

Clustal W

Introduction

- Clustal W is a general purpose multiple sequence alignment program for DNA or proteins.
- It produces biologically meaningful multiple sequence alignments of divergent sequence.
- It calculate the best match for the selected sequence ,and lines them up so that the identiies.

Algorithm

- Clustal W is a matrix-based algorithm, whereas tools like T-coffee are consistency-based.
- Clustal W has a fairly efficient algorithm that competes well against in other software.
- The first step to the algorithm is computing a rough distance matrix between each pair of sequence, also known as pairwise sequence alignment.

Accuracy and Results

- The algorithm ClustalW uses provides a close –to-optimal result almost every time.
- However, it does exceptionally well when the data set contains sequences with varied degrees of divergence.

Clustal Omega

- Clustal omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile –profile technique to generate alignment between three or more sequence.
- Ideally, might think to build up multiple alignments through weighted sum of pairs (pairwise scores)
- But this is too computationally intensive
 - And doesn't make much biological sense

Basic Steps

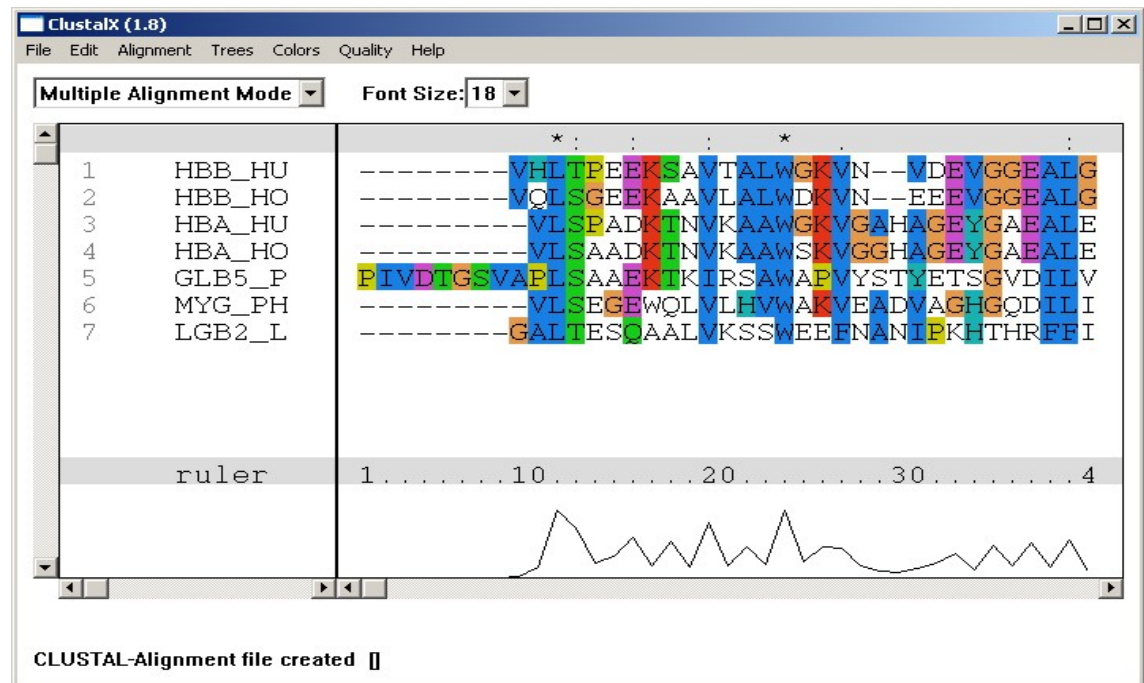
- Compute the pairwise alignments for all against all sequences.
- The similarities are stored in a matrix
- Convert the sequence similarity matrix values to distance measures , reflecting evolutionary distance between each pair of sequence.
- Construct a tree for the order in which pairs of sequence are to be aligned and combined with previous alignments.
- Progressively align the sequence together into each branch point of the guide tree, starting with the least distance pairs of sequence.

Uses of MSA

- Functional prediction
- Phylogeny
- Structural prediction
- Protein analysis
- To distinguish between orthology and paralogy

- Ideally, might think to build up multiple alignments through weighted sum of pairs (pairwise scores)
- But this is too computationally intensive
 - And doesn't make much biological sense
- So use heuristics and progressive alignment methods

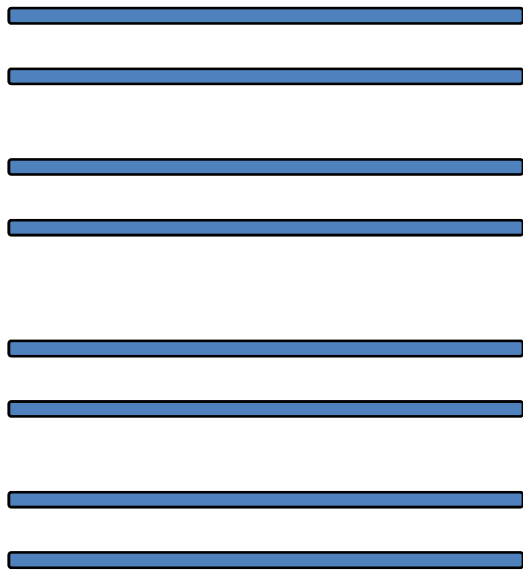
ClustalW



www.clustal.org

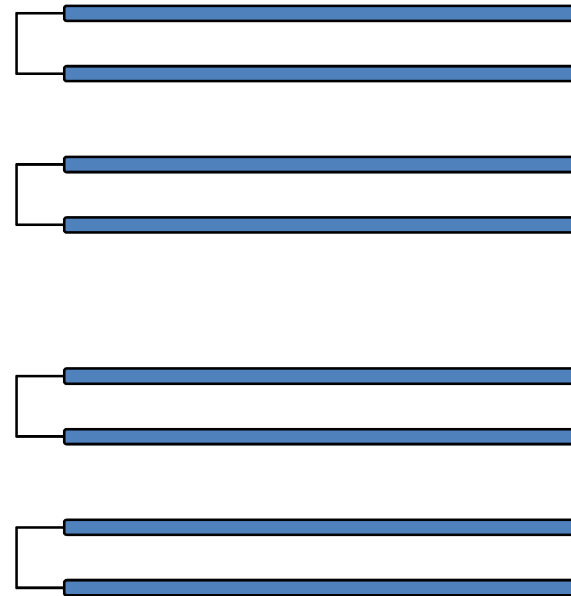
CLUSTAL

- Quick, pairwise alignment of all sequences
- Line up pairs, with the most similar first



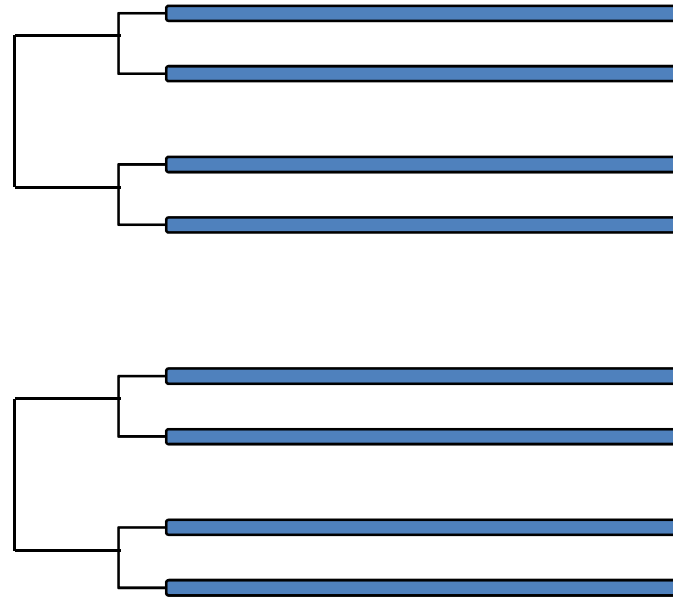
CLUSTAL

- Fix the alignment between pairs and treat as one sequence



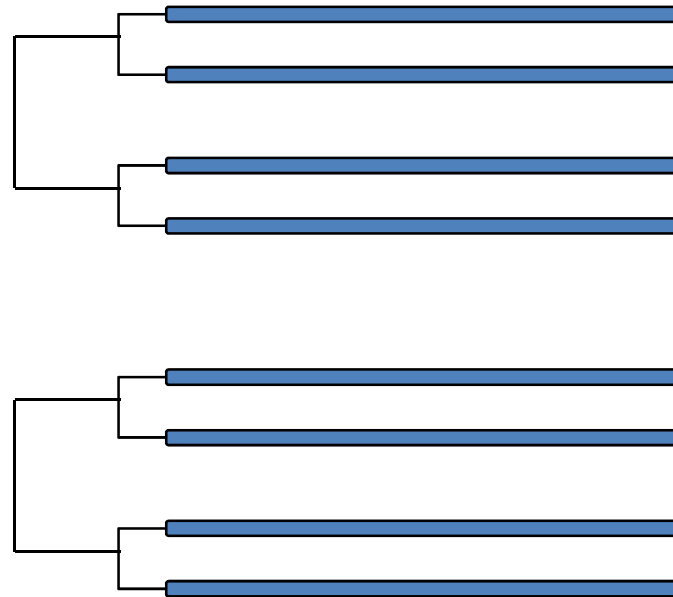
CLUSTAL

- Align your fixed pairs with each other

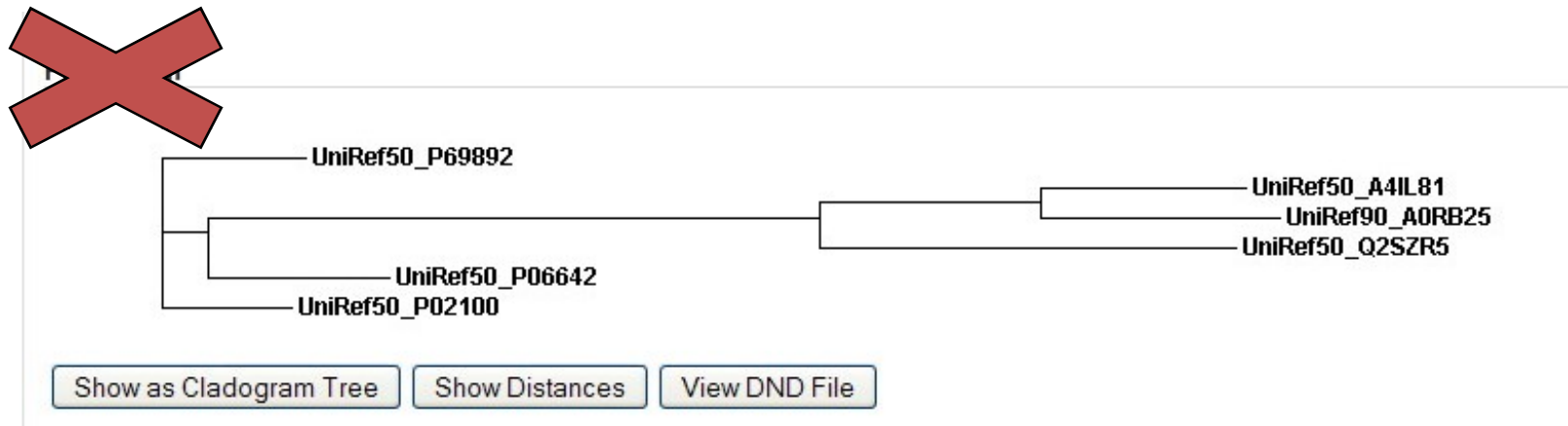


CLUSTAL

- Align your fixed pairs with each other



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



ClustalW at the EBI

The screenshot shows the EMBL-EBI website interface. At the top, there is a search bar with the text "Enter Text Here" and a dropdown menu set to "All Databases". Below the search bar is a navigation menu with tabs for "Databases", "Tools", "EBI Groups", "Training", "Industry", "About Us", and "Help". The "Tools" tab is selected, and a sidebar on the left lists various tool categories. The main content area displays a list of tools under the "Sequence Analysis" category. The "ClustalW2" tool is highlighted with a red circle. The tool's description is partially visible, mentioning "Structural analysis tools we have a".

Tools	EBI Groups	Training	Industry	About Us	Help
Tools Index					
ID Mapping		Sequence Analysis			
Literature		Sequence Analysis			
Microarray Analysis		Sequence Analysis			
Protein Functional Analysis		Sequence Analysis			
Proteomic Services		Sequence Analysis			
Sequence Analysis		Sequence Analysis			
Similarity & Homology		Sequence Analysis			
Structural Analysis		Sequence Analysis			
Tools - Miscellaneous		Sequence Analysis			
Web Services		Sequence Analysis			
Downloads		Sequence Analysis			
CENSOR		Sequence Analysis			
ClustalW2		Sequence Analysis			
CpG Plot/CpGreport		Sequence Analysis			
Dna Block Aligner Form		Sequence Analysis			
GeneWise		Sequence Analysis			
Kalign		Sequence Analysis			
MAFFT		Sequence Analysis			
MUSCLE		Sequence Analysis			
Pepstats/Pepwindow/Pepinfo		Sequence Analysis			
PromoterWise		Sequence Analysis			
SAPS		Sequence Analysis			
T-Coffee		Sequence Analysis			
Transeq		Sequence Analysis			

ClustalW

CLUSTAL W (1.83) multiple sequence alignment

```

beta globin  -----MVHLTPEEKSAVTALWGKVNVDD--EVGGEALGRLLVVYPWTQRRFFESFG- 47
myoglobin   -----MGLSDGEWQLVLNVWVKVEADIPGHGQEVLIIRLFKGGHPETLEKFDKFK- 48
neuroglobin -----MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR 47
soybean     -----MVAFTTEKQDALVSSSFQAFKANI PQYSVVFYTSILEKAPAAKDLFSFLA- 49
rice        MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKI FEVAPSASQMFSFLR- 59
              :   :   :   :   . . .   .   : :   *   *
              |   |   |   |   |   |   |   |   |   |   |   |   |
beta globin  DLSTPDAVMGNPKVKAHGKKVLGAFSDG LAHLDNLKGTFFATLS-----ELHCDKLHVDPE 102
myoglobin   HLKSEDEMKA SEDLKKHGATVLTALGGI LKKKGHHEAEIKPLA-----QSHATKHKIPVK 103
neuroglobin QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLA S---LGRKHRAVGVKLS 104
soybean     --NGVDPT--NPKLTGHA EKLFALVRDSAGQLKASGTVVADAA----LGSVHAQKAVTDP 101
rice        --NSDVPLEKNPKLKT HAMS VFVMTCEAA AQLRKAGKVTVRDTTLKRLGATHLKYGVGDA 117
              .   . . . *   . : :   :   :   :   :   :
beta globin  NFRL LGNVLV CVLAH HF--GKEFTPPVQAAYQKV VAGVANALAHKYH----- 147
myoglobin   YLEFI SECI IQVLQSKH--PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
neuroglobin SFSTV GESLLYMLEKCL--GPAFTPATRAAWSQLYGAVVQAMSRGWDGE---- 151
soybean     QFVVVKEALLKTIKAAV--GDKWSD ELSRAWEVAYDELAAA IKA----- 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  -----MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFESFG- 47
myoglobin   -----MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIIRLFKGHPEQVLEKFDKFK- 48
neuroglobin -----MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALPDLPLFQYNCR 47
soybean     -----MVAFTKQDALVSSSFSAFKANIPQYSVVFYTSILEKAPAAKDLFSFLA- 49
rice        MALVEDNNAVAVSFSEEQDALVLSWAILKKDSANIALRFFLKIIFEVAPSASQMFSLR- 59
           :   :   :   :   . . .   .   : :   *   * .

           :   :   :   :   . . .   .   : :   *   * .
           |   |   |   |   |   |   |   |   |   |   |   |   |   |
beta globin  DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLNPKGTFATLS-----ELHCDKLHVDPE 102
myoglobin   HLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLA-----QSHATKHKIPVK 103
neuroglobin QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLAS---LGRKHRAVGVKLS 104
soybean     --NGVDPT--NPKLTGHAEKLFALVRDSAGQLKASGTVVADAA----LGSVHAQKAVTDP 101
rice        --NSDVPLEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA 117
           .   . . . *   . : :   :   :   :   :   :   :   :

beta globin  NFRLLGNVLVCVLAHHF--GKEFTPPVQAAYQKVVAGVANALAHKYH----- 147
myoglobin   YLEFISECIIQVLQSKH--PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
neuroglobin SFSTVGESLLYMLEKCL--GPAFTPATRAAWSQLYGAVVQAMSRGWDGE---- 151
soybean     QFVVVKEALLKTIKAAV--GDKWSDELSRAWEVAYDELAAAIKKA----- 144
rice        HFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE--- 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).

THANK
YOU