

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 !

```
>AT1G09780 | 1 | training
GTGGAGTAGAAGAATTGAGAGCCTTATCAG
TTTTGAAGAGAGGGCTGAAACTCTCTAGT
TATCTTTGTGCTTTTCTAATAATAAGAG
TTTACACACAG
```

Part 1
Part 2
Part 3

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

The screenshot shows a BLAST search results page with the following table of sequences:

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
<input checked="" type="checkbox"/> PREDICTED: tropomyosin alpha-1 chain isoform X1 [Callithrix jacchus]	531	531	100%	0.0	99.65%	XP_002753250.2
<input checked="" type="checkbox"/> PREDICTED: tropomyosin alpha-1 chain isoform X5 [Chlorocebus sabaeus]	531	531	100%	0.0	99.65%	XP_008014544.1
<input checked="" type="checkbox"/> PREDICTED: tropomyosin alpha-1 chain isoform X1 [Macaca fascicularis]	531	531	100%	0.0	99.65%	XP_005559773.1
<input checked="" type="checkbox"/> tropomyosin alpha-1 chain isoform Tpm1.1st [Homo sapiens]	528	528	100%	0.0	100.00%	NP_001018005.1
<input checked="" type="checkbox"/> tropomyosin alpha-1 chain isoform 16 [Homo sapiens]	527	527	100%	0.0	99.65%	NP_001352708.1
<input checked="" type="checkbox"/> tropomyosin alpha-1 chain [Oryctolagus cuniculus]	526	526	100%	0.0	99.65%	NP_001099158.1
<input checked="" type="checkbox"/> tropomyosin alpha-1 chain isoform X2 [Lagenorhynchus obliquidens]	526	526	100%	0.0	99.65%	XP_026979007.1
<input checked="" type="checkbox"/> tropomyosin alpha striated muscle isoform [Homo sapiens]	526	526	100%	0.0	99.65%	AAT88285.1
<input checked="" type="checkbox"/> Chain A_Tropomyosin [Oryctolagus cuniculus]	526	526	100%	0.0	99.30%	2TMA_A
<input checked="" type="checkbox"/> tropomyosin alpha-1 chain isoform X2 [Heterocephalus glaber]	525	525	100%	0.0	99.30%	XP_004855748.1
<input checked="" type="checkbox"/> PREDICTED: tropomyosin alpha-4 chain isoform X6 [Chrysochloris asiatica]	525	525	100%	0.0	99.30%	XP_006831632.1
<input checked="" type="checkbox"/> tropomyosin alpha-1 chain isoform X1 [Balaenoptera acutorostrata scammonii]	525	525	100%	0.0	99.30%	XP_007166029.2
<input checked="" type="checkbox"/> PREDICTED: tropomyosin alpha-1 chain isoform X7 [Sorex araneus]	524	524	100%	0.0	99.30%	XP_004616749.1
<input checked="" type="checkbox"/> 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

Tools > Sequence Similarity Searching > FASTA

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

Summary Table Tool Output Visual Output Functional Predictions Submission Details

Selection:

Select All Invert Clear

Apply to selection:

Annotations:

Show Hide

Alignments:

Show Hide

Entries:

Download In

fasta

format

Tools:

Launch

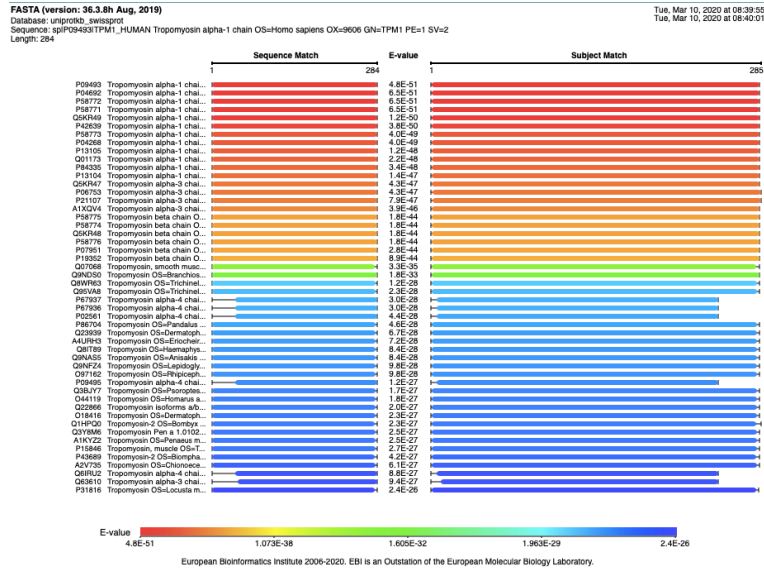
Clustal Omega

Align.	DB:ID	Source	Length	Score (Bits)	Identities %	Positives %	E()
1	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
2	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
3	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
4	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-l20200310-104708-0114-18168141-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      -----MDAIKKRMLAMKMEKENAIDRAEQMEQVDRDVEETFNKLEEEFNFLQKKFSNLQ      54
UNIPROT:TPM1_CAEEI      -----MDAIKKRQAMKIEKDNALDRADAEEKVRQITKLERVEEELRDTPKKMTQTG      54
UNIPROT:TPM1_ANISI      -----MDAIKKRQAMKIEKDNALDRADAEEKVRQITKLERVEEELRDTPKKMMQTE      54
UNIPROT:TPM1_TRICO      -----MDAIKKRQAMKIEKDNALDRADAEEKVRQITKLERVEEELRDTPKKMMQTE      54
UNIPROT:TPM1_TRIPS      -----MDAIKKRQAMKIEKDNAMDRADAEEKARQQQERVEKLEELERDTPKKMMQVE      54
UNIPROT:TPM1_TRISP      -----MDAIKKRQAMKIEKDNAMDRADAEEKARQQQERVEKLEELERDTPKKMMQVE      54
UNIPROT:TPM2_BONMO      -----MDAIKKRQAMKIEKDNALDRAMCEQQAKDANLRAEKAEERARLQKKIQTIE      54
UNIPROT:TPM1_LOCHI      -----MDAIKKRQAMKIEKDNALDRALLCEQQAKDANLRAEKAEERARLQKKIQTIE      54
UNIPROT:TPM1_PANBO      -----MDAIKKRQAMKIEKDNAMDRADTLEQQNREANNRAEKSEEEVFLQKKLQOLE      54
UNIPROT:TPM1_FENHO      -----MDAIKKRQAMKIEKDNAMDRADTLEQQNREANNRAEKSEEEVHNLQKKMQOLE      54
UNIPROT:TPM1_FENAT      -----MDAIKKRQAMKIEKDNAMDRADTLEQQNREANNRAEKSEEEVHNLQKKMQOLE      54
UNIPROT:TPM1_CHIOP      -----MDAIKKRQAMKIEKDNAMDRADTLEQQNREANNRAEKSEEEVHNLQKKMQOLE      54
UNIPROT:TPM1_ERISI      -----MDAIKKRQAMKIEKDNAMDRADTLEQQNREANNRAEKSEEEVHNLQKKMQOLE      54
UNIPROT:TPM1_HOMAM      -----MDAIKKRQAMKIEKDNAMDRADTLEQQNREANNRAEKSEEEVHNLQKKMQOLE      54
UNIPROT:TPM1_LEPDS      -----MEAIKRNQAMKLEKDNADRAEIAEQSRDANLRAEKSEEEVRLQKKIQQIE      54
UNIPROT:TPM1_DERPT      -----MEAIKRNQAMKLEKDNADRAEIAEQSRDANLRAEKSEEEVRLQKKIQQIE      54
UNIPROT:TPM1_DERFA      -----MEAIKRNQAMKLEKDNADRAEIAEQSRDANLRAEKSEEEVRLQKKIQQIE      54
UNIPROT:TPM1_FSOOV      -----MEAIKRNQAMKLEKDNADRAEIAEQSRDANLRAEKSEEEVRLQKKIQQIE      54
UNIPROT:TPM1_HAELO      -----MDAIKKRQAMKLEKDNADRAEIAEQSRDANLRAEKSEEEVRLQKKIQQIE      54
UNIPROT:TPM1_RHIMP      -----MEAIKKRQAMKLEKDNADRAEIAEQSRDANLRAEKSEEEVRLQKKIQQIE      54
UNIPROT:TPM3_RAT        MAGSTTIEAVKRRIQVLQQA-----                21
UNIPROT:TPM4_RAT        MAGLNSLEAVKRRIQALQQA-----                21
UNIPROT:TPM4_FIG        MAGLNSLEAVKRRIQALQQA-----                21
UNIPROT:TPM4_HUMAN      MAGLNSLEAVKRRIQALQQA-----                21
UNIPROT:TPM4_HORSE      MAGLNSLEAVKRRIQALQQA-----                21
UNIPROT:TPM4_MOUSE      MAGLNSLEAVKRRIQALQQA-----                21
UNIPROT:TPM1_CIOIN      -----MEAIKKRMTMLKLDKENAIDRAEQMETDRKSAEDRATGLEELQGLQKRLKATE      54
  
```

12

The old version presentation of the ClustalW

```

TBD_1265/493-734 493 TRLRQALERNELVLYHPIVEELASGRIVGGELVLRWEDPERGLVMPSAFI PA AEDTGLIVALSDWVLEACQTQLRAWQQQG573
YahA7-246      7 EAILSALLENHEFKPWIOPVFCAQTVLTCGEVLVLRWEHPQTGIIPDQFIP LAESSGLIVIMTRQLMKATADILMPVKH - - 85
FimK2/7-242   7 SELVHAIQNGOVYPVFCPIVDIHL-HIKGIEVLSRWRKGDV-VLLPTFELPNIQSEAIWFSLTAFVLC EAVQGINRYQG - - 83
CKO_03715/1-236 1 REFIIHAHISQOVFPVFCPI TDGHL-RLQGV EILSRWRGDN-VLLPGEFLPQIIHAEYAWLLTAFVLC IAIQNIHQHG - - 77
FimK7-242     7 QEWVQAIHDRQVFPVFCPIVDSRS-QLQGV EILSRWRGDN-VLLPQTFLPHFRADYTWL L TAFVLC EAVQNI NEYPG - - 83
PigX/7-240    7 T LLEHTLSRGGPRLYQKPAITREG-EVHHR E LISR IYDGSQ-ELLAAEYMP LVRQLGLTASYDRQLITRSIALTVSWP - - 82
MrkJ/7-230    7 EDNILSRNDIAVRYVFCMFSPOG-TLVAVECLSRFD-- -NLSI SPEDFFRHAT - - - - -AAVRERIFLEQLALI EKHKAA - - 76

TBD_1265/493-734 574 RAADDLTLSVNI STQFEGEHLTRAVDRLARSGLRDCLELEITENVMLVMTDEVRC L D A L R A R G V R L A L D D F G T G Y S 654
YahA7-246      86 LLFDNFHIGINVSAGCF LAAGFEKECLNLVNLKLGNDKIKLVLELTERNP I BVTREARAIFDSLHOHNI TFALDDFGTGYAT166
FimK2/7-242   84 EFYFTVNIPTCIAHHHLICLME TAWLGLHNP LWAD - -CLVLEFAETVDLTQQGNTIANMRKI QERGFRI FLDDCF SQNSV162
CKO_03715/1-236 78 KFWFSINI PPCI ANHENLRRMME TARQQLQQPQWSG - -RLVLEFAETVNLHQQGR TAENMDKIQRQGFRI FLDDCF SHS V156
FimK7-242     84 TFYFSVNI PSSLADSDSLLRMV E AARQQLRQPEGVA - -RLVLEYAETIDFRHQSRSAAHVAQLQRAGVRVMLDDCF SQSSV162
PigX/7-240    83 EAVLALPITVD SLLQRPF LHWLRETLLC P K K Q R Q R - - - I F F E L A E A D V Q O Y I G R L R P I L S L I S G L G C R L A V T Q A G L T L V S 160
MrkJ/7-230    77 -WFLRNHISATINVDDHILNLLRQKDI KAKVAALTC - - - V H F E V T N A E N L L H N S L A A W Q S P Q - - - D T S L W L D D F G S G Y A G 150

TBD_1265/493-734 655 LSYLSQLPFHGLKIDOSFVRKI PAHPSETQIVT T I L A L A R G L E M E V V A E G I E T A C Q V A F L R D R G C E F G O G N L M S T P Q A A D 734
YahA7-246      167 YRYLQAFPVDFIKIDKSFVQMASVDEISGHIVDNIVELARKPGLSIVAEQVETQE QADLMI GKGVHFLQGYLYSPPVPGN 246
FimK2/7-242   163 IPIRLARFCGYKLDKSIINDFQRDPHAMALMKSLIYYCQLTQSDCIAEGVDSL EKFNKLGKMGLVFFQGYLFSQPVEL 242
CKO_03715/1-236 157 MFPVRTIRFSGYKLDMSI V N D F Q R D P H A L A L I K S L L Y C Q L T Q S R C I A E G V D S L E K F N Q L K A L G V D R F Q G Y L F S P P I T H 236
FimK7-242     163 IFPARRLHFNAKLDMSI V N D A Q H D P K A L A L I K S L A Y Y C Q L S G S R C V A E G V D S L A K F T Q L K S L G I D R F O G Y L F S P P M R R E 242
PigX/7-240    161 I T Y I K S L Q I E I I K L H P G L V R S L E K R L E N Q L F V G S L E A C K G T H V K V F A V G V R T K S E W Q T L L D K G V C G G G D F F A S S E V G 240
MrkJ/7-230    151 I N A I R G Y H E D Y V X I D K D F F W H L M R K E S G R Q L M D A L V T F L S R N H H N V I I E G V E S E A H K E W L O G M E W F A I Q G H Y W R E V S I E Q 230

```

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

S=sequence

```

S1   S - T I S C T G - S - N I
S2   L - T I - C N G S S - N I
S3   L R T I S C S G F S Q N I

```

Induced pairwise alignment of S₁, S₂ :

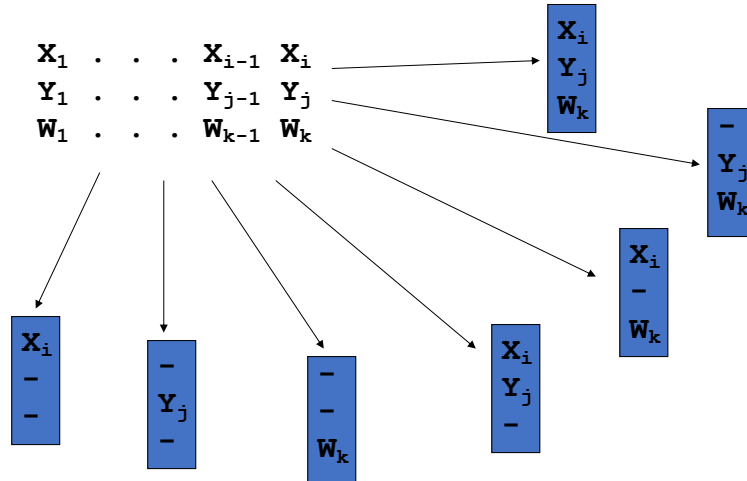
```

S1   S T I S C T G - S N I
S2   L T I - C N G S S N I

```

18

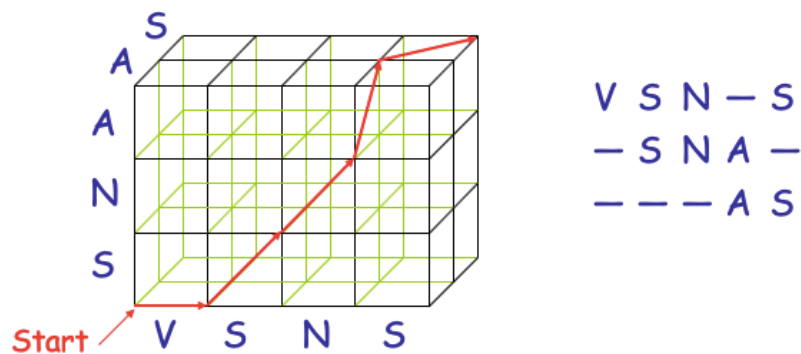
**How many possibilities are made from 3 sequences?
MSA: 7 ways alignment can end for 3 sequences**



19

Simulate the Dynamic programming for three sequences?

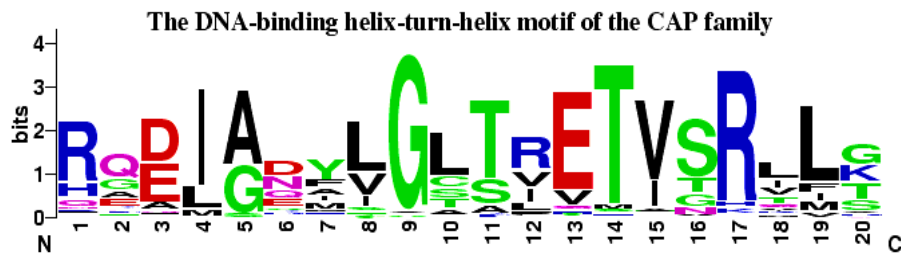
Each alignment is a path through the dynamic programming matrix



20

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

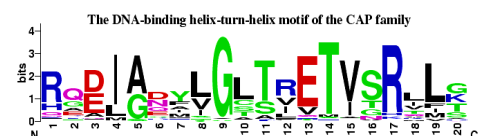
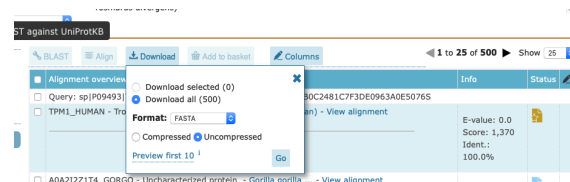
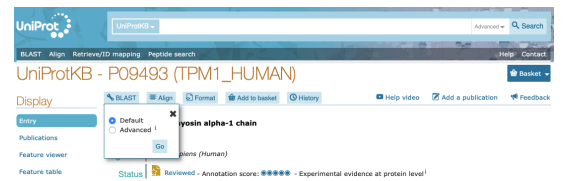
- Sequence Logos and conservativity can be found using
- <http://weblogo.berkeley.edu/>
- Sequence logos are based on Multiple Sequence Alignments
- Very useful to visualise Sequence profiles and motifs.



21

Tutorial

- Find a TPM1 (tropomyosin) gene for human by typing it in www.uniprot.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