Multiple Sequence Alignment
&
Phylogeny

Yazdan Asgari

2019

Multiple Sequence Alignment

What Is a Multiple Sequence Alignment ?

- The alignment of more than two sequences
- MSAs = multiple-sequence alignments

- The goal of an MSA is twofold:
  - Aligning corresponding regions of the sequences
  - Revealing positions that are conserved

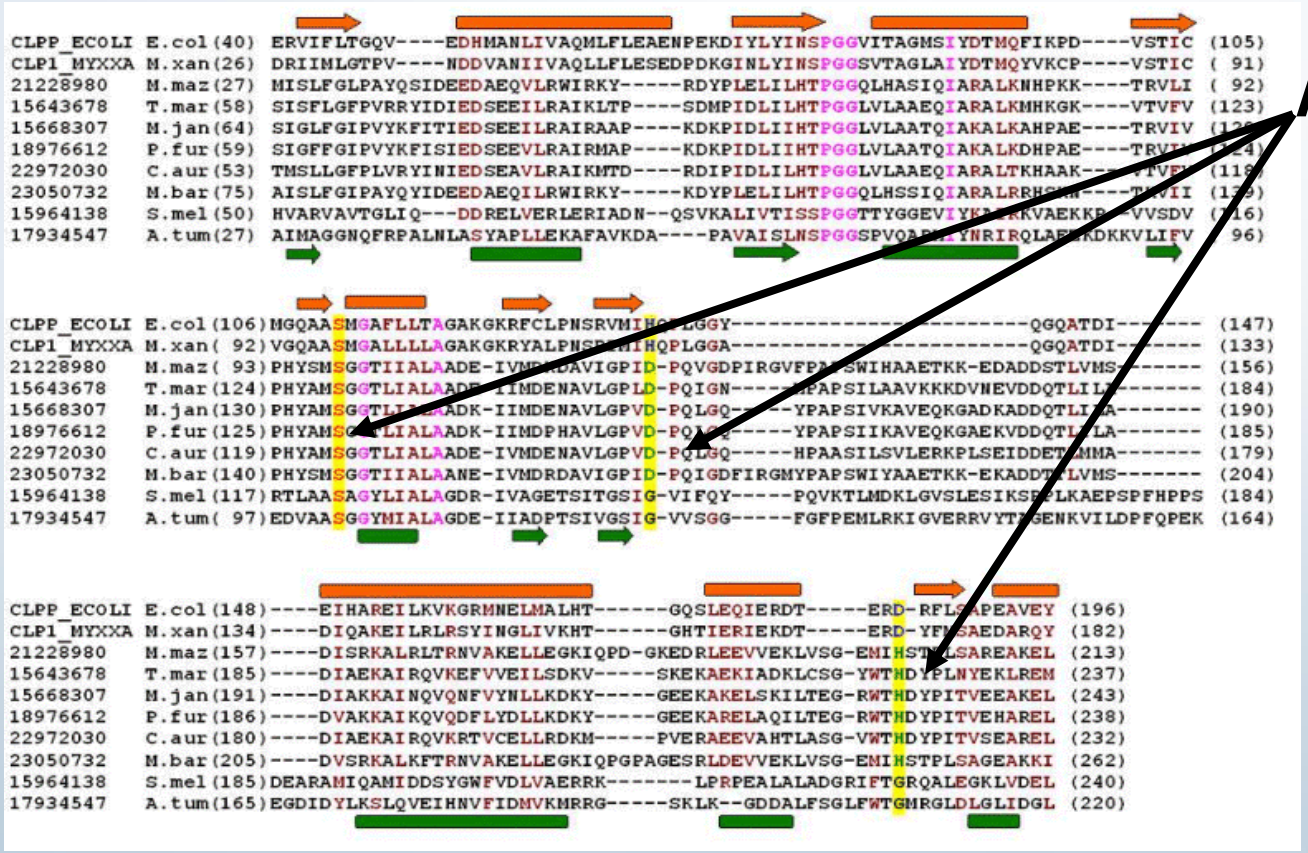
- The main steps to a useful MSA require
  - Choosing the right sequences
  - Choosing the right MSA method
  - Interpreting the alignment

Evolution

- Amino acids mutate randomly
- Mutations are then selected (accepted) or counter-selected (rejected)
- If a mutation is harmful, it is counter-selected
  - It disappears from the genome
  - You never see it
- Mutations of important positions (such as active sites) are almost always harmful
- You can recognize important positions because they never mutate!

- MSAs reveal these *conserved* positions

An Example of Conserved Positions: The Serine Proteases Active Site



Bioinformatics: Life Sciences, Robert Lessick

MSA – Example Applications

Application	*Procedure*
Extrapolation	Determine the **function** of your protein
Phylogenetic analysis	Build a Phylogenetic **tree**
Pattern identification	Discover **important positions**
Domain identification	Turn your alignment into a **domain profile**

MSA – Example Applications

DNA regulatory elements identification	Use your alignment to discover **promoters**
Structure prediction	Predict the **secondary structure** of **proteins** and **RNA** molecules
nsSNP analysis	Discover important **allelic variations** in human and other animals (nsSNP: non-synonymous single-nucleotide polymorphisms)
PCR analysis	Select your **PCR primers**

Choosing the Right Sequences

- When building an alignment, it is your job to select the sequences
- Two main factors when selecting sequences:
  - Number of sequences
  - Nature of the sequences
- A reasonable number of sequences: 20 to 50
  - Ideal for most methods
  - Small alignments are easy to display and analyze
- Types of sequences
  - Well-selected sequences ⇔ informative alignment

Choosing Sequences that are Different Enough

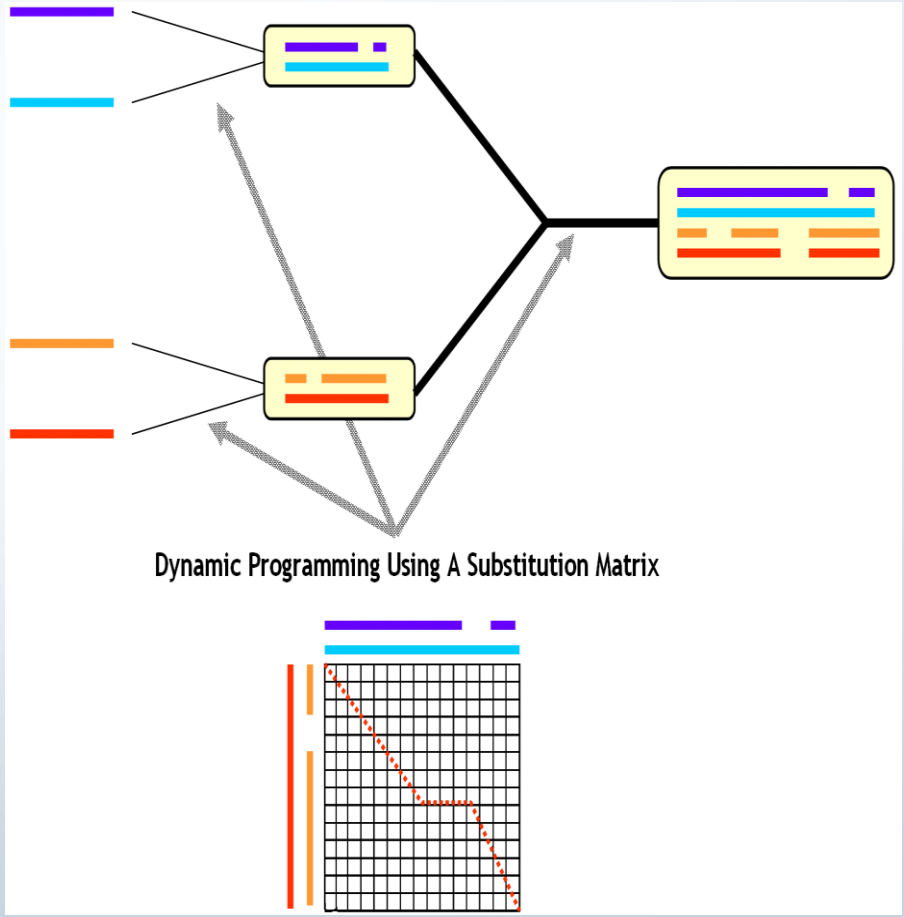
- An alignment is useful if . . .
  - The sequences are correctly aligned
  - It can be used to produce trees, profiles, and structure predictions
- To obtain this result, the sequences must be
  - Not too similar
  - Not too different
- Sequences that are very similar . . .
  - Are easy to align correctly
  - Are not informative ⇔ useless trees and profiles, bad predictions
- Sequences that are very different . . .
  - Are difficult to align
  - Are very informative ⇔ good trees and profiles, good predictions

DNA or Proteins?

- DNA sequences are harder to align than proteins
  - DNA-comparison models are less sophisticated
- Most methods work for both DNA and proteins
  - The results are less useful for DNA
- If your DNA is coding, work on the translated proteins
- If sequences are homologous . . .
  - Along their entire length ⇔ use progressive alignment methods
  - In terms of local similarity ⇔ use motif-discovery methods

MSA Methods: Progressive Algorithm

- Sequences are grouped by similarity (guide tree)
- Sequences are aligned 2 by 2
- The intermediate alignments are then aligned 2 by 2
- You align 2 sequences by using dynamic programming



Bioinformatics: Life Sciences, Robert Lessick

MSA Methods: Progressive Algorithm

- Its main strength is its speed
- Its main weakness is its greed
  - Sequences aligned at the beginning are never realigned
  - Early mistakes cannot be corrected
- Assemble datasets with lots of intermediate sequences
- Imagine each sequence is part of a stone bridge across a river:
  - Doesn't matter how wide the river is, if the stones are close enough together
  - Doesn't matter how diverse your sequences are, if each sequence has a close relative

https://www.tes.com

Selecting a Method

- Many alternative methods exist for MSAs
- Most of them use the **progressive algorithm**
- They all are approximate methods
- None is guaranteed to deliver the best alignments
- All existing methods have pros and cons
  - **ClustalW** is the most popular
  - T-Coffee/M-Coffee and ProbCons are more accurate but slower
  - MUSCLE is very fast, ideal for very large datasets

Aligning Your Sequences Correctly

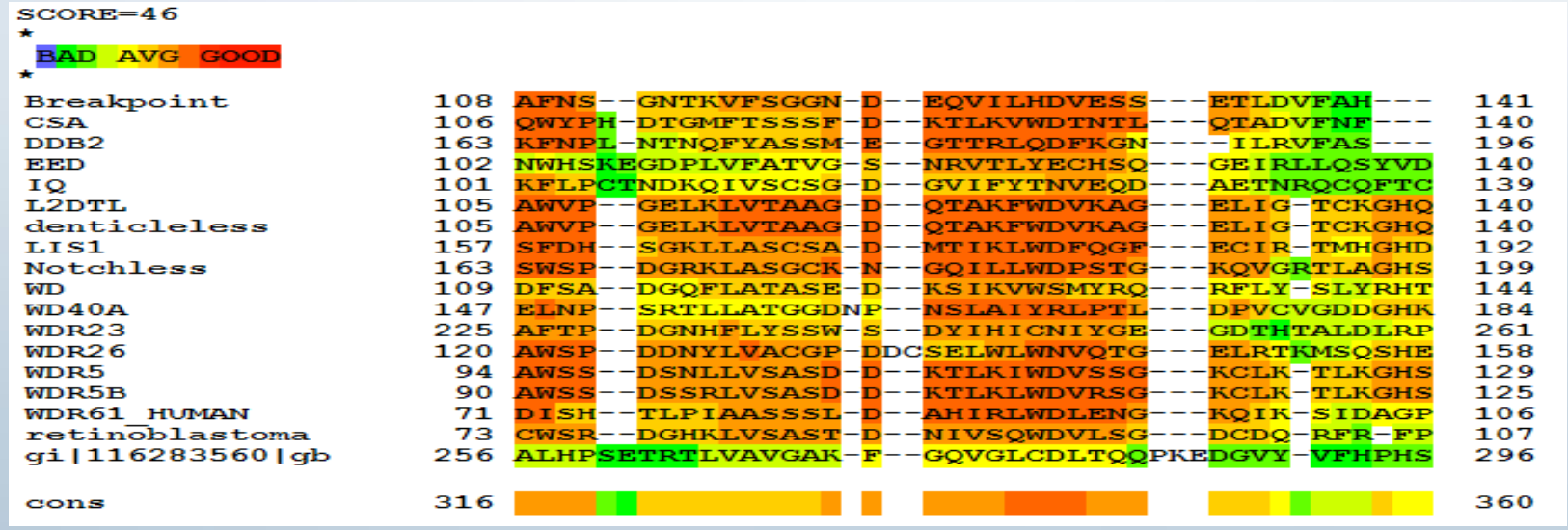
- It can be difficult to align sequences correctly
  - They evolve too fast
- For proteins, the best alternative is to use 3D structure
  - 3D Structures change slower than sequences
- Unfortunately few sequences have a known structure
- Expresso lets you find the structures that correspond to your sequences and use them to build an MSA

Using the EXPRESSO Web server

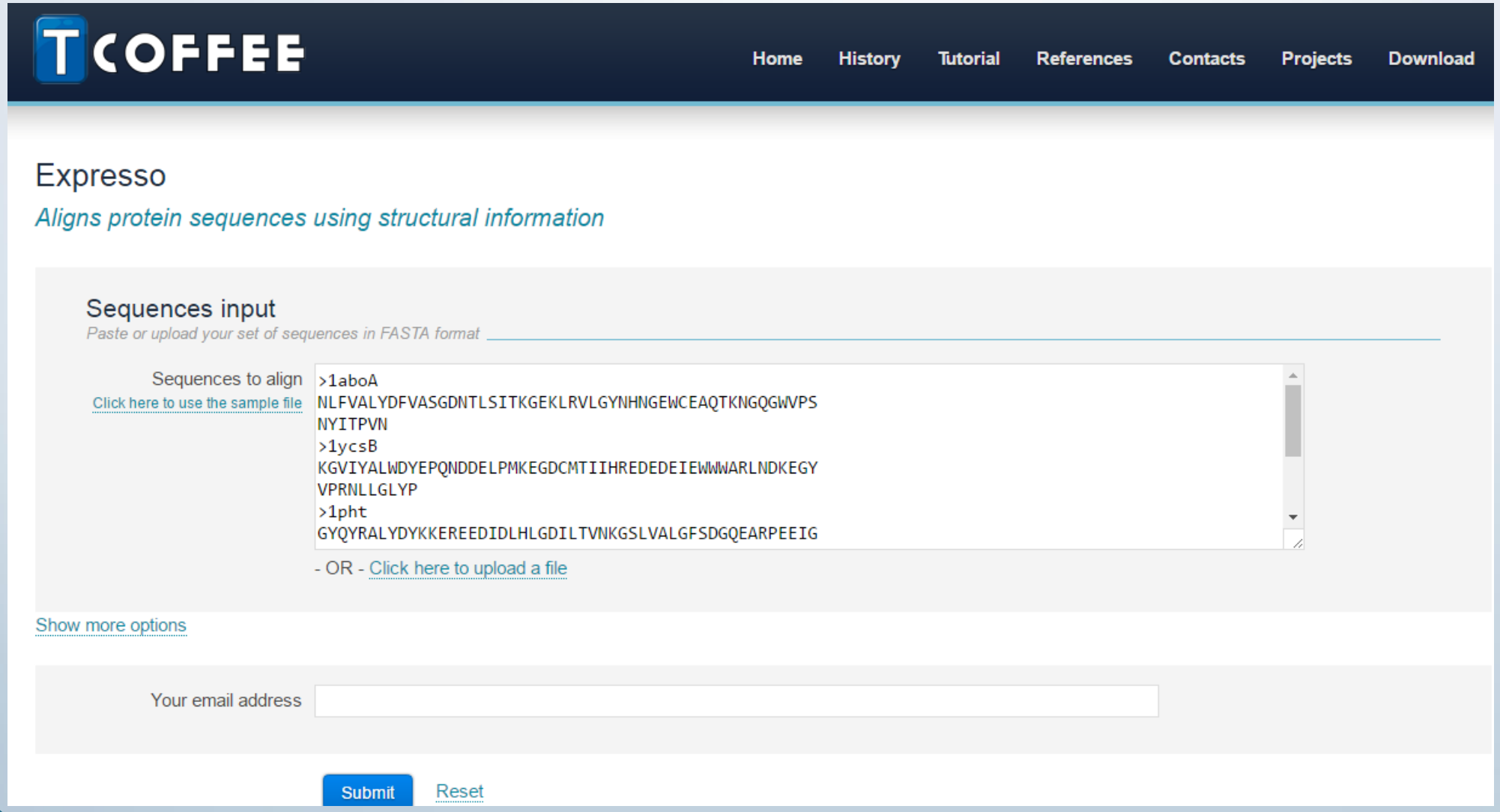
- The EXPRESSO Web server aligns sequences using 3D information
- The EXPRESSO tool . . .
  - Looks in PDB for the structure of your sequences
  - Aligns your sequences using structural information
  - Returns a multiple-sequence alignment based on structure
- If your sequences have a known structure, EXPRESSO is the most accurate MSA method

EXPRESSO Web server – Colored Output

- Red and Orange residues are probably well aligned
- Yellow should be treated with caution
- Green and blue are probably incorrectly aligned



EXPRESSO Web server - Example



http://tcoffee.crg.cat/apps/tcoffee/do:expresso

EXPRESSO Web server - Example

Expresso alignment result

MSA

The multiple sequence alignment result as produced by T-coffee.

```
T-COFFEE, Version_11.00.d625267 (2016-01-11 15:25:41 - Revision d625267 - Build 507)
Cedric Notredame
SCORE=91
*
 BAD AVG GOOD
*
1aboA    :  92
1ycsB    :  92
1pht     :  88
1vie     :  91
1ihvA    :  91
cons     :   9

1aboA   -NLFVALYDFVASGDNTLSITKGEKLRVL-----------GYNHN---GEWCEAQTK---NGQGWVPSNYITPVN
1ycsB   KGVIYALWDYEPQNDDELPMKEGDCMTII-----------HRED-EDEIEWWWARLN---DKEGYVPRNLLGLYP
1pht    GYQYRALYDYKKEREEDIDLHLGDILTVNKGSLVALGFSDGQEARPEEIGWLNGYNETT-GERGDFPGTYVEYIG
1vie    -DRVRKKSG------------AAWQGQIV-----------GWYCTNLTPEGYAVESEAHPGSVQIYPVAALERIN
1ihvA   NFRVYYRDS--------RDPVWKGPAKLL-----------WKG-----EGAVVIQDN---SDIKVVPRRKAKIIR

cons        .          :           :    .     *
```

Local Multiple Comparison Methods

- Most MSA programs assume your sequences are related along their whole length
- When this assumption is not true, the progressive approach will not work
- The only alternative is to compare multiple sequences locally

Local Multiple-Comparison Methods

- Gibbs Sampler
  - Will make a local multiple alignment
  - Will ignore unrelated segments of your sequences
  - Ideal for finding DNA patterns such as promoters
- Motif discovery methods
  - Will look for motifs conserved in your sequences
  - The sequences do not need to be aligned
- The most popular motif-discovery methods:
  - TEIRESIAS, MEME, SMILE, PRATT

MSA - Summary

- Assembling MSAs is a bit of an art
- Experience is a key factor
- Most methods are now available online
- Make sure you know which method to use:
  - ClustalW-like method to align homologous sequences
  - Motif method to look for conserved regions

Phylogeny

Why Build a Phylogenetic Tree ?

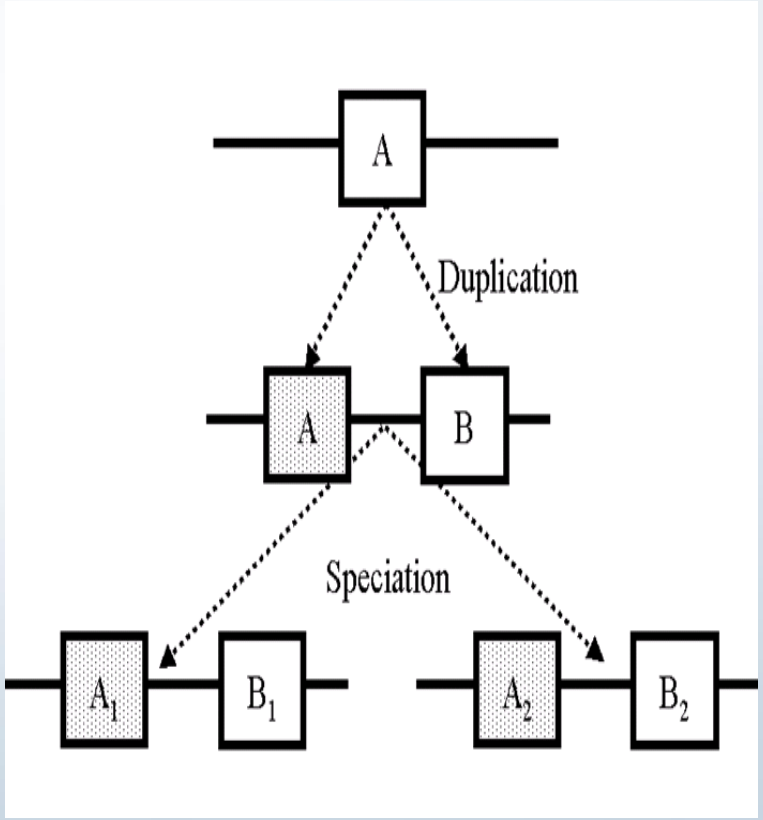
- Phylogenetic trees reconstruct the evolutionary history of your sequences
- They tell you who is closer to whom in the big tree of life
- Phylogenetic trees are based on sequence similarity rather than morphologic characters

3 applications to use your constructed Tree

- Finding the closest relative of your organism
  - Usually done with a tree based on the ribosomal RNA
- Discovering the function of a gene
  - Finding the orthologues of your gene
- Finding the origin of your gene
  - Finding whether your gene comes from another species

Orthology and Paralogy

- Orthologous genes
  - Separated by speciation
  - Often have the same function
- Paralogous genes
  - Separated by duplications
  - Can have different functions
- In the graph:
  - A is paralogous with B
  - A1 is orthologous with A2



Bioinformatics: Life Sciences, Robert Lessick

Working on the Right Data

- Garbage in ⇨ Garbage out
- The quality of your tree depends on the quality of the data
- Your first task is to assemble a very accurate MSA

Again: DNA or Proteins

- Most phylogenetic methods work on **Proteins** and **DNA** sequences
- If possible, always compute a multiple-sequence alignment on the protein sequences
  - Translate the sequences if the DNA is coding
  - Align the sequences
  - Thread the DNA sequences back onto the protein MSA
- If your DNA sequences are coding and have more than 70% identity . . .
  - Compute the **tree** on the **DNA** multiple-sequence alignment
- If your DNA sequences are coding and have less than 70% identity . . .
  - Compute the **tree** on the **protein** multiple-sequence alignment

Which Sequences ?

- Orthologous sequences
  - Produce a species tree
  - Show how the considered species have diverged
- Paralogous sequences
  - Produce a gene tree
  - Show the evolution of a protein family

Building the Right Tree

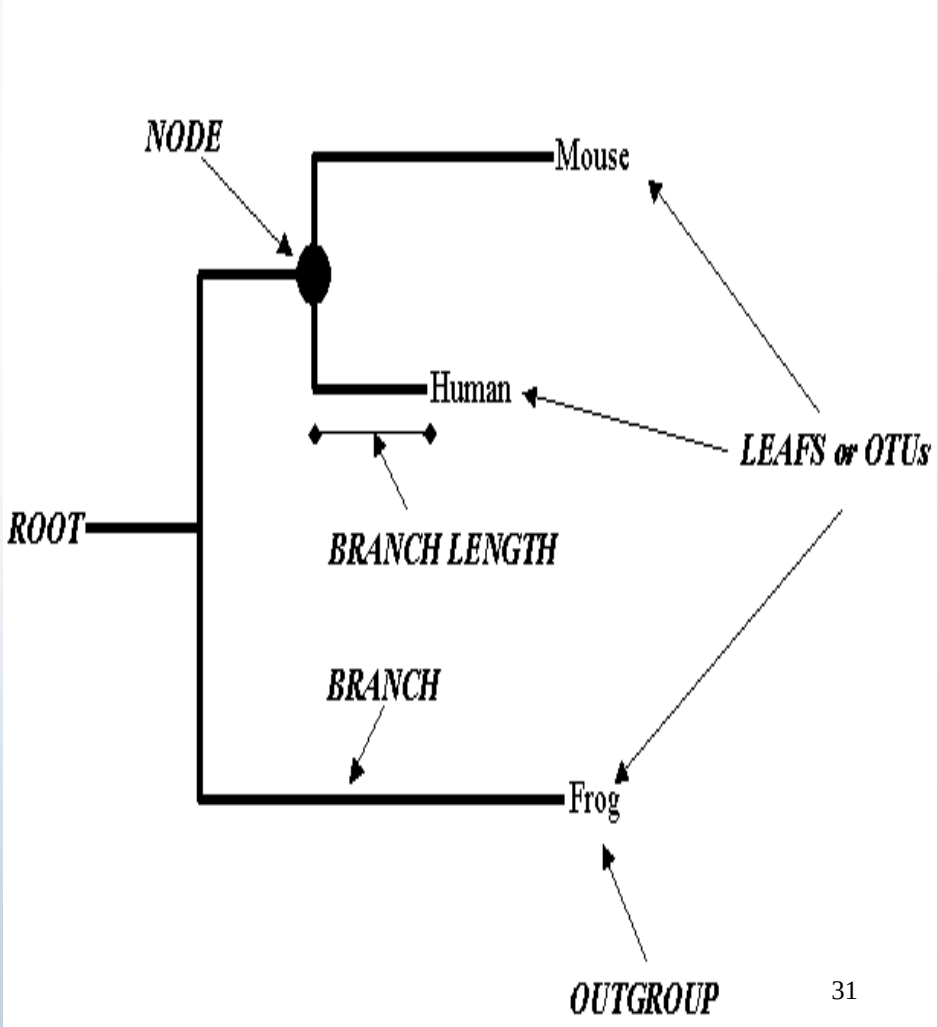
- There are two types of tree-reconstruction methods
  - Distance-based methods
  - Statistical methods
    - Statistical methods are the most accurate
      - Maximum likelihood of success
      - Parsimony
    - Statistical methods take more time
    - Limited to small datasets

Distance-based Methods for Tree Reconstruction

- Distance-based methods are the most popular
  - Neighbor Joining (NJ)
  - UPGMA
- Distance-based methods involve 2 steps:
  - Measure the distances between pairs of sequences in the MSA
  - Transform the distance matrix into a tree
- The two most popular packages for making trees are
  - **Clustalw**: very simple, not very sophisticated
  - **Phylip**: very powerful, less convivial

Reading Your Tree

- There's a lot of vocabulary in a tree
- **Nodes** correspond to common ancestors
- R**oot** is the oldest ancestor
  - Often artificial
  - Only meaningful with a good outgroup
- Trees can be un-rooted
- Branch lengths are only meaningful when the tree is scaled
  - **Cladograms** have the same branch lengths
  - **Phylograms** have real branch lengths



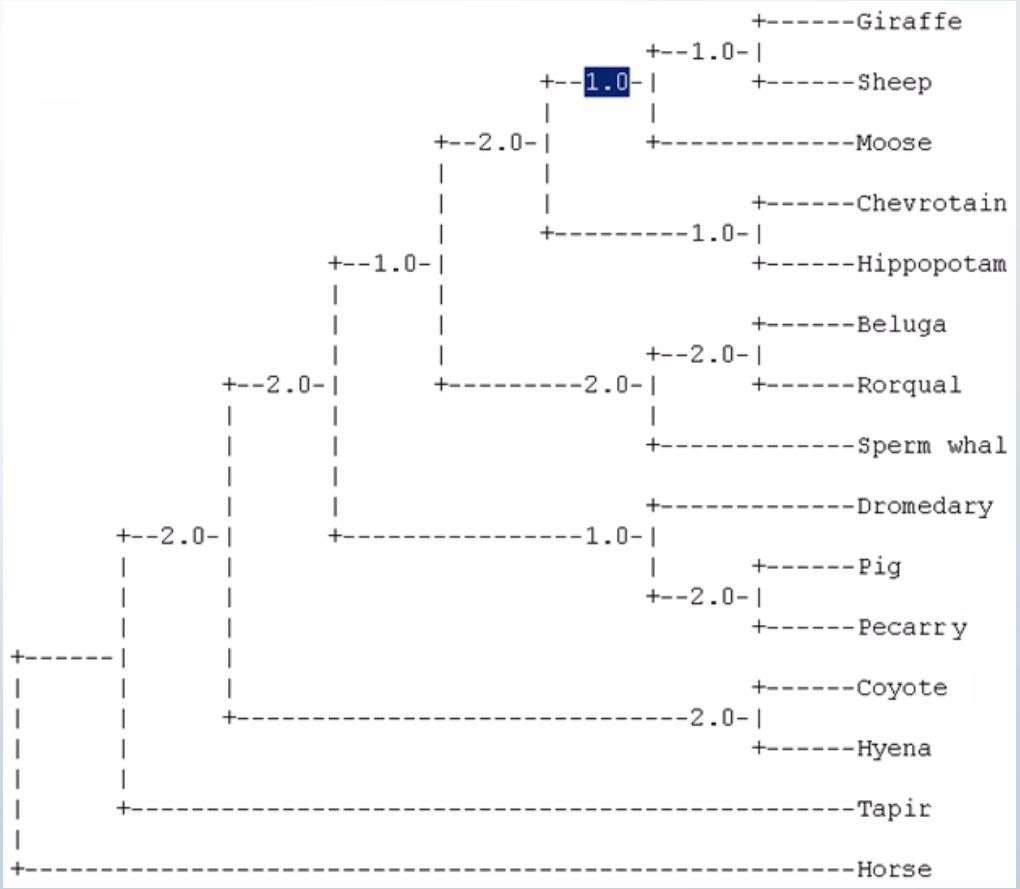
Bioinformatics: Life Sciences, Robert Lessick

Bootstrapping

- Use bootstrapping to verify the solidity of each node
- ClustalW and Phylip do bootstrap operations automatically
- Bootstrapping involves these steps:
  - Select a subset of your MSA
  - Redo the tree
  - Repeat this operation N times (100 or 1000 times if you can)
  - Compute a consensus tree of the N trees
  - Measure how many of the N trees agree with the consensus tree on each node
- Each node gets a bootstrap figure between 0 and N
- High bootstrap ⇔ good node

A Bootstrapped Tree

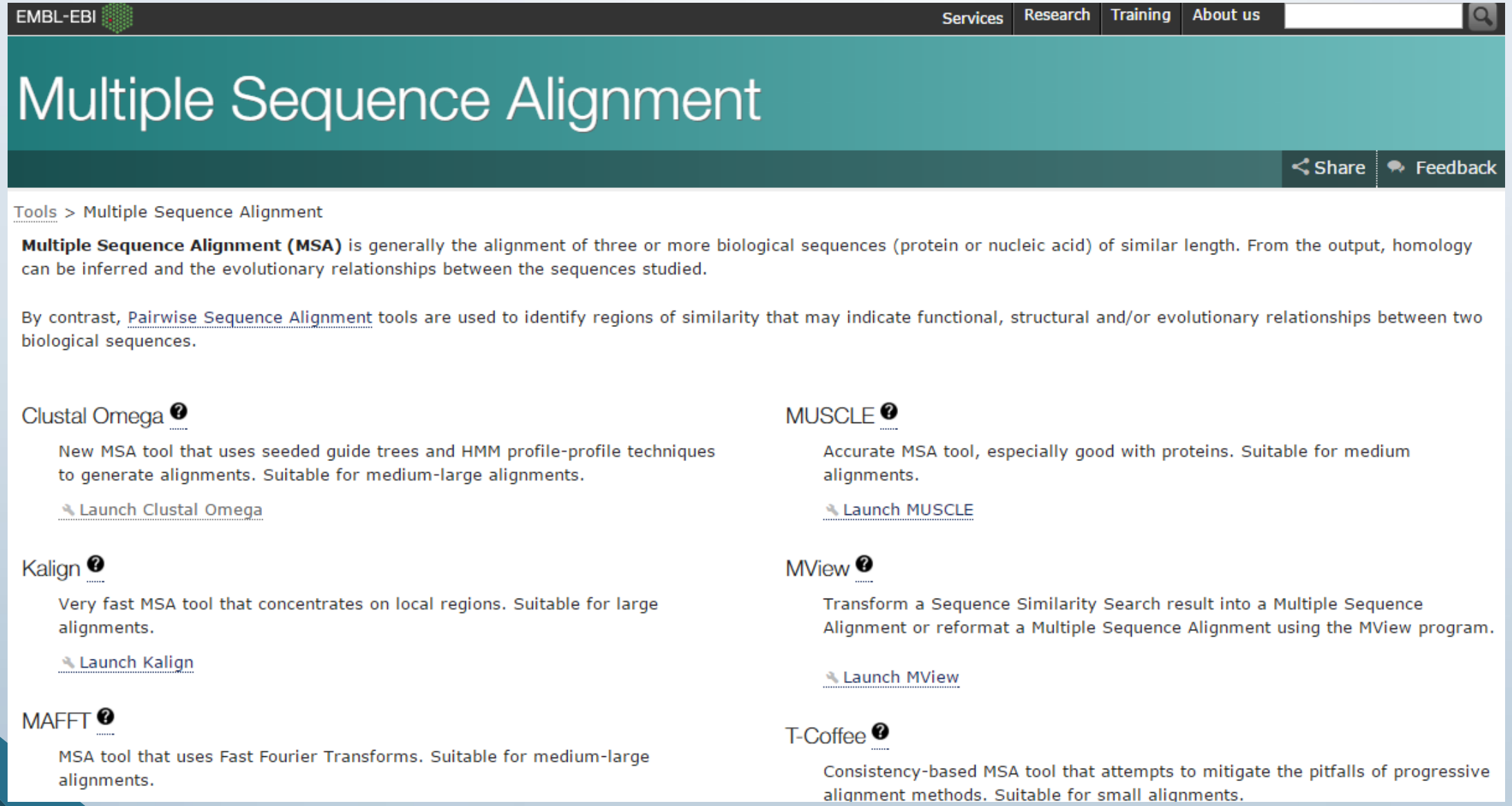
- This tree was produced with 2 bootstrap cycles
- It shows some nodes as more robust than others
- In practice, always use more than 100 cycles



Bioinformatics: Life Sciences, Robert Lessick

Practical Session

EMBL-EBI



http://www.ebi.ac.uk/Tools/msa/

Example: Human TNF-alpha orthologous

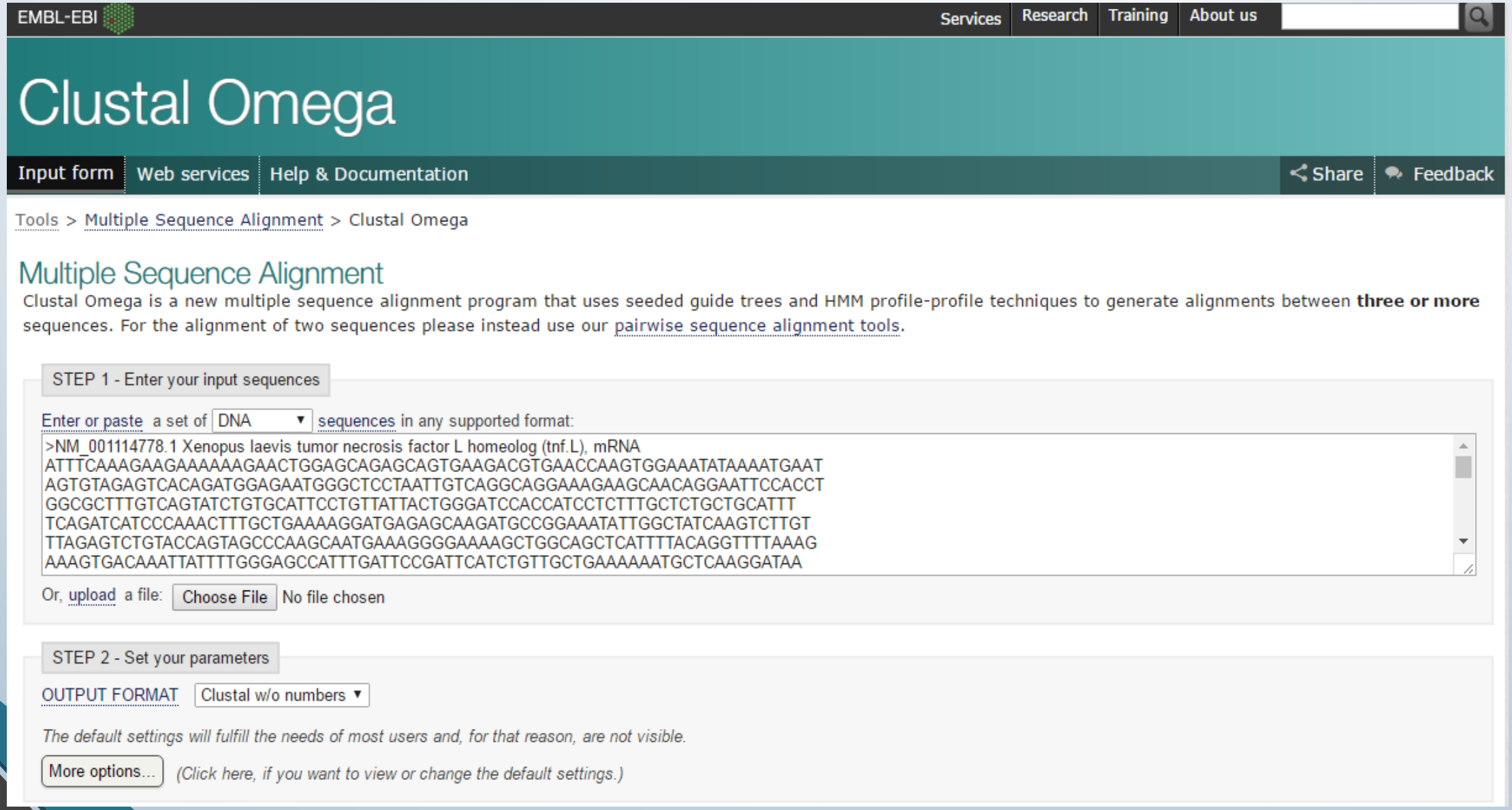
TNF tumor necrosis factor [*Homo sapiens* (human)]

Gene ID: 7124, updated on 6-Dec-2016

Summary

Official Symbol	TNF provided by HGNC
Official Full Name	tumor necrosis factor provided by HGNC
Primary source	HGNC:HGNC:11892
See related	Ensembl:ENSG00000232810 MIM:191160; Vega:OTTHUMG00000031194
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	DIF; TNFA; TNFSF2; TNLG1F; TNF-alpha
Summary	This gene encodes a multifunctional proinflammatory cytokine that belongs to the tumor necrosis factor (TNF) superfamily. This cytokine is mainly secreted by macrophages. It can bind to, and thus functions through its receptors TNFRSF1A/TNFR1 and TNFRSF1B/TNFBR. This cytokine is involved in the regulation of a wide spectrum of biological processes including cell proliferation, differentiation, apoptosis, lipid metabolism, and coagulation. This cytokine has been implicated in a variety of diseases, including autoimmune diseases, insulin resistance, and cancer. Knockout studies in mice also suggested the neuroprotective function of this cytokine. [provided by RefSeq, Jul 2008]
Orthologs	mouse all

Clustal Omega / DNA



Results

Results for job clustalo-I20161215-140559-0800-79720144-pg

Alignments | Result Summary | Phylogenetic Tree | Submission Details

Download Alignment File | Send to Simple_Phylogeny

```
CLUSTAL O(1.2.3) multiple sequence alignment


NM_001024447.1      ------------------------------------------------------------
NM_001200172.1      ------------------------------------------------------------
NM_001114778.1      ATTTCAAAGAAGAAAAAAGAACTGGAGCAGAGCAGTGAAGACGTGAACCAAGTGGAAATA
NM_001113671.1      --AAAGAAAGAAAGAAAGAAAGAAAAGCAGAGCGGTGAAGGCGTGAACCAAGTGAAAATA
NM_214022.1         ----------------------CCCAGAGTGAGGACACCAGGGGACCAGCCAGGAGAGA
NM_001286442.1      ------------------------------------------------------------
NM_173966.3         ------------AACAGAAGCTCCCAGAGCGGGGACACCAGGGGACCAGCCAGGAGAGA
NM_012675.3         ------------------------------------CACCAAGGGACCAGCCAGGAGGGA
NM_001082263.1      -----------------AAGCTCCCTCAGTGAG--GACACGGGCACCAGCTAGGAGGGA
NM_001003244.4      ------------------------------------------------------------
NM_001257261.1      ------------------------------------------------------------A
NM_001285277.1      ------------------------------------------------------------
NM_000594.3         --------------CAGACGCTCCCTCAGCAAGGACAGCAGAGGACCAGCTAAGAGGGA
NM_001045511.1      ------------------------------------------------------------
```

Results for job clustalo-I20161215-140559-0800-79720144-pg

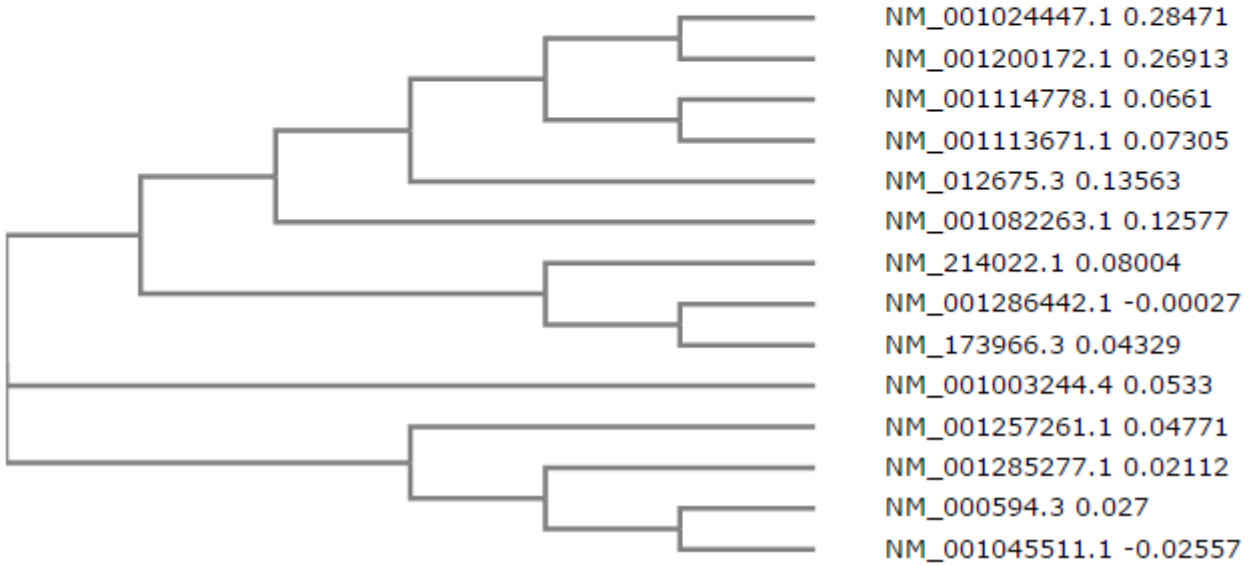
Alignments | Result Summary | Phylogenetic Tree | Submission Details

Phylogenetic Tree

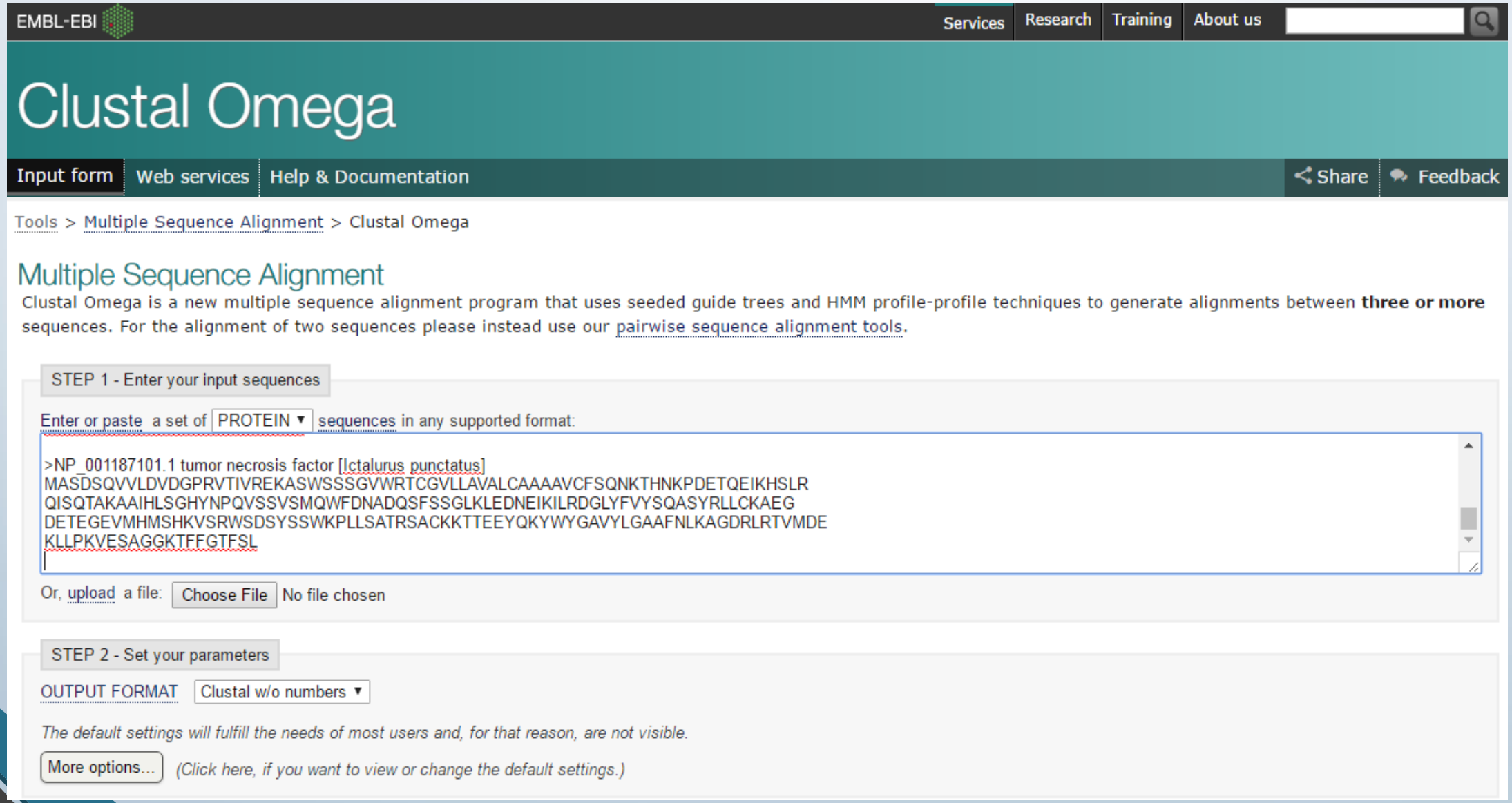
This is a Neighbour-joining tree without distance corrections.

Download Phylogenetic Tree Data

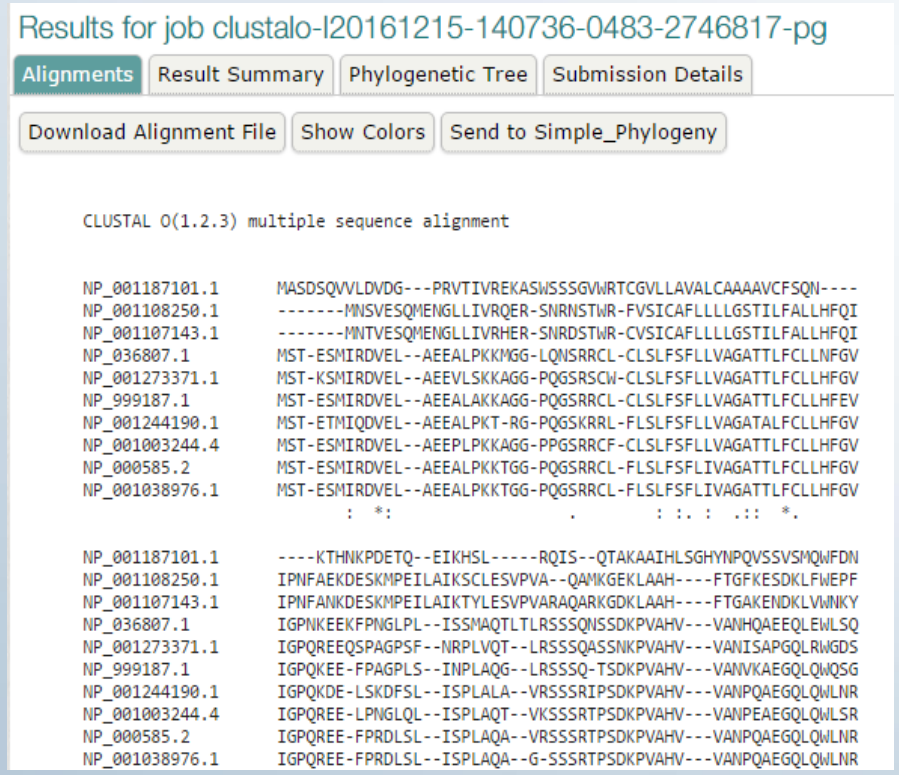
Branch length: ◉ Cladogram ○ Real

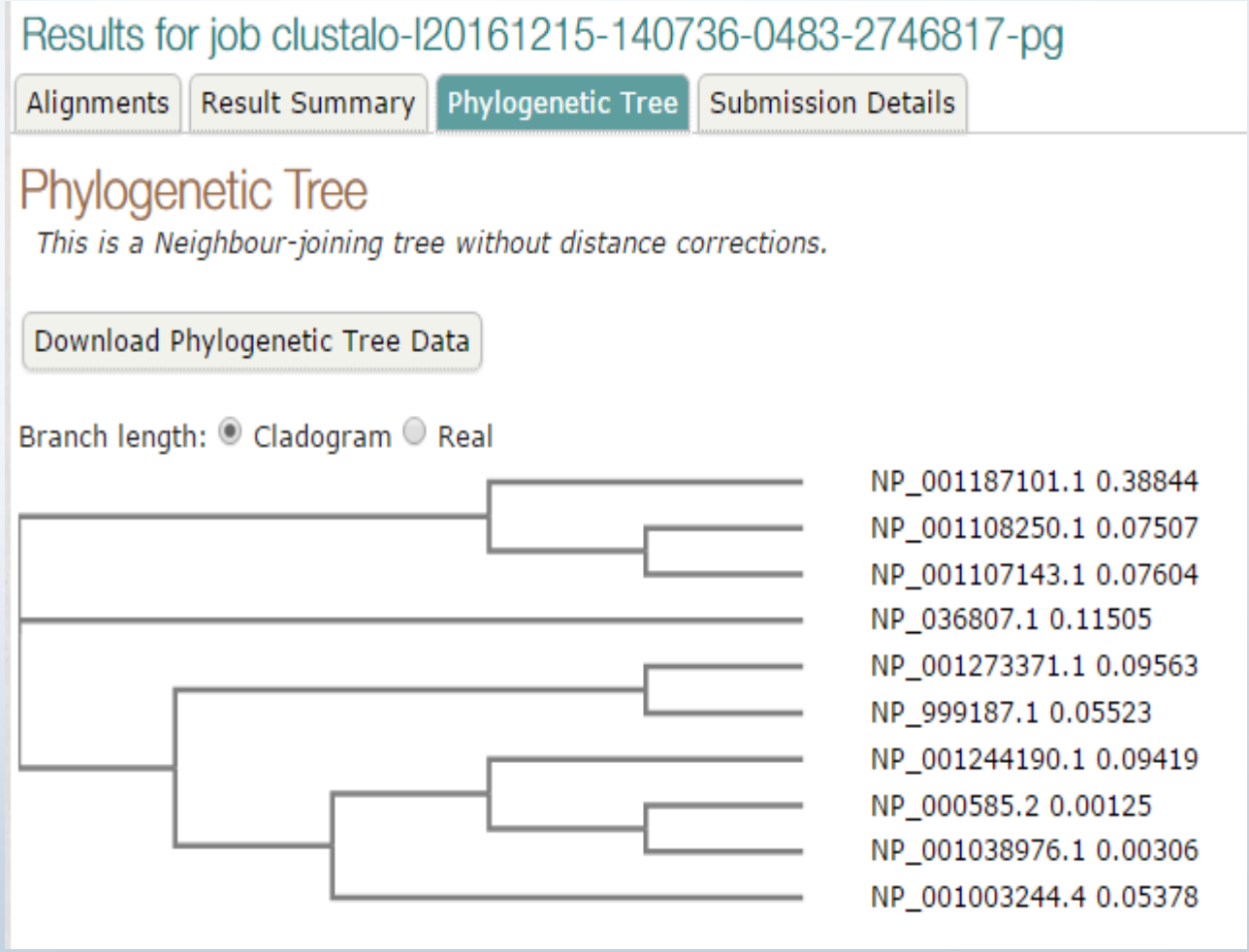


Clustal Omega / Protein

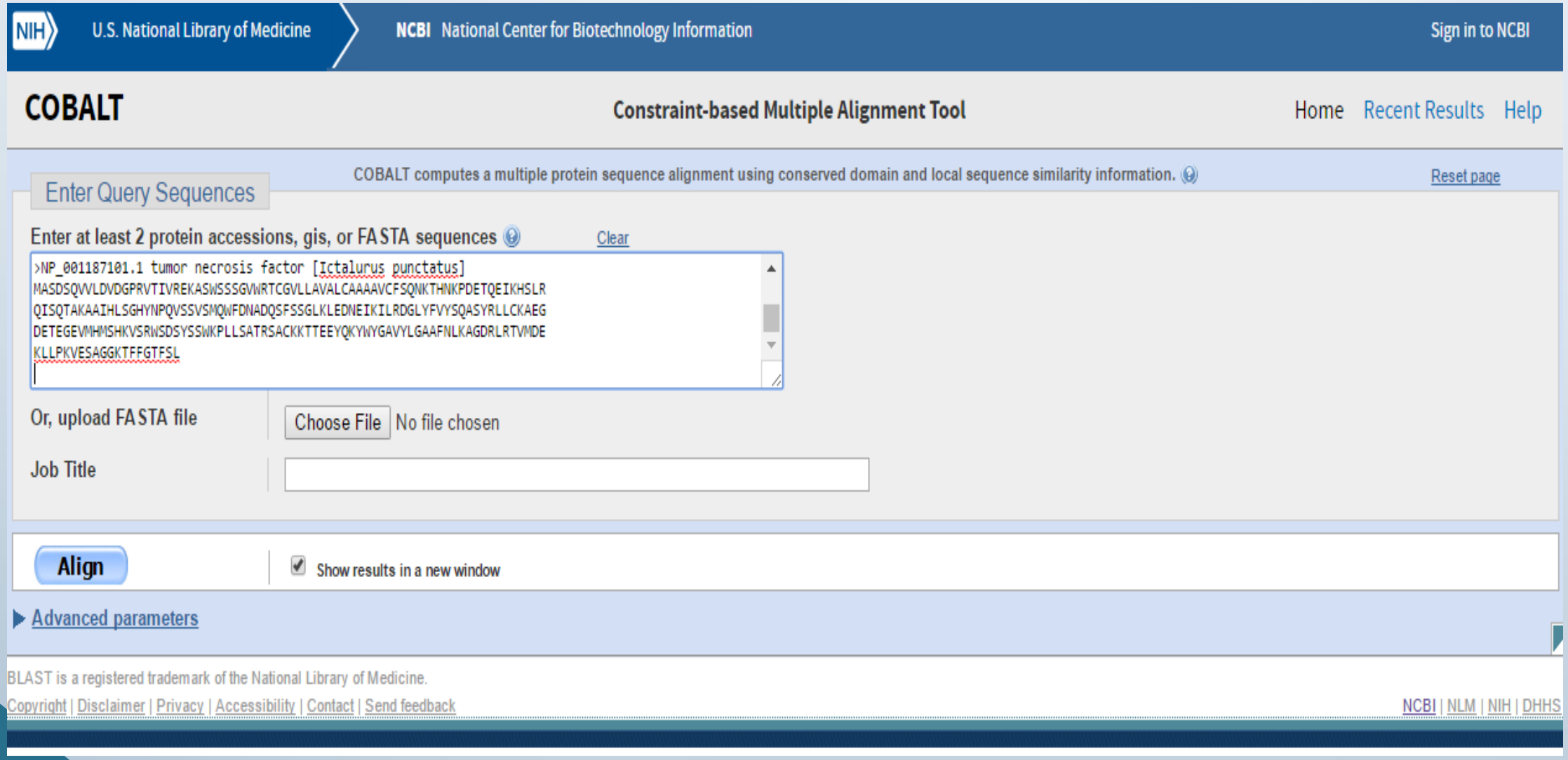


Results


Results for job clustalo-I20161215-140736-0483-2746817-pg
Alignments
Result Summary
Phylogenetic Tree
Submission Details
Download Alignment File
Show Colors
Send to Simple_Phylogeny
CLUSTAL O(1.2.3) multiple sequence alignment
NP_001187101.1 MASDSQVVLDVDG---PRVTIVREKASWSSSGVWRTCGVLLAVALCAAAAVCFSQN----
NP_001108250.1 -------MNSVESQMENGLLIVRQER-SNRNSTWR-FVSICAFLLLLGSTILFALLHFQI
NP_001107143.1 -------MNTVESQMENGLLIVRHER-SNRDSTWR-CVSICAFLLLLGSTILFALLHFQI
NP_036807.1 MST-ESMIRDVEL--AEEALPKKMGG-LQNSRRCL-CLSLFSFLLVAGATTLFCLLNFGV
NP_001273371.1 MST-KSMIRDVEL--AEEVLSKKAGG-PQGSRSCW-CLSLFSFLLVAGATTLFCLLHFGV
NP_999187.1 MST-ESMIRDVEL--AEEALAKKAGG-PQGSRRCL-CLSLFSFLLVAGATTLFCLLHFEV
NP_001244190.1 MST-ETMIQDVEL--AEEALPKT-RG-PQGSKRRL-FLSLFSFLLVAGATALFCLLHFGV
NP_001003244.4 MST-ESMIRDVEL--AEEPLPKKAGG-PPGSRRCF-CLSLFSFLLVAGATTLFCLLHFGV
NP_000585.2 MST-ESMIRDVEL--AEEALPKKTGG-PQGSRRCL-FLSLFSFLIVAGATTLFCLLHFGV
NP_001038976.1 MST-ESMIRDVEL--AEEALPKKTGG-PQGSRRCL-FLSLFSFLIVAGATTLFCLLHFGV
: *: . : :. : .:: *.
NP_001187101.1 ----KTHNKPDETQ--EIKHSL-----RQIS--QTAKAAIHLSGHYNPQVSSVSMQWFDN
NP_001108250.1 IPNFAEKDESKMPEILAIKSCLESVPVA--QAMKGEKLAAH----FTGFKESDKLFWEPF
NP_001107143.1 IPNFANKDESKMPEILAIKTYLESVPVARAQARKGDKLAAH----FTGAKENDKLVWNKY
NP_036807.1 IGPNKEEKFPNGLPL--ISSMAQTLTLRSSSQNSSDKPVAHV---VANHQAEEQLEWLSQ
NP_001273371.1 IGPQREEQSPAGPSF--NRPLVQT--LRSSSQASSNKPVAHV---VANISAPGQLRWGDS
NP_999187.1 IGPQKEE-FPAGPLS--INPLAQG--LRSSSQ-TSDKPVAHV---VANVKAEGQLQWQSG
NP_001244190.1 IGPQKDE-LSKDFSL--ISPLALA--VRSSSRIPSDKPVAHV---VANPQAEGQLQWLNR
NP_001003244.4 IGPQREE-LPNGLQL--ISPLAQT--VKSSSRTPSDKPVAHV---VANPEAEGQLQWLSR
NP_000585.2 IGPQREE-FPRDLSL--ISPLAQA--VRSSSRTPSDKPVAHV---VANPQAEGQLQWLNR
NP_001038976.1 IGPQREE-FPRDLSL--ISPLAQA--G-SSSRTPSDKPVAHV---VANPQAEGQLQWLNR


Results for job clustalo-I20161215-140736-0483-2746817-pg
Alignments
Result Summary
Phylogenetic Tree
Submission Details
Phylogenetic Tree
This is a Neighbour-joining tree without distance corrections.
Download Phylogenetic Tree Data
Branch length: Cladogram Real
NP_001187101.1 0.38844
NP_001108250.1 0.07507
NP_001107143.1 0.07604
NP_036807.1 0.11505
NP_001273371.1 0.09563
NP_999187.1 0.05523
NP_001244190.1 0.09419
NP_000585.2 0.00125
NP_001038976.1 0.00306
NP_001003244.4 0.05378

NCBI



https://www.ncbi.nlm.nih.gov/tools/cobalt/re_cobalt.cgi

Results

Descriptions Select All Re-align ▸Alignment parameters

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

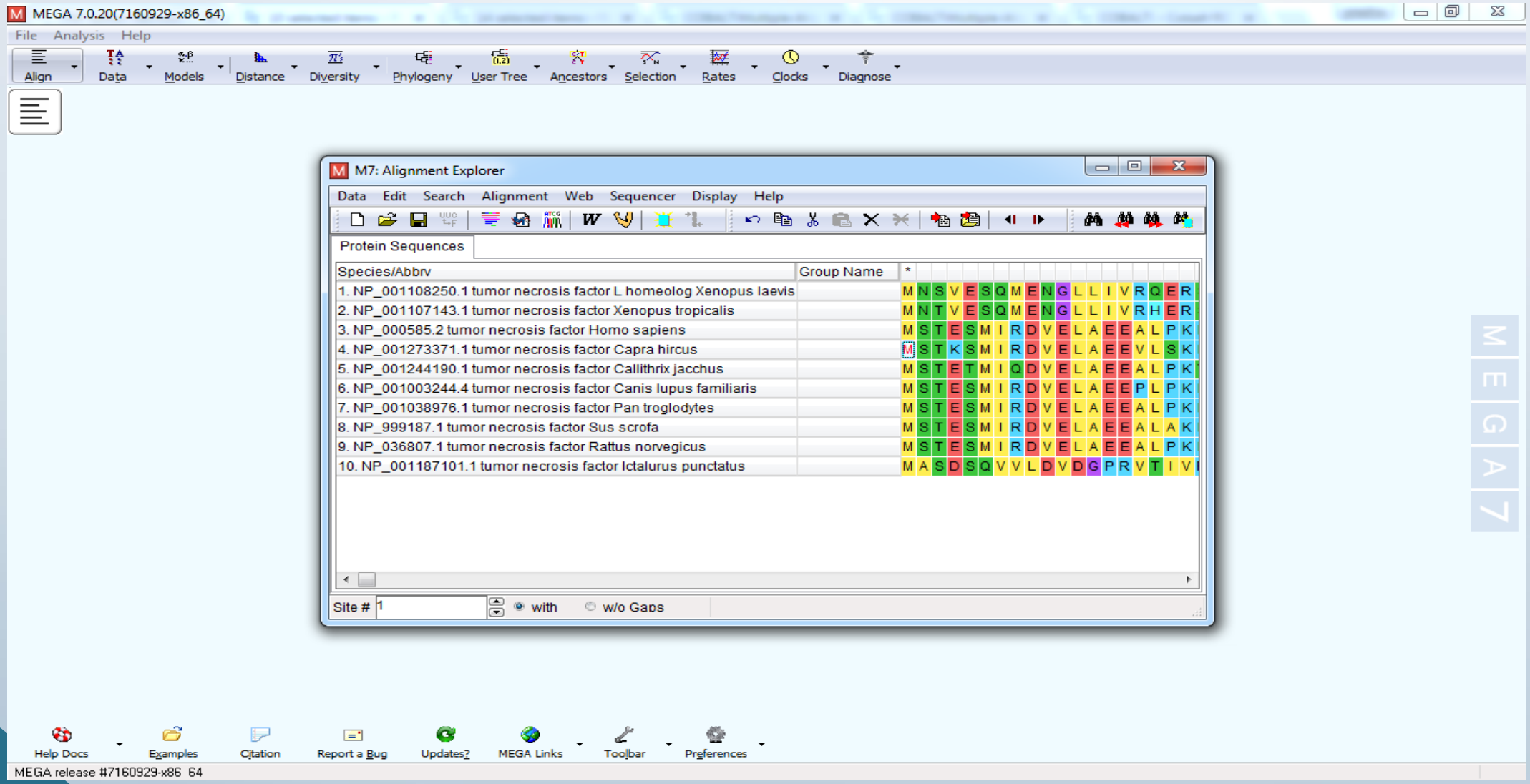
Accession	Description
lcl|Query_10001	NP_001108250.1 tumor necrosis factor L homeolog [Xenopus laevis]
lcl|Query_10002	NP_001107143.1 tumor necrosis factor [Xenopus tropicalis]
lcl|Query_10003	NP_000585.2 tumor necrosis factor [Homo sapiens]
lcl|Query_10004	NP_001273371.1 tumor necrosis factor [Capra hircus]
lcl|Query_10005	NP_001244190.1 tumor necrosis factor [Callithrix jacchus]
lcl|Query_10006	NP_001003244.4 tumor necrosis factor [Canis lupus familiaris]
lcl|Query_10007	NP_001038976.1 tumor necrosis factor [Pan troglodytes]
lcl|Query_10008	NP_999187.1 tumor necrosis factor [Sus scrofa]
lcl|Query_10009	NP_036807.1 tumor necrosis factor [Rattus norvegicus]
lcl|Query_10010	NP_001187101.1 tumor necrosis factor [Ictalurus punctatus]

Alignments Select All Re-align Mouse over the sequence identifer for sequence title

View Format: Compact Conservation Setting: 2 Bits

```
Query_10001   1   MNSvESQMENGLL----IVRQERSNRNSTWR--FVSICAFLLLLGSTILFALLHFQIIpnfAEKDESKMPEILAIKSCLE   74
Query_10002   1   MNTvESQMENGLL----IVRHERSNRDSTWR--CVSICAFLLLLGSTILFALLHFQIIpnfANKDESKMPEILAIKTYLE   74
Query_10003   1   MST-ESMIRDVELAE-EALPKKTGGPQGSRRCLFLSLFSFLIVAGATTLFCLLHFGVI---GPQREE-FPRDLSLISPLA   74
Query_10004   1   MST-KSMIRDVELAE-EVLSKKAGGPQGSRSCWCLSLFSFLLVAGATTLFCLLHFGVI---GPQREEQSPAGPSFNRPLV   75
Query_10005   1   MST-ETMIQDVELAE-EALPK-TRGPQGSKRRLFLSLFSFLLVAGATALFCLLHFGVI---GPQKDE-LSKDFSLISPLA   73
Query_10006   1   MST-ESMIRDVELAE-EPLPKKAGGPPGSRRCFCLSLFSFLLVAGATTLFCLLHFGVI---GPQREE-LPNGLQLISPLA   74
Query_10007   1   MST-ESMIRDVELAE-EALPKKTGGPQGSRRCLFLSLFSFLIVAGATTLFCLLHFGVI---GPQREE-FPRDLSLISPLA   74
Query_10008   1   MST-ESMIRDVELAE-EALAKKAGGPQGSRRCLCLSLFSFLLVAGATTLFCLLHFEVI---GPQKEE-FPAGPLSINPLA   74
Query_10009   1   MST-ESMIRDVELAE-EALPKKMGGLQNSRRCLCLSLFSFLLVAGATTLFCLLNFGVI---GPNKEEKFPNGLPLISSMA   75
```

MEGA Software



Results

M7: Pairwise Distances (C:\Users\yazdan\AppData\Local\Temp\PhyloAnalysis.m...

File Display Average Caption Help

	1	2	3	4	5
1. NP 001108250.1 tumor necrosis factor L homeolog Xenopus laevis					
2. NP 001107143.1 tumor necrosis factor Xenopus tropicalis	0.182				
3. NP 000585.2 tumor necrosis factor Homo sapiens	1.224	1.224			
4. NP 001273371.1 tumor necrosis factor Capra hircus	1.175	1.159	0.231		
5. NP 001244190.1 tumor necrosis factor Callithrix jacchus	1.207	1.191	0.125	0.321	
6. NP 001003244.4 tumor necrosis factor Canis lupus familiaris	1.241	1.207	0.092	0.224	0.1
7. NP 001038976.1 tumor necrosis factor Pan troglodytes	1.241	1.241	0.005	0.231	0.1
8. NP 999187.1 tumor necrosis factor Sus scrofa	1.191	1.159	0.148	0.171	0.2
9. NP 036807.1 tumor necrosis factor Rattus norvegicus	1.224	1.159	0.243	0.328	0.3
10. NP 001187101.1 tumor necrosis factor Ictalurus punctatus	1.447	1.386	1.367	1.329	1.3

[1,1] (NP 001108250.1 tumor necrosis factor L homeolog Xenopus laevis-NP 001108250.1 tun

M7: Tree Explorer: PhyloAnalysis.meg

File Search Image Subtree View Compute Caption Help