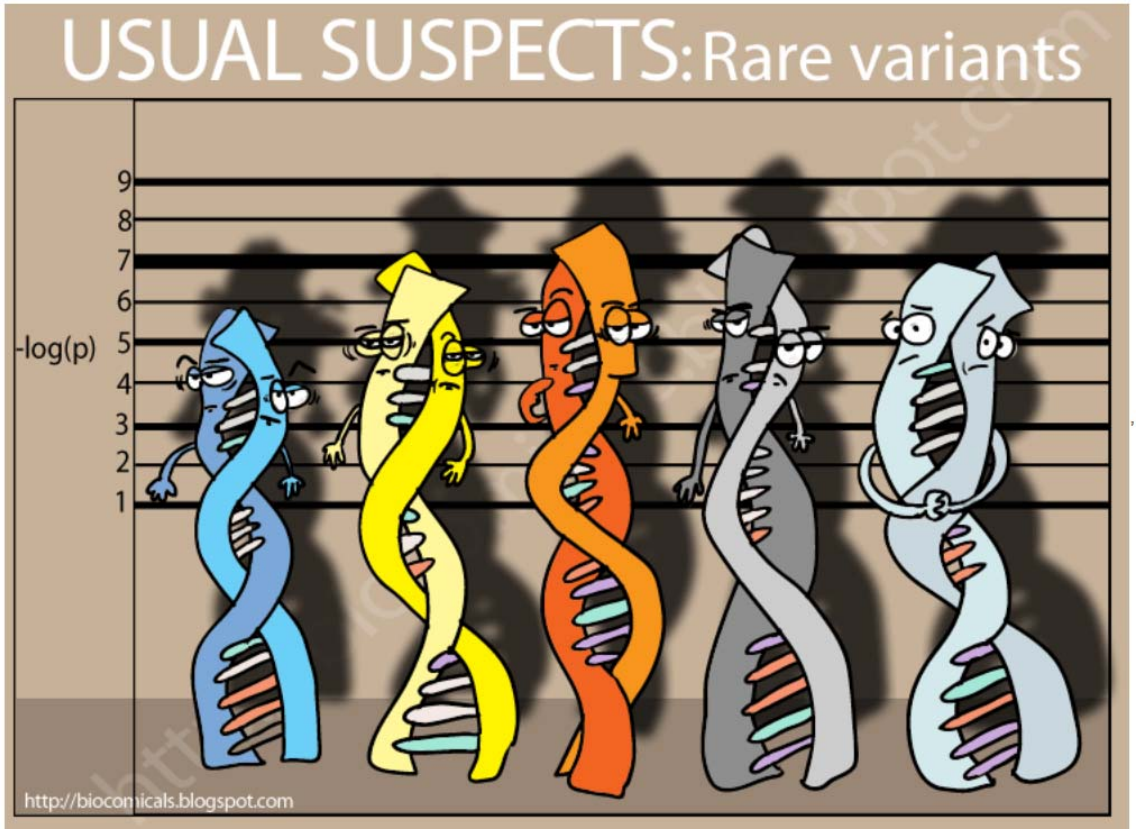
CS123A Bioinformatics Module 2 – Week 6 – Presentation 1

Leonard Wesley

Computer Science Dept

San Jose State Univ



Agenda

- Needleman-Wunsch optimal global alignment
- Multiple Alignment
- Dendrogram Representation Of Related Sequences

Similarity/Match Score Not The Same as % Similarity

- % Similarity is the # of exact matches in two sequences divided by the length of the sequences.

  M A I H W A A sim.% = 5/7x100% =71.4%

  M T I H M A A

- Similarity/match score is

  5+0+4-1+4+4 = 16

- Similarity/match score captures notions of conservation, evolution,…etc. % similarity does not.

Needleman-Wunsch: An Optimal Global Alignment Algorithm

- Consider

  I S A L I G N E D

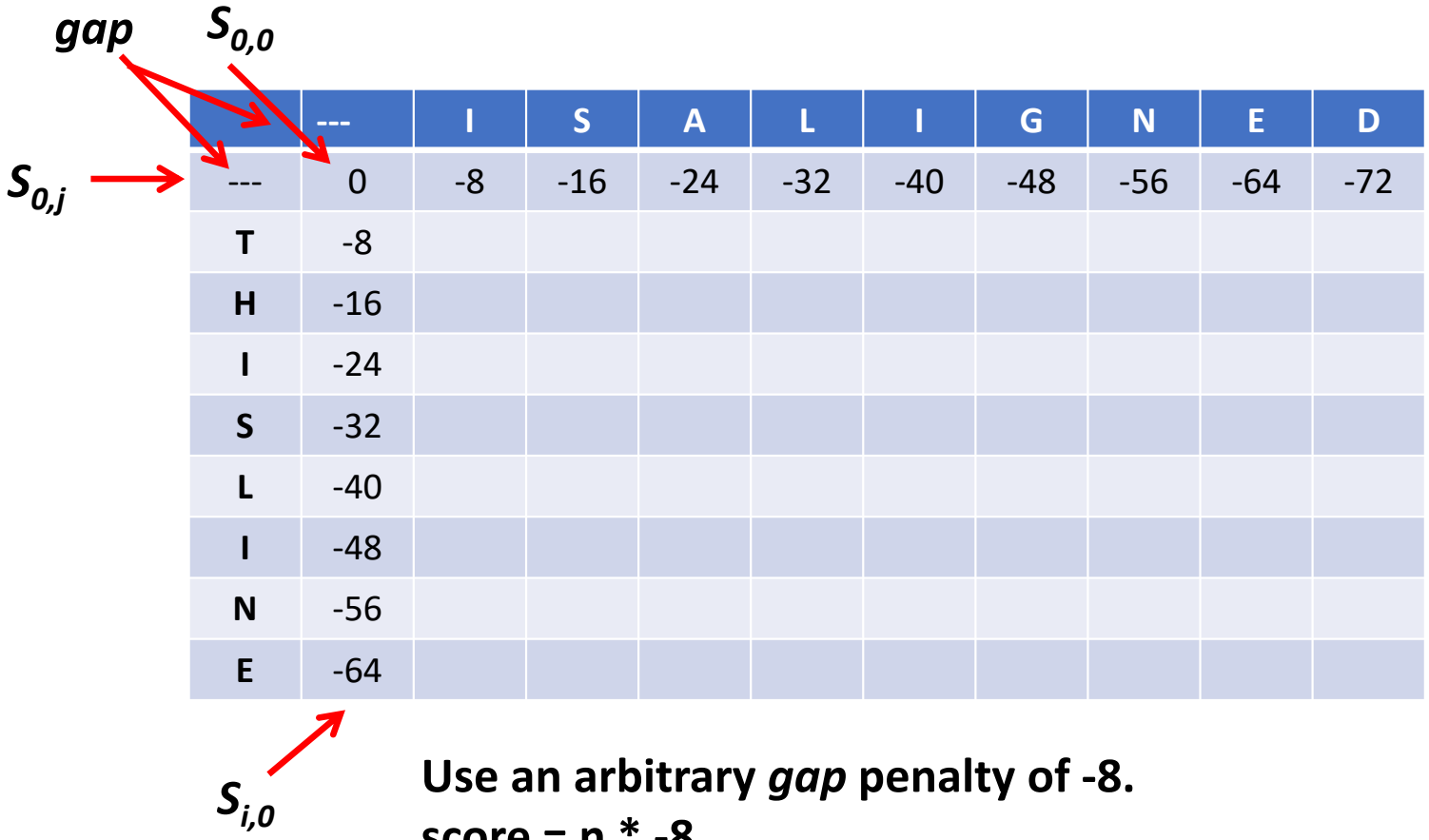
  T H I S L I N E * score = -1-2-1-2+2-4+6+5-4 = -1

- BUT we can see visually that a better alignment is

  - - **I S** A **L I** G **N E** D

  T H **I S** - **L I** - **N E** - What is its score?

Build Dynamic Programming Matrix



	---	I	S	A	L	I	G	N	E	D
---	0	-8	-16	-24	-32	-40	-48	-56	-64	-72
T	-8									
H	-16									
I	-24									
S	-32									
L	-40									
I	-48									
N	-56									
E	-64									

Use an arbitrary *gap* penalty of -8.
score = n * -8

Fill In $S_{i,j}$ Scores

Max [(0 + -1 = **-1**), (-8 + -8 = -16), (-8 +-8 = -16)] = **-1**

Use an arbitrary *gap* penalty of -8. score = n*8

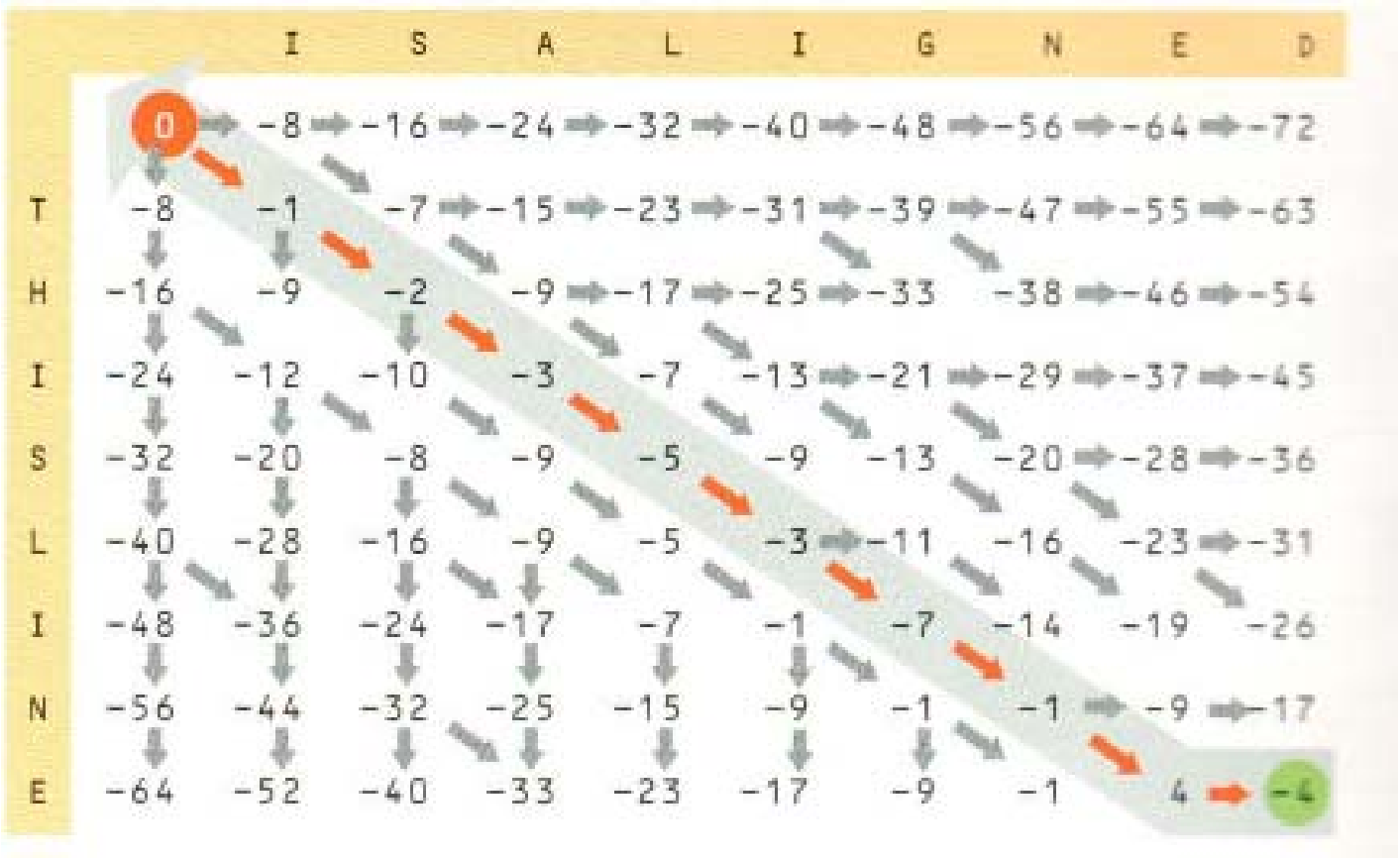
	---	I	S	A	L	I	G	N	E	D
---	0	-8	-16	-24	-32	-40	-48	-56	-64	-72
T	-8	**-1**								
H	-16									
I	-24									
S	-32									
L	-40									
I	-48									
N	-56									
E	-64									

$$S_{i,j} = \max \begin{cases} S_{i-1,j-1} + s(x_i, y_j) \\ S_{i-1,\,j} + g \\ S_{i,\,j-1} + g \end{cases}$$

BLOSUM68 Matrix: https://www.ncbi.nlm.nih.gov/Class/FieldGuide/BLOSUM62.txt

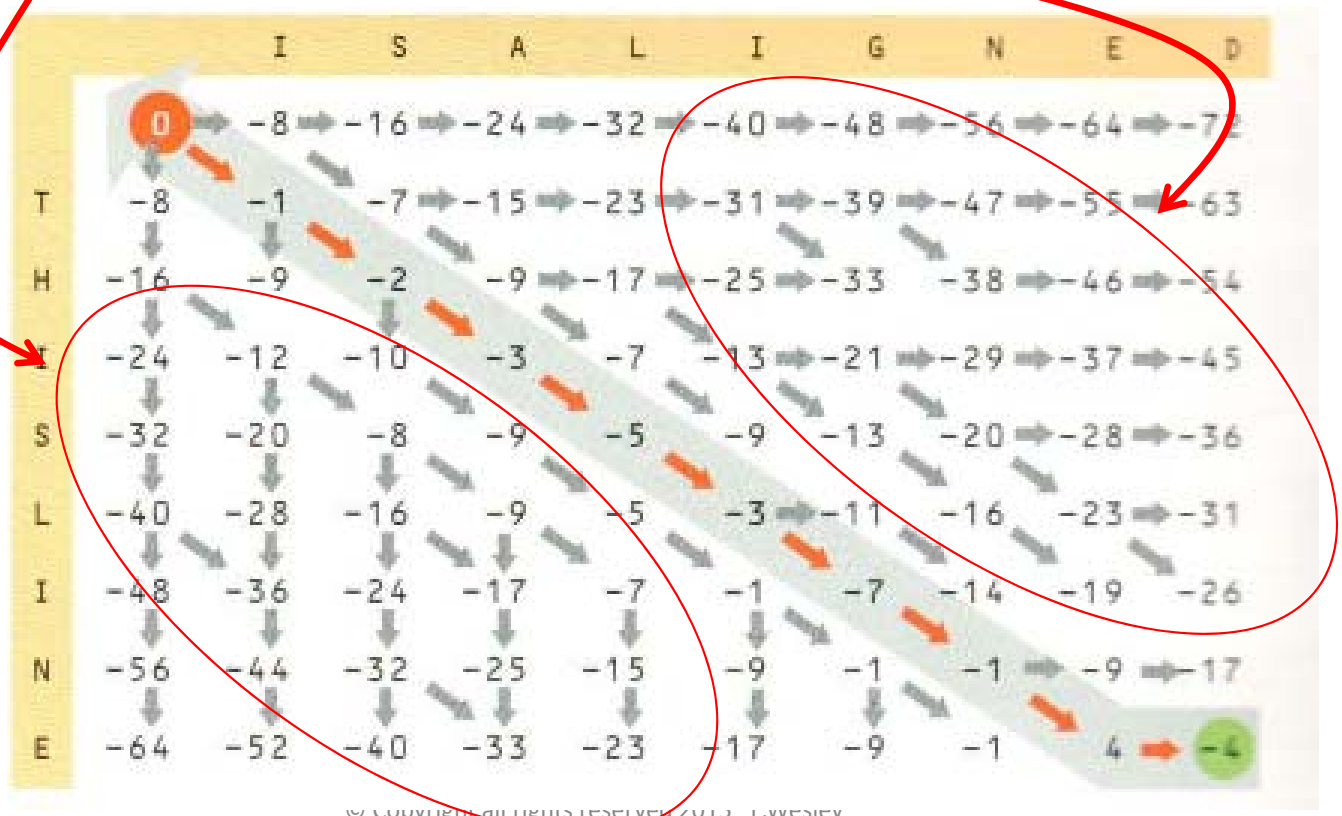
Score For Example Alignment

I S A L I G N E D T H I S L I N E -



Some Things To Note …

If gap penalty is too high, will tend to get larger negative values the further away you go from the diagonal.



Insert Gaps To Explore Achieving A Better Score

- -**I S** A L I G N E D
T H**I S** L I N E

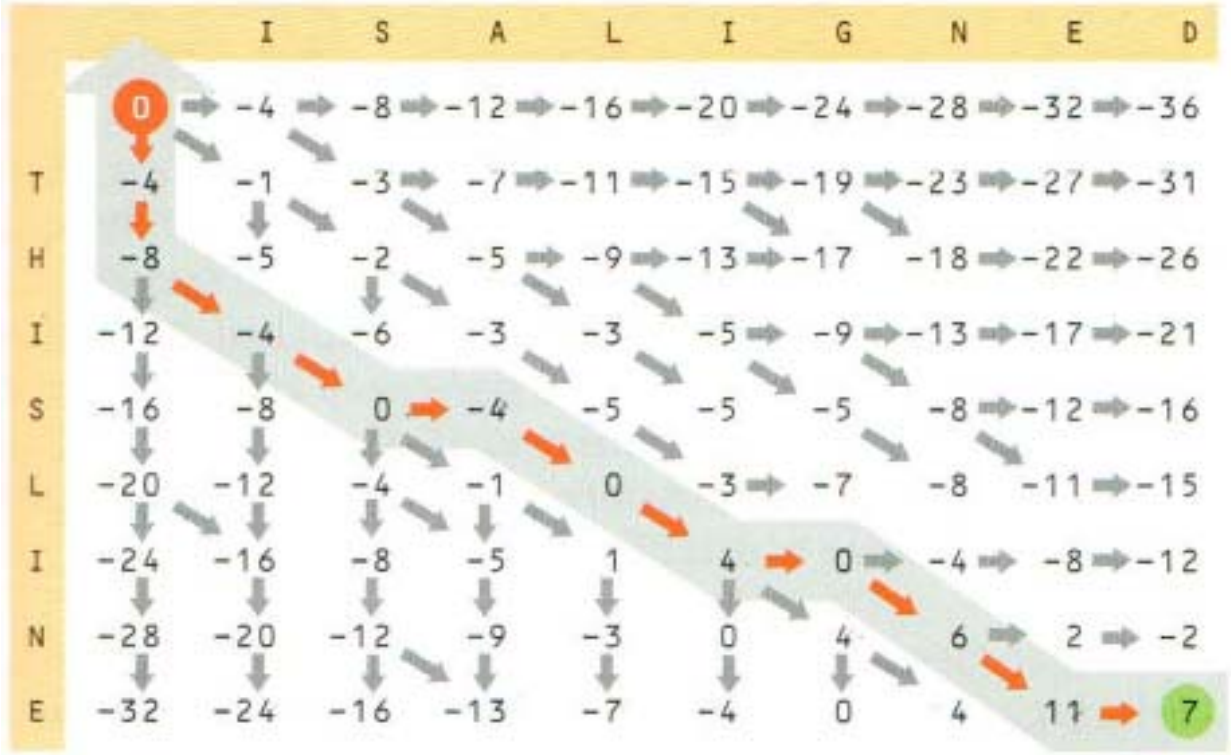
- -**I S** A **L I** G N E D
T H**I S - L I** N E

- -**I S** A **L I** G **N E** D
T H**I S - L I** - N E

- -**I S** A **L I** G **N E** D
T H**I S** - **L I** - **N E** - What is its score?

Score For Better Alignment

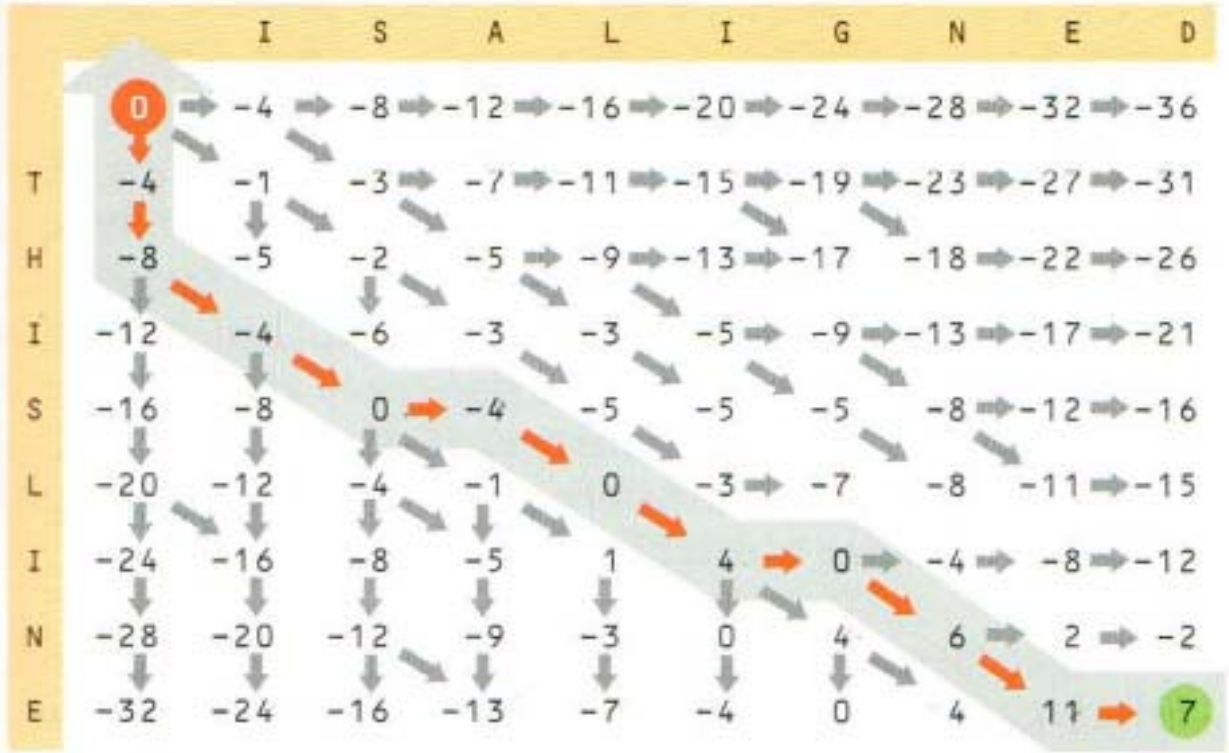
- - I S A L I G N E D
T H I S - L I - N E -



What Is The Best Alignment?
What Is The Best Alignment Score?

- - I S A L I G N E D
T H I S - L I - N E -

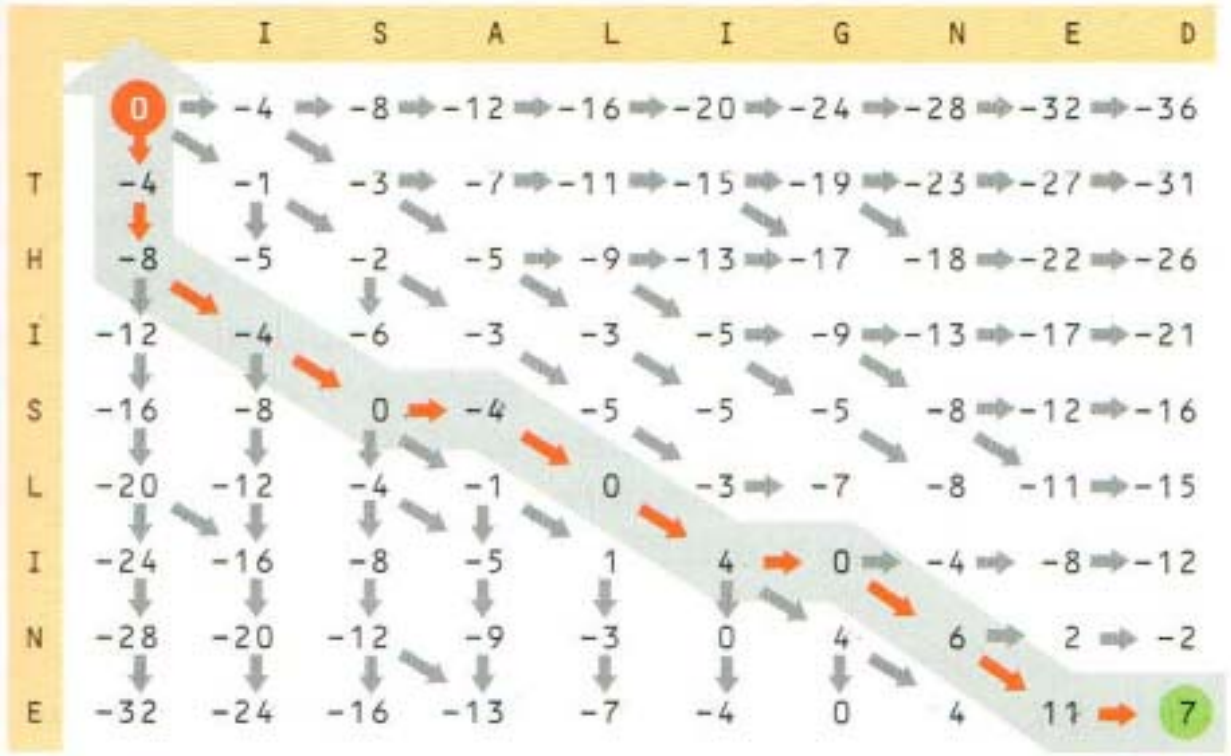
← ????



What Is The Best Alignment?
What Is The Best Alignment Score?

- - I S A L I G N E D
T H I S - L I - N E -

← Score = 7 + 11 + 6 + 0 + 4 + 0 + -4 + 0 + -4 + -8 + -4 + 0 = 8



Example of How Needleman-Wunsch Works

- http://experiments.mostafa.io/public/needleman-wunsch/

- Do not use for the lecture exercises or homework. Use just to get an idea of how the algorithm works.

Lecture Exercise

- Calculate the best Needleman-Wunsch score and alignment for

  REF SEQ: H E A G A W G H E E
  QUERY: P A W H E A E

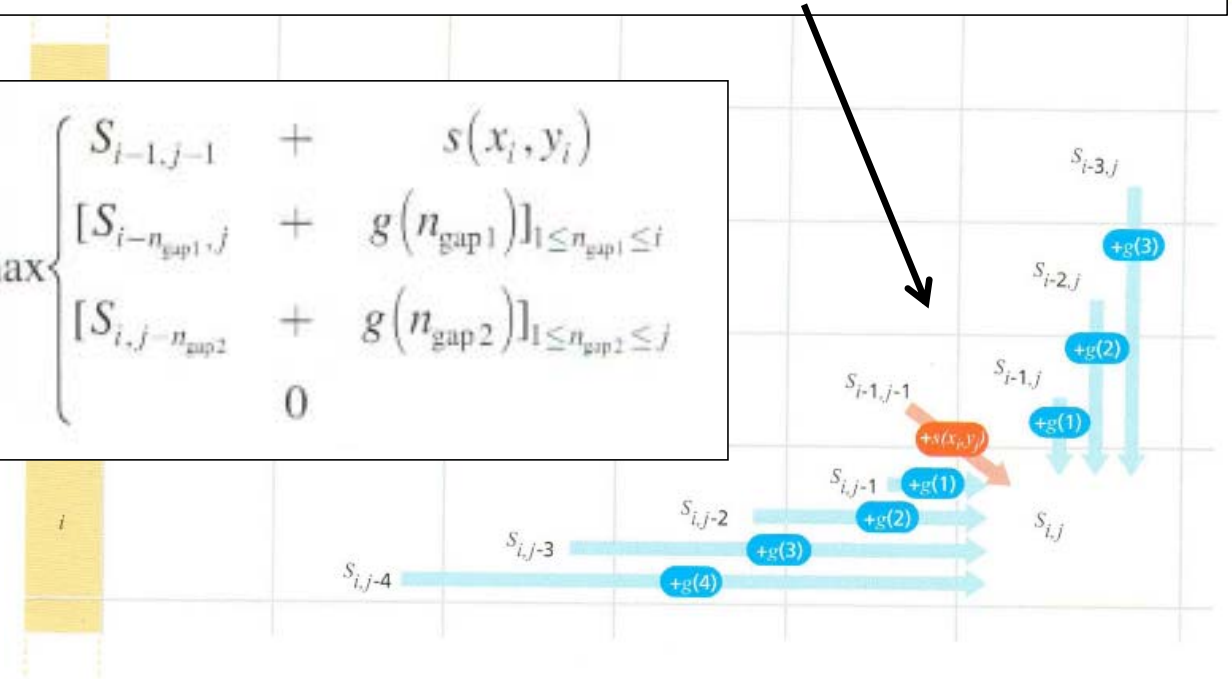
Use the dynamic programming template named "Dynamic_Programming_Score_Template.docx" in the Canvas Files -> Module 2 Alignment -> Week 6 -> Slides folder to build your score matrix. Upload your answer to Canvas -> Assignments -> Lecture Exercise 1 under the Lecture Exercises category.

Use the BLOSUM62 scoring matrix in the file named "Blosum62_Matrix.pdf" in the Canvas -> Files -> Module 2 Alignment -> Week 6 -> Slides folder.

So What Is An Acceptable Gap Penalty?

- Lower the initial gap penalty, but increase the penalty for consecutive gaps.

$$S_{i,j} = \max \begin{cases} S_{i-1,j-1} + s(x_i, y_i) \\ [S_{i-n_{gap1},j} + g(n_{gap1})]_{1 \le n_{gap1} \le i} \\ [S_{i,j-n_{gap2}} + g(n_{gap2})]_{1 \le n_{gap2} \le j} \\ 0 \end{cases}$$

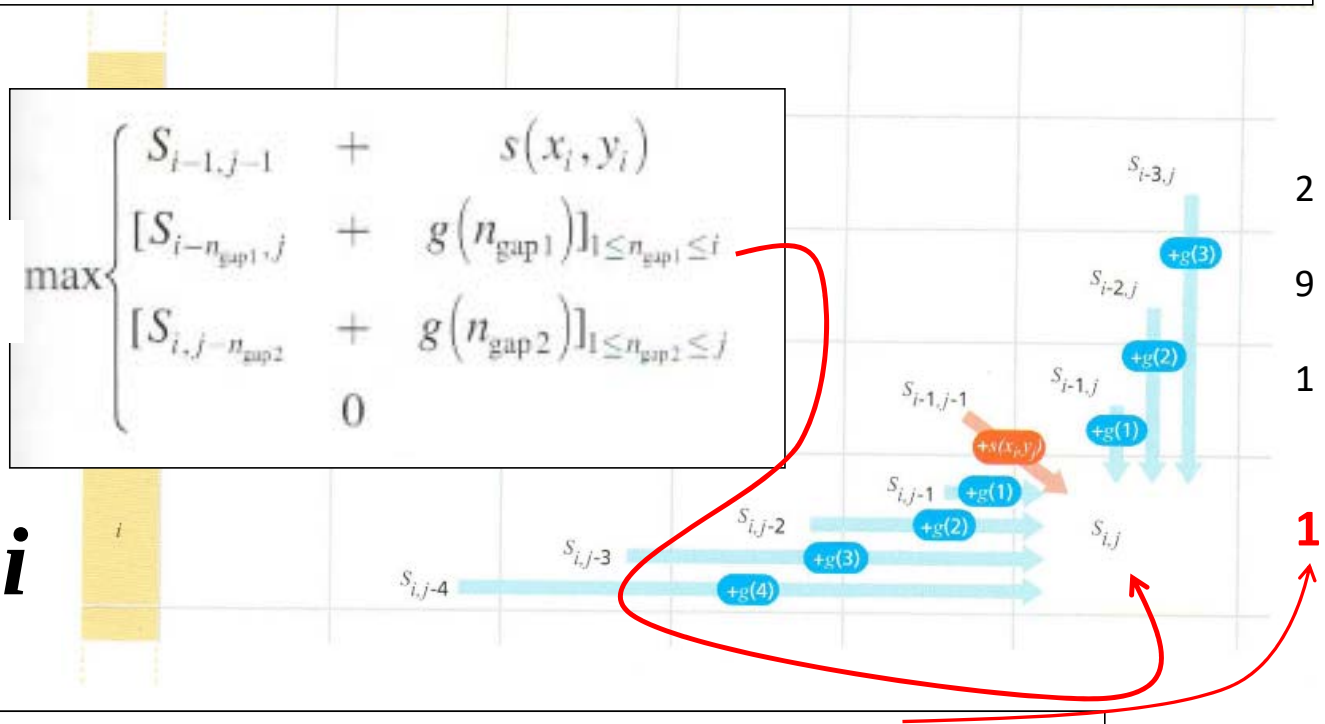


Example Calculation

Use gap penalty = -4, lets say we want to

compute $S_{i,j}$ *for* $i = 4, j = 6$

$$S_{i,j} = \max \begin{cases} S_{i-1,j-1} + s(x_i, y_i) \\ [S_{i-n_{gap1},j} + g(n_{gap1})]_{1 \le n_{gap1} \le i} \\ [S_{i,j-n_{gap2}} + g(n_{gap2})]_{1 \le n_{gap2} \le j} \\ 0 \end{cases}$$



MAX *i* [(2+(3*-4)=-10), (9+(2*-4)= 1), (1 + (1*-4)=-3)] = 1 Must do for *j*

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 and RNA

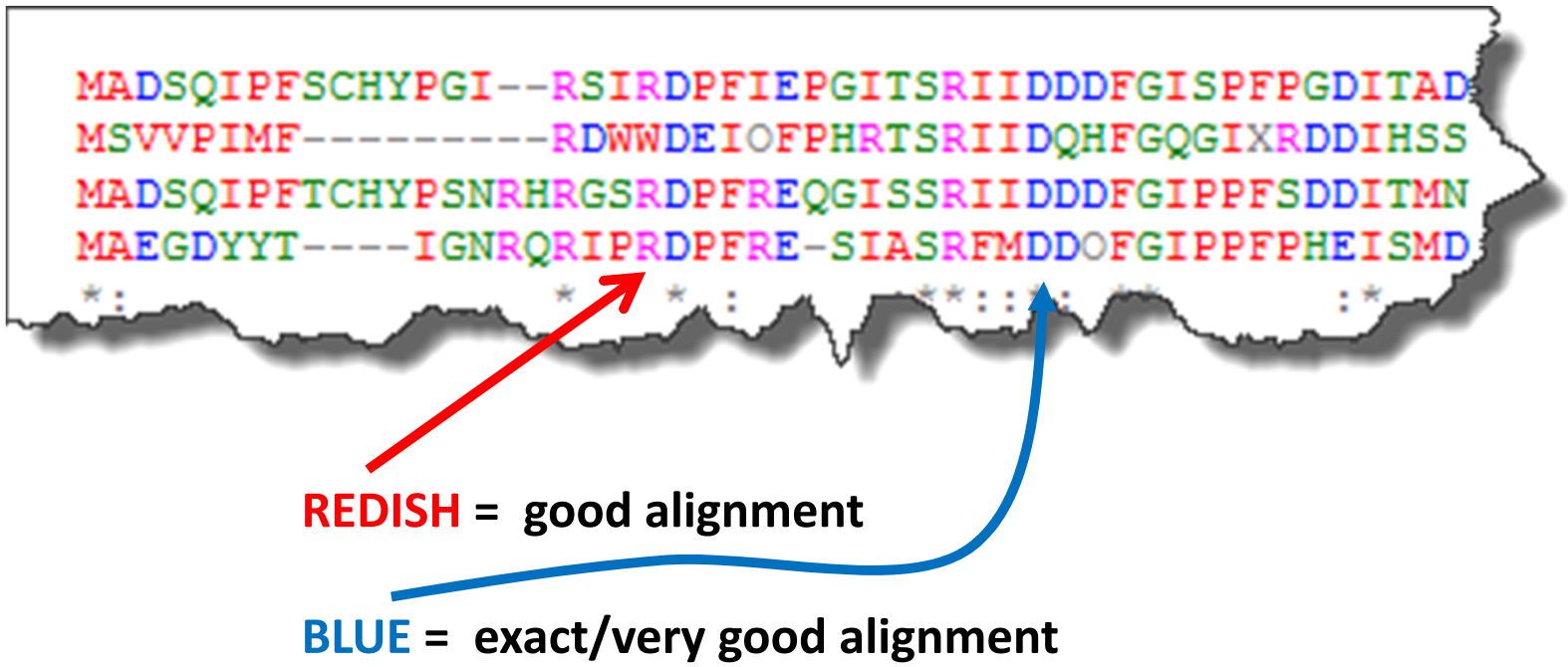
Advantages of MSA

- Multiple alignment helps improve accuracy of alignment between sequence pairs.

- Can reveal areas/patterns of conserved residues not readily found in pair wise alignment.

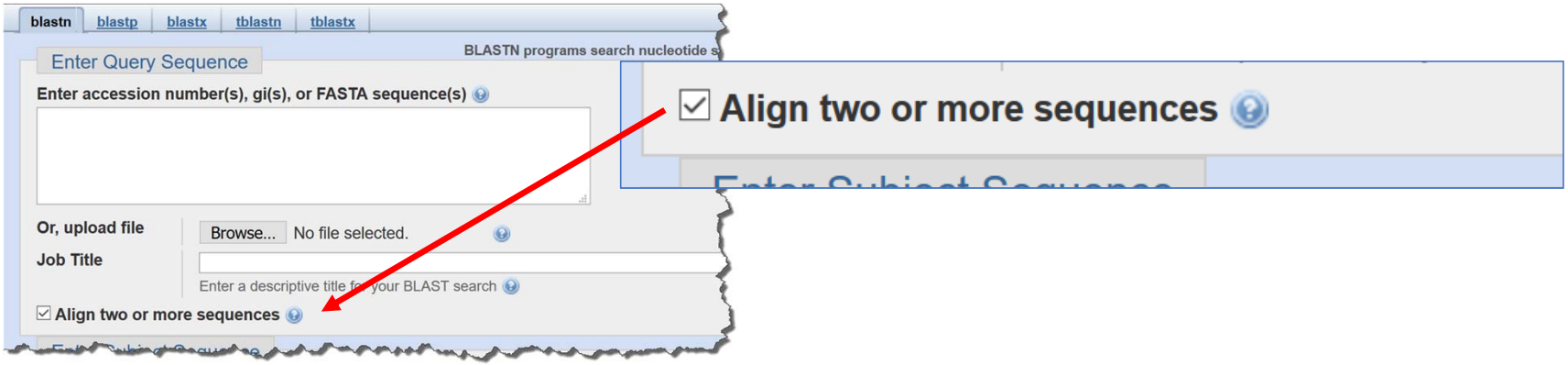
Example MSA From TCoffee

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



There Are Many MSA Tools

- NCBI – BLAST: 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



- ExPASy: http://www.expasy.org/genomics/sequence_alignment
- STRAP: http://www.bioinformatics.org/strap/
- NCBI: COBALT: http://www.st-va.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 ~100 times.
- The mean $\bar{S}$ and the standard deviation σ 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.

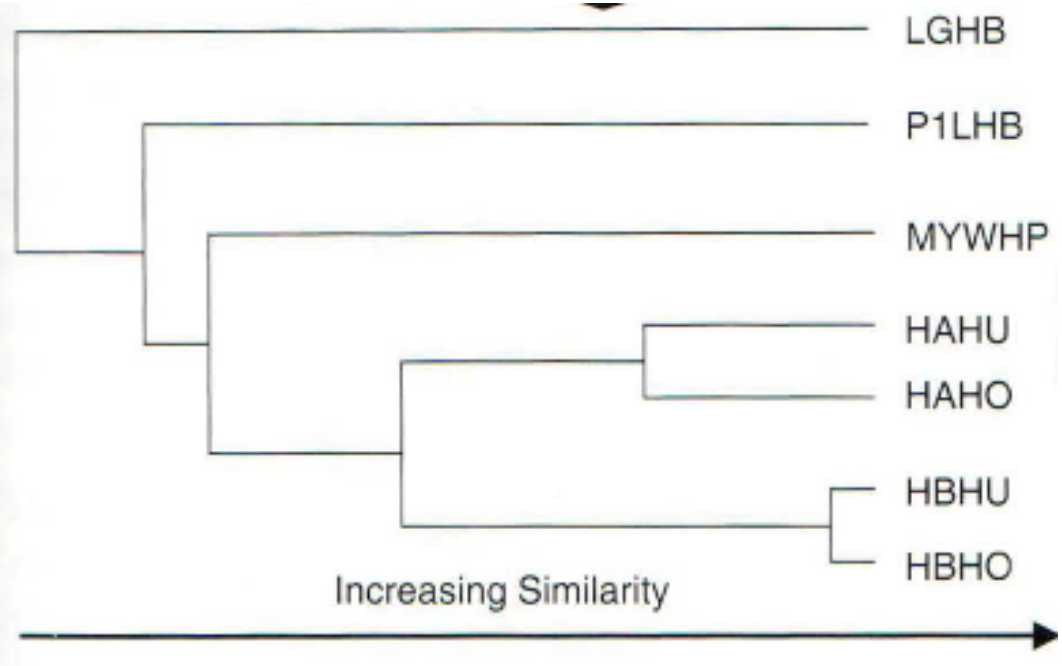
- However, some alignments with Z score < 6 can be correct. If and when this happens, one needs to consider the possibility that sequence similarity might have diverged faster than structural or functional similarity.

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	

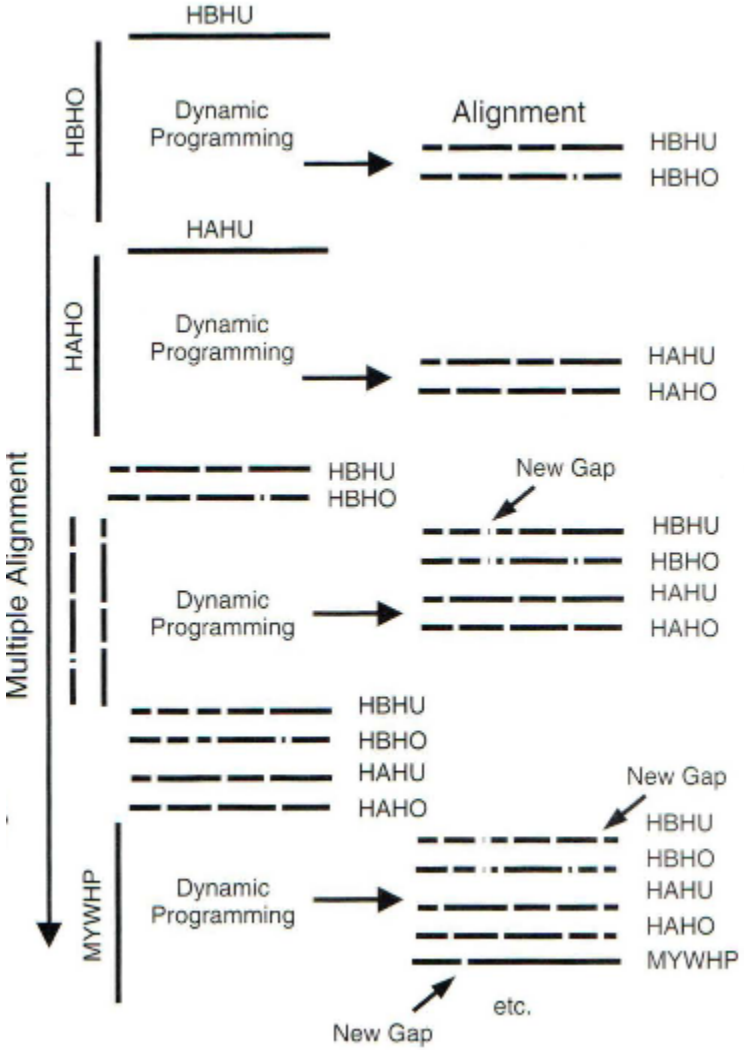
Pairwise Z-scores for comparison of each sequence pair.
Higher numbers mean greater similarity

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 I 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 2 Alignment -> Week 6 -> 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 form 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. Cane 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.