DELTA-BLAST Hands-On

BLAST (http://blast.ncbi.nlm.nih.gov/), **B**asic **L**ocal **A**lignment **S**earch tool, is a suite of programs for finding similarities between biological sequences. **DELTA-BLAST** is a new, sensitive algorithm, intended for detection of distant protein homologs. Like "BLAST", "DELTA" is an acronym and it stands for **D**omain **E**nhanced **L**ookup **T**ime **A**ccelerated.

Outline:

- Compare DELTA-BLAST searches with blastp.
- Search for homologs for a human RAS oncogene family.
- Find which of the human RAS oncogene family are conserved throughout the Eukaryotes.
- Use the longest protein isoforms of these genes to search for homologs in Bacteria.
- Search against a few bacterial genera in the RefSeq protein database.

Step 1: Homologene search

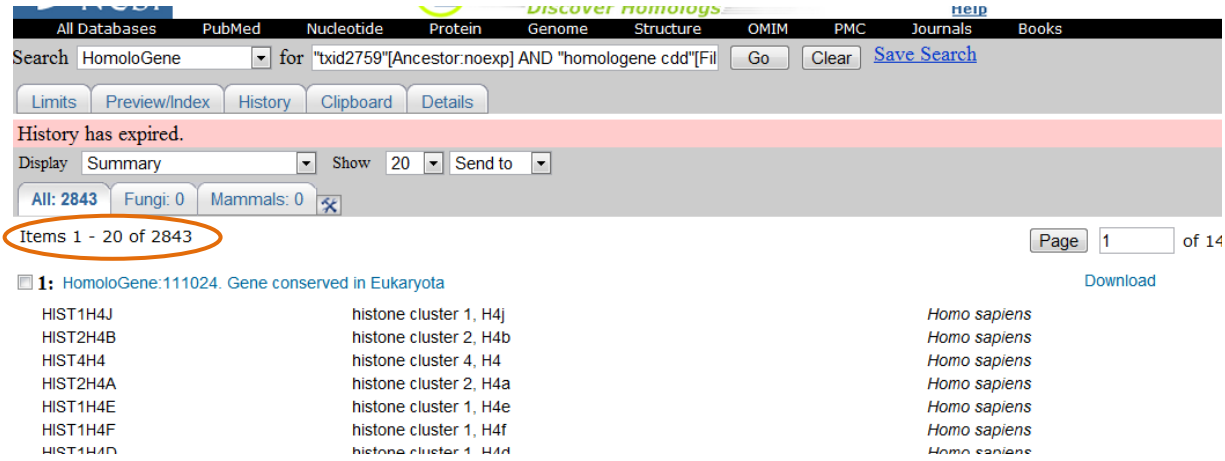
http://www.ncbi.nlm.nih.gov/homologene/

Search for clusters with the following requirements:

- conserved in Eukaryotes.
- proteins in the cluster have identifiable conserved domains.
- human gene should be included in the cluster.

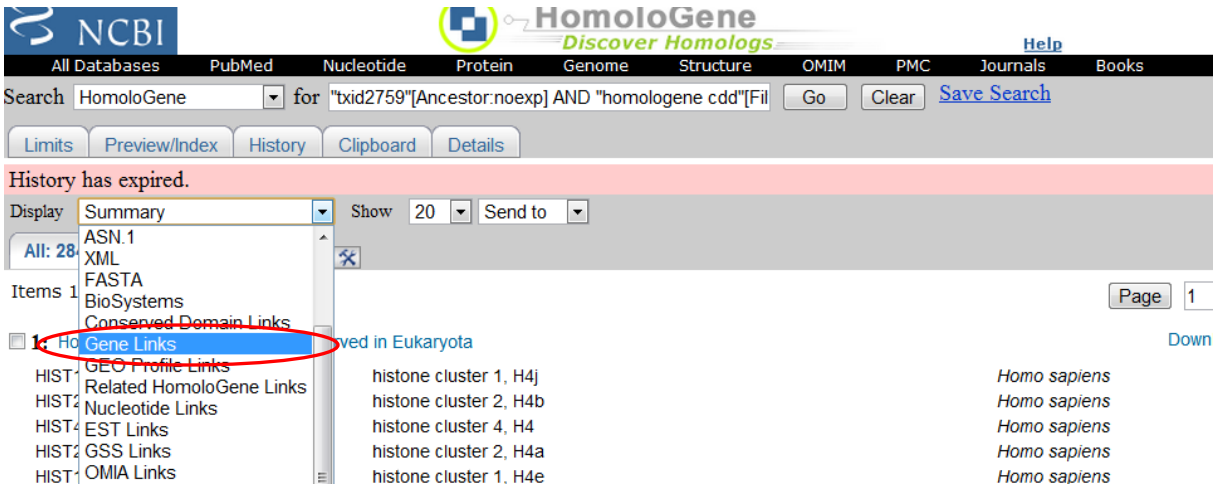
The search term for such requirements is:

"txid2759"[Ancestor:noexp] AND "homologene cdd"[Filter] AND human[organism]

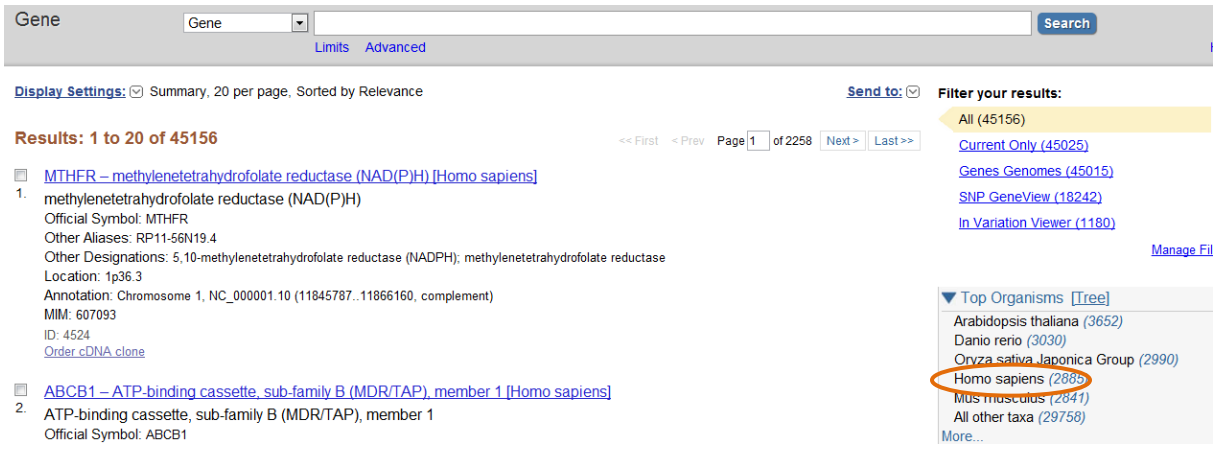


Step 2: The Gene database search

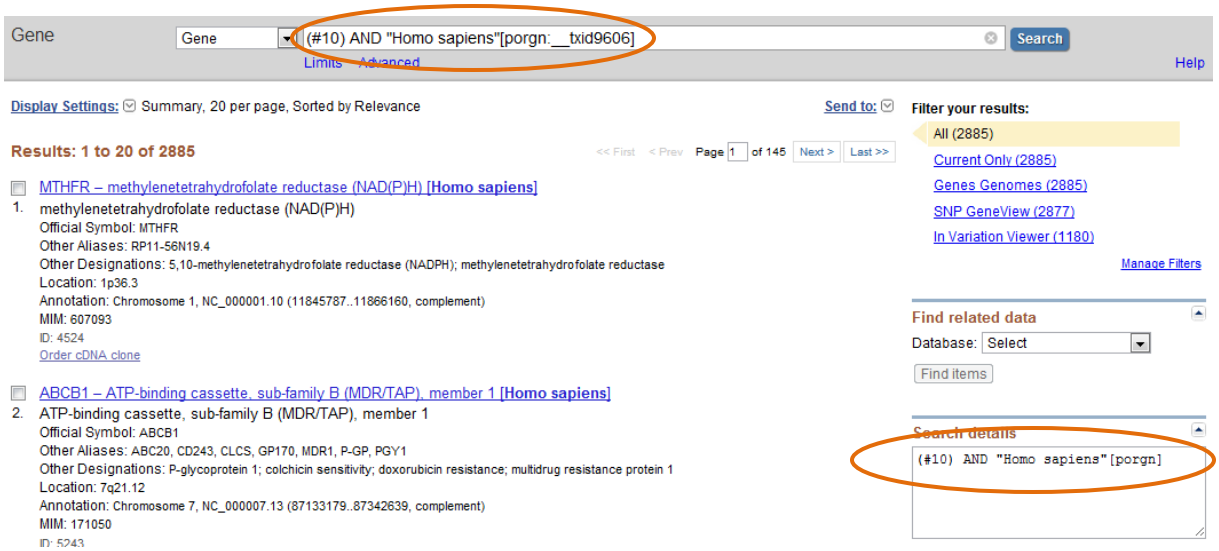
From the Homologene search result page, link to the Gene database.



Select the human records.

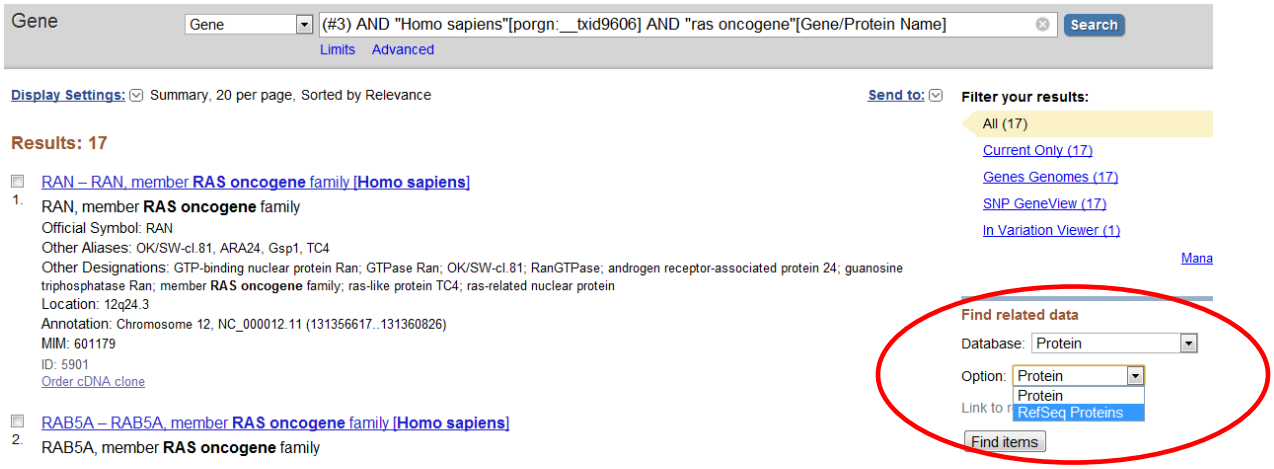


Note the search term in the Gene search box, once the page reloads:
Designation **#10** comes from the Homologene link. It will be different each time.
The **"Homo sapiens"[porgn:__txid9606]** term has been added from selecting the human organism from the taxonomy sorter. This is also reflected in the search details on the right side of the screen.



Add **"ras oncogene"[Gene/Protein Name]** to the search term in the query box and run the search. The result of the search is 17 human Gene records for genes that have "ras oncogene" in their name. According to our selection, they all have identified conserved domains and they are conserved throughout Eukaryotic kingdom.

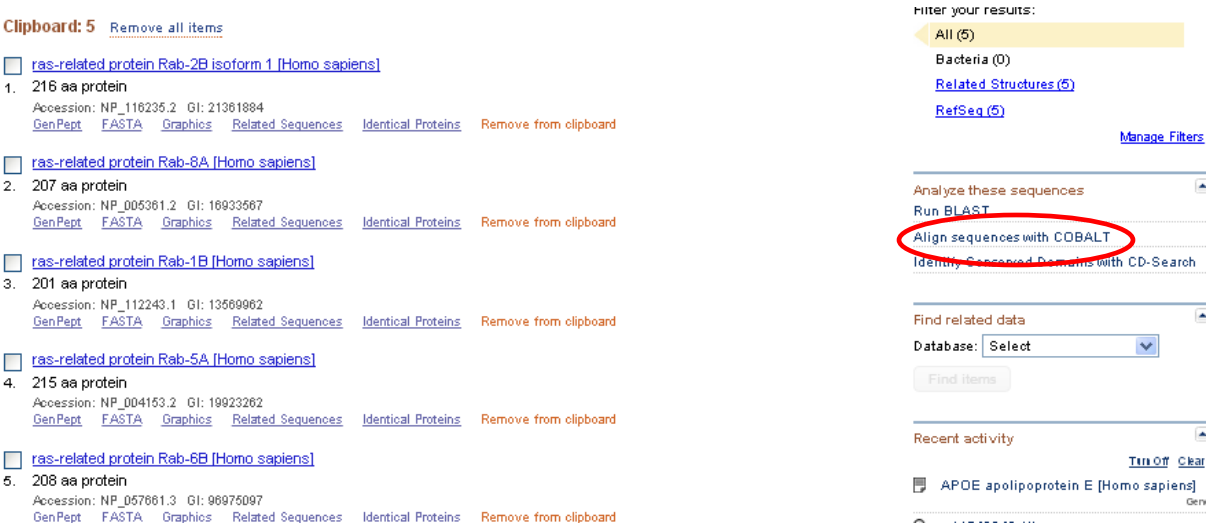
In find related data, select protein database.
Link to the Reference Sequence proteins for these genes.



Step 3: Protein records and protein alignment

After selecting find items. Some of the Genes have more than one transcript. These splice variants result in different protein isoforms.
You can select the longest isoforms of each gene by simply looking at the reported protein length on the result summary page. (You can send your section to clipboard and save the items in My NCBI as a collection).
Check those below:

NP_004153.2 (215 aa), NP_005361.2 (207 aa), NP_112243.1 (201 aa), NP_116235.2 (216 aa), NP_057661.3 (208 aa)

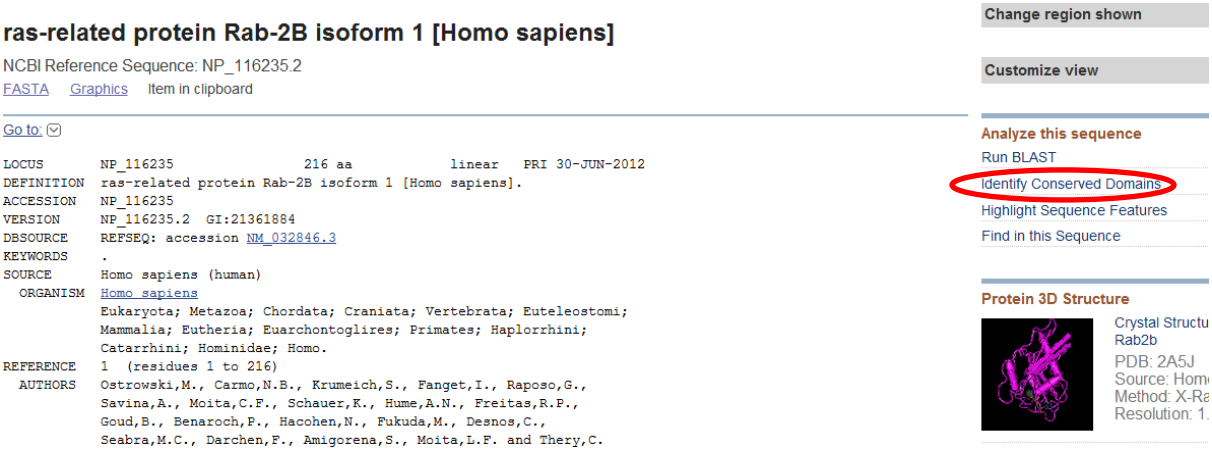


Select the link to the COBALT tool to see their multiple alignments.

Result of their alignment, Select NP_116235 protein.

Step 4: Conserved domain and PSSM

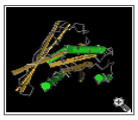
Go to the protein record of Rab-2B and select the link to Conserved Domains to see its corresponding conserved domain.



The domain is Rab2 and it belongs to the P-loop_NTPase superfamily, link to the domain.
You can see active GTP/Mg2+ binding site on the aligned sequences if you scroll down the record.
You can also view the PSSM for the CD (link indicated).

A PSSM, or Position-Specific Scoring Matrix, is a type of scoring matrix used in protein BLAST searches in which amino acid substitution scores are given separately for each position in a protein multiple sequence alignment.

cd01866: **Rab2, with user query added**



Rab GTPase family 2 (Rab2)

Rab2 is localized on cis-Golgi membranes and interacts with Golgi matrix proteins. Rab2 is also implicated in the maturation of vesicula tubular clusters (VTCs), which are microtubule-associated intermediates in transport between the ER and Golgi apparatus. In plants, I regulates vesicle trafficking between the ER and the Golgi bodies and is important to pollen tube growth. GTPase activating proteins (GAPs) interact with GTP-bound Rab and accelerate the hydrolysis of GTP to GDP. Guanine nucleotide exchange factors (GEFs) intera with GDP-bound Rabs to promote the formation of the GTP-bound state. Rabs are further regulated by guanine nucleotide dissociation inhibitors (GDIs), which facilitate Rab recycling by masking C-terminal lipid binding and promoting cytosolic localization. Most Rab GTPa contain a lipid modification site at the C-terminus, with sequence motifs CC, CXC, or CCX. Lipid binding is essential for membrane attachment, a key fea of most Rab proteins. Due to the presence of truncated sequences in this CD, the lipid modification site is not available for annotation.

Links

Source:	cd00154
Taxonomy:	Eukaryota
PubMed:	19 links
Book:	4 links
Protein:	Representatives
	Specific Protein
	Related Protein
	Related Structure
	Architectures
Superfamily:	cl09099
BioSystems:	239 links

Statistics

PSSM-Id:	206658
View PSSM:	cd01866
Aligned:	21 rows
Threshold Bit Score:	321.294

Conserved Features/Sites | PubMed References | Book References

GTP/Mg2+ | putative GEF | putative GDI | putative | Switch I region

Feature 1: GTP/Mg2+ binding site [chemical binding site]

Evidence:

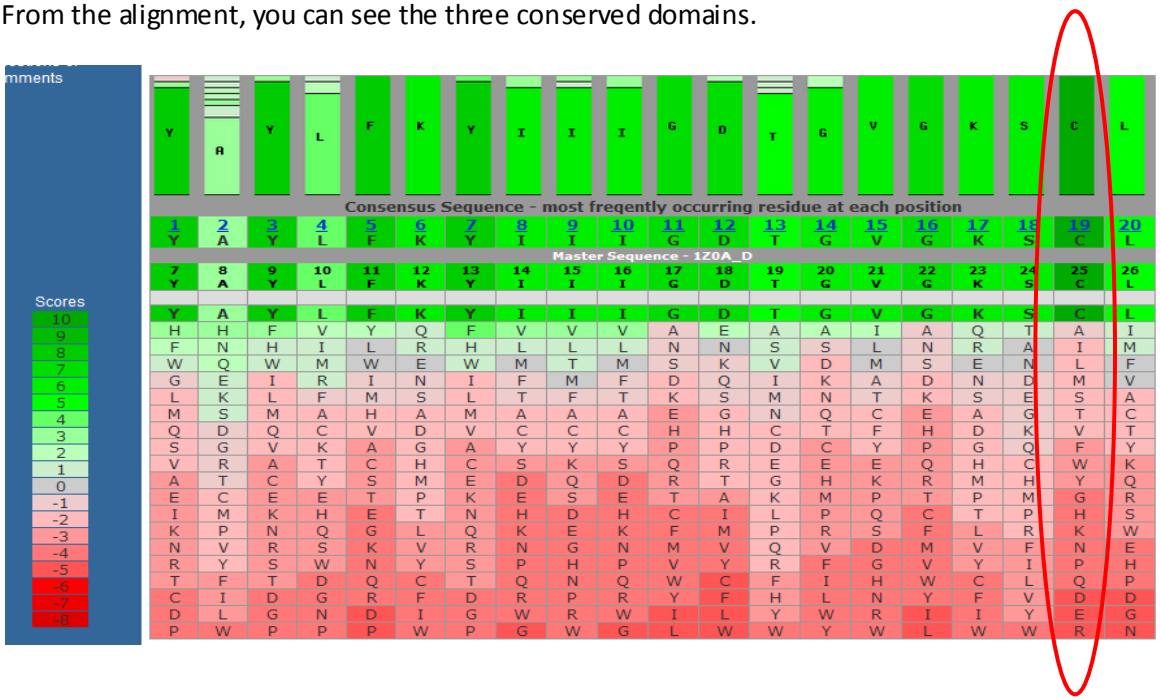
- **Comment:** The active conformation of Rab is stabilized by interations between the gamma phosphate of GTP and two critically conserved residues, Thr in switch I and Gly in switch II
- **Citation:** PMID 10196122
- **Structure:** 2A5J: Human Rab2b binds GDP and Mg2+, defined using 3.5 A contacts - View structure with Cn3D

Download Cn3D for Viewing 3D Structure | Scroll to Sequence Alignment Display

cd01866 is part of a hierarchy of related CD models.
Use the graphical representation to navigate this hierarchy.
cd01866 is a member of the superfamily cl09099.

```
Feature 1                    ######          #
1Z0A_D        7 YAYLFKYIIIGDTGVGKSCLLLQFTDKRFQPVHDLTIGVEFGARMITIDGKQIKLQIWDTAGQESFRSITRSYYRGAAGA 86
query         3 YAYLFKYIIIGDTGVGKSCLLLQFTDKRFQPVHDLTIGVEFGARMVNIDGKQIKLQIWDTAGQESFRSITRSYYRGAAGA 82
2A5J_A       18 GSYLFKYIIIGDTGVGKSCLLLQFTDKRFQPVHDLTIGVEFGARMVNIDGKQIKLQIWDTAGQESFRSITRSYYRGAAGA 97
gi 13397937   4 YEYLFKYIIIGDTGVGKSCLLLQFTDKRFRADHDLTIGVEFGARLVNLDNKQIKLQIWDTAGQESFRSITRSYYRGAAGA 83
gi 28556900   3 YAYLFKYIIIGDTGVGKSCLLLQFTDKRFQPVHDLTIGVEFGARMITIDGRQIKLQIWDTAGQESFRSITRSYYRGAAGA 82
gi 30023652   3 YAYLFKYIIIGDTGVGKSCLLLQFTDKRFQPVHDLTIGVEFGARMITIDNKPIKLQIWDTAGQESFRSITRSYYRGAAGA 82
gi 45382561   3 YAYLFKYIIIGDTGVGKSCLLLQFTDKRFQPVHDLTIGVEFGARMITIDGKQIKLQIWDTAGQESFRSITRSYYRGAAGA 82
gi 27881912   3 YAYLFKYIIIGDTGVGKSCLLLQFTDKRFQPVHDLTIGVEFGARMITIDGKQIKLQIWDTAGQESFRSITRSYYRGAAGA 82
gi 21645117   3 YAYLFKYIIIGDTGVGKSCLLLQFTDKRFQPVHDLTIGVEFGARMITIDGKQIKLQIWDTAGQEAFRSITRSYYRGAAGA 82
gi 49257196   3 YAYLFKYIIIGDTGVGKSCLLLQFTDKRFQPVHDLTIGVEFGARMINIDGKPIKLQIWDTAGQESFRSITRSYYRGAAGA 82

Feature 1                                        ## #                        ###
1Z0A_D       87 LLVYDITRRDTFNHLTTWLEDARQHSNSNMVIMLIGNKSDL.[3].REVKKEEGEAFAREHGLIFMETSAKTASNVEEAF 164
query        83 LLVYDITRRETFNHLTSWLEDARQHSSSNMVIMLIGNKSDL.[3].RDVKREEGEAFAREHGLIFMETSAKTACNVEEAF 160
2A5J_A       98 LLVYDITRRETFNHLTSWLEDARQHSSSNMVIMLIGNKSDL.[3].RDVKREEGEAFAREHGLIFMETSAKTACNVEEAF 175
gi 13397937  84 LLVYDITRRETFNHLNRWLDEVRQNSNPHMAIILVGNKCDL.[2].REVSAEEGAQFARQNGLIFLETSAKTAKNVEEAF 160
gi 28556900  83 LLVYDITRRDTFNHLTTWLEDARQHSSSNMVIMLIGNKSDL.[3].REVKKEEGEAFAREHGLIFMETSAKTAANVEEAF 160
gi 30023652  83 LLVYDITRRETFNHLASWLEDARQHANANMTIMLIGNKCDL.[3].RAVSTEEGEQFAKEHGLIFMEASAKTAQNVEEAF 160
gi 45382561  83 LLVYDITRRDTFNHLTTWLEDARQHSNSNMVIMLIGNKSDL.[3].REVKKEEGEAFAREHGLIFMETSAKTASNVEEAF 160
gi 27881912  83 LLVYDITRRDTFNHLTTWLEDARQHSNSNMVIMLIGNKSDL.[3].REVKKEEGEAFAREHGLIFMETSAKTASNVEEAF 160
gi 21645117  83 LLVYDITRRETFNHLTTWLEDARQHSNSNMVIMLIGNKSDL.[3].REVKKEEGEAFAREHGLVFMETSARTAANVEEAF 160
gi 49257196  83 LLVYDITRRETFSHLTSWLEDARQHSSSNMVIILIGNKSDL.[3].RDVSREEGEAFAREHGLIFMETSAKTSANVEEAF 160
```

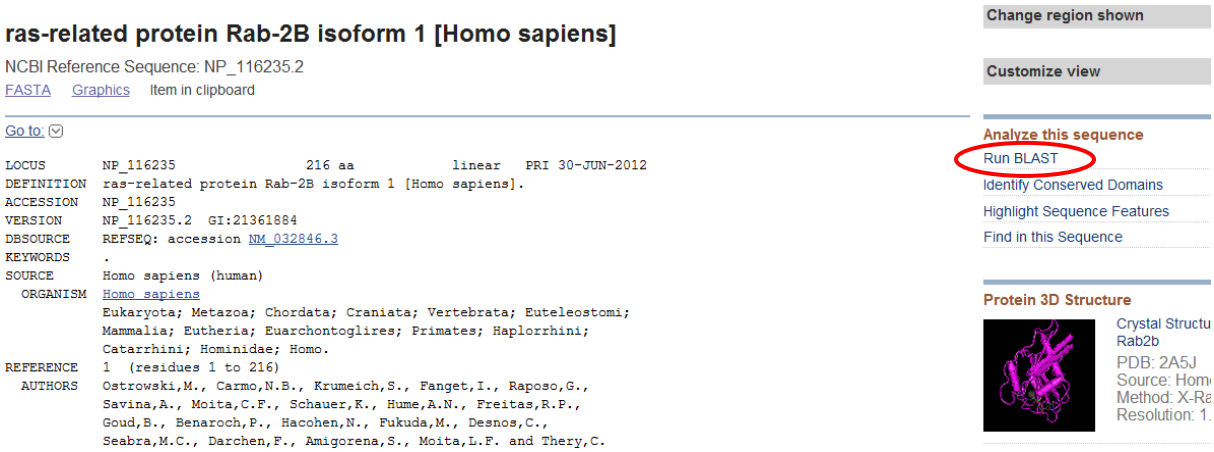
From the alignment, you can see the three conserved domains.



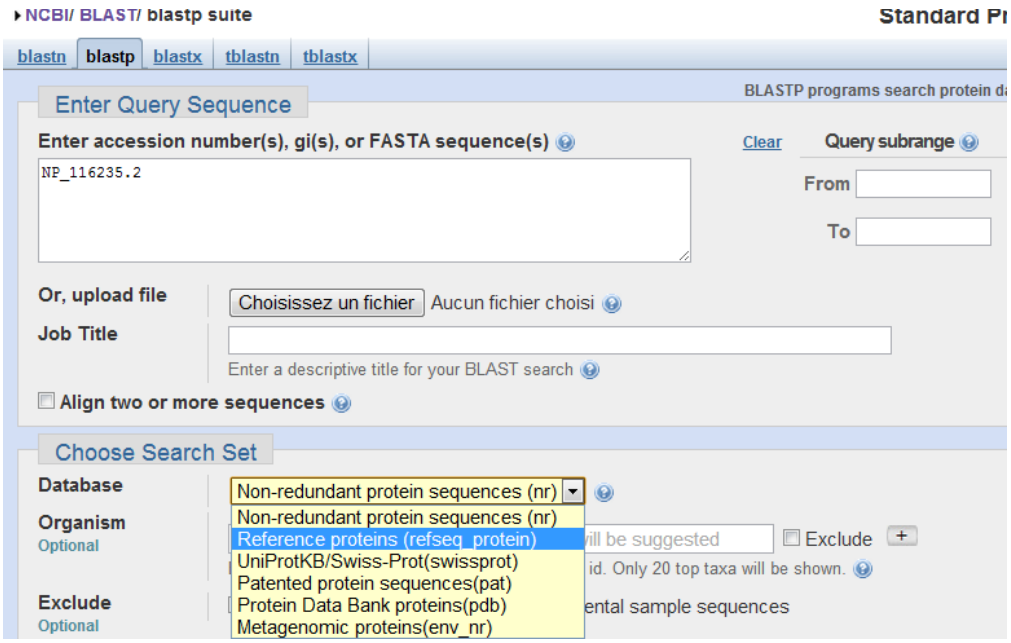
Cysteine at position 19 has score 10.

Step 5: BLAST searches

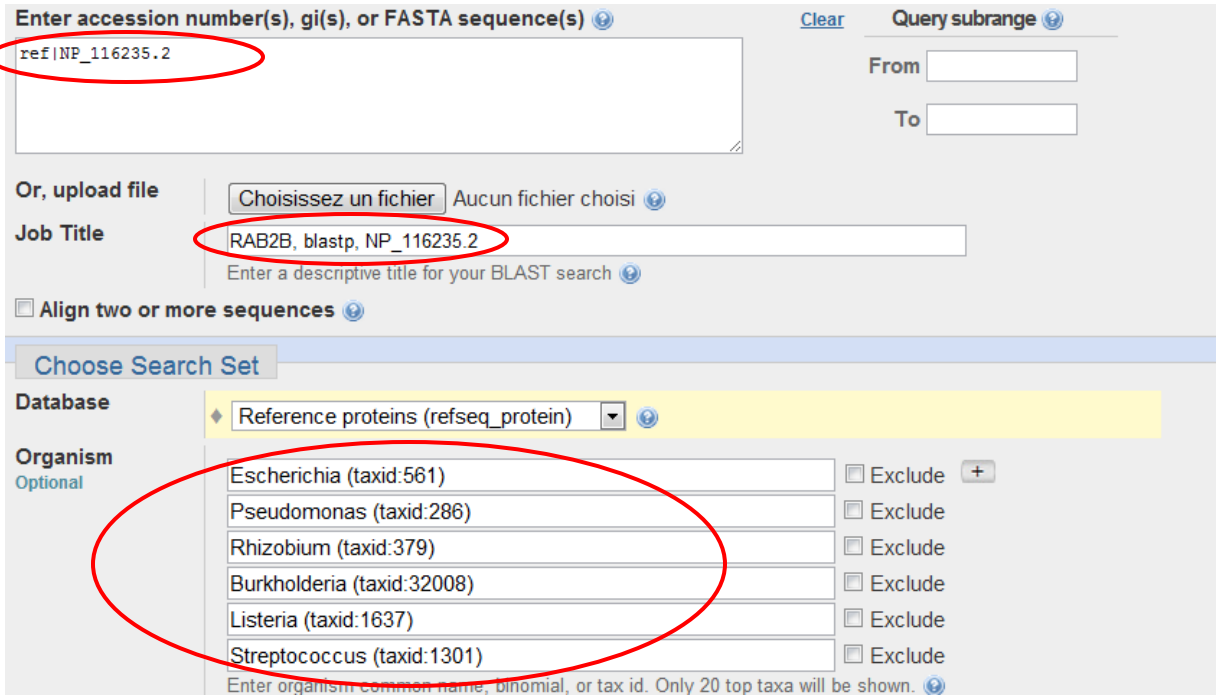
Select the link to the BLAST page from the protein record.



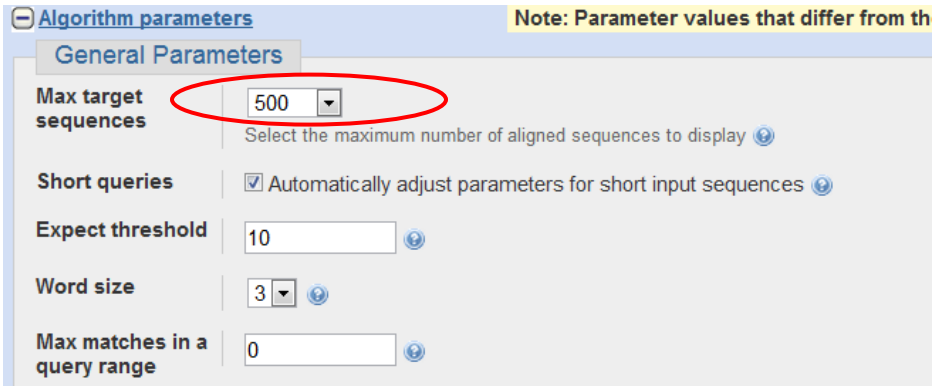
The link leads to the blastp page with the query accession added.
Change the database to search to the Reference proteins.



NP_116235.2 is the accession number of the query for the Rab2 isoform.
Add all the information, as you can see in the picture below.

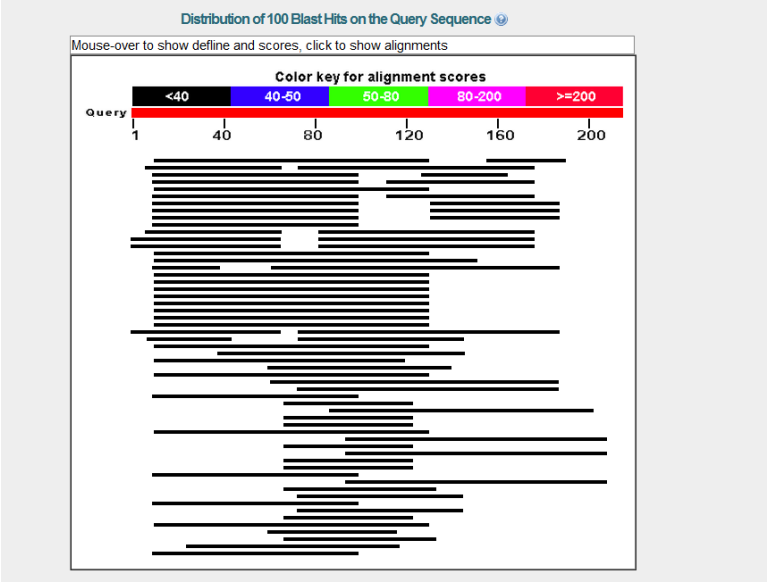


Increase the Max target sequences to 500.



BLASTp Output for RAB2:

Alignment result.

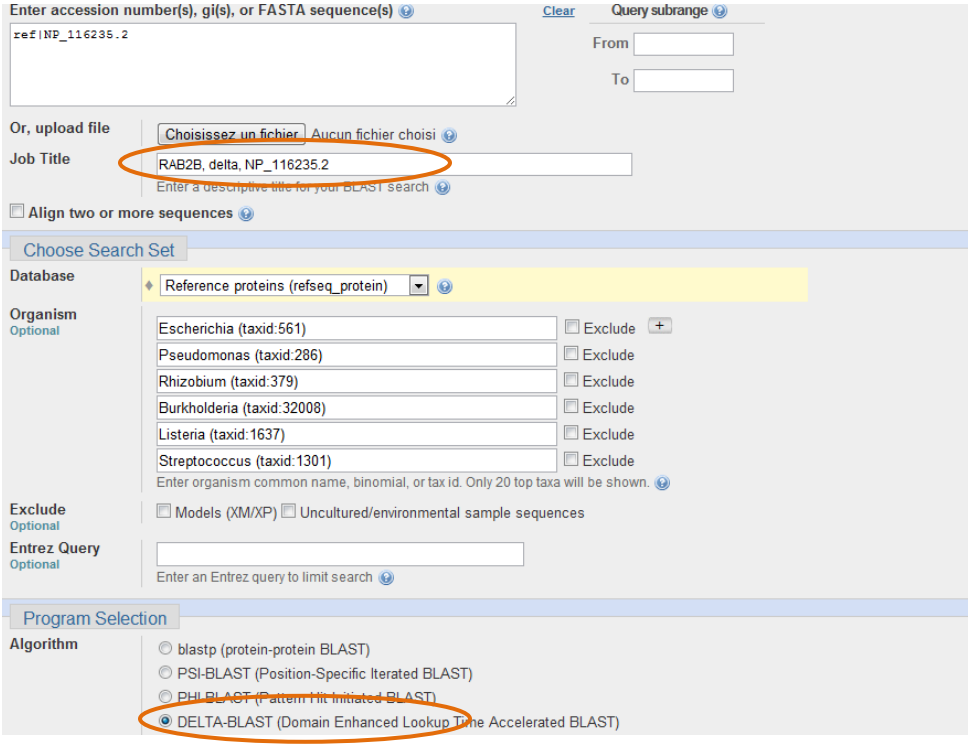



```
            DLE+R D+ R
Sbjct  152  DLETRLDINR  161


>ref|ZP_05638623.1|  twitching motility protein [Pseudomonas syringae pv. tabaci str.
ATCC 11528]
Length=389

 Score = 34.7 bits (78),  Expect = 3e-04, Method: Compositional matrix adjust.
 Identities = 25/91 (27%), Positives = 39/91 (43%), Gaps = 0/91 (0%)

Query  10   IIIGDTGVGKSCLLQFTDKRFQPVHDLTIGVEFGARMVNIDGKQIKLQIWDTAGQESFR  69
            + +G TG GKS  L     D R +       I +E     ++    K I  Q       SFR
Sbjct  127  LFVGATGSGKSTSLAALIDYRNRNASGHIITIEDPVEFIHRHKKSIVNQREVGVDTRSFR  186

Query  70   SITRSYYRGAAGALLVYDITRRETFNHLTSW  100
            +  ++  R A   +L+ +I  RET  H  ++
Sbjct  187  AALKNTLRQAPDVILIGEIRDRETMEHALAF  217
```

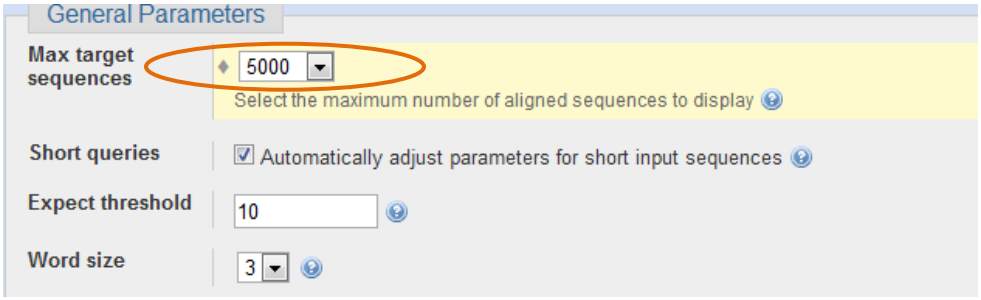
The blastp alignment covers only one of the active site.

DELTA-BLAST for RAB2:

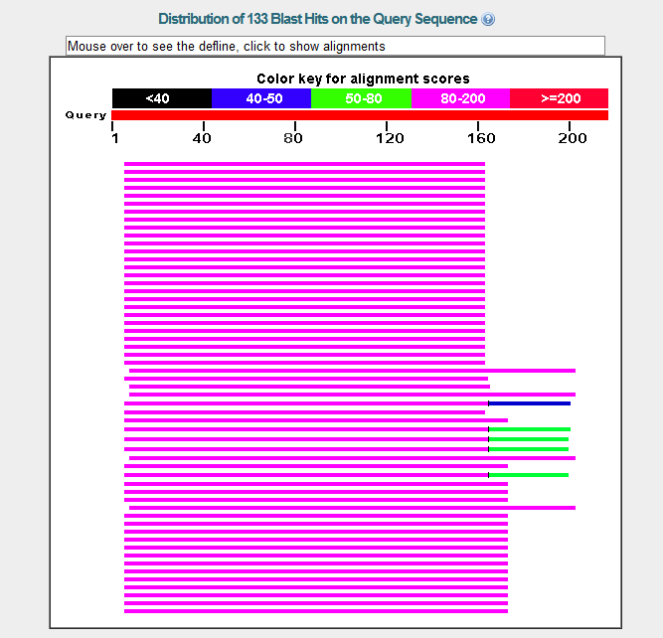
Change the job title, replace blastp by delta.
Select Delta-Blast for algorithm.



Increase the Max target sequences to 5000.



The result:



GENE ID: 915398 trmE | tRNA modification GTPase TrmE
[Escherichia coli O157:H7 str. Sakai] (10 or fewer PubMed links)

Score = 92.0 bits (227), Expect = 7e-22, Method: Composition-based stats.
Identities = 37/166 (22%), Positives = 60/166 (36%), Gaps = 19/166 (11%)

```
Query  7    FKYIIIGDTGVGKSCLLLQFTDKRFQPVHD-LTIGVEFGARMVNIDGKQIKLQIWDTAGQ  65
            K +I G    GKS LL     +    V D      +     ++IDG  + L I DTAG
Sbjct  217  MKVVIAGRPNAGKSSLLNALAGREAAIVTDIAGTTRDVLREHIHIDG--MPLHIIDTAGL  274

Query  66   ESFRS------ITRSY--YRGAAGALLVYDITRRETFNHLTSWLEDARQHSSSNMVIMLI  117
                       I R++     A   L + D T  +  +    W  +      + + I ++
Sbjct  275  REASDEVERIGIERAWQEIEQADRVLFMVDGTTDAVDPAEIW-PEFIARLPAKLPITVV  333

Query  118  GNKSDLESRREVKREEGEAFAREHGLIFMETSAKTACNVEEAFINT  163
             NK+D+         E    +G    SAKT   V+    +
Sbjct  334  RNKADI--------TGETLGMSEVNGHALIRLSAKTGEGVDVLRNHL  372
```

```
Feature 1                    ######          #
1Z0A_D        7 YAYLFKYIIIGDTGVGKSCLLLQFTDKRFQPVHDLTIGVEFGARMITIDGKQIKLQIWDTAGQESFRSITRSYYRGAAGA 86
query         3 YAYLFKYIIIGDTGVGKSCLLLQFTDKRFQPVHDLTIGVEFGARMVNIDGKQIKLQIWDTAGQESFRSITRSYYRGAAGA 82
2A5J_A       18 GSYLFKYIIIGDTGVGKSCLLLQFTDKRFQPVHDLTIGVEFGARMVNIDGKQIKLQIWDTAGQESFRSITRSYYRGAAGA 97
gi 13397937   4 YEYLFKYIIIGDTGVGKSCLLLQFTDKRFRADHDLTIGVEFGARLVNLDNKQIKLQIWDTAGQESFRSITRSYYRGAAGA 83
gi 28556900   3 YAYLFKYIIIGDTGVGKSCLLLQFTDKRFQPVHDLTIGVEFGARMITIDGRQIKLQIWDTAGQESFRSITRSYYRGAAGA 82
gi 30023652   3 YAYLFKYIIIGDTGVGKSCLLLQFTDKRFQPVHDLTIGVEFGARMITIDNKPIKLQIWDTAGQESFRSITRSYYRGAAGA 82
gi 45382561   3 YAYLFKYIIIGDTGVGKSCLLLQFTDKRFQPVHDLTIGVEFGARMITIDGKQIKLQIWDTAGQESFRSITRSYYRGAAGA 82
gi 27881912   3 YAYLFKYIIIGDTGVGKSCLLLQFTDKRFQPVHDLTIGVEFGARMITIDGKQIKLQIWDTAGQESFRSITRSYYRGAAGA 82
gi 21645117   3 YAYLFKYIIIGDTGVGKSCLLLQFTDKRFQPVHDLTIGVEFGARMITIDGKQIKLQIWDTAGQEAFRSITRSYYRGAAGA 82
gi 49257196   3 YAYLFKYIIIGDTGVGKSCLLLQFTDKRFQPVHDLTIGVEFGARMINIDGKPIKLQIWDTAGQESFRSITRSYYRGAAGA 82

Feature 1                                          ## #                          ###
1Z0A_D       87 LLVYDITRRDTFNHLTTWLEDARQHSNSNMVIMLIGNKSDL.[3].REVKKEEGEAFAREHGLIFMETSAKTACNVEEAF 164
query        83 LLVYDITRRETFNHLTSWLEDARQHSSSNMVIMLIGNKSDL.[3].RDVKREEGEAFAREHGLIFMETSAKTACNVEEAF 160
2A5J_A       98 LLVYDITRRETFNHLTSWLEDARQHSSSNMVIMLIGNKSDL.[3].RDVKREEGEAFAREHGLIFMETSAKTACNVEEAF 175
gi 13397937  84 LLVYDITRRETFNHLNRWLDEVRQNSNPHMAIILVGNKCDL.[2].REVSAEEGAQFARQNGLIFLETSAKTAKNVEEAF 160
gi 28556900  83 LLVYDITRRDTFNHLTTWLEDARQHSSSNMVIMLIGNKSDL.[3].REVKKEEGEAFAREHGLIFMETSAKTAANVEEAF 160
gi 30023652  83 LLVYDITRRETFNHLASWLEDARQHANANMTIMLIGNKCDL.[3].RAVSTEEGEQFAKEHGLIFMEASAKTAQNVEEAF 160
gi 45382561  83 LLVYDITRRDTFNHLTTWLEDARQHSNSNMVIMLIGNKSDL.[3].REVKKEEGEAFAREHGLIFMETSAKTASNVEEAF 160
gi 27881912  83 LLVYDITRRDTFNHLTTWLEDARQHSNSNMVIMLIGNKSDL.[3].REVKKEEGEAFAREHGLIFMETSAKTASNVEEAF 160
gi 21645117  83 LLVYDITRRETFNHLTTWLEDARQHSNSNMVIMLIGNKSDL.[3].REVKKEEGEAFAREHGLVFMETSAKTAANVEEAF 160
gi 49257196  83 LLVYDITRRETFSHLTSWLEDARQHSSSNMVIILIGNKSDL.[3].RDVSREEGEAFAREHGLIFMETSAKTSANVEEAF 160
```

For DELTA-BLAST, all three active sites are in the alignment.