Genome Annotation

MODULE 2

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**) 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**

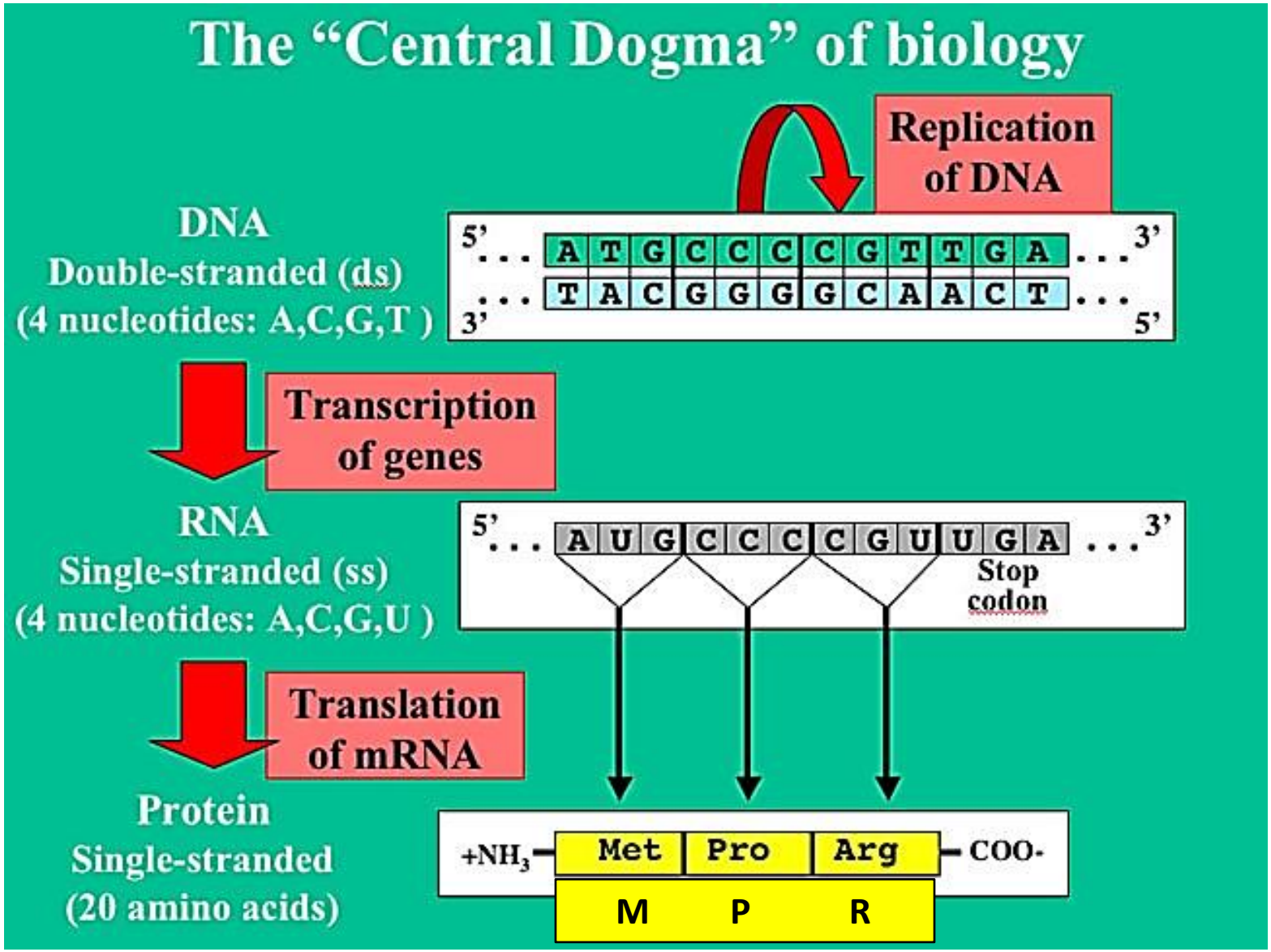
Tree based Consistency Objective Function For alignment Evaluation (T-Coffee) is a multiple sequence alignment program that aligns a set of homologous (similar) sequences

4. **WebLogo**

WebLogo is a program that enables easy creation of sequence logos from the multiple sequence alignments

Relationship

DNA ----------→ RNA ----------→ PROTEIN



Bruce Fouke (2006)

Using amino acid sequence (proteins) and not DNA sequence (Gene) in similarity searches WHY?

Amino Acids and Their Symbols			Codons
G	Gly	Glycine	GGA; GGC; GGG; GGU
H	His	Histidine	CAC; CAU
I	Ile	Isoleucine	AUA; AUC; AUU
K	Lys	Lysine	AAA; AAG
L	Leu	Leucine	UUA; UUG; CUA; CUC; CUG; CUU
M	Met	Methionine	AUG
N	Asn	Asparagine	AAC; AAU
P	Pro	Proline	CCA; CCC; CCG; CCU
Q	Gln	Glutamine	CAA; CAG
R	Arg	Arginine	AGA; AGG; CGA; CGC; CGG; CGU
S	Ser	Serine	AGC; AGU; UCA; UCC; UCG; UCU
T	Thr	Threonine	ACA; ACC; ACG; ACU
V	Val	Valine	GUA; GUC; GUG; GUU
W	Trp	Tryptophan	UGG
Y	Tyr	Tyrosine	UAC; UAU

Redundancy of codons

ANSWER: MORE than one codon or triplet can code for a particular amino acid. A lot of variation exists in DNA sequences that code for the same amino acid.

Comparing amino acid sequences is a more reliable to use for similarity between two sequences than comparing nucleotide sequences. Removes the redundancy issue.

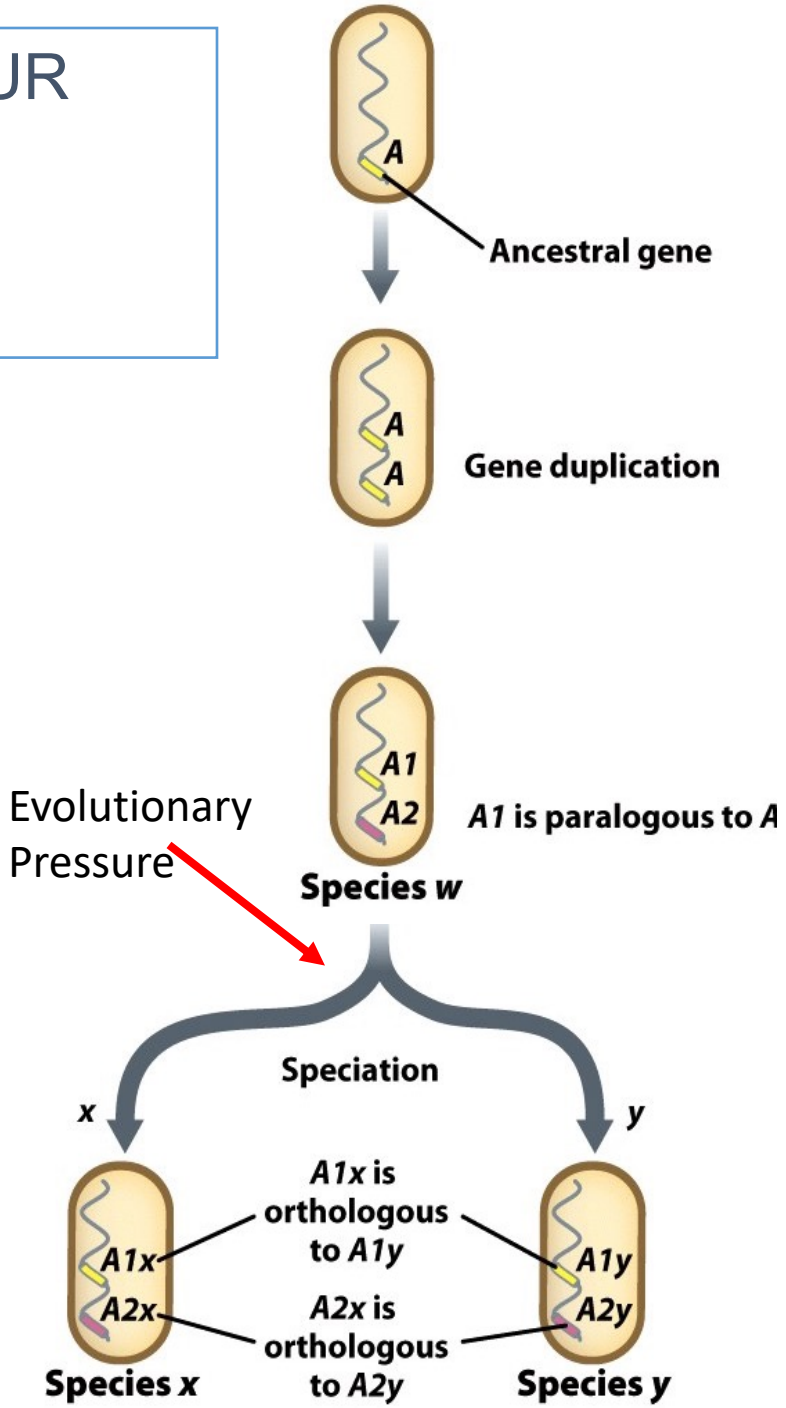
BLAST: Searches for similarity of YOUR protein sequence to all known protein sequences from all organisms in the databases.

We are looking for

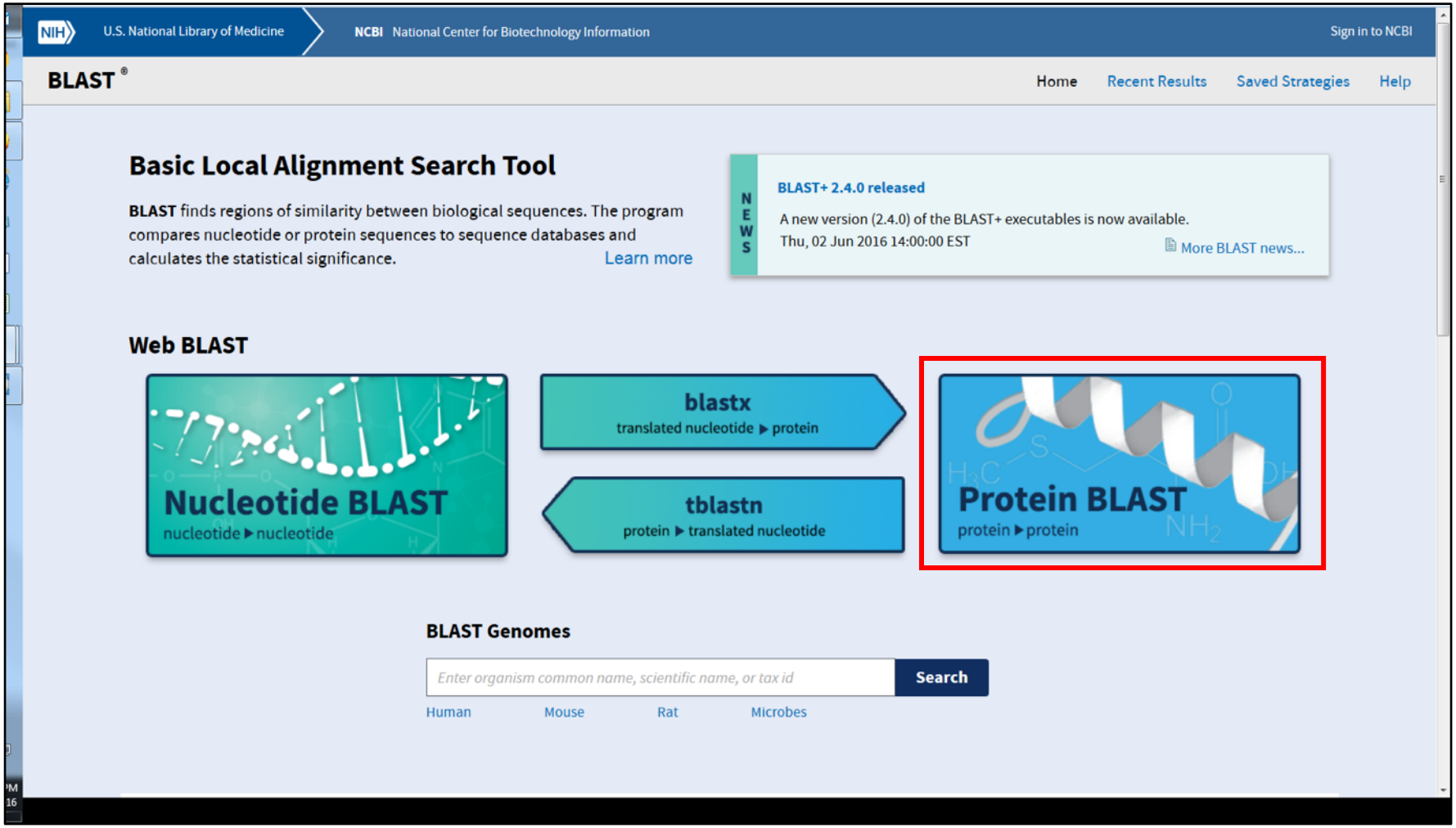
Orthologous genes/proteins

- What are orthologs?

  Genes are duplicated with appearance of new species. They code for proteins that have similar function in different organisms.

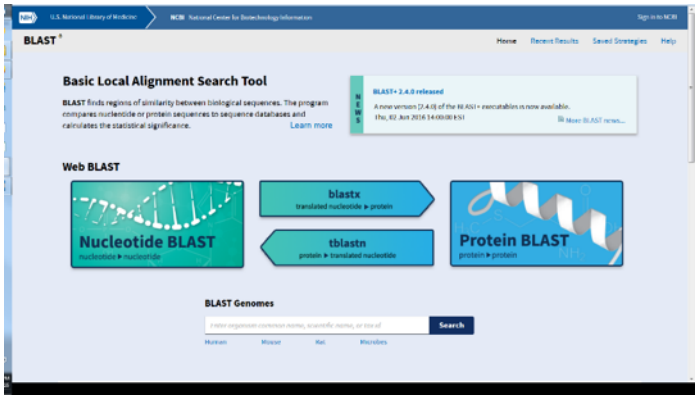


Basic Local Alignment Search Tool
BLAST (NCBI)



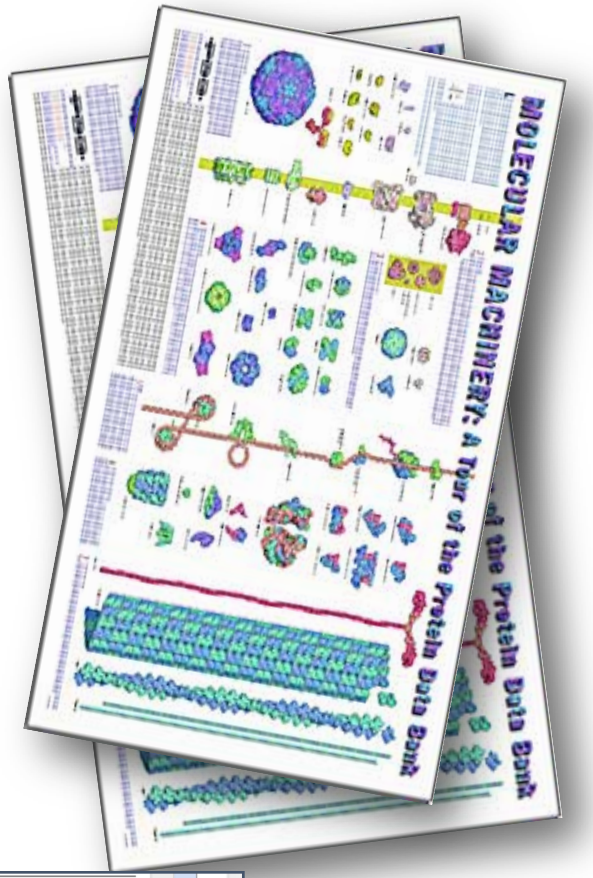
Altschul, Stephen; Gish, Warren; Miller, Webb; Myers, Eugene; Lipman, David "Basic local alignment search tool". *Journal of Molecular Biology* ***215*** *(3): 403–410 (1990) One of the highest cited papers >50,000 times*

What is protein BLAST?



Protein Databases (Subject)

BLAST



The algorithms in this tool finds regions of similarity between amino acid sequence of your proteins (Query) with those in databases and calculates the statistical significance.

Putative Function

Alignments | Download | GenPept | Graphics | Distance tree of results | Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
RecName: Full=Chromosomal replication initiator protein DnaA	517	517	97%	1e-177	56%	A6W3V4.1
RecName: Full=Chromosomal replication initiator protein DnaA	498	498	99%	7e-171	52%	A1T102.1
RecName: Full=Chromosomal replication initiator protein DnaA	495	495	99%	2e-169	51%	A0PKB2.1
RecName: Full=Chromosomal replication initiator protein DnaA	494	494	99%	4e-169	51%	A0R7K1.1
RecName: Full=Chromosomal replication initiator protein DnaA	493	493	99%	1e-168	50%	B2HI46.1
RecName: Full=Chromosomal replication initiator protein DnaA	490	490	99%	9e-168	51%	Q1BG61.1
RecName: Full=Chromosomal replication initiator protein DnaA	488	488	99%	7e-167	50%	B1MDH6.1

Basic Local Alignment Search Tool
BLAST

- Widely used similarity search algorithm by scientist
- Searches for similarity of protein sequence (under study) in FASTA format to ALL protein sequences from ALL organisms in the database
- Is able to identify regions of similarity within two sequences thus finding local alignments (some portion of 2 sequences) as opposed to global alignment (alignment of 2 sequences over their full length)
- **BENEFITS:**
  - **SPEED**
  - **USER FRIENDLY**
  - **STATISTICAL RIGOR**
  - **SENSITIVE**

Altschul, Stephen; Gish, Warren; Miller, Webb; Myers, Eugene; Lipman, David "Basic local alignment search tool".
*Journal of Molecular Biology **215** (3): 403–410 (1990) One of the highest cited papers >50,000 times*

(NCBI (National council for Biotechnology Information from NIH)
http://www.ncbi.nlm.nih.gov/blast

Key Aspects of FASTA format- QUERY AA SEQUENCE- Protein

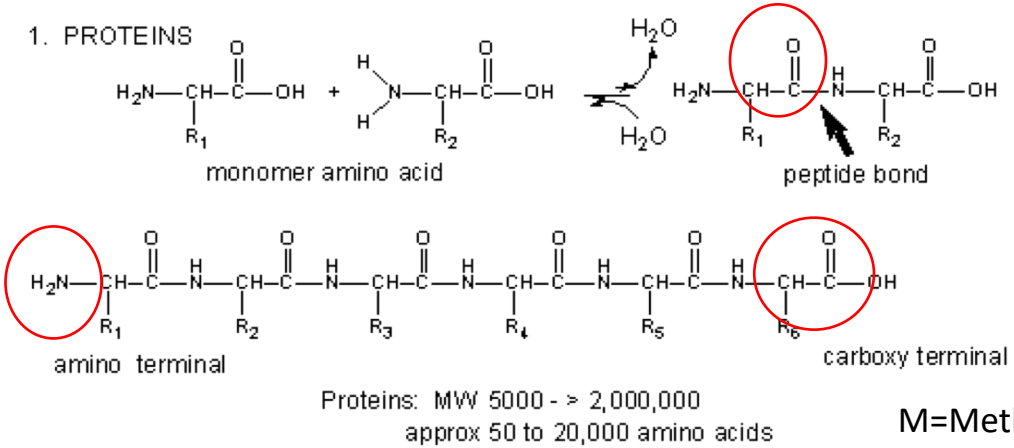
”description line” (not read as sequence data)

- Begins with >
- Ends with a hard return

```
>Ksed_00010 amino acid

MSQTPDDHATAIWQEAMVHLQGAGLA
PRDIGVLRLATLVGLLEGTALLAVKY
DHVKDAVEGHLREDVSTALAEVLDRD
IRLAVSVDPDAVSAAQEEAAPPAPSP
ADEDDPATGEGPLSTAVDGAVEKH
```

N-terminal end or amino term end

C- or carboxy terminal end

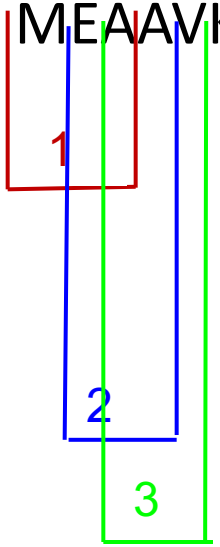


Sequence data-Protein
(amino acid sequence)

M=Methionine, S= Serine, Q= Glutamine, T= Threonine …….

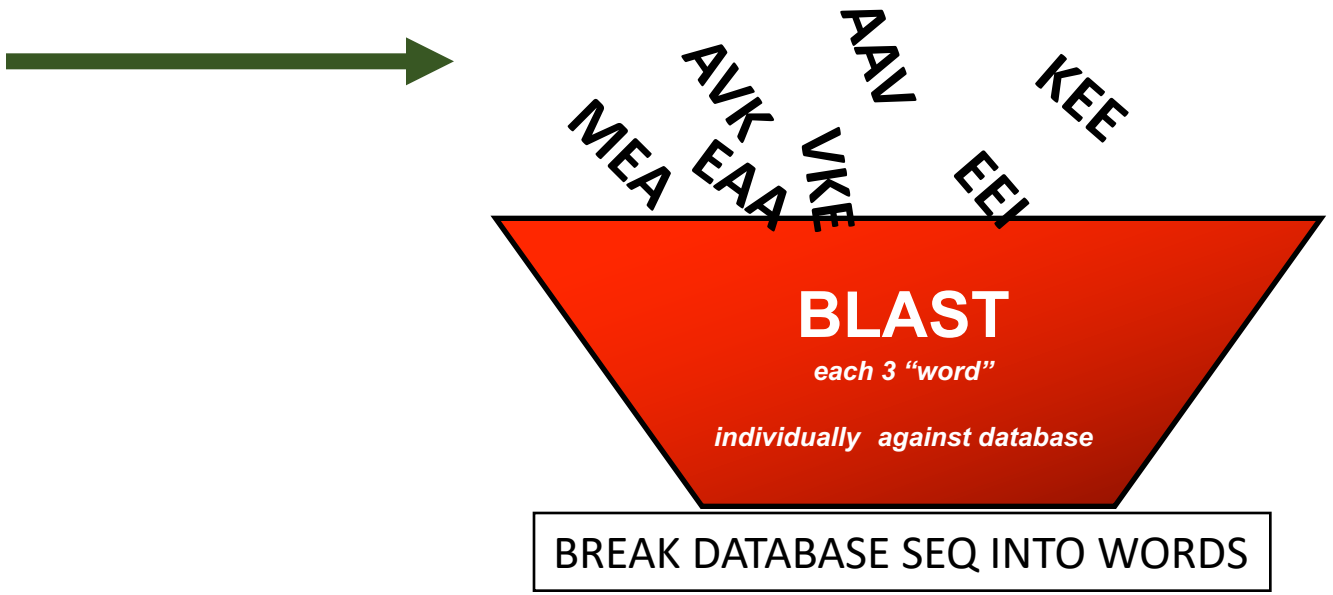
BLAST chops your query amino acid seq (protein) into "word" length pieces, amino acids (3 letters) Input sequences are in FASTA format

MEAAVKEEISVEDEAVDK……- Query sequence



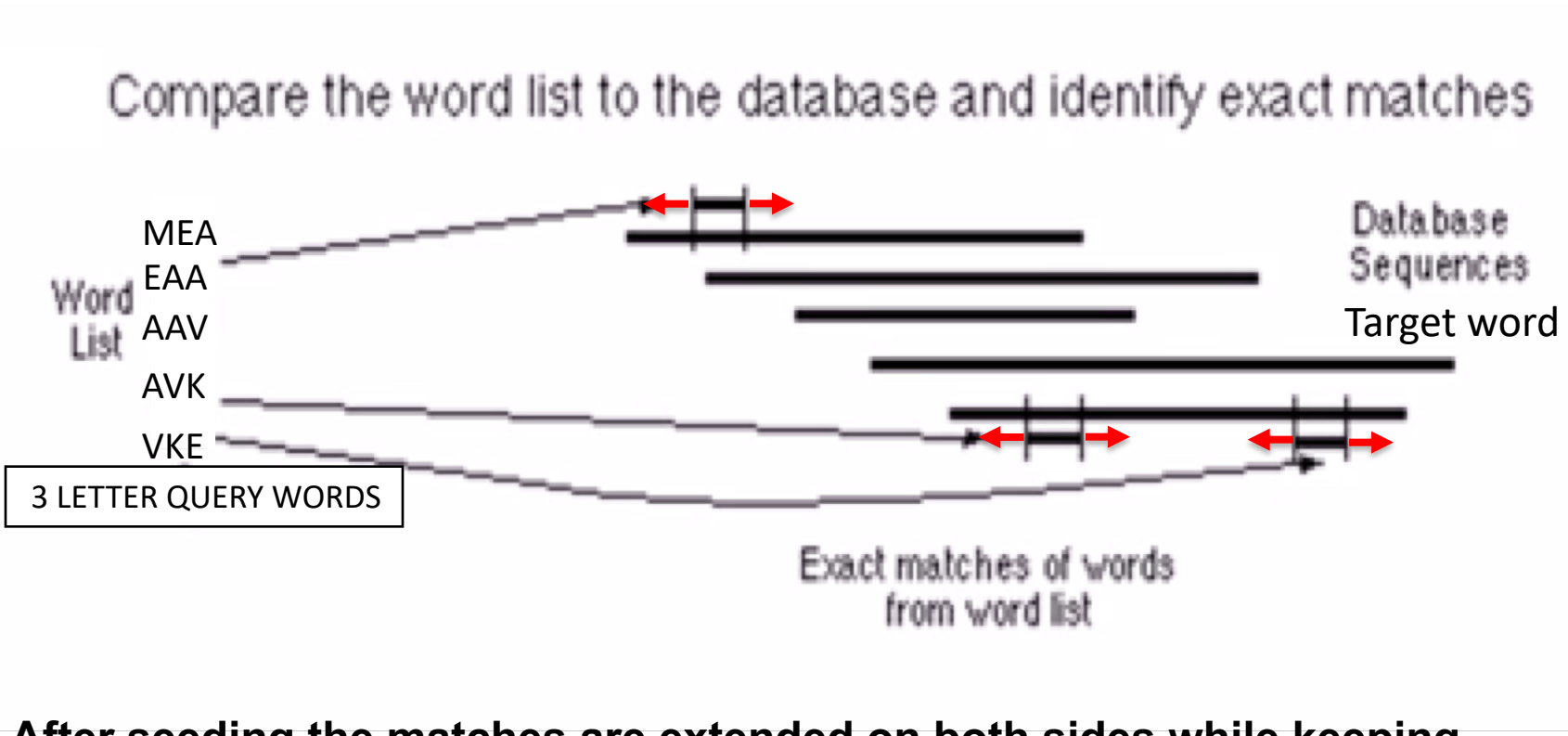
MEA
EAA
AAV
AVK
VKE

BREAK QUERY INTO WORDS



BLAST finds similar sequences by locating short lengths of exact matches between the two sequences in the database chosen.

M= Methionine, E = Gultamic acid, A = Alanine



1. **After seeding the matches are extended on both sides while keeping score**
2. **Each extension impacts the score (up or down)**
3. **The score is added up continuously and must reach a minimum threshold T value.**
4. **The threshold score T determines whether or not a particular word will be included in the alignment and carried forward.**

Example of a nr-BLAST Alignment
Interpreting a Match

Download GenPept Graphics

chromosomal replication initiation protein [Ornithinimicrobium pekingense]

Sequence ID: ref|WP_022920049.1| Length: 490 Number of Matches: 1

Range 1: 3 to 490 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
610 bits(1574)	0.0	Compositional matrix adjust.	315/503(63%)	376/503(74%)	15/503(2%)

```
Query  2    SQTPDDHATAIWQEAMVHLQGAGLAPRDIGVLRLATLVGLLEGTALLAVKYDHVKDAVEG  61
            SQ+P + A  +WQ  +  L+  G+  RD   LRL  LVGLL+ TALLAV Y H K+ +E
Sbjct  3    SQSPAESA-EVWQRVVSQLESQGVTARDRAFLRLTQLVGLLDTTALLAVPYQHTKETLET  61

Query  62   HLREDVSTALAEVLDRDIRLAVSVDPDAVSAAQEEAAPPAPSPADEDDPATGEGPLSTAV  121
             LR+ +  ALA  L  D+RLA++VD D     ++E  P AP PA      T + P  +
Sbjct  62   TLRQPIVDALAGELGHDVRLAITVDEDLRRQVEDEGDP-APGPA-----VTEQVP--SDP  113

Query  122  DGAVEKHEGSSPARAGESVAPATTASLTATNSSPGVERDYSALNHKYTFDTFVLGSSNRF  181
            D    +  G+ P   GE   P +    T + +      + + LN KYTFDTFV GSSNRF
Sbjct  114  DRTPYRSNGAGP---GE---PRSDGHRTPSGAVQTASAEDARLNPKYTFDTFVSGSSNRF  167

Query  182  AHAAATAVAEAPARAYNPLFIYGGSGLGKTHLLHAIGHYARTLDSSVRVKYVNSEEFTNQ  241
            AHAA+ AVAE+PARAYNPLFIYG SGLGKTHLLHAIGHYAR+L   VRV+YVNSEEFTN
Sbjct  168  AHAASLAVAESPARAYNPLFIYGESGLGKTHLLHAIGHYARSLYPGVRVRYVNSEEFTND  227

Query  242  FINAVSAGQANAFQRQYRDVDVLLIDDIQFLQGKEQTMEEFFHTFNTLHNSEKQIVITSD  301
            FIN++   +A AFQR+YR+VD LL+DDIQFLQGKEQT+EEFFHTFNTLHNSEKQ+VITSD
Sbjct  228  FINSIRDDKAGAFQRRYRNVDFLLVDDIQFLQGKEQTVEEFFHTFNTLHNSEKQVVITSD  287

Query  302  QPPKKLSGFAERMRSRFEWGLLTDVQPPDLETRIAILRRKAAADKLDIPDDVLHLIASKI  361
            QPPK+LSGFAERMRSRFEWGLLTDVQPPDLETRIAIL++KAA + + +PD+VL LI SKI
Sbjct  288  QPPKRLSGFAERMRSRFEWGLLTDVQPPDLETRIAILKKKAAQEGMQLPDEVLELIGSKI  347

Query  362  SSNIRELEGALTRVTAFASLSGSPLDEYLARTVLKDVMPGGDSGQITPTMILEETAGYFV  421
            S+NIRELEGAL RVTAFASLS +P D  LA  VLKD++P  +S  IT   I+ E A YF
Sbjct  348  STNIRELEGALIRVTAFASLSSTPPDAALASHVLKDIIPNSESAAITVPTIMAEVADYFQ  407

Query  422  ISVEEIQGASRSRNLTRARQIAMYLCRELTDLSLPKIGKEFGGRDHTTVMHAERKIKQLL  481
            IS +++ G SRSR L  ARQIAMYLCRELTDLSLPKIG+EFGGRDHTTVMHAERKI+QL+
Sbjct  408  ISNDDLCGTSRSRTLVNARQIAMYLCRELTDLSLPKIGQEFGGRDHTTVMHAERKIRQLI  467

Query  482  GEDRRVYDEVSELTSIIRKKAAR  504
            GE R +YD+++ELT IIRK +AR
Sbjct  468  GERRALYDQITELTGIIRKASAR  490
```

Query : Your gene/protein –aa seq
Sbject: BLAST match in database

The line between these two sequences will tell the extent of match.

EXACT MATCH: the same amino acid is indicated

Similar biochemical properties: + indicated.

Total mismatch: No letter

To get better alignment BLAST can also introduce gaps, indicated by a series of –
Why do gaps exist?

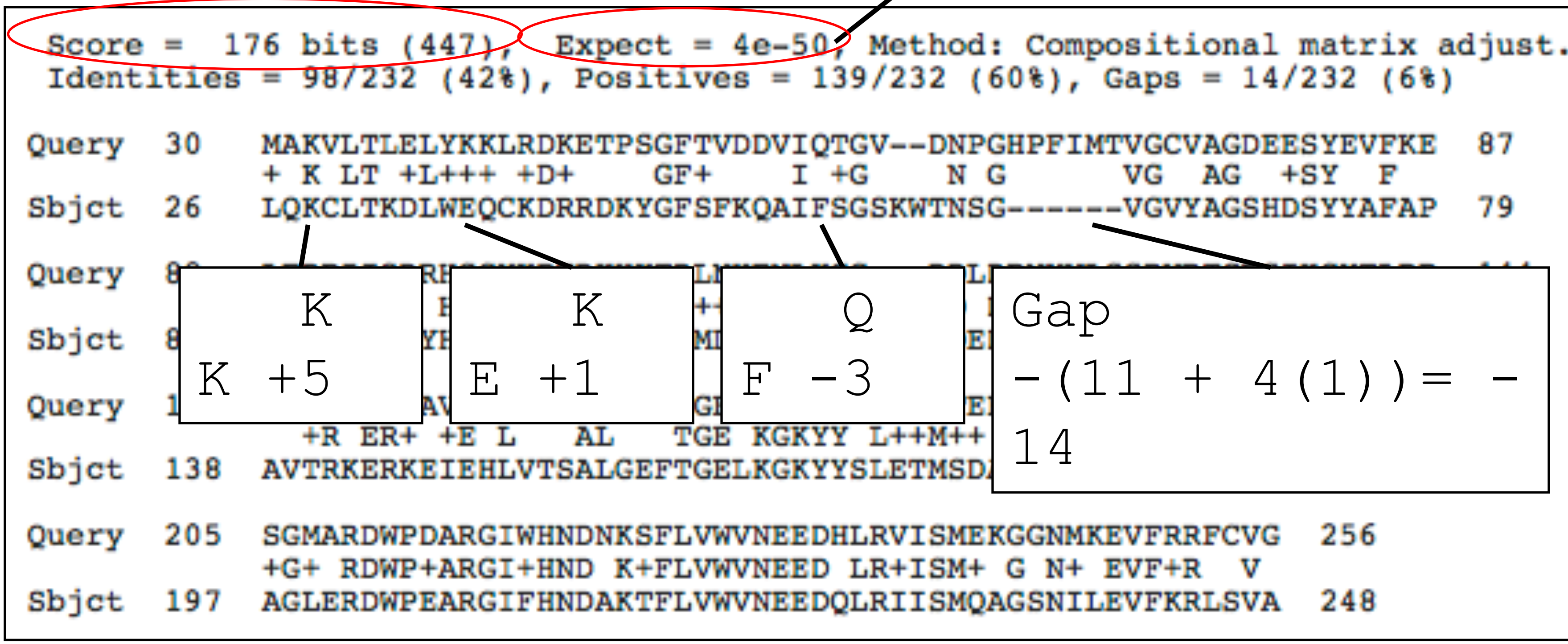
T=Threonine ; S=serine (first +)
D=Aspartic acid; E = Glutamic acid (second +)
Q=Glutamine

The gaps might represent insertion or deletion mutations that have occurred over evolution in one or the other protein.

Random Example-Scoring

Query : Your gene /aa seq
Sbject: BLAST match in database

Number of Chance Alignments = 4 X 10^{-50}



K=Lysine, E= Glutamic acid, Q= Glutamine, F= Phenylalanine,

Scores from BLOSUM62, a position independent matrix

Statistical significance

The Expect value (E)

- A statistical value that shows whether the matches that BLAST found is "expected" to be observed by chance.

  -Takes into act. total number of residues of the query seq and total number of residues in the database, scores, alignment.

- The lower the E-value, **or the closer it is to zero,** the more "significant" the match because the more unlikely that the match is **simply by chance.**

- The E-value cut off for this course is E-03 – SIGNIFICANT

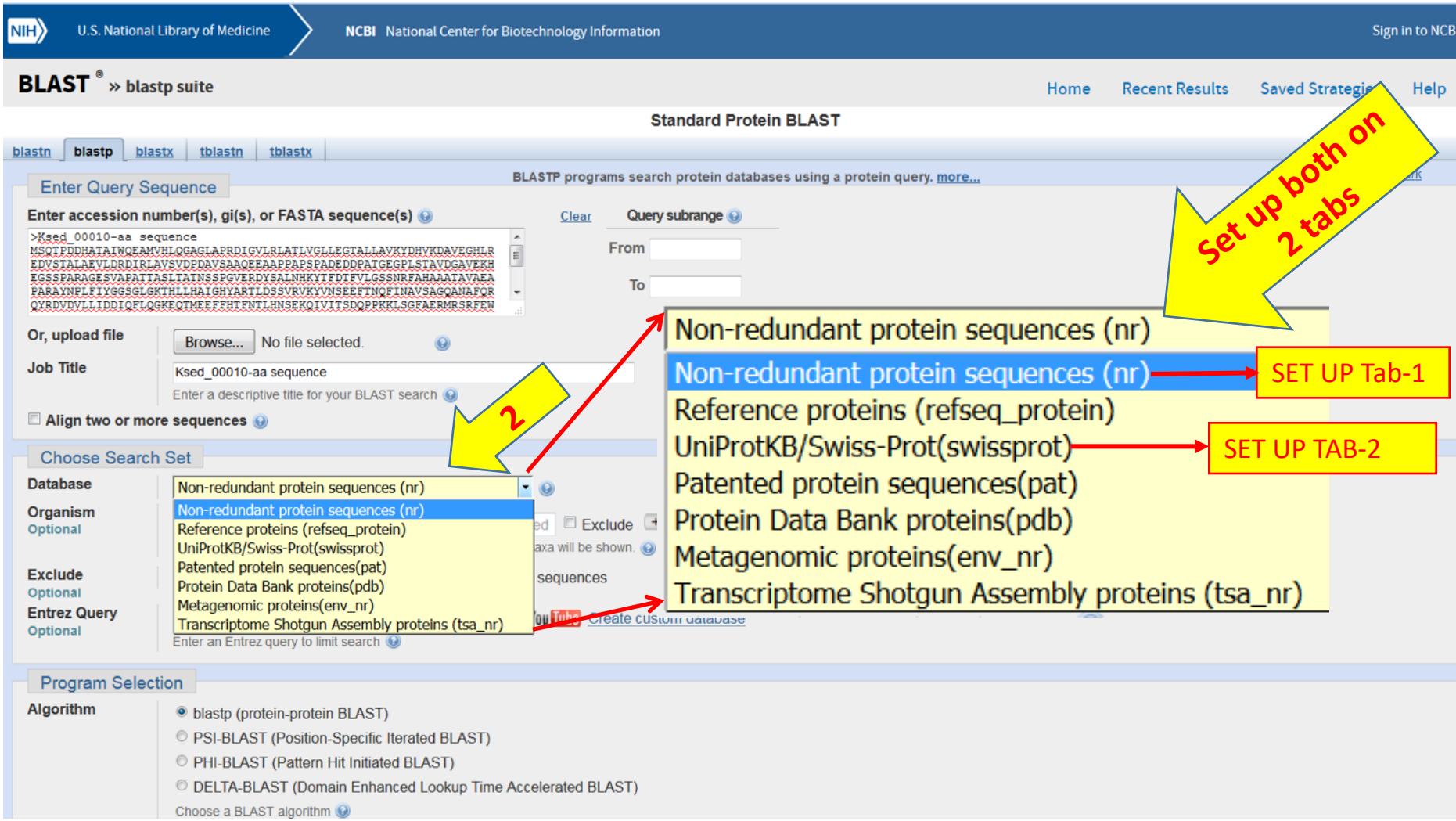
- E = $1x10^{-3}$ or is = 0.001 is thus expected to occur by chance 1 in 1,000 times

- E value equal to or less than 10^{-15} may indicate good match.

Be CAREFUL of mindless BLAST

- Believing that E tells the whole story.
- Ignoring length of match since calculation of the E value takes into account the length of the query sequence.

2 Databases to use in BLAST searches WHY?



HANDS ON…….

Compare the results obtained from both the nr and Swiss-Prot searches. Things to keep in mind as you compare the results are:

Things to keep in mind as you compare the results are:

A. Do both searches give significant results (as indicated by low E-values and high scores described below)?
B. Are the names of the significant hits in both searches identical or very similar?
    i. If the answer to both a and b above are yes, then you should use only the Swiss-Prot results to record in your notebook.
    ii. If no significant hits are found using SwissProt, but are found in nr, record that fact in your notebook and use the nr database.
    iii. If significant hits are found in BOTH databases, but the names given to each seem to be different, or the e-values and scores are significantly better in the nr database than in the Swiss-Prot database, you should record results for the top 2 BLAST hits in Swiss-Prot and nr in the lab notebook.

Notebook

Sequence-based Similarity Data Module

Module Instructions

BLAST

go to http://www.ncbi.nlm.nih.gov/blast

Gene product name (*top hit*)

Organism

`save for annotation report`

Alignment length

`save for annotation report`

Score

`score`

E-Value

`save for annotation report`

Alignment of the top hit and the query sequence

`alignment`

Gene product name (*second hit*)

`save for annotation report`

Organism

`save for annotation report`

Alignment length

`save for annotation report`

BLAST RESULTS - Swissprot database
Top of the results page

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]

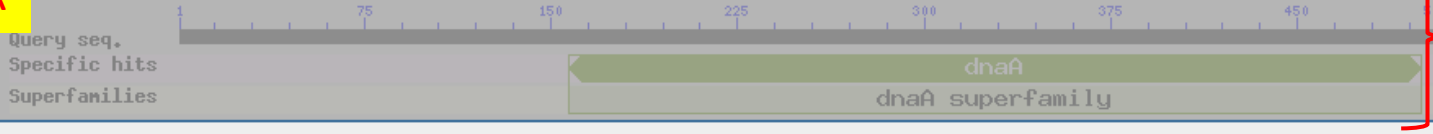
New Analyze your query with SmartBLAST

Graphic Summary

Show Conserved Domains

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

LATER



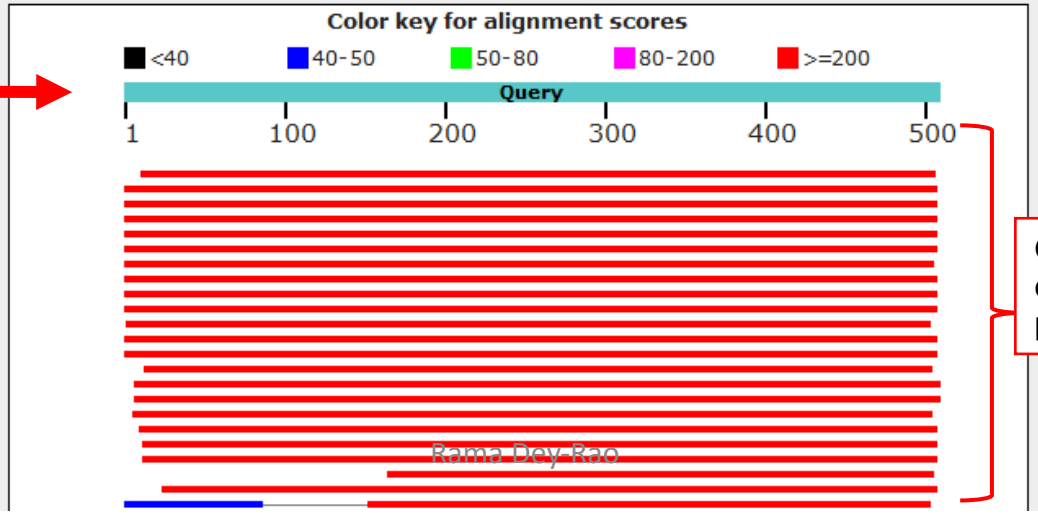
CDD search (conserved domain database)

residues in sequence in scale below the teal line labeled query.

A high score and close to 100% coverage would indicate a high quality alignment. This indicates several significantly similar proteins (orthologs) in different organisms that were found in the SP database (studied in the lab).

Distribution of the top 106 Blast Hits on 100 subject sequences

Mouse over to see the title, click to show alignments



Quick scan: Visual rep of coverage in orthologous proteins

Questions/comments

BLAST RESULTS - Swissprot database
Top of the results page

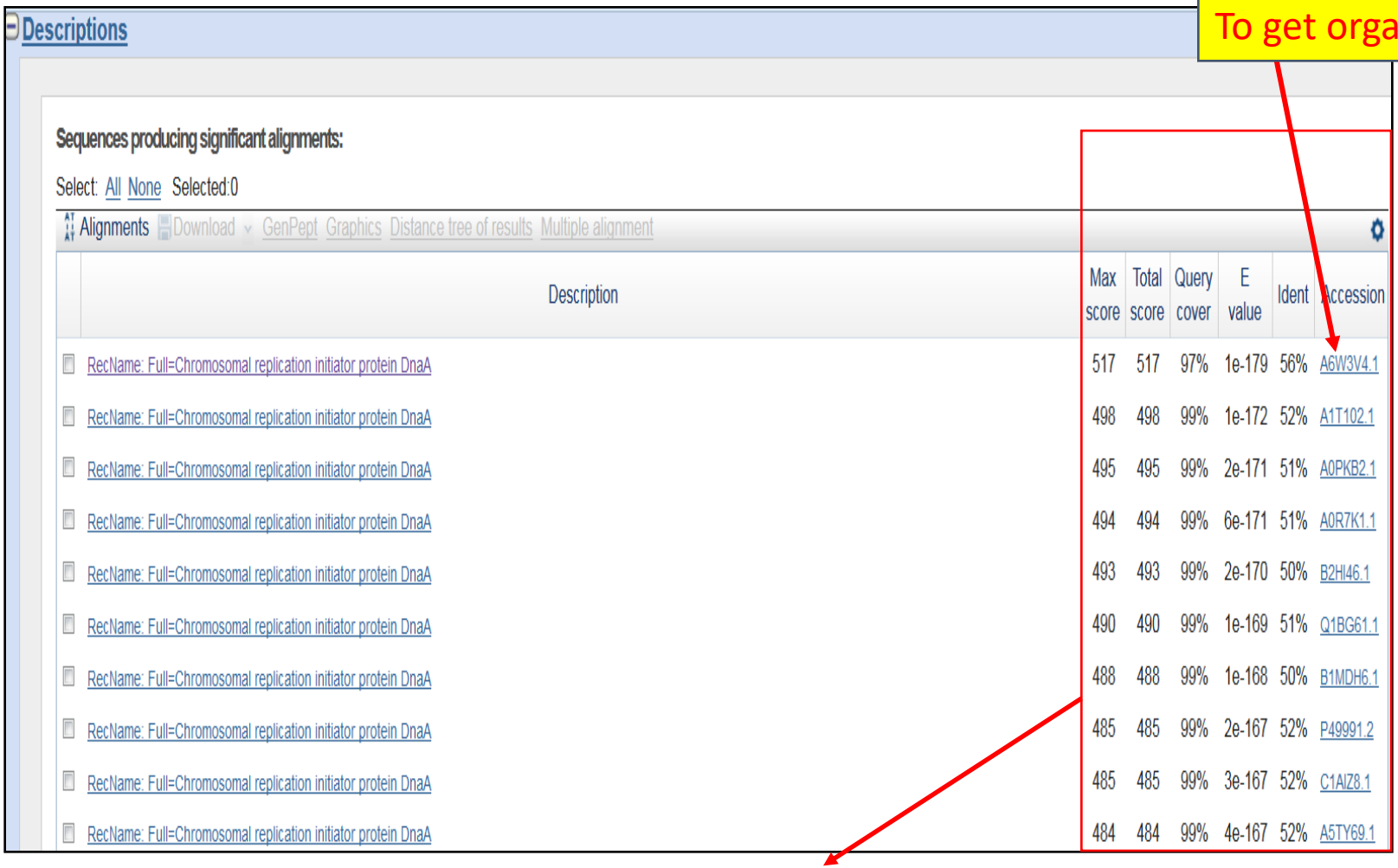
Mouse over to see the title, click to show alignments

SCROLL BELOW

Scroll below to see the top BLAST hits (second section of results page)

Swissprot database

Click
To get organism



Descriptions

Sequences producing significant alignments:

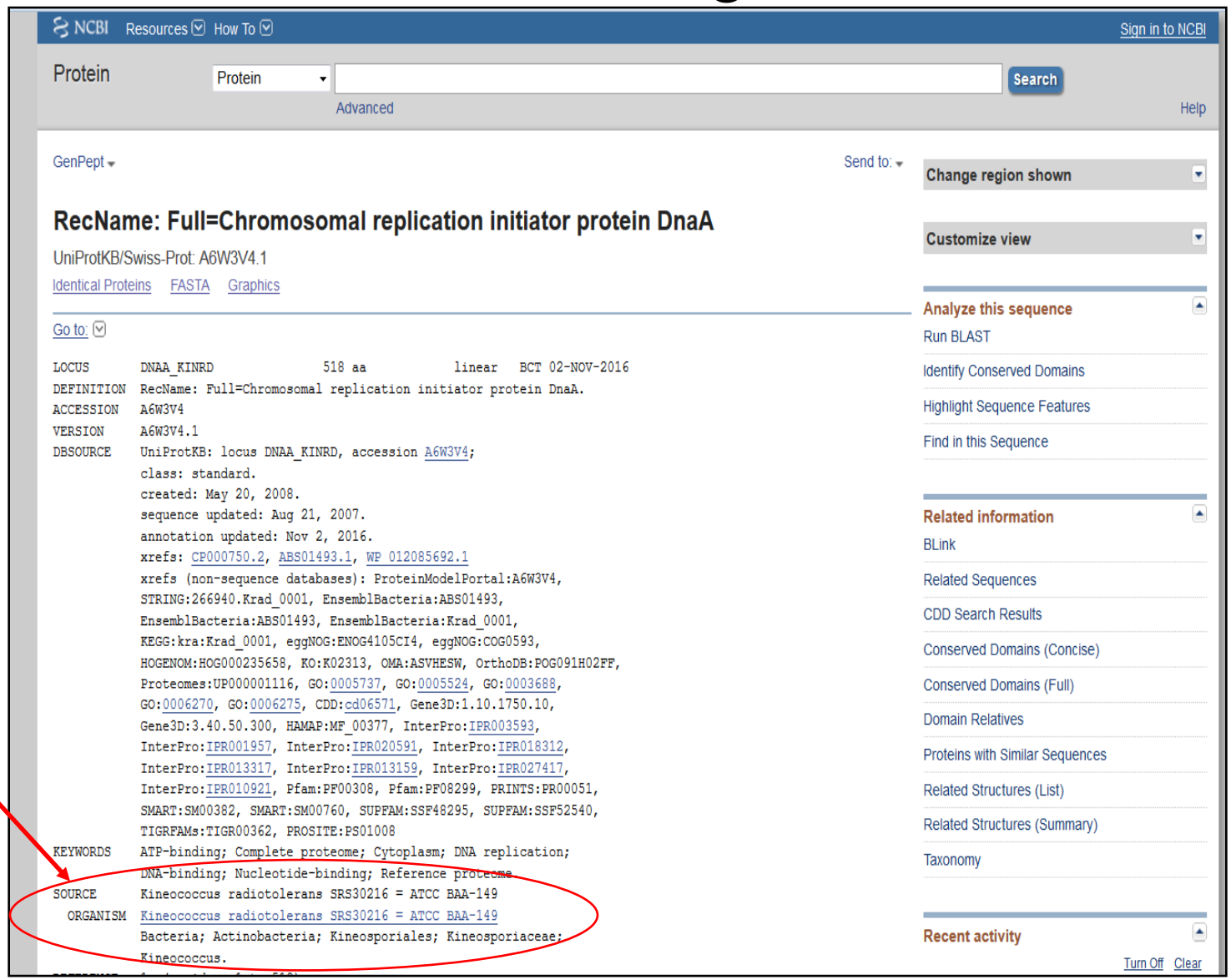
Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

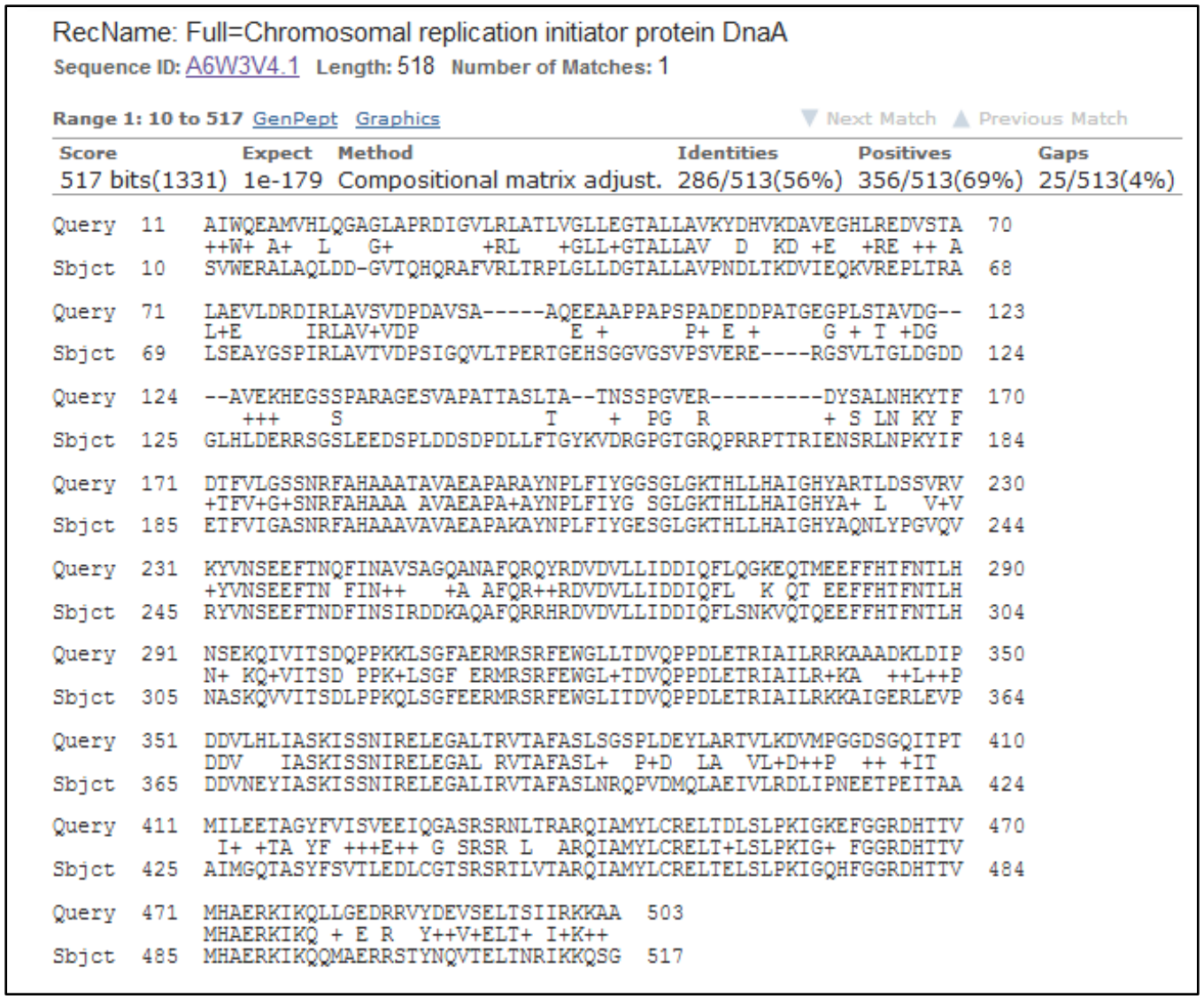
Description	Max score	Total score	Query cover	E value	Ident	Accession
RecName: Full=Chromosomal replication initiator protein DnaA	517	517	97%	1e-179	56%	A6W3V4.1
RecName: Full=Chromosomal replication initiator protein DnaA	498	498	99%	1e-172	52%	A1T102.1
RecName: Full=Chromosomal replication initiator protein DnaA	495	495	99%	2e-171	51%	A0PKB2.1
RecName: Full=Chromosomal replication initiator protein DnaA	494	494	99%	6e-171	51%	A0R7K1.1
RecName: Full=Chromosomal replication initiator protein DnaA	493	493	99%	2e-170	50%	B2HI46.1
RecName: Full=Chromosomal replication initiator protein DnaA	490	490	99%	1e-169	51%	Q1BG61.1
RecName: Full=Chromosomal replication initiator protein DnaA	488	488	99%	1e-168	50%	B1MDH6.1
RecName: Full=Chromosomal replication initiator protein DnaA	485	485	99%	2e-167	52%	P49991.2
RecName: Full=Chromosomal replication initiator protein DnaA	485	485	99%	3e-167	52%	C1AIZ8.1
RecName: Full=Chromosomal replication initiator protein DnaA	484	484	99%	4e-167	52%	A5TY69.1

The score, % coverage of the query and E value are shown, along with a hyperlink to the file describing the hit (Accession column).

Click A6W3V4.1 for Source organism



BLAST Alignment Details



RecName: Full=Chromosomal replication initiator protein DnaA

Sequence ID: A6W3V4.1 **Length:** 518 **Number of Matches:** 1

Range 1: 10 to 517 GenPept Graphics ▼ Next Match ▲ Previous Match

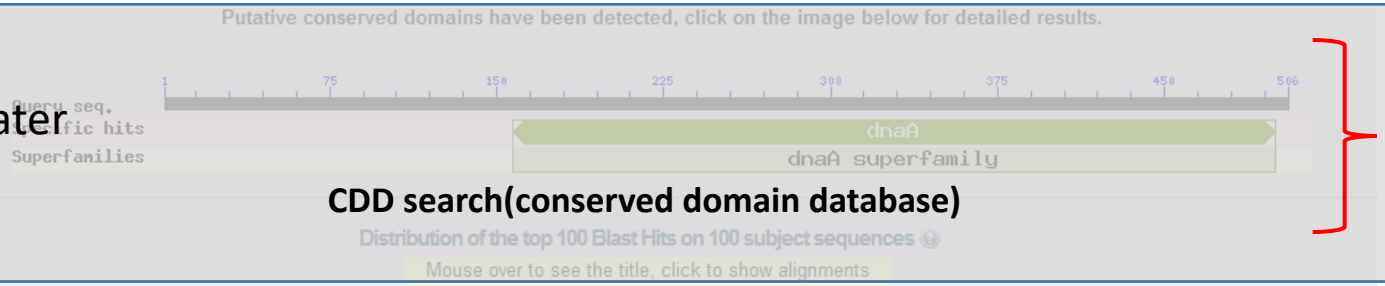
Score	Expect	Method	Identities	Positives	Gaps
517 bits(1331)	1e-179	Compositional matrix adjust.	286/513(56%)	356/513(69%)	25/513(4%)

```
Query  11   AIWQEAMVHLQGAGLAPRDIGVLRLATLVGLLEGTALLAVKYDHVKDAVEGHLREDVSTA  70
            ++W+ A+  L   G+       +RL   +GLL+GTALLAV  D  KD +E  +RE ++ A
Sbjct  10   SVWERALAQLDD-GVTQHQRAFVRLTRPLGLLDGTALLAVPNDLTKDVIEQKVREPLTRA  68

Query  71   LAEVLDRDIRLAVSVDPDAVSA-----AQEEAAPPAPSPADEDDPATGEGPLSTAVDG--  123
            L+E     IRLAV+VDP             E +      P+ E +     G + T +DG
Sbjct  69   LSEAYGSPIRLAVTVDPSIGQVLTPERTGEHSGGVGSVPSVERE----RGSVLTGLDGDD  124

Query  124  --AVEKHEGSSPARAGESVAPATTASLTA--TNSSPGVER---------DYSALNHKYTF  170
               +++    S                T    +  PG  R          + S LN KY F
Sbjct  125  GLHLDERRSGSLEEDSPLDDSDPDLLFTGYKVDRGPGTGRQPRRPTTRIENSRLNPKYIF  184

Query  171  DTFVLGSSNRFAHAAATAVAEAPARAYNPLFIYGGSGLGKTHLLHAIGHYARTLDSSVRV  230
            +TFV+G+SNRFAHAAA AVAEAPA+AYNPLFIYG SGLGKTHLLHAIGHYA+ L   V+V
Sbjct  185  ETFVIGASNRFAHAAAVAVAEAPAKAYNPLFIYGESGLGKTHLLHAIGHYAQNLYPGVQV  244

Query  231  KYVNSEEFTNQFINAVSAGQANAFQRQYRDVDVLLIDDIQFLQGKEQTMEEFFHTFNTLH  290
            +YVNSEEFTN FIN++    +A AFQR++RDVDVLLIDDIQFL   K QT EEFFHTFNTLH
Sbjct  245  RYVNSEEFTNDFINSIRDDKAQAFQRRHRDVDVLLIDDIQFLSNKVQTQEEFFHTFNTLH  304

Query  291  NSEKQIVITSDQPPKKLSGFAERMRSRFEWGLLTDVQPPDLETRIAILRRKAAADKLDIP  350
            N+ KQ+VITSD PPK+LSGF ERMRSRFEWGL+TDVQPPDLETRIAILR+KA  ++L++P
Sbjct  305  NASKQVVITSDLPPKQLSGFEERMRSRFEWGLITDVQPPDLETRIAILRKKAIGERLEVP  364

Query  351  DDVLHLIASKISSNIRELEGALTRVTAFASLSGSPLDEYLARTVLKDVMPGGDSGQITPT  410
            DDV   IASKISSNIRELEGAL RVTAFASL+  P+D  LA  VL+D++P  ++ +IT
Sbjct  365  DDVNEYIASKISSNIRELEGALIRVTAFASLNRQPVDMQLAEIVLRDLIPNEETPEITAA  424

Query  411  MILEETAGYFVISVEEIQGASRSRNLTRARQIAMYLCRELTDLSLPKIGKEFGGRDHTTV  470
             I+ +TA YF +++E++ G SRSR L  ARQIAMYLCRELT+LSLPKIG+ FGGRDHTTV
Sbjct  425  AIMGQTASYFSVTLEDLCGTSRSRTLVTARQIAMYLCRELTELSLPKIGQHFGGRDHTTV  484

Query  471  MHAERKIKQLLGEDRRVYDEVSELTSIIRKKAA  503
            MHAERKIKQ + E R  Y++V+ELT+ I+K++
Sbjct  485  MHAERKIKQQMAERRSTYNQVTELTNRIKKQSG  517
```

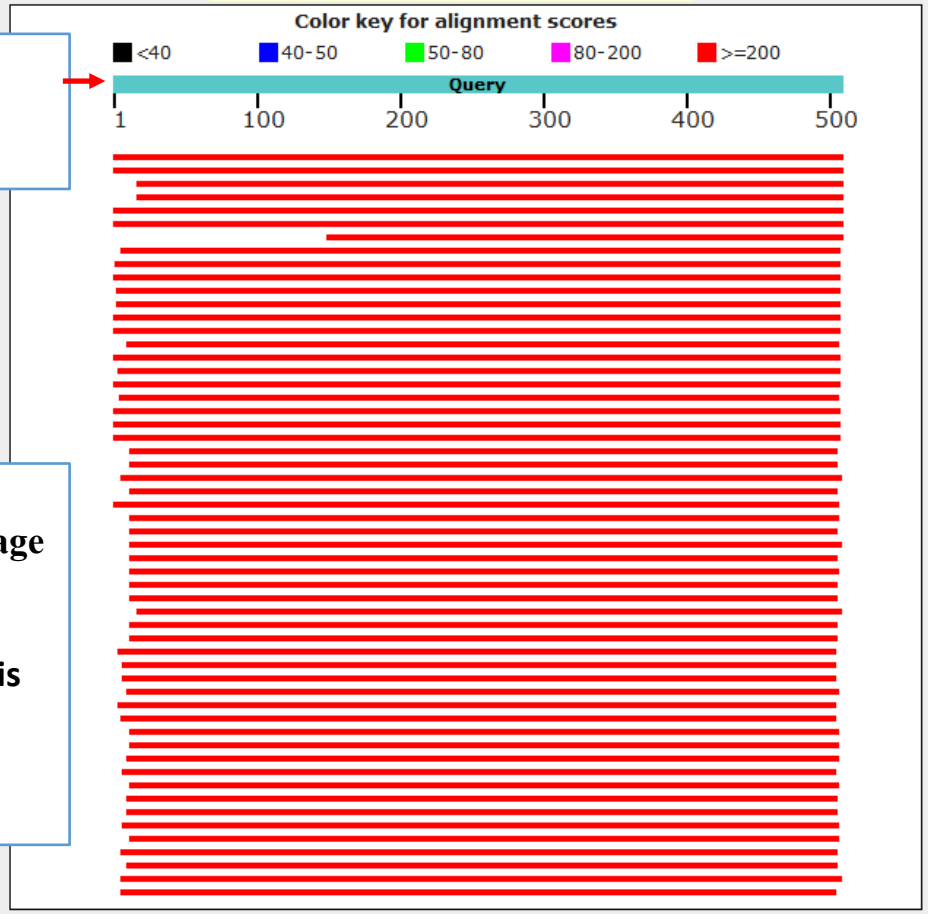
BLAST RESULTS PAGE- nr database

Both a Conserved Domain Database (CDD Results) and BLAST searches are done simultaneously.

Will do later



CDD search(conserved domain database)

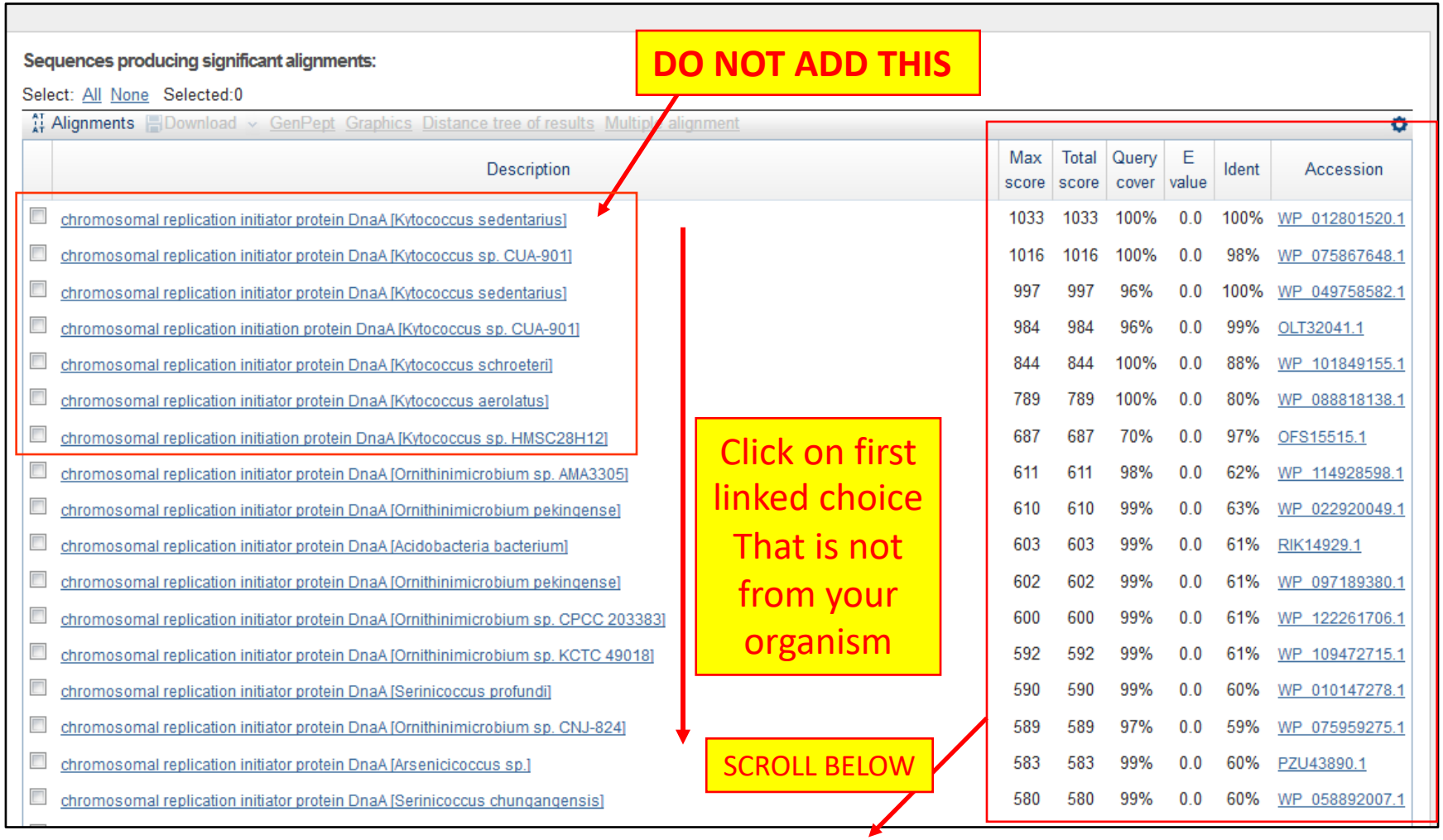


residues in sequence in scale below the teal line labeled query.

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

Scroll below to see the top BLAST hits (nr database)



Description	Max score	Total score	Query cover	E value	Ident	Accession
chromosomal replication initiator protein DnaA [Kytococcus sedentarius]	1033	1033	100%	0.0	100%	WP_012801520.1
chromosomal replication initiator protein DnaA [Kytococcus sp. CUA-901]	1016	1016	100%	0.0	98%	WP_075867648.1
chromosomal replication initiator protein DnaA [Kytococcus sedentarius]	997	997	96%	0.0	100%	WP_049758582.1
chromosomal replication initiation protein DnaA [Kytococcus sp. CUA-901]	984	984	96%	0.0	99%	OLT32041.1
chromosomal replication initiator protein DnaA [Kytococcus schroeteri]	844	844	100%	0.0	88%	WP_101849155.1
chromosomal replication initiator protein DnaA [Kytococcus aerolatus]	789	789	100%	0.0	80%	WP_088818138.1
chromosomal replication initiation protein DnaA [Kytococcus sp. HMSC28H12]	687	687	70%	0.0	97%	OFS15515.1
chromosomal replication initiator protein DnaA [Ornithinimicrobium sp. AMA3305]	611	611	98%	0.0	62%	WP_114928598.1
chromosomal replication initiator protein DnaA [Ornithinimicrobium pekingense]	610	610	99%	0.0	63%	WP_022920049.1
chromosomal replication initiator protein DnaA [Acidobacteria bacterium]	603	603	99%	0.0	61%	RIK14929.1
chromosomal replication initiator protein DnaA [Ornithinimicrobium pekingense]	602	602	99%	0.0	61%	WP_097189380.1
chromosomal replication initiator protein DnaA [Ornithinimicrobium sp. CPCC 203383]	600	600	99%	0.0	61%	WP_122261706.1
chromosomal replication initiator protein DnaA [Ornithinimicrobium sp. KCTC 49018]	592	592	99%	0.0	61%	WP_109472715.1
chromosomal replication initiator protein DnaA [Serinicoccus profundi]	590	590	99%	0.0	60%	WP_010147278.1
chromosomal replication initiator protein DnaA [Ornithinimicrobium sp. CNJ-824]	589	589	97%	0.0	59%	WP_075959275.1
chromosomal replication initiator protein DnaA [Arsenicicoccus sp.]	583	583	99%	0.0	60%	PZU43890.1
chromosomal replication initiator protein DnaA [Serinicoccus chungangensis]	580	580	99%	0.0	60%	WP_058892007.1

The score, % coverage of the query and E value are shown, along with a hyperlink to the Genbank file describing the hit (Accession column).

Scroll below to see BLAST hits (nr database)

Ksed_00010
OID : 644990317

Score

Gene product name (*top hit*)

chromosomal replication initiation protein DnaA [Ornithinimicrobium pekingense]

Organism

Sequence ID: WP_022920049.1 Length: 490 Number of Matches: 1

Range 1: 3 to 490 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
610 bits(1574)	0.0	Compositional matrix adjust.	315/503(63%)	376/503(74%)	15/503(2%)

```
Query  2    SQTPDDIATAIWQEAMVHLQGAGLAPRDIGVLRLATLVGLLEGTALLAVKYDHVKDAVEG  61
            SQ+P + A  +WQ  +  L+  G+  RD   LRL  LVGLL+ TALLAV Y H K+ +E
Sbjct  3    SQSPAESA-FVWQRVVSQLESQGVTARDRAFLRLTQLVGLLDTTALLAVPYQHTKETLET  61

Query  62   HLREDVSTALAEVLDRDIRLAVSVDPDAVSAAQEEAAPPAPSPADEDDPATGEGPLSTAV  121
             LR+ +  ALA  L  D+RLA++VD D      ++E  P AP PA      T + P  +
Sbjct  62   TLRQPIVDALAGELGHDVRLAITVDEDLRRQVEDEGDP-APGPA-----VTEQVP--SDP  113

Query  122  DGAVEKHEGSSPARAGESVAPATTASLTATNSSPGVERDYSALNHKYTFDTFVLGSSNRF  181
            D    +  G+ P   GE   P +    T + +      + + LN KYTFDTFV GSSNRF
Sbjct  114  DRTPYRSNGAGP---GE---PRSDGHRTPSGAVQTASAEDARLNPKYTFDTFVSGSSNRF  167

Query  182  AHAAATAVAEAPARAYNPLFIYGGSGLGKTHLLHAIGHYARTLDSSVRVKYVNSEEFTNQ  241
            AHAA+ AVAE+PARAYNPLFIYG SGLGKTHLLHAIGHYAR+L   VRV+YVNSEEFTN
Sbjct  168  AHAASLAVAESPARAYNPLFIYGESGLGKTHLLHAIGHYARSLYPGVRVRYVNSEEFTND  227

Query  242  FINAVSAGQANAFQRQYRDVDVLLIDDIQFLQGKEQTMEEFFHTFNTLHNSEKQIVITSD  301
            FIN++   +A AFQR+YR+VD LL+DDIQFLQGKEQT+EEFFHTFNTLHNSEKQ+VITSD
Sbjct  228  FINSIRDDKAGAFQRRYRNVDFLLVDDIQFLQGKEQTVEEFFHTFNTLHNSEKQVVITSD  287

Query  302  QPPKKLSGFAERMRSRFEWGLLTDVQPPDLETRIAILRRKAAADKLDIPDDVLHLIASKI  361
            QPPK+LSGFAERMRSRFEWGLLTDVQPPDLETRIAIL++KAA + + +PD+VL LI SKI
Sbjct  288  QPPKRLSGFAERMRSRFEWGLLTDVQPPDLETRIAILKKKAAQEGMQLPDEVLELIGSKI  347

Query  362  SSNIRELEGALTRVTAFASLSGSPLDEYLARTVLKDVMPGGDSGQITPTMILEETAGYFV  421
            S+NIRELEGAL RVTAFASLS +P D  LA  VLKD++P  +S  IT   I+ E A YF
Sbjct  348  STNIRELEGALIRVTAFASLSSTPPDAALASHVLKDIIPNSESAAITVPTIMAEVADYFQ  407

Query  422  ISVEEIQGASRSRNLTRARQIAMYLCRELTDLSLPKIGKEFGGRDHTTVMHAERKIKQLL  481
            IS +++ G SRSR L  ARQIAMYLCRELTDLSLPKIG+EFGGRDHTTVMHAERKI+QL+
Sbjct  408  ISNDDLCGTSRSRTLVNARQIAMYLCRELTDLSLPKIGREFGGRDHTTVMHAERKIRQLM  467

Query  482  GEDRRVYDEVSELTSIIRKKAAR  504
            GE R +YD++++LT IIRK +AR
Sbjct  468  GERRALYDQITELTGIIRKASAR  490
```

Copy

Alignment Length
(Last Query #) – (Ist Query #) +1
In this case 504-2+1=503

Number of Chance Alignments =0

Problem: nr blast only gives you a list of the same bacteria or bacteria from the same genus.

- If you find that all the top blast hits are from the same organism that you are investigating, or all are from the same genus you will need to set up an ”exclusion blast” to exclude those very closely related or identical hits.
- This may happen if you are annotating a gene from a “clinically significant” (disease causing) bacterium. Such bacteria are likely to have many variants or isolates sequenced and in the database.
- Other non-clinically significant bacteria that are commonly studied may also have many strains sequenced in the database.

- Finding a match to protein an identical bacterium or very closely related species can happen simply because of the fact such organisms are evolutionarily closely related.
- We would prefer to find matches to proteins in bacteria other than those above mentioned above to be able to identify conserved domains or regions that might be important to protein function in multiple species.
- Instructions for doing an exclusion blast b can be found at the following link: https://drive.google.com/file/d/1PUKj_v8vYPxyG7h6cXSHZtpMfV1QonbA/view

Example of a nr-BLAST Alignment
Interpreting a match

Download GenPept Graphics

chromosomal replication initiation protein [Ornithinimicrobium pekingense]

Sequence ID: ref|WP_022920049.1| Length: 490 Number of Matches: 1

Range 1: 3 to 490 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
610 bits(1574)	0.0	Compositional matrix adjust.	315/503(63%)	376/503(74%)	15/503(2%)

```
Query  2    SQTPDDHATAIWQEAMVHLQGAGLAPRDIGVLRLATLVGLLEGTALLAVKYDHVKDAVEG  61
            SQ+P + A  +WQ  +  L+  G+  RD   LRL  LVGLL+ TALLAV Y H K+ +E
Sbjct  3    SQSPAESA-EVWQRVVSQLESQGVTARDRAFLRLTQLVGLLDTTALLAVPYQHTKETLET  61

Query  62   HLREDVSTALAEVLDRDIRLAVSVDPDAVSAAQEEAAPPAPSPADEDDPATGEGPLSTAV  121
             LR+ +  ALA  L  D+RLA++VD D     ++E  P AP PA      T + P  +
Sbjct  62   TLRQPIVDALAGELGHDVRLAITVDEDLRRQVEDEGDP-APGPA-----VTEQVP--SDP  113

Query  122  DGAVEKHEGSSPARAGESVAPATTASLTATNSSPGVERDYSALNHKYTFDTFVLGSSNRF  181
            D    +  G+ P   GE   P +    T + +      + + LN KYTFDTFV GSSNRF
Sbjct  114  DRTPYRSNGAGP---GE---PRSDGHRTPSGAVQTASAEDARLNPKYTFDTFVSGSSNRF  167

Query  182  AHAAATAVAEAPARAYNPLFIYGGSGLGKTHLLHAIGHYARTLDSSVRVKYVNSEEFTNQ  241
            AHAA+ AVAE+PARAYNPLFIYG SGLGKTHLLHAIGHYAR+L   VRV+YVNSEEFTN
Sbjct  168  AHAASLAVAESPARAYNPLFIYGESGLGKTHLLHAIGHYARSLYPGVRVRYVNSEEFTND  227

Query  242  FINAVSAGQANAFQRQYRDVDVLLIDDIQFLQGKEQTMEEFFHTFNTLHNSEKQIVITSD  301
            FIN++   +A AFQR+YR+VD LL+DDIQFLQGKEQT+EEFFHTFNTLHNSEKQ+VITSD
Sbjct  228  FINSIRDDKAGAFQRRYRNVDFLLVDDIQFLQGKEQTVEEFFHTFNTLHNSEKQVVITSD  287

Query  302  QPPKKLSGFAERMRSRFEWGLLTDVQPPDLETRIAILRRKAAADKLDIPDDVLHLIASKI  361
            QPPK+LSGFAERMRSRFEWGLLTDVQPPDLETRIAIL++KAA + + +PD+VL LI SKI
Sbjct  288  QPPKRLSGFAERMRSRFEWGLLTDVQPPDLETRIAILKKKAAQEGMQLPDEVLELIGSKI  347

Query  362  SSNIRELEGALTRVTAFASLSGSPLDEYLARTVLKDVMPGGDSGQITPTMILEETAGYFV  421
            S+NIRELEGAL RVTAFASLS +P D  LA  VLKD++P  +S  IT   I+ E A YF
Sbjct  348  STNIRELEGALIRVTAFASLSSTPPDAALASHVLKDIIPNSESAAITVPTIMAEVADYFQ  407

Query  422  ISVEEIQGASRSRNLTRARQIAMYLCRELTDLSLPKIGKEFGGRDHTTVMHAERKIKQLL  481
            IS +++ G SRSR L  ARQIAMYLCRELTDLSLPKIG+EFGGRDHTTVMHAERKI+QL+
Sbjct  408  ISNDDLCGTSRSRTLVNARQIAMYLCRELTDLSLPKIGQEFGGRDHTTVMHAERKIRQLI  467

Query  482  GEDRRVYDEVSELTSIIRKKAAR  504
            GE R +YD+++ELT IIRK +AR
Sbjct  468  GERRALYDQITELTGIIRKASAR  490
```

Query : Your gene/protein –aa seq
Sbject: BLAST match in database

The line between these two sequences will tell the extent of match.

EXACT MATCH: the same amino acid is indicated

Similar biochemical properties: + indicated.

Total mismatch: No letter

To get better alignment BLAST can also introduce gaps, indicated by a series of –

T=Threonine ; S=serine (first +)
D=Aspartic acid; E = Glutamic acid (second +)
Q=Glutamine

The gaps might represent insertion or deletion mutations that have occurred over evolution in one or the other protein.

Example of a BLAST Alignment

Score

Gene product name (*top hit*)

Ksed_00010 or RS 00010
OID : 644990317

Download GenPept Graphics

chromosomal replication initiation protein [Ornithinimicrobium pekingense]

Sequence ID: ref|WP_022920049.1| Length: 490 Number of Matches: 1

Range 1: 3 to 490 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
610 bits(1574)	0.0	Compositional matrix adjust.	315/503(63%)	376/503(74%)	15/503(2%)

```
Query  2    SQTPDDHATAIWQEVMVHLQGAGLAPRDIGVLRLATLVGLLEGTALLAVKYDHVKDAVEG  61
            SQ+P + A  +WQ  +  L+  G+  RD   LRL  LVGLL+ TALLAV Y H K+ +E
Sbjct  3    SQSPAESA-EVWQRVVSQLESQGVTARDRAFLRLTQLVGLLDTTALLAVPYQHTKETLET  61

Query  62   HLREDVSTALAEVLDRDIRLAVEVDPDAVSAAQEEAAPPAPSPADEDDPATGEGPLSTAV  121
             LR+ +  ALA  L  D+RLA++VE D      ++E  P AP PA      T + P  +
Sbjct  62   TLRQPIVDALAGELGHDVRLAITVDEDRRQVEDEGDP-APGPA-----VTEQVP--SDP  113

Query  122  DGAVEKHEGSSPARAGESVAPATTASLTATNSSPGVERDYSALNHKYTFDTFVLGSSNRF  181
            D    +  G+ P   GE   P +    T + +       + + LN KYTFDTFV GSSNRF
Sbjct  114  DRTPYRSNGAGP---GE---PRSDGHRTPSGAVQTASAEDARLNPKYTFDTFVSGSSNRF  167

Query  182  AHAAATAVAEAPARAYNPLFIYGGSGLGKTHLLHAIGHYARTLDSSVRVKYVNSEEFTNQ  241
            AHAA+ AVAE+PARAYNPLFIYG SGLGKTHLLHAIGHYAR+L   VRV+YVNSEEFTN
Sbjct  168  AHAASLAVAESPARAYNPLFIYGESGLGKTHLLHAIGHYARSLRPGVRVRYVNSEEFTND  227

Query  242  FINAVSAGQANAFQRQYRDVDVLLIDDIQFLQGKEQTMEEFFHTFNTLHNSEKQIVITSD  301
            FIN++   +A AFQR+YR+VD LL+DDIQFLQGKEQT+EEFFHTFNTLHNSEKQ+VITSD
Sbjct  228  FINSIRDDKAGAFQRRYRNVDFLLVDDIQFLQGKEQTVEEFFHTFNTLHNSEKQVVITSD  287

Query  302  QPPKKLSGFAERMRSRFEWGLLTDVQPPDLETRIAILRRKAAADKLDIPDDVLHLIASKI  361
            QPPK+LSGFAERMRSRFEWGLLTDVQPPDLETRIAIL++KAA + + +PD+VL LI S+I
Sbjct  288  QPPKRLSGFAERMRSRFEWGLLTDVQPPDLETRIAILKKKAAQEGMQLPDEVLELIGSKI  347

Query  362  SSNIRELEGALIRVTAFASLSGSPLDEYLARTVLKDVMPGGDSGQITPTMILEETAGYFV  421
            S+NIRELEGAL RVTAFASLS +P D  LA  VLKD++P  +S  IT   I+ E A YF
Sbjct  348  STNIRELEGALIRVTAFASLSSTPPDAALASHVLKDIIPNSESAAITVPTIMAEVADYFQ  407

Query  422  ISVEEIQGASRSRNLTRARQIAMYLCRELTDLSLPKIGKEFGGRDHTTVMHAERKIKQLL  481
            IS +++ G SRSR L  ARQIAMYLCRELTDLSLPKIG+EFGGRDHTTVMHAERKI+QL+
Sbjct  408  ISNDDLCGTSRSRTLVNARQIAMYLCRELTDLSLPKIGQEFGGRDHTTVMHAERKIRQLI  467

Query  482  GEDRRVYDEVSELTSIIRKKAAR  504
            GE R +YD+++ELT IIRK +AR
Sbjct  468  GERRALYDQITELTGIIRKASAR  490
```

Organism

E-value

Alignment Length

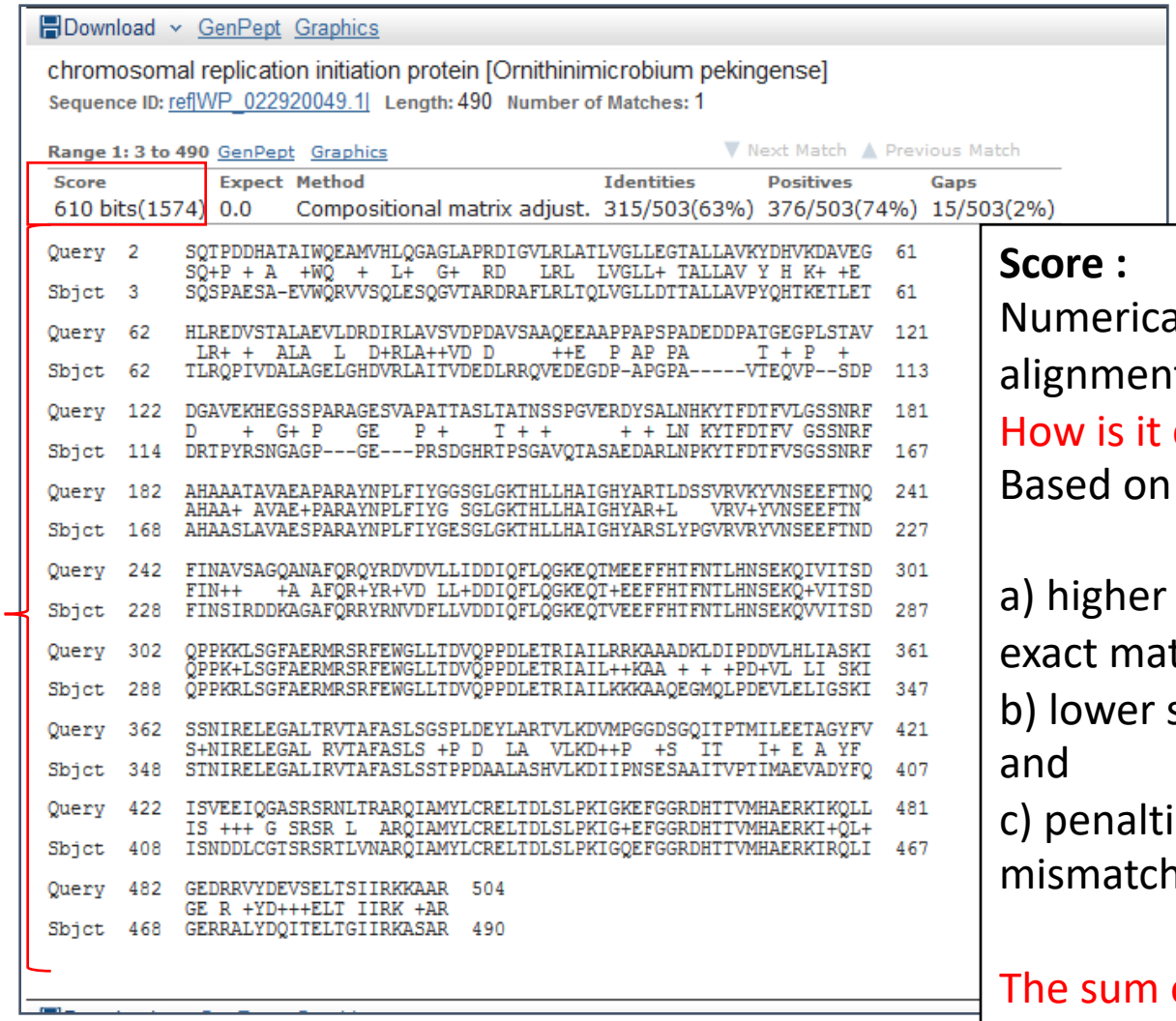
(Last Query #) – (Ist Query #) +1
In this case 504-2+1=503

Use Snipping or GRAB tool to frame and cut picture. Save as .png file and upload

>35% identity to experimentally characterized protein (especially in conserved regions) can be considered good evidence for function
E-value→ less than 10^{-3} is significant ; equal to or less than 10^{-15} may indicate good match

SCORE

#2 after *Kytococcus sedentarius*



Download GenPept Graphics

chromosomal replication initiation protein [Ornithinimicrobium pekingense]

Sequence ID: ref|WP_022920049.1| Length: 490 Number of Matches: 1

Range 1: 3 to 490 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
610 bits(1574)	0.0	Compositional matrix adjust.	315/503(63%)	376/503(74%)	15/503(2%)

```
Query  2    SQTPDDHATAIWQEAMVHLQGAGLAPRDIGVLRLATLVGLLEGTALLAVKYDHVKDAVEG  61
            SQ+P + A  +WQ  +  L+  G+  RD   LRL  LVGLL+ TALLAV Y H K+ +E
Sbjct  3    SQSPAESA-EVWQRVVSQLESQGVTARDRAFLRLTQLVGLLDTTALLAVPYQHTKETLET  61

Query  62   HLREDVSTALAEVLDRDIRLAVSVDPDAVSAAQEEAAPPAPSPADEDDPATGEGPLSTAV  121
             LR+ +  ALA  L  D+RLA++VD D     ++E  P AP PA      T + P  +
Sbjct  62   TLRQPIVDALAGELGHDVRLAITVDEDLRRQVEDEGDP-APGPA-----VTEQVP--SDP  113

Query  122  DGAVEKHEGSSPARAGESVAPATTASLTATNSSPGVERDYSALNHKYTFDTFVLGSSNRF  181
            D    +  G+ P   GE   P +    T + +      + + LN KYTFDTFV GSSNRF
Sbjct  114  DRTPYRSNGAGP---GE---PRSDGHRTPSGAVQTASAEDARLNPKYTFDTFVSGSSNRF  167

Query  182  AHAAATAVAEAPARAYNPLFIYGGSGLGKTHLLHAIGHYARTLDSSVRVKYVNSEEFTNQ  241
            AHAA+ AVAE+PARAYNPLFIYG SGLGKTHLLHAIGHYAR+L   VRV+YVNSEEFTN
Sbjct  168  AHAASLAVAESPARAYNPLFIYGESGLGKTHLLHAIGHYARSLYPGVRVRYVNSEEFTND  227

Query  242  FINAVSAGQANAFQRQYRDVDVLLIDDIQFLQGKEQTMEEFFHTFNTLHNSEKQIVITSD  301
            FIN++   +A AFQR+YR+VD LL+DDIQFLQGKEQT+EEFFHTFNTLHNSEKQ+VITSD
Sbjct  228  FINSIRDDKAGAFQRRYRNVDFLLVDDIQFLQGKEQTVEEFFHTFNTLHNSEKQVVITSD  287

Query  302  QPPKKLSGFAERMRSRFEWGLLTDVQPPDLETRIAILRRKAAADKLDIPDDVLHLIASKI  361
            QPPK+LSGFAERMRSRFEWGLLTDVQPPDLETRIAIL++KAA + + +PD+VL LI SKI
Sbjct  288  QPPKRLSGFAERMRSRFEWGLLTDVQPPDLETRIAILKKKAAQEGMQLPDEVLELIGSKI  347

Query  362  SSNIRELEGALTRVTAFASLSGSPLDEYLARTVLKDVMPGGDSGQITPTMILEETAGYFV  421
            S+NIRELEGAL RVTAFASLS +P D   LA  VLKD++P  +S  IT   I+ E A YF
Sbjct  348  STNIRELEGALIRVTAFASLSSTPPDAALASHVLKDIIPNSESAAITVPTIMAEVADYFQ  407

Query  422  ISVEEIQGASRSRNLTRARQIAMYLCRELTDLSLPKIGKEFGGRDHTTVMHAERKIKQLL  481
            IS +++ G SRSR L  ARQIAMYLCRELTDLSLPKIG+EFGGRDHTTVMHAERKI+QL+
Sbjct  408  ISNDDLCGTSRSRTLVNARQIAMYLCRELTDLSLPKIGQEFGGRDHTTVMHAERKIRQLI  467

Query  482  GEDRRVYDEVSELTSIIRKKAAR  504
            GE R +YD+++ELT IIRK +AR
Sbjct  468  GERRALYDQITELTGIIRKASAR  490
```

Score :

Numerical representation of quality of alignment

How is it calculated ? :

Based on how well the sequences match

a) higher numerical values assigned for exact matches

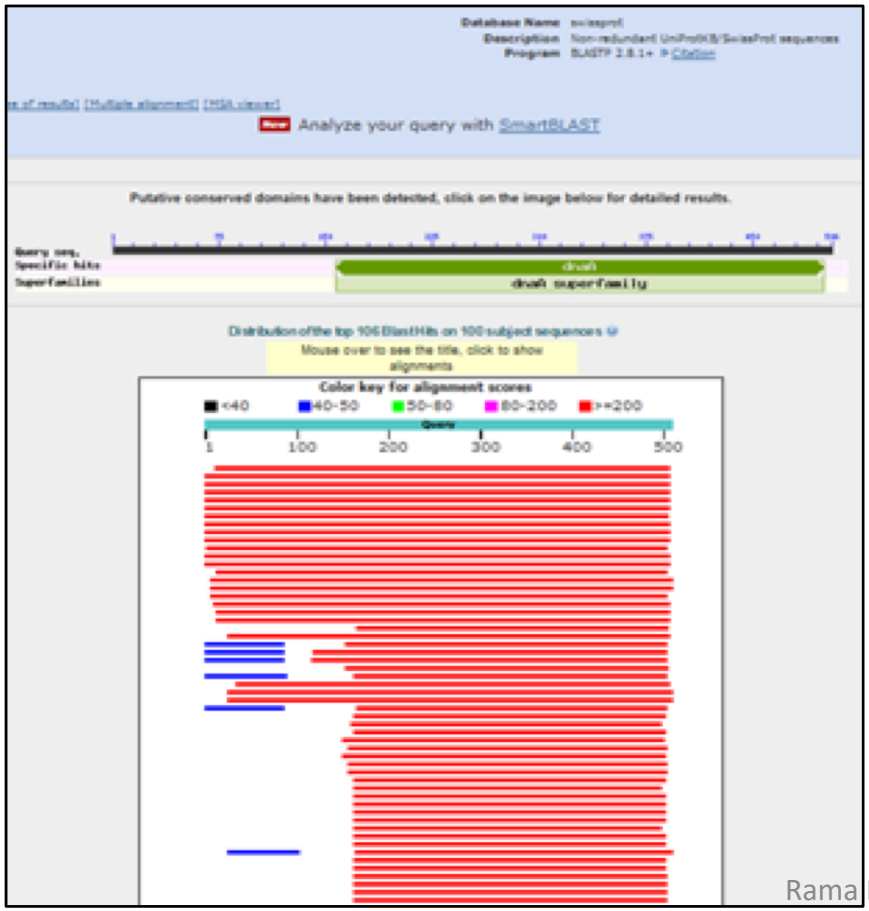
b) lower scores for "similar" amino acids and

c) penalties assigned for gaps and for mismatches.

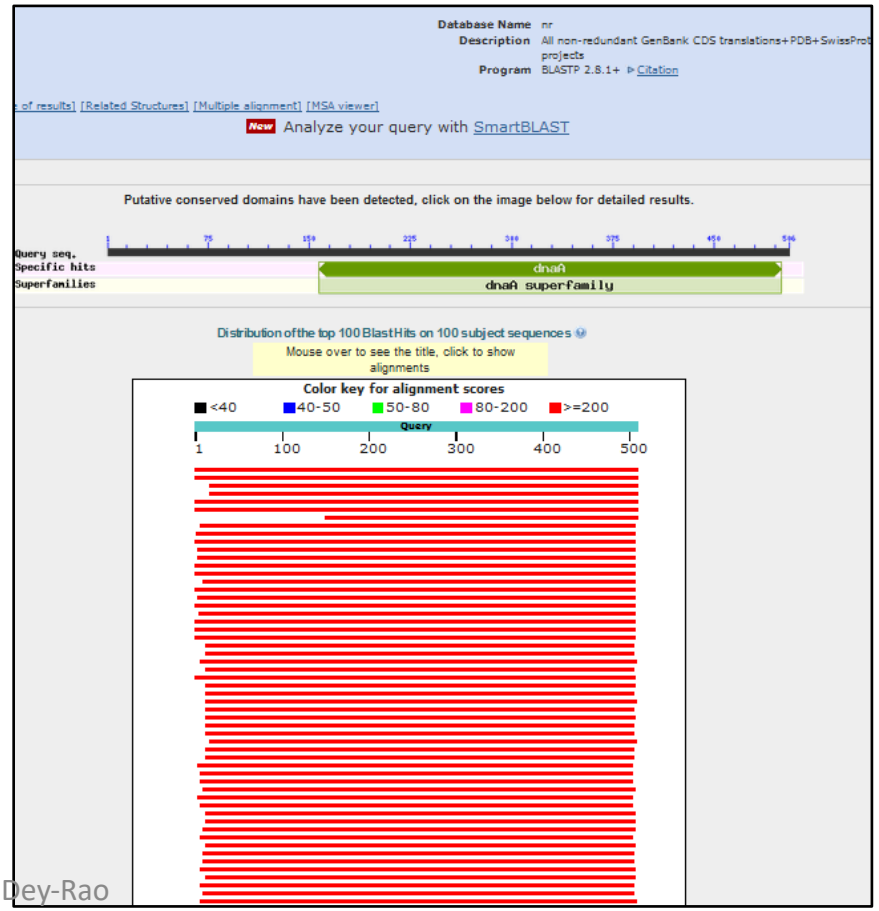
The sum of these numbers is the score. The higher the score, the more likely the alignment is significant.

Results:
Swissprot and nr databases

Swissprot database

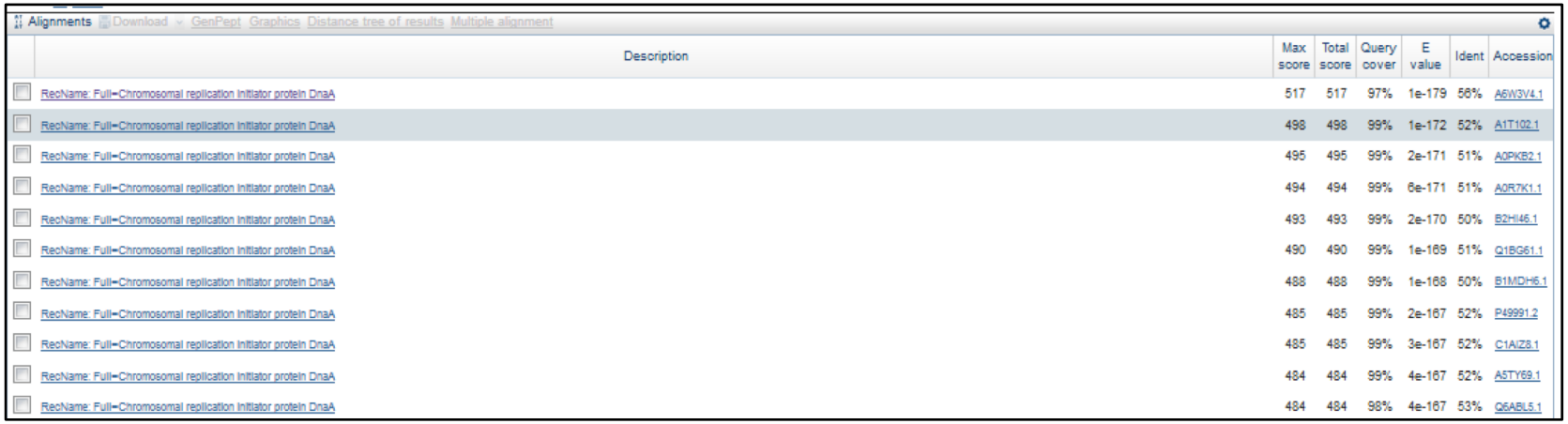


nr database



Results:
Swissprot and nr databases

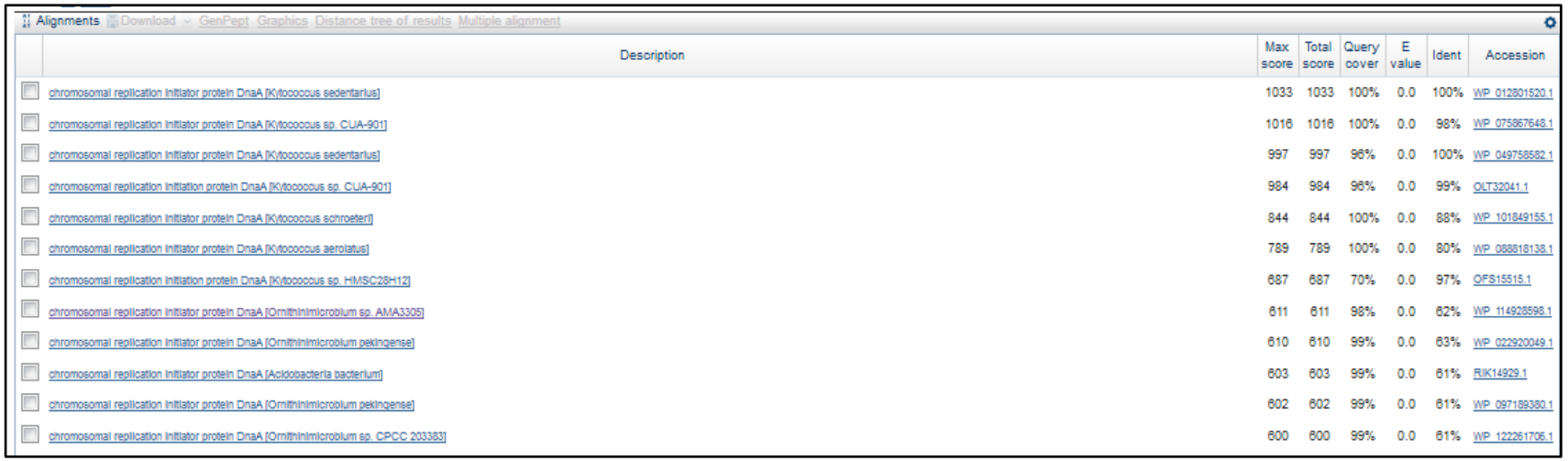
Swissprot database



Alignments | Download | GenPept | Graphics | Distance tree of results | Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
RecName: Full=Chromosomal replication initiator protein DnaA	517	517	97%	1e-179	56%	A6W3V4.1
RecName: Full=Chromosomal replication initiator protein DnaA	498	498	99%	1e-172	52%	A1T102.1
RecName: Full=Chromosomal replication initiator protein DnaA	495	495	99%	2e-171	51%	A0PKB2.1
RecName: Full=Chromosomal replication initiator protein DnaA	494	494	99%	6e-171	51%	A0R7K1.1
RecName: Full=Chromosomal replication initiator protein DnaA	493	493	99%	2e-170	50%	B2HI46.1
RecName: Full=Chromosomal replication initiator protein DnaA	490	490	99%	1e-169	51%	Q1BG61.1
RecName: Full=Chromosomal replication initiator protein DnaA	488	488	99%	1e-168	50%	B1MDH6.1
RecName: Full=Chromosomal replication initiator protein DnaA	485	485	99%	2e-167	52%	P49991.2
RecName: Full=Chromosomal replication initiator protein DnaA	485	485	99%	3e-167	52%	C1AIZ8.1
RecName: Full=Chromosomal replication initiator protein DnaA	484	484	99%	4e-167	52%	A5TY69.1
RecName: Full=Chromosomal replication initiator protein DnaA	484	484	98%	4e-167	53%	Q6ABL5.1

nr database



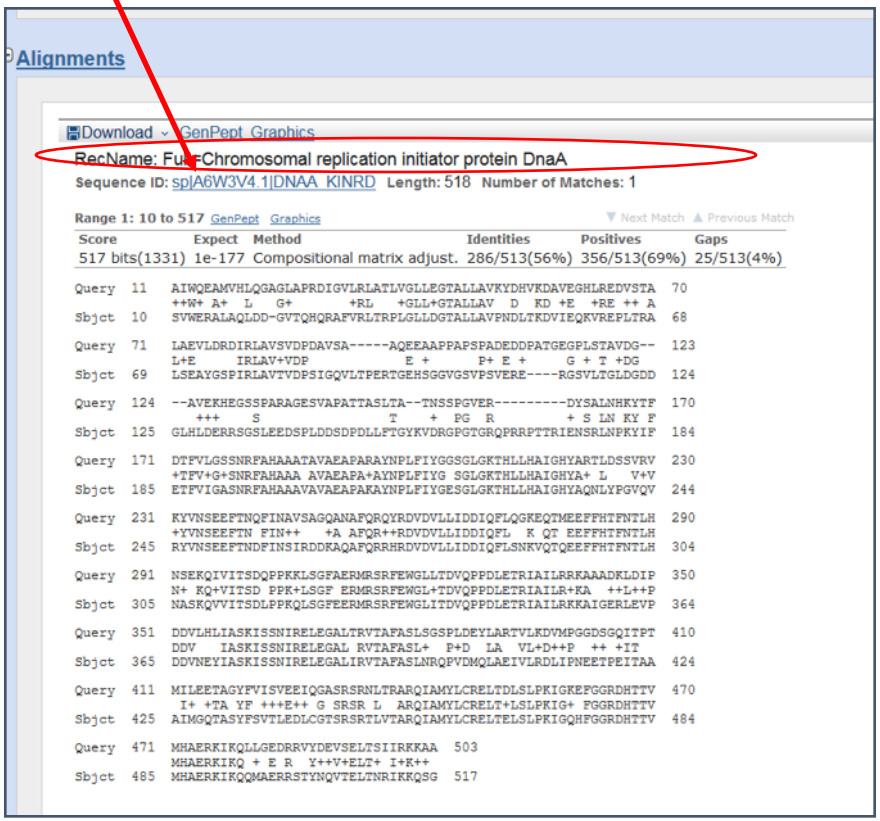
Alignments | Download | GenPept | Graphics | Distance tree of results | Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
chromosomal replication initiator protein DnaA [Kytococcus sedentarius]	1033	1033	100%	0.0	100%	WP_012801520.1
chromosomal replication initiator protein DnaA [Kytococcus sp. CUA-901]	1016	1016	100%	0.0	98%	WP_075867648.1
chromosomal replication initiator protein DnaA [Kytococcus sedentarius]	997	997	96%	0.0	100%	WP_049758582.1
chromosomal replication initiation protein DnaA [Kytococcus sp. CUA-901]	984	984	96%	0.0	99%	OLT32041.1
chromosomal replication initiator protein DnaA [Kytococcus schroeteri]	844	844	100%	0.0	88%	WP_101849155.1
chromosomal replication initiator protein DnaA [Kytococcus aerolatus]	789	789	100%	0.0	80%	WP_088818138.1
chromosomal replication initiation protein DnaA [Kytococcus sp. HMSC28H12]	687	687	70%	0.0	97%	OFS15515.1
chromosomal replication initiator protein DnaA [Ornithinimicrobium sp. AMA3305]	611	611	98%	0.0	62%	WP_114928598.1
chromosomal replication initiator protein DnaA [Ornithinimicrobium pekingense]	610	610	99%	0.0	63%	WP_022920049.1
chromosomal replication initiator protein DnaA [Acidobacteria bacterium]	603	603	99%	0.0	61%	RIK14929.1
chromosomal replication initiator protein DnaA [Ornithinimicrobium pekingense]	602	602	99%	0.0	61%	WP_097189380.1
chromosomal replication initiator protein DnaA [Ornithinimicrobium sp. CPCC 203383]	600	600	99%	0.0	61%	WP_122261706.1

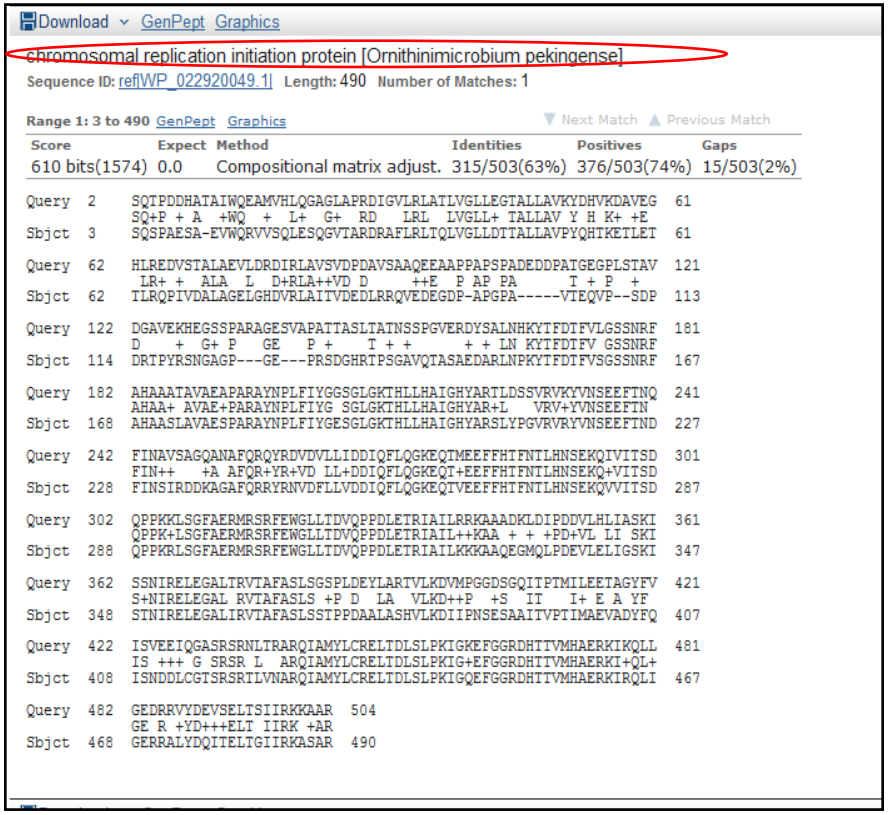
Results:
nr and Swissprot databases

Click on link

Swissprot database



nr database



Conclusion: The top alignment by BLAST by both nr and Swissprot database is the same
The top hit is- chromosomal Replication initiator protein DnaA.

What if you get no significant BLAST hit? SwissProt and NR database searches

If there are no BLAST hits with E-values lower than 1 x 10-3

1. Make that notation in notebook (do not leave empty)
2. Move onto the next module.

A finding of no significant BLAST hits would indicate that no sequence in the database has any homology to query protein

Could be due to:

a) Dealing with a newly discovered protein
b) The protein has been called in error and does not really exist.