Genome Annotation

MODULE 2

PART -II

Rama Dey-Rao, PhD

Clinical Assistant Professor

Biotechnical and Clinical Lab Sciences

Senior Scientist

Department of Microbiology & Immunology, SUNY at Buffalo

dey@buffalo.edu

Sequence-based Similarity

4 TOOLS

1. BLAST	The Basic Local Alignment Search Tool (**BLAST**) tool finds regions of local similarity between sequences and calculates the statistical significance of matches
2. CDD	Conserved Domain Database Search (**CDD**) finds sequence similarity with genes in conserved orthologous groups (COGs).
3. T-Coffee	Tree based Consistency Objective Function For alignment Evaluation (T-Coffee) is a multiple sequence alignment program that aligns a set of homologous (similar) sequences
4. WebLogo	WebLogo is a program that enables easy creation of sequence logos from the multiple sequence alignments

Conserved Domain Database Search

CDD

click on the CDD search results at the top of the BLAST results page

COG number (*top hit*)

```
save for annotation report
```

COG name

```
save for annotation report
```

Score

```
score
```

E-value

```
save for annotation report
```

Significant COG number (*second hit*)

```
save for annotation report
```

COG name

```
save for annotation report
```

Score

```
score
```

E-value

```
save for annotation report
```

T-Coffee

go to http://www.ebi.ac.uk/Tools/msa/tcoffee/

Sequences used for alignment

```
sequences
```

Multiple sequence alignment

```
alignment
```

WebLogo

go to http://weblogo.berkeley.edu/

Sequence logo

Protein Domains
COG (Clusters of Orthologous Groups)
Conserved Domain Database

- Sequence –based domains in proteins have a particular structure that is related to function.
- Can be seen as building blocks put together in different ways in different proteins.
- Parts of the proteins with similar and vital functions are conserved –clusters of ortholog groups in conserved domain database.
- When a very significant COG hit is observed for the query gene it can be interpreted as a strong likelihood that the protein has the same function.

Bauer et al. CDD: Nucleic Acids Res. 2011 Jan;39 (Database issue):D225-9. Epub 2010 Nov 24.

BLAST RESULTS PAGE- Swissprot database

Both a Conserved Domain Database
and BLAST searches are done simultaneously.

RID 631JYCW7016 (Expires on 02-12 21:27 pm)

Query ID lcl|Query_44273
Description KSED_RS00005- Ksed_00010-aa sequence
Molecule type amino acid
Query Length 506

Database Name swissprot
Description Non-redundant UniProtKB/SwissProt sequences
Program BLASTP 2.8.1+ ▹Citation

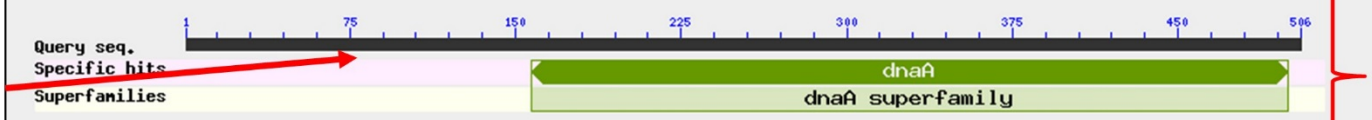
Other reports: ▹Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment] [MSA viewer]

New Analyze your query with SmartBLAST

Graphic Summary

Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.



Distribution of the top 106 Blast Hits on 100 subject sequences

CDD search (conserved domain database)

Double Click anywhere

A high score and close to 100% coverage would indicate a high quality alignment, suggesting this sequence is highly conserved in a number of different organsims.

Quick scan
Visual rep
of coverage

Questions/comment

Done

COG – Clusters of Orthologous Groups

NCBI

Conserved Domains

HOME | SEARCH | GUIDE | NewSearch | Structure Home | 3D Macromolecular Structures | Conserved Domains | Pubchem | BioSystems

Conserved domains on [gi|256687299|gb|ACV05101|]

View: Standard Results

chromosomal replication initiator protein DnaA [Kytococcus sedentarius DSM 20547]

Protein Classification

chromosomal replication initiator protein DnaA (domain architecture ID 11478209)
chromosomal replication initiator protein DnaA plays a key role in the initiation and regulation of chromosomal replication

Graphical summary

Zoom to residue level | show extra options »

Query seq. (1, 75, 150, 225, 300, 375, 450, 506)

DnaA box-binding interface

Specific hits: dnaA; DnaA; DnaA; Bac_DnaA; Bac_DnaA_C; Bac_DnaA_C

Superfamilies: Bac_DnaA_C superfamily; P-loop_NTPase superfamily; dnaA superfamily

Click [+]

Search for similar domain architectures | Refine search

List of domain hits

[+]	Name	Accession	Description	Interval	E-value
[+]	dnaA	PRK00149	chromosomal replication initiation protein; Reviewed	157-500	0e+00
[+]	DnaA	TIGR00362	chromosomal replication initiator protein DnaA; DnaA is involved in DNA biosynthesis; ...	33-498	0e+00
[+]	DnaA	COG0593	Chromosomal replication initiation ATPase DnaA [Replication, recombination and repair];	147-496	2.48e-153
[+]	Bac_DnaA	pfam00308	Bacterial dnaA protein;	164-381	2.73e-115
[+]	Bac_DnaA_C	cd06571	C-terminal domain of bacterial DnaA proteins. The DNA-binding C-terminal domain of DnaA ...	409-498	1.32e-37
[+]	Bac_DnaA_C	smart00760	Bacterial dnaA protein helix-turn-helix domain; Could be involved in DNA-binding.	408-476	3.34e-30

COG hit

List of domain hits

	Name	Accession	Description	Interval	E-value
[+]	dnaA	PRK00149	chromosomal replication initiation protein; Reviewed	157-500	0e+00
[+]	DnaA	TIGR00362	chromosomal replication initiator protein DnaA; DnaA is involved in DNA biosynthesis; ...	33-498	0e+00
[-]	DnaA	COG0593	Chromosomal replication initiation ATPase DnaA [Replication, recombination and repair];	147-496	2.48e-153

Chromosomal replication initiation ATPase DnaA [Replication, recombination and repair];

COG name

Pssm-ID: 223666 Cd Length: 408 Bit Score: 447.50 E-value: 2.48e-153

Length, bit score, and E-value

```
                    10        20        30        40        50        60        70        80
            ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
gi 256687299 147 SLTATNSSPGVERDYSALNHKYTFDTFVLGSSNRFAHAAATAVAEAPARAYNPLFIYGGSGLGKTHLLHAIGHYARTLDS 226
Cdd:COG0593   63 KVEVRASAPAQLPLPSGLNPKYTFDNFVVGPSNRLAYAAAKAVAENPGGAYNPLFIYGGVGLGKTHLLQAIGNEALANGP 142

                    90       100       110       120       130       140       150       160
            ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
gi 256687299 227 SVRVKYVNSEEFTNQFINAVSAGQANAFQRQYrDVDVLLIDDIQFLQGKEQTMEEFFHTFNTLHNSEKQIVITSDQPPKK 306
Cdd:COG0593  143 NARVVYLTSEDFTNDFVKALRDNEMEKFKEKY-SLDLLLIDDIQFLAGKERTQEEFFHTFNALLENGKQIVLTSDRPPKE 221

                   170       180       190       200       210       220       230       240
            ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
gi 256687299 307 LSGFAERMRSRFEWGLLTDVQPPDLETRIAILRRKAAADKLDIPDDVLHLIASKISSNIRELEGALTRVTAFASLSGSPL 386
Cdd:COG0593  222 LNGLEDRLRSRLEWGLVVEIEPPDDETRLAILRKKAEDRGIEIPDEVLEFLAKRLDRNVRELEGALNRLDAFALFTKRAI 301

                   250       260       270       280       290       300       310       320
            ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
```

If there are no hits, write “no significant hits” in notebook

If there are several hits, click the **[+]** sign next to the hits and record

COG – Clusters of Orthologous Groups

NCBI

Conserved Domains

HOME | SEARCH | GUIDE | NewSearch | Structure Home | 3D Macromolecular Structures | Conserved Domains | Pubchem | BioSystems

Conserved domains on [gi|256687299|gb|ACV05101|]

View: Standard Results

chromosomal replication initiator protein DnaA [Kytococcus sedentarius DSM 20547]

Protein Classification

chromosomal replication initiator protein DnaA (domain architecture ID 11478209)
chromosomal replication initiator protein DnaA plays a key role in the initiation and regulation of chromosomal replication

Graphical summary

Zoom to residue level | show extra options »

Query seq. (1, 75, 150, 225, 300, 375, 450, 506)

DnaA box-binding interface

Specific hits: dnaA; DnaA; DnaA; Bac_DnaA; Bac_DnaA_C; Bac_DnaA_C

Superfamilies: Bac_DnaA_C superfamily; P-loop_NTPase superfamily; dnaA superfamily

Click

Search for similar domain architectures | Refine search

List of domain hits

	Name	Accession	Description	Interval	E-value
[+]	dnaA	PRK00149	chromosomal replication initiation protein; Reviewed	157-500	0e+00
[+]	DnaA	TIGR00362	chromosomal replication initiator protein DnaA; DnaA is involved in DNA biosynthesis; ...	33-498	0e+00
[+]	DnaA	COG0593	Chromosomal replication initiation ATPase DnaA [Replication, recombination and repair];	147-496	2.48e-153
[+]	Bac_DnaA	pfam00308	Bacterial dnaA protein;	164-381	2.73e-115
[+]	Bac_DnaA_C	cd06571	C-terminal domain of bacterial DnaA proteins. The DNA-binding C-terminal domain of DnaA ...	409-498	1.32e-37
[+]	Bac_DnaA_C	smart00760	Bacterial dnaA protein helix-turn-helix domain; Could be involved in DNA-binding.	408-476	3.34e-30

COG – Clusters of Orthologous Groups

NCBI

Conserved Protein Domain Family
DnaA

HOME | SEARCH | SITE MAP | Entrez | CDD | Structure | Protein | Help

COG0593: DnaA ?

Chromosomal replication initiation ATPase DnaA [Replication, recombination and repair]

Links ?
- **Source:** COG
- **Taxonomy:** Bacteria
- **Protein:** Representatives, Specific Protein, Related Protein, Related Structure, Architectures
- **Superfamily:** cl07055

Statistics ?

Structure ?

COG0593 is a member of the superfamily cl07055.

Click

NCBI

Conserved Protein Domain Family
Bac_DnaA_C

HOME | SEARCH | SITE MAP | Entrez | CDD | Structure | Protein | Help

cl07055: Bac_DnaA_C Superfamily ?

C-terminal domain of bacterial DnaA proteins. The DNA-binding C-terminal domain of DnaA contains a helix-turn-helix motif that specifically interacts with the DnaA box, a 9-mer motif that occurs repetitively in the replication origin oriC. Multiple copies of DnaA, which is an ATPase, bind to 9-mers at the origin and form an initial complex in which the DNA strands are being separated in an ATP-dependent step.

Sequence-based Similarity Data Module

4 TOOLS

1. BLAST	The Basic Local Alignment Search Tool (**BLAST**) finds regions of local similarity between sequences and calculates the statistical significance of matches
2. CDD	Conserved Domain Database Search (**CDD**) finds sequence similarity with genes in conserved orthologous groups (COGs).
3. T-Coffee	Tree based Consistency Objective Function For alignment Evaluation (T-Coffee) is a multiple sequence alignment program that aligns a set of homologous (similar) sequences
4. WebLogo	WebLogo is a program that enables easy creation of sequence logos from the multiple sequence alignments

T-Coffee

Tree-based Consistency Objective Function for alignment Evaluation

Multiple sequence alignment tool

- Across evolution amino acids are likely to be conserved because they are important for structure and function.

- One way to measure conservation is to align multiple similar protein sequences from related organisms (orthologs)

T-Coffee: A novel method for multiple sequence alignments. Notredame, Higgins, Heringa, **JMB** 302(205-217) **2000**

T-Coffee is a freeware open source package distributed under the GNU public license

T-Coffee Server is hosted by the **Centre for Genomic Regulation** (CRG) of Barcelona, SPAIN

In Notebook

T-Coffee

go to http://www.ebi.ac.uk/Tools/msa/tcoffee/

Sequences used for alignment

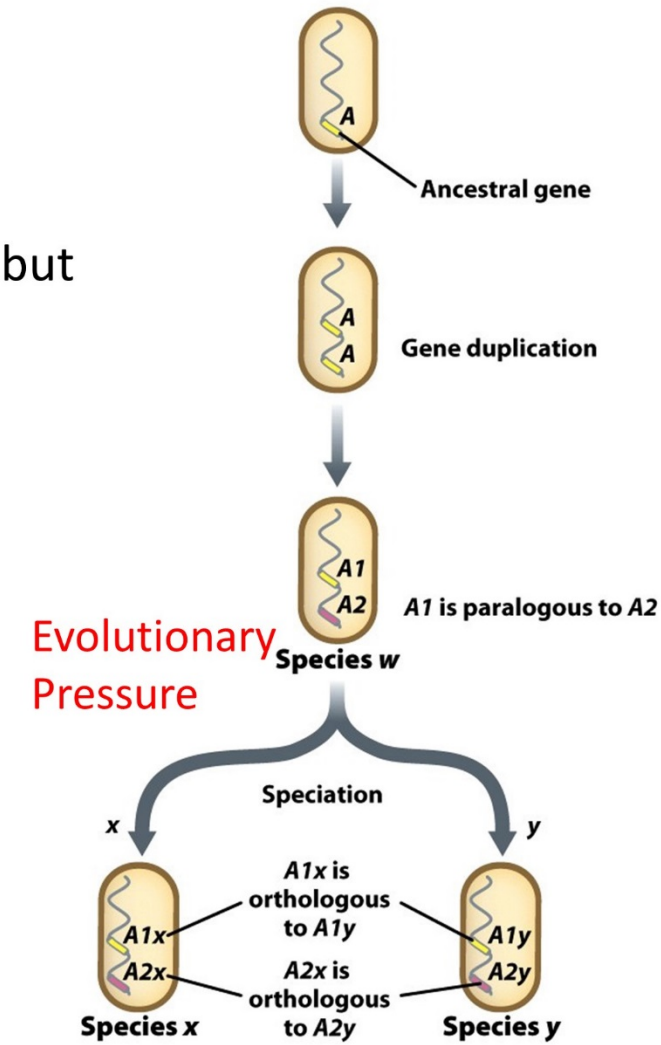
```
sequences
```

Multiple sequence alignment

```
alignment
```

RECALL: What are orthologs?

- Homologs
  - Orthologs
    - Genes that share similarity in function but are found in different organisms
  - Paralogs
    - Genes duplicated within a species
      - Perform slightly different tasks in cell
        - » Can develop new capabilities
        - » Can become pseudogene if functionality lost but sequence similarity retained



How and where are the orthologs to compare?

Go back to BLAST Search from nr database

Sequences producing significant alignments:

Select: All None Selected:10

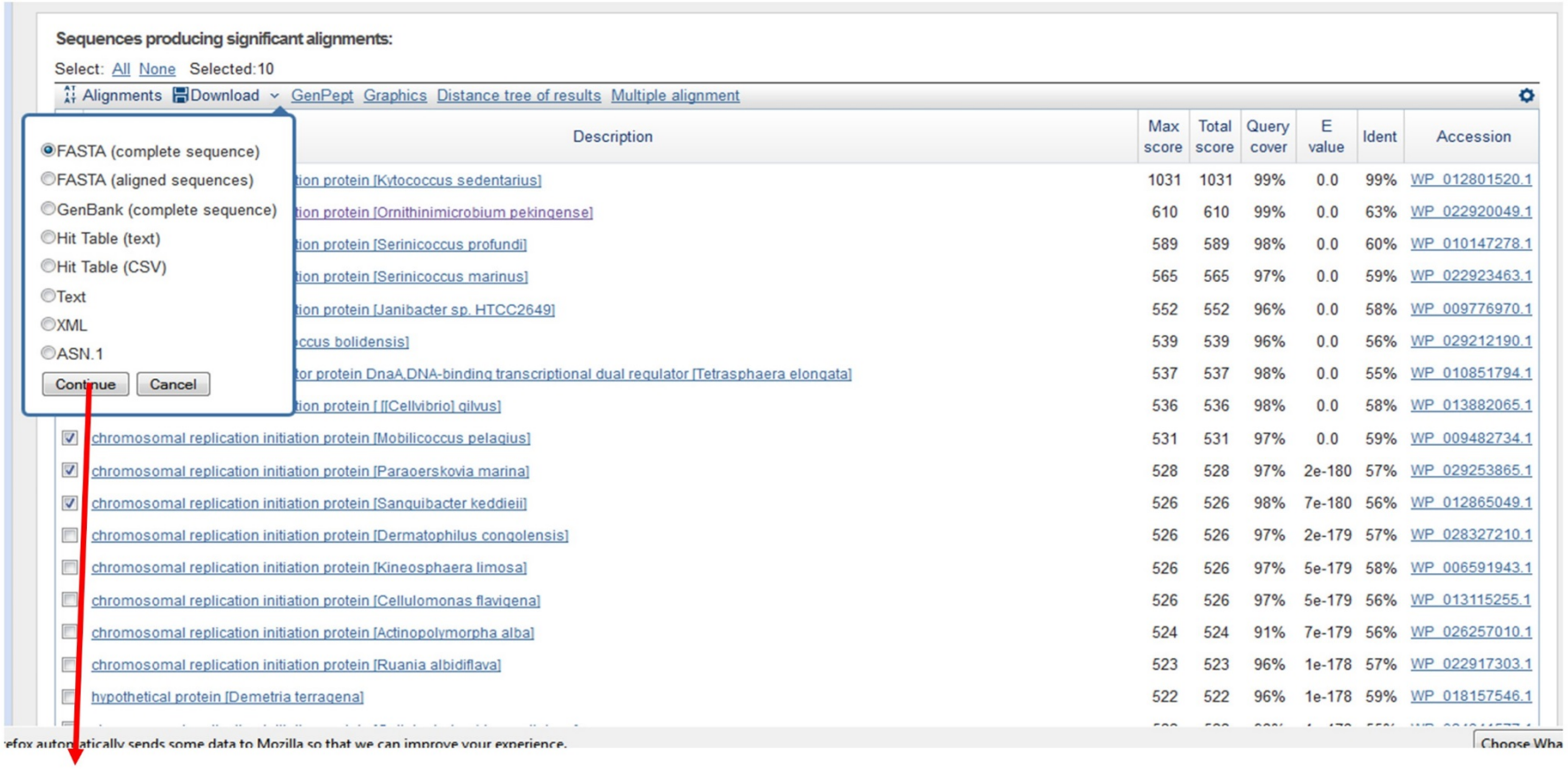
Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
chromosomal replication initiation protein [Kytococcus sedentarius]	1031	1031	99%	0.0	99%	WP_012801520.1
chromosomal replication initiation protein [Ornithinimicrobium pekingense]	610	610	99%	0.0	63%	WP_022920049.1
chromosomal replication initiation protein [Serinicoccus profundi]	589	589	98%	0.0	60%	WP_010147278.1
chromosomal replication initiation protein [Serinicoccus marinus]	565	565	97%	0.0	59%	WP_022923463.1
chromosomal replication initiation protein [Janibacter sp. HTCC2649]	552	552	96%	0.0	58%	WP_009776970.1
hypothetical protein [Arsenicicoccus bolidensis]	539	539	96%	0.0	56%	WP_029212190.1
chromosomal replication initiator protein DnaA,DNA-binding transcriptional dual regulator [Tetrasphaera elongata]	537	537	98%	0.0	55%	WP_010851794.1
chromosomal replication initiation protein [[[Cellvibrio] gilvus]	536	536	98%	0.0	58%	WP_013882065.1
chromosomal replication initiation protein [Mobilicoccus pelagius]	531	531	97%	0.0	59%	WP_009482734.1
chromosomal replication initiation protein [Paraoerskovia marina]	528	528	97%	2e-180	57%	WP_029253865.1
chromosomal replication initiation protein [Sanguibacter keddieii]	526	526	98%	7e-180	56%	WP_012865049.1

Click on 10 different significant orthologs from the list
Do NOT use the top 10 without checking out if they are from different organisms if available
Sometimes the same organism appear multiple times (different strains of the same organism)

Remember: you may need to do an “exclusion blast” if you don’t find enough different organisms in your routine nr blast search!

BLAST Search from nr database- Why?



Click the Download pull down menu at the top of the page and
make sure the FASTA (complete sequence) link is clicked –Continue
Copy and paste all 10 sequences including *Kytococcus sedentarius* in a word document.
Change font to Courier New 10 if needed. Edit *Kytococcus* FASTA header

Paste all sequences in your notebook and EDIT

```
>WP_012801520.1 chromosomal replication initiator protein DnaA [Kytococcus sedentarius]
MSQTPDDHATAIWQEAMVHLQGAGLAPRDIGVLRLATLVGLLEGTALLAVKYDHVKDAVEGHLREDVSTALAEVLDRDIR
LAVSVDPDAVSAAQEEAAPPAPSPADEDDPATGEGPLSTAVDGAVEKHEGSSPARAGESVAPATTASLTATNSSPGVERD
YSALNHKYTFDTFVLGSSNRFAHAAATAVAEAPARAYNPLFIYGGSGLGKTHLLHAIGHYARTLDSSVRVKYVNSEEFTN
QFINAVSAGQANAFQRQYRDVDVLLIDDIQFLQGKEQTMEEFFHTFNTLHNSEKQIVITSDQPPKKLSGFAERMRSRFEW
GLLTDVQPPDLETRIAILRRKAAADKLDI
GGDSGQITPTMILEETAGYFVISVEEIQ
LGEDRRVYDEVSELTSIIRKKAARGR
>WP_022920049.1 chromosomal
pekingense]
MTSQSPAESAEVWQRVVSQLESQGVTARDRAFLRLTQLVGLLDTTALLAVPYQHTKETLETTLRQPIVDALAGELGHDVR
LAITVDEDLRRQVEDEGDPAPGPAVTEQVPSDPDRTPYRSNGAGPGEPRSDGHRTPSGAVQTASAEDARLNPKYTFDTFV
SGSSNRFAHAASLAVAESPARAYNPLFIYGESGLGKTHLLHAIGHYARSLYPGVRVRYVNSEEFTNDFINSIRDDKAGAF
QRRYRNVDFLLVDDIQFLQGKEQTVEEFFHTFNTLHNSEKQVVITSDQPPKRLSGFAERMRSRFEWGLLTDVQPPDLETR
IAILKKKAAQEGMQLPDEVLELIGSKISTNIRELEGALIRVTAFASLSSTPPDAALASHVLKDIIPNSESAAITVPTIMA
EVADYFQISNDDLCGTSRSRTLVNARQIAMYLCRELTDLSLPKIGQEFGGRDHTTVMHAERKIRQLIGERRALYDQITEL
TGIIRKASAR
>RIK14929.1 chromosomal replication initiator protein DnaA [Acidobacteria bacterium]
MTHDPSPAASAEVWERVVAELDQGVTARDRAFLRLTRLVGLLDGTVLLAVPYQHTKDTLETTLRQPIVSALAEELGHDVR
LAITVDESLRQELKAEEGAVTPPQVAPAGGSTPYPVEVEPTSVPPVAEPTPRRAGATQGTGPDEARLNPKYSFDTFVSGS
SNRFAHAASLAVAESPARAYNPLFIYGESGLGKTHLLHAIGHYARSLYPGVRVRYVNSEEFTNDFINSIRNQEAGAFQRR
YRNVDFLLIDDIQFMQGKEQTVEEFFHTFNTLHNSEKQVVITSDQPPKRLSGFAERMRSRFEWGLLTDVQPPDLETRIAI
LRKKAAQEGMNTPDEVLELIASRITTNIRELEGALIRVTAFASLSSEPLTAELAAHVLKDIIPSGEAAAIGVPTIIAEVS
DYFQITRDELCGTSRSRSLVNARQIAMYLCRELTELSLPKIGQEFGGRDHTTVMHAERKIRQLMGERRALYDQITDLTGI
IRKASAR
>WP_010147278.1 chromosomal replication initiator protein DnaA [Serinicoccus profundi]
MSQPSTDSGDTWRRVVSELEDKGLGAREKAFLRLTTMVGVLDSTVLLAVPYPHTKEMLETTLRQPIVDLLSRELDREVRL
AITVDDDVRQRVEDEADDEADEDAQTRESLTRPASQPSSSAGAGVPGPSGNGIPRPATPAGPAVTGAADEARLNPKYSFD
TFVSGPSNRFAHAASLAVAESPARAYNPLFIYGESGLGKTHLLHAIGHYARKLYPGVRVRYVNSEEFTNDFINSIRDDKA
GAFQRRYRNVDFLLVDDIQFLQGKEQTVEEFFHTFNTLHNSEKQVVITSDQPPKRLSGFAERMRSRFEWGLLTDVQPPDL
ETRIAILRKKAAQEGMQLPDEVLEHIASRITTNIRELEGALIRVTAFASLSSQPADADLAAHVLKDIVPGSDTAQITVST
IIREVSEYFQISIDELCGTSRSRTLVNARQIAMYLCRELTDLSLPKIGQEFGGRDHTTVMHAERKIRAQIGERRALYDQI
AELTGTIRRASQR
```

Select all letters before the word *Kytococcus* and delete

The headers should look like this

```
>[Kytococcus sedentarius]
MSQTPDDHATAIWQEAMVHLQGAGLAPRDIGVLRLATLVGLLEGTALLAVKYDHVKDAVEGHLREDVSTALAEVLDRDIR
LAVSVDPDAVSAAQEEAAPPAPSPADEDDPATGEGPLSTAVDGAVEKHEGSSPARAGESVAPATTASLTATNSSPGVERD
YSALNHKYTFDTFVLGSSNRFAHAAATAVAEAPARAYNPLFIYGGSGLGKTHLLHAIGHYARTLDSSVRVKYVNSEEFTN
QFINAVSAGQANAFQRQYRDVDVLLIDDIQFLQGKEQTMEEFFHTFNTLHNSEKQIVITSDQPPKKLSGFAERMRSRFEW
GLLTDVQPPDLETRIAILRRKAAADKLDIPDDVLHLIASKISSNIRELEGALTRVTAFASLSGSPLDEYLARTVLKDVMP
GGDSGQITPTMILEETAGYFVISVEEIQGASRSRNLTRARQIAMYLCRELTDLSLPKIGKEFGGRDHTTVMHAERKIKQL
LGEDRRVYDEVSELTSIIRKKAARGR
>[Ornithinimicrobium pekingense]
MTSQSPAESAEVWQRVVSQLESQGVTARDRAFLRLTQLVGLLDTTALLAVPYQHTKETLETTLRQPIVDALAGELGHDVR
LAITVDEDLRRQVEDEGDPAPGPAVTEQVPSDPDRTPYRSNGAGPGEPRSDGHRTPSGAVQTASAEDARLNPKYTFDTFV
SGSSNRFAHAASLAVAESPARAYNPLFIYGESGLGKTHLLHAIGHYARSLYPGVRVRYVNSEEFTNDFINSIRDDKAGAF
QRRYRNVDFLLVDDIQFLQGKEQTVEEFFHTFNTLHNSEKQVVITSDQPPKRLSGFAERMRSRFEWGLLTDVQPPDLETR
IAILKKKAAQEGMQLPDEVLELIGSKISTNIRELEGALIRVTAFASLSSTPPDAALASHVLKDIIPNSESAAITVPTIMA
EVADYFQISNDDLCGTSRSRTLVNARQIAMYLCRELTDLSLPKIGQEFGGRDHTTVMHAERKIRQLIGERRALYDQITEL
TGIIRKASAR
>[Acidobacteria bacterium]
MTHDPSPAASAEVWERVVAELDQGVTARDRAFLRLTRLVGLLDGTVLLAVPYQHTKDTLETTLRQPIVSALAEELGHDVR
LAITVDESLRQELKAEEGAVTPPQVAPAGGSTPYPVEVEPTSVPPVAEPTPRRAGATQGTGPDEARLNPKYSFDTFVSGS
SNRFAHAASLAVAESPARAYNPLFIYGESGLGKTHLLHAIGHYARSLYPGVRVRYVNSEEFTNDFINSIRNQEAGAFQRR
YRNVDFLLIDDIQFMQGKEQTVEEFFHTFNTLHNSEKQVVITSDQPPKRLSGFAERMRSRFEWGLLTDVQPPDLETRIAI
LRKKAAQEGMNTPDEVLELIASRITTNIRELEGALIRVTAFASLSSEPLTAELAAHVLKDIIPSGEAAAIGVPTIIAEVS
DYFQITRDELCGTSRSRSLVNARQIAMYLCRELTELSLPKIGQEFGGRDHTTVMHAERKIRQLMGERRALYDQITDLTGI
IRKASAR
>[Serinicoccus profundi]
MSQPSTDSGDTWRRVVSELEDKGLGAREKAFLRLTTMVGVLDSTVLLAVPYPHTKEMLETTLRQPIVDLLSRELDREVRL
AITVDDDVRQRVEDEADDEADEDAQTRESLTRPASQPSSSAGAGVPGPSGNGIPRPATPAGPAVTGAADEARLNPKYSFD
TFVSGPSNRFAHAASLAVAESPARAYNPLFIYGESGLGKTHLLHAIGHYARKLYPGVRVRYVNSEEFTNDFINSIRDDKA
GAFQRRYRNVDFLLVDDIQFLQGKEQTVEEFFHTFNTLHNSEKQVVITSDQPPKRLSGFAERMRSRFEWGLLTDVQPPDL
ETRIAILRKKAAQEGMQLPDEVLEHIASRITTNIRELEGALIRVTAFASLSSQPADADLAAHVLKDIVPGSDTAQITVST
IIREVSEYFQISIDELCGTSRSRTLVNARQIAMYLCRELTDLSLPKIGQEFGGRDHTTVMHAERKIRAQIGERRALYDQI
AELTGTIRRASQR
```

Click on Link: T-Coffee tools

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

T-Coffee

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

Tools > Multiple Sequence Alignment > T-Coffee

Multiple Sequence Alignment

T-Coffee is a multiple sequence alignment program. Its main characteristic is that it will allow you to comb

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:

```
>Kytococcus sedentarius]
MSQTPDDHATAIWQEAMVHLQGAGLAPRDIGVLRLATLVGLLEGTALLAVKYDHVKDAVEGHLREDVSTALAEVLDRDIR
LAVSVDPDAVSAAQEEAAPPAPSPADEDDPATGEGPLSTAVDGAVEKHEGSSPARAGESVAPATTASLTATNSSPGVERD
YSALNHKYTFDTFVLGSSNRFAHAAATAVAEAPARAYNPLFIYGGSGLGKTHLLHAIGHYARTLDSSVRVKYVNSEEFJN
QFINAVSAGQANAFQRQYRDVDVLLIDDIQFLQGKEQTMEEFFHTFNTLHNSEKQIVITSDQPPKKLSGFAERMRSRFEW
GLLTDVQPPDLETRIAILRRKAAADKLDIPDDVLHLIASKISSNIRELEGALTRVTAFASLSGSPLDEYLARTVLKDVMP
GGDSGQITPTMILEETAGYFVISVEEIQGASRSRNLTRARQIAMYLCRELTDLSLPKIGKEFGGRDHTTVMHAERKIKQL
LGEDRRVYDEVSELTSIIRKKAARGR
```

Or upload a file: Browse... No file selected.

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

STEP 2 - Set your Parameters

OUTPUT FORMAT:

ClustalW

The default settings will fulfill the needs of most users.

More options... *(Click here, if you want t*

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

Copy / paste all the 10 amino acid sequences in FASTA format into the query window.

CLICK SUBMIT

Multiple Sequence Alignment (msa)

Copy / paste this alignment into your lab notebook-
Courier New 9

Results for job tcoffee-l20190215-165604-0351-23150635-p1m

Alignments | Result Summary | Guide Tree | Phylogenetic Tree | Submission Detail

Download Alignment File | Show Colors | View result with Jalview | Send to Simple

```
CLUSTAL W (1.83) multiple sequence alignment

[Acidobacteria       MTHDPSPAASAEVWERVVAELD--QGVTARDRAFLRLTRLVGLLDGTVLL
[Arsenicicoccus      MSQ-PST-DTGDTWRRVVSELED-KGLGAREKAFLRLTTMVGVLDSTVLL
[Cellulomonas        MSG-QDD-QLSQVWSAAMAQLEVSPDITPRQLAFVRLAKPLGLLDGTMLL
[Janibacter          MSE-PSE-DLDAVWRSILAAVSH-DGVPAPHRAFLSLARFVGLLDGTALV
[Knoellia            MD---------QIWRTTLDALDS-DGIPVQQRAFLSLAKLVGLLDETALI
[Kribbia             MGG-TDE-DFAQIWHNILDTLDA-DGVPVTERAFLQLGKLVGLLDGTAVI
[Kytococcus          MSQTPDD-HATAIWQEAMVHLQG-AGLAPRDIGVLRLATLVGLLEGTALL
[Ornithinimicrobium  MTSQSPA-ESAEVWQRVVSQLES-QGVTARDRAFLRLTQLVGLLDTTALL
[Serinicoccus        MSQ-PST-DSGDTWRRVVSELED-KGLGAREKAFLRLTTMVGVLDSTVLL
[Tetrasphaera        MD---------QIWRTTLDALDN-DGIPVQQRAFLSLARLVGLLDDTALI
                     *           *   :  :.   .:   . ..: *   :*:*: * ::

[Acidobacteria       AVPYQHTKDTLETTLRQPIVSALAEELGHDVRLAITVDESLRQELKAE-E
[Arsenicicoccus      AVPYPHTKEMLETTLRQPIVDLLSRELDREVRLAITVDDDVRQRVEDE-A
[Cellulomonas        AVGNDLTKDYLETRVRQEVTDALAAALGRDARFAITVDPSLDGAGDPS-L
[Janibacter          AVPNNYCKTYVERALRVPVTQAFSAHYGQDVRLAVTVDPDLDDTEDEL-P
[Knoellia            AVPNDFTKDIVETRLRDRVTETLRSQLGHDVRLAVTVDPSLGDAPVLVPA
[Kribbia             AVPNDFSKQFVEHRLRQHVTSALSAQLGSEVRLAVTVDSSLAEGGDTD-T
[Kytococcus          AVKYDHVKDAVEGHLREDVSTALAEVLDRDIRLAVSVDPDAVSAAQEE-A
[Ornithinimicrobium  AVPYQHTKETLETTLRQPIVDALAGELGHDVRLAITVDEDLRRQVEDE-G
[Serinicoccus        AVPYPHTKEMLETTLRQPIVDLLSRELDREVRLAITVDDDVRQRVEDE-A
[Tetrasphaera        AVPNDFTKDIVETRLRERVTETLSSQLGHDVRLAVTVDQSLADAPAPE-A
                     **     *  :*  :*  :   :    . : *:*::** .
```

Repeating blocks of 50 amino acid stretches

Asterix : (*), conserved amino acid in all sequences

Colon (:), A position of the MSA composed of residues having the same physicochemical properties

Dot (.), indicates the column of MSA for which semi-conserved substitutions are observed

Return to T-COFFEE Results - show color

```
CLUSTAL W (1.83) multiple sequence alignment

gi|497130903|ref|WP_009482734.1|  MSVSGESSTTPSEPGRIWGATLRALDQ-AGIPAPQRAFLRQATLVGVLDT
gi|497462772|ref|WP_009776970.1|  M-------------DQIWRTTLDALDS-DGIPVQQRAFLSLAKLVGLLDE
gi|497833122|ref|WP_010147278.1|  MSQ------PSTDSGDTWRRVVSELED-KGLGAREKAFLRLTTMVGVLDS
gi|499072896|ref|WP_010851794.1|  MAD--------ASMTSVWVRILRALDR-EGVSHQERAFLSITRLAGVLDE
gi|502479361|ref|WP_012801520.1|  MSQT-----PDDHATAIWQEAMVHLQG-AGLAPRDIGVLRLATLVGLLEG
gi|502628385|ref|WP_012865049.1|  MAT------TDDNISEIWKQAIAELEASPDITPRQLAFVKLAKPLGLFDG
gi|503647989|ref|WP_013882065.1|  MAQ-------DEELSRVWGHVVTTLEESPDITQRQLAFVRLAQPLGLLDG
gi|551300082|ref|WP_022920049.1|  MTSQ-----SPAESAEVWQRVVSQLES-QGVTARDRAFLRLTQLVGLLDT
gi|656266264|ref|WP_029212190.1|  MTD------AQVDVPRVWRDTLRALES-GGISAQHRGFLRLSRLVGLLEG
gi|656321871|ref|WP_029253865.1|  MPA------AEVSIDEVWEQTIATLGSNPHMTRRQMGYVKMAKPRAVFEG
                                  *               *   :  *     :   . . :  :   .:::

gi|497130903|ref|WP_009482734.1|  TALIAVPDDFTKEIVESRARDRLVRALTDQVGHEVRLAVTVDASLREQFA
gi|497462772|ref|WP_009776970.1|  TALIAVPNDFTKDIVETRLRDRVTETLSSQLGHDVRLAVTVDHSLADVPV
gi|497833122|ref|WP_010147278.1|  TVLLAVPYPHTKEMLETTLRQPIVDLLSRELDREVRLAITVDDDVRQRVE
gi|499072896|ref|WP_010851794.1|  TALIAVPNDFSKDIVETRLRGRISGHLTAELDRPLRLAVTVDPSLAEAEP
gi|502479361|ref|WP_012801520.1|  TALLAVKYDHVKDAVEGHLREDVSTALAEVLDRDIRLAVSVDPDAVSAAQ
gi|502628385|ref|WP_012865049.1|  TVIIAVANDHTRDFLETRVRAEVVQALSNALGRDARFAITVDPELGFDEE
gi|503647989|ref|WP_013882065.1|  TIILAVGNEYTKEYLETKVRAEVTSALGSALGRDGRFAITVDPSLVDDAP
gi|551300082|ref|WP_022920049.1|  TALLAVPYQHTKETLETTLRQPIVDALAGELGHDVRLAITVDEDLRRQVE
gi|656266264|ref|WP_029212190.1|  TALIAVPNDYTRDIVEKRIRTELVAALQEQLGRDVRLAVTVDSSLELSEA
gi|656321871|ref|WP_029253865.1|  NVFLAVPADHVRTFIESSLRDDLVEALTSVLGTEVRFAISVEPDMDVQPP
                                  . ::**   . :  :*   *  :   *   :.   *:*::*: .
```

Amino Acid Residue	3-Letter Code	1-Letter Code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic Acid	Asp	D
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamic Acid	Glu	E
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

AVFPMILW	RED	Small (small+ hydrophobic (incl.aromatic -Y))
DE	BLUE	Acidic
RK	MAGENTA	Basic
STYHCNGQ	GREEN	Hydroxyl + Amine + Basic - Q
Others	Gray	

The colors give information about the amino acid

Sequence-based Similarity Data Module

4 TOOLS

1. BLAST	The Basic Local Alignment Search Tool (**BLAST**) finds regions of local similarity between sequences and calculates the statistical significance of matches
2. CDD	Conserved Domain Database Search (**CDD**) finds sequence similarity with genes in conserved orthologous groups (COGs).
3. T-Coffee	Tree based Consistency Objective Function For alignment Evaluation (T-Coffee) is a multiple sequence alignment program that aligns a set of homologous (similar) sequences
4. WebLogo	WebLogo is a program that enables easy creation of sequence logos from the multiple sequence alignments

http://weblogo.berkeley.edu/

- This is a program designed to enable easy creation of sequence logos from multiple sequence alignments.
- One simple graphic is generated.
- At least 10 sequences should be used.
- Save image as .png and attach to notebook.

Crooks GE, Hon G, Chandonia JM, Brenner SE WebLogo: A sequence logo generator, *Genome Research*, 14:1188-1190, (2004)
Created by Computational Genomics Research Group,, Department of Plant and Microbial Biology, University of California, Berkeley

WebLogo

WebLogo

go to http://weblogo.berkeley.edu/

Sequence logo

logo image

Comments/observations about the Multiple Sequence Alignment WebLogo

comments

Provide an overall summary of your findings from the Sequence Similarity Module in the box below.

Be sure to save this document after completing the sequence-based information module!

Home page for WebLogo

WEBLOGO

· about · create · examples ·

Version 2.8.2 (2005-09-08)

(⇒ WebLogo 3)

References

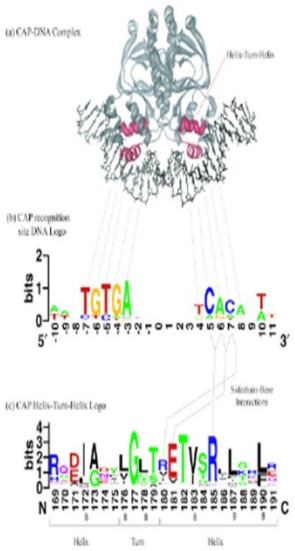
Crooks GE, Hon G, Chandonia JM, Brenner SE WebLogo: A sequence logo generator, *Genome Research*, 14:1188-1190, (2004) [Full Text]

Schneider TD, Stephens RM. 1990. Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Res. 18*:6097-6100

Introduction

WebLogo is a web based application designed to make the generation of sequence logos as easy and painless as possible. Click **here** to create your own sequence logos.

Sequence logos are a graphical representation of an amino acid or nucleic acid multiple sequence alignment developed by Tom Schneider and Mike Stephens. Each logo consists of stacks of symbols, one stack for each position in the sequence. The overall height of the stack indicates the sequence conservation at that position, while the height of symbols within the stack indicates the relative frequency of each amino or nucleic acid at that position. In general, a sequence logo provides a richer and more precise description of, for example, a binding site, than would a consensus sequence.



WebLogo Source Code

The WebLogo source code is available for download. See the README file for installation instructions and release notes, and LICENSE for the open source license.

Paste multiple alignment with top and bottom line intact

CLICK 32 so that aa are chopped in 32 bits and shown

File Edit View History Bookmarks Tools Help

Autoimmune vitiligo is ass... | In Reply to Ho. - PubMed -... | geni-act | WebLogo - Create Sequence L...

weblogo.berkeley.edu/logo.cgi

Most Visited | Blackboard Learn | BMC Genomics | Catalog of Published ... | COE sharepoint | Facebook | GeneCards - Human G... | geni-act | Getting Started | Google | HGNC database of hu... | IPA Launch | MITF: master regulator... | nsf.gov - Publications ... | file:///C:/Users/dey.DE...

WEBLOGO · about · create · examples ·

Multiple Sequence Alignment

```
CLUSTAL W (1.83) multiple sequence alignment

Kseed_00010                          VSQTPD-----------------------------
DHATAIWQEAMVH
gi|1187706|sp|P21173.1|DNAA_MICLU    MVADQ--------------------------------
AVLSSWRSVVGS
gi|123144805|sp|Q0SAG7.1|DNAA_RHOSR  MNDDPN-------------------------------
ALARIWIDVVAD
gi|123774818|sp|Q47U23.1|DNAA_THEFY  MSEGQ--------------------------------
INLAMVWSRVLDN
gi|166214685|sp|A1T102.1|DNAA_MYCVP  MTTDPD-------------------------------
```

Upload Sequence Data: Browse... No file selected.

Image Format & Size

Image Format: PNG (bitmap) | Logo Size per Line: 18 X 5 cm

Create Logo | Reset

Advanced Logo Options

Sequence Type: amino acid / DNA / RNA / Automatic Detection

First Position Number: 1 | Logo Range: -

Small Sample Correction: ☑ | Frequency Plot: ☐

Multiline Logo (Symbols per Line): ☑ (32)

Advanced Image Options

Bitmap Resolution: 96 pixels/inch (dpi) | Antialias Bitmaps: ☑

Title: | Y-Axis Height: (bits)

Show Y-Axis: ☑ | Y-Axis Label: bits

Show X-Axis: ☑ | X-Axis Label:

Show Error Bars: ☐ | Label Sequence Ends: ☑

Boxed / Boxed Shrink Factor: ☐ / 0.5 | Outline Symbols: ☐

Show fine print: ☑ | Y-Axis Tic Spacing: 1 (bits)

Colors

Color Scheme: Default / Black & White / Custom (See Below.)

Create →Paste multiple alignment with top and bottom line intact

CLICK 32

aa are shown in repeating blocks of 32

Most Visited | Blackboard Learn | BMC Genomics | Catalog of Published ... | COE sharepoint | Facebook | GeneCards - Human G... | geni-act | Getting Started | Google | HGNC database of hu... | IPA Launch | MITF: master regulator... | nsf.gov - Publications ...

```
gi|189044597|sp|A6W3V4.1|DNAA_KINRD  RSTYNQVTELTNRIKKQ-----SGA
gi|189044633|sp|B0RH69.1|DNAA_CLAMS  RSIYNQVTELTSRIKQNHR--YGKM
gi|226735850|sp|B1VPF0.1|DNAA_STRGG  RSIYNQVTELTNRIKN--------G
gi|254777911|sp|C1B7S7.1|DNAA_RHOOB  RKVYDQVQELTARIKQRS-----KR
gi|61212561|sp|Q6ABL5.1|DNAA_PROAC   RQIFNQVSEITNRIKQ--------Y
gi|61212563|sp|Q6AHN6.1|DNAA_LEIXX   RSIYNQVTELTSRIKQN------HR
                                     *  ::.* *:*  *:.
```

Upload Sequence Data: Browse... No file selected.

Image Format & Size

Image Format: PNG (bitmap) | Logo Size per Line: 18 X 5 cm

Create Logo | Reset

Advanced Logo Options

Sequence Type: amino acid / DNA / RNA / Automatic Detection

First Position Number: 1 | Logo Range: -

Small Sample Correction: ☑ | Frequency Plot: ☐

Multiline Logo (Symbols per Line): ☑ (32)

Advanced Image Options

Bitmap Resolution: 96 pixels/inch (dpi) | Antialias Bitmaps: ☑

Title: | Y-Axis Height: (bits)

Show Y-Axis: ☑ | Y-Axis Label: bits

Show X-Axis: ☑ | X-Axis Label:

Show Error Bars: ☐ | Label Sequence Ends: ☑

Boxed / Boxed Shrink Factor: ☐ / 0.5 | Outline Symbols: ☐

Show fine print: ☑ | Y-Axis Tic Spacing: 1 (bits)

Colors

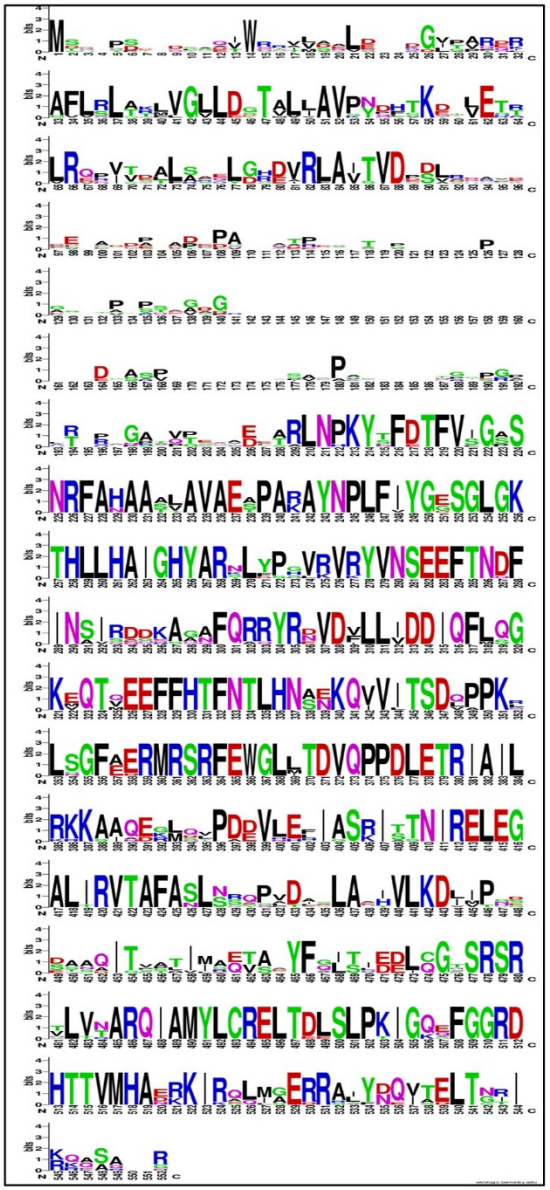
Color Scheme: Default / Black & White / Custom (See Below.)

Symbols	Color	RGB	Symbols	Color	RGB
KRH	green			purple	
DE	blue			orange	
AVLIPWFM	red			black	
	black		Other	black	

Create Logo | Reset

CLICK

Weblogo of the entire protein
Save as a .png file. Upload into notebok



COLOR CODE

Polar amino acids (G,S,T,Y,C,Q,N) : green

Basic (K,R,H) blue,

Acidic (D,E) red

Hydrophobic (A,V,L,I,P,W,F,M) : **black.**

The most common residue at each position in the alignment will be the largest letter at that position.

The relative height of the stack of letters will be proportional to the % conservation.

The relative widths of the stacks indicate the proportion of valid readings of nucleic bases or amino acids at that position.

The more gaps in the sequence at a specific position means a thinner stack.