

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

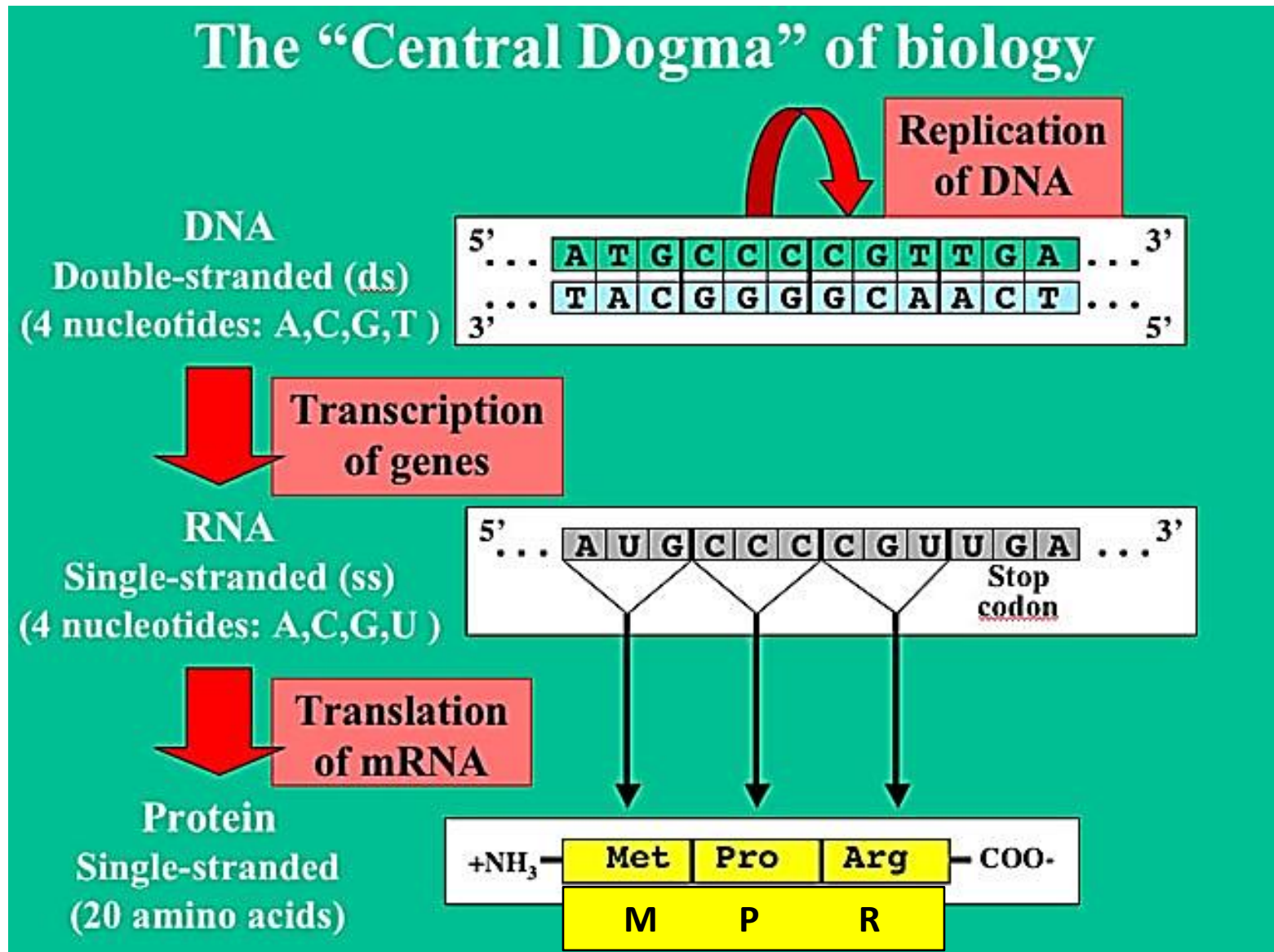
Tree based Consistency Objective Function For alignment Evaluation (T-Coffee) is a multiple sequence alignment program that aligns a set of homologous (similar) sequences

4. WebLogo

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

Relationship

DNA -----> RNA -----> PROTEIN



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.

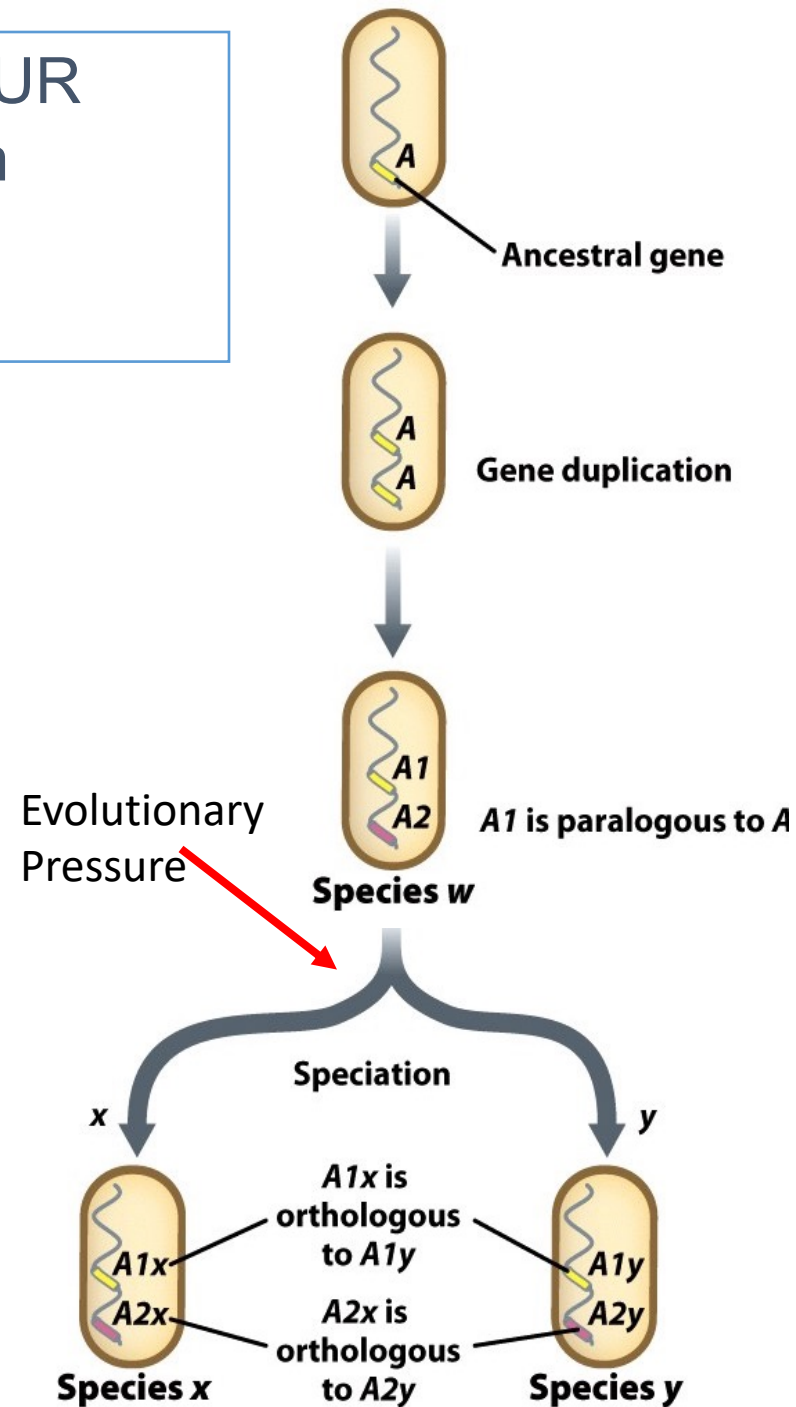
Comparing amino acid sequences is a more reliable to use for similarity between two sequences than comparing nucleotide sequences. Removes the redundancy issue.

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)**

The screenshot shows the NCBI BLAST website. At the top, there is a navigation bar with the NIH logo, "U.S. National Library of Medicine", "NCBI National Center for Biotechnology Information", and a "Sign in to NCBI" link. Below this is the "BLAST" logo and navigation links for "Home", "Recent Results", "Saved Strategies", and "Help".

The main content area is titled "Basic Local Alignment Search Tool". It includes a brief description: "BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance." A "Learn more" link is provided.

To the right of the description is a "NEWS" box titled "BLAST+ 2.4.0 released". It states: "A new version (2.4.0) of the BLAST+ executables is now available. Thu, 02 Jun 2016 14:00:00 EST" and includes a link for "More BLAST news...".

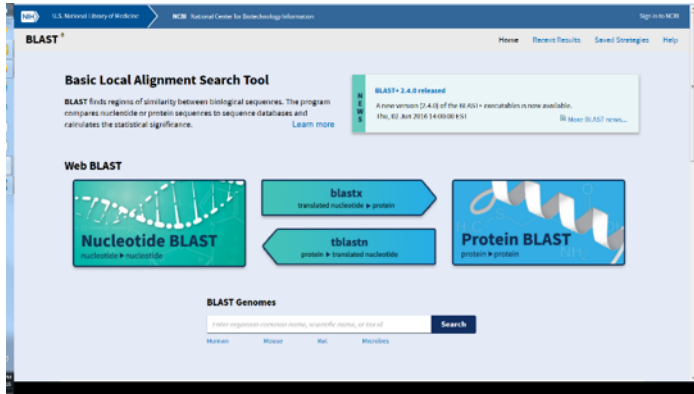
Below the main text is the "Web BLAST" section. It features four buttons for different search types:

- Nucleotide BLAST**: nucleotide ► nucleotide
- blastx**: translated nucleotide ► protein
- tblastn**: protein ► translated nucleotide
- Protein BLAST**: protein ► protein (This button is highlighted with a red border in the image)

At the bottom of the page is the "BLAST Genomes" section, which includes a search input field with the placeholder text "Enter organism common name, scientific name, or tax id" and a "Search" button. Below the input field are links for "Human", "Mouse", "Rat", and "Microbes".

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?

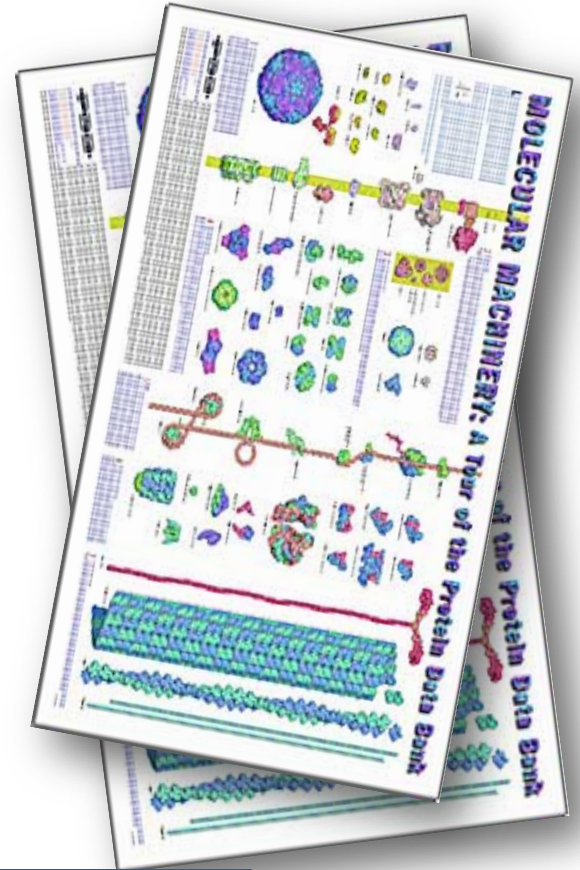


Protein
Databases
(**Subject**)

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



Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/> RecName: Full=Chromosomal replication initiator protein DnaA	517	517	97%	1e-177	56%	A6W3V4.1
<input type="checkbox"/> RecName: Full=Chromosomal replication initiator protein DnaA	498	498	99%	7e-171	52%	A1T102.1
<input type="checkbox"/> RecName: Full=Chromosomal replication initiator protein DnaA	495	495	99%	2e-169	51%	A0PKB2.1
<input type="checkbox"/> RecName: Full=Chromosomal replication initiator protein DnaA	494	494	99%	4e-169	51%	A0R7K1.1
<input type="checkbox"/> RecName: Full=Chromosomal replication initiator protein DnaA	493	493	99%	1e-168	50%	B2H46.1
<input type="checkbox"/> RecName: Full=Chromosomal replication initiator protein DnaA	490	490	99%	9e-168	51%	Q1BG61.1
<input type="checkbox"/> 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

N-terminal end or amino term end

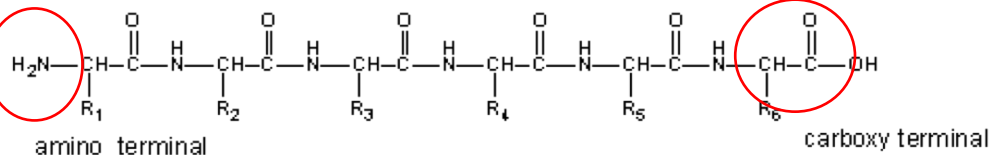
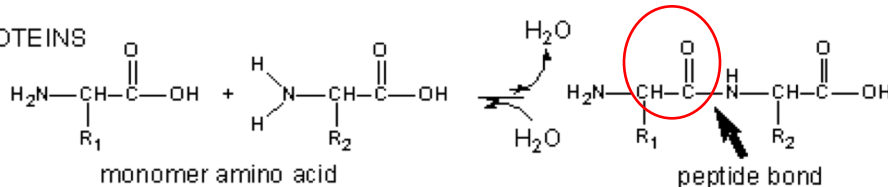
```
>Ksed_00010 amino acid
```

```
MSQTPDDHATAIWQEAMVHLQGAGLA  
PRDIGVLRRLATLVGLLLEGTALLAVKY  
DHVKDAVEGHLREDVSTALAEVLDLRD  
IRLAVSVDPDAVSAAQEEAAPPAPSP  
ADEDDPATGEGPLSTAVDGAVEKH
```

C- or carboxy terminal end

Sequence data-Protein
(amino acid sequence)

1. PROTEINS

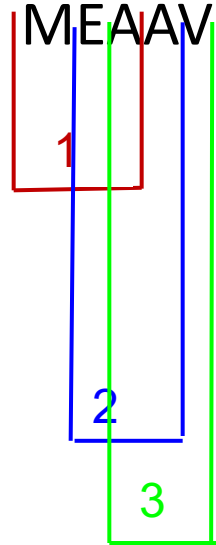


Proteins: MW 5000 - > 2,000,000
approx 50 to 20,000 amino acids

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



BLAST

each 3 “word”

individually against database

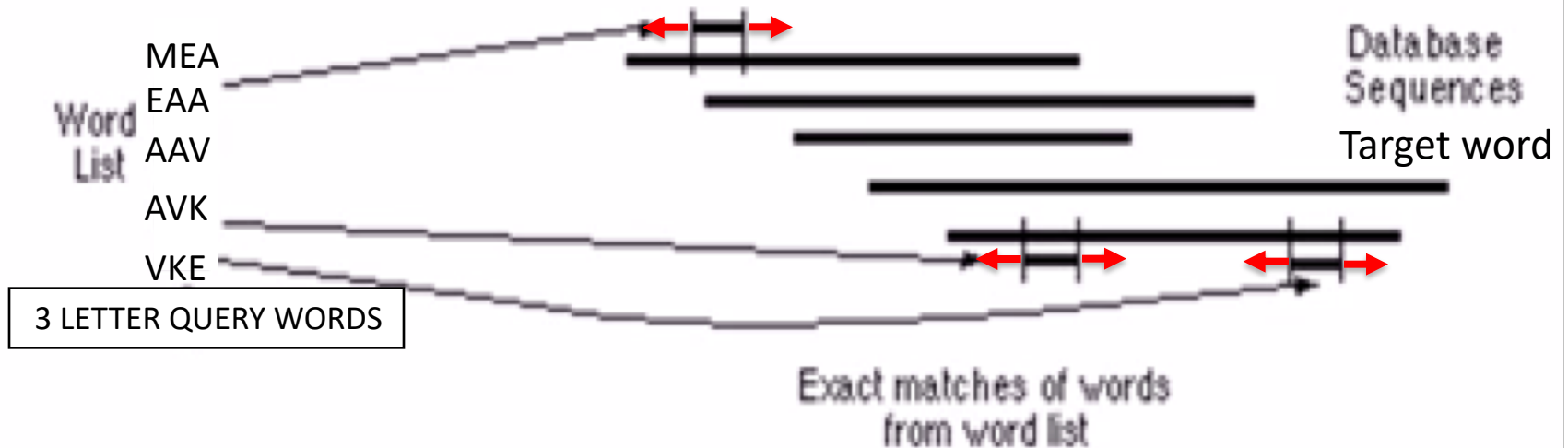
BREAK DATABASE SEQ INTO WORDS

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.

Compare the word list to the database and identify exact matches



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 v 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	SQTPDDHATAIWQEAMVHLQAGLAPRDIGVLRRLATLVGLLEG TALLAVKYD HVKDAVEG	61
Sbjct	3	SQ+P + A +WQ + L+ G+ RD LRL LVGLL+ TALLAV Y H K+ +E SQSPAESA-EVWQRVVSQLESQGVTIARDRAFLRLTQLVGLLDTTALLAVPYQHTKETLET	61
Query	62	HLREDVSTALAEVLDLDRDIRLAVSVDPDAVSAAQEEAAPPAPSPAEDDDPATGEGPLSTAV	121
Sbjct	62	LR+ + ALA L D+RLA++VD D ++E P AP PA T + P + TLRQPIVDALAGELGHDVRLAITVDEDLRRQVEDEGDP-APGPA-----VTEQVP--SDP	113
Query	122	DGAVEKHEGSSPARAGESVAPATTASLTAINSSPGVERDYSALNHKTYFDTFVLGSSNRF	181
Sbjct	114	D + G+ P GE P + T + + + + LN KYTFDTFV GSSNRF DRTPYRSNGAGP---GE---PRSDGHRTPSGAVQTASAEDARLNPKYTFDTFVSGSSNRF	167
Query	182	AHAATAVAEAPARAYNPLFIYGGSGLGKTHLLHAIGHYARTLDDSSVRVKYVNSEEF TNQ	241
Sbjct	168	AHAA+ AVAE+PARAYNPLFIYG SGLGKTHLLHAIGHYAR+L VRV+YVNSEEF TN AHAASLAVAESPARAYNPLFIYGESGLGKTHLLHAIGHYARS LYPGVRVRYVNSEEF TND	227
Query	242	FINAVSAGQANAFQRQYRDVDVLLIDDIQFLQGKEQIMEEFFHTFNILHNSEKQIVITSD	301
Sbjct	228	FIN++ +A AFQR+YR+VD LL+DDIQFLQGKEQ+EEFFHTFNILHNSEKQ+VITSD FINSIRDDKAGAFQRRYRNVDFLLVDDIQFLQGKEQTVEEFFHTFNILHNSEKQVITSD	287
Query	302	QPPKKLSGFAERMRSRFEWGLLTDVQPPDLETRIAILRRKAAADKLDIPDDVLHLIASKI	361
Sbjct	288	QPPK+LSGFAERMRSRFEWGLLTDVQPPDLETRIAIL++KAA + + +PD+VL LI SKI QPPKRLSGFAERMRSRFEWGLLTDVQPPDLETRIAILKKKAAQEGMQLPDEVLELIGSKI	347
Query	362	SSNIRELEGALRVTAFAASLSGSPLEYLARTVLKDVMPGGDSGQITPTMILEETAGYFV	421
Sbjct	348	S+NIRELEGAL RVTAFAASL+P D LA VLKD++P +S IT I+ E A YF STNIRELEGALRVTAFAASLSSTPPDAALASHVLKDIIIPNSESAAITVPTIMAEVADYFQ	407
Query	422	ISVEEIQGASRSRNLTARQIAMYLCRELTDLSLPKIGKEFGGRDHTVMHAERKIKQLL	481
Sbjct	408	IS+++ G SRSL L ARQIAMYLCRELTDLSLPKIG+EFGRDHTVMHAERKI+QL+ ISNDDLCGTSRSRTLVNARQIAMYLCRELTDLSLPKIGQEFGRDHTVMHAERKIRQLI	467
Query	482	GEDRRVYDEVSELTSIIRKKAAR 504	
Sbjct	468	GE R +YD+++ELT IIRK +AR 490 GERRALYDQITELTGIIRKASAR	

Query : Your gene/protein –aa seq

Subject: 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.

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
- $E = 1 \times 10^{-3}$ 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?**

The screenshot shows the NCBI BLAST Standard Protein BLAST interface. The 'Enter Query Sequence' section contains a FASTA sequence for 'Ksed_00010-aa sequence'. The 'Choose Search Set' section has a dropdown menu open, showing a list of databases. A yellow callout box with the number '2' points to the dropdown menu. A larger yellow callout box with the text 'Set up both on 2 tabs' points to the 'Non-redundant protein sequences (nr)' option in the dropdown. Two red arrows point from the 'Non-redundant protein sequences (nr)' option to yellow boxes labeled 'SET UP Tab-1' and 'SET UP TAB-2'. The 'Program Selection' section shows 'blastp (protein-protein BLAST)' selected.

U.S. National Library of Medicine | NCBI National Center for Biotechnology Information | Sign in to NCBI

BLAST® » blastp suite | Home | Recent Results | Saved Strategies | Help

Standard Protein BLAST

blastn | **blastp** | blastx | tblastn | tblastx

BLASTP programs search protein databases using a protein query. [more...](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Query subrange

>Ksed_00010-aa sequence
MSQTFFDQHATAIWOEAMVHLOGAGLAPRDICVLRATI VCLLEG TALLAVKYDHWKDAVEGHLR
EDVSTALAEVLDRLIRLAVSVDDPAVSAAQEEAAPPAPSPADEDDPAIGEGPLSTAVDGAVEKH
EGSSPARAGESVAPATTASLTATNSSPGVERDYSALNHKYTFDI FVLGSSNRFAHAAATAVAEA
PARAYNPLFIYGGSLGKTHLLHAIGHYARTLDSVSRVKYVNSEFTNQFINAVSAGQANAFOR
QYRDVVDLLDDIQFLOGKEQTMEEFFHIFNTLHNSEKQIVITSDQPPKLSGFAERMRSRFEW

Or, upload file No file selected.

Job Title: Ksed_00010-aa sequence
Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database: **Non-redundant protein sequences (nr)**

Organism: Optional

Exclude: Optional

Entrez Query: Optional

Program Selection

Algorithm: blastp (protein-protein BLAST)
 PSI-BLAST (Position-Specific Iterated BLAST)
 PHI-BLAST (Pattern Hit Initiated BLAST)
 DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

Annotations:

- Yellow callout '2' points to the 'Choose Search Set' dropdown menu.
- Yellow callout 'Set up both on 2 tabs' points to the 'Non-redundant protein sequences (nr)' option.
- Red arrows point from 'Non-redundant protein sequences (nr)' to yellow boxes labeled 'SET UP Tab-1' and 'SET UP TAB-2'.

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

Alignment length

Score

E-Value

Alignment of the top hit and the query sequence

Gene product name (*second hit*)

Organism

Alignment length

BLAST RESULTS - Swissprot database

Top of the results page

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

Query ID [Id|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).

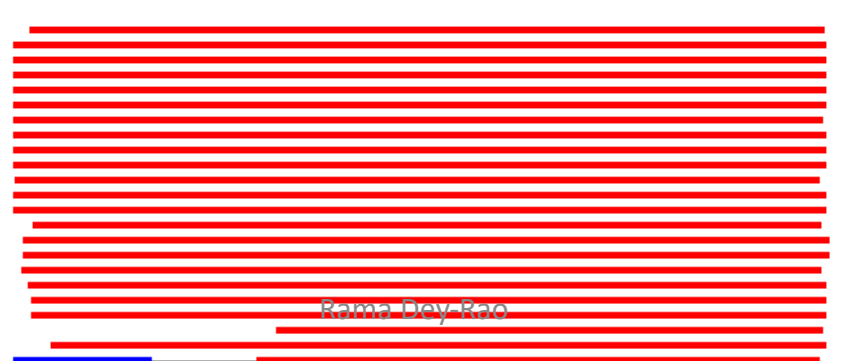
Distribution of the top 106 Blast Hits on 100 subject sequences

Mouse over to see the title, click to show alignments

Color key for alignment scores

■ <40 ■ 40-50 ■ 50-80 ■ 80-200 ■ >=200

1 100 200 300 400 500



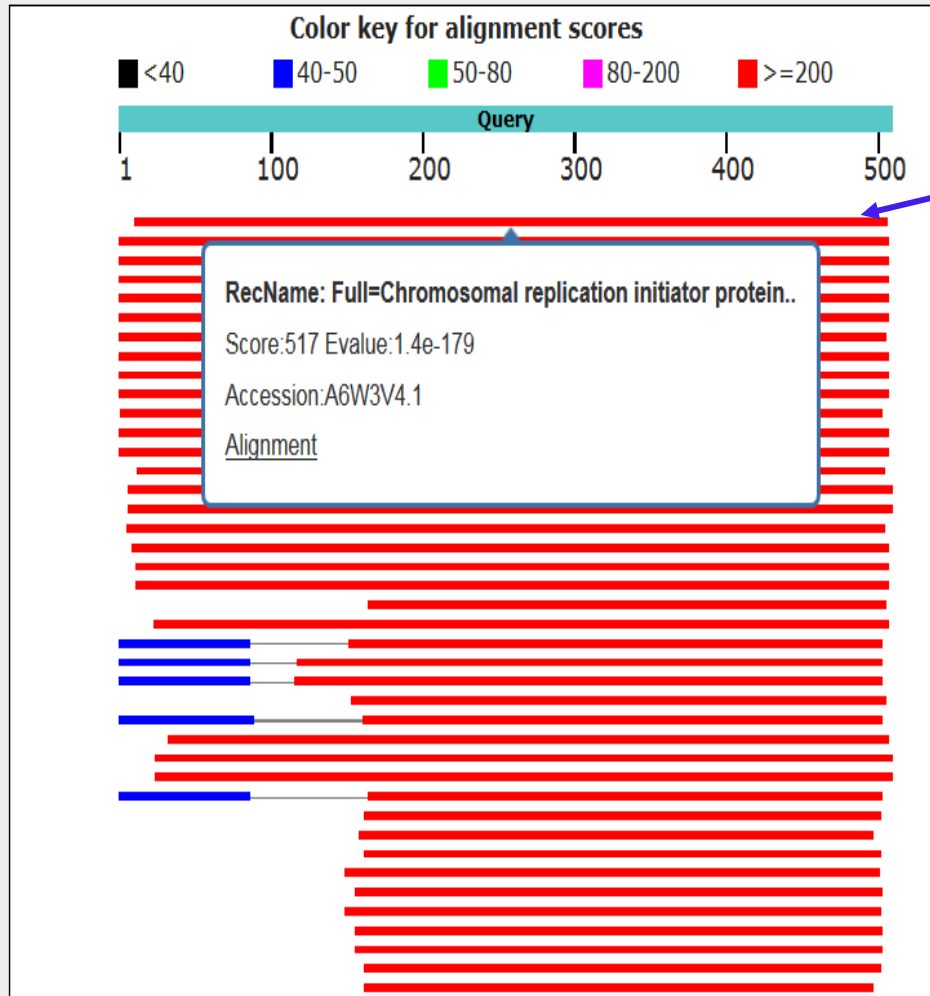
Quick scan: Visual rep of
coverage in orthologous
proteins

BLAST RESULTS - Swissprot database

Top of the results page

Distribution of the top 106 Blast Hits on 100 subject sequences

Mouse over to see the title, click to show alignments



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
<input type="checkbox"/> RecName: Full=Chromosomal replication initiator protein DnaA	517	517	97%	1e-179	56%	A6W3V4.1
<input type="checkbox"/> RecName: Full=Chromosomal replication initiator protein DnaA	498	498	99%	1e-172	52%	A1T102.1
<input type="checkbox"/> RecName: Full=Chromosomal replication initiator protein DnaA	495	495	99%	2e-171	51%	A0PKB2.1
<input type="checkbox"/> RecName: Full=Chromosomal replication initiator protein DnaA	494	494	99%	6e-171	51%	A0R7K1.1
<input type="checkbox"/> RecName: Full=Chromosomal replication initiator protein DnaA	493	493	99%	2e-170	50%	B2HI46.1
<input type="checkbox"/> RecName: Full=Chromosomal replication initiator protein DnaA	490	490	99%	1e-169	51%	Q1BG61.1
<input type="checkbox"/> RecName: Full=Chromosomal replication initiator protein DnaA	488	488	99%	1e-168	50%	B1MDH6.1
<input type="checkbox"/> RecName: Full=Chromosomal replication initiator protein DnaA	485	485	99%	2e-167	52%	P49991.2
<input type="checkbox"/> RecName: Full=Chromosomal replication initiator protein DnaA	485	485	99%	3e-167	52%	C1AI28.1
<input type="checkbox"/> 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

NCBI Resources How To Sign in to NCBI

Protein Protein Search Advanced Help

GenPept Send to: Change region shown Customize view

RecName: Full=Chromosomal replication initiator protein DnaA

UniProtKB/Swiss-Prot: A6W3V4.1

[Identical Proteins](#) [FASTA](#) [Graphics](#)

Go to: [v]

LOCUS DNAA_KINRD 518 aa linear BCT 02-NOV-2016

DEFINITION RecName: Full=Chromosomal replication initiator protein DnaA.

ACCESSION A6W3V4

VERSION A6W3V4.1

DBSOURCE UniProtKB: locus DNAA_KINRD, accession [A6W3V4](#);
class: standard.
created: May 20, 2008.
sequence updated: Aug 21, 2007.
annotation updated: Nov 2, 2016.
xrefs: [CP000750.2](#), [ABS01493.1](#), [WP_012085692.1](#)
xrefs (non-sequence databases): ProteinModelPortal:A6W3V4,
STRING:266940.Krad_0001, EnsemblBacteria:ABS01493,
EnsemblBacteria:ABS01493, EnsemblBacteria:Krad_0001,
KEGG:kra:Krad_0001, eggNOG:ENOG4105CI4, eggNOG:COG0593,
HOGENOM:HOG000235658, KO:K02313, OMA:ASVHESW, OrthoDB:POG091H02FF,
Proteomes:UP000001116, GO:0005737, GO:0005524, GO:0003688,
GO:0006270, GO:0006275, CDD:cd06571, Gene3D:1.10.1750.10,
Gene3D:3.40.50.300, HAMAP:MF_00377, InterPro:IPR003593,
InterPro:IPR001957, InterPro:IPR020591, InterPro:IPR018312,
InterPro:IPR013317, InterPro:IPR013159, InterPro:IPR027417,
InterPro:IPR010921, Pfam:PF00308, Pfam:PF08299, PRINTS:PR00051,
SMART:SM00382, SMART:SM00760, SUPFAM:SSF48295, SUPFAM:SSF52540,
TIGRFAMs:TIGR00362, PROSITE:PS01008

KEYWORDS ATP-binding; Complete proteome; Cytoplasm; DNA replication;
DNA-binding; Nucleotide-binding; Reference proteome

SOURCE Kineococcus radiotolerans SRS30216 = ATCC BAA-149

ORGANISM [Kineococcus radiotolerans SRS30216 = ATCC BAA-149](#)
Bacteria; Actinobacteria; Kineosporiales; Kineosporiaceae;
Kineococcus.

Analyze this sequence [v]

Run BLAST

Identify Conserved Domains

Highlight Sequence Features

Find in this Sequence

Related information [v]

BLink

Related Sequences

CDD Search Results

Conserved Domains (Concise)

Conserved Domains (Full)

Domain Relatives

Proteins with Similar Sequences

Related Structures (List)

Related Structures (Summary)

Taxonomy

Recent activity [v]

Turn Off Clear

BLAST Alignment Details

RecName: Full=Chromosomal replication initiator protein DnaA

Sequence ID: [A6W3V4.1](#) Length: 518 Number of Matches: 1

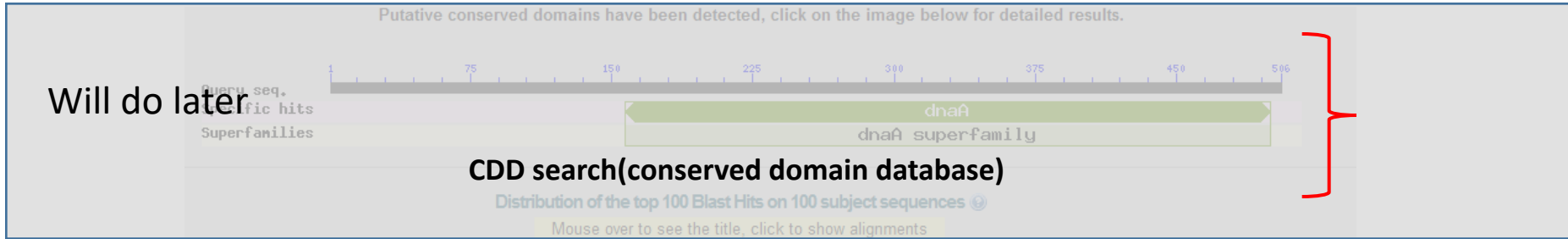
Range 1: 10 to 517 [GenPept](#) [Graphics](#)

▼ Next Match ▲ Previous Match

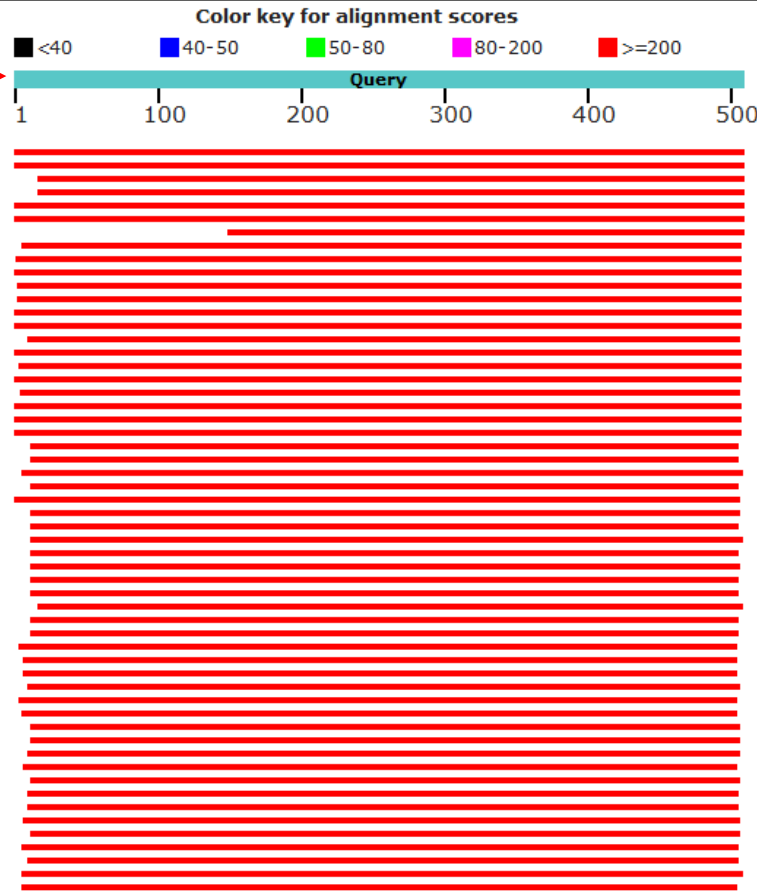
Score	Expect	Method	Identities	Positives	Gaps
517 bits(1331)	1e-179	Compositional matrix adjust.	286/513(56%)	356/513(69%)	25/513(4%)
Query 11	AIWQEAMVHLQGAGLAPRDIGVLRLATLVGLLEG TALLAVKYDHSV KDAVEGHLREDVSTA				70
	++W+ A+ L G+ +RL +GLL+GTALLAV D KD +E +RE ++ A				
Sbjct 10	SVWERALAQ LDD-GVTQHQR AFVRLTRPLGLLDGTALLAVPNDLTKDVIEQKVREPLTRA				68
Query 71	LAEVLDRDIRLAVSVDPDAVSA-----AQEEAAPAPSPADEDDPATGEGPLSTAVDG--				123
	L+E IRLAV+VDP E + P+ E + G + T +DG				
Sbjct 69	LSEAYGSPIRLAVIVDPSIGQVLTPERTGEGSGGVSVPSVERE-----RGSVLTGLDGDD				124
Query 124	--AVEKHEGSSPARAGESVAPATTASLTA--INSSPGVER-----DYSALNHKYTF				170
	+++ S T + PG R + S LN KY F				
Sbjct 125	GLHLDERRSGLSEEDSPLDDSDPDL LFTGYKVD RGP GTGRQPRRP TTRIENSRLNPKYIF				184
Query 171	DTFVLGSSNRF AHAATAVAEAPARAYNPLFIYGG SGLGKTHLLHAIGHYARTLDSSVRV				230
	+TFV+G+SNRF AHAATAVAEAPARAYNPLFIYGG SGLGKTHLLHAIGHYA+ L V+V				
Sbjct 185	ETFVIGASNRF AHAATAVAEAPAKAYNPLFIYGESGLGKTHLLHAIGHYAQNLYPGVQV				244
Query 231	KYVNSEEF T NQFINAVSAGQANAFQ RQYRDVDVLLIDDIQFLQGKEQTMEEFFHTFNTLH				290
	+YVNSEEF T N FIN++ +A AFQR++RDVDVLLIDDIQFL K QT EEFFHTFNTLH				
Sbjct 245	RYVNSEEF T NDFINSIRDDKAQAFQRRHRDVDVLLIDDIQFLSNK VQTQE EEFFHTFNTLH				304
Query 291	NSEKQIVITSDQPPK KLSGFAERMRSRFEWGLLTDVQPPDLETRIAILRRKAAADKLDIP				350
	N+ KQ+VITSD PPK+LSGF ERMRSRFEWGL+TDVQPPDLETRIAILR+KA ++L++P				
Sbjct 305	NASKQVVITSDLPKQLSGFEERMRSRFEWGLITDVQPPDLETRIAILRKKAI GERLEVP				364
Query 351	DDVLHLIASKISSNIRELEGALTRVTA FASLSGSP LDEYLARTIVLKDVMPPGGDSGQITPT				410
	DDV IASKISSNIRELEGAL RVTA FASL+ P+D LA VL+D++P ++ +IT				
Sbjct 365	DDVNEYIASKISSNIRELEGALIRVTA FASLNRQPVD MQLAEIVLRDLIPNEETPEITAA				424
Query 411	MILEETAGYFVISVEEIQGASRSRNLTRARQIAMYLCRELTDLSLPKIGKEFGGRDHTTV				470
	I+ +IA YF +++E++ G SRSR L ARQIAMYLCRELT+LSLPKIG+ FGGRDHTTV				
Sbjct 425	AIMGQTASYF SVILEDL CGTSRSRTLVTARQIAMYLCRELTEL SLPKIGQHFGGRDHTTV				484
Query 471	MHAERKIKQLLGEDRRVYDEVSELT SIIRKKA 503				
	MHAERKIKQ + E R Y++V+ELT+ I+K++				
Sbjct 485	MHAERKIKQMAERRSTYNQVTELTNRIKKQSG 517				

BLAST RESULTS PAGE- nr database

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



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 organisms.

**Quick scan
Visual rep
of coverage**

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

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
<input type="checkbox"/> chromosomal replication initiator protein DnaA [Kytococcus sedentarius]	1033	1033	100%	0.0	100%	WP_012801520.1
<input type="checkbox"/> chromosomal replication initiator protein DnaA [Kytococcus sp. CUA-901]	1016	1016	100%	0.0	98%	WP_075867648.1
<input type="checkbox"/> chromosomal replication initiator protein DnaA [Kytococcus sedentarius]	997	997	96%	0.0	100%	WP_049758582.1
<input type="checkbox"/> chromosomal replication initiation protein DnaA [Kytococcus sp. CUA-901]	984	984	96%	0.0	99%	OLT32041.1
<input type="checkbox"/> chromosomal replication initiator protein DnaA [Kytococcus schroeteri]	844	844	100%	0.0	88%	WP_101849155.1
<input type="checkbox"/> chromosomal replication initiator protein DnaA [Kytococcus aerolatus]	789	789	100%	0.0	80%	WP_088818138.1
<input type="checkbox"/> chromosomal replication initiation protein DnaA [Kytococcus sp. HMSC28H12]	687	687	70%	0.0	97%	OFS15515.1
<input type="checkbox"/> chromosomal replication initiator protein DnaA [Ornithinimicrobium sp. AMA3305]	611	611	98%	0.0	62%	WP_114928598.1
<input type="checkbox"/> chromosomal replication initiator protein DnaA [Ornithinimicrobium pekinense]	610	610	99%	0.0	63%	WP_022920049.1
<input type="checkbox"/> chromosomal replication initiator protein DnaA [Acidobacteria bacterium]	603	603	99%	0.0	61%	RIK14929.1
<input type="checkbox"/> chromosomal replication initiator protein DnaA [Ornithinimicrobium pekinense]	602	602	99%	0.0	61%	WP_097189380.1
<input type="checkbox"/> chromosomal replication initiator protein DnaA [Ornithinimicrobium sp. CPCC 203383]	600	600	99%	0.0	61%	WP_122261706.1
<input type="checkbox"/> chromosomal replication initiator protein DnaA [Ornithinimicrobium sp. KCTC 49018]	592	592	99%	0.0	61%	WP_109472715.1
<input type="checkbox"/> chromosomal replication initiator protein DnaA [Serinicoccus profundii]	590	590	99%	0.0	60%	WP_010147278.1
<input type="checkbox"/> chromosomal replication initiator protein DnaA [Ornithinimicrobium sp. CNJ-824]	589	589	97%	0.0	59%	WP_075959275.1
<input type="checkbox"/> chromosomal replication initiator protein DnaA [Arsenicococcus sp.]	583	583	99%	0.0	60%	PZU43890.1
<input type="checkbox"/> chromosomal replication initiator protein DnaA [Serinicoccus chungangensis]	580	580	99%	0.0	60%	WP_058892007.1

DO NOT ADD THIS

Click on first linked choice That is not from your organism

SCROLL BELOW

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	SQTPDDNATAIWQEAMVHLQAGLAPRDIGVLRRLATLVGLLEGTTALLAVKYDHSVKADEG				61
Sbjct 3	SQSPAESA-IVWQQRVVSQLESQGVTDRAFLRLTQLVGLLDTTALLAVPYQHTKETLET				61
Query 62	HLREDVSTALAEVLLDRDIRLAVSVDPDAVSAAQEEAAPPAPSPAEDDDPATGEGPLSTAV				121
Sbjct 62	TLRQPIVDALAGELGHDVRLAITVDEDLRRQVEDEGDF-APGPA-----VTEQVP--SDP				113
Query 122	DGAVEKHEGSSPARAGESVAPATTASLTATNSSPGVERDYSALNHKYTFDFTVLGSSNRF				181
Sbjct 114	DRTPYRSNGAGP---GE---PRSDGHRTPSGAVQTASAEDARLNPKYTFDFTVSGSSNRF				167
Query 182	AHAAATAVAEAPARAYNPLFIYCGSSGLGKTHLLHAIGHYARTLDSSVRVKYVNSEEFTNQ				241
Sbjct 168	AHAASLAVAESPAPARAYNPLFIYCGESGLGKTHLLHAIGHYARSPLYGVRVRYVNSEEFTND				227
Query 242	FINAVSAGQANAFQRQYRDVDVLLIDDIQFLQKQEQTMEEFFHTFNTLHNSKQIVITSD				301
Sbjct 228	FINSIRDKAGAFQRRYRNVDVFLVDDIQFLQKQEQTMEEFFHTFNTLHNSKQVVTSD				287
Query 302	QPPKLSGFAERMRSRFEWGLLTDVQPPDLETRIAILRRKAAADKLDIPDDVLHLIASKI				361
Sbjct 288	QPPKLSGFAERMRSRFEWGLLTDVQPPDLETRIAILRRKAAQEGMQLPDEVLELIGSKI				347
Query 362	SSNIRELEGALTRVTAFAFASLSGSPLEDEYLARTVLKDVMPGDSGQITPTMILEETAGYFV				421
Sbjct 348	STNIRELEGALIRVTAFAFASLSSTPPDAALASHVLKDIIPNSESAAITVPTIMAEVADYFQ				407
Query 422	ISVEEIQGASRSRNLTRARQIAMYLCRELTDSLPLKIGKEFGGRDHTVMHAERKIKQLL				481
Sbjct 408	ISNDDLCGTSRSRTLNRQIAMYLCRELTDSLPLKIGKEFGGRDHTVMHAERKI+QL+				
Query 482	GEDRRVYDEVSELTSIIRKKAAR	504			
Sbjct 468	GERRALYDQITELTGIIRKASAR	490			

Copy

Alignment Length
(Last Query #) - (1st 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 in 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 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 v 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	SQTPDDHATAIWQEAMVHLQAGLAPRDIGVLRRLATLVGLLEG TALLAVKYD HVKDAVEG	61
Sbjct	3	SQ+P + A +WQ + L+ G+ RD LRL LVGLL+ TALLAV Y H K+ +E SQSPAESA-EVWQRVVSQLESQGVTIARDRAFLRLTQLVGLLDTTALLAVPYQHTKETLET	61
Query	62	HLREDVSTALAEVLDLDRDIRLAVSVDPDAVSAAQEEAAPPAPSPAEDDDPATGEGPLSTAV	121
Sbjct	62	LR+ + ALA L D+RLA++VD D ++E P AP PA T + P + TLRQPIVDALAGELGHVRLAITVDEDLRRQVEDEGDP-APGPA-----VTEQVP--SDP	113
Query	122	DGAVEKEHEGSSPARAGESVAPATTASLTAINSSPGVERDYSALNHKTYFDTFVLGSSNRF	181
Sbjct	114	D + G+ P GE P + T + + + + LN KYTFDTFV GSSNRF DRTPYRSNGAGP---GE---PRSDGHRTPSGAVQTASAEDARLNPKYTFDTFVSGSSNRF	167
Query	182	AHAATAVAEAPARAYNPLFIYGGSGLGKTHLLHAIGHYARTLDSSVRVKYVNSEEFTNQ	241
Sbjct	168	AHAA+ AVAE+PARAYNPLFIYG SGLGKTHLLHAIGHYAR+L VRV+YVNSEEFNT AHAASLAVAESPARAYNPLFIYGESGLGKTHLLHAIGHYARSLYPGVRVRYVNSEEFTND	227
Query	242	FINAVSAGQANAFQQRQYRDVDVLLIDDIQFLQGKEQIMEEFFHTFNTILHNSEKQIVITSD	301
Sbjct	228	FIN++ +A AFQR+YR+VD LL+DDIQFLQGKEQ+EEFFHTFNTILHNSEKQ+VITSD FINSIRDDKAGAFQRRYRNVDFLLVDDIQFLQGKEQTVEEFFHTFNTILHNSEKQVITSD	287
Query	302	QPPKLSGFAERMRSRFEWGLLTDVQPPDLETRIAILRRKAAADKLDIPDDVLHLIASKI	361
Sbjct	288	QPPK+LSGFAERMRSRFEWGLLTDVQPPDLETRIAIL++KAA + + +PD+VL LI SKI QPPKRLSGFAERMRSRFEWGLLTDVQPPDLETRIAILKKKAAQEGMQLPDEVLELIGSKI	347
Query	362	SSNIRELEGALRVTAFAASLSGSPLEYLARTVLKDVMPGGDSGQITPTMILEETAGYFV	421
Sbjct	348	S+NIRELEGAL RVTAFAASL+P D LA VLKD++P +S IT I+ E A YF STNIRELEGALRVTAFAASLSSTPPDAALASHVLKDIIIPNSESAAITVPTIMAEVADYFQ	407
Query	422	ISVEEIQGASRSRNLTRARQIAMYLCRELTDLSLPKIGKEFGGRDHTVMHAERKIKQLL	481
Sbjct	408	IS+++ G SRSR L ARQIAMYLCRELTDLSLPKIG+EFGRDHTVMHAERKI+QL+ ISNDDLCGTSRSRTLVNARQIAMYLCRELTDLSLPKIGQEFGRDHTVMHAERKIRQLI	467
Query	482	GEDRRVYDEVSELTISIIRKKAAR 504	
Sbjct	468	GE R +YD+++ELT IIRK +AR GERRALYDQITELTGIIRKASAR 490	

Query : Your gene/protein –aa seq

Subject: 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|VVP_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	SQTPDDHATAIWQELMVHLQAGLAPRDIGVLRRLATLVGLLEGTTALLAVKYDVKDAVEG		61		
Sbjct 3	SQSPAESAEVWQVRVVSQLESQVGTARDRAFLRLTQLVGLLDTTALLAVPYQHTKETLET		61		
Query 62	HLREDVSTALAEVLDRLAVVNDPDAVSAQEEAAPPAPSPADEDDPATGEGPLSTAV		121		
Sbjct 62	TLRQPIVDALAGELGHDVRLAIVDELRRQVEDEGDP-APGPA-----VTEQVP--SDP		113		
Query 122	DGAVEKHEGSSPARAGESVAPATTASLTATNSPGVERDYSALNHKYYTDFDFVLGSSNRF		181		
Sbjct 114	DRTFYRSNGAGP---GE---PRSDGHRTPSGAVQTSAAEDARLNPKYTFDTFVSGSSNRF		167		
Query 182	AHAAATAVAEAPARAYNPLFIYGGSGLGKTHLLHAIGHYARSLDSSVVRVYVNSEEFNQ		241		
Sbjct 168	AHAA+AVAE+PARAYNPLFIYG SGLGKTHLLHAIGHYAR+VRV+YVNSEEFN		227		
Query 242	FINAVSAGQANAFQQRQYRDVVDLLIDDIQFLQKQEQTMEFFHTFNTLNSEKQIVITSD		301		
Sbjct 228	FINSIRDDKAGAFQRRYRNVDVFLVDDIQFLQKQEQTVEEFFHTFNTLNSEKQIVITSD		287		
Query 302	QPPKLSGFAERMRSRFEWGLLTDVQPPDLETRIAILRRKAAADKLDIPDDVHLIASKI		361		
Sbjct 288	QPPKLSGFAERMRSRFEWGLLTDVQPPDLETRIAILLKAA+ + +PD+VL LI SAI		347		
Query 362	SSNIRELEGALTRVIAFASLSGSPLEYLARTVVLKDVMPGGDSEQIITPMILEETAGYFV		421		
Sbjct 348	STNIRELEGALIRVIAFASLSSTPPDAALASHVLKDIIPNSESAAITVPTIIMAEVADYFQ		407		
Query 422	ISVEEIQGASRSRNLTRARQIAMYLCRELTDLSLPKIGKEFGGRDHTTMHAERKIKQLL		481		
Sbjct 408	IS+++G SRSR L ARQIAMYLCRELTDLSLPKIG+EFGGRDHTTMHAERKI+QL+		467		
Query 482	GEDRRVYDEVSELTISIIRKKAAR 504				
Sbjct 468	GERRALYDQITELTGIIRKASAR 490				

Organism

Alignment Length

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

E-value

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	SQTPDDHATAIWQEMVHLQAGLAPRDIGVLRLATLVGLLEG TALLAVKYDVKDAVEG				61
Sbjct 3	SQSPAESAEVWQVRVVSQLESQGV TARDRAFLRLTQLVGLLDTTALLAVPYQHTKETLET				61
Query 62	HLREDVSTALAEVLDRDIRLAVSVDPDAVSAAQEEAAPPAPSPAEDDDPATGEGPLSTAV				121
Sbjct 62	TLRQPIVDALAGELGHDVRLAITVDEDLRRQVEDEGDP-APGPA-----VTEQVP--SDP				113
Query 122	DGAVEKHEGSSPARAGESVAPATTASLTATNSSPGVERDYSALNHKYYFDITFVLGSSNRF				181
Sbjct 114	DRTPYRSNGAGP---GE---PRSDGHRTPSGAVQTASAEDARLNPKYTFDITFVSGSSNRF				167
Query 182	AHAAATAVAEAPARAYNPLFIYGGSGLGKTHLLHAIGHYARTLDSSVRVKYVNSEEF T NQ				241
Sbjct 168	AHAA+ AVAE+PARAYNPLFIY G SGLGKTHLLHAIGHYAR+L VRV+YVNSEEF T N				227
Query 242	FINAVSAGQANAFQQRVDDVLLIDDIQFLQGKEQTMEEFFHTFNTLHNSEKQIVITSD				301
Sbjct 228	FINSIRDDKAGAFQRRYRNVDFLLVDDIQFLQGKEQTVEEFFHTFNTLHNSEKQVVITSD				287
Query 302	QPPKLSGFAERMRSRFEWGLLTDVQPPDLETRIAILLRKAADKLDIPDDVLHLIASKI				361
Sbjct 288	QPPK+LSGFAERMRSRFEWGLLTDVQPPDLETRIAILL++KAA++ +PD+VL LI SKI				347
Query 362	SSNIRELEGALRVTAFAFASLSGSPLEYLARIVLKDVMPPGGDSGQITPTMILEETAGYFV				421
Sbjct 348	STNIRELEGALRVTAFAFASLSSTPPDAALASHVLDIIPNSESAAITVPTIMAEVADYFQ				407
Query 422	ISVEEIQGASRSRNLTRARQIAMYLCRELTDLSPKIGKEFGGRDHTVMHAERKIKQLL				481
Sbjct 408	ISNDLCGTSRSTLVNARQIAMYLCRELTDLSPKIGKEFGGRDHTVMHAERKIRQLI				467
Query 482	GEDRRVYDEVSELTSIIRKKAAR 504				
Sbjct 468	GE R +YD+++ELT IIRK +AR GERRALYDQITELTGIIRKASAR 490				

Score :

Numerical representation of quality of alignment

How is it calculated ? :

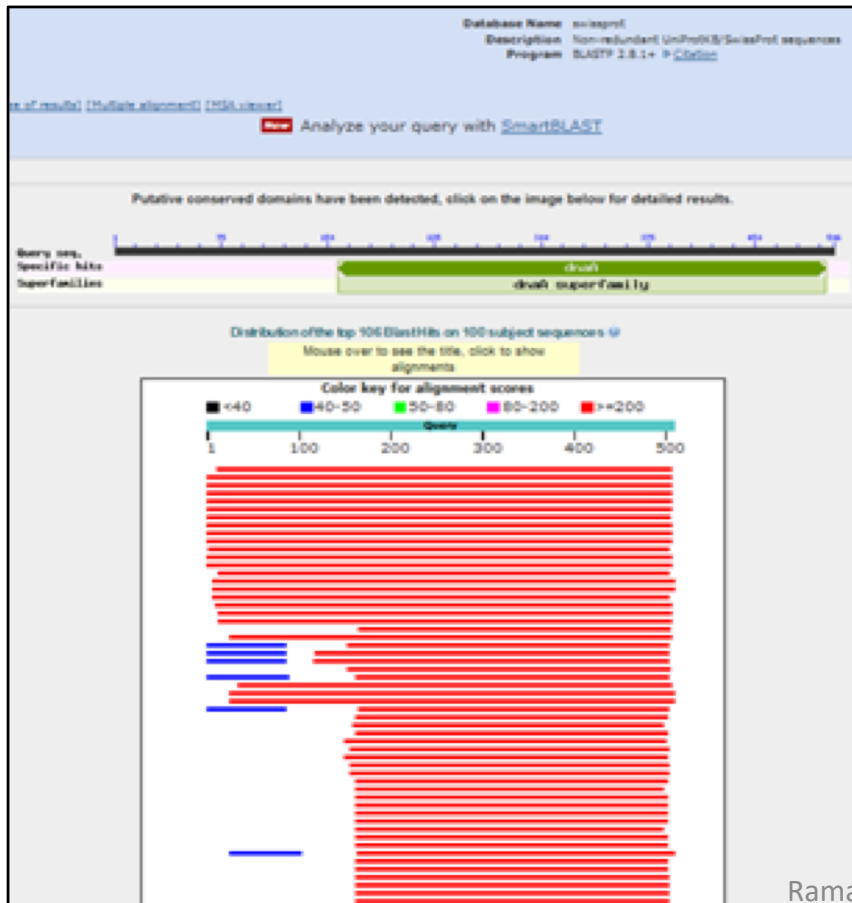
Based on how well the sequences match

- higher numerical values assigned for exact matches
- lower scores for “similar” amino acids and
- penalties assigned for gaps and for mismatches.

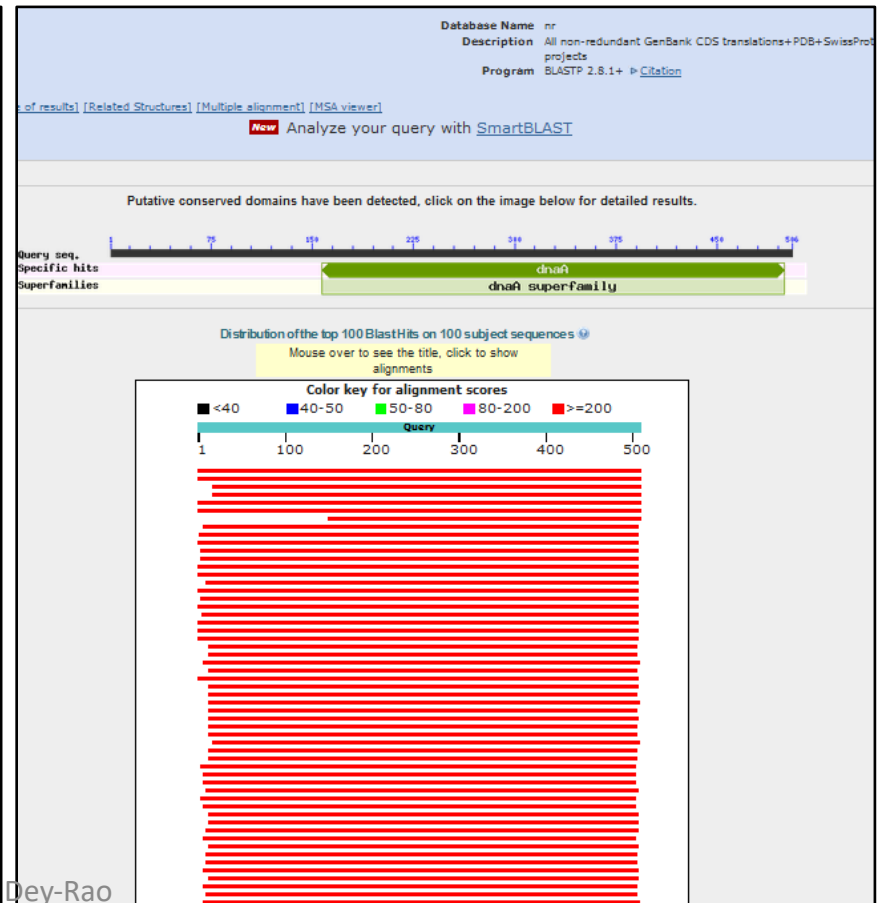
The sum of these numbers is the score.
 The higher the score, the more likely the alignment is significant.

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
<input type="checkbox"/> RecName: Full-Chromosomal replication Initiator protein DnaA	517	517	97%	1e-179	56%	A6W3V4.1
<input type="checkbox"/> RecName: Full-Chromosomal replication Initiator protein DnaA	498	498	99%	1e-172	52%	A1T102.1
<input type="checkbox"/> RecName: Full-Chromosomal replication Initiator protein DnaA	495	495	99%	2e-171	51%	A0PKB2.1
<input type="checkbox"/> RecName: Full-Chromosomal replication Initiator protein DnaA	494	494	99%	6e-171	51%	A0R7K1.1
<input type="checkbox"/> RecName: Full-Chromosomal replication Initiator protein DnaA	493	493	99%	2e-170	50%	B2HI46.1
<input type="checkbox"/> RecName: Full-Chromosomal replication Initiator protein DnaA	490	490	99%	1e-169	51%	Q1BG61.1
<input type="checkbox"/> RecName: Full-Chromosomal replication Initiator protein DnaA	488	488	99%	1e-168	50%	B1MDH6.1
<input type="checkbox"/> RecName: Full-Chromosomal replication Initiator protein DnaA	485	485	99%	2e-167	52%	P49991.2
<input type="checkbox"/> RecName: Full-Chromosomal replication Initiator protein DnaA	485	485	99%	3e-167	52%	C1AI28.1
<input type="checkbox"/> RecName: Full-Chromosomal replication Initiator protein DnaA	484	484	99%	4e-167	52%	A5TY69.1
<input type="checkbox"/> 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
<input type="checkbox"/> chromosomal replication Initiator protein DnaA [K]toococcus sedentarius]	1033	1033	100%	0.0	100%	WP_012801520.1
<input type="checkbox"/> chromosomal replication Initiator protein DnaA [K]toococcus sp. CUA-901]	1016	1016	100%	0.0	98%	WP_075867648.1
<input type="checkbox"/> chromosomal replication Initiator protein DnaA [K]toococcus sedentarius]	997	997	96%	0.0	100%	WP_049758582.1
<input type="checkbox"/> chromosomal replication Initiator protein DnaA [K]toococcus sp. CUA-901]	984	984	96%	0.0	99%	OLT32041.1
<input type="checkbox"/> chromosomal replication Initiator protein DnaA [K]toococcus schroeteri]	844	844	100%	0.0	88%	WP_101849155.1
<input type="checkbox"/> chromosomal replication Initiator protein DnaA [K]toococcus aerolatus]	789	789	100%	0.0	80%	WP_088618138.1
<input type="checkbox"/> chromosomal replication Initiator protein DnaA [K]toococcus sp. HMSC28H12]	687	687	70%	0.0	97%	QFS15515.1
<input type="checkbox"/> chromosomal replication Initiator protein DnaA [O]mithinimicrobium sp. AMA3305]	611	611	98%	0.0	62%	WP_114928598.1
<input type="checkbox"/> chromosomal replication Initiator protein DnaA [O]mithinimicrobium pekinqense]	610	610	99%	0.0	63%	WP_022920049.1
<input type="checkbox"/> chromosomal replication Initiator protein DnaA [A]ckobacteria bacterium]	603	603	99%	0.0	61%	RfK14929.1
<input type="checkbox"/> chromosomal replication Initiator protein DnaA [O]mithinimicrobium pekinqense]	602	602	99%	0.0	61%	WP_097189380.1
<input type="checkbox"/> chromosomal replication Initiator protein DnaA [O]mithinimicrobium sp. CPC2 203383]	600	600	99%	0.0	61%	WP_122261706.1

Results: nr and Swissprot databases

Click on link

Swissprot database

nr database

Alignments

Download GenPept Graphics

RecName: Full=Chromosomal replication initiator protein DnaA

Sequence ID: [sp|A6W3V4.1|DNAA_KINRD](#) 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-177	Compositional matrix adjust.	286/513(56%)	356/513(69%)	25/513(4%)
Query 11	AIWQEMVHLQAGLAPRDIGVRLATLVGLLEGSTALLAVKYDHVKDAVEGHLREDVSTA	70			
Sbjct 10	++N+ A+ L G+ +RL +GLL+GTALLAV D KD +E +RE ++ A SVWERALQQLDD-GVTQHQAFAVRLTRPLGLLDDGTALLAVENDLTKDVIQKVRPLTRA	68			
Query 71	LAEVLDRDIRLAVSVDPDAVSA-----AQEEAAPAPSPADEDDPATGEGPLSTAVD--	123			
Sbjct 69	L+E IRLAV+YDP E + P+ E + G + T +DG LSEAYGSPIRLAVTVDFSIGVLTPERTGEHSGGVSVFSVERE----RGSVLTLGLDGD	124			
Query 124	--AVEKHEGSSPARAGESVAPATTASLTA--TNSSPGVER-----DYSALNHKYTF	170			
Sbjct 125	+++ S T + PG R + S LN KY F GLHLDERRSGLSEEDSPLDDSDPDLFLTGKVDKRGPGTGRQFRPTTRIENTRSLNPKYIF	184			
Query 171	DTFVLGSSNRFAHAATAVAEAPARAYNPLFIYGGSGLGKTHLLHAIGHYARTLDSVSV	230			
Sbjct 185	+TFV+G+SNRFMAHAA AVAEAPFA+AYNPLFIYG SGLGKTHLLHAIGHYA+ L V+V ETPFVIGASNRFAHAAVAVAEAPAKAYNPLFIYGESGLGKTHLLHAIGHYAQNLYPGVQV	244			
Query 231	KYVNSEEFTNQFINAVSAGQANAFQRQYRDVLLIDDIQFLQGKQTMEEFFHTFNTLH	290			
Sbjct 245	+YVNSEEFTN FIN++ +A AFQR++RDVVDLLIDDIQFL K QT EEFHTFNTLH RYVNSEEFTNDFINSIRDDKAQAFQRHRDVLIDDIQFLSNKVGQTEEFFHTFNTLH	304			
Query 291	NSEKQIVITSDQPPKLSGFAERMRSRFEWGLLTDVQPPDLETRIAILRRKAAADKIDP	350			
Sbjct 305	N+ RQ+VITSD PFK+LSGF ERMRSRFEWGL+TDVQPPDLETRIAILR+EA ++L+P+ NASKQVVITSDLFPKQLSGFEERMRSRFEWGLITDVQPPDLETRIAILRKAIGERLEVP	364			
Query 351	DDVHLHIAKISSNIRELEGALRVTAFAASL+P+D LA VL+D+P++ +IT	410			
Sbjct 365	DDVNEIYAKISSNIRELEGALRVTAFAASLNQFVDMQLAEIQLRDLIPMEETFEITAA	424			
Query 411	MILEETAGYFVIVVEEIQGASRNRNLTARQIAMYLCRELTDLSLPKIKGFFGGRDHTV	470			
Sbjct 425	I+ +TA YF +++E++ G SRSL R ARQIAMYLCRELT+LSLPKIG+ FGGRDHTV AIMGQTASYFVITLEDLCGTSRSRSLVTARQIAMYLCRELTLSLPKIQHFGGRDHTV	484			
Query 471	MHAERKIRQLLEGDRRVYDEVSELTSIIRKKA 503				
Sbjct 485	MHAERKIRQ + E R Y++V+ELT+ I+K++ MHAERKIQQMAERRSTYNTVTELTNRIKQSG 517				

Download GenPept Graphics

chromosomal replication initiation protein [Ornithinimicrobium pekingense]

Sequence ID: [ref|WP_022920049.1](#) Length: 490 Number of Matches: 1

Range 1: 1 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	SQTPDDHATAIWEAMVHLQAGLAPRDIGVRLATLVGLLEGSTALLAVKYDHVKDAVEG	61			
Sbjct 3	SQ+P + A +WQ + L+ G+ RD LRL LVGLL+ TALLAV Y H K+ +E SQSPAESA-EVWQRVVSQLESQGVARDRAFLRLTQLVGLLDTTALLAVPYQHTKETLET	61			
Query 62	HLREDVSTALAEVLDRDIRLAVSVDPDAVSAQAEEAAPAPSPADEDDPATGEGPLSTAV	121			
Sbjct 62	TLRQPIVDALAGELGHDVRLAITVDEDLRRQVEDEGDP-APGPA-----VTEQVP--SDP	113			
Query 122	DGAVEKHEGSSPARAGESVAPATTASLTAITNSSPGVREDYSALNHKTYFDITFVLGSSNRF	181			
Sbjct 114	DRTPYRNSGAGP---GE---PRSDGHRTPSGAVQTASAEARLNPKYTFDITFVSGSSNRF	167			
Query 182	AHAATAVAEAPARAYNPLFIYGGSGLGKTHLLHAIGHYARTLDSVSVKVNSEEFFNQ	241			
Sbjct 168	AHAASLVAESPARAYNPLFIYGESGLGKTHLLHAIGHYARSYPGRVRYVNSEEFFND	227			
Query 242	FINAVSAGQANAFQRQYRDVLLIDDIQFLQGKQTMEEFFHTFNTLHNSKQIVITSD	301			
Sbjct 228	FINSIRDDKAGAFQRRYRNVDVLLIDDIQFLQGKQTMEEFFHTFNTLHNSKQIVITSD	287			
Query 302	QPPKLSGFAERMRSRFEWGLLTDVQPPDLETRIAILRRKAAADKIDPDDVHLHIAKISS	361			
Sbjct 288	QPPKLSGFAERMRSRFEWGLLTDVQPPDLETRIAILRKAQAQGMQLPDEVLELIGSKI	347			
Query 362	SNIRELEGALRVTAFAASL+P+D LA VLK+D+P +S II I+ E A YF	421			
Sbjct 348	STNIRELEGALRVTAFAASLSTTTPDAALASHVLKDIIPNSEAAITVPTMAEVADYFQ	407			
Query 422	ISVEEIQGASRNRNLTARQIAMYLCRELTDLSLPKIKGFFGGRDHTVMAERKIKQLL	481			
Sbjct 408	I+ ++ G SRSL R ARQIAMYLCRELTDLSLPKIG+ EFGGRDHTVMAERKI+QL+ ISNDLDCGTSRSRSLVNARQIAMYLCRELTDLSLPKIQEFGGRDHTVMAERKIRQLI	467			
Query 482	GEDRRVYDEVSELTSIIRKKAAR 504				
Sbjct 468	GE R +YD+++ELI IIRK +AR GERRALYDQITELTGIIRKASAR 490				

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 1×10^{-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.