# 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 parology

- 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

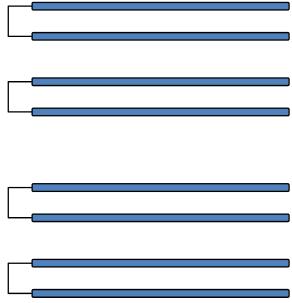
## ClustalW

ClustalX (		
	lignment Trees Colors Alignment Mode	Font Size: 18
1 2 3 4 5 6 7	HBB_HU HBB_HO HBA_HU HBA_HO GLB5_P MYG_PH LGB2_L	*: : * : * · · · VDEVGGEALG VQLSGEEKAAVLALWDKVNEEEVGGEALG VLSPADKTNVKAAWGKVGAHAGEYGAEALE VLSAADKTNVKAAWSKVGGHAGEYGAEALE PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILV VLSEGEWQLVLHVWAKVEADVAGHGQDILI GALTESQAALVKSSWEEENANIPKHTHRFFI
CLUSTAL	ruler • -Alignment file create	1

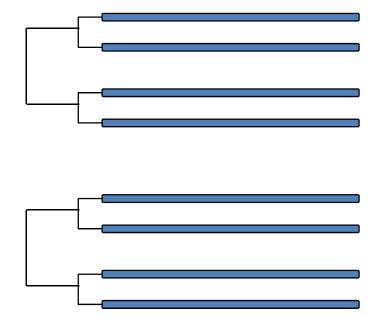
#### www.clustal.org

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

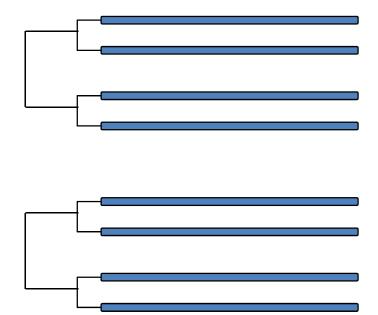
• Fix the alignment between pairs and treat as one sequence



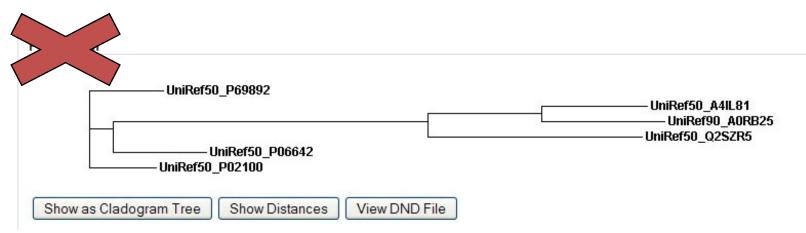
• Align your fixed pairs with each other



• Align your fixed pairs with each other



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



#### ClustalW at the EBI

MBL-EBI	EB Sea	All Datab	ases 🔽 Ent	er Text Here	
atabases	Tools	EBI Grou	ps Training	Industry About Us Help	
Fools Home	Tools Index ID Mapping	•	Sequence Analysis		
Tools A-Z ID Mapping Literature Microarray An Analysis		•	e Analysis		
		nalysis 🕨	alysis encompasses the use of various bioinformatic met or structure of genes and the proteins they code for.		
		tional 🕨 🕨			
	Proteomic Services		s <u>Transeq</u> can help determine the protein coding regions (		
Analysis	Sequence A	nalysis 🕨 🕨	Align	in order to elucidate their related	
Proteomic Se	Similarity & r	Homology	OTHOUR		
Sequence An	Structural Analysis		ClustalW2	uctura analysis tools we have :	
Align	Tools - Miscellaneous		CoC Plot/ CoGreport		
CENSOR	Web Services		Dna Block Aligner For	m Desc	
ClustalW2	Downloads		GeneWise	wise global and local alignment tool	
CpG Plot/ C	pGreport	CENSOR (	Kalign	een query sequences against a ref	
Dna Block A	ligner	ClustalW2	MAFFT	iple sequence alignments.	
Form GeneWise Kalign		CpG Plot/Cp	MUSCLE	Island finder and plotting tool (EMB	
		Dna Block A	Pepstats/ Pepwindow/	/ pares two DNA sequences assum	
		GeneWise	Pepinfo	hpares a protein sequence or a prot	
MAFFT		Kalign (i)	PromoterWise	ast and accurate multiple sequence	
MUSCLE Pepstats/ Pepwindow/ Pepinfo		MAFFT ()	SAPS	FFT (Multiple Alignment using Fast F	
			T-Coffee	iple sequence alignment program.	
PromoterW	Constant and Constant	MUSCLE	Transeq	tiple Sequence Comparison by Log	
SAPS				better average accuracy and better spe depending on the chosen options.	
T-Coffee		Pepstats/Pe	pwindow/Pepinfo (i)	EMBOSS programs for basic protein se	
Transaa					

#### **ClustalW**

CLUSTAL W (1.83) multiple sequence alignment

beta globin	MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFG- 47	1						
myoglobin	MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFK- 48	}						
neuroglobin	MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR 47	1						
soybean	MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFLA- 49	,						
rice	MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSFLR- 59	,						
	: : : : :: * *.							
	$\nabla$							
beta globin	DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPE 10	2						
myoglobin	HLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVK 10	3						
neuroglobin	QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLASLGRKHRAVGVKLS 10	)4						
soybean	NGVDPTNPKLTGHAEKLFALVRDSAGQLKASGTVVADAALGSVHAQKAVTDP 10	)1						
rice	NSDVPLEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA 11	.7						
	* .:: : : :							
beta globin	NFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHKYH 147							
myoglobin	YLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154							
neuroglobin	SESTVCESTLYMLERCL-CDAETDATRAAWSOLVCAWOOAMSBCWDCE 151							

neuroglobin SFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRGWDGE---- 151 soybean QFVVVKEALLKTIKAAV-GDKWSDELSRAWEVAYDELAAAIKKA----- 144 rice HFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE--- 166 : : :: :: : \* . . :

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

#### Homology

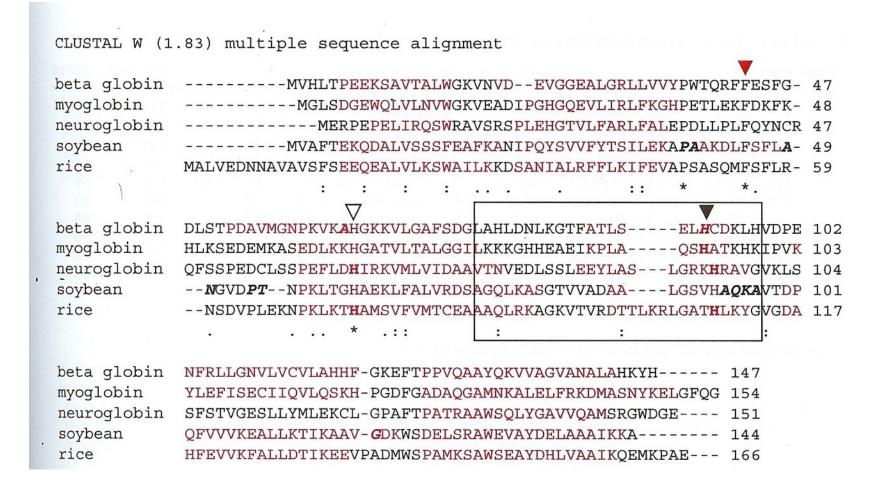
VS.

# 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



Pevsner, Jonathan. Bioinformatics and Functional Genomics. 2nd. Baltimore Maryland: Wiley-Blackwell, 2009. 179-207. Print.

# 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