

MSA

MSA

- **Multiple sequence alignment** is a tool used to study closely related genes or proteins in order to find the evolutionary relationships between genes and to identify shared patterns among functionally or structurally related genes.

Introduction

- The principle of dynamic programming in pairwise alignment can be extended to multiple sequences. Unfortunately, the time required grows exponentially exponentially with the number of sequences and sequence lengths, this turns out to be impractical. Algorithms in use are heuristic and most are progressive/hierarchical.

Approaches to MSA

- Progressive alignment methods
- Iterative refinement methods
- **Progressive alignment methods**
- This approach is the most commonly used in MSA. Two sequences are chosen and aligned by standard pairwise alignment; this alignment is fixed. A third sequence is chosen and aligned to the first alignment. This process is iterated until all sequences have been aligned.

Progressive Alignment

- Progressive alignment (Feng and Doolittle, 1987) is a heuristic for multiple sequence alignment that does not optimize any obvious alignment score. The idea is to do a succession of pairwise alignments, starting with the most similar pairs of sequences and proceeding to less similar ones

Progressive Alignment Methods

This approach was applied in a number of algorithms, which differ in , How to choose the order to do the alignment , Whether the progression involves only alignment of sequences to a single growing alignment or whether subfamilies are built up on a tree structure and, at certain points, alignments are aligned to alignments . Procedure used to align and score sequences or alignments against existing alignments.

Progressive Alignment Methods

- Advantages
- Fast
- Efficient
- The resulting alignments are reasonable in many cases

- Disadvantages
- Heuristic
- Accuracy is very important
- Errors are propagated into the progressive steps.

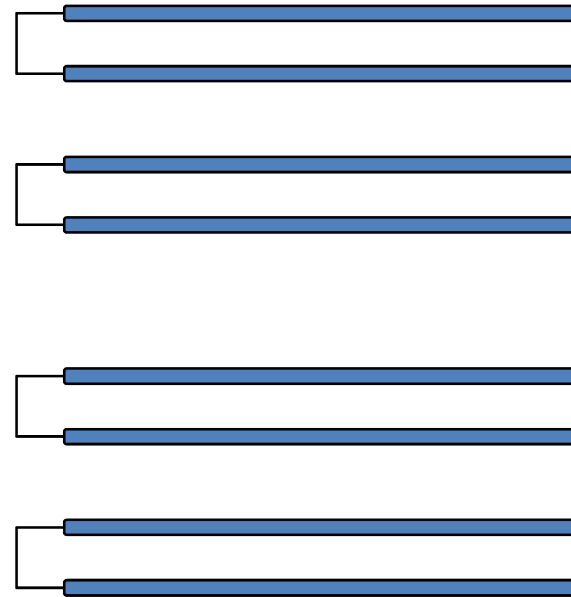
- Iterative methods
- A set of methods to produce MSAs while reducing the errors inherent in progressive methods are classified as "iterative" because they work similarly to progressive methods but repeatedly realign the initial sequences as well as adding new sequences to the growing MSA

- One reason progressive methods are so strongly dependent on a high-quality initial alignment is the fact that these alignments are always incorporated into the final result — that is, once a sequence has been aligned into the MSA, its alignment is not considered further. This approximation improves efficiency at the cost of accuracy.

- By contrast, iterative methods can return to previously calculated pairwise alignments or sub-MSAs incorporating subsets of the query sequence as a means of optimizing a general objective form such as finding a high-quality alignment score.

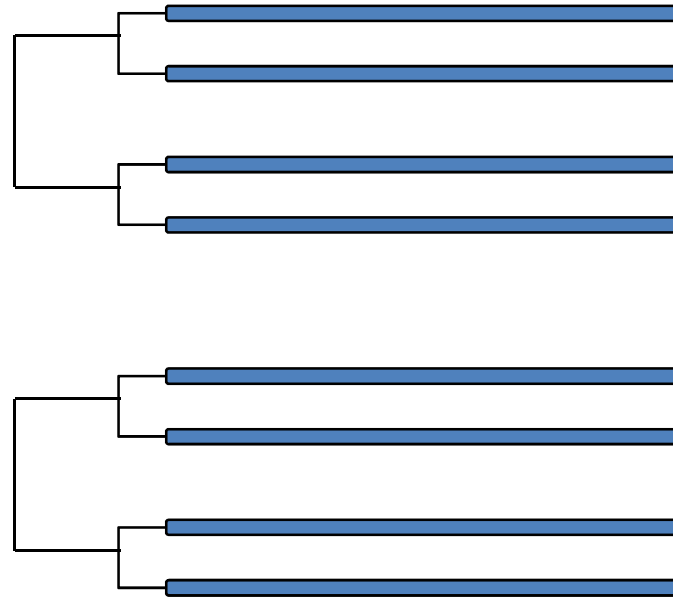
CLUSTAL

- Fix the alignment between pairs and treat as one sequence

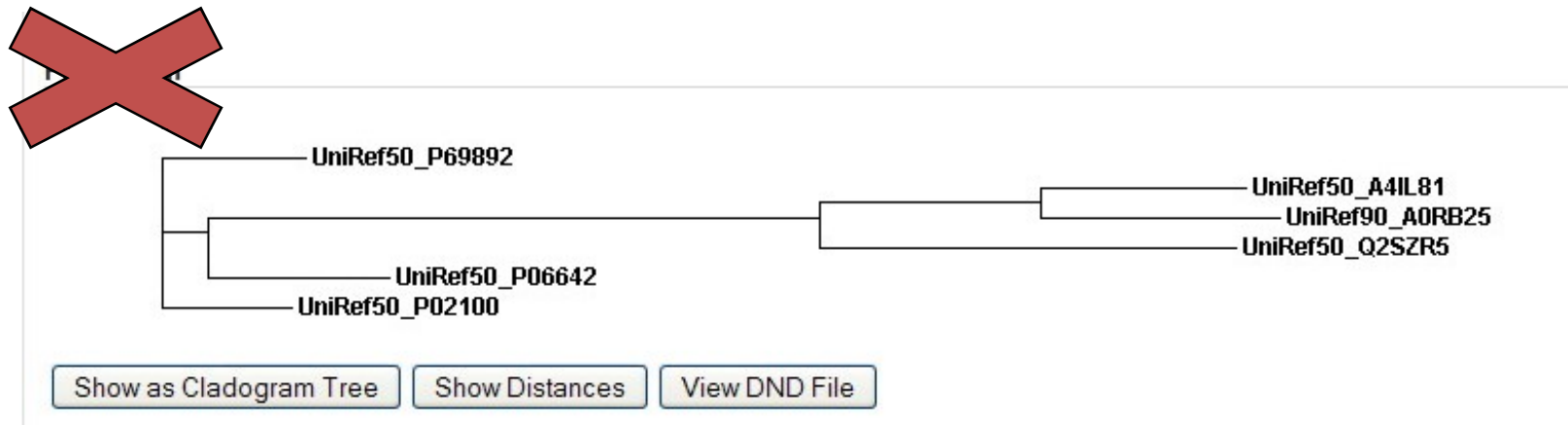


CLUSTAL


- Align your fixed pairs with each other



- Note, this is not a phylogram!
- Only a guide tree for the alignment



ClustalW at the EBI

EMBL-EBI  **EB-eye Search** All Databases

Databases	Tools	EBI Groups	Training	Industry	About Us	Help
■ Tools Home	Tools Index					
■ Tools A-Z	ID Mapping	Sequence Analysis				
■ ID Mapping	Literature	Sequence Analysis				
■ Literature	Microarray Analysis	Sequence Analysis encompasses the use of various bioinformatic methods for the analysis of sequence data in order to elucidate their relationships.				
■ Microarray Analysis	Protein Functional Analysis	Protein Functional Analysis encompasses the use of various bioinformatic methods for the analysis of protein structure and function.				
■ Protein Functional Analysis	Proteomic Services	Proteomic Services includes tools like Transeq which can help determine the protein coding regions of a sequence.				
■ Proteomic Services	Sequence Analysis	Align				
■ Sequence Analysis	Similarity & Homology	ClustalW2				
■ Align	Structural Analysis	CENSOR				
■ CENSOR	Tools - Miscellaneous	CpG Plot/CpGreport				
■ ClustalW2	Web Services	Dna Block Aligner Form				
■ CpG Plot/CpGreport	Downloads	GeneWise				
■ Dna Block Aligner Form		Kalign				
■ GeneWise		ClustalW2				
■ Kalign		MAFFT				
■ MAFFT		MUSCLE				
■ MUSCLE		Pepstats/ Pepwindow/ Pepinfo				
■ Pepstats/ Pepwindow/ Pepinfo		PromoterWise				
■ PromoterWise		SAPS				
■ SAPS		T-Coffee				
■ T-Coffee		Transeq				
■ Transeq						

ClustalW2 is highlighted with a red circle in the original image.

ClustalW

CLUSTAL W (1.83) multiple sequence alignment

```

beta globin  -----MVHLTPEEKSAVTALWGKVNVDD--EVGGEALGRLLVVYPWTQRFFESFG- 47
myoglobin    -----MGLSDGEWQLVLNVWVGKVEADIPGHGQEVLRIRLFKGHPEETLEKFDKFK- 48
neuroglobin  -----MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALPDLLPLFQYNCR 47
soybean      -----MVAFTTEKQDALVSSSFEEAFKANIPQYSVVFYTSILEKAPAAKDLFSFLA- 49
rice         MALVEDNNAVAVSFSFEEQEALVLKSWAILKKDSANIALRFFLKI FEVAPSASQMFSFLR- 59
              :   :   :   :   .   .   .   :   :   *   *   .

              ▽
beta globin  DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLS-----ELHCDKLHVDPE 102
myoglobin    HLKSEDEMKA SEDLKKHGATVLTALGGILKKKGHHEAEIKPLA-----QSHATKHKIPVK 103
neuroglobin  QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLA S---LGRKHRAVGVKLS 104
soybean      --NGVDPT--NPKLTGHA EKLFALVRDSAGQLKASGTVVADAA----LGSVHAQKAVTDP 101
rice         --NSDVPLEKNPKLKTHAMSVFVMTCEAA AQLRKAGKVTVRDTTLKRLGATHLKYGVGDA 117
              .   .   .   *   .::   :   :   :

beta globin  NFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHKYH----- 147
myoglobin    YLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
neuroglobin  SFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRGWDGE---- 151
soybean      QFVVVKEALLKTIKAAV-GDKWSD ELSRAWEVAYDELAAAIKKA----- 144
rice         HFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE--- 166
              :   :   ::   :   :   *   .   .   :
  
```

Note how the region of a conserved histidine (▼) varies depending on which of five prominent algorithms is used

Homology

vs.

Similarity

- Presence of similar features because of common decent
 - Cannot be observed since the ancestors are not anymore
 - Is inferred as a conclusion based on ‘similarity’
 - Homology is like pregnancy: Either one is or one isn’t! (Gribskov – 1999)
- Quantifies a ‘likeness’
 - Uses statistics to determine ‘significance’ of a similarity
 - Statistically significant similar sequences are considered ‘homologous’

ClustalW example

CLUSTAL W (1.83) multiple sequence alignment

```

beta globin  -----MVHLTPEEKSAVTALWGKVNVDD--EVGGEALGRLLVVYPWTQRFFESFG- 47
myoglobin    -----MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIIRLFKGHPEPLEKFDKFK- 48
neuroglobin  -----MERPEPELIRQSWRAVSRSPLEHGTIVLFARLFALEPDLLPLFQYNCR 47
soybean      -----MVAFTKQDALVSSSFSAFKANIPQYSVVFYTSILEKAPAAKDLFSFLA- 49
rice         MALVEDNNAVAVSFSEEQDALVLSWAILKKDSANIALRFFLKIFEVAPSASQMFSLR- 59
              :   :   :   :   .   .   .   :   *   *   .
              ▽
beta globin  DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLS-----ELHCDKLHVDPE 102
myoglobin    HLKSEDEMKAEDLKKHGATVLTALGGILKKKGHHEAEIKPLA-----QSHATKHKIPVK 103
neuroglobin  QFSSPEDCLSSPEFLDHIRKVMLVIDAANTNVEDLSSLEEYLA-----LGRKHRAVGVKLS 104
soybean      --NGVDPT--NPKLTGHAEKLFALVRDSAGQLKASGTVVADAA-----LGSVHAQKAVTDP 101
rice         --NSDVPLEKNPKLKTHTAMSVFVMTCEAAQRLKAGKVTVRDITLRLGATHLKYGVGDA 117
              .   .   .   *   .   :   :   :
              :
beta globin  NFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHKYH----- 147
myoglobin    YLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
neuroglobin  SFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRGWDGE----- 151
soybean      QFVVVKEALLKTIKAAV-GDKWSELSRAWEVAYDELAALAIKKA----- 144
rice         HFEVVKFALLDTIKKEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE--- 166
  
```

Summary of ClustalW method

- For years, ClustalW was the best method available.
- It's still a solid performer, provided that sequences are closely related and do not have significant structural differences
- Distinguishing characteristics:
 - Progressive alignment based on a guide tree
 - Gap parameters informed by hydrophobicity of amino acids and by previously inserted gaps
 - Amino acid substitution matrices derived from observed sequence divergence (different matrices for different groups).