CLUSTAL W Tool (Practical)
Value added Course
Lecture 8

Er. Brijendra Singh

Assistant Professor, Department of Bioinformatics

UIET, CSJMU University, Kanpur

Sequence data mining using CLUSTAL W

- Sequential pattern mining is a topic of data mining concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence. It is usually presumed that the values are discrete, and thus time series mining is closely related, but usually considered a different activity.

- Biological sequences define the sequences of nucleotides or amino acids. Biological sequence analysis compares, aligns, indexes, and study biological sequences and therefore plays an essential role in bioinformatics and current biology.

- Sequence alignment depends on the fact that all living organisms are associated by development. This indicate that the nucleotide (DNA, RNA) and protein sequences of species that are nearer to each other in evolution must exhibit higher similarities. An alignment is the procedure of lining up sequences to obtain a maximal identity level, which also defines the degree of similarity among sequences.

Conti..

- ClustalW is a tool for aligning multiple protein or nucleotide sequences. The alignment is achieved via three steps: pairwise alignment, guide-tree generation and progressive alignment.
- Clustal W is a general purpose multiple alignment program for DNA or proteins. The sensitivity of the commonly used progressive multiple sequence alignment method has been greatly improved for the alignment of divergent protein sequences.
- The objective of Clustal W to align three or more sequences to find out structural and functional relationship between these sequences.
- **Conserved regions:** In biology, during the evolutionary time there may be some regions called group of bases or a sequence of nucleotides preserved as such in DNA, those sequences or a region, if seen in next generations called as Conserved regions.
- **Consensus Sequence:** In a Nucleotide or an amino acid sequence, each base pair (an amino acid or a nucleotide) may occur more frequently at a particular region in different sequences of nature.

Summary of MSA Programs

Summary of MSA programs that we consider to be the best currently available

Program	Advantages	Cautions
CLUSTALW	Uses less memory than other programs	Less accurate or scalable than modern programs
DIALIGN	Attempts to distinguish between alignable and non-alignable regions	Less accurate than CLUSTALW on global benchmarks
MAFFT, MUSCLE	Faster and more accurate than CLUSTALW; good trade-off of accuracy and computational cost. Options to run even faster, with lower average accuracy, for high-throughput applications.	For very large data sets (say, more than 1000 sequences) select time- and memory-saving options
PROBCONS	Highest accuracy score on several benchmarks	Computation time and memory usage is a limiting factor for large alignment problems (>100 sequences)
ProDA	Does not assume global alignability; allows repeated, shuffled and absent domains.	High computational cost and less accurate than CLUSTALW on global benchmarks
T-COFFEE	High accuracy and the ability to incorporate heterogeneous types of information	Computation time and memory usage is a limiting factor for large alignment problems (>100 sequences)

CLUSTALW / CLUSTAL Omega

- Pair wise sequence alignment has been approached with dynamic programming between nucleotide or amino acid sequences. The same approach can be used for alignment of ‘n’ number of sequences. But this program is limited to pair wise, since there will be exponential increase in memory, number of steps with respect to number of sequences. Because of such limitations with dynamic programming, researchers came up with an approach called *'progressive method'* to align three or more sequences.
- Progressive method was first suggested by Feng and Doolittle in 1987. It compares only a pair of sequences together at a time using the following steps:
- Using the standard dynamic programming algorithm on each pair, we can calculate the (N*(N-1))/2 (N is total number of sequences) distances between the sequence pairs.
- From the distance matrix obtained using the clustering algorithm, construct a guide tree.
- From the tree obtained, align the first node to the second node. After fixing the alignment, add another sequence or the third node. Iterate the step until all the sequences are aligned. When a sequence is aligned to a group or when there is alignment in between the two groups of sequences, the alignment is performed that had the highest alignment score. The gap symbols in the alignment replaced with a neutral character. Where it helps to guide the alignment of sequence- alignment and alignment –alignment.

Conti…

- CLUSTALW uses the progressive algorithm, by adding the sequence one by one until all the sequences are completely aligned.
- Steps for CLUSTAL algorithm

i) Calculate all possible pairwise alignments, record the score for each pair.

ii) Calculate a guide tree based on the pairwise distances (algorithm: Neighbor Joining).

iii) Find the two most closely related sequences

iv) Align the sequences by progressive method

a) Calculate a consensus of this alignment.

b) Replace the two sequences with the consensus.

c) Find the two next-most closely related sequences.

d) Iterate until all sequences have been aligned

v) Expand the consensus sequences with the (gapped) original sequences

vi) Report the multiple sequence alignment

What are the types of mining sec | schdule - brijendra.bioinfo@gma | ClustalW - Google Search

google.com/search?q=ClustalW&rlz=1C1CHWL_enIN963IN963&biw=1280&bih=913&sxsrf=ALiCzsYzD5K0glXoASLCtbGqDUTJcvtPcQ%3A1655...

Google ClustalW

All Books Images Videos News More Tools

About 36,50,000 results (0.38 seconds)

https://www.genome.jp › tools-bin › clustalw

Multiple Sequence Alignment - CLUSTALW - GenomeNet

Multiple Sequence Alignment by **CLUSTALW**. ETE3, MAFFT, **CLUSTALW**, PRRN. Help. General Setting Parameters: Output Format: CLUSTAL, GCG(MSF), GDE, PIR, Phylip ...
Phylogenetic analysis pipeline... · Mafft · Help

https://www.ebi.ac.uk › Tools › msa › clustalo

Clustal Omega < Multiple Sequence Alignment < EMBL-EBI

Clustal Omega is a multiple sequence alignment program. It produces biologically meaningful multiple sequence alignments of divergent sequences.
Pairwise Sequence · Help & Documentation · Bioinformatics Tools FAQ
You've visited this page 2 times. Last visit: 18/6/22

https://www.ebi.ac.uk › Tools › msa › clustalw2

ClustalW2 < Multiple Sequence Alignment < EMBL-EBI

ClustalW2 is a general purpose multiple sequence alignment program for DNA or proteins. It attempts to calculate the best match for the selected sequences, ...

People also ask

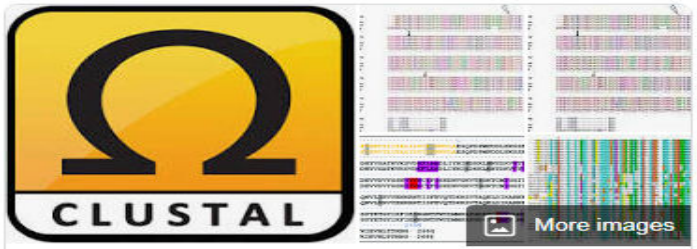
What is ClustalW used for?

What is ClustalW algorithm?

What is ClustalW format?

What is a Clustal alignment?

Feedback



Clustal

Computer program

Clustal is a series of widely used computer programs used in bioinformatics for multiple sequence alignment. There have been many versions of Clustal over the development of the algorithm that are listed below. The analysis of each tool and its algorithm are also detailed in their respective categories. Wikipedia

Developer(s): Des Higgins; Fabian Sievers; David Dineen; Andreas Wilm (all at the Conway Institute, UCD)

Operating system: UNIX, Linux, MacOS, MS-Windows, FreeBSD, Debian

Stable release: 1.2.2 / 1 July 2016; 5 years ago

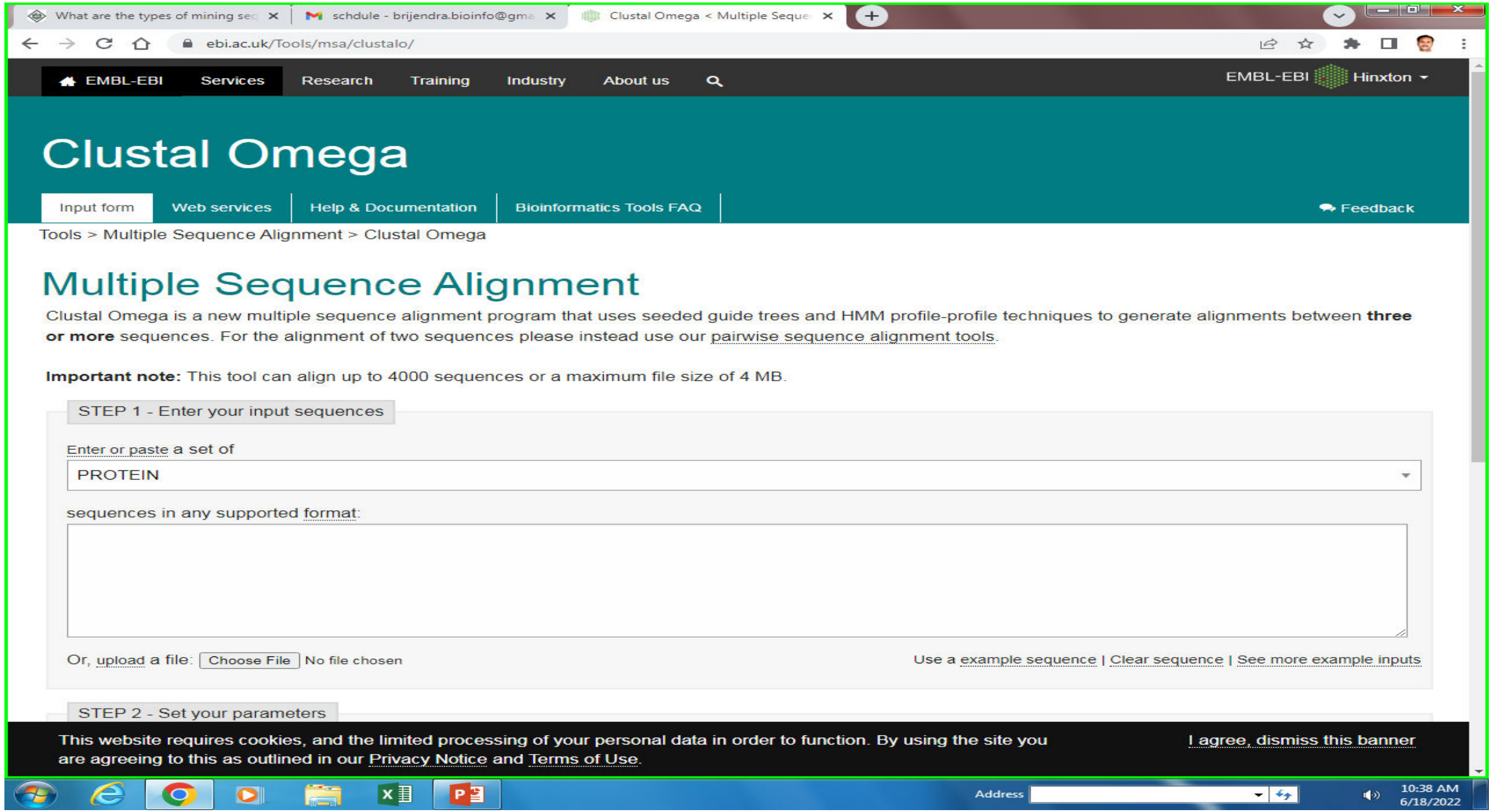
Written in: C++

Purpose

Command line

Address

10:37 AM
6/18/2022


What are the types of mining sec
schdule - brijendra.bioinfo@gma
Clustal Omega < Multiple Seque
ebi.ac.uk/Tools/msa/clustalo/
EMBL-EBI
Services
Research
Training
Industry
About us
EMBL-EBI
Hinxton
Clustal Omega
Input form
Web services
Help & Documentation
Bioinformatics Tools FAQ
Feedback
Tools > Multiple Sequence Alignment > Clustal Omega
Multiple Sequence Alignment
Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.
Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB.
STEP 1 - Enter your input sequences
Enter or paste a set of
PROTEIN
sequences in any supported format:
Or, upload a file:
Choose File
No file chosen
Use a example sequence | Clear sequence | See more example inputs
STEP 2 - Set your parameters
This website requires cookies, and the limited processing of your personal data in order to function. By using the site you are agreeing to this as outlined in our Privacy Notice and Terms of Use.
I agree, dismiss this banner
Address
10:38 AM
6/18/2022

An official website of the United States government Here's how you know

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

All Databases | Search

- NCBI Home
- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News & Blog

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

Popular Resources
- PubMed
- Bookshelf
- PubMed Central
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI News & Blog

Save the Date: NCBI at the Bioinformatics Open Science Conference (BOSC), July 2022
16 Jun 2022

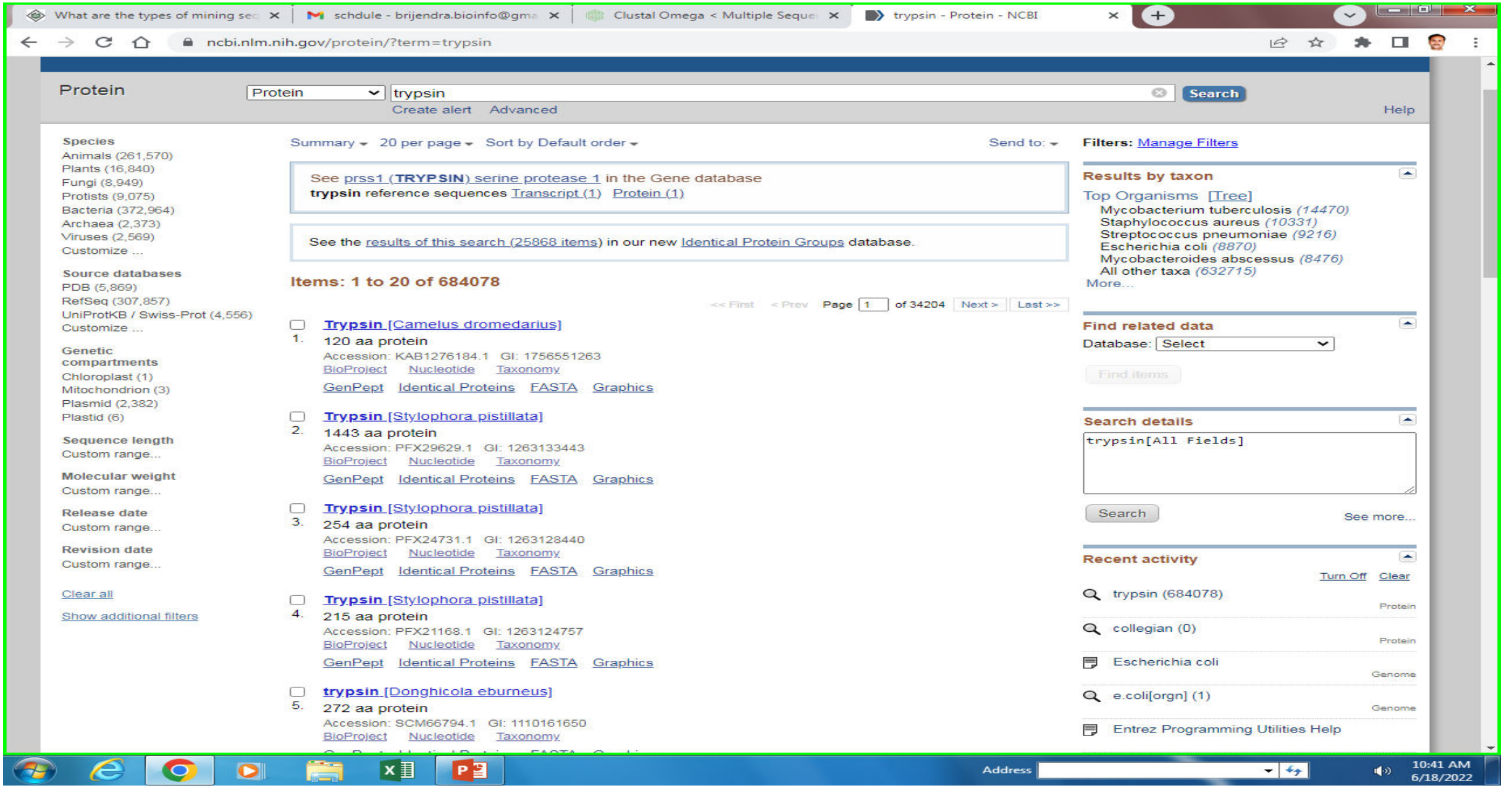
June 15 Webinar: What's new with NCBI Virus?
06 Jun 2022

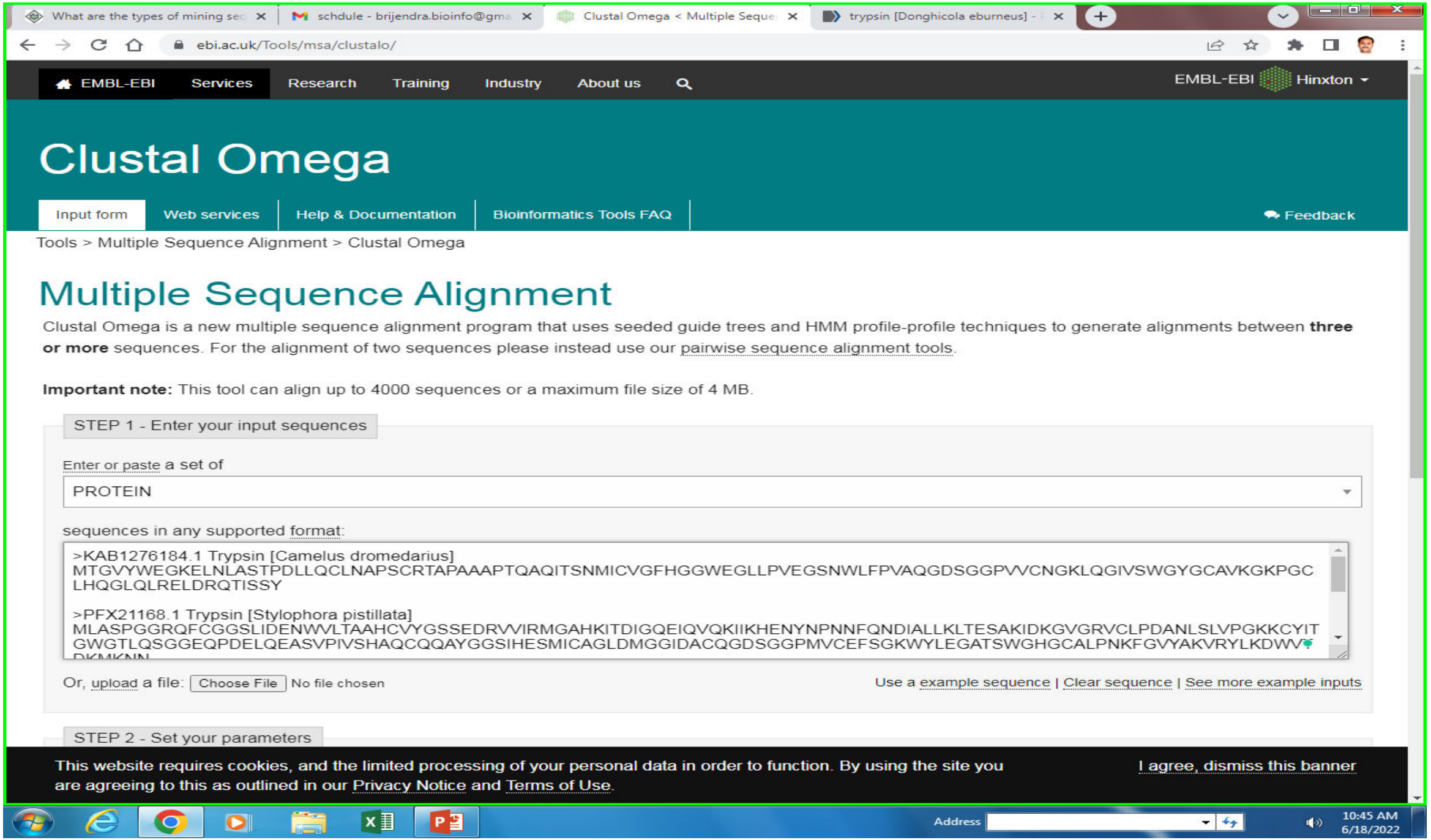
Join us on June 15 , 2022 at 12PM US

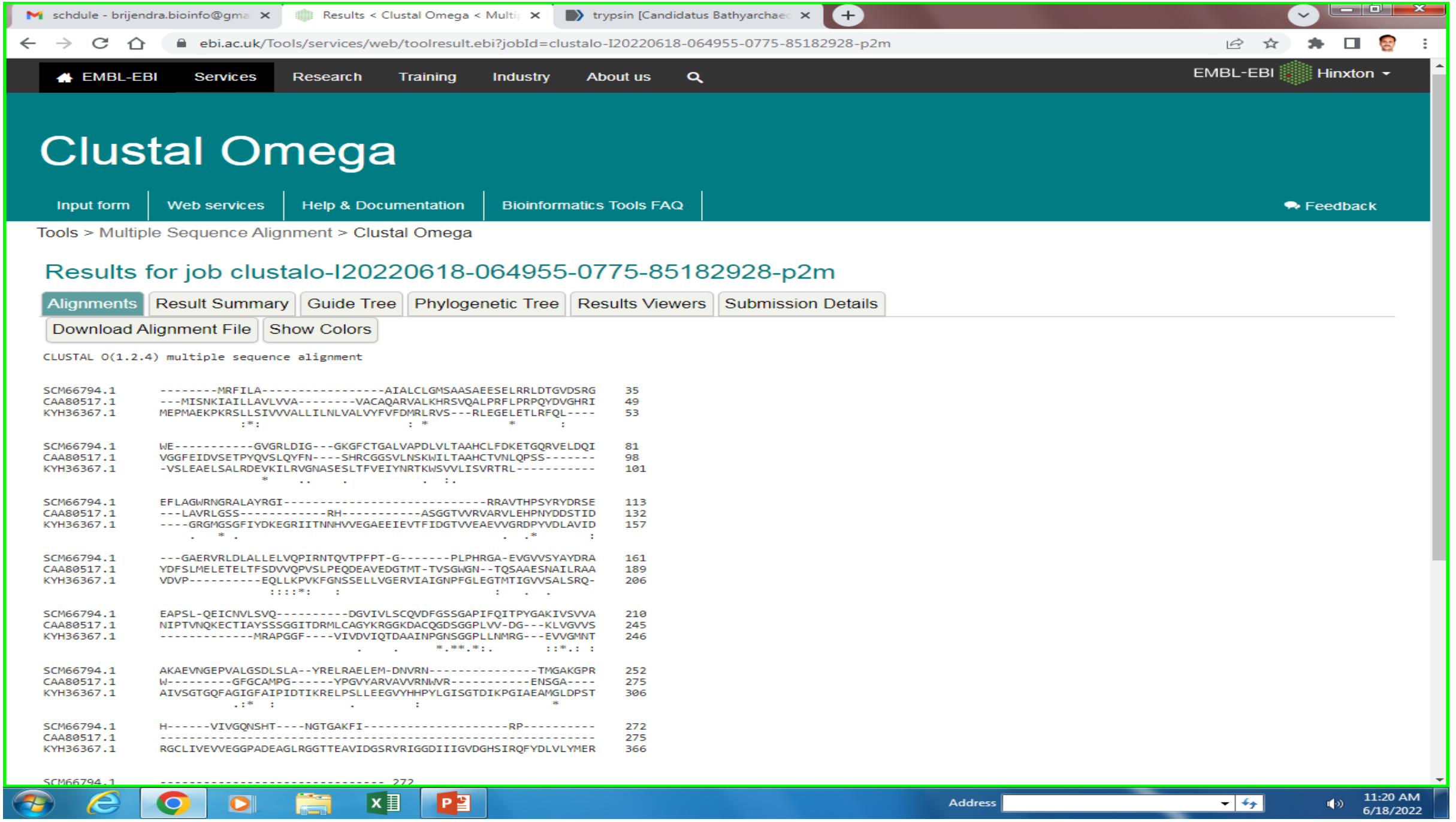
Come see NCBI at the ASM Microbe Conference 2022
31 May 2022

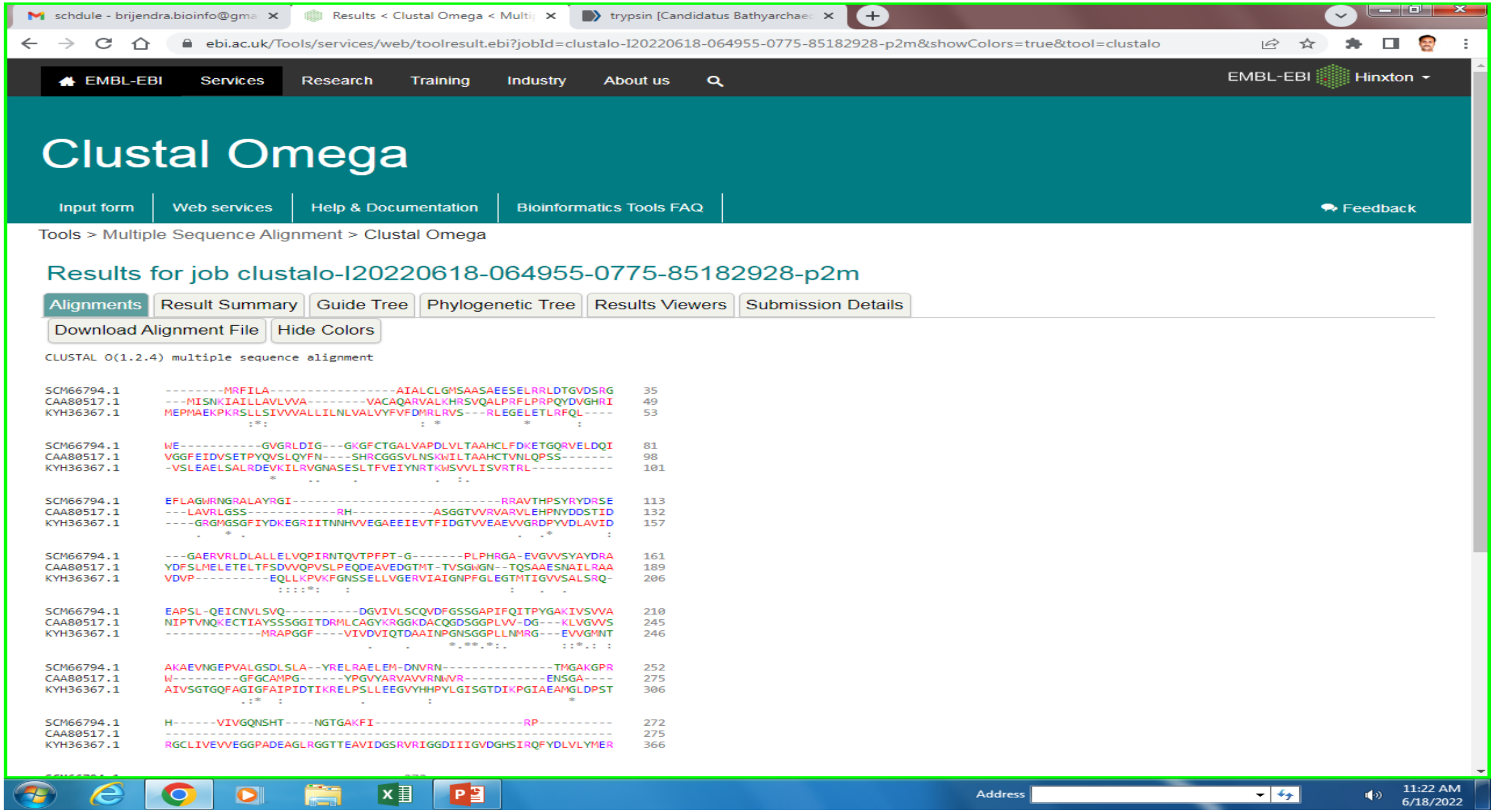
The American Society of Microbiology

More...


What are the types of mining sec
schdule - brijendra.bioinfo@gma
Clustal Omega < Multiple Seque
trypsin - Protein - NCBI
ncbi.nlm.nih.gov/protein/?term=trypsin
Protein
Protein
trypsin
Search
Create alert
Advanced
Help
Species
Animals (261,570)
Plants (16,840)
Fungi (8,949)
Protists (9,075)
Bacteria (372,964)
Archaea (2,373)
Viruses (2,569)
Customize ...
Source databases
PDB (5,869)
RefSeq (307,857)
UniProtKB / Swiss-Prot (4,556)
Customize ...
Genetic compartments
Chloroplast (1)
Mitochondrion (3)
Plasmid (2,382)
Plastid (6)
Sequence length
Custom range...
Molecular weight
Custom range...
Release date
Custom range...
Revision date
Custom range...
Clear all
Show additional filters
Summary
20 per page
Sort by Default order
Send to:
See prss1 (TRYPSIN) serine protease 1 in the Gene database
trypsin reference sequences Transcript (1) Protein (1)
See the results of this search (25868 items) in our new Identical Protein Groups database.
Items: 1 to 20 of 684078
<< First
< Prev
Page 1 of 34204
Next >
Last >>
1. Trypsin [Camelus dromedarius]
120 aa protein
Accession: KAB1276184.1 GI: 1756551263
BioProject Nucleotide Taxonomy
GenPept Identical Proteins FASTA Graphics
2. Trypsin [Stylophora pistillata]
1443 aa protein
Accession: PFX29629.1 GI: 1263133443
BioProject Nucleotide Taxonomy
GenPept Identical Proteins FASTA Graphics
3. Trypsin [Stylophora pistillata]
254 aa protein
Accession: PFX24731.1 GI: 1263128440
BioProject Nucleotide Taxonomy
GenPept Identical Proteins FASTA Graphics
4. Trypsin [Stylophora pistillata]
215 aa protein
Accession: PFX21168.1 GI: 1263124757
BioProject Nucleotide Taxonomy
GenPept Identical Proteins FASTA Graphics
5. trypsin [Donghicola eburneus]
272 aa protein
Accession: SCM66794.1 GI: 1110161650
BioProject Nucleotide Taxonomy
Filters: Manage Filters
Results by taxon
Top Organisms [Tree]
Mycobacterium tuberculosis (14470)
Staphylococcus aureus (10331)
Streptococcus pneumoniae (9216)
Escherichia coli (8870)
Mycobacteroides abscessus (8476)
All other taxa (632715)
More...
Find related data
Database: Select
Find items
Search details
trypsin[All Fields]
Search
See more...
Recent activity
Turn Off Clear
trypsin (684078) Protein
collegian (0) Protein
Escherichia coli Genome
e.coli[orgn] (1) Genome
Entrez Programming Utilities Help
Address
10:41 AM
6/18/2022


What are the types of mining sec
schdule - brijendra.bioinfo@gma
Clustal Omega < Multiple Seque
trypsin [Donghicola eburneus] -
ebi.ac.uk/Tools/msa/clustalo/
EMBL-EBI
Services
Research
Training
Industry
About us
EMBL-EBI Hinxton
Clustal Omega
Input form
Web services
Help & Documentation
Bioinformatics Tools FAQ
Feedback
Tools > Multiple Sequence Alignment > Clustal Omega
Multiple Sequence Alignment
Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.
Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB.
STEP 1 - Enter your input sequences
Enter or paste a set of
PROTEIN
sequences in any supported format:
>KAB1276184.1 Trypsin [Camelus dromedarius]
MTGVYWEGKELNLASTPDLLQCLNAPSCRTAPAAAPTQAQITSNMICVGFHGGWEGLLPVEGSNWLFPVAQGDSGGPVVCNGKLQGIVSWGYGCAVKGKPGC
LHQGLQLRELDRQTISSY
>PFX21168.1 Trypsin [Stylophora pistillata]
MLASPGGRQFCGGSLIDENWVLTAAHCVYGSSEDRVVIRMGAHKITDIGQEIQVQKIIKHENYNPNNFQNDIALLKLTESAKIDKGVGRVCLPDANLSLVPGKKCYIT
GWGTLQSGGEQPDELQEASVPIVSHAQCQQAYGGSIHESMICAGLDMGGIDACQGDSGGPMVCEFSGKWYLEGATSWGHGCALPNKFGVYAKVRYLKDWV
Or, upload a file: Choose File No file chosen
Use a example sequence | Clear sequence | See more example inputs
STEP 2 - Set your parameters
This website requires cookies, and the limited processing of your personal data in order to function. By using the site you are agreeing to this as outlined in our Privacy Notice and Terms of Use.
I agree, dismiss this banner
Address
10:45 AM
6/18/2022


schdule - brijendra.bioinfo@gma
Results < Clustal Omega < Multi
trypsin [Candidatus Bathyarchaec
ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=clustalo-I20220618-064955-0775-85182928-p2m
EMBL-EBI
Services
Research
Training
Industry
About us
EMBL-EBI Hinxton
Clustal Omega
Input form
Web services
Help & Documentation
Bioinformatics Tools FAQ
Feedback
Tools > Multiple Sequence Alignment > Clustal Omega
Results for job clustalo-I20220618-064955-0775-85182928-p2m
Alignments
Result Summary
Guide Tree
Phylogenetic Tree
Results Viewers
Submission Details
Download Alignment File
Show Colors
CLUSTAL O(1.2.4) multiple sequence alignment
SCM66794.1 --------MRFILA-----------------AIALCLGMSAASAEESELRRLDTGVDSRG 35
CAA80517.1 ---MISNKIAILLAVLVVA--------VACAQARVALKHRSVQALPRFLPRPQYDVGHRI 49
KYH36367.1 MEPMAEKPKRSLLSIVVVALLILNLVALVYFVFDMRLRVS---RLEGELETLRFQL---- 53
SCM66794.1 WE-----------GVGRLDIG---GKGFCTGALVAPDLVLTAAHCLFDKETGQRVELDQI 81
CAA80517.1 VGGFEIDVSETPYQVSLQYFN----SHRCGGSVLNSKWILTAAHCTVNLQPSS------- 98
KYH36367.1 -VSLEAELSALRDEVKILRVGNASESLTFVEIYNRTKWSVVLISVRTRL----------- 101
SCM66794.1 EFLAGWRNGRALAYRGI----------RRAVTHPSYRYDRSE 113
CAA80517.1 ---LAVRLGSS----------RH----------ASGGTVVRVARVLEHPNYDDSTID 132
KYH36367.1 ----GRGMGSGFIYDKEGRIITNNHVVEGAEEIEVTFIDGTVVEAEVVGRDPYVDLAVID 157
SCM66794.1 ---GAERVRLDLALLELVQPIRNTQVTPFPT-G-------PLPHRGA-EVGVVSYAYDRA 161
CAA80517.1 YDFSLMELETELTFSDVVQPVSLPEQDEAVEDGTMT-TVSGWGN--TQSAAESNAILRAA 189
KYH36367.1 VDVP----------EQLLKPVKFGNSSELLVGERVIAIGNPFGLEGTMTIGVVSALSRQ- 206
SCM66794.1 EAPSL-QEICNVLSVQ----------DGVIVLSCQVDFGSSGAPIFQITPYGAKIVSVVA 210
CAA80517.1 NIPTVNQKECTIAYSSSGGITDRMLCAGYKRGGKDACQGDSGGPLVV-DG---KLVGVVS 245
KYH36367.1 ------------MRAPGGF----VIVDVIQTDAAINPGNSGGPLLNMRG---EVVGMNT 246
SCM66794.1 AKAEVNGEPVALGSDLSLA--YRELRAELEM-DNVRN---------------TMGAKGPR 252
CAA80517.1 W---------GFGCAMPG------YPGVYARVAVVRNWVR-----------ENSGA---- 275
KYH36367.1 AIVSGTGQFAGIGFAIPIDTIKRELPSLLEEGVYHHPYLGISGTDIKPGIAEAMGLDPST 306
SCM66794.1 H------VIVGQNSHT----NGTGAKFI--------------------RP---------- 272
CAA80517.1 ---------- 275
KYH36367.1 RGCLIVEVVEGGPADEAGLRGGTTEAVIDGSRVRIGGDIIIGVDGHSIRQFYDLVLYMER 366
SCM66794.1 ---------- 272
Address
11:20 AM
6/18/2022


schdule - brijendra.bioinfo@gma
Results < Clustal Omega < Multi
trypsin [Candidatus Bathyarchae
ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=clustalo-I20220618-064955-0775-85182928-p2m&showColors=true&tool=clustalo
EMBL-EBI
Services
Research
Training
Industry
About us
EMBL-EBI Hinxton
Clustal Omega
Input form
Web services
Help & Documentation
Bioinformatics Tools FAQ
Feedback
Tools > Multiple Sequence Alignment > Clustal Omega
Results for job clustalo-I20220618-064955-0775-85182928-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
SCM66794.1 --------MRFILA-----------------AIALCLGMSAASAEESELRRLDTGVDSRG 35
CAA80517.1 ---MISNKIAILLAVLVVA--------VACAQARVALKHRSVQALPRFLPRPQYDVGHRI 49
KYH36367.1 MEPMAEKPKRSLLSIVVVALLILNLVALVYFVFDMRLRVS---RLEGELETLRFQL---- 53
SCM66794.1 WE-----------GVGRLDIG---GKGFCTGALVAPDLVLTAAHCLFDKETGQRVELDQI 81
CAA80517.1 VGGFEIDVSETPYQVSLQYFN----SHRCGGSVLNSKWILTAAHCTVNLQPSS------- 98
KYH36367.1 -VSLEAELSALRDEVKILRVGNASESLTFVEIYNRTKWSVVLISVRTRL----------- 101
SCM66794.1 EFLAGWRNGRALAYRGI----------------------------RRAVTHPSYRYDRSE 113
CAA80517.1 ---LAVRLGSS-------------RH-----------ASGGTVVRVARVLEHPNYDDSTID 132
KYH36367.1 ----GRGMGSGFIYDKEGRIITNNHVVEGAEEIEVTFIDGTVVEAEVVGRDPYVDLAVID 157
SCM66794.1 ---GAERVRLDLALLELVQPIRNTQVTPFPT-G-------PLPHRGA-EVGVVSYAYDRA 161
CAA80517.1 YDFSLMELETELTFSDVVQPVSLPEQDEAVEDGTMT-TVSGWGN--TQSAAESNAILRAA 189
KYH36367.1 VDVP----------EQLLKPVKFGNSSELLVGERVIAIGNPFGLEGTMTIGVVSALSRQ- 206
SCM66794.1 EAPSL-QEICNVLSVQ----------DGVIVLSCQVDFGSSGAPIFQITPYGAKIVSVVA 210
CAA80517.1 NIPTVNQKECTIAYSSSGGITDRMLCAGYKRGGKDACQGDSGGPLVV-DG---KLVGVVS 245
KYH36367.1 -------------MRAPGGF----VIVDVIQTDAAINPGNSGGPLLNMRG---EVVGMNT 246
SCM66794.1 AKAEVNGEPVALGSDLSLA--YRELRAELEM-DNVRN---------------TMGAKGPR 252
CAA80517.1 W----------GFGCAMPG------YPGVYARVAVVRNWVR-----------ENSGA---- 275
KYH36367.1 AIVSGTGQFAGIGFAIPIDTIKRELPSLLEEGVYHHPYLGISGTDIKPGIAEAMGLDPST 306
SCM66794.1 H------VIVGQNSHT----NGTGAKFI--------------------RP---------- 272
CAA80517.1 ---------- 275
KYH36367.1 RGCLIVEVVEGGPADEAGLRGGTTEAVIDGSRVRIGGDIIIGVDGHSIRQFYDLVLYMER 366
Address
11:22 AM
6/18/2022

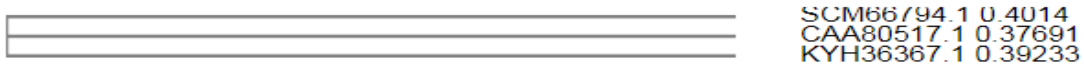
Clustal Omega

Results for job clustalo-I20220618-064955-0775-85182928-p2m

Phylogenetic Tree

This is a Neighbour-joining tree without distance corrections.

Branch length: Cladogram Real



Tree Data

```
(
SCM66794.1:0.40140,
CAA80517.1:0.37691,
KYH36367.1:0.39233);
```

Result Analysis

- The out put of Clustal W you can see that the last lines contains seemingly cabalistic signs such as (*), (:), (.). These three symbols have very precise meanings.
- (*) A star indicates an entirely conserved column.
- (:) A colon indicates columns where all the residues have roughly the same size and the same hydropathy.
- (.) A period indicates columns where the size or the hydropathy has been preserved in the course of evolution.

Clustal Omega

Results for job clustalo-I20220618-064955-0775-85182928-p2m

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

Input Sequences

clustalo-I20220618-064955-0775-85182928-p2m.input

Tool Output

clustalo-I20220618-064955-0775-85182928-p2m.output

Alignment in CLUSTAL format with base/residue numbering

clustalo-I20220618-064955-0775-85182928-p2m.clustal_num

Guide Tree

clustalo-I20220618-064955-0775-85182928-p2m.dnd

Phylogenetic Tree

clustalo-I20220618-064955-0775-85182928-p2m.ph

Percent Identity Matrix

clustalo-I20220618-064955-0775-85182928-p2m.pim

Main Criteria for Building a MSA

1. Structural similarity
2. Evolutionary similarity
3. Functional similarity
4. Sequence similarity

Main Application of Multiple Sequence Alignment

1. Phylogenetic analysis
2. Pattern Identification
3. Extrapolation
4. Domain Identification
5. DNA Regulatory Elements
6. Structure Prediction
7. PCR analysis.

- Reference:-

Bioinformatics “A Beginner’s Guide” by Jean – Michel (Wiley publication).

THANK YOU