

Bioinformatics I

Multiple Sequence Alignment MSA using Dynamic Programming

Dr Manaf A Guma

University Of Anbar- college of Applied sciences-Hit

Department of applied 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 do we use dynamic programming in MSA?

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

1. BLAST.

FASTA format) do you remember it !

2. FASTA.

```
>AT1G09780 | 1 | training
GTGGAGTAGAAGAATTGAGAGCCTTATCAG
TTTTTGAAGAGAGGGCTGAAACTCTCTAGT
TATCTTTTGTGCTTTTCTAATAATAAGAG
TTTACACACAG
```

Part 1

Part 2

Part 3

3. ClustalW.

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 matrices used to generate the comparison.

4

Can we get MSA from 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

The screenshot shows a BLAST search results page with the following components:

- Navigation tabs:** Descriptions (selected), Graphic Summary, Alignments, Taxonomy.
- Actions:** Download, Manage Columns, Show 100, and a help icon.
- Summary:** 100 sequences selected.
- Table:** A table with columns: Description, Max Score, Total Score, Query Cover, E value, Per. Ident, and Accession.

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 scammonii]	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 [Otlemur 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

Tools > Sequence Similarity Searching > FASTA

Results for job fasta-l20200310-083954-0267-59723302-p2m

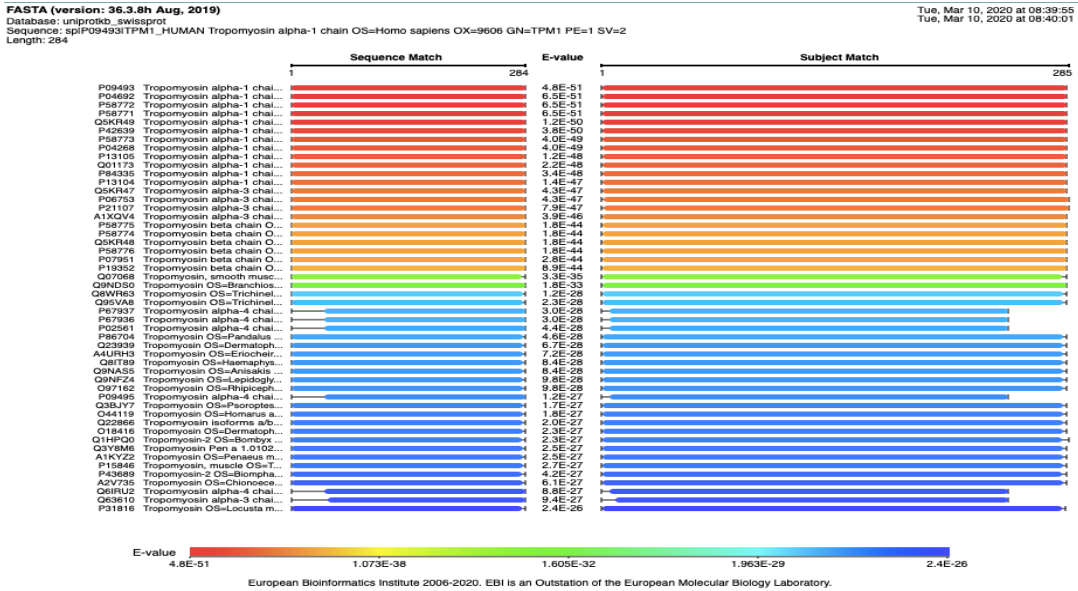
[Summary Table](#)
[Tool Output](#)
[Visual Output](#)
[Functional Predictions](#)
[Submission Details](#)

Align.	DB:ID	Source	Length	Score (Bits)	Identities %	Positives %	E()
<input checked="" type="checkbox"/>	SP:P09493	Tropomyosin alpha-1 chain OS=Homo sapiens OX=9606 GN=TPM1 PE=1 SV=2 <i>Cross-references and related information in:</i> ▶ Gene expression ▶ Bioactive molecules ▶ Nucleotide sequences ▶ Genomes & metagenomes ▶ Literature ▶ Samples & ontologies ▶ Diseases ▶ Molecular interactions ▶ Protein families ▶ Macromolecular structures ▶ Protein expression data ▶ Reactions & pathways ▶ Protein sequences	284	202.6	100.0	100.0	4.8E-51
<input checked="" type="checkbox"/>	SP:P04692	Tropomyosin alpha-1 chain OS=Rattus norvegicus OX=10116 GN=Tpm1 PE=1 SV=3 <i>Cross-references and related information in:</i> ▶ Gene expression ▶ Bioactive molecules ▶ Nucleotide sequences ▶ Genomes & metagenomes ▶ Literature ▶ Samples & ontologies ▶ Molecular interactions ▶ Protein families ▶ Macromolecular structures ▶ Protein expression data ▶ Reactions & pathways ▶ Protein sequences	284	202.1	99.6	100.0	6.5E-51
<input checked="" type="checkbox"/>	SP:P58772	Tropomyosin alpha-1 chain OS=Oryctolagus cuniculus OX=9986 GN=TPM1 PE=1 SV=1 <i>Cross-references and related information in:</i> ▶ Bioactive molecules ▶ Nucleotide sequences ▶ Literature ▶ Samples & ontologies ▶ Molecular interactions ▶ Protein families ▶ Macromolecular structures ▶ Protein expression data ▶ Protein sequences	284	202.1	99.6	100.0	6.5E-51
<input checked="" type="checkbox"/>	SP:P58771	Tropomyosin alpha-1 chain OS=Mus musculus OX=10090	284	202.1	99.6	100.0	6.5E-51

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 last 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      -----MDAIKKRMLAMKMEKENAIDRAEQMEQKVRDVEETFNKLEEFNNLQKFFSNLQ      54
UNIPROT:TPM1_CAEL      -----MDAIKKRMOAMKIEKDNALDRADAEEKVRQITKLERVEEELRDTPKRMQTG      54
UNIPROT:TPM1_ANISI      -----MDAIKKRMOAMKIEKDNALDRADAEEKVRQMTDKLERVEEELRDTPKRMQTE      54
UNIPROT:TPM1_TRICO      -----MDAIKKRMOAMKIEKDNALDRADAEEKVRQITKLERVEEELRDTPKRMQTE      54
UNIPROT:TPM1_TRIPS      -----MDAIKKRMOAMKIEKDNAMDRADAEEKARQQQERVKLEEEELRDTPKRMQVE      54
UNIPROT:TPM2_TRISP      -----MDAIKKRMOAMKIEKDNAMDRADAEEKARQQQERVKLEEEELRDTPKRMQVE      54
UNIPROT:TPM2_BOHMO      -----MDAIKKRMOAMKLEKDNALDRAMCEQQAKDANLRAEKAEERARLQKKIQTIE      54
UNIPROT:TPM1_LOCHI      -----MDAIKKRMOAMKLEKDNALDRALLCEQQAKDANLRAEKAEERARLQKKIQTIE      54
UNIPROT:TPM1_FANBO      -----MDAIKKRMOAMKLEKDNAMDRADTLEQQNKEANNRAEKSEEEVFLQKKLQOLE      54
UNIPROT:TPM1_FENMO      -----MDAIKKRMOAMKLEKDNAMDRADTLEQQNKEANNRAEKSEEEVHNLQKRMQOLE      54
UNIPROT:TPM1_FENAT      -----MDAIKKRMOAMKLEKDNAMDRADTLEQQNKEANNRAEKSEEEVHNLQKRMQOLE      54
UNIPROT:TPM1_CHIOP      -----MDAIKKRMOAMKLEKDNAMDRADTLEQQNKEANNRAEKTEEEIRANQKKSQVLE      54
UNIPROT:TPM1_ERISI      -----MDAIKKRMOAMKLEKDNAMDRADTLEQQNKEANNRAEKTEEEIRATQKRMQVE      54
UNIPROT:TPM1_HOMAM      -----MDAIKKRMOAMKLEKDNAMDRADTLEQQNKEANNRAEKTEEEIRITHKRMQVE      54
UNIPROT:TPM1_LEPDS      -----MEALNNMQAMKLEKDNADRAEIAEQSRDANLRAEKSEEEVRLQKKIQOLE      54
UNIPROT:TPM1_DERFT      -----MEALNNMQAMKLEKDNADRAEIAEQKARDANLRAEKSEEEVRLQKKIQOLE      54
UNIPROT:TPM1_DERFA      -----MEALNNMQAMKLEKDNADRAEIAEQKARDANLRAEKSEEEVRLQKKIQOLE      54
UNIPROT:TPM1_PSOOV      -----MEAIKKRMOAMKLEKDNADRAEIAEQKARDANLRAEKSEEEVRLQKKIQOLE      54
UNIPROT:TPM1_HAELO      -----MDAIKKRMOAMKLEKDNADRAEIAEQSRDANLRAEKSEEEVRLQKKIQOLE      54
UNIPROT:TPM1_RHIMP      -----MEAIKKRMOAMKLEKDNADRAEIAEQSRDANLRAEKSEEEVRLQKKIQOLE      54
UNIPROT:TPM3_RAT      MAGSTTIEAVKRRIOVLQQA-----                21
UNIPROT:TPM4_RAT      MAGLNSLEAVKRRIQALQQA-----                21
UNIPROT:TPM4_PIG      MAGLNSLEAVKRRIQALQQA-----                21
UNIPROT:TPM4_HUMAN      MAGLNSLEAVKRRIQALQQA-----                21
UNIPROT:TPM4_HORSE      MAGLNSLEAVKRRIQALQQA-----                21
UNIPROT:TPM4_MOUSE      MAGLNSLEAVKRRIQALQQA-----                21
UNIPROT:TPM1_CIOIN      -----MEAIKKRMTMLKLDKENAIDRAEQMETDRKSAEDKATGLEELQGLQKRLKATE      54
    
```

12

The old version presentation of the ClustalW

```

TBD_1265/493-734 493 TR LROALERNELVLHYQPIVELASGRIVGGEALVRWEDPERGLVMPSAFIPAAEDTGLIVALSDWVLEACQTQLRAWQQQ573
YahA7-246      7 EAILSALLENHFKPWIQPVFCAQTVLTVGCEVLVLRWEHPQTGIIPDPGFIPLAESSGLIYIMTRQLMKGTADILMPVKH--85
FimK2/7-242   7 SELVHAIONGGVYVPFQPIVDIHL-HIKGIEVLRSWRKQGV-VLLPTFFLNIOSEA|WFSLTAFVLEAVQGINRYOG--83
CKO_03715/1-236 1 REFIIHAHSQGVFPVFPITDGHILRLQGVIELSRWRGDN-VLLPGEFLPQIHAEYAWLLTAFVLEIAIONIQHQG--77
FimK7-242     7 QEWQAIHDRGVFPVFPQIVDSRS-QLQGVIELIWRWRHQ-VLHRQTFPHFRADYTWLLELAFVLEAVQNINEYPG--83
PigX/7-240    7 TLEHTLSRGGPRLYQKPAITREG-EVHHRLELSRIYDGSQ-ELLAAEYMP|LVRQLGLTASYDRQLITRSIALVSWP--82
MrkJ7-230     7 ENILSRNDIARVYVFKMFSPOG-TLVAVCELSRFD---NLSEISPEDFFRHAT-----AAVRERIFLQQLALEKHKAA--76

TBD_1265/493-734 574 RAADDLTLSVNI STROFEGEHLTRAVDRA LARSGLRFDCELEIITENVMLVMTDEVRTCLDALRARGVRLALDDFGTGYSS654
YahA7-246      86 LLPDNFHI GINVSAGFLAAGFKECELNLVNKLGNDKIKLVLELTERNP|PVTPEARAFDLSLHQQNITFALDDFGTGYAT166
FimK2/7-242   84 EYFTVNI PTCIAHHHLICLMEAWLQLHNP LWAD--CLVLEFAETVDLTQQGNTIANMRKIQERGFRIFLDDCFQNSV162
CKO_03715/1-236 78 KFWFSINI PPCI ANHENL LRMMETARQQLQQPQWSE--RLVLEFAETVNLHQQGRTAENMDKIQRQGFRIFLDDCFQNSV156
FimK7-242     84 TFYFSVNI PSSLADSDSLRMYEAARQQLRQPEGVA--RLVLEFAETIDFRHQSRSAAHVAQLQRAGVRLVMDGCFQNSV162
PigX/7-240    83 EAVLALPITVDSL LQRPF LHWLRETLLCCKKQRQR---IFFELAEADVQCYIQRRLPILSLISGLGCR|LAVTQAGLTLVS160
MrkJ7-230     77 -WFLRNHI SATINVDDHILNLLRQKDIKAKVAALTC---VHFEVTENAENLHNSLAAWQSPQ---DLSLWLDGFGSYAG150

TBD_1265/493-734 655 LSYLSQLPFHGLKIDQSFVRKI PAHPSETQIVTILALARGLMEVVAEGIETADQYAFLRDRGCEFGGGLMSTPQAAD 734
YahA7-246      167 YRYLQAFPVDFIKIDKSFVQMASVDEISGHI VDNIVELARKKPLSIVAEGETQEQADLMI GKGVHFLOGYLYSPPVPGN 246
FimK2/7-242   163 IPIRLARFCGYKLDKSIINDFQRP HAMAALMKS LIYYQQLTQSDCIAEGVDSLKFNKLGKGLVFFQGYLFSQPVEL 242
CKO_03715/1-236 157 MFPVTRIRFSGYKLDMSI VNDFQRP HAPALIKS LLYYQQLTQSRCIAEGVDSLKFNKLGKGLVDFRQGYLFSPPITHD 236
FimK7-242     163 IFPARRLHFNAYKLDMSI VNDACHDPKALALIKS LAYYQQLTQSRCIAEGVDSLKFNKLGKGLVDFRQGYLFSPPMRR 242
PigX/7-240    161 ITYIKSLQIEIKLHPGLVRSLEKRLLENQLFVQSLTEACKGTHVKVFAVGVRTKSEWQTL LDKGVCGGGDF FASSEVVG 240
MrkJ7-230     151 INAIRGYPHFYKIKDFFWHLMRKESGRQLMDALVITFLSRNHHNVIIEGVSEAHKEWLQGMWFALQGHYWRVSI EQ 230

```

13

What other programs 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

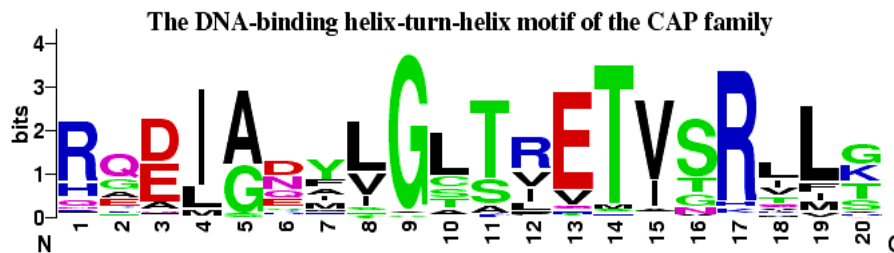
4. Genome assembly: how many gene in this genome.

5. MSA is to build a phylogenetic analysis.

15

How 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 **Multiple Sequence Alignments**
- Very useful to visualise Sequence profiles and motifs.



16

Tutorial

- Find a TPM1 (tropomyosin) gene for human by typing it in www.uniprot.com. Type the gene name
- Go inside the 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>

