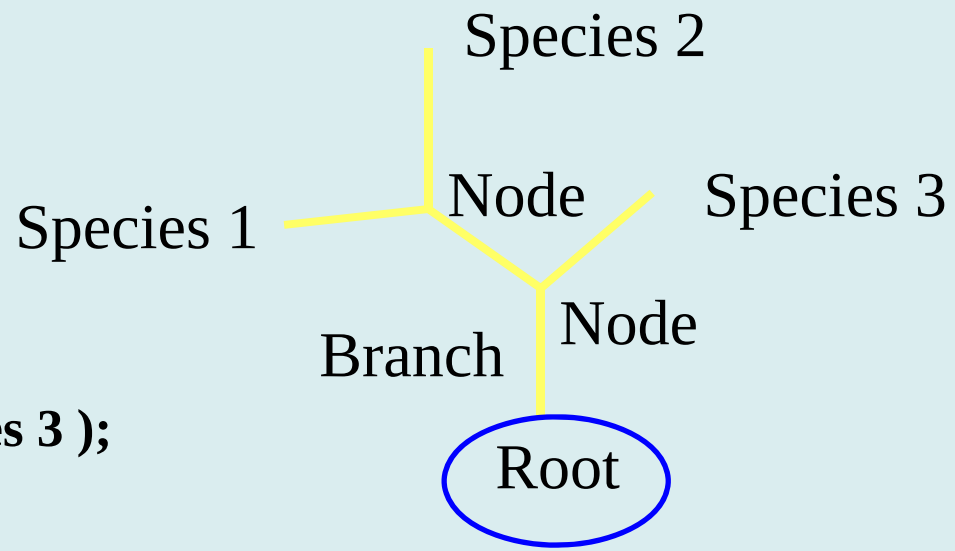




PHYLOGENETIC ANALYSIS. PROPERTIES

A tree is characterised by LEAVES, NODES, and BRANCHES

LEAVES (vertex) represent comparison among species or sequences.
NODES (vertex) are junctions and represent diferential events in species, from hypothetical ancertor sequences.
BRANCHES (edges) are always lineal and represent the sequence diversity and also the evolutive distance.
ROOT is optional and represents the hypothetical ancestor.



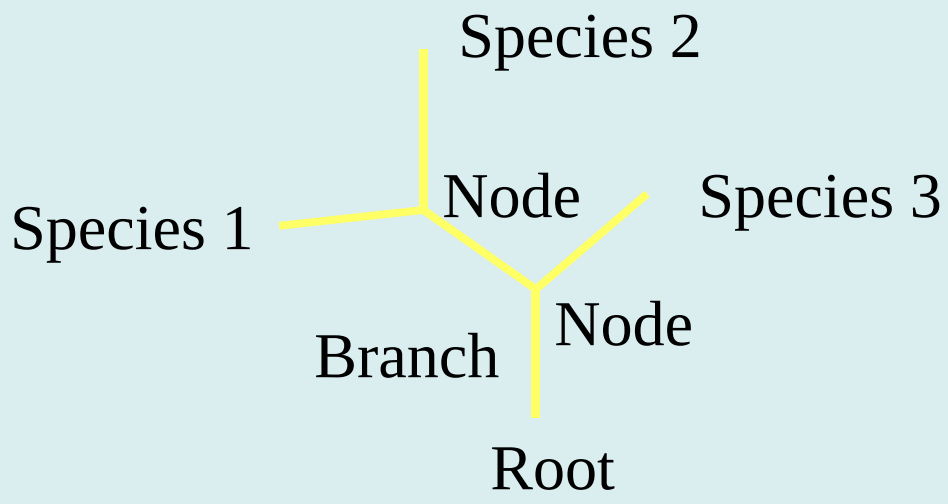
Lineal tree form:

((Species 1, Species 2), Species 3);

PHILOGENETIC ANALYSIS. PROPERTIES

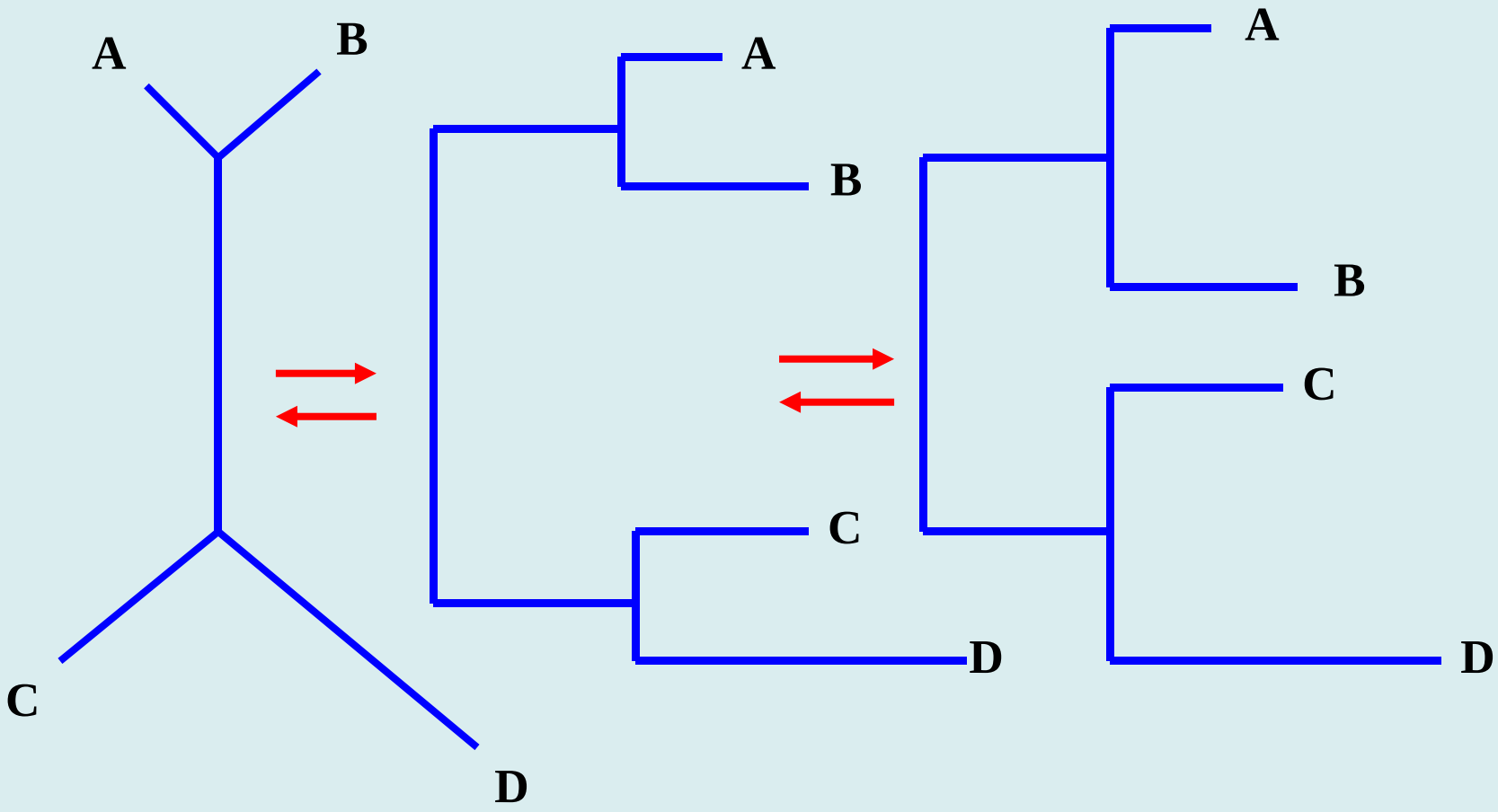
Phylogeny shows relation among species.
The presence of a root indicates de direction of evolution.

Branches length: Related to the substitutions in DNA or protein, and direct consequence of evolution time.



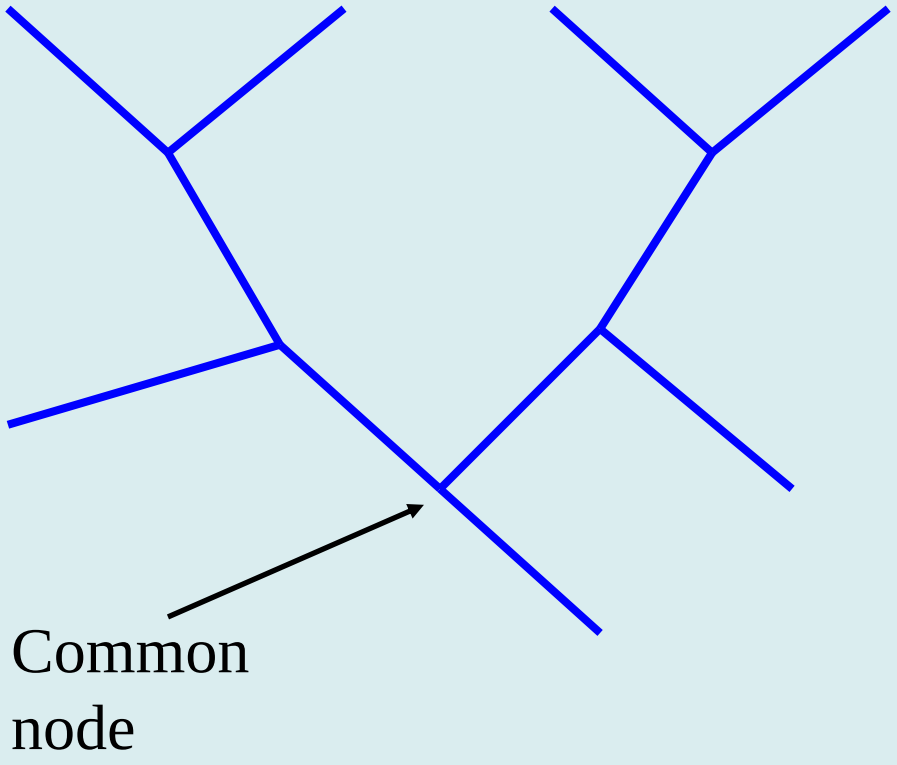
Lineal tree form with branches length:
((Species 1: 0.05, Species 2: 0.08), Species 3: 0.02): 0.03;

RADIAL TREE AND DENDOGRAM



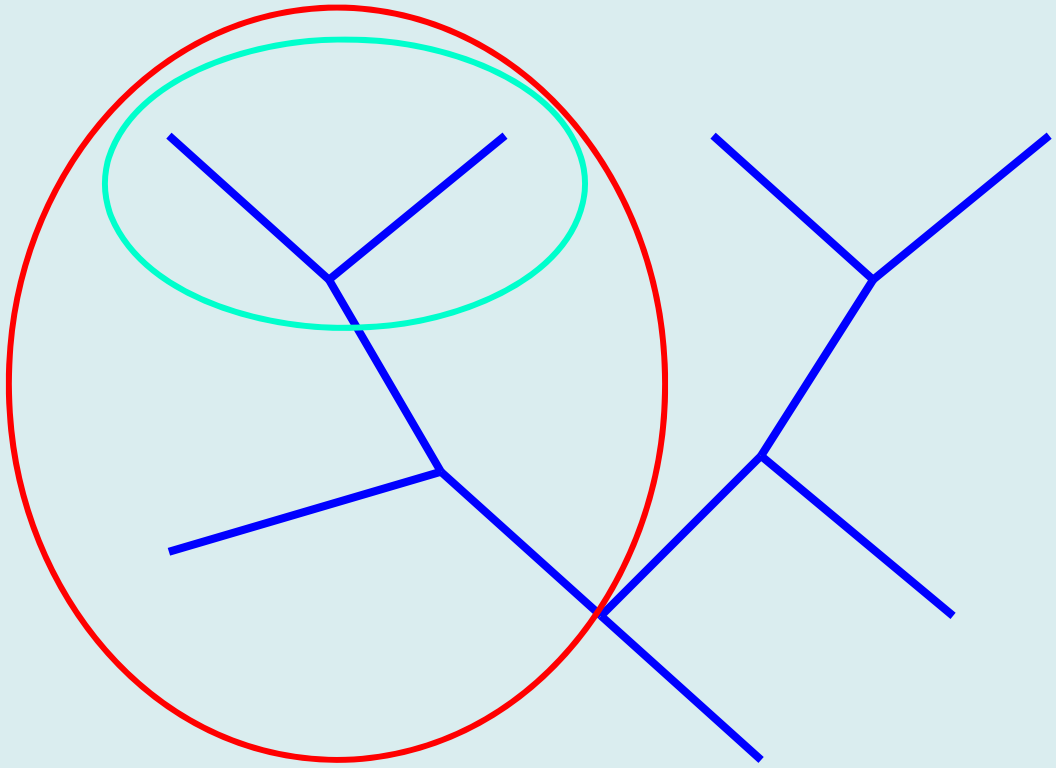
MORPHOGENETIC GROUP

A morphogenetic group is characterised by a common descentant from all the members, and all the members share a common node.



MONOPHILETIC GROUP

A monophiletic group is characterised by a common descendant for all the members (at least two) and all the members share a node.



PHILOGENETIC ANALYSIS METHODS

The best statistical methods are the slowest and analyse less sequences

```
ndvfull7b  CTTGCTATGG CTTGGGAATA ATACCCTCGA TCAGATGAGA GCCACTACAA
ndvfull7d  CTTGCTATGG CTTGGGAATA ATACCCTCGA TCAGATGAGA GCCACTACAA
ndvfull7c  CTTACTATGG CTTGGGAATA ACACCCTCGA TCAGATGAGA GCCACTACAA
ndvfull7a  CTTACTATGG CTTGGGAATA ATACCCTCGA TCAGATGAGA GCCACCACAA
 ndvfull8  CTTACTATGG CTTGGGAATA ATACCCTCGA TCAGATGAGA
```

Distance matrix

Distant matrix calculation

↓

Tree arranged by grouping methods, ex. Neighbour Joining.

Tree with branch length

Maximum parsimony

One or more trees with the same number of pieces

↓ ↓ ↓

Consensus evaluation

↓

Tree without branch length

Maximum likelihood

Trees arranged according to probability

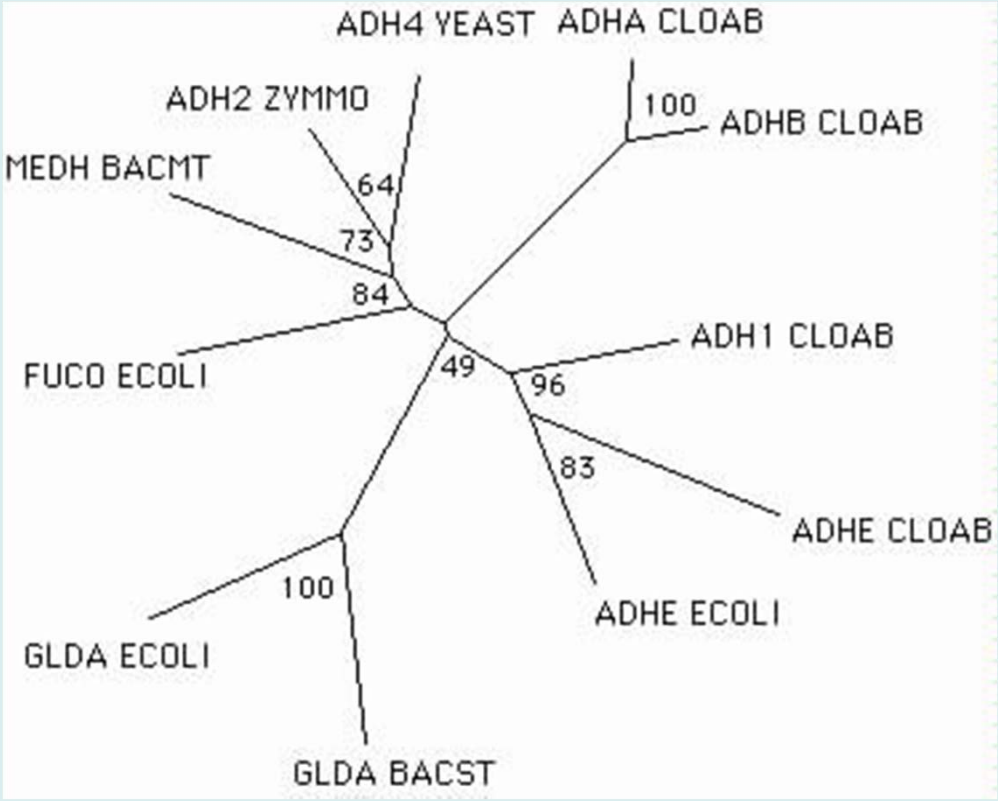
↓ ↓ ↓

Tree chosen with higher probability

↓

Tree with branch length

ESTATISTIC SIGNIFICANCE



Estimation of tree statistical significance: **Bootstrap values**

Shows the number of times each junction has been observed in a definite number of trials (Ex.: 100). The higher the fraction, the more confidence on the fact that the sequences which are within a branch form a group.