BHARATHIDASAN UNIVERSITY

Tiruchirappalli- 620024,
Tamil Nadu, India.

Programme : M.Sc., Biomedical Science

Course Title : Bioinformatics

Course Code : BM35S1BI

Unit-III

TOPIC: MULTIPLE SEQUENCE ALIGNMENT

- CLSTALW
- TCOFFEE

Dr. P. JEGANATHAN
Guest Lecturer
Department of Biomedical Science

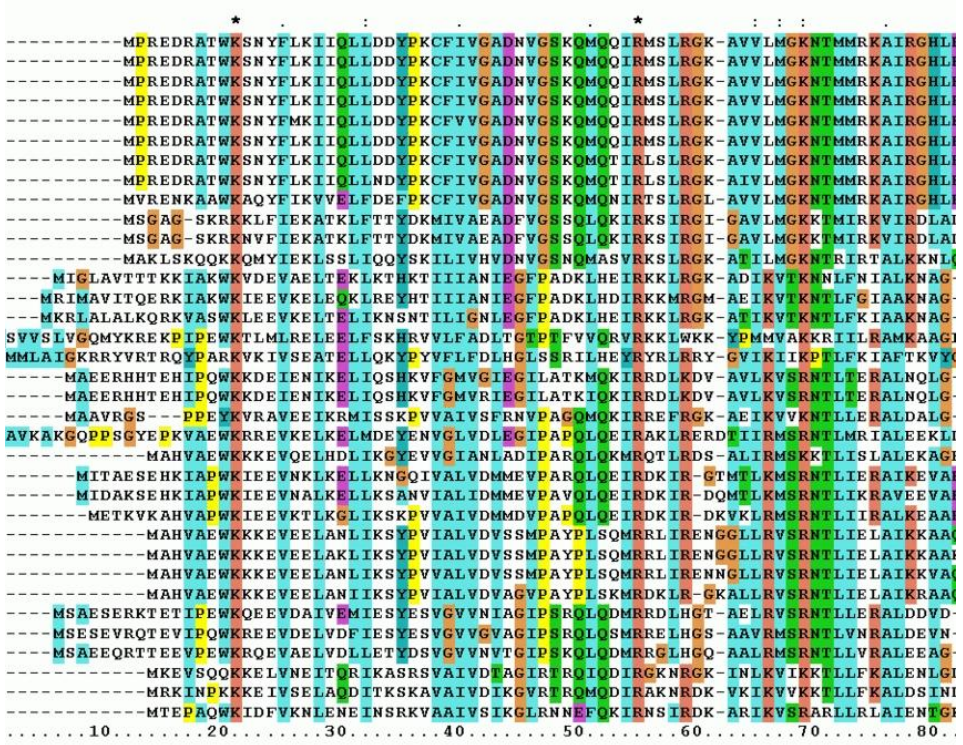
MULTIPLE SEQUENCE ALIGNMENT

- CLSTALW
- TCOFFEE

,

MULTIPLE SEQUENCE ALIGNMENT

- In bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA or protein to identify regions of similarity
- The reason to perform sequence alignment, is to find out the regions of similarity, which refers to functional equivalence and evolutionary relationship between sequences.
- Now, Multiple sequence alignment is basically an alignment of more than two sequences



Approaches

To carry out MSA , with dynamic programming, multidimensional matrix is needed- the amount of computing time and computer memory it requires increases exponentially as the number of sequence increases. So it cannot be applied for more than 10 sequences.

Hence , Heuristic approaches which engage pairwise dynamic programming algorithms are most often used.

It falls into three categories:

1. Progressive alignment type(ex- Clustal tool)
2. Iterative alignment type
3. Block- based alignment type

STEPS INVOLVED IN MULTIPLE SEQUENCE ALIGNMENT

1. Align the new sequence to each of the previous sequence
2. Create a distance matrix / function for each sequence pair
3. Create a phylogenetic guide tree from the matrices placing the sequences at the terminal nodes.
4. Use the guide tree to determine the next sequence to be added to the alignment.

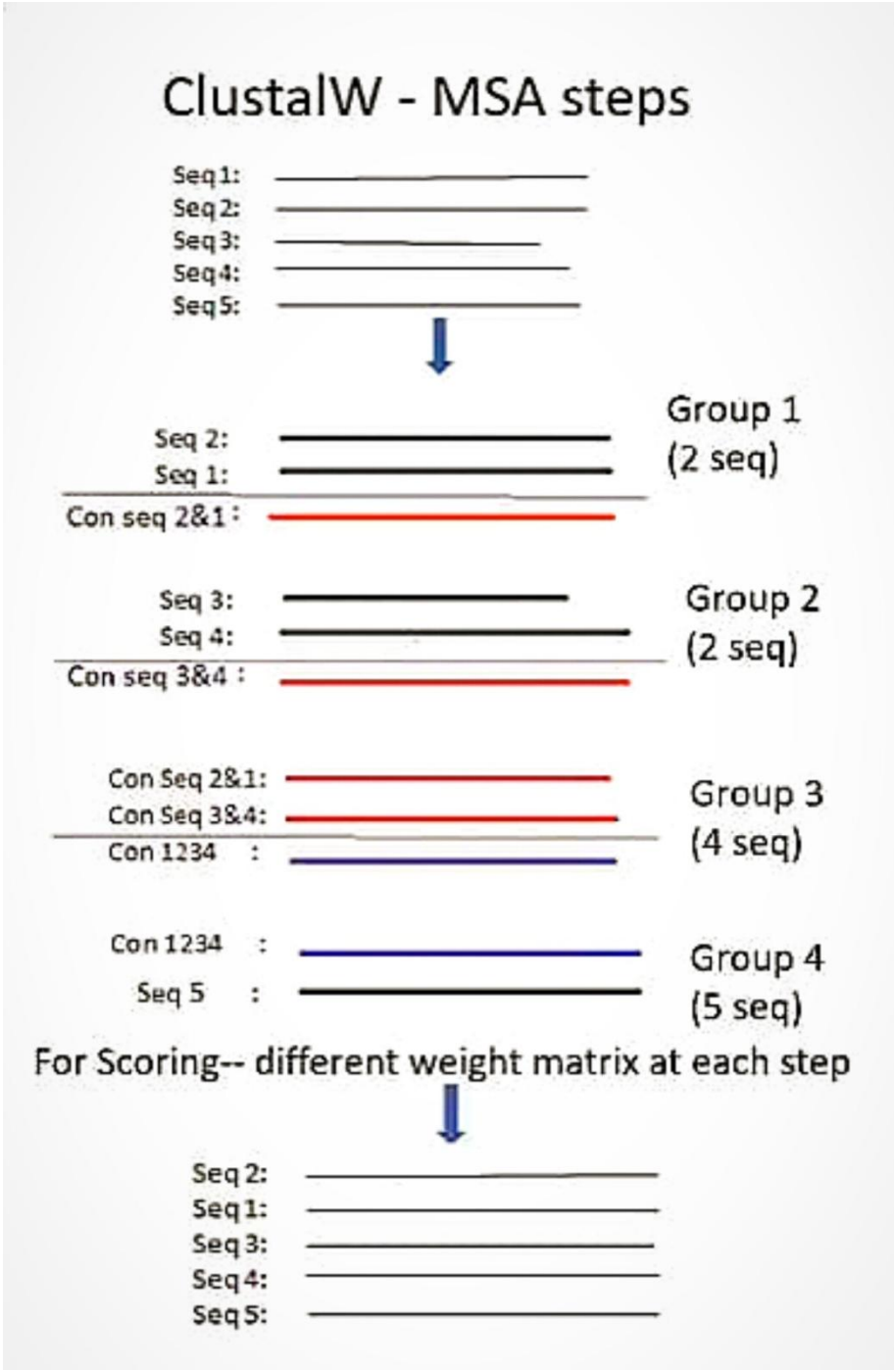
TOOLS OF MSA

- Clustal W
- Clustal W2
- Clustal Omega
- MAFFT
- MUSCLE
- M View
- T- Coffee
- Web PRANK
- MACAW

ClustalW

- CLUSTALW (eg.https://www.genome.jp/tools-bin/clustalw) is a progressive MSA program, which follows a heuristic approach.
- ClustalW is produced by Julie D. Thompson, Toby Gibson of European Molecules Biology Laboratory.
- ClustalW can create multiple alignment , manipulate exciting alignment, do profile analysis and create phylogenetic trees.
- Alignment can be done by 2 methods;
    1. Slow/accurate
    2. Fast/approximate
- 'W' stands for 'weighted' (sequences are weighted differently).

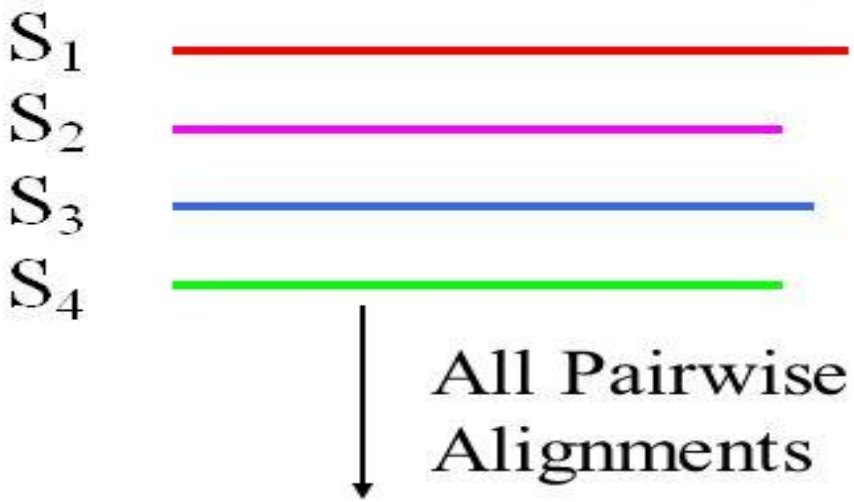
- The Clustal software align sequence using a heuristic that progressively builds a multiple sequence alignment from series of pairwise alignment.
- This method works by analysing the sequence as a whole, then utilizing the UPGMA /Neighbor-joining methods to generate a distance matrix.
- A guide tree is then calculated from the scores of the sequences in the matrix, then progressively align the sequence in order of similarity
- Essentially, Clustal creates multiple sequence alignment through three main steps:



The ClustalW Algorithm

- Step 1 : Determined all pairwise alignment between sequence and determine degree of similarity between each pair.
- Step 2 : Construct a similarity tree.
- Step 3 : Combine the alignment starting from the most closely related group to the most distantly related groups using the “once a gap always a gap” rule.

ClustalW steps

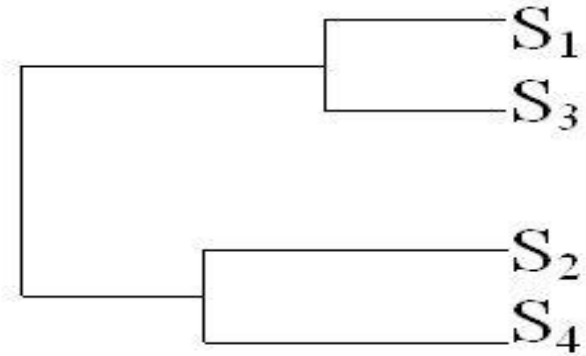


All Pairwise Alignments

Similarity Matrix

	S_1	S_2	S_3	S_4
S_1		4	9	4
S_2			4	7
S_3				4
S_4				

Cluster Analysis



Distance

Dendrogram

Multiple Alignment Step:
1. Aligning S_1 and S_3
2. Aligning S_2 and S_4
3. Aligning (S_1,S_3) with (S_2,S_4).

From Higgins(1991) and Thompson(1994).

Multiple Sequence Alignment by CLUSTALW

ETE3 | MAFFT | CLUSTALW | PRRN

Help

General Setting Parameters:

Output Format: CLUSTAL

Pairwise Alignment: **FAST/APPROXIMATE** **SLOW/ACCURATE**

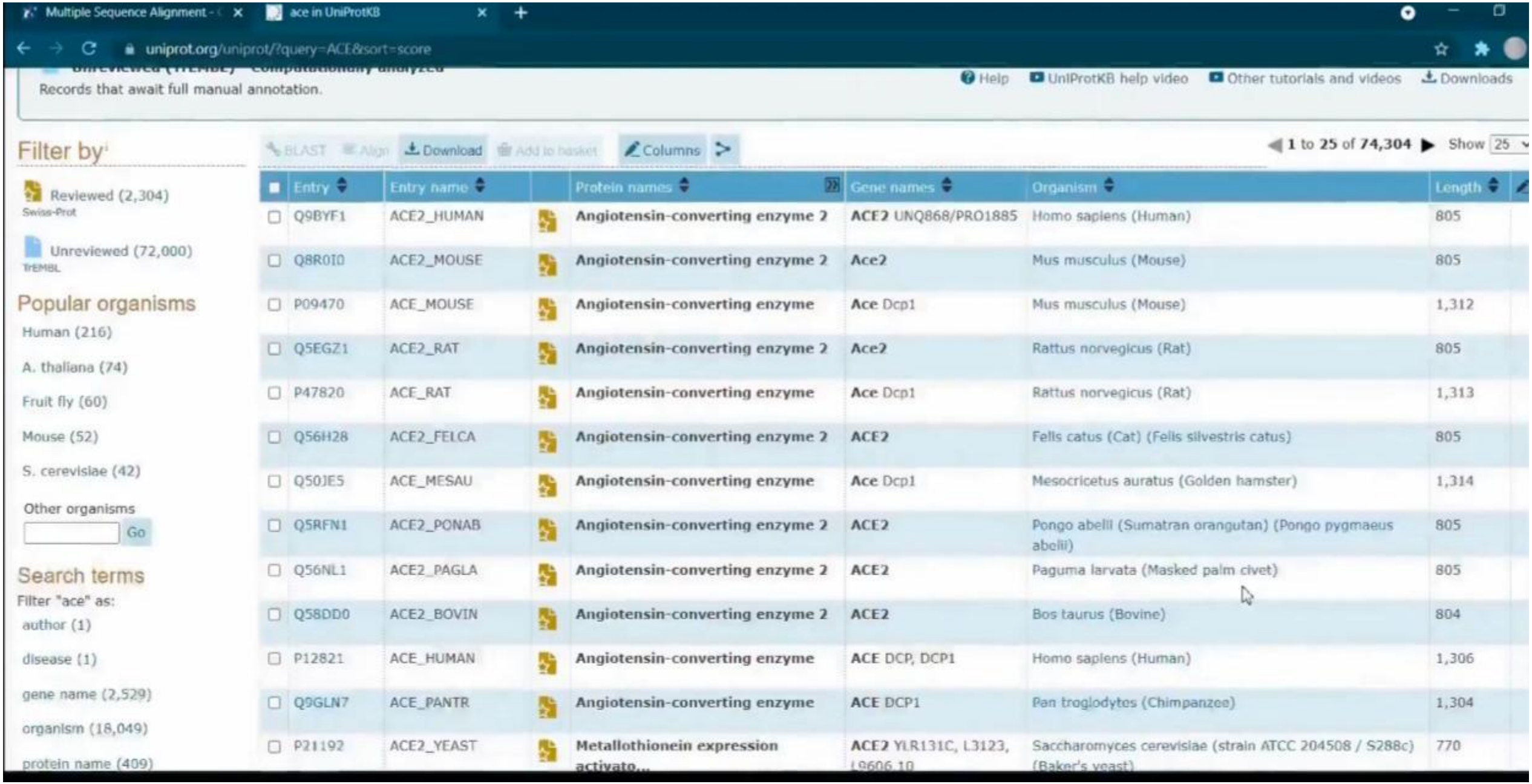
Enter your sequences (with labels) below (copy & paste): **PROTEIN** **DNA**

Support Formats: FASTA (Pearson), NBRF/PIR, EMBL/Swiss Prot, GDE, CLUSTAL, and GCG/MSF

Or give the file name containing your query

Choose file No file chosen

Execute Multiple Alignment Reset


Multiple Sequence Alignment -
ace in UniProtKB
uniprot.org/uniprot/?query=ACE&sort=score
Records that await full manual annotation.
Help
UniProtKB help video
Other tutorials and videos
Downloads
Filter by
Reviewed (2,304)
Swiss-Prot
Unreviewed (72,000)
TrEMBL
Popular organisms
Human (216)
A. thaliana (74)
Fruit fly (60)
Mouse (52)
S. cerevisiae (42)
Other organisms
Go
Search terms
Filter "ace" as:
author (1)
disease (1)
gene name (2,529)
organism (18,049)
protein name (409)
BLAST
Align
Download
Add to basket
Columns
1 to 25 of 74,304
Show 25
Entry
Entry name
Protein names
Gene names
Organism
Length
Q9BYF1 ACE2_HUMAN Angiotensin-converting enzyme 2 ACE2 UNQ868/PRO1885 Homo sapiens (Human) 805
Q8R0I0 ACE2_MOUSE Angiotensin-converting enzyme 2 Ace2 Mus musculus (Mouse) 805
P09470 ACE_MOUSE Angiotensin-converting enzyme Ace Dcp1 Mus musculus (Mouse) 1,312
Q5EGZ1 ACE2_RAT Angiotensin-converting enzyme 2 Ace2 Rattus norvegicus (Rat) 805
P47820 ACE_RAT Angiotensin-converting enzyme Ace Dcp1 Rattus norvegicus (Rat) 1,313
Q56H28 ACE2_FELCA Angiotensin-converting enzyme 2 ACE2 Felis catus (Cat) (Felis silvestris catus) 805
Q50JE5 ACE_MESAU Angiotensin-converting enzyme Ace Dcp1 Mesocricetus auratus (Golden hamster) 1,314
Q5RFN1 ACE2_PONAB Angiotensin-converting enzyme 2 ACE2 Pongo abelii (Sumatran orangutan) (Pongo pygmaeus abelii) 805
Q56NL1 ACE2_PAGLA Angiotensin-converting enzyme 2 ACE2 Paguma larvata (Masked palm civet) 805
Q58DD0 ACE2_BOVIN Angiotensin-converting enzyme 2 ACE2 Bos taurus (Bovine) 804
P12821 ACE_HUMAN Angiotensin-converting enzyme ACE DCP, DCP1 Homo sapiens (Human) 1,306
Q9GLN7 ACE_PANTR Angiotensin-converting enzyme ACE DCP1 Pan troglodytes (Chimpanzee) 1,304
P21192 ACE2_YEAST Metallothionein expression activato... ACE2 YLR131C, L3123, L9606.10 Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast) 770

More Detail Parameters...

Pairwise Alignment Parameters:

For FAST/APPROXIMATE:

K-tuple(word) size: 1, Window size: 5, Gap Penalty: 3

Number of Top Diagonals: 5, Scoring Method: PERCENT

For SLOW/ACCURATE:

Gap Open Penalty: 10.0, Gap Extension Penalty: 0.1

Select Weight Matrix: BLOSUM (for PROTEIN)

(Note that only parameters for the algorithm specified by the above "Pairwise Alignment" are valid.)

Multiple Alignment Parameters:

Gap Open Penalty: 10, Gap Extension Penalty: 0.05

Weight Transition: ○ **YES (Value:** 0.5 **),** ◉ **NO**

Hydrophilic Residues for Proteins: GPSNDQERK

Hydrophilic Gaps: ◉ **YES** ○ **NO**

Select Weight Matrix: BLOSUM (for PROTEIN)

Type additional options (delimited by whitespaces) below:

(**-options** for help)

Execute Multiple Alignment | Reset

Pairwise Alignment Parameters:

For FAST/APPROXIMATE:

K-tuple(word) size: 1 , Window size: 5 , Gap Penalty: 3

Number of Top Diagonals: 5 , Scoring Method: ABSOLUTE

Wilbur & Lipman algorithm
(Heuristic approach)

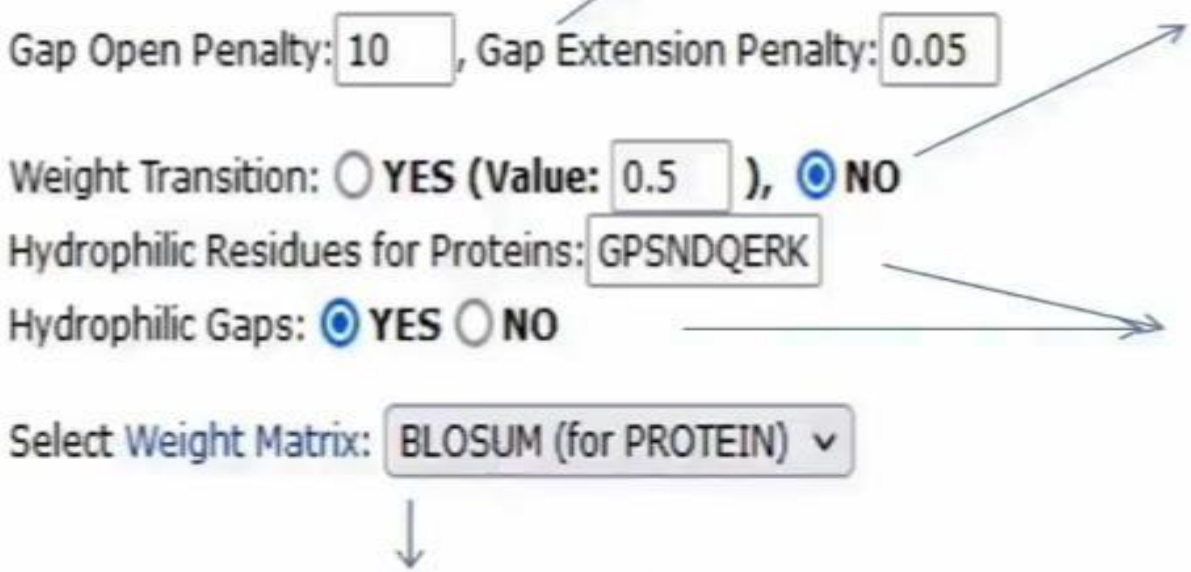
For SLOW/ACCURATE:

Gap Open Penalty: 10 , Gap Extension Penalty: 0.1

Select Weight Matrix: BLOSUM (for PROTEIN)

Dynamic programming algorithm
Choose this for less number of short sequences.

Multiple alignment parameters



During MSA to control gaps. No penalty for terminal gaps

DNA transition score: 0-1
For closely related sequence it should be near to 0 and vice versa.

For Loops and random coil regions where insertions and deletions are common.

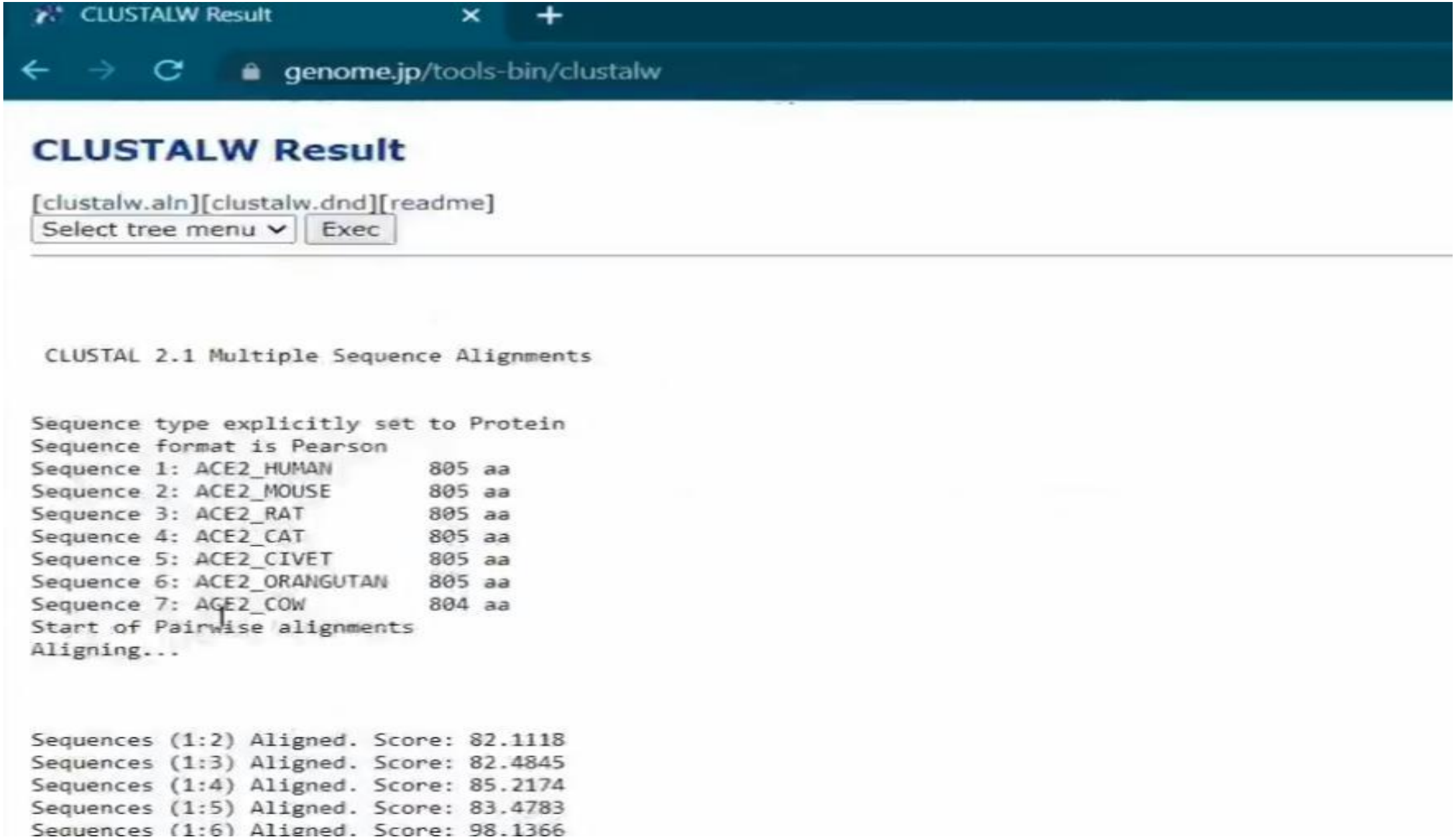
Includes gaps where 5 or more hydrophillic residues are seen consecutively.

Can alter the residues in option.

For proteins- series of matrices are used depending upon the similarity of sequences aligned at each step.

For DNA –Single matrix is used.

Result


CLUSTALW Result
genome.jp/tools-bin/clustalw
CLUSTALW Result
[clustalw.aln][clustalw.dnd][readme]
Select tree menu
Exec
CLUSTAL 2.1 Multiple Sequence Alignments
Sequence type explicitly set to Protein
Sequence format is Pearson
Sequence 1: ACE2_HUMAN 805 aa
Sequence 2: ACE2_MOUSE 805 aa
Sequence 3: ACE2_RAT 805 aa
Sequence 4: ACE2_CAT 805 aa
Sequence 5: ACE2_CIVET 805 aa
Sequence 6: ACE2_ORANGUTAN 805 aa
Sequence 7: ACE2_COW 804 aa
Start of Pairwise alignments
Aligning...
Sequences (1:2) Aligned. Score: 82.1118
Sequences (1:3) Aligned. Score: 82.4845
Sequences (1:4) Aligned. Score: 85.2174
Sequences (1:5) Aligned. Score: 83.4783
Sequences (1:6) Aligned. Score: 98.1366

CLUSTALW Result

genome.jp/tools-bin/clustalw

```
Sequence 6: ACE2_ORANGUTAN   805 aa
Sequence 7: ACE2_COW         804 aa
Start of Pairwise alignments
Aligning...


Sequences (1:2) Aligned. Score: 82.1118
Sequences (1:3) Aligned. Score: 82.4845
Sequences (1:4) Aligned. Score: 85.2174
Sequences (1:5) Aligned. Score: 83.4783
Sequences (1:6) Aligned. Score: 98.1366
Sequences (1:7) Aligned. Score: 80.8458
Sequences (2:3) Aligned. Score: 90.4348
Sequences (2:4) Aligned. Score: 81.7391
Sequences (2:5) Aligned. Score: 81.6149
Sequences (2:6) Aligned. Score: 81.7391
Sequences (2:7) Aligned. Score: 80.2239
Sequences (3:4) Aligned. Score: 81.6149
Sequences (3:5) Aligned. Score: 81.118
Sequences (3:6) Aligned. Score: 81.9876
Sequences (3:7) Aligned. Score: 80.0995
Sequences (4:5) Aligned. Score: 93.2919
Sequences (4:6) Aligned. Score: 84.9689
Sequences (4:7) Aligned. Score: 82.9602
Sequences (5:6) Aligned. Score: 83.2298
Sequences (5:7) Aligned. Score: 81.592
Sequences (6:7) Aligned. Score: 81.0945
Guide tree file created:   [clustalw.dnd]

There are 6 groups
Start of Multiple Alignment

Aligning...
Group 1: Sequences:   2      Score:13303
Group 2: Sequences:   2      Score:13025
Group 3: Sequences:   4      Score:12290
Group 4: Sequences:   2      Score:12748
Group 5: Sequences:   6      Score:12194
Group 6: Sequences:   7      Score:12107
Alignment Score 92252

CLUSTAL-Alignment file created  [clustalw.aln]
```

```
clustalw.aln

CLUSTAL 2.1 multiple sequence alignment
```

CLUSTALW Result

genome.jp/tools-bin/clustalw

```
 CLUSTAL 2.1 Multiple Sequence Alignments


Sequence type explicitly set to Protein
Sequence format is Pearson
Sequence 1: ACE2_HUMAN       805 aa
Sequence 2: ACE2_MOUSE       805 aa
Sequence 3: ACE2_RAT         805 aa
Sequence 4: ACE2_CAT         805 aa
Sequence 5: ACE2_CIVET       805 aa
Sequence 6: ACE2_ORANGUTAN   805 aa
Sequence 7: ACE2_COW         804 aa
Start of Pairwise alignments
Aligning...


Sequences (1:2) Aligned. Score: 82.1118
Sequences (1:3) Aligned. Score: 82.4845
Sequences (1:4) Aligned. Score: 85.2174
Sequences (1:5) Aligned. Score: 83.4783
Sequences (1:6) Aligned. Score: 98.1366
Sequences (1:7) Aligned. Score: 80.8458
Sequences (2:3) Aligned. Score: 90.4348
Sequences (2:4) Aligned. Score: 81.7391
Sequences (2:5) Aligned. Score: 81.6149
Sequences (2:6) Aligned. Score: 81.7391
Sequences (2:7) Aligned. Score: 80.2239
Sequences (3:4) Aligned. Score: 81.6149
Sequences (3:5) Aligned. Score: 81.118
Sequences (3:6) Aligned. Score: 81.9876
Sequences (3:7) Aligned. Score: 80.0995
Sequences (4:5) Aligned. Score: 93.2919
Sequences (4:6) Aligned. Score: 84.9689
Sequences (4:7) Aligned. Score: 82.9602
Sequences (5:6) Aligned. Score: 83.2298
Sequences (5:7) Aligned. Score: 81.592
Sequences (6:7) Aligned. Score: 81.0945
Guide tree file created:   [clustalw.dnd]

There are 6 groups
Start of Multiple Alignment

Aligning...
Group 1: Sequences:   2      Score:13303
Group 2: Sequences:   2      Score:13025
```

guidetree.docx - Word — Abhishek Mudaliar

File | Home | Insert | Design | Layout | References | Mailings | Review | View | Help | Tell me

Paste | Clipboard | Calibri (Body) | 11 | Font | Paragraph | AaBbCcDd ¶ Normal | AaBbCcDd ¶ No Spac... | AaBbC Heading 1 | Styles

---

clustalw.dnd

```
(
(
(
ACE2_HUMAN:0.00820,
ACE2_ORANGUTAN:0.01043)
:0.07143,
(
ACE2_CAT:0.02784,
ACE2_CIVET:0.03924)
:0.04347)
:0.00480,
(
ACE2_MOUSE:0.04728,
ACE2_RAT:0.04837)
:0.05047,
ACE2_COW:0.10008);
```

CLUSTALW Result

genome.jp/tools-bin/clustalw

```
Sequence 6: ACE2_ORANGUTAN   805 aa
Sequence 7: ACE2_COW         804 aa
Start of Pairwise alignments
Aligning...



Sequences (1:2) Aligned. Score: 82.1118
Sequences (1:3) Aligned. Score: 82.4845
Sequences (1:4) Aligned. Score: 85.2174
Sequences (1:5) Aligned. Score: 83.4783
Sequences (1:6) Aligned. Score: 98.1366
Sequences (1:7) Aligned. Score: 80.8458
Sequences (2:3) Aligned. Score: 90.4348
Sequences (2:4) Aligned. Score: 81.7391
Sequences (2:5) Aligned. Score: 81.6149
Sequences (2:6) Aligned. Score: 81.7391
Sequences (2:7) Aligned. Score: 80.2239
Sequences (3:4) Aligned. Score: 81.6149
Sequences (3:5) Aligned. Score: 81.118
Sequences (3:6) Aligned. Score: 81.9876
Sequences (3:7) Aligned. Score: 80.0995
Sequences (4:5) Aligned. Score: 93.2919
Sequences (4:6) Aligned. Score: 84.9689
Sequences (4:7) Aligned. Score: 82.9602
Sequences (5:6) Aligned. Score: 83.2298
Sequences (5:7) Aligned. Score: 81.592
Sequences (6:7) Aligned. Score: 81.0945
Guide tree file created:   [clustalw.dnd]

There are 6 groups
Start of Multiple Alignment

Aligning...
Group 1: Sequences:   2      Score:13303
Group 2: Sequences:   2      Score:13025
Group 3: Sequences:   4      Score:12290
Group 4: Sequences:   2      Score:12748
Group 5: Sequences:   6      Score:12194
Group 6: Sequences:   7      Score:12107
Alignment Score 92252

CLUSTAL-Alignment file created  [clustalw.aln]
```

clustalw.aln

```
CLUSTAL 2.1 multiple sequence alignment
```

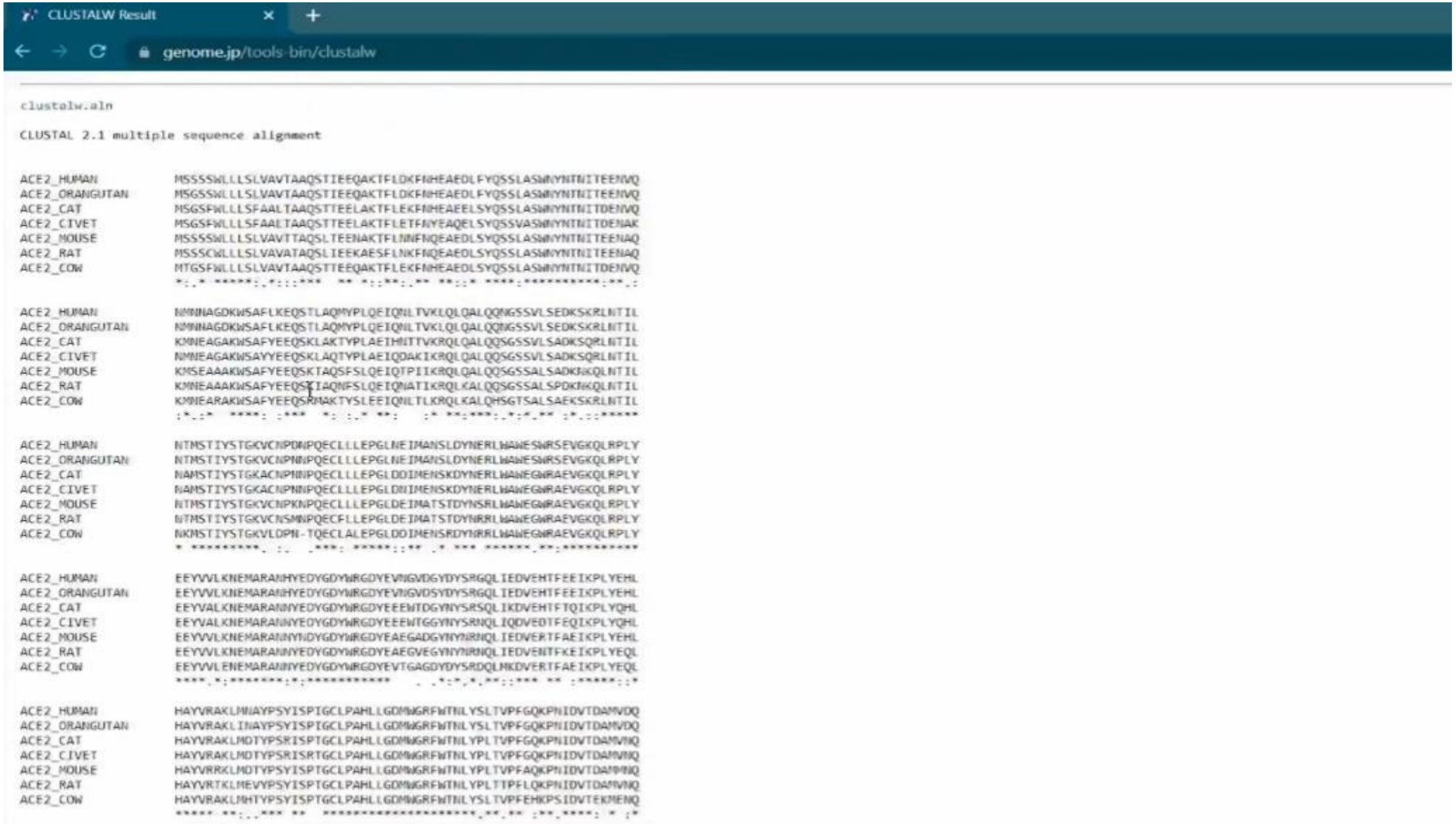
guidetree.docx - Word — Abhishek Mudaliar

File Home Insert Design Layout References Mailings Review View Help Tell me

Paste | Clipboard | Calibri (Body) 11 | Font | Paragraph | AaBbCcDd ¶ Normal | AaBbCcDd ¶ No Spac... | AaBbC Heading 1 | Styles | Editi

clustalw.dnd

```
(
(
(
ACE2_HUMAN:0.00820,
ACE2_ORANGUTAN:0.01043)
:0.07143,
(
ACE2_CAT:0.02784,
ACE2_CIVET:0.03924)
:0.04347)
:0.00480,
(
ACE2_MOUSE:0.04728,
ACE2_RAT:0.04837)
:0.05047,
ACE2_COW:0.10008);
```


CLUSTALW Result
genome.jp/tools-bin/clustalw
clustalw.aln
CLUSTAL 2.1 multiple sequence alignment
ACE2_HUMAN MSSSSWLLLSLVAVTAAQSTIEEQAKTFLDKFNHEAEDLFYQSSLASWNYNTNITEENVQ
ACE2_ORANGUTAN MSGSSWLLLSLVAVTAAQSTIEEQAKTFLDKFNHEAEDLFYQSSLASWNYNTNITEENVQ
ACE2_CAT MSGSFWLLLSFAALTAAQSTTEELAKTFLEKFNHEAEELSYQSSLASWNYNTNITDENVQ
ACE2_CIVET MSGSFWLLLSFAALTAAQSTTEELAKTFLETFNYEAQELSYQSSVASWNYNTNITDENAK
ACE2_MOUSE MSSSSWLLLSLVAVTTAQSLEENAKTFLNNFNQEAEDLSYQSSLASWNYNTNITEENAQ
ACE2_RAT MSSSCWLLLSLVAVATAQSLIEEKAESFLNKFNQEAEDLSYQSSLASWNYNTNITEENAQ
ACE2_COW MTGSFWLLLSLVAVTAAQSTTEEQAKTFLEKFNHEAEDLSYQSSLASWNYNTNITDENVQ
ACE2_HUMAN NMNNAGDKWSAFLKEQSTLAQMYPLQEIQNLTVKLQLQALQQNGSSVLSEDKSKRLNTIL
ACE2_ORANGUTAN NMNNAGDKWSAFLKEQSTLAQMYPLQEIQNLTVKLQLQALQQNGSSVLSEDKSKRLNTIL
ACE2_CAT KMNEAGAKWSAFYEEQSKLAKTYPLAEIHNTTVKRQLQALQQSGSSVLSADKSQRLNTIL
ACE2_CIVET NMNEAGAKWSAYYEEQSKLAQTYPLAEIQDAKIKRQLQALQQSGSSVLSADKSQRLNTIL
ACE2_MOUSE KMSEAAAKWSAFYEEQSKTAQSFSLQEIQTPIIKRQLQALQQSGSSALSADKNKQLNTIL
ACE2_RAT KMNEAAAKWSAFYEEQSKIAQNFSLQEIQNATIKRQLKALQQSGSSALSPDKNKQLNTIL
ACE2_COW KMNEARAKWSAFYEEQSRMAKTYSLEEIQNLTLKRQLKALQHSGTSALSAEKSKRLNTIL
ACE2_HUMAN NTMSTIYSTGKVCNPDNPQECLLLEPGLNEIMANSLDYNERLWAWESWRSEVGKQLRPLY
ACE2_ORANGUTAN NTMSTIYSTGKVCNPNNPQECLLLEPGLNEIMANSLDYNERLWAWESWRSEVGKQLRPLY
ACE2_CAT NAMSTIYSTGKACNPNNPQECLLLEPGLDDIMENSKDYNERLWAWEGWRAEVGKQLRPLY
ACE2_CIVET NAMSTIYSTGKACNPNNPQECLLLEPGLDNIMENSKDYNERLWAWEGWRAEVGKQLRPLY
ACE2_MOUSE NTMSTIYSTGKVCNPKNPQECLLLEPGLDEIMATSTDYNSRLWAWEGWRAEVGKQLRPLY
ACE2_RAT NTMSTIYSTGKVCNSMNPQECFLLEPGLDEIMATSTDYNRRLWAWEGWRAEVGKQLRPLY
ACE2_COW NKMSTIYSTGKVLDPN-TQECLALEPGLDDIMENSRDYNRRLWAWEGWRAEVGKQLRPLY
ACE2_HUMAN EEYVVLKNEMARANHYEDYGDYWRGDYEVNGVDGYDYSRGQLIEDVEHTFEEIKPLYEHL
ACE2_ORANGUTAN EEYVVLKNEMARANHYEDYGDYWRGDYEVNGVDSYDYSRGQLIEDVEHTFEEIKPLYEHL
ACE2_CAT EEYVALKNEMARANNYEDYGDYWRGDYEEEWTDGYNYSRSQLIKDVEHTFTQIKPLYQHL
ACE2_CIVET EEYVALKNEMARANNYEDYGDYWRGDYEEEWTGGYNYSRNQLIQDVEDTFEQIKPLYQHL
ACE2_MOUSE EEYVVLKNEMARANNYNDYGDYWRGDYEAEGADGYNYNRNQLIEDVERTFAEIKPLYEHL
ACE2_RAT EEYVVLKNEMARANNYEDYGDYWRGDYEAEGVEGYNYNRNQLIEDVENTFKEIKPLYEQL
ACE2_COW EEYVVLENEMARANNYEDYGDYWRGDYEVTGAGDYDYSRDQLMKDVERTFAEIKPLYEQL
ACE2_HUMAN HAYVRAKLMNAYPSYISPIGCLPAHLLGDMWGRFWTNLYSLTVPFGQKPNIDVTDAMVDQ
ACE2_ORANGUTAN HAYVRAKLINAYPSYISPIGCLPAHLLGDMWGRFWTNLYSLTVPFGQKPNIDVTDAMVDQ
ACE2_CAT HAYVRAKLMDTYPSRISPTGCLPAHLLGDMWGRFWTNLYPLTVPFGQKPNIDVTDAMVNQ
ACE2_CIVET HAYVRAKLMDTYPSRISRTGCLPAHLLGDMWGRFWTNLYPLTVPFGQKPNIDVTDAMVNQ
ACE2_MOUSE HAYVRRKLMDTYPSYISPTGCLPAHLLGDMWGRFWTNLYPLTVPFAQKPNIDVTDAMVNQ
ACE2_RAT HAYVRTKLMEVYPSYISPTGCLPAHLLGDMWGRFWTNLYPLTTPFLQKPNIDVTDAMVNQ
ACE2_COW HAYVRAKLMNTYPSYISPTGCLPAHLLGDMWGRFWTNLYSLTVPFEHKPSIDVTEKMENQ

ClustalW output - Symbol

- ‘ * ‘ indicates positions which have a single fully conserved residue
- ‘ : ‘ indicates that one of the following ‘strong’ groups is fully conserved

  STA

  NEQK

  NHQK

  NDEQ

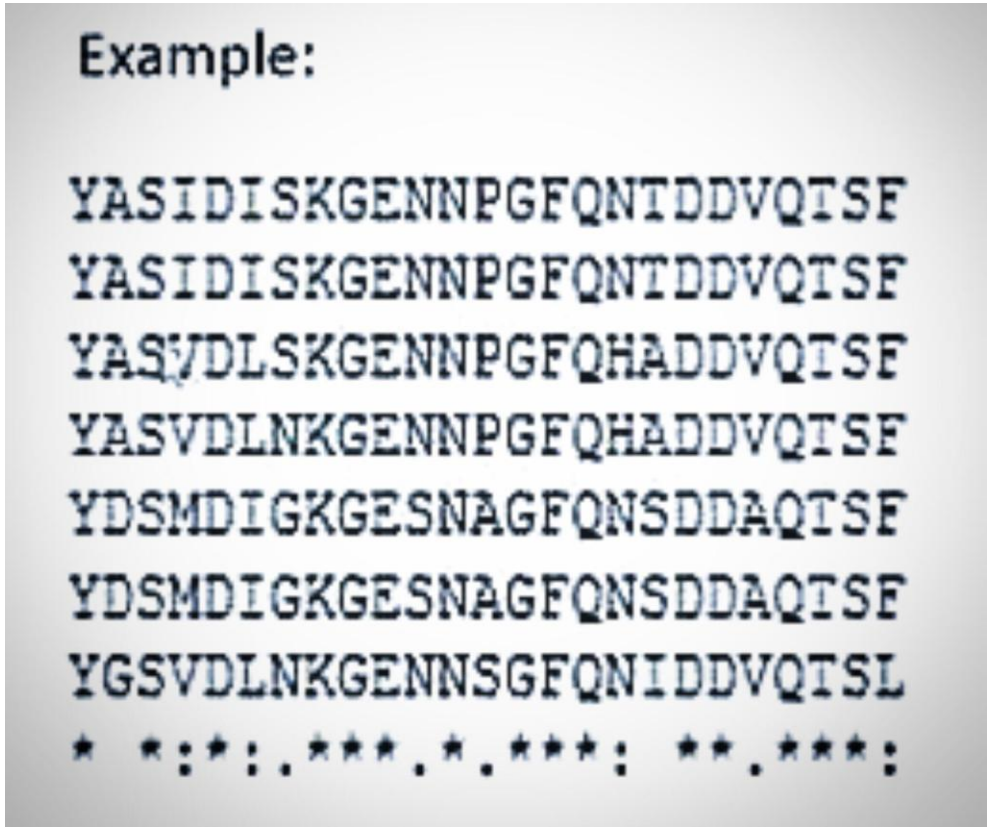
  QHRK

  MILV

  MILF

  HY

  FYW



Cont,

- ‘. ‘indicates that one of the following ‘weaker’ groups is fully conserved

  CSA

  ATV

  SAG

  STNK

  STPA

  SGND

  SNDEQK

  NDEQRK

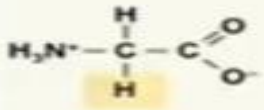
  NEQHRK

  FVLIM

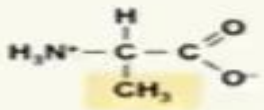
  HFY

Amino acid Substitutions

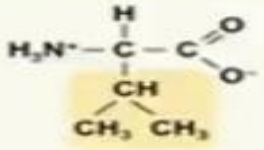
Hydrophobic amino acid (Non-Polar)



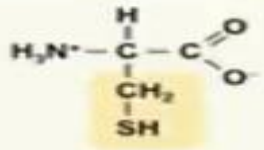
Glycine
(Gly / G)



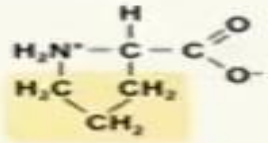
Alanine
(Ala / A)



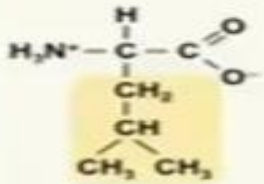
Valine
(Val / V)



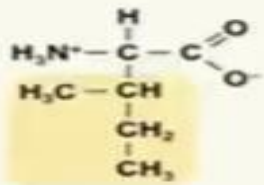
Cysteine
(Cys / C)



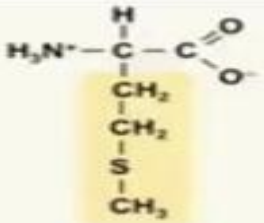
Proline
(Pro / P)



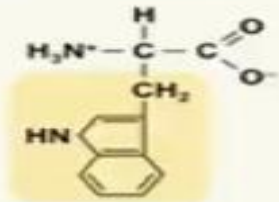
Leucine
(Leu / L)



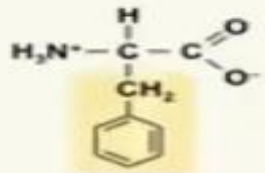
Isoleucine
(Ile / I)



Methionine
(Met / M)

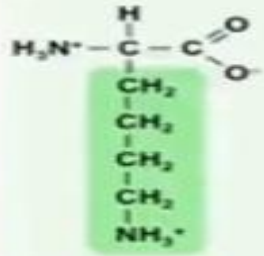


Tryptophan
(Trp / W)

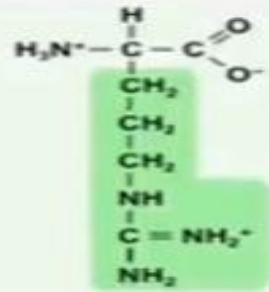


Phenylalanine
(Phe / F)

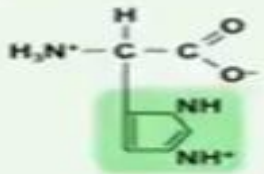
Basic (+ charge), Polar



Lysine
(Lys / K)

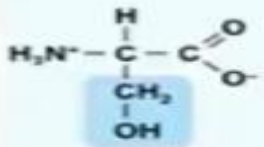


Arginine
(Arg / R)

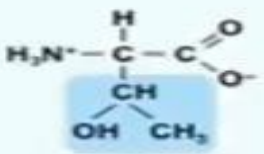


Histidine
(His / H)

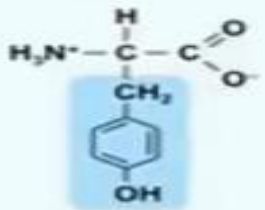
Less polar



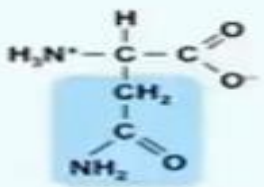
Serine
(Ser / S)



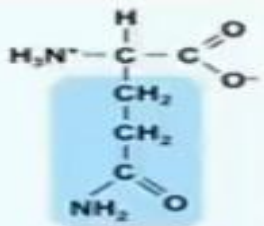
Threonine
(Thr / T)



Tyrosine
(Tyr / Y)

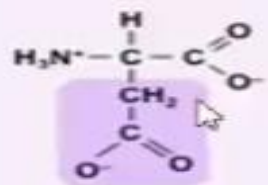


Asparagine
(Asn / N)

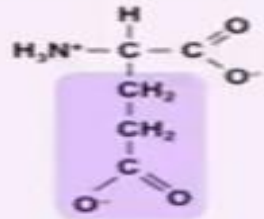


Glutamine
(Gln / Q)

Acidic (- charge), Polar



Aspartic Acid
(Asp / D)



Glutamic Acid
(Glu / E)

1. Aliphatic , aromatic
2. Size
3. Acidic, Basic (charged), Neutral (uncharged)
4. Hydrophobic (non-polar), hydrophilic (polar)

T-COFFEE

- T-COFFEE (Tree –based Consistency Objectives Function For AlignmEnt Evaluation) is a multiple sequence alignment software using a progressive approach.
- It generates a library of pairwise alignment to guide the multiple sequence alignment.
- It has advanced features to evaluate the quality of alignments and capacity for identifying occurrence of motifs.
- The main characteristic of T-coffee is that it will combine results obtained with several alignment methods. This tool can align up to 500 sequences or a maximum file size of 1 MB.
- T coffee integrates different Pair-wise alignment techniques and combines different multiple alignment methods and sequence alignment methods and plug in user knowledge

ALGORITHMS

How coffee works?

- T-Coffee alignment utilizes heterogeneous data sources and provide simple and flexible means of generating multiple alignment.
- Create a library of pairwise alignment for each possible pairs of sequences and compare each pair of aligned residues in the MSA to its counterpart in the library.
- T-coffee can complete multiple alignment using a library that was generated using a mixture of local and global pair-wise alignment.
- Weight alignment by the percentage of identical residues.
- Residues that consistently match up, end up with very high weight, this is what is meant by CONSISTENCY-BASED SCORING.

The overall consistency score is equal to the number of pairs that occurs in both MSA and the library, divided by the total no. Of pairs in MSA.

T- Coffee and Consistency.....

1. With T-coffee we pre -process a data set of all pair -wise alignment between the Sequences.
2. This provides us with a library of alignment information that can used to guide the progressive alignment.
3. Intermediate alignment are based not only on the sequence to be aligned next, and also on how all of the sequence align with each other.

T-Coffee

Multiple Sequence Alignment

T-Coffee is a multiple sequence alignment program. Its main characteristic is that it will allow you to combine results obtained with several alignment methods.

Important note: This tool can align up to 500 sequences or a maximum file size of 1 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of

PROTEIN

sequences in any supported format:

>Seq1

>Seq5
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAVDDLLLSPDDIAQWFIEDPGPDEAPRMSEAASPVDPAPAAP
IPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAM
AIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNS
SCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGRDRRTEEENFRKKGEPHHELPPGSTKRALPNNTSSSPQPKK
KPLDGEYFTLQDQTSFQKENC

Or upload a file: Choose File No file chosen

Use a example sequence | Clear sequence | See more example inputs

STEP 2 - Set your Parameters

OUTPUT FORMAT:

HTML

The default settings will fulfill the needs of most users.

More options... *(Click here, if you want to view or change the default settings.)*

STEP 3 - Submit your job

Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Results < T-Coffee < Multiple Se

ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=tcoffee-I20220819-113257-0191-82986554-p1m

T-Coffee

Input form | Web services | Help & Documentation | Bioinformatics Tools FAQ | Feedback

Tools > Multiple Sequence Alignment > T-Coffee

Your job is currently running... please be patient

The result of your job will appear in this browser window.

Job ID: tcoffee-I20220819-113257-0191-82986554-p1m

Please note the following

- You may press Shift+Refresh or Reload on your browser at any time to check if results are ready.
- You may bookmark this page to view your results later if you wish.
- Results are stored for 7 days.

Waiting for www.ebi.ac.uk...

Type here to search

ENG

T-Coffee

Results for job tcoffee-I20220819-113257-0191-82986554-p1m

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

Download Alignment File

```
T-COFFEE, Version_13.45.0.4846264 (2020-10-15 17:52:11 - Revision 5becd5d - Build 620)
Cedric Notredame
CPU TIME:0 sec.
SCORE=655
*
 BAD AVG GOOD
*
AC105292.3_1867    :  54
AE014298.5_9585    :  68
CP023329.1_9479    :  68
CP023335.1_9479    :  68
NM_001272455.2_    :  68
cons               :  65

AC105292.3_1867    GA----TTTGCTT-CTCTGTTGTTTGGTTCA-ATCGTCT--TTAT
AE014298.5_9585    ATGCCATTTGTGGACCCCTCAGCGTCGCACATATACACGCCATAT
CP023329.1_9479    ATGCCATTTGTGGACCCCTCAGCGTCGCACATATACACGCCATAT
CP023335.1_9479    ATGCCATTTGTGGACCCCTCAGCGTCGCACATATACACGCCATAT
NM_001272455.2_    ATGCCATTTGTGGACCCCTCAGCGTCGCACATATACACGCCATAT

cons                   ****    * *    *  * *  ** **   *    ***


AC105292.3_1867    TTTGCTTCACTGTTGACAAACATTCCATTT-GTTTATATCATTTA
AE014298.5_9585    C-TCCAACCATGCCGCCCCAAAACCGATTTTCAGTTTTTGACCTT
```

```
CPU TIME:0 sec.
SCORE=655
*
 BAD AVG GOOD
*
AC105292.3_1867    :  54
AE014298.5_9585    :  68
CP023329.1_9479    :  68
CP023335.1_9479    :  68
NM_001272455.2_    :  68
cons               :  65

AC105292.3_1867    GA----TTTGCTT-CTCTGTTGTTTGGTTCA-ATCGTCT--TTAT
AE014298.5_9585    ATGCCATTTGTGGACCCCTCAGCGTCGCACATATACACGCCATAT
CP023329.1_9479    ATGCCATTTGTGGACCCCTCAGCGTCGCACATATACACGCCATAT
CP023335.1_9479    ATGCCATTTGTGGACCCCTCAGCGTCGCACATATACACGCCATAT
NM_001272455.2_    ATGCCATTTGTGGACCCCTCAGCGTCGCACATATACACGCCATAT

cons                   ****     * *     *   * *   ** **    *     ***


AC105292.3_1867    TTTGCTTCACTGTTGACAAACATTCCATTT-GTTTATATCATTTA
AE014298.5_9585    C-TCCAACCATGCCGCCCCAAAACCGATTTTCAGTTTTTGACCTT
CP023329.1_9479    C-TCCAACCATGCCGCCCCAAAACCGATTTTCAGTTTTTGACCTT
CP023335.1_9479    C-TCCAACCATGCCGCCCCAAAACCGATTTTCAGTTTTTGACCTT
NM_001272455.2_    C-TCCAACCATGCCGCCCCAAAACCGATTTTCAGTTTTTGACCTT

cons                * *   *   **   * *   * *   * ****     * * * *   *
```

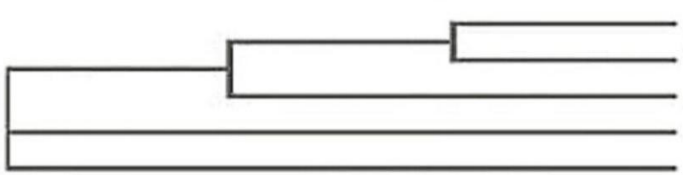
Alignments | Result Summary | Guide Tree | Phylogenetic Tree | Results Viewers | Submission Details

Download Phylogenetic Tree Data

Phylogenetic Tree

This is a Neighbour-joining tree without distance corrections.

Branch length: ◉ Cladogram ○ Real



AC105292.3_18679-22817 0.52693
AE014298.5_9585675-9589813 0
CP023329.1_9479705-9483843 0
CP023335.1_9479704-9483842 0
NM_001272455.2_1-4139 0

Tree Data

```
(
(
(
AC105292.3_18679-22817:0.52693,
AE014298.5_9585675-9589813:0.00000)
:0.00000,
CP023329.1_9479705-9483843:0.00000)
:0.00000,
CP023335.1_9479704-9483842:0.00000,
NM_001272455.2_1-4139:0.00000);
```

Applications of MSA

- Detecting similarities between sequences(closely or distinctly related).
- Detecting conserved regions or motifs in sequences.
- Detecting of structural homologies.
- Thus, assisting the improved prediction of secondary and tertiary structures of proteins.
- An important step for phylogenetic analysis.
- Useful in designing experiments to test and modify the function of specific proteins and also in predicting the function and structure of proteins, and in identifying new members of protein families.

References

1. Statistics for Bioinformatics

-Julie Dawn Thompson

2. Multiple Sequence Alignment methods

-David J.Russell

3. Bioinformatics for DUMMIES (2nd Edition)

-Jean – Micha Claverie, Cedric Notredame


THANK YOU