STRUCTURAL BIOINFORMATICS ASSIGNMENT 3

Name 1: Shruti Verma – 7022659

Name 2: Umutcan Ünaldı – 7025677

Name 3: Nazlıgül Keske - 7025902

THEORITICAL EXERCISES

Q1. Using the simple Markov Model shown in Figure 1.1, write the transition probability matrix and calculate the probability that the weather for the next five days will be Rainy, Rainy, Cloudy, Sunny, Sunny, given that today is Sunny. (2 points) You can use numpy or other tools to do actual calculations (in this case you still need to provide a transition matrix and describe steps to solve this task).

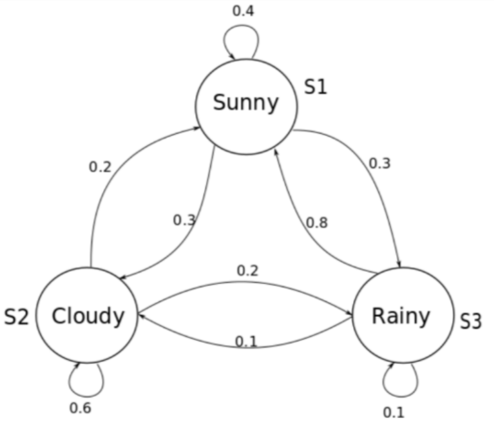


Figure 1.1: Simple Markov Model

$$\begin{array}{c|ccc} & S & R & C \\ \hline S & 0.4 & 0.3 & 0.3 \\ R & 0.8 & 0.1 & 0.1 \\ C & 0.2 & 0.2 & 0.6 \end{array}$$ => Transition probability matrix, where S= Sunny, R= Rainy, C= Cloudy

Calculating initial/ stationary state

Since the present day is 'Sunny' so we are supposing $\pi = \begin{array}{ccc} S & R & C \\ [1 & 0 & 0] \end{array}$

Solving the equation $\pi A=\pi$

$$[1 \quad 0 \quad 0]\begin{bmatrix}0.4 & 0.3 & 0.3\\0.8 & 0.1 & 0.1\\0.2 & 0.2 & 0.6\end{bmatrix} = [0.4 \quad 0.3 \quad 0.3]$$

$$[0.4 \quad 0.3 \quad 0.3]\begin{bmatrix}0.4 & 0.3 & 0.3\\0.8 & 0.1 & 0.1\\0.2 & 0.2 & 0.6\end{bmatrix} = [0.46 \quad 0.21 \quad 0.33]$$

$$[0.46 \quad 0.21 \quad 0.33]\begin{bmatrix}0.4 & 0.3 & 0.3\\0.8 & 0.1 & 0.1\\0.2 & 0.2 & 0.6\end{bmatrix} = [0.418 \quad 0.225 \quad 0.357]$$

$$[0.418 \quad 0.225 \quad 0.357]\begin{bmatrix}0.4 & 0.3 & 0.3\\0.8 & 0.1 & 0.1\\0.2 & 0.2 & 0.6\end{bmatrix} =[0.4186 \quad 0.2193 \quad 0.3621]$$

$$[0.4186 \quad 0.2193 \quad 0.3621]\begin{bmatrix}0.4 & 0.3 & 0.3\\0.8 & 0.1 & 0.1\\0.2 & 0.2 & 0.6\end{bmatrix} =[0.4153 \quad 0.21993 \quad 0.36477]$$

$$[0.4153 \quad 0.21993 \quad 0.36477]\begin{bmatrix}0.4 & 0.3 & 0.3\\0.8 & 0.1 & 0.1\\0.2 & 0.2 & 0.6\end{bmatrix} =[0.415 \quad 0.219 \quad 0.365]$$

Since, **πA=π**

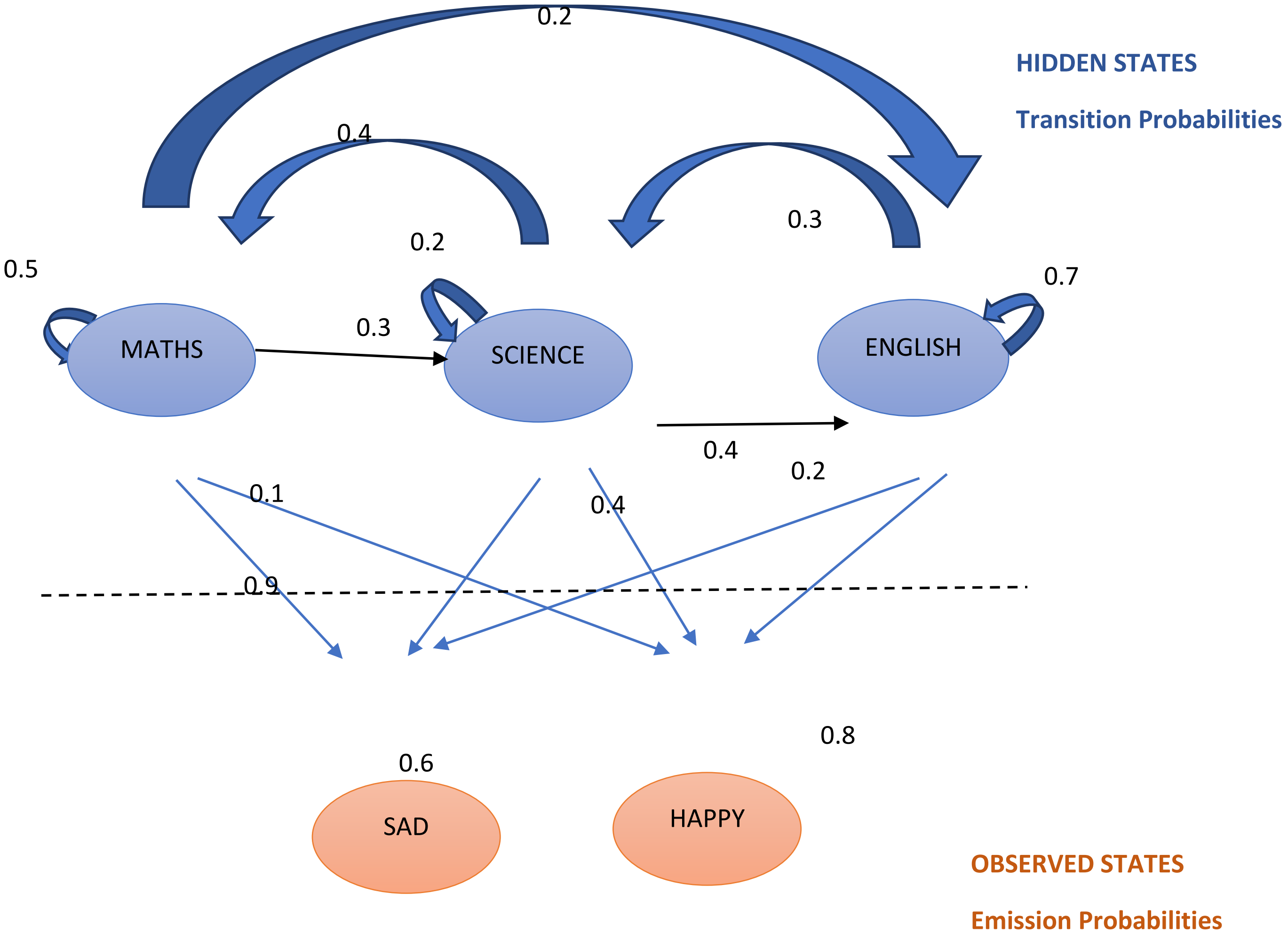
$$\text{Therefore,} \boldsymbol{\pi} = [\mathbf{0.415} \quad \mathbf{0.219} \quad \mathbf{0.365}]$$

"To find the probability of the given sequence "Rainy, Rainy, Cloudy, Sunny, Sunny" **given that today is Sunny,**

- ⇨ P(S) x [P(R) | S] x [P(R) | R] x [P(C) | R] x [P(S) | C] x [P(S) | S]
- ⇨ 0.415 x 0.3 x 0.1 x 0.1 x 0.2 x 0.4
- ⇨ 0.0000996

Therefore, the probability of the sequence given is **0.0000996**

Q2. Explain Hidden Markov Models (HMM) architecture with a sample diagram (don't use the one from the lecture). Make sure to mark all of the components in the diagram along with their probabilities and write down the total number of parameters.



$$\begin{array}{c|ccc} & M & Sc & E \\ \hline M & 0.5 & 0.3 & 0.2 \\ Sc & 0.4 & 0.2 & 0.4 \\ E & 0.0 & 0.3 & 0.7 \end{array}$$

→ **Transition Matrix**

$$\begin{array}{c|cc} & S & H \\ \hline M & 0.9 & 0.1 \\ Sc & 0.6 & 0.4 \\ E & 0.2 & 0.8 \end{array}$$

→ **Emission Matrix**

PARAMETERS: -

X = Hidden States = Maths (M), Science (Sc), English (E)

Y = Observed States = Happy (H). Sad (S)

A = Transition probability, which is:

$$\begin{array}{c|ccc} & M & Sc & E \\ \hline M & 0.5 & 0.3 & 0.2 \\ Sc & 0.4 & 0.2 & 0.4 \\ E & 0.0 & 0.3 & 0.7 \end{array}$$

B = Output/ Emission probability, which is:

$$\begin{array}{c|cc} & S & H \\ \hline M & 0.9 & 0.1 \\ S & 0.6 & 0.4 \\ E & 0.2 & 0.8 \end{array}$$

Let's consider the example below for better explanation:-

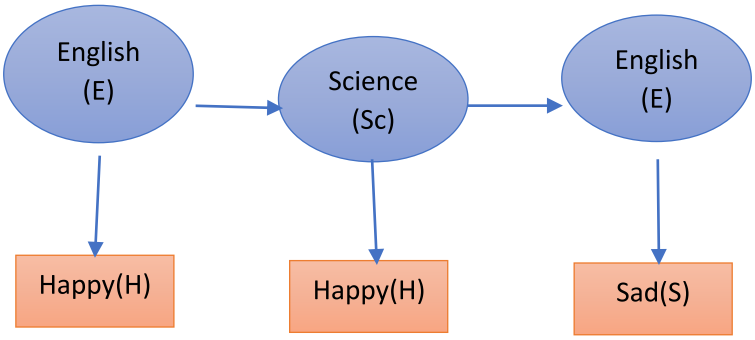


Figure 2

Finding the probability of the above sequence.

For reference:

$$\begin{array}{c} \begin{array}{ccc}\text{M} & \text{Sc} & \text{E}\end{array} \\ [\boldsymbol{A} \quad \boldsymbol{B} \quad \boldsymbol{C}] \end{array} \qquad \begin{array}{cc} & \begin{array}{ccc}\text{M} & \text{Sc} & \text{E}\end{array} \\ \begin{array}{c}\text{M}\\ \text{Sc}\\ \text{E}\end{array} & \begin{bmatrix} a & d & g \\ b & e & h \\ c & f & i \end{bmatrix} \end{array}$$

Step 1: We have to find Initial/stationary state first.

Since today the student is studying English (E), so we are supposing that $\boldsymbol{\pi} = \begin{array}{ccc}\text{M} & \text{Sc} & \text{E}\\ [\mathbf{0} & \mathbf{0} & \mathbf{1}]\end{array}$

Solving the equation πA=π

$$\begin{array}{cc} & \begin{array}{ccc}\text{M} & \text{Sc} & \text{E}\end{array} \\ [\mathbf{0} \quad \mathbf{0} \quad \mathbf{1}] & \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.4 & 0.2 & 0.4 \\ 0.0 & 0.3 & 0.7 \end{bmatrix} = [\mathbf{0} \quad \mathbf{0.3} \quad \mathbf{0.7}] \end{array}$$

$$[\mathbf{0} \quad \mathbf{0.3} \quad \mathbf{0.7}] \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.4 & 0.2 & 0.4 \\ 0.0 & 0.3 & 0.7 \end{bmatrix} = [\mathbf{0.12} \quad \mathbf{0.27} \quad \mathbf{0.61}]$$

$$[\mathbf{0.12} \quad \mathbf{0.27} \quad \mathbf{0.61}] \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.4 & 0.2 & 0.4 \\ 0.0 & 0.3 & 0.7 \end{bmatrix} = [\mathbf{0.168} \quad \mathbf{0.273} \quad \mathbf{0.559}]$$

$$[\mathbf{0.168} \quad \mathbf{0.273} \quad \mathbf{0.559}] \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.4 & 0.2 & 0.4 \\ 0.0 & 0.3 & 0.7 \end{bmatrix} = [\mathbf{0.1932} \quad \mathbf{0.2727} \quad \mathbf{0.5341}]$$

$$[\mathbf{0.1932} \quad \mathbf{0.2727} \quad \mathbf{0.5341}] \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.4 & 0.2 & 0.4 \\ 0.0 & 0.3 & 0.7 \end{bmatrix} = [\mathbf{0.20568} \quad \mathbf{0.27273} \quad \mathbf{0.52159}]$$

$$[\mathbf{0.20568} \quad \mathbf{0.27273} \quad \mathbf{0.52159}] \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.4 & 0.2 & 0.4 \\ 0.0 & 0.3 & 0.7 \end{bmatrix} =$$

$$[\mathbf{0.211932} \quad \mathbf{0.272727} \quad \mathbf{0.51534}]$$

$$[\mathbf{0.211932} \quad \mathbf{0.272727} \quad \mathbf{0.51534}] \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.4 & 0.2 & 0.4 \\ 0.0 & 0.3 & 0.7 \end{bmatrix} = [\mathbf{0.218} \quad \mathbf{0.273} \quad \mathbf{0.51}]$$

Since, **πA=π**

$\text{Therefore,} \boldsymbol{\pi} = [\mathbf{0.218} \quad \mathbf{0.273} \quad \mathbf{0.51}]$

P(Y = H, H, S ; X = E, Sc, E)

P(X_1= E) × P(Y_1= H | X_1 = E) × P(Y_2=H | $X_2 = Sc$) × P(Y_3= S | X_3 = E) × P(X_2 = Sc | X_1=E) × P(X_3 = E | X_2 = Sc)

- ⇨ 0.51× 0.8 x 0.4 x 0.2 x 0.3 x 0.4
- ⇨ 0.00391

Therefore, probability for the above sequence is **0.00391**

Q3. Explain sequence logos and compare them to sequence profiles. Where can they be used? (1 point)

Sequence logos are representation of conserved sequence along a series of multiple sequence alignment of genetically similar sequences in a single graphical form. It is created by a consensus sequence of all the given sequences. The higher the frequency of occurrence of a certain base pair in a given sequence alignment of DNA, RNA or protein, higher it will be the height of that particular base in the graphical representation. This way we can easily see which bases or amino acids are more common compared to others in an alignment.

On the other hand, sequence profiles give much more information about consensus sequences. This is due to information coming from occurrences. In addition to consensus sequence, sequence profiles also calculate the occurrence of amino acids or nucleotides for each of the position as exact numbers. This gives an estimate prediction of probability of which amino acid or nucleotide can occur on that position.

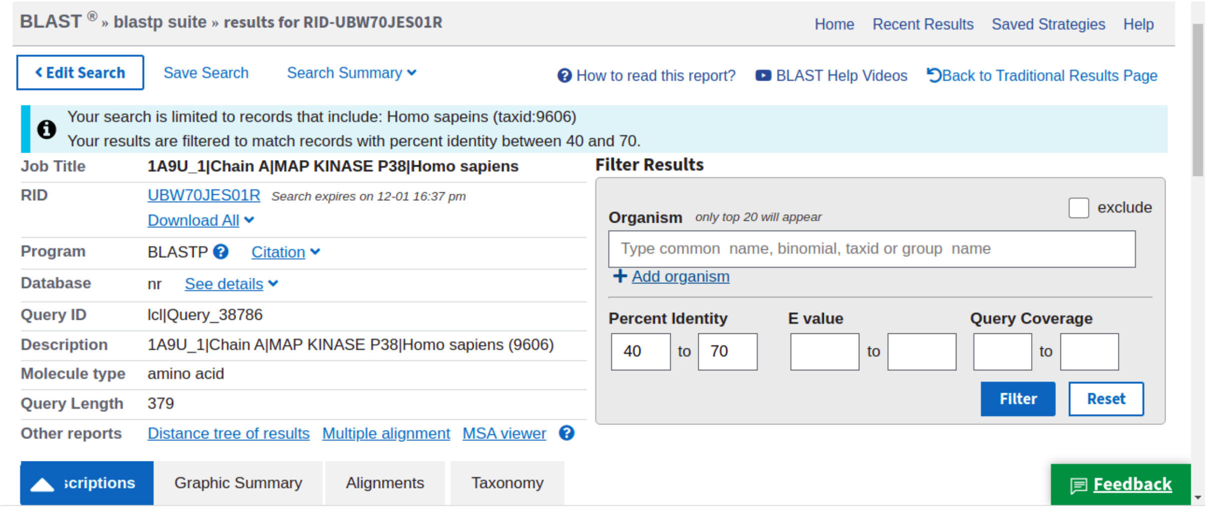
Both methods can be used any of the multiple sequence alignments. They both can make it easier for users to detect the abundance of gene or amino acids on particular positions.

Q4. What are the differences between the traditional artificial neural networks and deep learning? (1 point)

	Traditional Artificial neural network	**Deep learning**
1.	An artificial neural network imitates the actual human brain neurons and trains itself to give best possible output according to the real world trained based on assorted sets of algorithms designed to mimic how neurons perceive and react to sensory data	Deep learning imitates the data processing techniques of human brain just like ANN but it is rather trained based on representation of data which is unstructured unlike ANN.
2.	A neural network basically consists of 3 layers usually with 1 hidden layer along with 1 input layer and 1 output layer.	A Deep learning system can be referred to as an ANN with more than 3 layers including1 input layer and 1 output layer
3.	Traditional artificial neural networks are a subset of the field of Machine learning.	Deep learning is a subset of the field of Artificial neural networks in the area of Machine learning.
4.	It is applicable to detect objects and facial character recognition. Identification and classification of text recognition according to the relevant certain categories.	It can be in speech and textual recognition and categorization and for generation of HD videos depending om the observations on low quality image and footages which can further be used for development of high-quality images of historic data. It can also be sued for digital marketing by showing advertisements based on the surfer’s previous search history.

PROGRAMMING EXERCISES

1- 1A9U.pdb is used for this assignment.



Your search is limited to records that include: Homo sapiens (taxid:9606)
Your results are filtered to match records with percent identity between 40 and 70.

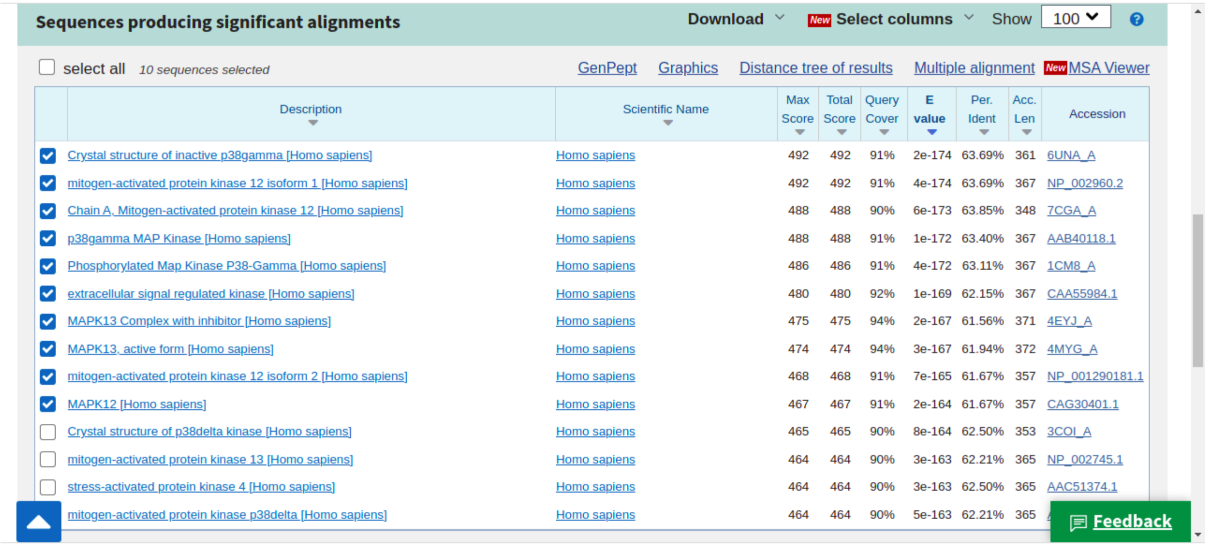
Job Title	1A9U_1|Chain A|MAP KINASE P38|Homo sapiens
RID	UBW70JES01R *Search expires on 12-01 16:37 pm* Download All
Program	BLASTP Citation
Database	nr See details
Query ID	lcl|Query_38786
Description	1A9U_1|Chain A|MAP KINASE P38|Homo sapiens (9606)
Molecule type	amino acid
Query Length	379
Other reports	Distance tree of results Multiple alignment MSA viewer

Filter Results

Organism *only top 20 will appear* □ exclude
Type common name, binomial, taxid or group name
+ Add organism

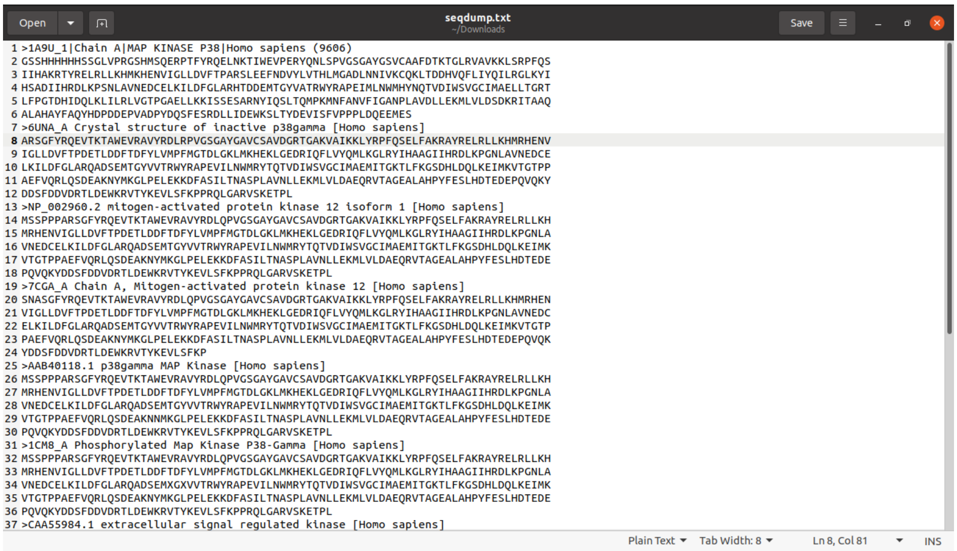
Percent Identity 40 to 70
E value to
Query Coverage to
Filter Reset

10 proteins with 40 to 70 percent identity is chosen for multiple alignment.



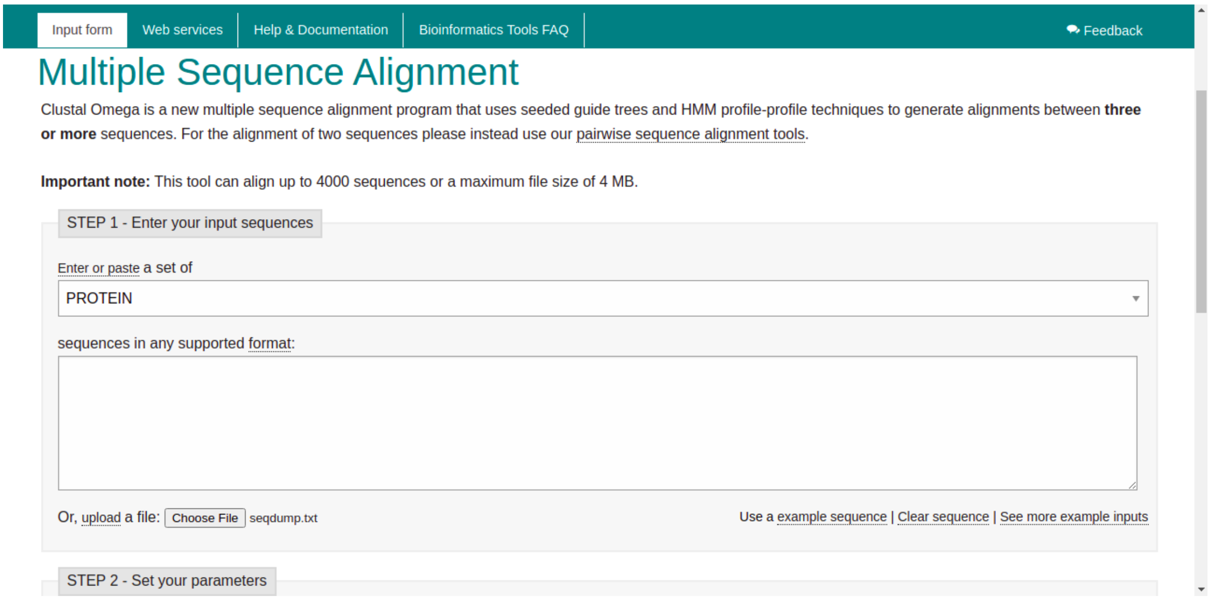
	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Crystal structure of inactive p38gamma [Homo sapiens]	Homo sapiens	492	492	91%	2e-174	63.69%	361	6UNA_A
☑	mitogen-activated protein kinase 12 isoform 1 [Homo sapiens]	Homo sapiens	492	492	91%	4e-174	63.69%	367	NP_002960.2
☑	Chain A, Mitogen-activated protein kinase 12 [Homo sapiens]	Homo sapiens	488	488	90%	6e-173	63.85%	348	7CGA_A
☑	p38gamma MAP Kinase [Homo sapiens]	Homo sapiens	488	488	91%	1e-172	63.40%	367	AAB40118.1
☑	Phosphorylated Map Kinase P38-Gamma [Homo sapiens]	Homo sapiens	486	486	91%	4e-172	63.11%	367	1CM8_A
☑	extracellular signal regulated kinase [Homo sapiens]	Homo sapiens	480	480	92%	1e-169	62.15%	367	CAA55984.1
☑	MAPK13 Complex with inhibitor [Homo sapiens]	Homo sapiens	475	475	94%	2e-167	61.56%	371	4EYJ_A
☑	MAPK13, active form [Homo sapiens]	Homo sapiens	474	474	94%	3e-167	61.94%	372	4MYG_A
☑	mitogen-activated protein kinase 12 isoform 2 [Homo sapiens]	Homo sapiens	468	468	91%	7e-165	61.67%	357	NP_001290181.1
☑	MAPK12 [Homo sapiens]	Homo sapiens	467	467	91%	2e-164	61.67%	357	CAG30401.1
□	Crystal structure of p38delta kinase [Homo sapiens]	Homo sapiens	465	465	90%	8e-164	62.50%	353	3COI_A
□	mitogen-activated protein kinase 13 [Homo sapiens]	Homo sapiens	464	464	90%	3e-163	62.21%	365	NP_002745.1
□	stress-activated protein kinase 4 [Homo sapiens]	Homo sapiens	464	464	90%	3e-163	62.50%	365	AAC51374.1
	mitogen-activated protein kinase p38delta [Homo sapiens]	Homo sapiens	464	464	90%	5e-163	62.21%	365	A

2- The file is downloaded as SeqDump.txt and FASTA of original .pdb is added on top of it.



```
>1A9U_1|Chain A|MAP KINASE P38|Homo sapiens (9606)
GSSHHHHHHSSGLVPRGSHMSQERPTFYRQELNKTIWEVPERYQNLSPVGSGAYGSVCAAFDTKTGLRVAVKKLSRPFQS
IIHAKRTYRELRLLKHMKHENVIGLLDVFTPARSLEEFNDVYLVTHLMGADLNNIVKCQKLTDDHVQFLIYQILRGLKYI
HSADIIHRDLKPSNLAVNEDCELKILDFGLARHTDDEMTGYVATRWYRAPEIMLNWMHYNQTVDIWSVGCIMAELLTGRT
LFPGTDHIDQLKLILRLVGTPGAELLKKISSESARNYIQSLTQMPKMNFANVFIGANPLAVDLLEKMLVLDSDKRITAAQ
ALAHAYFAQYHDPDDEPVADPYDQSFESRDLLIDEWKSLTYDEVISFVPPPLDQEEMES
>6UNA_A Crystal structure of inactive p38gamma [Homo sapiens]
ARSGFYRQEVTKTAWEVRAVYRDLRPVGSGAYGAVCSAVDGRTGAKVAIKKLYRPFQSELFAKRAYRELRLLKHMRHENV
IGLLDVFTPDETLDDFTDFYLVMPFMGTDLGKLMKHEKLGEDRIQFLVYQMLKGLRYIHAAGIIHRDLKPGNLAVNEDCE
LKILDFGLARQADSEMTGYVVTRWYRAPEVILNWMRYTQTVDIWSVGCIMAEMITGKTLFKGSDHLDQLKEIMKVTGTPP
AEFVQRLQSDEAKNYMKGLPELEKKDFASILTNASPLAVNLLEKMLVLDAEQRVTAGEALAHPYFESLHDTEDEPQVQKY
DDSFDDVDRTLDEWKRVTYKEVLSFKPPRQLGARVSKETPL
>NP_002960.2 mitogen-activated protein kinase 12 isoform 1 [Homo sapiens]
MSSPPPARSGFYRQEVTKTAWEVRAVYRDLQPVGSGAYGAVCSAVDGRTGAKVAIKKLYRPFQSELFAKRAYRELRLLKH
MRHENVIGLLDVFTPDETLDDFTDFYLVMPFMGTDLGKLMKHEKLGEDRIQFLVYQMLKGLRYIHAAGIIHRDLKPGNLA
VNEDCELKILDFGLARQADSEMTGYVVTRWYRAPEVILNWMRYTQTVDIWSVGCIMAEMITGKTLFKGSDHLDQLKEIMK
VTGTPPAEFVQRLQSDEAKNYMKGLPELEKKDFASILTNASPLAVNLLEKMLVLDAEQRVTAGEALAHPYFESLHDTEDE
PQVQKYDDSFDDVDRTLDEWKRVTYKEVLSFKPPRQLGARVSKETPL
>7CGA_A Chain A, Mitogen-activated protein kinase 12 [Homo sapiens]
SNASGFYRQEVTKTAWEVRAVYRDLQPVGSGAYGAVCSAVDGRTGAKVAIKKLYRPFQSELFAKRAYRELRLLKHMRHEN
VIGLLDVFTPDETLDDFTDFYLVMPFMGTDLGKLMKHEKLGEDRIQFLVYQMLKGLRYIHAAGIIHRDLKPGNLAVNEDC
ELKILDFGLARQADSEMTGYVVTRWYRAPEVILNWMRYTQTVDIWSVGCIMAEMITGKTLFKGSDHLDQLKEIMKVTGTP
PAEFVQRLQSDEAKNYMKGLPELEKKDFASILTNASPLAVNLLEKMLVLDAEQRVTAGEALAHPYFESLHDTEDEPQVQK
YDDSFDDVDRTLDEWKRVTYKEVLSFKP
>AAB40118.1 p38gamma MAP Kinase [Homo sapiens]
MSSPPPARSGFYRQEVTKTAWEVRAVYRDLQPVGSGAYGAVCSAVDGRTGAKVAIKKLYRPFQSELFAKRAYRELRLLKH
MRHENVIGLLDVFTPDETLDDFTDFYLVMPFMGTDLGKLMKHEKLGEDRIQFLVYQMLKGLRYIHAAGIIHRDLKPGNLA
VNEDCELKILDFGLARQADSEMTGYVVTRWYRAPEVILNWMRYTQTVDIWSVGCIMAEMITGKTLFKGSDHLDQLKEIMK
VTGTPPAEFVQRLQSDEAKNNMKGLPELEKKDFASILTNASPLAVNLLEKMLVLDAEQRVTAGEALAHPYFESLHDTEDE
PQVQKYDDSFDDVDRTLDEWKRVTYKEVLSFKPPRQLGARVSKETPL
>1CM8_A Phosphorylated Map Kinase P38-Gamma [Homo sapiens]
MSSPPPARSGFYRQEVTKTAWEVRAVYRDLQPVGSGAYGAVCSAVDGRTGAKVAIKKLYRPFQSELFAKRAYRELRLLKH
MRHENVIGLLDVFTPDETLDDFTDFYLVMPFMGTDLGKLMKHEKLGEDRIQFLVYQMLKGLRYIHAAGIIHRDLKPGNLA
VNEDCELKILDFGLARQADSEMXGXVVTRWYRAPEVILNWMRYTQTVDIWSVGCIMAEMITGKTLFKGSDHLDQLKEIMK
VTGTPPAEFVQRLQSDEAKNYMKGLPELEKKDFASILTNASPLAVNLLEKMLVLDAEQRVTAGEALAHPYFESLHDTEDE
PQVQKYDDSFDDVDRTLDEWKRVTYKEVLSFKPPRQLGARVSKETPL
>CAA55984.1 extracellular signal regulated kinase [Homo sapiens]
```

Input is selected for multiple sequence alignment.



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

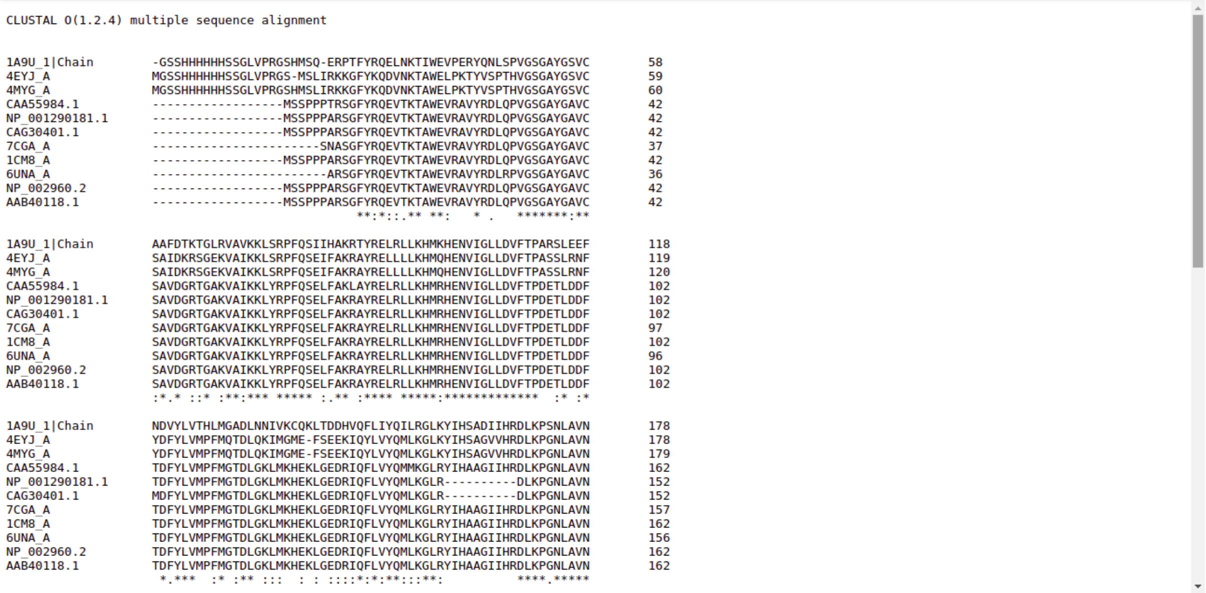
sequences in any supported format:

Or, upload a file: Choose File seqdump.txt

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

STEP 2 - Set your parameters

Result of alignment.



```
CLUSTAL O(1.2.4) multiple sequence alignment


1A9U_1|Chain        -GSSHHHHHHSSGLVPRGSHMSQ-ERPTFYRQELNKTIWEVPERYQNLSPVGSGAYGSVC        58
4EYJ_A              MGSSHHHHHHSSGLVPRGS-MSLIRKKGFYKQDVNKTAWELPKTYVSPTHVGSGAYGSVC        59
4MYG_A              MGSSHHHHHHSSGLVPRGSHMSLIRKKGFYKQDVNKTAWELPKTYVSPTHVGSGAYGSVC        60
CAA55984.1          ------------------MSSPPPTRSGFYRQEVTKTAWEVRAVYRDLQPVGSGAYGAVC        42
NP_001290181.1      ------------------MSSPPPARSGFYRQEVTKTAWEVRAVYRDLQPVGSGAYGAVC        42
CAG30401.1          ------------------MSSPPPARSGFYRQEVTKTAWEVRAVYRDLQPVGSGAYGAVC        42
7CGA_A              -----------------------SNASGFYRQEVTKTAWEVRAVYRDLQPVGSGAYGAVC        37
1CM8_A              ------------------MSSPPPARSGFYRQEVTKTAWEVRAVYRDLQPVGSGAYGAVC        42
6UNA_A              ------------------------ARSGFYRQEVTKTAWEVRAVYRDLRPVGSGAYGAVC        36
NP_002960.2         ------------------MSSPPPARSGFYRQEVTKTAWEVRAVYRDLQPVGSGAYGAVC        42
AAB40118.1          ------------------MSSPPPARSGFYRQEVTKTAWEVRAVYRDLQPVGSGAYGAVC        42
                                                **:*::.** **:   * .   *******:**

1A9U_1|Chain        AAFDTKTGLRVAVKKLSRPFQSIIHAKRTYRELRLLKHMKHENVIGLLDVFTPARSLEEF        118
4EYJ_A              SAIDKRSGEKVAIKKLSRPFQSEIFAKRAYRELLLLKHMQHENVIGLLDVFTPASSLRNF        119
4MYG_A              SAIDKRSGEKVAIKKLSRPFQSEIFAKRAYRELLLLKHMQHENVIGLLDVFTPASSLRNF        120
CAA55984.1          SAVDGRTGAKVAIKKLYRPFQSELFAKLAYRELRLLKHMRHENVIGLLDVFTPDETLDDF        102
NP_001290181.1      SAVDGRTGAKVAIKKLYRPFQSELFAKRAYRELRLLKHMRHENVIGLLDVFTPDETLDDF        102
CAG30401.1          SAVDGRTGAKVAIKKLYRPFQSELFAKRAYRELRLLKHMRHENVIGLLDVFTPDETLDDF        102
7CGA_A              SAVDGRTGAKVAIKKLYRPFQSELFAKRAYRELRLLKHMRHENVIGLLDVFTPDETLDDF        97
1CM8_A              SAVDGRTGAKVAIKKLYRPFQSELFAKRAYRELRLLKHMRHENVIGLLDVFTPDETLDDF        102
6UNA_A              SAVDGRTGAKVAIKKLYRPFQSELFAKRAYRELRLLKHMRHENVIGLLDVFTPDETLDDF        96
NP_002960.2         SAVDGRTGAKVAIKKLYRPFQSELFAKRAYRELRLLKHMRHENVIGLLDVFTPDETLDDF        102
AAB40118.1          SAVDGRTGAKVAIKKLYRPFQSELFAKRAYRELRLLKHMRHENVIGLLDVFTPDETLDDF        102
                    :*.* ::* :**:*** ***** :.** :**** *****:*************  :* :*

1A9U_1|Chain        NDVYLVTHLMGADLNNIVKCQKLTDDHVQFLIYQILRGLKYIHSADIIHRDLKPSNLAVN        178
4EYJ_A              YDFYLVMPFMQTDLQKIMGME-FSEEKIQYLVYQMLKGLKYIHSAGVVHRDLKPGNLAVN        178
4MYG_A              YDFYLVMPFMQTDLQKIMGME-FSEEKIQYLVYQMLKGLKYIHSAGVVHRDLKPGNLAVN        179
CAA55984.1          TDFYLVMPFMGTDLGKLMKHEKLGEDRIQFLVYQMMKGLRYIHAAGIIHRDLKPGNLAVN        162
NP_001290181.1      TDFYLVMPFMGTDLGKLMKHEKLGEDRIQFLVYQMLKGLR----------DLKPGNLAVN        152
CAG30401.1          MDFYLVMPFMGTDLGKLMKHEKLGEDRIQFLVYQMLKGLR----------DLKPGNLAVN        152
7CGA_A              TDFYLVMPFMGTDLGKLMKHEKLGEDRIQFLVYQMLKGLRYIHAAGIIHRDLKPGNLAVN        157
1CM8_A              TDFYLVMPFMGTDLGKLMKHEKLGEDRIQFLVYQMLKGLRYIHAAGIIHRDLKPGNLAVN        162
6UNA_A              TDFYLVMPFMGTDLGKLMKHEKLGEDRIQFLVYQMLKGLRYIHAAGIIHRDLKPGNLAVN        156
NP_002960.2         TDFYLVMPFMGTDLGKLMKHEKLGEDRIQFLVYQMLKGLRYIHAAGIIHRDLKPGNLAVN        162
AAB40118.1          TDFYLVMPFMGTDLGKLMKHEKLGEDRIQFLVYQMLKGLRYIHAAGIIHRDLKPGNLAVN        162
                     *.***  :* :** :::  : : ::::*:*:**:::**:          ****.*****
```

3-

```
# Panda is called for DataFrame
import pandas as pd
# Bio and Bio.Align is specifically used for consensus calculations
from Bio import AlignIO as al
from Bio.Align import AlignInfo as alI

# Alignment is called with the format "clustal"
alignment = al.read("clustalo-E20211130-094437-0140-87358764-
p2m.clustal_num", "clustal")
```

```
# Summary info is assigned
align = alI.SummaryInfo(alignment)
# Dumb_consensus function of the AlignInfo module is used to generate
consensus
consensus = align.dumb_consensus()
print("Consensus of the clustal is:" + "\n" + consensus)
```

Gives the output as:

```
Consensus of the clustal is:

MGSSHHHHHHSSGLVPRGXSXXXXXRSGFYRQEVTKTAWEVRAVYRDLXPVGSGA
YGAVCSAVDGRTGAKVAIKKLYRPFQSELFAKRAYRELRLLKHMRHENVIGLLDVF
TPDETLDDFXDFYLVMPFMGTDLGKLMKHEKLGEDRIQFLVYQMLKGLRYIHXAGI
IHRDLKPGNLAVNEDCELKILDFGLARQADSEMTGYVVTRWYRAPEVILNWMXYT
QTVDIWSVGCIMAEMITGKTLFKGSDHLDQLKEIMKVTGTPPAEFVQRLQSDEAKN
YMKGLPELEKKDFASILTNASPLAVNLLEKMLVLDAXXRXTAGEALAHPYFESLHD
TEDEPQAVQKYDDSFDXXDRTLDEWKRVTYKEVLSFKPPRQLGARVSKETPL
```