Experiment - 2

Multiple sequence alignment and pattern determination using PROSITE

a) Multiple sequence alignment :-

Aim :- To carry out multiple sequence alignment MSA of given sequence from various organisms using CLUSTAL W2/ CLUSTAL X.

Introduction :-

The clustal programs are widely used for carrying out automatic multiple alignment of nucleotide or amino acid sequences. The most familiar version is clustal W, which uses a simple text menu system that is portable to more or use all computer systems clustal x features a graphical user interface and some powerful graphical utilities for adding the interpretation of alignment and is the preferred version for interactive usage. Users may run clustal remotely from several sites using the web or the programs may be downloaded and run locally on PC's, maintain or unix computers. The protocols discuss how to use clustal x and clustal W to construct an alignment and create profile alignment by merging existing alignments.

Query:- Protein uridyltransferase protein Tumor protein :p53 (Papain)

URL : www.ebi.ac.uk/CLUSTAL w/

Procedure:-

→ Multiple sequence alignment is done with the help of CLUSTAL W tool available at EBI database.

→ Login to NCBI and retrieve more than two sequence of protein / nucleotide in FASTA format.

→ Login to www.ebi.ac.uk/CLUSTAL w/ and click on clustal omega < launch clustal omega>.

Nucleotide:- Paste the retrieved sequence in sequence box and observe the results by clicking on the submit option

protein:- Paste the retrieved amino acid sequence in the sequence box and click on submit option to order to observe results.

RESULT and Discussion:-

Multiple sequence alignment is done with the help of C

CLUSTAL -w tool and following results were obtained

NUCLEOTIDE SEQUENCE:-

The nucleotide sequence for the given query "uridyl transferase" is obtained as follows

-TCAGG TAA CGACC - - - - GAP - - - - TG

GAA AGG TTA AAG CGCTGG TATA TTG

GCCCG GTG AATCCC A AA CATCTGAG

* *   *   *

└→ Match   └→ Mismatch

B) PATTERN DETERMINATION USING PROSITE TOOL :-

Query :- protein - Elastin
URL : www. expasy. ch/prosite/
Procedure :-

→ Login to NICBI and retrieve more than two protein sequences in FASTA format login to www. expasy. ch/prosite/

→ PROSITE webpage will be opened, option called PRATT will be present, click on it to enter the tool and paste the sequences in the sequence box and click on RUN PRATT option.

→ pattern will be displayed (phase 1 results will be displayed at the top most position within the box.

→ In the result page of PROSITE select graphical view the output format. Below that click on the start scan button the results will be in the graphical view, represents the presence of same pattern in all displayed hits.

Results and Discussion

Pattern search is done for the given amino acid query "Elastin" with the help of PRATT. which is present in PROSITE and inturn present in EXPASY.

In the first phase of results following pattern is obtained
{ P-G-V-P-P-G-G-x(0,1),-G-V-x[AP]-K-A-P-G-[GV]-G-G-A-F-A-G-V-G-P-F-G-G[PA]-G-P-G-V-P-L-G-Y-P-G-G }

[ ] → indicate the list of acceptance aminoacid for given position
X → it is used to indicates any amino acid is accepted in position
x (0,1) → any amino acid present 1 or 2 times in particular position .

In second phase results "12 hits is obtained".

Elastin (Homosapiens) _____ (860.a)

Elastin (Mus musculus) _____ (1300.a)

Elastin (SUS scrofa) _____ (759 a.a)

P - 10
C - 10
V - 10
R - 20
—————
50

MAG TPA: Papain [Myoviridae sp. ct5lt7]

GenBank: DAJ21293.1

GenPept Identical Proteins Graphics

```
>DAJ21293.1 MAG TPA: Papain [Myoviridae sp. ct5lt7]
MKRFYGYIKTDNLESYKEYNPEVCELPKEYKLKDLGEVWDQGSKGSCVSCSISEMYNFYQISHGRKLDIP
YTYVYDKRSDKSLDGMQPAEGFKILSDEGRIKMFSRISTLDSLKQSIALANGPALIAMTVRDTDRTDFWNG
PTAQNIGHAVAVVGYNQDSLIIKNSWGYSYGEYGYWYLPFNDFQVVKEAWTIIK
```

papain [synthetic construct]

GenBank: AAA72774.1

GenPept Identical Proteins Graphics

```
>AAA72774.1 papain [synthetic construct]
MAMIPSISKLLFVAICLFVYMGLSFGDFSIVGYSQNDLTSTERLIQLFESWMLKHNKIYKNIDEKIYRFE
IFKDNLKYIDETNKKNNSYWLGLNVFADMSNDEFKEKYTGSIAGNYTTTELSYEEVLNDGDVNIPEYVDW
RQKGAVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTGNLNEYSEQELLDCDRRSYGCNGGYPWSALQLVAQ
YGIHYRNTYPYEGVQRYCRSREKGPYAAKTDGVRQVQPYNEGALLYSIANQPVSVVLEAAGKDFQLYRGG
IFVGPCGNKVDHAVAAVGYGPNYILIKNSWGTGWGENGYIRIKRGTGNSYGVCGLYTSSFYPVKN
```

papain [Nanobdella aerobiophila]

GenBank: BBL45284.1

GenPept Identical Proteins Graphics

```
>BBL45284.1 papain [Nanobdella aerobiophila]
MKRVQINLLNVGIILVIVGLAIGIYYIYIYNLKTSQQNIEFQNFVQNIQNLENTITTFQGNYFIRLSVFN
DVIIYYANNSIVLNYNNNNYSITFNNYNISLYNNNSEINYLLYPNEIYFIKSNNHIYITNNYNESEYLID
NNTYNMIINSNNSNNIISSNDCLYNYTGQTSWNWGDVNGDSYLPPIGNQGTCGDCWAWSAAYAIASNYMI
YNNLPNLSLYLSPTEIAVNCNVGNPNSPGYCQDVMGCLGGLPYQGIQFSEEYGIGTDNNYNTYYQELADC
STNISNYPGDVCGQSYSIPNNNIYVPDSEQYLASGFLDSEEAIGPGGQLAALPPSGDLSTQQIKQDLLCY
GPLSISGFLGGLSNQDTSYYLEATGHATLLVGFNDNSNICQEEYNQPGCWIIQNEWGQQSACIIVVYNTT
NYYDTTPLLIPNNMISIQSCSQLFLNQNFCNYYVQKVLGSGYNCGGVYWMMNGYIYVPYNSTPYGQSLPY
YIISINNVTYIS
```

Papain, partial [Noccaea caerulescens]

GenBank: JAU21028.1

GenPept Identical Proteins Graphics

```
>JAU21028.1 Papain, partial [Noccaea caerulescens]
```

```
AGITKLLVVVLAALSAAAVYRHVSRVSEKPSSGIPREVIALYSTWKNAHGKLYSTPAENDHRLRVFYSQK
QFVDKANVDYASAMLKRDGTVITENPFTLNGFADLAADEFTAKYTGEAPETHENAELAPEVNYVESADAG
LAQAGFDVRVRNQGSCGSCWAFSAVATYEKFWFLKTGQRLDLSQQQLVDCDTSDNGCNG
```

hypothetical protein L6164_015316 [Bauhinia variegata]

GenBank: KAI4336836.1

GenPept Identical Proteins Graphics

```
>KAI4336836.1 hypothetical protein L6164_015316 [Bauhinia variegata]
MSMTGFSSFSFSLTFLLLFTSSALSSCSDTSELFESWCKQHDKNYSSQEEKLYRFKVFEDNYDFVTRHNE
IGNSAYTLSLNAFADLTHHEFKASRLGLSPALLRFDHRLLQESDDLLLDIPSEIDWRKKGAVTQVKDQGN
CGACWSFSATGAIEGINKIFTGSLVSLSEQELVDCDRTYNNGCDGGLMDYAYQFVIDNNGIDTEEDYPYQ
GHQSTCNKGKLRRRVVTIDGYIDVPPNNENQLLKAAANQPVSVGICGSDRAFQLYSKGIFTGPCSTSLDH
AVLIVGYDSENGVDYWIVKNSWGTGWGMNGYIHMLRNNGDSDGLCGINRLASYPMKTIPNPPSPSPPSPT
RCDLFTYCSEGETCCCARRILGFCISWKCCGLASAVCCKDHHHCCPQDYPVCDTTRGQCLKRTANATRIQ
ALEKEGRSVIG
```

hypothetical protein L6164_036901 [Bauhinia variegata]

GenBank: KAI4296988.1

GenPept Identical Proteins Graphics

```
>KAI4296988.1 hypothetical protein L6164_036901 [Bauhinia variegata]
MAVTLAMKLIFAICIIFCTSAFQAMSRTLNAATIAAAHEQWMAKQGRTYADDAEKQKRFKIFAENLQHIE
SVNNAGNKSYKLGLNKFSDLTAQEFIASYTGLKISNLPKPSKVASFGPLNLDDVPTRLNWTEKGAVTPVK
DQGKCGSCWAFSTVAAVEGITQIKKGTLPSLSEQQLVDCVTDNHGCNGGVMDYAFKYIIDNQGIANENDY
PYQAKQGSCKSVSSTAATITGFNDVPDNEEQLLQAVTNQPVSVAISVCDDFYQYQEGVYSENCYRQLNHA
VTVVGYGTTDDGKKYWLIKNSWSESWGDNGYIKLWRDSGVPGGLIGIAKKASYPTI
```

hypothetical protein L6164_036818 [Bauhinia variegata]

GenBank: KAI4296898.1

GenPept Identical Proteins Graphics

```
>KAI4296898.1 hypothetical protein L6164_036818 [Bauhinia variegata]
MVVTVDMKHLAAICFALWACAFPVMSRTLNEDSVAATHEQWMTKFGRSYANEAEKAKRFKIFMENLLYIE
NFNNAGNQSYKMGLNKFSDLTKEEFLAFYTRPFKVSNQTNSSKMTSFFKQLDVSDVPESMDWIEKGAVTPI
KNQQSCGSCWAFATVAAVEGIVQITTGDLIPLSEQQLVDCDTTENNGCKGGLMDNGFEYIKENNGIASETD
YPYLGDSTGKCYSNKAAKHAAHITDYVVVKGEDQLLQAVAKQPVAIRVAAGDHFHQYGNGIFSGPCGSKL
NHEVTVVGYGTSEDGTKYWLMKNSWGEDWGEKGYIRMKMNVGDEGLCGIATHASYPTINA
```

papain

PRF: 1303270A

GenPept Identical Proteins Graphics

```
>prf||1303270A papain
```

papain

```
>prf||1303270A papain
```

```
MAMIPSISKLLFVAICLFVYMGLSFGDFSIVGYSQNDLTSTERLIQLFESWMLKHNKIYKNIDEKIYRFE
IFKDNLKYIDETNKKNNSYWLGLNVFADMSNDEFKEKYTGSIAGNYTTTELSYEEVLNDGDVNIPEYVDW
RQKGAVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTGNLNEYSEQELLDCDRRSYGCNGGYPWSALQLVAQ
YGIHYRNTYPYEGVQRYCRSREKGPYAAKTDGVRQVQPYNEGALLYSIANQPVSVVLEAAGKDFQLYRGG
IFVGPCGNKVDHAVAAVGYGPNYILIKNSWGTGWGENGYIRIKRGTGNSYGVCGLYTSSFYPVKN
```

Chain C, Papain

PDB: 3TNX_C

```
>pdb|3TNX|C Chain C, Papain
MHHHHHHSSGLVPRGSGMKETAAAKFERQHMDSPDLGTDDDDKMDFSIVGYSQNDLTSTERLIQLFESWM
LKHNKIYKNIDEKIYRFEIFKDNLKYIDETNKKNNSYWLGLNVFADMSNDEFKEKYTGSIAGNYTTTELS
YEEVLNDGDVNIPEYVDWRQKGAVTPVKNQGSCGSAWAFSAVSTIESIIKIRTGNLNEYSEQELLDCDRR
SYGCNGGYPWSALQLVAQYGIHYRNTYPYEGVQRYCRSREKGPYAAKTDGVRQVQPYNEGALLYSIANQP
VSVVLEAAGKDFQLYRGGIFVGPCGNKVDHAVAAVGYGPNYILIRNSWGTGWGENGYIRIKRGTGNSYGV
CGLYTSSFYPVKN
```

peptidase C1A papain [Haliscomenobacter hydrossis DSM 1100]

GenBank: AEE49325.1

```
>AEE49325.1 peptidase C1A papain [Haliscomenobacter hydrossis DSM 1100]
MKTPNVSTFSILKTLSFFALLLSACGIQAQVQVFQIYINREIQAPQNIKIELQQQRQFINNQKLFFAVGF
TSVSGKQIQQITGERELSFQQLQAVNNYLRVKQFTIDPGVLACNAGMAAYDARSKGQVTPVRDQGGCGSC
WAFGTLAAYETNYLKVNGGAANALNLSEQQILSCSGGGDCGGGMSNIVYQWWVDGNNSVVTEATYPYSSS
ATACQPKPATNYKAEAWGLADPAATYAGIASVAKIKEAICKYGSVASSVMVDGLFQNYVAGVFFGFASNY
GAPSSNHVVQIIGWDDAKGAWLIKNSWGTDWGMAGLMWIKYNANNIGRRAVWVKAKAQLVLVNPGIVVNP
GIAVNPGVIVAASDNWVNADPNTSSLTKINFLNAGANLNTWGKCSPQDCDWGKAKVDPLVGNPNYTHCAI
YNQGFVLRTIYFKKVGTTMNVFMVSKFNDARGTQNSAMTFKKS
```

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:

```
>DA/21293.1 MAG TPA  Papain [Mycovester sp. p5b7]
MKRFIGYIKTIALFSYKEYNFEVCELPKEYKLADLGEVWDQGSKGSCVSCSISEMYNFYQSHGRKLDIP
YTYVYDKRSOKSLDGMGPAEGFKLSDEGRKMFSRISTLDSLKGSILANGPALIAMTVRDTDRTDFWNG
PTAQNGHAVAVSYNGDSLIKNSWGYSYGEYGYWYLPFNDFQVVKEAWTIIt.
>AAA72774.1 papain [synthetic construct]
MAMIPSISKLLFVAICLFVYMGLSFGDFSIVGYSQNDLTSTERLIQLFESWMLKHNKIYKNIDEKIYRFE
IFKDNLKYIDETNKKNNSYWLGLNVFADMSNDEFKEKYTGSIAGNYTTTELSYEEVLNDGDVNIPEYVDW
```

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

ClustalW with character counts

The default settings will fulfil 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)

[Submit]

Multiple Sequence Alignment

Tools > Multiple Sequence Alignment

Service Announcement

We will be retiring these pages on the 31ˢᵗ of January, 2024. Please visit our new website at www.ebi.ac.uk/jdispatcher and share your feedback with us!

Multiple Sequence Alignment (MSA) is generally the alignment of three or more biological sequences (protein or nucleic acid) of similar length. From the output, homology can be inferred and the evolutionary relationships between the sequences studied

By contrast, Pairwise Sequence Alignment tools are used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences

Clustal Omega

New MSA tool that uses seeded guide trees and HMM profile-profile techniques to generate alignments. Suitable for medium-large alignments

↘Launch Clustal Omega

```
AEA45718.1     --IPESYFPNPGSQGQIGSCVAWSTTYALASFYFAAKNKWGSPKTTNMIMSPAFVYNQI- 111
BBL45284.1     DVNGDSYLPPIGNQGTCGDCWAWSAAYAIASNYMIYNNLPN----LSLYLS----PTEIA 227
AEE49325.1     S---KGQVTPVRDQGGCGSCWAFGTLAAYETNYLKVNGGAA----NALNLS----EQQI- 171
DAJ21293.1     D---LGE---VWDQGSKGSCVSCS-ISEMYNFYQISHG-------RKLDIP----YTYV- 74
JAU21028.1     Q---AGFDVRVRNQGSCGSCWAFSAVATYEKFWFLKTG-------QRLDLS----QQQL- 187
AAA72774.1     Q---KGAVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTG-------NLNEYS----EQEL- 186
prf||1303270A  Q---KGAVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTG-------NLNEYS----EQEL- 186
pdb|3TNX|C     Q---KGAVTPVKNQGSCGSAWAFSAVSTIESIIKIRTG-------NLNEYS----EQEL- 204
KAI4336836.1   E---KGAVTPVKDQGKCGSCWAFSTVAAVEGITQIKKG-------TLPSLS----EQQL- 647
KAI4296898.1   E---KGAVTPIKNQQSCGSCWAFATVAAVEGIVQITTG-------DLIPLS----EQQL- 177
                   .        .*   *.. :.                              :

AEA45718.1     RECT-----------CGPNCGT-YVHDALNLLK--TKGVVTWQTMPYDNAN--SCSVPS- 154
BBL45284.1     VNCNVGNPNSPGYCQDVMGCLGGLPYQGIQFSE--EYGIGTDNNYNTYYQELADCSTNI- 284
AEE49325.1     LSC-----------SGGGDCGGGMSNIVYQMWVDGNNSVVTEATYPYSSSA-TACQPKPA 219
DAJ21293.1     YDK-----------RSDKSLDGMQPAEGFKILSD-EGRIKNFS---------------- 105
JAU21028.1     VDCD-----------TSDNGCNG------------------------------------ 199
AAA72774.1     LDCD-----------RRSYGCNGGYPWSALQ-LVA-QYGIHYRNTYPYEGVQ-RYCRSREK 233
prf||1303270A  LDCD-----------RRSYGCNGGYPWSALQ-LVA-QYGIHYRNTYPYEGVQ-RYCRSREK 233
pdb|3TNX|C     LDCD-----------RRSYGCNGGYPWSALQ-LVA-QYGIHYRNTYPYEGVQ-RYCRSREK 251
KAI4336836.1   VDCV-----------TDNHGCNGGVMDYAFKYIID-NQGIANENDYPYQAKQ-GSCKSVS- 694
KAI4296898.1   LDCT-----------TENNGCKGGLMDNGFEYIKE-NNGIASETDYPYLGDSTGKCYSNKA 226
```