Bioinformatics II:

Multiple Sequence Alignment MSA using Dynamic Programming

Dr Manaf A Guma

University Of Anbar- college of applied sciences-Heet.

Department of chemistry

1

What is the MSA?

- It is an alignment of more than 2 sequences.
- Why do we do MSA? Or what is the purposes of MSA?

1. *To highlight conservation and variation. How? By identifying the regions of similarity among different species.*
2. *To find the relation among different species.*
3. *To find the profile of sequence from the database.*
4. *To know how to draw phylogenetic trees.*

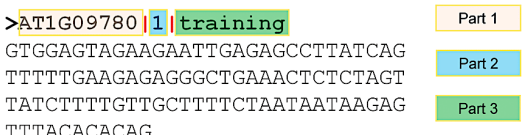
2

Why we do use dynamic programming in MSA?

- Because there is a huge database which make the comparison very difficult if we run MSA by hand.
- Which software and websites are commonly used to do MSA?

1. BLAST.
2. FASTA.
3. ClustalW.

FASTA format) do you remember it !



3

How do you use BLAST to run MSA? (Tutorial)

1. We have to have a specific sequence for (protein or DNA for a specific species) that we need to find the similarity with it.
2. If we do not have it, then we go to https://www.uniprot.org and then find the Protein seq.
3. Copy the seq (in a FASTA format) do you remember it !
4. Open https://blast.ncbi.nlm.nih.gov/Blast.cgi and find blast protein-protein.
5. Paste the seq in the box labeled with Enter Query Sequence:
6. Click on BLAST to find the similarities.
7. The result will show the comparison (the identity and the scoring of the similarity) of the protein to various proteins in the database.
8. It also show you the mattress used to generate the comparison.

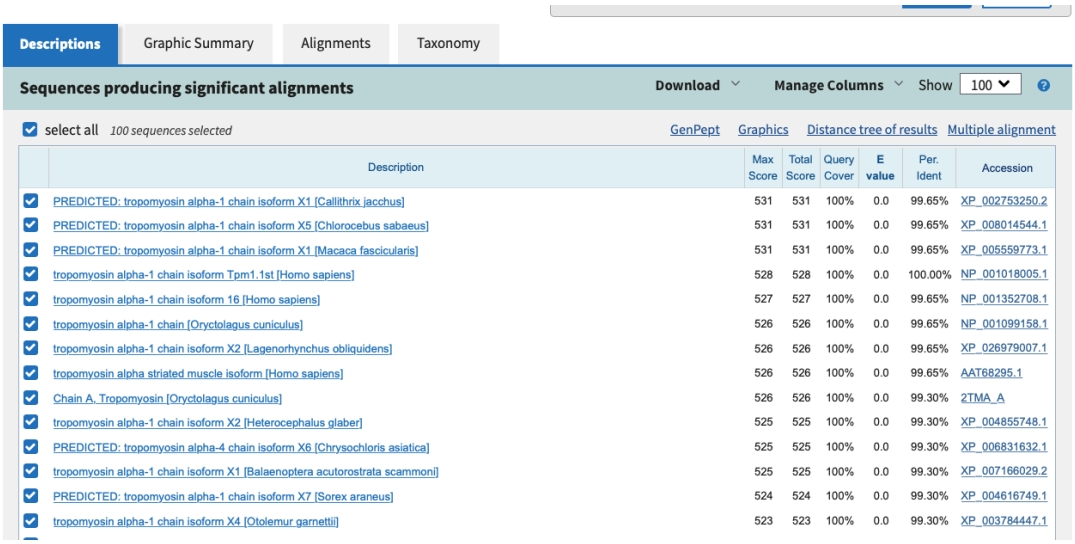
4

Can we get MSA form BLAST? What can we get?

- We can get only pairwise alignment using BLAST. (what is pairwise- do you remember?)
- But we can not get all of the sequences aligned together in the same screen using BLAST.
- We can get the profile of each sequence (the type of the species, the gene name and gene number etc.)

5

An example to see how BLAST works



Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments — Download — Manage Columns — Show 100

select all *100 sequences selected* — GenPept — Graphics — Distance tree of results — Multiple alignment

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
PREDICTED: tropomyosin alpha-1 chain isoform X1 [Callithrix jacchus]	531	531	100%	0.0	99.65%	XP_002753250.2
PREDICTED: tropomyosin alpha-1 chain isoform X5 [Chlorocebus sabaeus]	531	531	100%	0.0	99.65%	XP_008014544.1
PREDICTED: tropomyosin alpha-1 chain isoform X1 [Macaca fascicularis]	531	531	100%	0.0	99.65%	XP_005559773.1
tropomyosin alpha-1 chain isoform Tpm1.1st [Homo sapiens]	528	528	100%	0.0	100.00%	NP_001018005.1
tropomyosin alpha-1 chain isoform 16 [Homo sapiens]	527	527	100%	0.0	99.65%	NP_001352708.1
tropomyosin alpha-1 chain [Oryctolagus cuniculus]	526	526	100%	0.0	99.65%	NP_001099158.1
tropomyosin alpha-1 chain isoform X2 [Lagenorhynchus obliquidens]	526	526	100%	0.0	99.65%	XP_026979007.1
tropomyosin alpha striated muscle isoform [Homo sapiens]	526	526	100%	0.0	99.65%	AAT68295.1
Chain A, Tropomyosin [Oryctolagus cuniculus]	526	526	100%	0.0	99.30%	2TMA_A
tropomyosin alpha-1 chain isoform X2 [Heterocephalus glaber]	525	525	100%	0.0	99.30%	XP_004855748.1
PREDICTED: tropomyosin alpha-4 chain isoform X6 [Chrysochloris asiatica]	525	525	100%	0.0	99.30%	XP_006831632.1
tropomyosin alpha-1 chain isoform X1 [Balaenoptera acutorostrata scammoni]	525	525	100%	0.0	99.30%	XP_007166029.2
PREDICTED: tropomyosin alpha-1 chain isoform X7 [Sorex araneus]	524	524	100%	0.0	99.30%	XP_004616749.1
tropomyosin alpha-1 chain isoform X4 [Otolemur garnettii]	523	523	100%	0.0	99.30%	XP_003784447.1

6

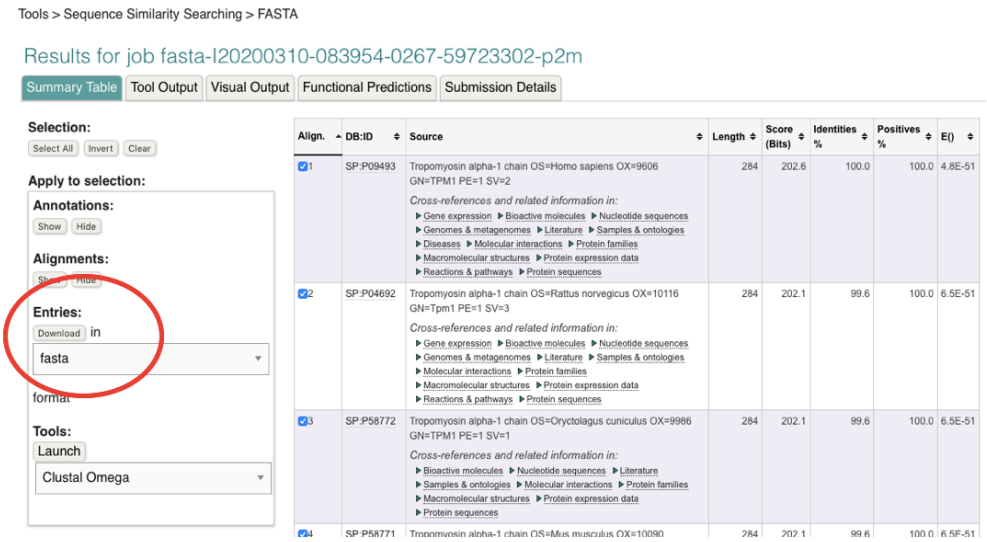
How do you use FASTA to run MSA?

1. Get the protein/DNA seq from https://www.uniprot.org.
2. copy the seq in FSATA format.
3. Open FASTA web page https://www.ebi.ac.uk/Tools/sss/fasta/.
4. Paste the seq.,
5. The results will show different choses to get various bioinformatic analysis in a table.
6. You can show the MSA by clicking on visual output.
7. You can also download the seq by clicking on Download

7

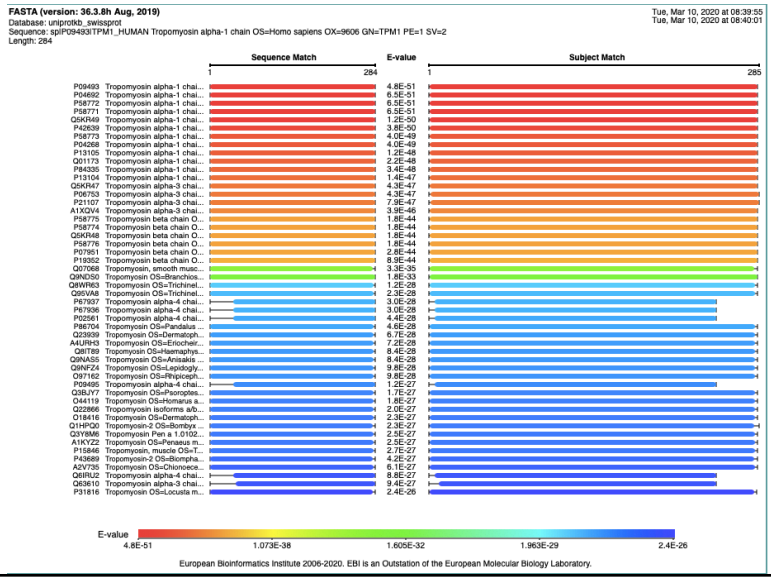
The tables of FASTA results: an example

You can download all the seq form here



8

An example to see how FASTA works



9

What is ClustalW ?

- **ClustalW is the "classic" MSA tool using C++ programming made by JD Thompson, DG Higgins, and TJ Gibson.**
- The original publication describing ClustalW is one of the 100 most cited publications in 'web of science'.
- How **CLUSTAL W deals with MSA?**

CLUSTAL W: deals with multiple sequence alignment through:

1. **Sequence weighting.**
2. **position-specific gap penalties**
3. **weight matrix choice.**

- **What is the lasts version of** ClustalW?
- ClustalW It is an old version, the version is Clustal Omega which is much faster and better tools are available. http://www.ebi.ac.uk/Tools/msa/

10

How do you use ClustalW to run MSA? (very common)

1. Get the protein/DNA seq from https://www.uniprot.org.
2. copy the seq in FSATA to download multiple seq.
3. Open FASTA web page https://www.ebi.ac.uk/Tools/sss/fasta/.
4. Paste the multiple seq in the box.
5. Run the FASTA omega. You can color it.
6. You see also the phylogenetic tree as well.

11

An example of ClustalW Omega

Results for job clustalo-I20200310-104708-0114-18168141-p2m

Alignments | Result Summary | Guide Tree | Phylogenetic Tree | Results Viewers | Submission Details

Download Alignment File | Hide Colors

```
CLUSTAL O(1.2.4) multiple sequence alignment


UNIPROT:TPM2_BIOGL      ------MDAIKKKMLAMKMEKENAIDRAEQMEQKVRDVEETKNKLEEEFNNLQKKFSNLQ    54
UNIPROT:TPM1_CAEEL      ------MDAIKKKMQAMKIEKDNALDRADAAEEKVRQITEKLERVEEELRDTQKKMTQTG    54
UNIPROT:TPM_ANISI       ------MDAIKKKMQAMKIEKDNALDRADAAEEKVRQMTDKLERIEEELRDTQKKMMQTE    54
UNIPROT:TPMM_TRICO      ------MDAIKKKMQAMKIEKDNALDRADAAEEKVRQITEKLERVEEELRDTQKKMMQTE    54
UNIPROT:TPM_TRIPS       ------MDAIKKKMQAMKIEKDNAMDRADAAEEKARQQQERVEKLEEELRDTQKKMMQVE    54
UNIPROT:TPM_TRISP       ------MDAIKKKMQAMKIEKDNAMDRADAAEEKARQQQERVEKLEEELRDTQKKMMQVE    54
UNIPROT:TPM2_BOMMO      ------MDAIKKKMQAMKLEKDNALDRAAMCEQQAKDANLRAEKAEEEARQLQKKIQTIE    54
UNIPROT:TPM_LOCMI       ------MDAIKKKMQAMKLEKDNALDRALLCEQQARDANLRAEKAEEEARALQKKIQTIE    54
UNIPROT:TPM_PANBO       ------MDAIKKKMQAMKLEKDNAMDRADTLEQQNKEANNRAEKSEEEVFGLQKKLQQLE    54
UNIPROT:TPM_PENMO       ------MDAIKKKMQAMKLEKDNAMDRADTLEQQNKEANNRAEKSEEEVHNLQKKMQQLE    54
UNIPROT:TPM_PENAT       ------MDAIKKKMQAMKLEKDNAMDRADTLEQQNKEANNRAEKSEEEVHNLQKKMQQLE    54
UNIPROT:TPM_CHIOP       ------MDAIKKKMQAMKLEKDNAMDKADTLEQQNKEANLRAEKTEEEIRANQKKSQLVE    54
UNIPROT:TPM_ERISI       ------MDAIKKKMQAMKLEKDNAMDRADTLEQQNKEANNRAEKTEEEIRATQKKMQQVE    54
UNIPROT:TPM_HOMAM       ------MDAIKKKMQAMKLEKDNAMDRADTLEQQNKEANIRAEKTEEEIRITHKKMQQVE    54
UNIPROT:TPM_LEPDS       ------MEAIKNKMQAMKLEKDNAIDRAEIAEQKSRDANLRAEKSEEEVRGLQKKIQQIE    54
UNIPROT:TPM_DERPT       ------MEAIKNKMQAMKLEKDNAIDRAEIAEQKARDANLRAEKSEEEVRALQKKIQQIE    54
UNIPROT:TPM_DERFA       ------MEAIKKKMQAMKLEKDNAIDRAEIAEQKARDANLRAEKSEEEVRALQKKIQQIE    54
UNIPROT:TPM_PSOOV       ------MEAIKKKMQAMKLEKDNAIDRAEIAEQKARDANLRAEKSEEEVRGLQKKIQQIE    54
UNIPROT:TPM_HAELO       ------MDAIKKKMQAMKLEKDNAVDRAETAEQQSRDAALRAEKAEEEVRSLQKKIQQIE    54
UNIPROT:TPM_RHIMP       ------MEAIKKKMQAMKLEKDNAVDRAETAEQQSREAALRAEKAEEEVRSLQKKIQQIE    54
UNIPROT:TPM3_RAT        MAGSTTIEAVKRKIQVLQQQA---------------------------------------    21
UNIPROT:TPM4_RAT        MAGLNSLEAVKRKIQALQQQA---------------------------------------    21
UNIPROT:TPM4_PIG        MAGLNSLEAVKRKIQALQQQA---------------------------------------    21
UNIPROT:TPM4_HUMAN      MAGLNSLEAVKRKIQALQQQA---------------------------------------    21
UNIPROT:TPM4_HORSE      MAGLNSLEAVKRKIQALQQQA---------------------------------------    21
UNIPROT:TPM4_MOUSE      MAGLNSLEAVKRKIQALQQQA---------------------------------------    21
UNIPROT:TPM1_CIOIN      ------MEAIKKKMTMLKLDKENAIDRAEQAETDKKSAEDKATGLEEELQGLQKRLKATE    54
```

12

The old version presentation of the ClustalW

13

What other programmes used for MSA?

Because Often multiple sequence alignments require manual editing:

1. ***Jalview*** is a powerful MSA-editor for MSA. see http://www.jalview.org/index.html
2. ***Muscle:*** ***https://www.ebi.ac.uk/Tools/msa/muscle/***
3. PRANK: https://www.ebi.ac.uk/research/goldman/software/prank.
4. MAFFT: https://mafft.cbrc.jp/alignment/software/

14

What are the benefits of MSA?

1. Find out which parts "do the same thing"

   Similar genes are conserved across widely divergent species, often performing similar functions.

2. Structure prediction

   Use knowledge of structure of one or more members of a protein MSA to predict structure of other members

3. Create "profiles" for protein families

   Allow us to search for other members of the family

4. Genome assembly: how many gene in this genome.

5. MSA is to build a phylogenetic analysis.

15

How does MSA do matching?

- By Bringing the greatest number of similar characters into the same column of the alignment
- Similar to alignment of two sequences.

Pairwise alignment:

Query:	DALCD	
Hit1:	DALCR	**Both hits are**
Hit2:	RALCD	**equally probable**

... a profile is buit from a MSA

polar/charged hydrophobic small

DALCD
KAIGD
EALAD

... A profile search allows you to rank the two hits

Hit1:	DALCR	DALCR
Hit2:	RALCD	RALCD
	weak match	**strong match**

16

How MSA is calculated ?

- **Using: (do you remember!)**

1. **Optimal Global Alignments** -Dynamic programming
2. **Global Progressive Alignments** - Match closely-related sequences first using a guide tree
3. **Global Iterative Alignments** - Multiple re-building attempts to find best alignment
4. **Local alignments**

17

How does software work?

If you have multiple seq: there are more possibilities to align :

We need to re-arrange the sequences to find the best similarities.

Score of multiple alignment = $\sum_{i<j}$ score(S_i,S_j) where score(S_i,S_j) = score of induced pairwise alignment

S=sequence

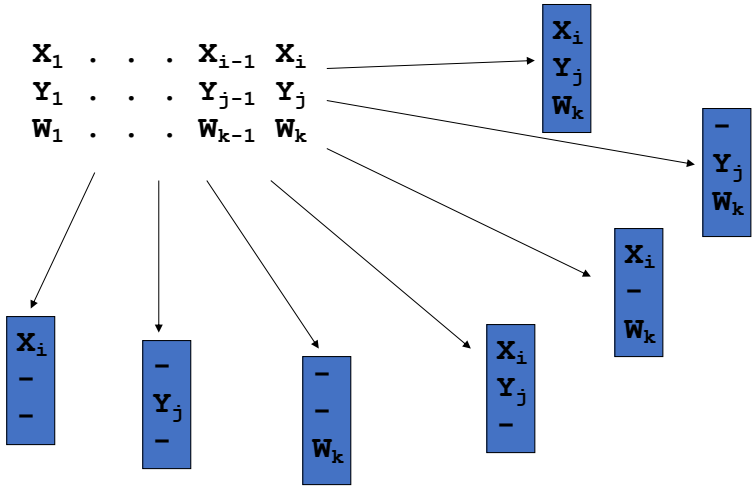
S_1	S	-	T	I	S	C	T	G	-	S	-	N	I
S_2	L	-	T	I	-	C	N	G	S	S	-	N	I
S_3	L	R	T	I	S	C	S	G	F	S	Q	N	I

Induced pairwise alignment of S_1, S_2:

S_1	S	T	I	S	C	T	G	-	S	N	I
S_2	L	T	I	-	C	N	G	S	S	N	I

18

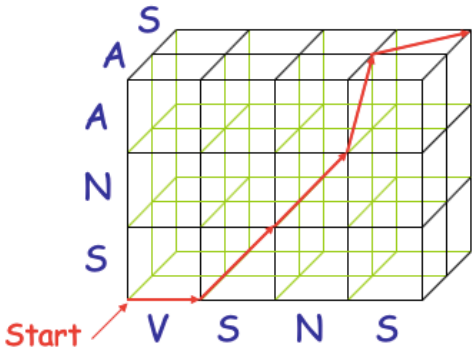
How many possibilites are made from 3 sequences?
MSA: 7 ways alignment can end for 3 sequences



19

Simulate the Dynamic programming for three sequences?

Each alignment is a path through the dynamic programming matrix

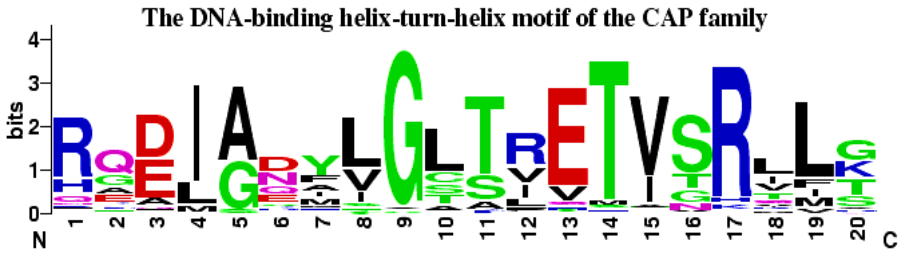


V S N – S
– S N A –
– – – A S

20

How to find the most conservative amino acid in a seq among multiple species

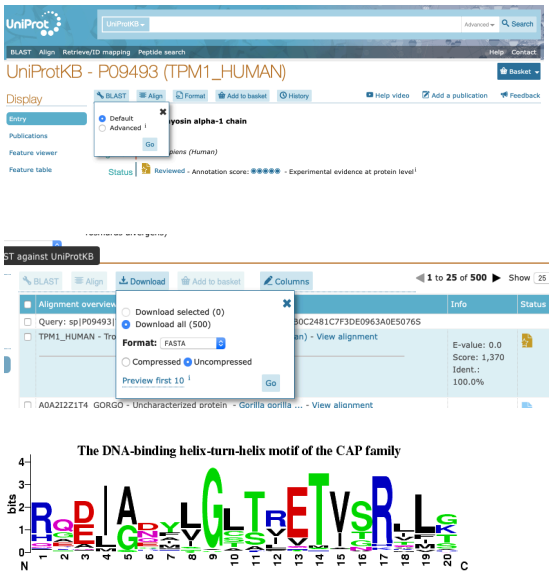
- **Sequence Logos and conservativity can be found using**
- http://weblogo.berkeley.edu/
- Sequence logos are based on **M**ultiple **S**equence **A**lignments
- Very useful to visualise Sequence profiles and motifs.



21

Tutorial

- Find a TPM1 (tropomyosin) gene for human by typing it in www.uniport.com. Type the gene name
- Go inside the its page do click alignment.
- The job will take time.
- Download seq, paste it in https://www.uniprot.org/blast/uniprot/B202003208BC4D7ADE02784B0C2481C7F3DE0963A0E5076S
- Download the whole seq, paste it in http://weblogo.berkeley.edu/logo.cgi



22