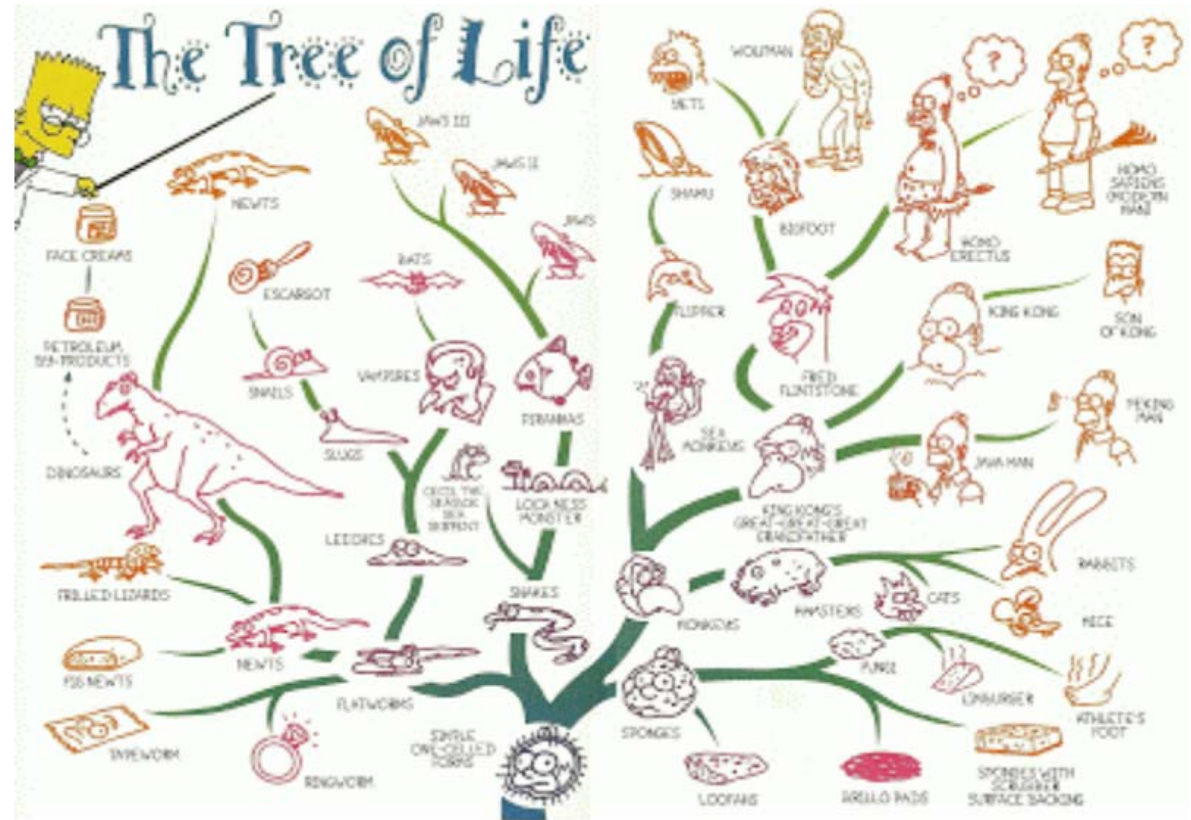


# CS123A

## Bioinformatics

### Module 3 – Week 8 – Presentation 2

Leonard Wesley  
Computer Science Dept  
San Jose State Univ



Bart Simpson's Tree of Life  
© Matt Groening

# Agenda

- BLOSUM vs PAM Matrices
- Multiple Sequence Alignment (MSA)
- Building Phylogenetic Trees

# BLOSUM vs PAM Matrices

# Why Align Protein Sequences?

- At times they can provide more & better information (e.g., homologous regions or domains) about organisms & species than just nucleotide sequences.
- ASSUMPTIONS:
  - Sequence similarity → similarity in function (and/or structure)
  - Almost always true for similarity > ~30%, 20% to 30% is the “twilight” zone.
- HOWEVER: Function is carried out at the level of a folded protein, i.e., 3D structure, BUT sequence conservation occurs at the level of 1D sequence.
- THEREFORE: Structural similarity –X-> Sequence similarity (or even homology)

# Convergent Evolution



Courtesy of [Matthew Field](#). License: CC-BY.



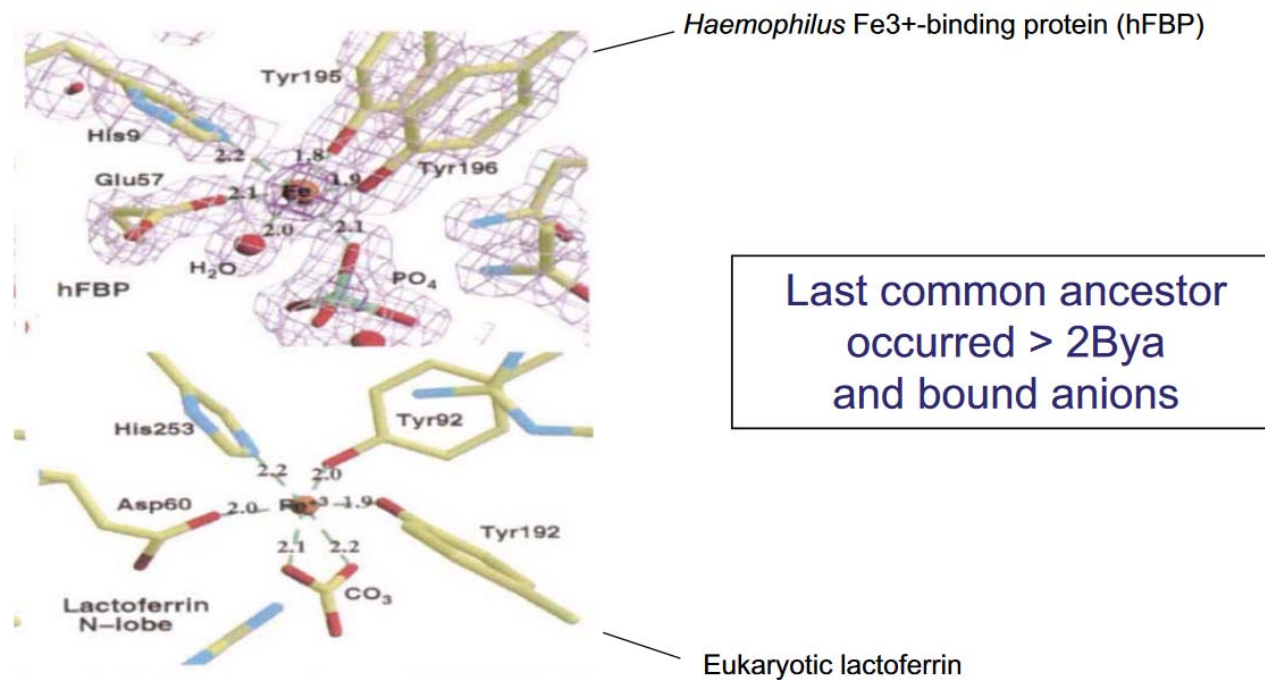
© Dave Green at [Butterfly Conservation](#). All rights reserved.  
This content is excluded from our Creative Commons license. For  
more information, see <http://ocw.mit.edu/help/faq-fair-use/>.

Leonard Wesley (c) 2019

Last common ancestor  
lived > 500 Mya and  
lacked wings (and  
probably legs and eyes)

Same idea for proteins  
- can result in similar  
structures with no  
significant similarity in  
sequence

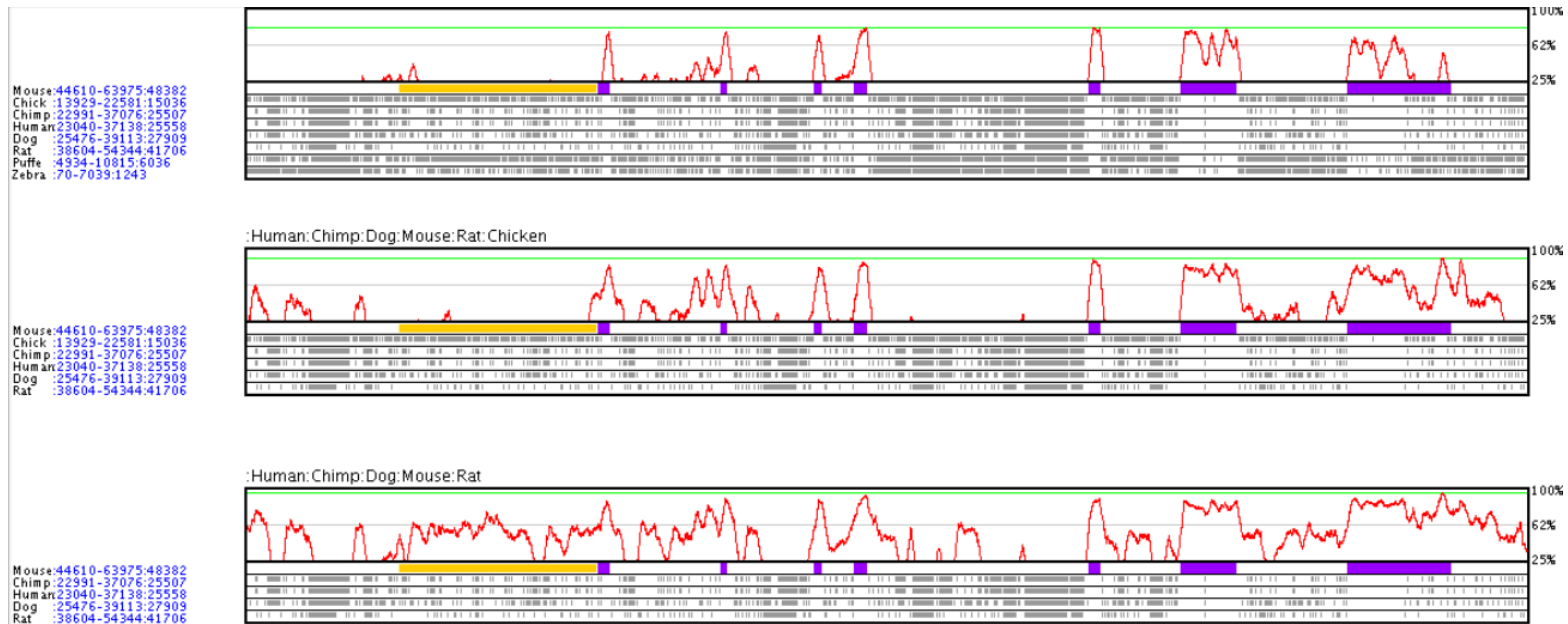
# Convergent Evolution Of $\text{Fe}^{3+}$ Binding Proteins



Courtesy of Nature Publishing Group. Used with permission.  
Source: Bruns, Christopher M., Andrew J. Nowalk, et al. "Structure of *Haemophilus influenzae* Fe<sup>3+</sup>-Binding Protein Reveals Convergent Evolution within a Superfamily." *Nature Structural & Molecular Biology* 4, no. 11 (1997): 919-24.

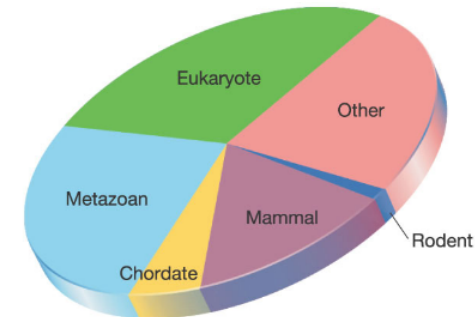
Bruns et al. *Nature Struct. Biol.* 1997

# Sequence Conservation Implies Function



## Alignment is the key to

- Finding important regions
- Determining function
- Uncovering evolutionary events



Leonard Weisley (c) 2013

# Substitution Matrices

- Two popular sets of matrices for protein sequences
  - –BLOSUM matrices [Henikoff & Henikoff, 1992]
  - –PAM (Point Accepted Mutation) matrices [Dayhoff et al., 1978]. They refer to the replacement of a single amino acid in a protein with a different amino acid. These mutations were identified by comparing highly similar sequences with at least 85% identity, and it is assumed that any observed substitutions were the result of a single mutation between the ancestral sequence and one of the present day sequences.
- BLOSUM & PAM matrices account for substitutions in very different ways, but they end up having comparable matrices.
- Both try to capture the relative substitutability of amino acid pairs in the context of evolution.

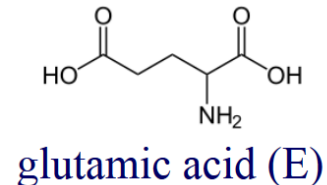
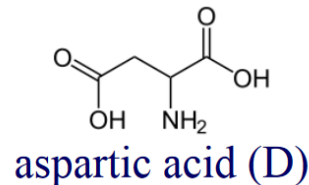


# More On Scoring Matches

- So far, we've discussed multiple gap penalty functions, but only one match-scoring scheme:

$$s(x_i, y_i) = \begin{array}{ll} +1 & \text{when } x_i = y_i \\ -1 & \text{when } x_i \neq y_i \end{array}$$

- For protein sequence alignment, some amino acids have similar structures and can be substituted in nature: aspartic acid (D) glutamic acid (E)



# BLOSUM62

[illegible]

# PAM Matrices

- PAM matrices define a time unit, where 1 PAM is the time in which 1/100 amino acids are expected to undergo a mutation.
- The PAM1 probability matrix shows the probability of the amino acid at column  $j$  being replaced by the amino acid at row  $i$ .
- It was calculated from Dayhoff's PAM counts, and rescaled to be 1 PAM unit of time.

	ORIGINAL AMINO ACID																			
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
A Ala	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R Arg	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N Asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D Asp	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C Cys	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q Gln	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E Glu	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G Gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H His	1	2	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I Ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L Leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K Lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M Met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F Phe	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P Pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S Ser	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T Thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W Trp	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y Tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

## PAM Matrices (cont.)

- To calculate the amino acid replacement probabilities for longer time durations, the matrix can be multiplied by itself the corresponding number of times.
- The PAM250 probability matrix, describing the replacement probabilities given 250 PAM units of time, was derived by raising the PAM1 probability matrix to the power 250 (all elements were scaled by 100 for legibility):

		ORIGINAL AMINO ACID																			
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
REPLACEMENT AMINO ACID	A Ala	13	6	9	9	5	8	9	12	5	8	6	7	7	4	11	11	11	2	4	9
	R Arg	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
	N Asn	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
	D Asp	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
	C Cys	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
	Q Gln	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
	E Glu	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
	G Gly	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
	H His	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
	I Ile	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
	L Leu	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
	K Lys	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
	M Met	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
	F Phe	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
	P Pro	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
	S Ser	9	5	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
	T Thr	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
	W Trp	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
	Y Tyr	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
	V Val	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

# PAM vs BLOSUM

- The replacement probabilities derived using this exponentiation correctly account for multiple substitutions. Not only are the off-diagonal probabilities proportionally larger as you would expect for a longer time duration, but they are flatter.
  - For example, the probability of a valine (V) to isoleucine (I) replacement is 33× larger than a V to histidine (H) replacement in the PAM1 matrix, but only 4.5× larger in the PAM250 matrix.
  - Score matrices can then be calculated from the probability matrices and observed base frequencies.
- The BLOSUM matrices, developed by Steven and Jorja Henikoff and published in [1992](#), take a very different approach. Whereas PAM is implicitly applying a stationary finite sites model of evolution using matrix exponentiation, the effect of multiple substitutions is dealt with **implicitly** in BLOSUM by constructing different score matrices for different time scales.

## PAM vs BLOSUM (cont.)

- Within multiple sequence alignments of homologous sequences, conserved contiguous blocks of amino acids are identified. Within each block, multiple sequences are clustered when their pairwise average sequence identity is higher than some threshold. The threshold is 80% for the BLOSUM80 matrix, 62% for BLOSUM62, 50% for BLOSUM50 and so on.
- This means that for BLOSUM80, blocks will have average pairwise identities no greater than 80%, for BLOSUM62 no greater than 62%, *et cetera*.
- Amino acid replacement probabilities for homologous sequences are calculated from pairwise comparisons between clusters. These probabilities will be the result of single and multiple substitutions, with multiple substitutions having greater influence at greater evolutionary distances. Therefore score matrices generated from pairwise comparisons between clusters of on average greater distance, like the BLOSUM50 matrix, will naturally account for the larger effect of multiple substitutions.

# Some Differences Between PAM & BLOSUM

PAM	BLOSUM
Based on global alignments of closely related proteins.	Based on local alignments.
PAM1 is the matrix calculated from comparisons of sequences with no more than 15% divergence but corresponds to 99% sequence identity.	BLOSUM 62 is a matrix calculated from comparisons of sequences with a pairwise identity of no more than 62%.
Other PAM matrices are extrapolated from PAM1.	Based on observed alignments; they are not extrapolated from comparisons of closely related proteins.
Higher numbers in matrices naming scheme denote larger evolutionary distance.	Larger numbers in matrices naming scheme denote higher sequence similarity and therefore smaller evolutionary distance. <sup>[19]</sup>

# Comparable PAM & BLOSUM Matrices

PAM	BLOSUM
PAM250	BLOSUM45
PAM160	BLOSUM62
PAM120	BLOSUM80

Below diagonal: BLOSUM 62  
Above diagonal: BLOSUM 62 - PAM 160

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	0	-1	1	0	2	1	1	2	1	2	0	0	2	4	1	5	1	2	-2	5	C
S		2	0	-2	0	-1	0	0	0	1	0	0	0	1	0	-1	-1	1	1	-1	S
T	9		2	-1	-1	-1	0	0	0	0	0	0	-1	0	-1	1	0	1	1	3	T
P	-1	4		2	-2	-1	-1	0	0	-1	-1	-1	1	1	0	-1	0	0	2	1	P
A	-1	1	5		2	-1	-2	-2	-1	0	0	1	1	0	0	1	0	1	1	2	A
G	-3	-1	-1	7		2	0	-1	-2	0	1	1	0	0	-1	0	-1	1	2	4	G
N	0	1	0	-1	4		3	-1	-1	0	0	1	-1	0	-1	0	-1	0	0	0	N
D	-3	0	-2	-2	0	6		2	-1	-1	-1	0	-1	0	0	0	0	2	1	3	D
E	-3	1	0	-2	-2	0	6		1	0	0	2	2	1	-1	0	0	2	2	4	E
Q	-3	0	-1	-1	-2	-1	1	6		0	-2	0	1	1	-1	0	0	1	3	3	Q
H	-4	0	-1	-1	-1	-2	0	2	5		2	-1	0	1	0	-1	0	1	2	-2	H
R	-3	0	-1	-1	-1	-2	0	0	2	5		-1	-1	0	-1	1	0	1	3	-4	R
K	-3	-1	-2	-2	-2	-2	1	-1	0	0	8		1	-2	-1	1	1	2	3	1	K
M	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5		-2	-1	-1	0	1	2	4	M
I	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5		-1	1	0	0	1	3	I
L	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5		-1	0	-1	1	2	L
V	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4		0	1	2	4	V
F	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4		-1	-2	1	F
Y	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4		-1	2	Y
W	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		-1	W
	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		
	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	



An abstract graphic featuring a large, solid green oval in the center. The oval is surrounded by several thin, curved lines in light gray and black, creating a sense of motion or a stylized orbit. The lines are more densely packed on the left side and more sparse on the right.

# Multiple Sequence Alignment

# Multiple Sequence Alignment (MSA)

## Simultaneously Compares 3 Or More Sequences

- Why MSA?
  - Need to identify regions of homology as well as orthologs.
  - Infer structural and functional properties of protein molecules.
  - Identify important residues. *Residues are the individual organic compounds called amino acids that comprise some of the building blocks of complete proteins.*
- MSA can be applied to DNA, RNA and Proteins

# Advantages of MSA

- Multiple alignment helps improve accuracy of alignment between sequence pairs.
- Can reveal areas/patterns of conserved residues not readily found in pair wise alignment.

# Example MSA From TCoffee

Tcoffee URL: <https://www.ebi.ac.uk/Tools/msa/tcoffee/>



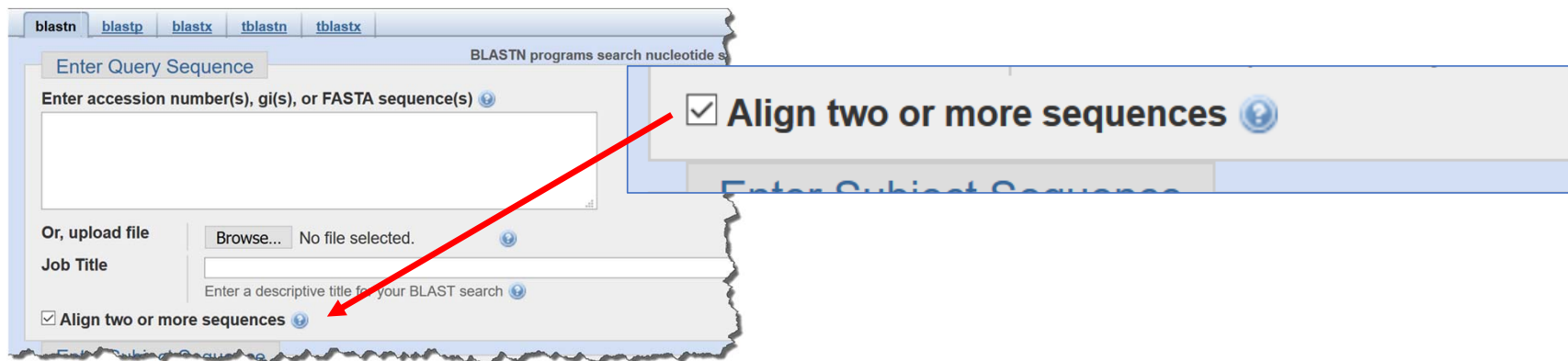
**REDISH** = good alignment

**BLUE** = exact/very good alignment

# There Are Many MSA Tools

- NCBI – BLAST:

[https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE\\_TYPE=BlastSearch&BLAST\\_SPEC=blast2seq&LINK\\_LOC=align2seq](https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&BLAST_SPEC=blast2seq&LINK_LOC=align2seq) Select the “Align two or more sequences” option

A screenshot of the NCBI BLAST search interface. The interface has a light blue header with tabs for 'blastn', 'blastp', 'blastx', 'tblastn', and 'tblastx'. Below the header, there's a section titled 'Enter Query Sequence' with a text input field for 'Enter accession number(s), gi(s), or FASTA sequence(s)'. To the right of this section, there's a checkbox labeled 'Align two or more sequences' which is checked. Below the query input, there's a section for 'Or, upload file' with a 'Browse...' button and 'No file selected.' text. Below that, there's a 'Job Title' input field with the placeholder text 'Enter a descriptive title for your BLAST search'. At the bottom left of the main form area, there's another checkbox labeled 'Align two or more sequences' which is also checked. A red arrow points from this checkbox to the one on the right. The interface is partially obscured by a torn paper effect on the right side.

- ExPASy: [http://www.expasy.org/genomics/sequence\\_alignment](http://www.expasy.org/genomics/sequence_alignment)
- STRAP: <http://www.bioinformatics.org/strap/>
- NCBI: COBALT: [http://www.st-vn.ncbi.nlm.nih.gov/tools/cobalt/re\\_cobalt.cgi?](http://www.st-vn.ncbi.nlm.nih.gov/tools/cobalt/re_cobalt.cgi?)
- ... *many, many others* ....

# Clustal: A well Known MSA Algorithm

- ClustalW: Thompson et al., 1994 – gives good alignments for sequences significantly similar and roughly the same length.
- ClustalW superseded by Clustal X and then Clustal Omega
  - Clustal -> Clustal IV -> ClustalW -> Clustal X -> Clustal Omega
- ClustalW uses a hierarchical MSA method.

# Hierarchical MSA Is A Multiple Step Process.

- Given 3 or more sequences to align
- Sometimes random unrelated sequences are given to a MSA algorithm. Must determine significance by performing a randomization test.
- Two sequences are pair-wise aligned and the score (S) recorded.
- Then amino acids/nucleic acids in the sequences are shuffled so order is changed but length kept the same.

## Hierarchical MSA Is A Multiple Step Process. *(cont. #1)*

- Shuffled sequences are compared again and scores ( $S$ ) recorded again. This is repeated  $\sim 100$  times.
- The mean  $\bar{S}$  and the standard deviation  $\sigma$  for the scores is calculated.
- A Z score =  $(S - \bar{S}) / \sigma$  provides an indication of the significance of the two sequences.



## Hierarchical MSA Is A Multiple Step Process. (*cont. #2*)

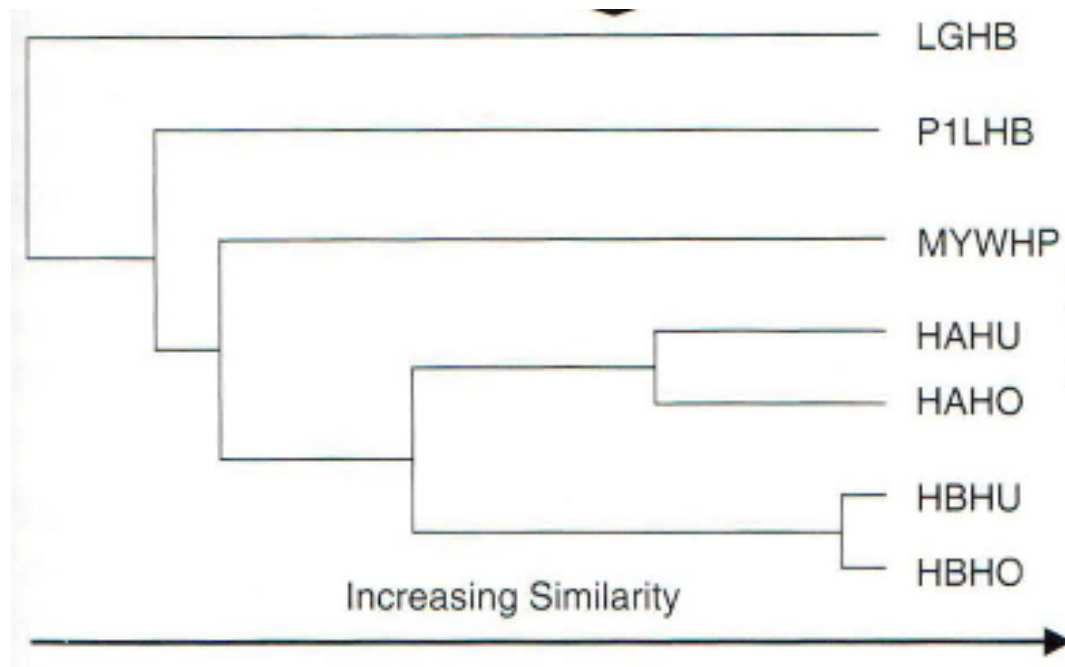
- A Z score  $> 6$  means high likelihood the two sequences can be aligned and aligned correctly in a way that can give insight into function, structure, ...and so forth.
- However, some alignments with Z score  $< 6$  can be correct. If and when this happens, one needs to consider the possibility that sequence similarity might have diverged faster than structural or functional similarity.

# Example Z Score Matrix

	HAHU	HBHU	HAHO	HBHO	MYWHP	P1LHB	LGHB
HAHU							
HBHU	21.1						
HAHO	32.9	19.7					
HBHO	20.7	39.0	20.4				
MYWHP	11.0	9.8	10.3	9.7			
P1LHB	9.3	8.6	9.6	8.4	7.0		
LGHB	7.1	7.3	7.5	7.4	7.3	4.3	

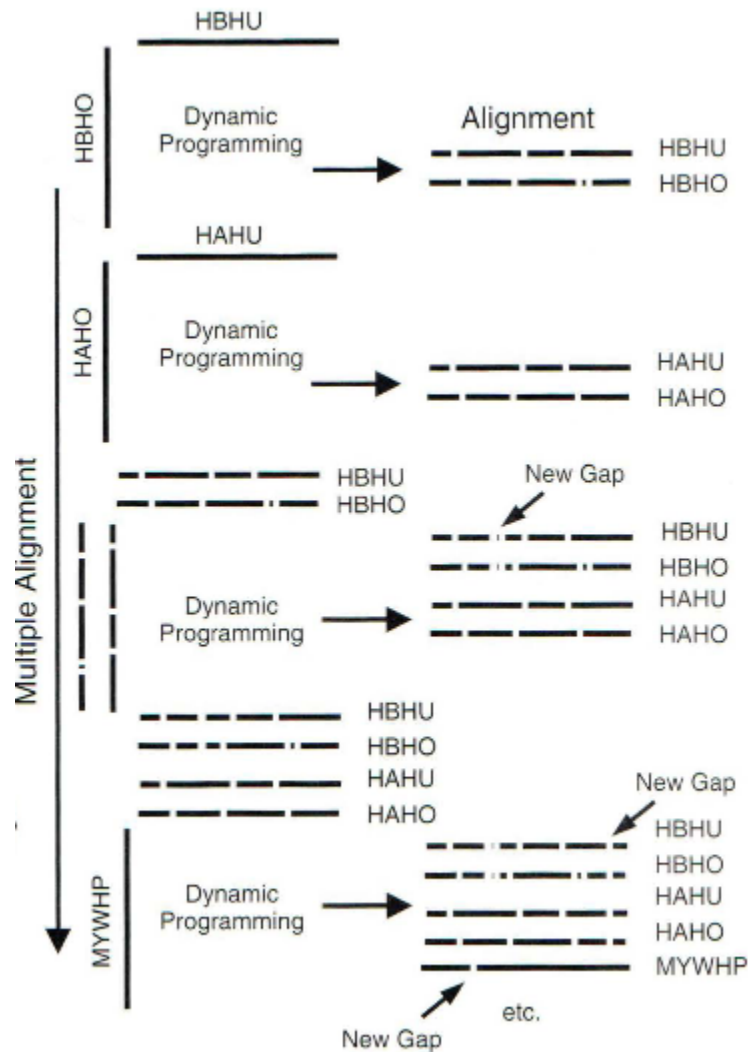
**Pairwise Z-scores for comparison of each sequence pair.  
Higher numbers mean greater similarity**

# Cluster Analysis



**Hierarchical cluster analysis of the Z-score table generates the dendrogram. Items joined toward the right of the tree are more similar than those linked at the left. Thus, LGHB is the sequence that is least similar to the other sequences in the set, whereas HBHU and HBHO are the most similar pair.**

# Building The Multiple Alignment



> The first two steps are pairwise alignments.

> The third step is a comparison of profiles from the two alignments generated in steps 1 and 2.

> The fourth step adds a single sequence (MYWHP) to the alignment generated at step 3.

> Further sequences are added in a similar manner.

# Other MSA Algorithms

- Hierarchical not guaranteed to find optimal alignment
- TCoffee: builds a library of pairwise alignments → inputs this to a hierarchical method
- PSI-BLAST: searches DB with a single sequence, high scores are retrieved and built into a MSA
- SCANPS: Similar to PSI-BLAST uses Smith-Waterman
- STAMP: Aligns multiple protein structures vs sequences.

## Example TCoffee MSA

- Go To <http://www.ebi.ac.uk/Tools/msa/tcoffee>
- Select “Use a example sequence” Then click “More options...” Then select BLOSUM
- Click Submit and then wait for the results.
- Then Select “Show Colors”. Look for good (Red) and Excellent (Blue) alignment regions. Then Select “Phylogenetic Tree”. Identify closely and distant organisms.

# MSA Lecture Exercise

- You came back from a trip to a jungle swamp after obtaining what you believe are DNA and/or protein samples of possibly known and/or unknown organisms. You want to know (1) If you have found evidence of existing or new organisms. If existing organisms, which one(s)?; (2) What part or structure of the organism's genome, if any, are we looking at?; and (3) What are related organisms ?
- The sequencing lab has provided you with a file that contains a protein sequence from the liquid sample that you gave them. The sequenced protein is contained in the file name "CS123A\_Example\_seq.txt" that is located in Canvas -> Files -> Module 3 Phylogenetic Trees -> Week 8 -> Slides folder.
- BLASTP the sequence to find possible best matches. In the "Organism" section type in "prokaryote" in the first window and select the (taxid:2) entry. Click on the "+" then enter and select the Rattus (taxid:10114) entry. Click "+" one last time and enter "Fish stool-associated RNA virus (taxid:2219050)". Click the BLAST button. Note the names of the top 4 "DIFFERENT" organisms. What are these organisms?
- Create and name .txt file. Get the FASTA sequence for the first 4 "DIFFERENT" matches you selected. You can get the FASTA sequence after clicking on each accession number and going to that web page. Then look for a link to the FASTA file. Click that link, then on the drop down tab in the upper left next to the word FASTA, select the "FASTA txt" option. Copy and paste the FASTA info into to the .txt file that you created and named at the start of this step.
- Copy each of the 4 FASTA sequences into your .txt file. Then do a MSA on the sequences. Use the dendrogram to determine which sequences are most closely related. Upload your answer to "which sequences are most closely related" to Canvas Lecture Exercise 2.

# Summary

- Sequence alignment is useful to identify novel and existing organisms from genomic sequences. MSA is helpful to identify homologous and conserved regions.
- BLAST & BLAST2: Performs local pairwise and multiple alignments for nucleotides, proteins, and from nucleotides to proteins and from proteins back to nucleotide. Score (S) and Expect (E) values used to help assess quality of match.
- Smith-Waterman: Uses dynamic programming to provide optimal local sequence pairwise alignment. Can be used by multiple sequence alignment (MSA) algorithms, SCANPS.
- Needleman-Wunsch: Uses dynamic programming to provide optimal global sequence pairwise alignment. Gaps can be inserted to optimal sequence scores and to make each sequence the same length. Can be used by MSA algorithms.



## Summary *(cont.)*

- Several good MSA tools: TCoffee: builds a library of pairwise alignments → inputs this to a hierarchical method.
- PSI-BLAST: searches DB with a single sequence, high scores are retrieved and built into a MSA.
- SCANPS: Similar to PSI-BLAST uses Smith-Waterman.
- STAMP: Aligns multiple protein structures vs sequences.

# Building Phylogenetic Trees From MSAs

- One Way:
  - Use the alignment score between each pair of sequences.
  - Use the similarity score between each pair of sequences.

	Seq1	Seq2	Seq3	Seq4
Seq1	-	3	1	4
Seq2	3	-		
Seq3	1	7	-	3
Seq4	4	0.5	3	-

- Use hierarchical clustering techniques to build dendrogram. See in-lecture example.

# Start/Continue Reading Assigned Reading In Chap 7