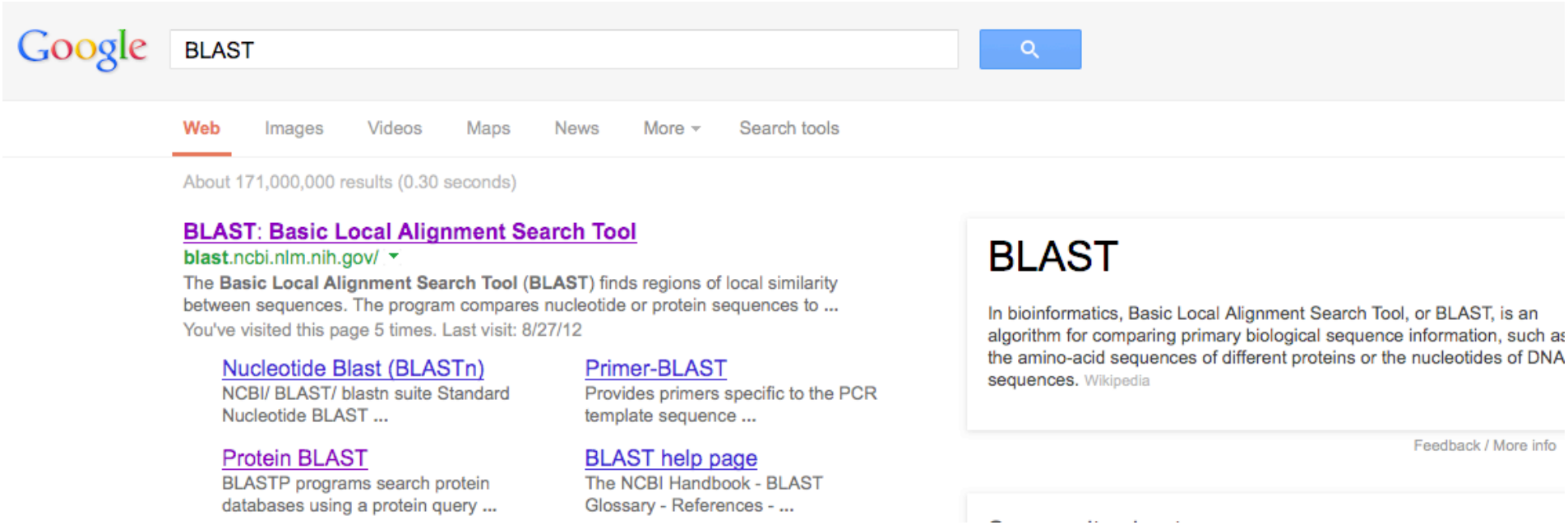
BLAST assignment instructions

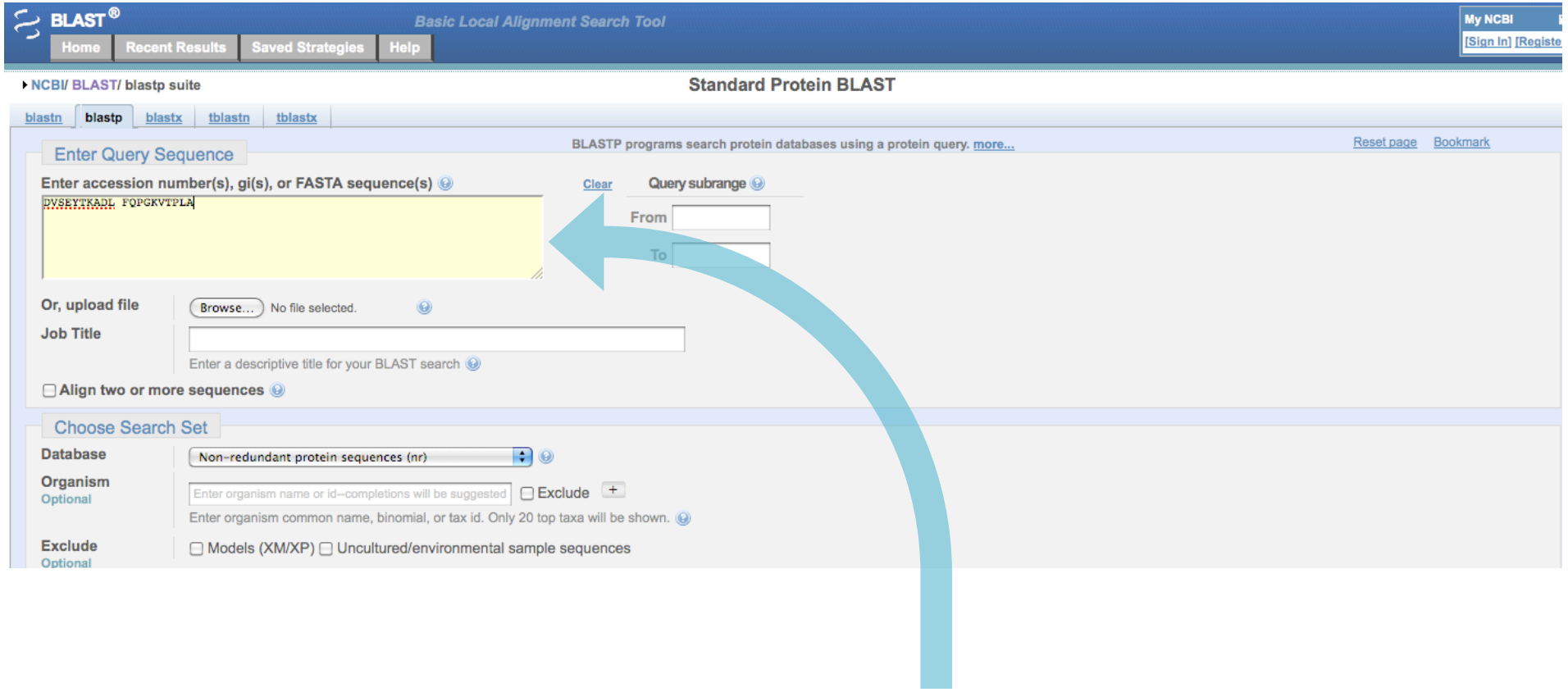
- BLAST 20-mer sequence
- Obtain your protein sequence
- Obtain DNA seq. that codes for protein
- Design primers to clone gene into vector
- Make a multiple sequence alignment
- Make an informative image of protein
- Write 0.5 page summary on protein

BLAST 20-mer sequence



Search for the **B**asic **L**ocal **A**lignment **S**earch **T**ool (BLAST) on the **N**ational **C**enter for **B**iotechnology **I**nformation (NCBI) website, and find the Protein BLAST tool

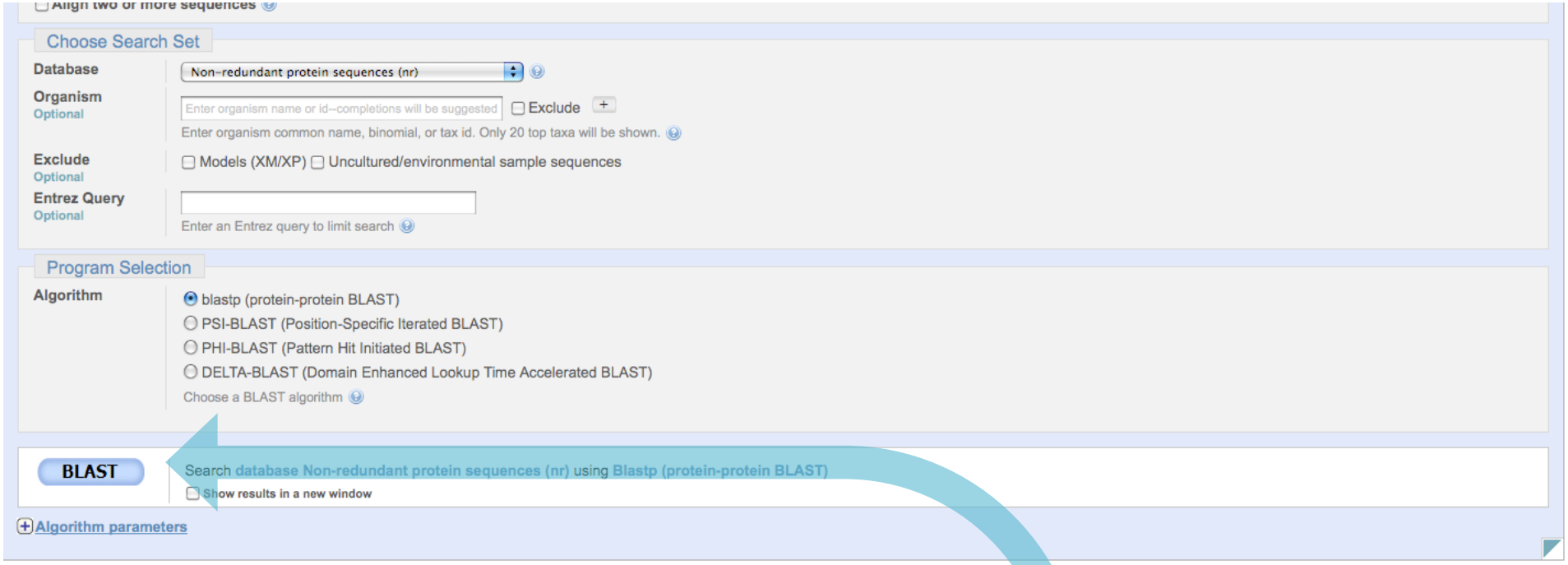
BLAST 20-mer sequence



Input your 20-mer sequence into the “Query Sequence” box …

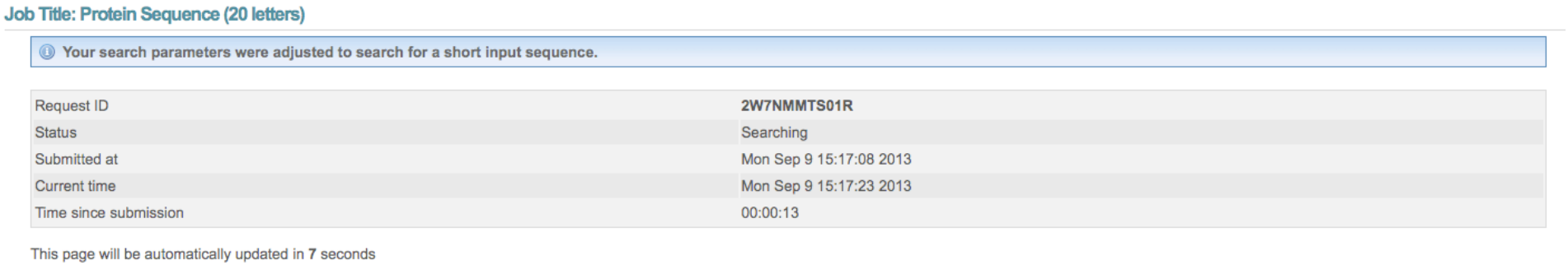
The sequence I put here is `DVSEYTKADL FQPGKVTPLA`

BLAST 20-mer sequence



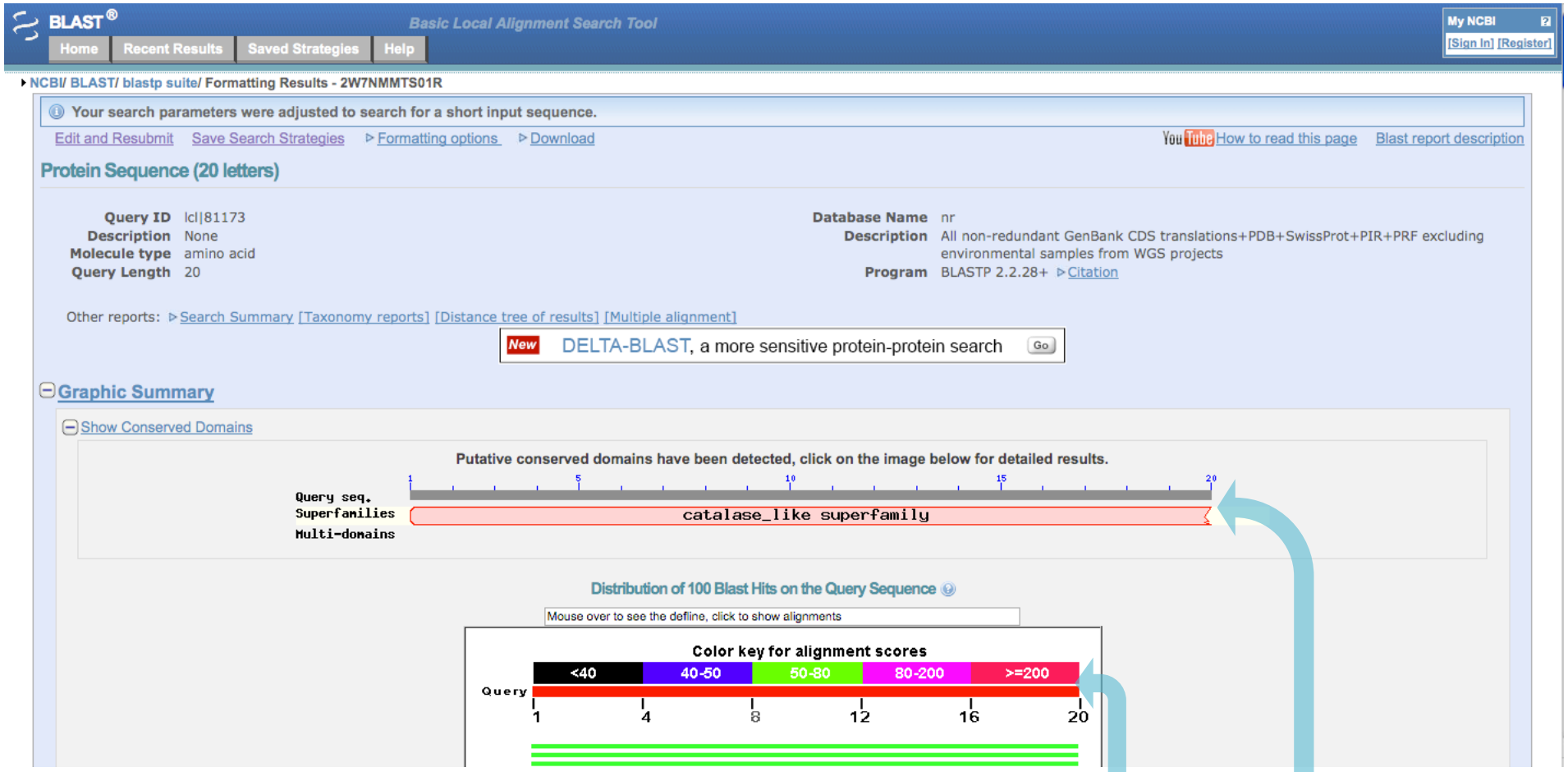
… and scroll to the bottom and hit the large, blue BLAST (!) button

BLAST 20-mer sequence



Be patient while it compares your 20-mer sequence to the entire database and ranks the matches for you. This can take a while during regular work hours.

BLAST 20-mer sequence



It has finished! It has recognized that it shares similarity to a protein superfamily (catalase in this instance) and has graphically shown the matches and their respective scores.

BLAST 20-mer sequence

Descriptions

Sequences producing significant alignments:

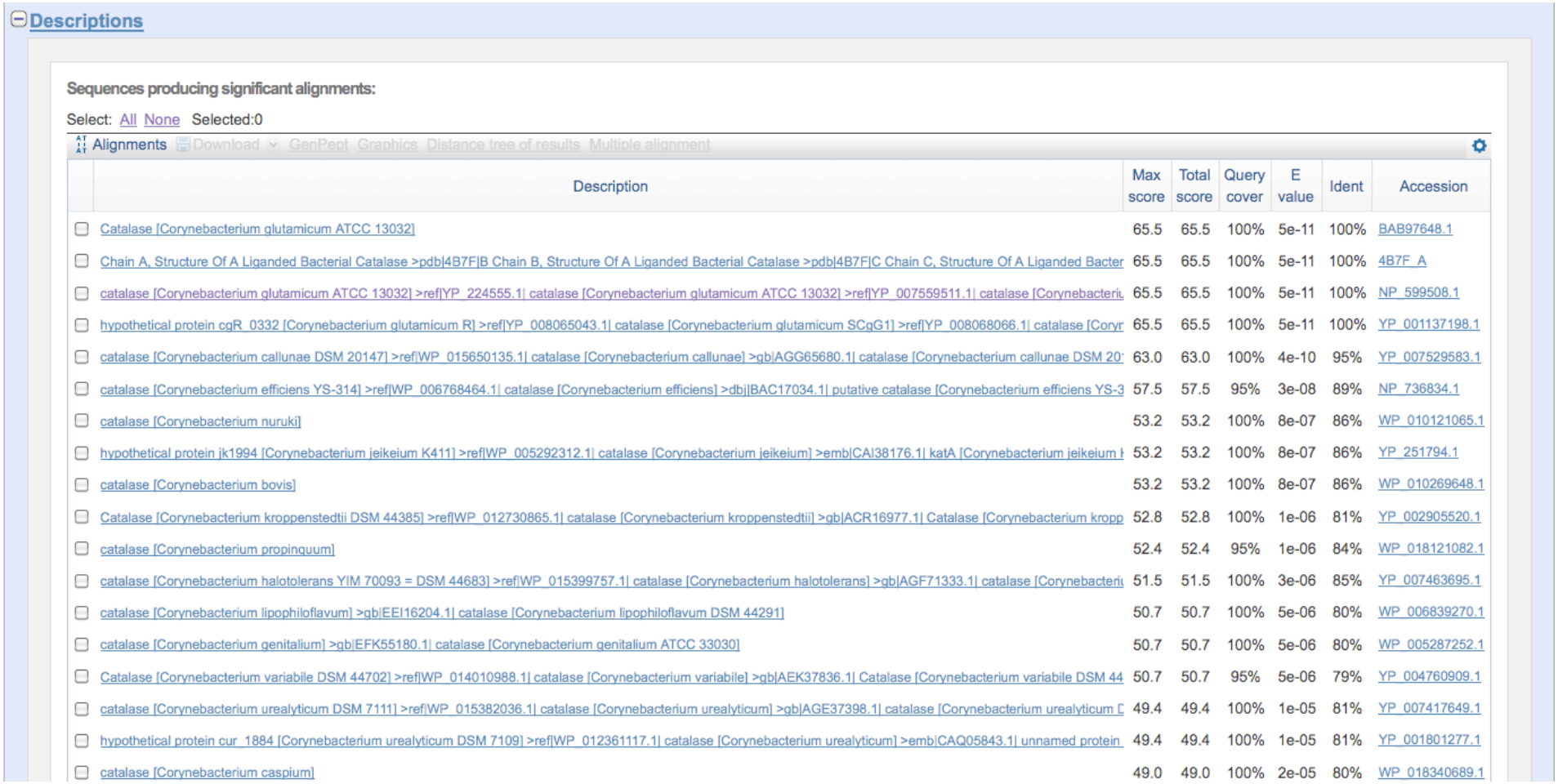
Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
Catalase [Corynebacterium glutamicum ATCC 13032]	65.5	65.5	100%	5e-11	100%	BAB97648.1
Chain A, Structure Of A Liganded Bacterial Catalase >pdb|4B7F|B Chain B, Structure Of A Liganded Bacterial Catalase >pdb|4B7F|C Chain C, Structure Of A Liganded Bacter	65.5	65.5	100%	5e-11	100%	4B7F_A
catalase [Corynebacterium glutamicum ATCC 13032] >ref|YP_224555.1| catalase [Corynebacterium glutamicum ATCC 13032] >ref|YP_007559511.1| catalase [Corynebacteriu	65.5	65.5	100%	5e-11	100%	NP_599508.1
hypothetical protein cgR_0332 [Corynebacterium glutamicum R] >ref|YP_008065043.1| catalase [Corynebacterium glutamicum SCgG1] >ref|YP_008068066.1| catalase [Coryr	65.5	65.5	100%	5e-11	100%	YP_001137198.1
catalase [Corynebacterium callunae DSM 20147] >ref|WP_015650135.1| catalase [Corynebacterium callunae] >gb|AGG65680.1| catalase [Corynebacterium callunae DSM 20	63.0	63.0	100%	4e-10	95%	YP_007529583.1
catalase [Corynebacterium efficiens YS-314] >ref|WP_006768464.1| catalase [Corynebacterium efficiens] >dbj|BAC17034.1| putative catalase [Corynebacterium efficiens YS-3	57.5	57.5	95%	3e-08	89%	NP_736834.1
catalase [Corynebacterium nuruki]	53.2	53.2	100%	8e-07	86%	WP_010121065.1
hypothetical protein jk1994 [Corynebacterium jeikeium K411] >ref|WP_005292312.1| catalase [Corynebacterium jeikeium] >emb|CAI38176.1| katA [Corynebacterium jeikeium K	53.2	53.2	100%	8e-07	86%	YP_251794.1
catalase [Corynebacterium bovis]	53.2	53.2	100%	8e-07	86%	WP_010269648.1
Catalase [Corynebacterium kroppenstedtii DSM 44385] >ref|WP_012730865.1| catalase [Corynebacterium kroppenstedtii] >gb|ACR16977.1| Catalase [Corynebacterium kropp	52.8	52.8	100%	1e-06	81%	YP_002905520.1
catalase [Corynebacterium propinquum]	52.4	52.4	95%	1e-06	84%	WP_018121082.1
catalase [Corynebacterium halotolerans YIM 70093 = DSM 44683] >ref|WP_015399757.1| catalase [Corynebacterium halotolerans] >gb|AGF71333.1| catalase [Corynebacteriu	51.5	51.5	100%	3e-06	85%	YP_007463695.1
catalase [Corynebacterium lipophiloflavum] >gb|EEI16204.1| catalase [Corynebacterium lipophiloflavum DSM 44291]	50.7	50.7	100%	5e-06	80%	WP_006839270.1
catalase [Corynebacterium genitalium] >gb|EFK55180.1| catalase [Corynebacterium genitalium ATCC 33030]	50.7	50.7	100%	5e-06	80%	WP_005287252.1
Catalase [Corynebacterium variabile DSM 44702] >ref|WP_014010988.1| catalase [Corynebacterium variabile] >gb|AEK37836.1| Catalase [Corynebacterium variabile DSM 44	50.7	50.7	95%	5e-06	79%	YP_004760909.1
catalase [Corynebacterium urealyticum DSM 7111] >ref|WP_015382036.1| catalase [Corynebacterium urealyticum] >gb|AGE37398.1| catalase [Corynebacterium urealyticum D	49.4	49.4	100%	1e-05	81%	YP_007417649.1
hypothetical protein cur_1884 [Corynebacterium urealyticum DSM 7109] >ref|WP_012361117.1| catalase [Corynebacterium urealyticum] >emb|CAQ05843.1| unnamed protein	49.4	49.4	100%	1e-05	81%	YP_001801277.1
catalase [Corynebacterium caspium]	49.0	49.0	100%	2e-05	80%	WP_018340689.1

Scroll down to see a description of the protein matches, an arbitrary score to rank them to each other, the query cover (the % of the input that was found to align with the produced sequence), the expectation value (this should be close to zero), and the % identical residues.

BLAST 20-mer sequence



Descriptions

Sequences producing significant alignments:

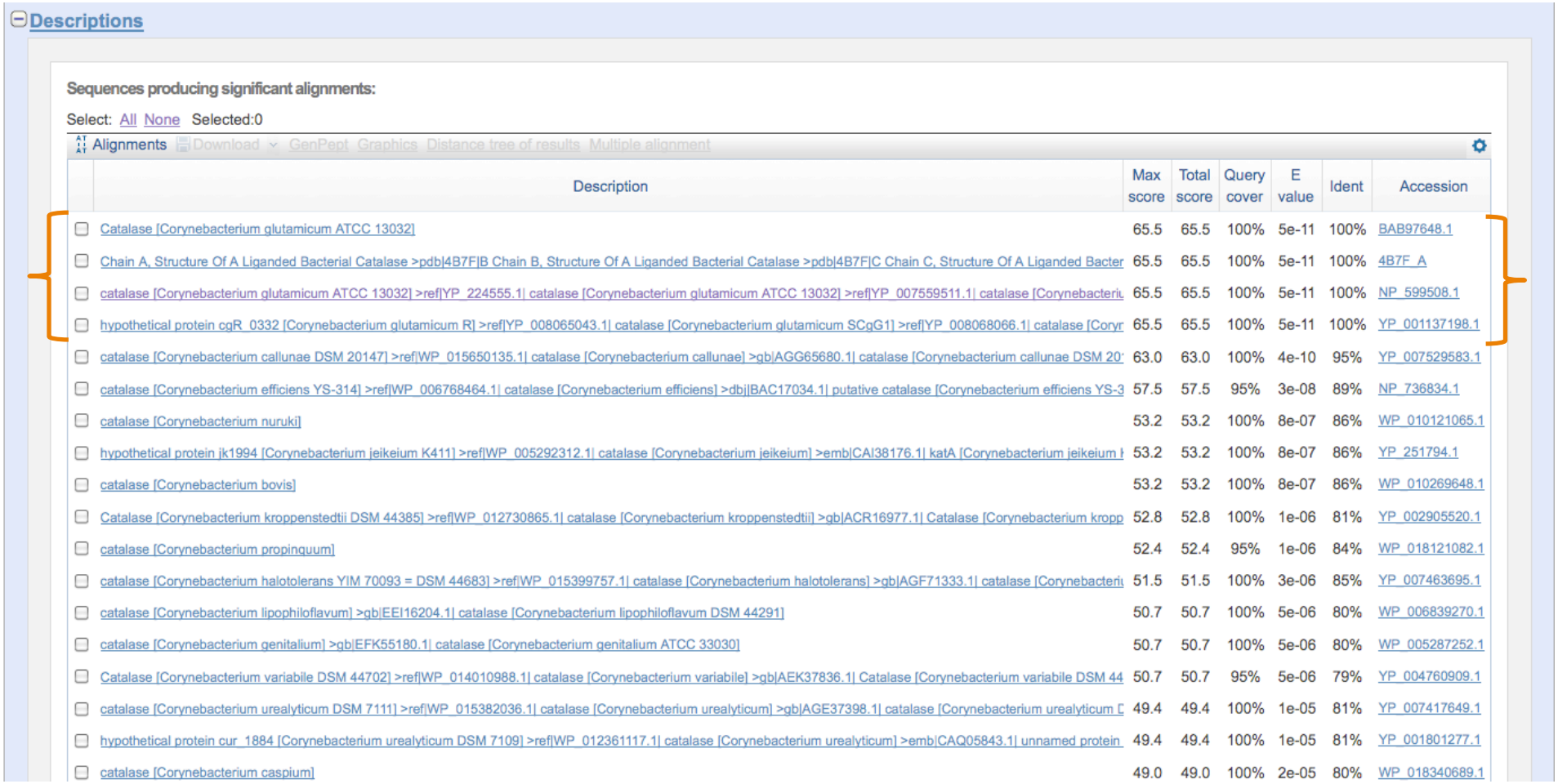
Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
Catalase [Corynebacterium glutamicum ATCC 13032]	65.5	65.5	100%	5e-11	100%	BAB97648.1
Chain A, Structure Of A Liganded Bacterial Catalase >pdb|4B7F|B Chain B, Structure Of A Liganded Bacterial Catalase >pdb|4B7F|C Chain C, Structure Of A Liganded Bacter	65.5	65.5	100%	5e-11	100%	4B7F_A
catalase [Corynebacterium glutamicum ATCC 13032] >ref|YP_224555.1| catalase [Corynebacterium glutamicum ATCC 13032] >ref|YP_007559511.1| catalase [Corynebacteriu	65.5	65.5	100%	5e-11	100%	NP_599508.1
hypothetical protein cgR_0332 [Corynebacterium glutamicum R] >ref|YP_008065043.1| catalase [Corynebacterium glutamicum SCgG1] >ref|YP_008068066.1| catalase [Coryr	65.5	65.5	100%	5e-11	100%	YP_001137198.1
catalase [Corynebacterium callunae DSM 20147] >ref|WP_015650135.1| catalase [Corynebacterium callunae] >gb|AGG65680.1| catalase [Corynebacterium callunae DSM 20	63.0	63.0	100%	4e-10	95%	YP_007529583.1
catalase [Corynebacterium efficiens YS-314] >ref|WP_006768464.1| catalase [Corynebacterium efficiens] >dbj|BAC17034.1| putative catalase [Corynebacterium efficiens YS-3	57.5	57.5	95%	3e-08	89%	NP_736834.1
catalase [Corynebacterium nuruki]	53.2	53.2	100%	8e-07	86%	WP_010121065.1
hypothetical protein jk1994 [Corynebacterium jeikeium K411] >ref|WP_005292312.1| catalase [Corynebacterium jeikeium] >emb|CAI38176.1| katA [Corynebacterium jeikeium K	53.2	53.2	100%	8e-07	86%	YP_251794.1
catalase [Corynebacterium bovis]	53.2	53.2	100%	8e-07	86%	WP_010269648.1
Catalase [Corynebacterium kroppenstedtii DSM 44385] >ref|WP_012730865.1| catalase [Corynebacterium kroppenstedtii] >gb|ACR16977.1| Catalase [Corynebacterium kropp	52.8	52.8	100%	1e-06	81%	YP_002905520.1
catalase [Corynebacterium propinquum]	52.4	52.4	95%	1e-06	84%	WP_018121082.1
catalase [Corynebacterium halotolerans YIM 70093 = DSM 44683] >ref|WP_015399757.1| catalase [Corynebacterium halotolerans] >gb|AGF71333.1| catalase [Corynebacteriu	51.5	51.5	100%	3e-06	85%	YP_007463695.1
catalase [Corynebacterium lipophiloflavum] >gb|EEI16204.1| catalase [Corynebacterium lipophiloflavum DSM 44291]	50.7	50.7	100%	5e-06	80%	WP_006839270.1
catalase [Corynebacterium genitalium] >gb|EFK55180.1| catalase [Corynebacterium genitalium ATCC 33030]	50.7	50.7	100%	5e-06	80%	WP_005287252.1
Catalase [Corynebacterium variabile DSM 44702] >ref|WP_014010988.1| catalase [Corynebacterium variabile] >gb|AEK37836.1| Catalase [Corynebacterium variabile DSM 44	50.7	50.7	95%	5e-06	79%	YP_004760909.1
catalase [Corynebacterium urealyticum DSM 7111] >ref|WP_015382036.1| catalase [Corynebacterium urealyticum] >gb|AGE37398.1| catalase [Corynebacterium urealyticum D	49.4	49.4	100%	1e-05	81%	YP_007417649.1
hypothetical protein cur_1884 [Corynebacterium urealyticum DSM 7109] >ref|WP_012361117.1| catalase [Corynebacterium urealyticum] >emb|CAQ05843.1| unnamed protein	49.4	49.4	100%	1e-05	81%	YP_001801277.1
catalase [Corynebacterium caspium]	49.0	49.0	100%	2e-05	80%	WP_018340689.1

You should pick a protein that has 100% query cover and 100% identity. The score and the E value don't matter much for this step. There may be many options that fit these categories (either very similar or identical proteins from multiple databases).

BLAST 20-mer sequence



Descriptions

Sequences producing significant alignments:

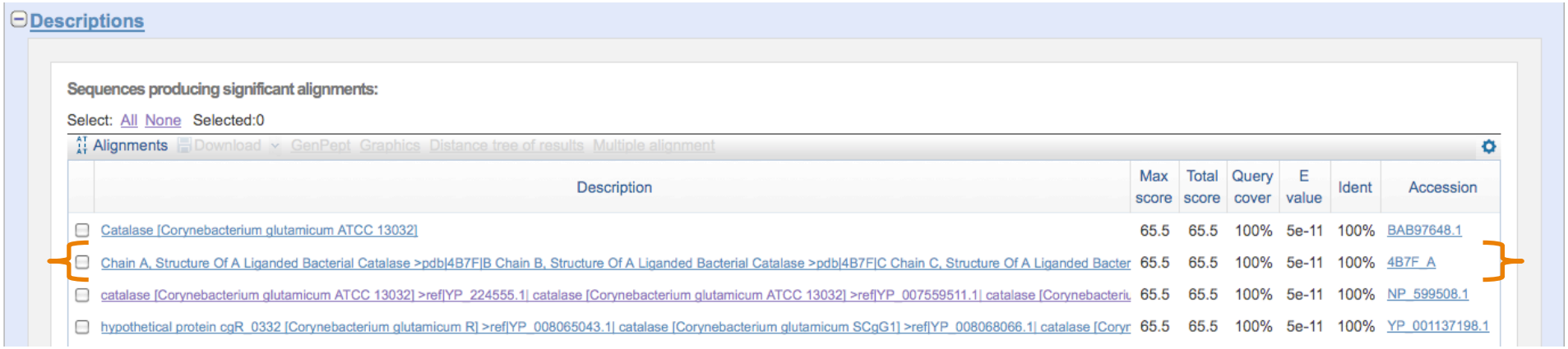
Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
Catalase [Corynebacterium glutamicum ATCC 13032]	65.5	65.5	100%	5e-11	100%	BAB97648.1
Chain A, Structure Of A Liganded Bacterial Catalase >pdb|4B7F|B Chain B, Structure Of A Liganded Bacterial Catalase >pdb|4B7F|C Chain C, Structure Of A Liganded Bacter	65.5	65.5	100%	5e-11	100%	4B7F_A
catalase [Corynebacterium glutamicum ATCC 13032] >ref|YP_224555.1| catalase [Corynebacterium glutamicum ATCC 13032] >ref|YP_007559511.1| catalase [Corynebacteriu	65.5	65.5	100%	5e-11	100%	NP_599508.1
hypothetical protein cgR_0332 [Corynebacterium glutamicum R] >ref|YP_008065043.1| catalase [Corynebacterium glutamicum SCgG1] >ref|YP_008068066.1| catalase [Coryr	65.5	65.5	100%	5e-11	100%	YP_001137198.1
catalase [Corynebacterium callunae DSM 20147] >ref|WP_015650135.1| catalase [Corynebacterium callunae] >gb|AGG65680.1| catalase [Corynebacterium callunae DSM 20	63.0	63.0	100%	4e-10	95%	YP_007529583.1
catalase [Corynebacterium efficiens YS-314] >ref|WP_006768464.1| catalase [Corynebacterium efficiens] >dbj|BAC17034.1| putative catalase [Corynebacterium efficiens YS-3	57.5	57.5	95%	3e-08	89%	NP_736834.1
catalase [Corynebacterium nuruki]	53.2	53.2	100%	8e-07	86%	WP_010121065.1
hypothetical protein jk1994 [Corynebacterium jeikeium K411] >ref|WP_005292312.1| catalase [Corynebacterium jeikeium] >emb|CAI38176.1| katA [Corynebacterium jeikeium K	53.2	53.2	100%	8e-07	86%	YP_251794.1
catalase [Corynebacterium bovis]	53.2	53.2	100%	8e-07	86%	WP_010269648.1
Catalase [Corynebacterium kroppenstedtii DSM 44385] >ref|WP_012730865.1| catalase [Corynebacterium kroppenstedtii] >gb|ACR16977.1| Catalase [Corynebacterium kropp	52.8	52.8	100%	1e-06	81%	YP_002905520.1
catalase [Corynebacterium propinquum]	52.4	52.4	95%	1e-06	84%	WP_018121082.1
catalase [Corynebacterium halotolerans YIM 70093 = DSM 44683] >ref|WP_015399757.1| catalase [Corynebacterium halotolerans] >gb|AGF71333.1| catalase [Corynebacteriu	51.5	51.5	100%	3e-06	85%	YP_007463695.1
catalase [Corynebacterium lipophiloflavum] >gb|EEI16204.1| catalase [Corynebacterium lipophiloflavum DSM 44291]	50.7	50.7	100%	5e-06	80%	WP_006839270.1
catalase [Corynebacterium genitalium] >gb|EFK55180.1| catalase [Corynebacterium genitalium ATCC 33030]	50.7	50.7	100%	5e-06	80%	WP_005287252.1
Catalase [Corynebacterium variabile DSM 44702] >ref|WP_014010988.1| catalase [Corynebacterium variabile] >gb|AEK37836.1| Catalase [Corynebacterium variabile DSM 44	50.7	50.7	95%	5e-06	79%	YP_004760909.1
catalase [Corynebacterium urealyticum DSM 7111] >ref|WP_015382036.1| catalase [Corynebacterium urealyticum] >gb|AGE37398.1| catalase [Corynebacterium urealyticum D	49.4	49.4	100%	1e-05	81%	YP_007417649.1
hypothetical protein cur_1884 [Corynebacterium urealyticum DSM 7109] >ref|WP_012361117.1| catalase [Corynebacterium urealyticum] >emb|CAQ05843.1| unnamed protein	49.4	49.4	100%	1e-05	81%	YP_001801277.1
catalase [Corynebacterium caspium]	49.0	49.0	100%	2e-05	80%	WP_018340689.1

Notice I have four options that have 100% query cover and 100% identity. They are probably identical sequences, as they come from the same organism, *Corynebacterium glutamicum*.

BLAST 20-mer sequence



Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

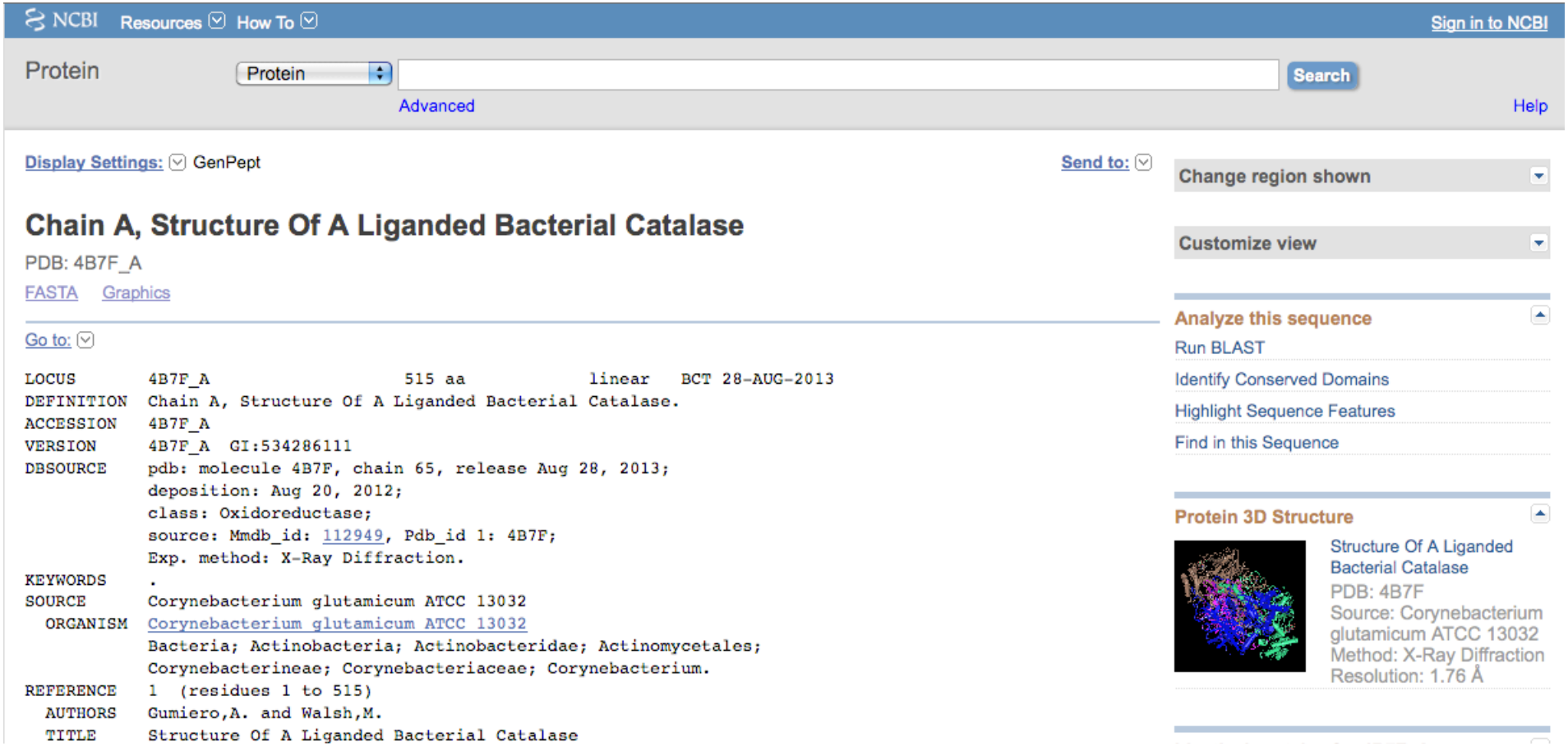
Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
Catalase [Corynebacterium glutamicum ATCC 13032]	65.5	65.5	100%	5e-11	100%	BAB97648.1
Chain A, Structure Of A Liganded Bacterial Catalase >pdb|4B7F|B Chain B, Structure Of A Liganded Bacterial Catalase >pdb|4B7F|C Chain C, Structure Of A Liganded Bacter	65.5	65.5	100%	5e-11	100%	4B7F_A
catalase [Corynebacterium glutamicum ATCC 13032] >ref|YP_224555.1| catalase [Corynebacterium glutamicum ATCC 13032] >ref|YP_007559511.1| catalase [Corynebacteriu	65.5	65.5	100%	5e-11	100%	NP_599508.1
hypothetical protein cgR_0332 [Corynebacterium glutamicum R] >ref|YP_008065043.1| catalase [Corynebacterium glutamicum SCgG1] >ref|YP_008068066.1| catalase [Coryr	65.5	65.5	100%	5e-11	100%	YP_001137198.1

Also notice that one reference has an associated 3D structure. References from this database have a similar accession number, so watch for alphanumeric codes like 4B7F_A

BEWARE!! There *should* be a file with a structure in your first BLAST search, but the sequence may not be identical to what is genomically coded. Watch out for partial or mutant proteins or a designation like E72A, which signifies residue #72, Glu, has been mutated to Ala.

Obtain your protein sequence



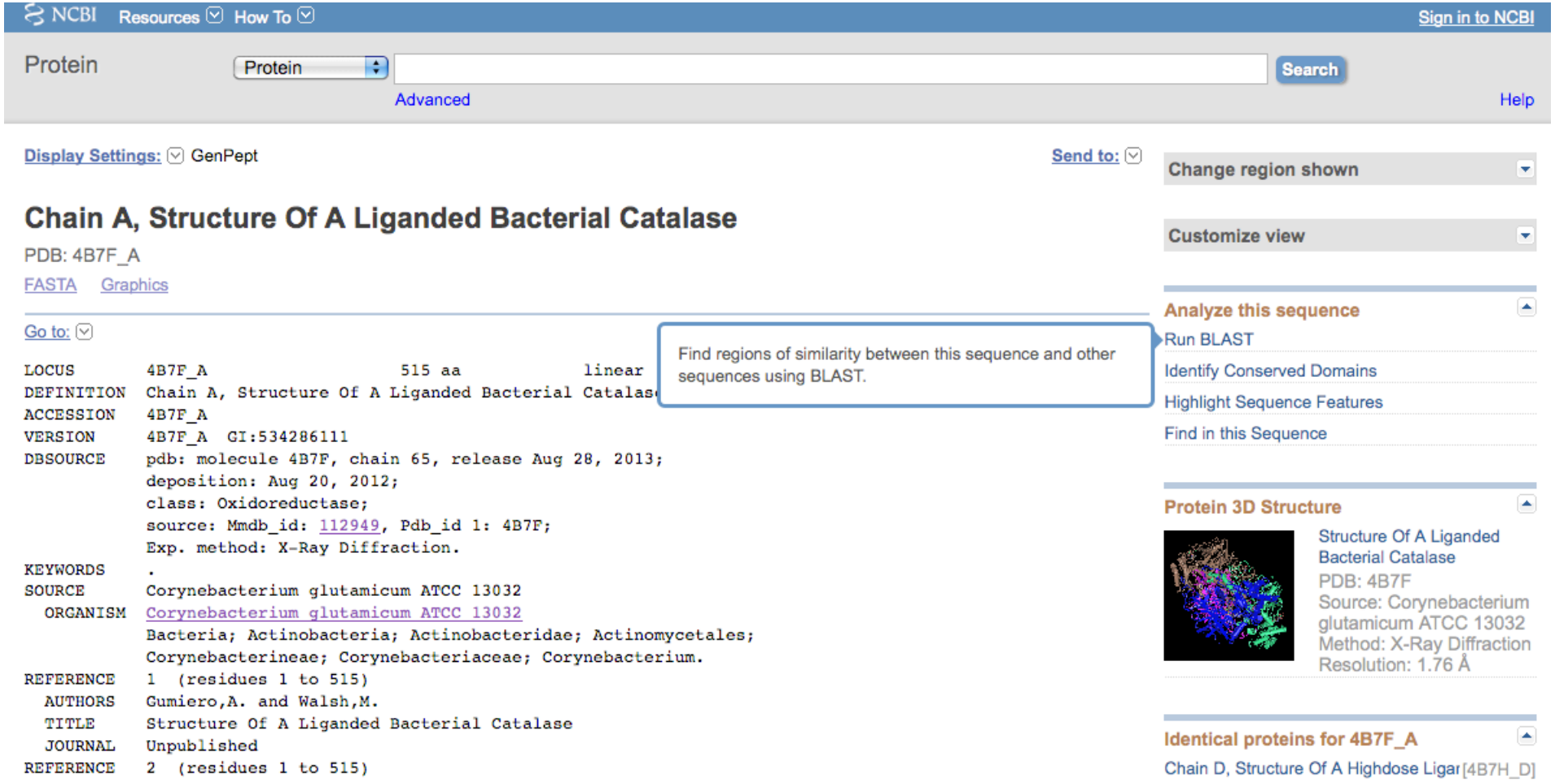
Clicking on the accession number will bring you to a webpage dedicated to everything associated with this protein in the database, such as the organism and any journal articles that use this protein.

Obtain your protein sequence

```
     Het             bond(352)
                     /heterogen="(HEM,1004 )"
     SecStr          454..461
                     /sec_str_type="helix"
                     /note="helix 8"
     SecStr          466..480
                     /sec_str_type="helix"
                     /note="helix 9"
     SecStr          488..498
                     /sec_str_type="helix"
                     /note="helix 10"
     SecStr          499..514
                     /sec_str_type="helix"
                     /note="helix 11"
ORIGIN
        1 seksaadqiv drgmrpklsg nttrhngapv psenisatag pqgpnvlndi hlieklahfn
       61 renvperiph akghgafgel hitedvseyt kadlfqpgkv tplavrfstv ageqgspdtw
      121 rdvhgfalrf yteegnydiv gnntptfflr dgmkfpdfih sqkrlnkngl rdadmqwdfw
      181 trapesahqv tylmgdrgtp ktsrhqdgfg shtfqwinae gkpvwvkyhf ktrqgwdcft
      241 daeaakvage nadyqredly naiengdfpi wdvkvqimpf edaenyrwnp fdltktwsqk
      301 dyplipvgyf ilnrnprnff aqieqialdp gnivpgvgls pdrmlqarif ayadqqryri
      361 ganyrdlpvn rpinevntys regsmqyifd aegepsyspn rydkgagyld ngtdsssnht
      421 sygqaddiyv npdphgtdlv raayvkhqdd ddfiqpgily revldegeke rladnisnam
      481 qgiseatepr vydywnnvde nlgarvkely lqkka
//
```

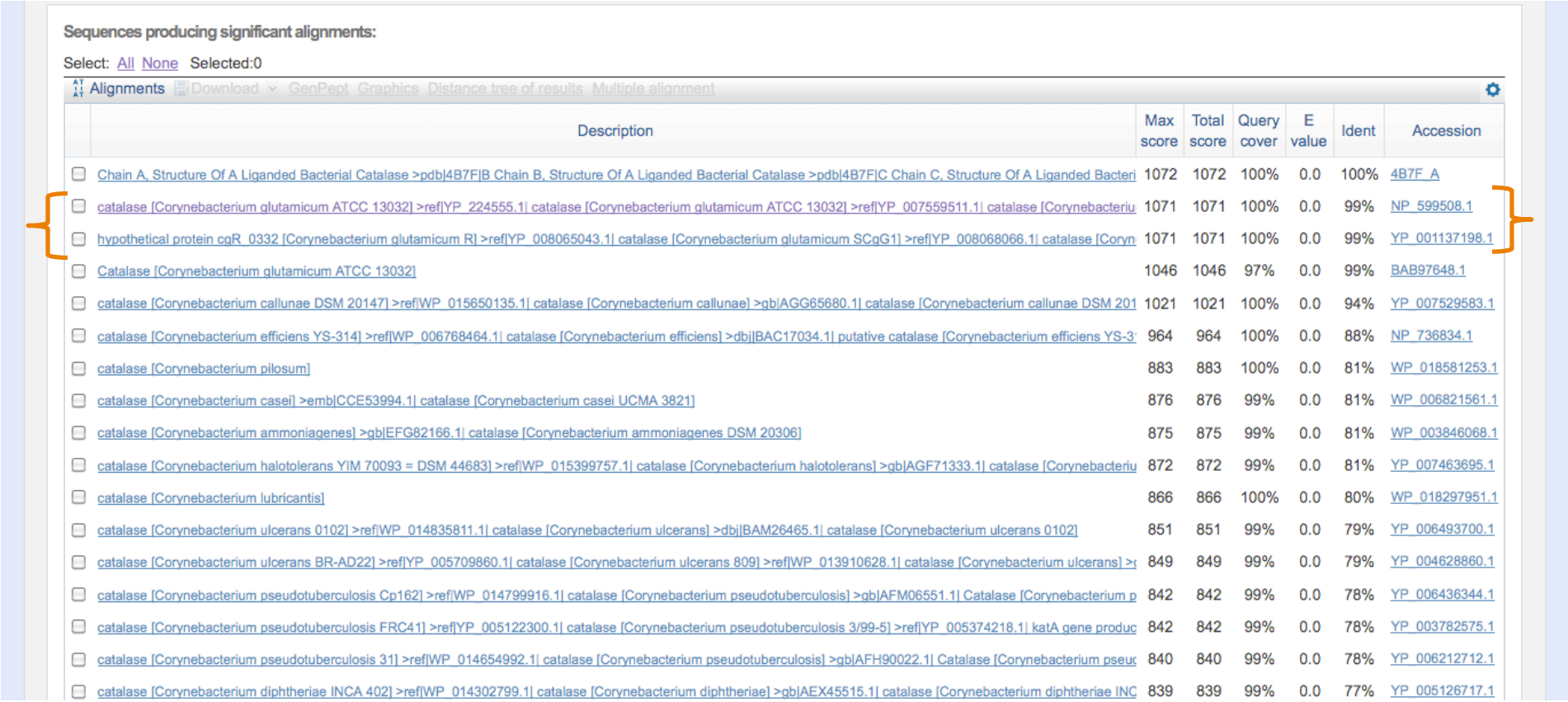
The bottom of this page lists the protein sequence. If you picked a structure accession number, notice that the N-terminal methionine is usually missing (and possibly other residues)

Obtain DNA seq. that codes for protein



Now we need to find the DNA sequence for the protein. You can directly BLAST this protein by clicking this link found at the top of the page.

Obtain DNA seq. that codes for protein



Sequences producing significant alignments:

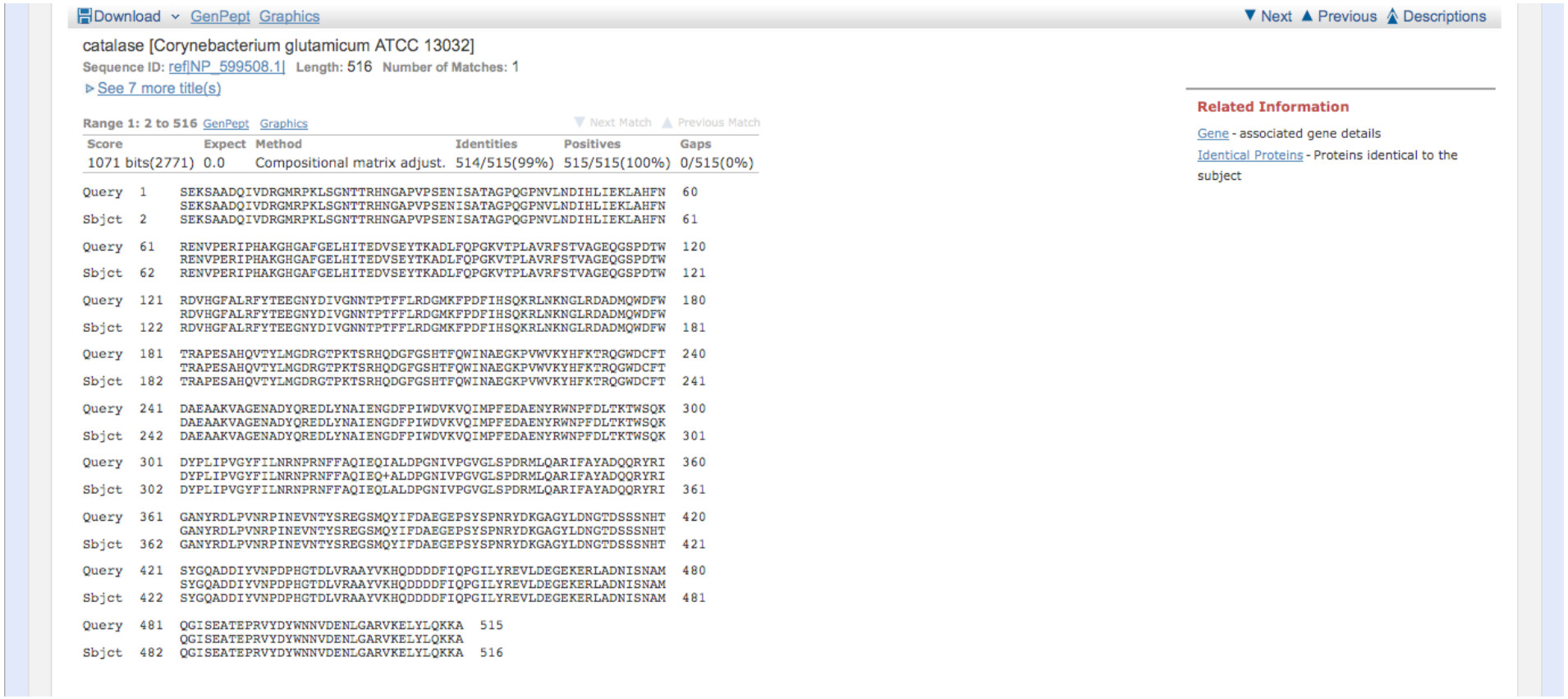
Select: All None Selected:0

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

Description	Max score	Total score	Query cover	E value	Ident	Accession
Chain A, Structure Of A Liganded Bacterial Catalase >pdb|4B7F|B Chain B, Structure Of A Liganded Bacterial Catalase >pdb|4B7F|C Chain C, Structure Of A Liganded Bacteri	1072	1072	100%	0.0	100%	4B7F_A
catalase [Corynebacterium glutamicum ATCC 13032] >ref|YP_224555.1| catalase [Corynebacterium glutamicum ATCC 13032] >ref|YP_007559511.1| catalase [Corynebacteriu	1071	1071	100%	0.0	99%	NP_599508.1
hypothetical protein cgR_0332 [Corynebacterium glutamicum R] >ref|YP_008065043.1| catalase [Corynebacterium glutamicum SCgG1] >ref|YP_008068066.1| catalase [Coryn	1071	1071	100%	0.0	99%	YP_001137198.1
Catalase [Corynebacterium glutamicum ATCC 13032]	1046	1046	97%	0.0	99%	BAB97648.1
catalase [Corynebacterium callunae DSM 20147] >ref|WP_015650135.1| catalase [Corynebacterium callunae] >gb|AGG65680.1| catalase [Corynebacterium callunae DSM 201	1021	1021	100%	0.0	94%	YP_007529583.1
catalase [Corynebacterium efficiens YS-314] >ref|WP_006768464.1| catalase [Corynebacterium efficiens] >dbj|BAC17034.1| putative catalase [Corynebacterium efficiens YS-3	964	964	100%	0.0	88%	NP_736834.1
catalase [Corynebacterium pilosum]	883	883	100%	0.0	81%	WP_018581253.1
catalase [Corynebacterium casei] >emb|CCE53994.1| catalase [Corynebacterium casei UCMA 3821]	876	876	99%	0.0	81%	WP_006821561.1
catalase [Corynebacterium ammoniagenes] >gb|EFG82166.1| catalase [Corynebacterium ammoniagenes DSM 20306]	875	875	99%	0.0	81%	WP_003846068.1
catalase [Corynebacterium halotolerans YIM 70093 = DSM 44683] >ref|WP_015399757.1| catalase [Corynebacterium halotolerans] >gb|AGF71333.1| catalase [Corynebacteriu	872	872	99%	0.0	81%	YP_007463695.1
catalase [Corynebacterium lubricantis]	866	866	100%	0.0	80%	WP_018297951.1
catalase [Corynebacterium ulcerans 0102] >ref|WP_014835811.1| catalase [Corynebacterium ulcerans] >dbj|BAM26465.1| catalase [Corynebacterium ulcerans 0102]	851	851	99%	0.0	79%	YP_006493700.1
catalase [Corynebacterium ulcerans BR-AD22] >ref|YP_005709860.1| catalase [Corynebacterium ulcerans 809] >ref|WP_013910628.1| catalase [Corynebacterium ulcerans] >(	849	849	99%	0.0	79%	YP_004628860.1
catalase [Corynebacterium pseudotuberculosis Cp162] >ref|WP_014799916.1| catalase [Corynebacterium pseudotuberculosis] >gb|AFM06551.1| Catalase [Corynebacterium p	842	842	99%	0.0	78%	YP_006436344.1
catalase [Corynebacterium pseudotuberculosis FRC41] >ref|YP_005122300.1| catalase [Corynebacterium pseudotuberculosis 3/99-5] >ref|YP_005374218.1| katA gene produc	842	842	99%	0.0	78%	YP_003782575.1
catalase [Corynebacterium pseudotuberculosis 31] >ref|WP_014654992.1| catalase [Corynebacterium pseudotuberculosis] >gb|AFH90022.1| Catalase [Corynebacterium pseuc	840	840	99%	0.0	78%	YP_006212712.1
catalase [Corynebacterium diphtheriae INCA 402] >ref|WP_014302799.1| catalase [Corynebacterium diphtheriae] >gb|AEX45515.1| catalase [Corynebacterium diphtheriae INC	839	839	99%	0.0	77%	YP_005126717.1

Here you have all the proteins that are nearly identical to yours! I BLASTed a protein with an associated structure, so notice that the next few matches are the identical protein without the N-terminal methionine, so it is 99% identical over 100% of the query.

Obtain DNA seq. that codes for protein



Download GenPept Graphics ▼ Next ▲ Previous Descriptions

catalase [Corynebacterium glutamicum ATCC 13032]

Sequence ID: ref|NP_599508.1| Length: 516 Number of Matches: 1

See 7 more title(s)

Range 1: 2 to 516 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
1071 bits(2771)	0.0	Compositional matrix adjust.	514/515(99%)	515/515(100%)	0/515(0%)

```
Query  1    SEKSAADQIVDRGMRPKLSGNTTRHNGAPVPSENISATAGPQGPNVLNDIHLIEKLAHFN  60
            SEKSAADQIVDRGMRPKLSGNTTRHNGAPVPSENISATAGPQGPNVLNDIHLIEKLAHFN
Sbjct  2    SEKSAADQIVDRGMRPKLSGNTTRHNGAPVPSENISATAGPQGPNVLNDIHLIEKLAHFN  61

Query  61   RENVPERIPHAKGHGAFGELHITEDVSEYTKADLFQPGKVTPLAVRFSTVAGEQGSPDTW  120
            RENVPERIPHAKGHGAFGELHITEDVSEYTKADLFQPGKVTPLAVRFSTVAGEQGSPDTW
Sbjct  62   RENVPERIPHAKGHGAFGELHITEDVSEYTKADLFQPGKVTPLAVRFSTVAGEQGSPDTW  121

Query  121  RDVHGFALRFYTEEGNYDIVGNNTPTFFLRDGMKFPDFIHSQKRLNKNGLRDADMQWDFW  180
            RDVHGFALRFYTEEGNYDIVGNNTPTFFLRDGMKFPDFIHSQKRLNKNGLRDADMQWDFW
Sbjct  122  RDVHGFALRFYTEEGNYDIVGNNTPTFFLRDGMKFPDFIHSQKRLNKNGLRDADMQWDFW  181

Query  181  TRAPESAHQVTYLMGDRGTPKTSRHQDGFGSHTFQWINAEGKPVWVKYHFKTRQGWDCFT  240
            TRAPESAHQVTYLMGDRGTPKTSRHQDGFGSHTFQWINAEGKPVWVKYHFKTRQGWDCFT
Sbjct  182  TRAPESAHQVTYLMGDRGTPKTSRHQDGFGSHTFQWINAEGKPVWVKYHFKTRQGWDCFT  241

Query  241  DAEAAKVAGENADYQREDLYNAIENGDFPIWDVKVQIMPFEDAENYRWNPFDLTKTWSQK  300
            DAEAAKVAGENADYQREDLYNAIENGDFPIWDVKVQIMPFEDAENYRWNPFDLTKTWSQK
Sbjct  242  DAEAAKVAGENADYQREDLYNAIENGDFPIWDVKVQIMPFEDAENYRWNPFDLTKTWSQK  301

Query  301  DYPLIPVGYFILNRNPRNFFAQIEQIALDPGNIVPGVGLSPDRMLQARIFAYADQQRYRI  360
            DYPLIPVGYFILNRNPRNFFAQIEQ+ALDPGNIVPGVGLSPDRMLQARIFAYADQQRYRI
Sbjct  302  DYPLIPVGYFILNRNPRNFFAQIEQLALDPGNIVPGVGLSPDRMLQARIFAYADQQRYRI  361

Query  361  GANYRDLPVNRPINEVNTYSREGSMQYIFDAEGEPSYSPNRYDKGAGYLDNGTDSSSNHT  420
            GANYRDLPVNRPINEVNTYSREGSMQYIFDAEGEPSYSPNRYDKGAGYLDNGTDSSSNHT
Sbjct  362  GANYRDLPVNRPINEVNTYSREGSMQYIFDAEGEPSYSPNRYDKGAGYLDNGTDSSSNHT  421

Query  421  SYGQADDIYVNPDPHGTDLVRAAYVKHQDDDDFIQPGILYREVLDEGEKERLADNISNAM  480
            SYGQADDIYVNPDPHGTDLVRAAYVKHQDDDDFIQPGILYREVLDEGEKERLADNISNAM
Sbjct  422  SYGQADDIYVNPDPHGTDLVRAAYVKHQDDDDFIQPGILYREVLDEGEKERLADNISNAM  481

Query  481  QGISEATEPRVYDYWNNVDENLGARVKELYLQKKA  515
            QGISEATEPRVYDYWNNVDENLGARVKELYLQKKA
Sbjct  482  QGISEATEPRVYDYWNNVDENLGARVKELYLQKKA  516
```

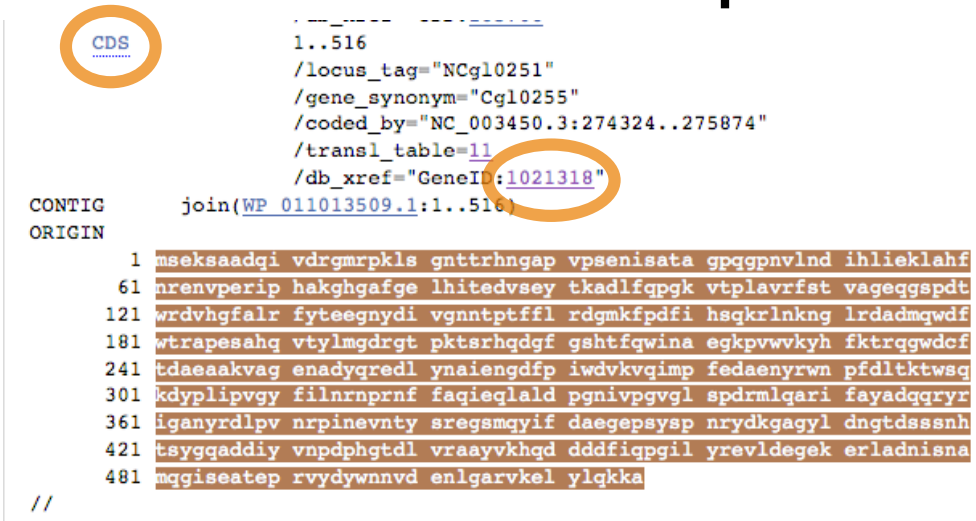
Related Information

Gene - associated gene details

Identical Proteins - Proteins identical to the subject

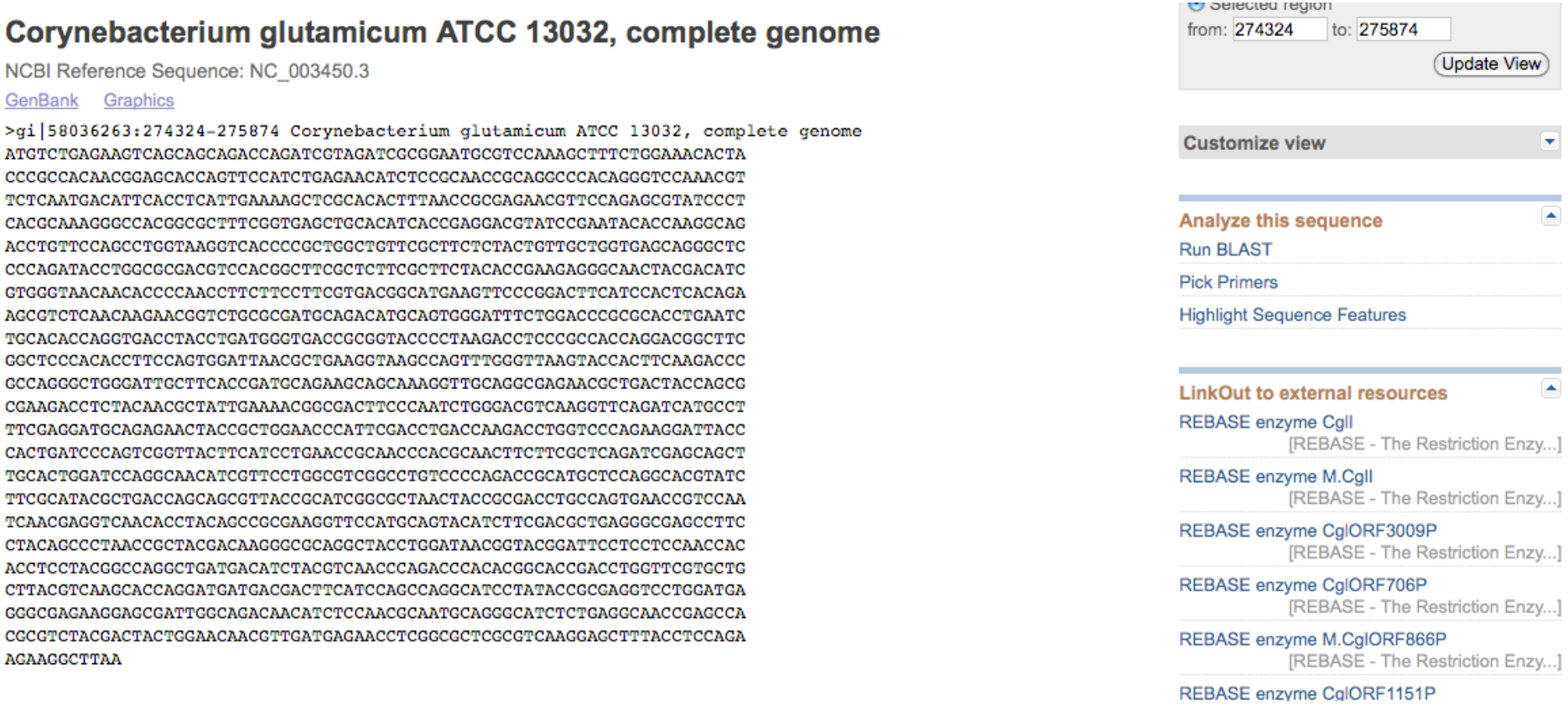
Obtaining the DNA sequence can be tricky, so it is a good idea to have a few entries to look through. Be sure that this sequence that you pick is identical to your 20-mer peptide sequence (scroll down to see pairwise alignments).

Obtain DNA seq. that codes for protein



If you click on the accession number, it will bring you to the new protein page and highlight the protein sequence. Clicking on CDS (**Co**ding **S**equence) or the GeneID to get a tool bar to display on the bottom of your screen. Click on FASTA.

Obtain DNA seq. that codes for protein



Corynebacterium glutamicum ATCC 13032, complete genome

NCBI Reference Sequence: NC_003450.3

GenBank Graphics

```
>gi|58036263:274324-275874 Corynebacterium glutamicum ATCC 13032, complete genome
ATGTCTGAGAAGTCAGCAGCAGACCAGATCGTAGATCGCGGAATGCGTCCAAAGCTTTCTGGAAACACTA
CCCGCCACAACGGAGCACCAGTTCCATCTGAGAACATCTCCGCAACCGCAGGCCCACAGGGTCCAAACGT
TCTCAATGACATTCACCTCATTGAAAAGCTCGCACACTTTAACCGCGAGAACGTTCCAGAGCGTATCCCT
CACGCAAAGGGCCACGGCGCTTTCGGTGAGCTGCACATCACCGAGGACGTATCCGAATACACCAAGGCAG
ACCTGTTCCAGCCTGGTAAGGTCACCCCGCTGGCTGTTCGCTTCTCTACTGTTGCTGGTGAGCAGGGCTC
CCCAGATACCTGGCGCGACGTCCACGGCTTCGCTCTTCGCTTCTACACCGAAGAGGGCAACTACGACATC
GTGGGTAACAACACCCCAACCTTCTTCCTTCGTGACGGCATGAAGTTCCCGGACTTCATCCACTCACAGA
AGCGTCTCAACAAGAACGGTCTGCGCGATGCAGACATGCAGTGGGATTTCTGGACCCGCGCACCTGAATC
TGCACACCAGGTGACCTACCTGATGGGTGACCGCGGTACCCCTAAGACCTCCCGCCACCAGGACGGCTTC
GGCTCCCACACCTTCCAGTGGATTAACGCTGAAGGTAAGCCAGTTTGGGTTAAGTACCACTTCAAGACCC
GCCAGGGCTGGGATTGCTTCACCGATGCAGAAGCAGCAAAGGTTGCAGGCGAGAACGCTGACTACCAGCG
CGAAGACCTCTACAACGCTATTGAAAACGGCGACTTCCCAATCTGGGACGTCAAGGTTCAGATCATGCCT
TTCGAGGATGCAGAGAACTACCGCTGGAACCCATTCGACCTGACCAAGACCTGGTCCCAGAAGGATTACC
CACTGATCCCAGTCGGTTACTTCATCCTGAACCGCAACCCACGCAACTTCTTCGCTCAGATCGAGCAGCT
TGCACTGGATCCAGGCAACATCGTTCCTGGCGTCGGCCTGTCCCCAGACCGCATGCTCCAGGCACGTATC
TTCGCATACGCTGACCAGCAGCGTTACCGCATCGGCGCTAACTACCGCGACCTGCCAGTGAACCGTCCAA
TCAACGAGGTCAACACCTACAGCCGCGAAGGTTCCATGCAGTACATCTTCGACGCTGAGGGCGAGCCTTC
CTACAGCCCTAACCGCTACGACAAGGGCGCAGGCTACCTGGATAACGGTACGGATTCCTCCTCCAACCAC
ACCTCCTACGGCCAGGCTGATGACATCTACGTCAACCCAGACCCACACGGCACCGACCTGGTTCGTGCTG
CTTACGTCAAGCACCAGGATGATGACGACTTCATCCAGCCAGGCATCCTATACCGCGAGGTCCTGGATGA
GGGCGAGAAGGAGCGATTGGCAGACAACATCTCCAACGCAATGCAGGGCATCTCTGAGGCAACCGAGCCA
CGCGTCTACGACTACTGGAACAACGTTGATGAGAACCTCGGCGCTCGCGTCAAGGAGCTTTACCTCCAGA
AGAAGGCTTAA
```

Selected region
from: 274324 to: 275874
Update View

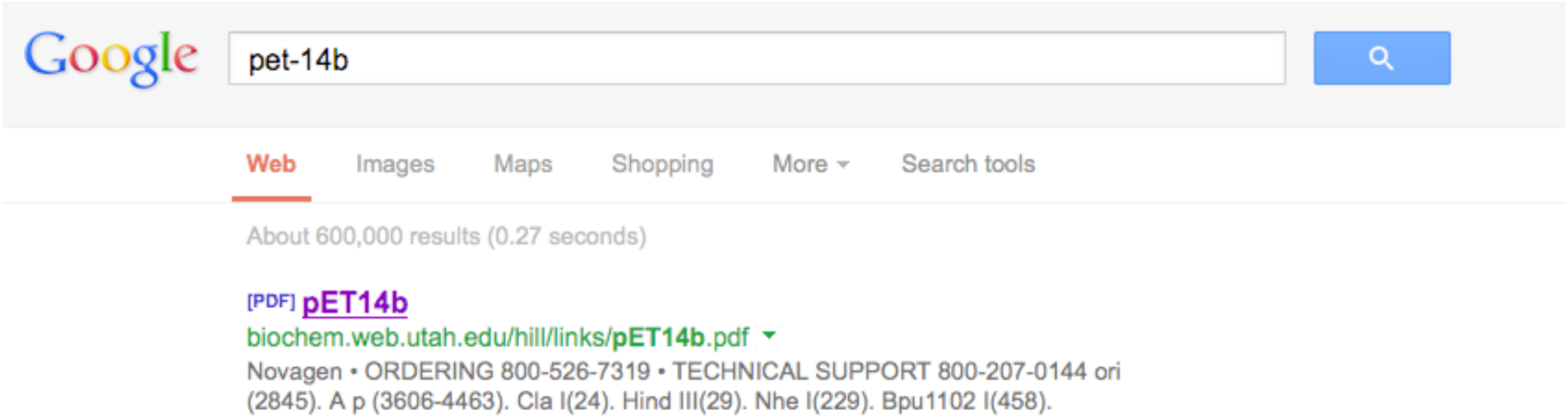
Customize view

Analyze this sequence
Run BLAST
Pick Primers
Highlight Sequence Features

LinkOut to external resources
REBASE enzyme CglI [REBASE - The Restriction Enzy...]
REBASE enzyme M.CglI [REBASE - The Restriction Enzy...]
REBASE enzyme CglORF3009P [REBASE - The Restriction Enzy...]
REBASE enzyme CglORF706P [REBASE - The Restriction Enzy...]
REBASE enzyme M.CglORF866P [REBASE - The Restriction Enzy...]
REBASE enzyme CglORF1151P

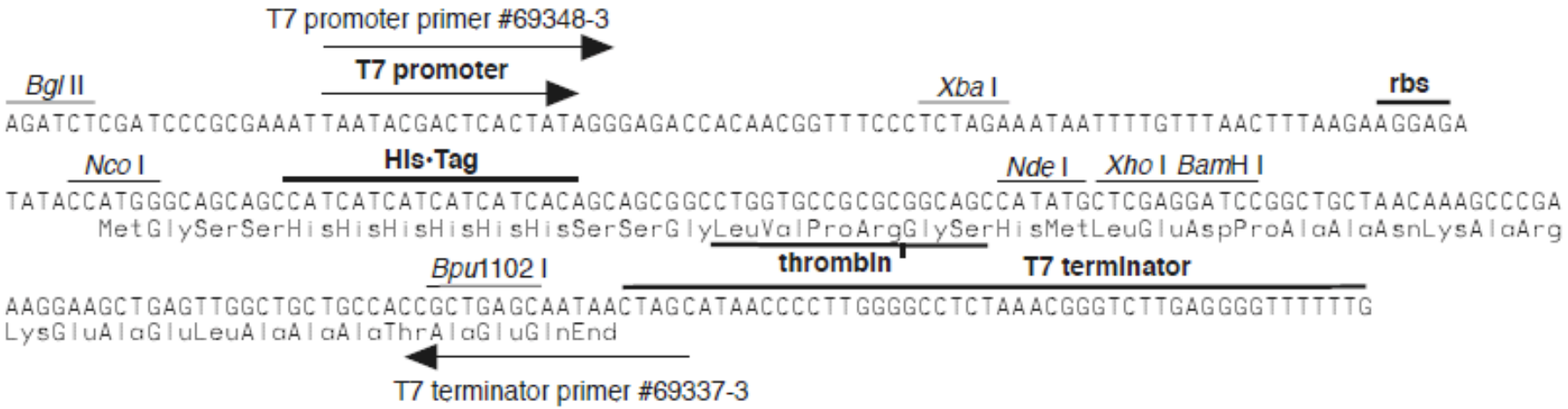
Here is the DNA sequence! Be sure that is begins with a start codon (ATG) and ends with a stop codon (TAA,TAG, or TGA)

Design primers to clone gene into vector



Now that we have the DNA sequence, let's find a vector suitable for protein expression. pET-14b is a very simple vector that allows for a His-tagged protein.

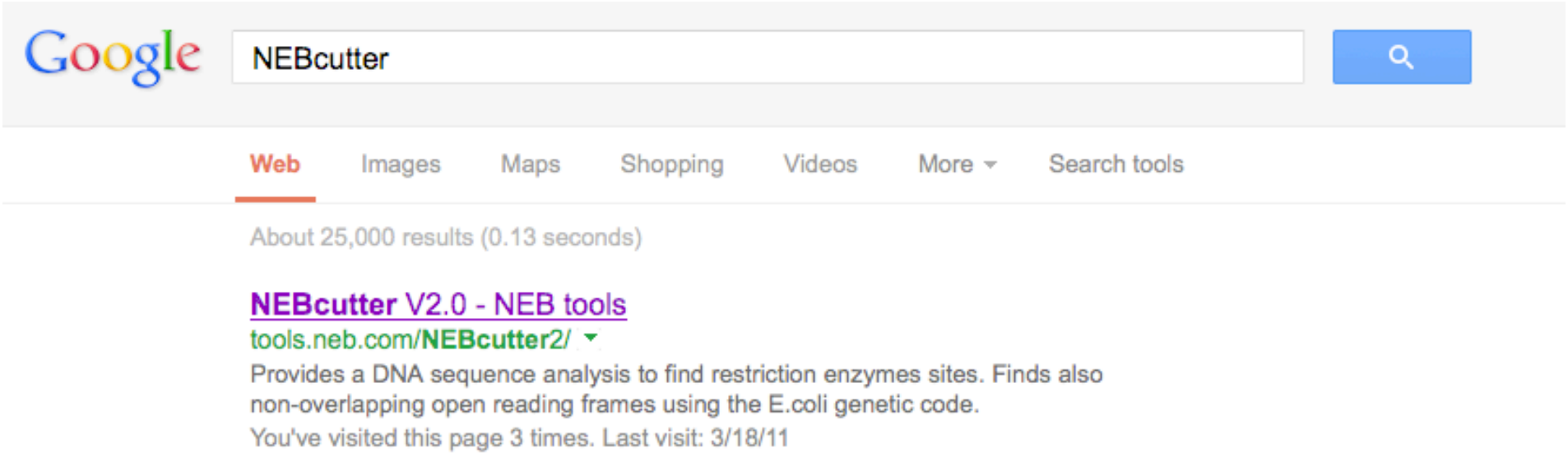
Design primers to clone gene into vector



pET-14b cloning/expression region

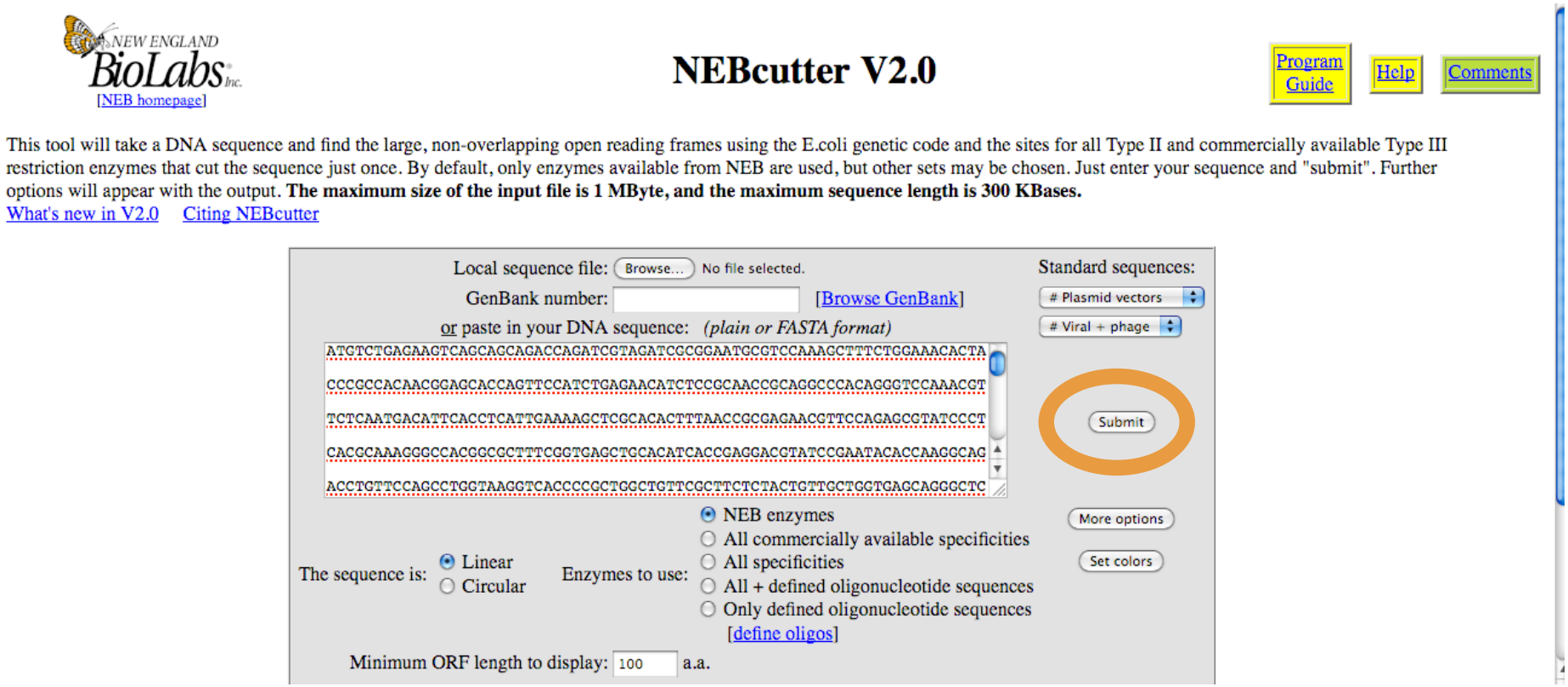
The link will open up a .pdf file, and above is a part of this file that shows the cloning/expression sequences of this vector. We want our protein to have a His-Tag, so we have to use two restriction sites after the His-Tag. This leaves NdeI, XhoI, and BamHI.

Design primers to clone gene into vector



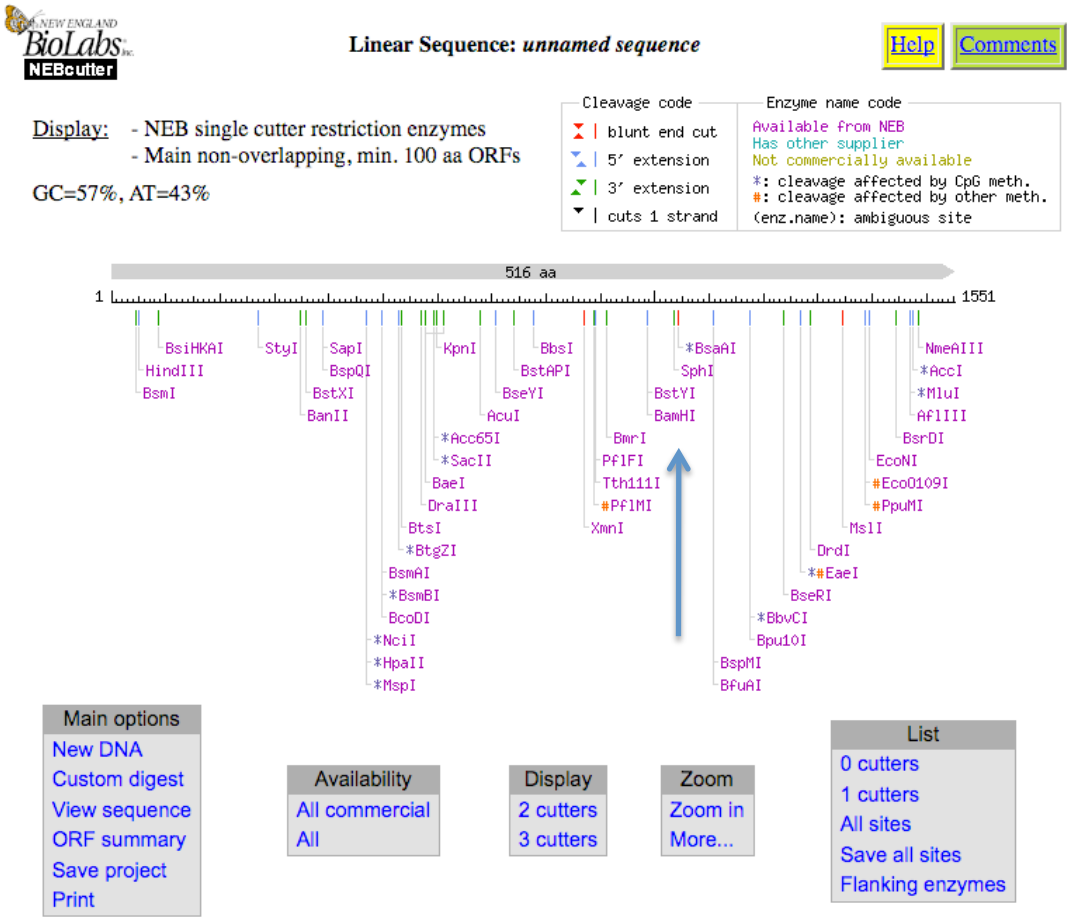
We are going to use restriction enzymes to clone our gene into pET-14b, so we have to be sure that these restriction enzymes won't cut our gene.

Design primers to clone gene into vector



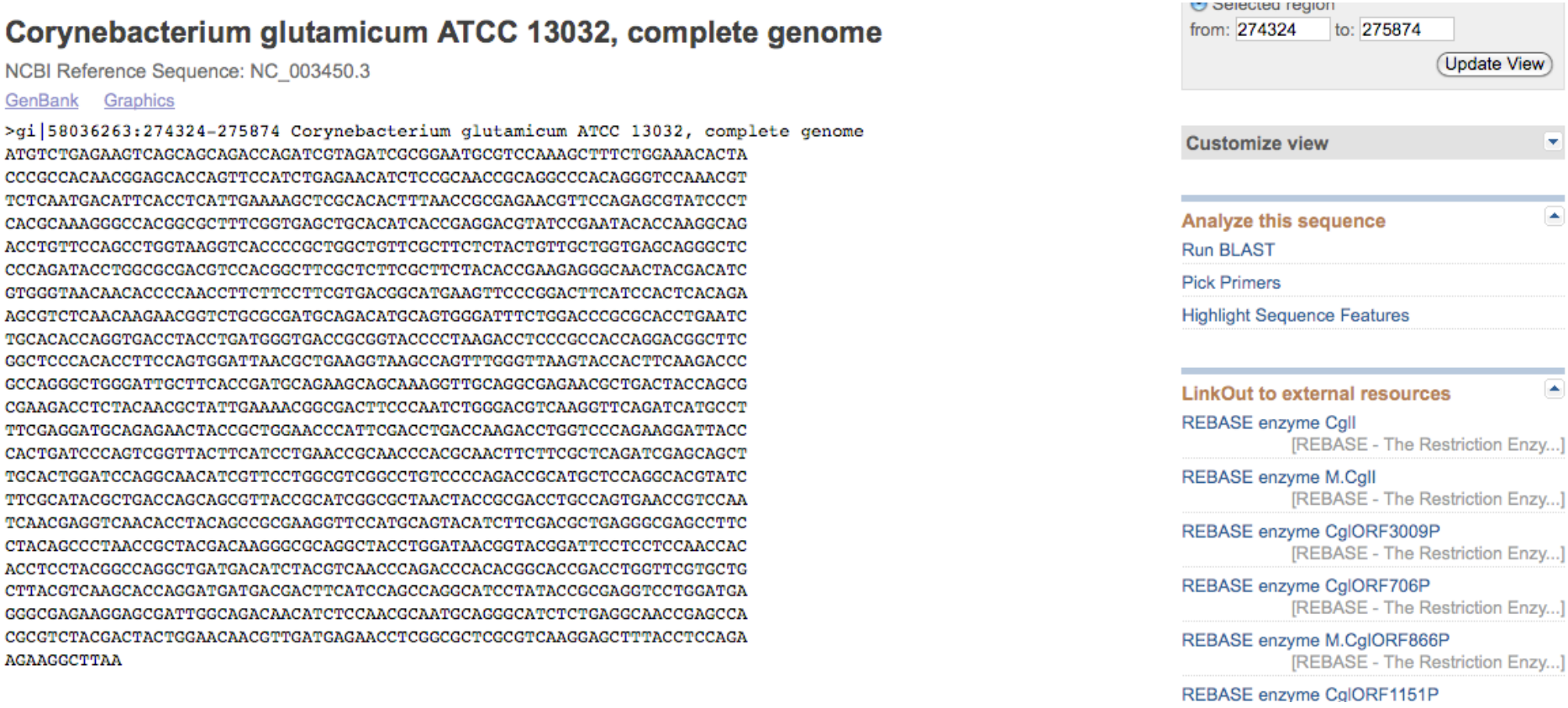
NEBcutter will analyze your gene sequence to find restrictions sites in your gene. Paste your sequence in the submission box and hit ‘Submit.’

Design primers to clone gene into vector



NdeI, XhoI, or BamHI anywhere? Don't use it. If you look carefully you see BamHI will cut this gene. NEBcutter will also list the enzymes that will and won't cut the gene. Do you have two sites in your gene? Time for a new vector. Try any from the pGEX or pMAL series, or pET21a.

Design primers to clone gene into vector



Corynebacterium glutamicum ATCC 13032, complete genome

NCBI Reference Sequence: NC_003450.3

GenBank Graphics

```
>gi|58036263:274324-275874 Corynebacterium glutamicum ATCC 13032, complete genome
ATGTCTGAGAAGTCAGCAGCAGACCAGATCGTAGATCGCGGAATGCGTCCAAAGCTTTCTGGAAACACTA
CCCGCCACAACGGAGCACCAGTTCCATCTGAGAACATCTCCGCAACCGCAGGCCCACAGGGTCCAAACGT
TCTCAATGACATTCACCTCATTGAAAAGCTCGCACACTTTAACCGCGAGAACGTTCCAGAGCGTATCCCT
CACGCAAAGGGCCACGGCGCTTTCGGTGAGCTGCACATCACCGAGGACGTATCCGAATACACCAAGGCAG
ACCTGTTCCAGCCTGGTAAGGTCACCCCGCTGGCTGTTCGCTTCTCTACTGTTGCTGGTGAGCAGGGCTC
CCCAGATACCTGGCGCGACGTCCACGGCTTCGCTCTTCGCTTCTACACCGAAGAGGGCAACTACGACATC
GTGGGTAACAACACCCCAACCTTCTTCCTTCGTGACGGCATGAAGTTCCCGGACTTCATCCACTCACAGA
AGCGTCTCAACAAGAACGGTCTGCGCGATGCAGACATGCAGTGGGATTTCTGGACCCGCGCACCTGAATC
TGCACACCAGGTGACCTACCTGATGGGTGACCGCGGTACCCCTAAGACCTCCCGCCACCAGGACGGCTTC
GGCTCCCACACCTTCCAGTGGATTAACGCTGAAGGTAAGCCAGTTTGGGTTAAGTACCACTTCAAGACCC
GCCAGGGCTGGGATTGCTTCACCGATGCAGAAGCAGCAAAGGTTGCAGGCGAGAACGCTGACTACCAGCG
CGAAGACCTCTACAACGCTATTGAAAACGGCGACTTCCCAATCTGGGACGTCAAGGTTCAGATCATGCCT
TTCGAGGATGCAGAGAACTACCGCTGGAACCCATTCGACCTGACCAAGACCTGGTCCCAGAAGGATTACC
CACTGATCCCAGTCGGTTACTTCATCCTGAACCGCAACCCACGCAACTTCTTCGCTCAGATCGAGCAGCT
TGCACTGGATCCAGGCAACATCGTTCCTGGCGTCGGCCTGTCCCCAGACCGCATGCTCCAGGCACGTATC
TTCGCATACGCTGACCAGCAGCGTTACCGCATCGGCGCTAACTACCGCGACCTGCCAGTGAACCGTCCAA
TCAACGAGGTCAACACCTACAGCCGCGAAGGTTCCATGCAGTACATCTTCGACGCTGAGGGCGAGCCTTC
CTACAGCCCTAACCGCTACGACAAGGGCGCAGGCTACCTGGATAACGGTACGGATTCCTCCTCCAACCAC
ACCTCCTACGGCCAGGCTGATGACATCTACGTCAACCCAGACCCACACGGCACCGACCTGGTTCGTGCTG
CTTACGTCAAGCACCAGGATGATGACGACTTCATCCAGCCAGGCATCCTATACCGCGAGGTCCTGGATGA
GGGCGAGAAGGAGCGATTGGCAGACAACATCTCCAACGCAATGCAGGGCATCTCTGAGGCAACCGAGCCA
CGCGTCTACGACTACTGGAACAACGTTGATGAGAACCTCGGCGCTCGCGTCAAGGAGCTTTACCTCCAGA
AGAAGGCTTAA
```

Selected region
from: 274324 to: 275874
Update View

Customize view

Analyze this sequence
Run BLAST
Pick Primers
Highlight Sequence Features

LinkOut to external resources
REBASE enzyme CglI [REBASE - The Restriction Enzy...]
REBASE enzyme M.CglI [REBASE - The Restriction Enzy...]
REBASE enzyme CglORF3009P [REBASE - The Restriction Enzy...]
REBASE enzyme CglORF706P [REBASE - The Restriction Enzy...]
REBASE enzyme M.CglORF866P [REBASE - The Restriction Enzy...]
REBASE enzyme CglORF1151P

My goal now is to design primers that are specific for this gene sequence, and to contain the NdeI and XhoI restriction sites so the gene can be cloned into the pET-14b vector.

Design primers to clone gene into vector

```
ATGTCTGAGAAGTCAGCAGCAGACCAGATCGTAGATCGCGGAATGCGTCCAAAGCTTTCTGGAAACACTA
CCCGCCACAACGGAGCACCAGTTCCATCTGAGAACATCTCCGCAACCGCAGGCCCACAGGGTCCAAACGT
TCTCAATGACATTCACCTCATTGAAAAGCTCGCACACTTTAACCGCGAGAACGTTCCAGAGCGTATCCCT
CACGCAAAGGGCCACGGCGCTTTCGGTGAGCTGCACATCACCGAGGACGTATCCGAATACACCAAGGCAG
ACCTGTTCCAGCCTGGTAAGGTCACCCCGCTGGCTGTTCGCTTCTCTACTGTTGCTGGTGAGCAGGGCTC
CCCAGATACCTGGCGCGACGTCCACGGCTTCGCTCTTCGCTTCTACACCGAAGAGGGCAACTACGACATC
GTGGGTAACAACACCCCAACCTTCTTCCTTCGTGACGGCATGAAGTTCCCGGACTTCATCCACTCACAGA
AGCGTCTCAACAAGAACGGTCTGCGCGATGCAGACATGCAGTGGGATTTCTGGACCCGCGCACCTGAATC
TGCACACCAGGTGACCTACCTGATGGGTGACCGCGGTACCCCTAAGACCTCCCGCCACCAGGACGGCTTC
GGCTCCCACACCTTCCAGTGGATTAACGCTGAAGGTAAGCCAGTTTGGGTTAAGTACCACTTCAAGACCC
GCCAGGGCTGGGATTGCTTCACCGATGCAGAAGCAGCAAAGGTTGCAGGCGAGAACGCTGACTACCAGCG
CGAAGACCTCTACAACGCTATTGAAAACGGCGACTTCCCAATCTGGGACGTCAAGGTTCAGATCATGCCT
TTCGAGGATGCAGAGAACTACCGCTGGAACCCATTCGACCTGACCAAGACCTGGTCCCAGAAGGATTACC
CACTGATCCCAGTCGGTTACTTCATCCTGAACCGCAACCCACGCAACTTCTTCGCTCAGATCGAGCAGCT
TGCACTGGATCCAGGCAACATCGTTCCTGGCGTCGGCCTGTCCCCAGACCGCATGCTCCAGGCACGTATC
TTCGCATACGCTGACCAGCAGCGTTACCGCATCGGCGCTAACTACCGCGACCTGCCAGTGAACCGTCCAA
TCAACGAGGTCAACACCTACAGCCGCGAAGGTTCCATGCAGTACATCTTCGACGCTGAGGGCGAGCCTTC
CTACAGCCCTAACCGCTACGACAAGGGCGCAGGCTACCTGGATAACGGTACGGATTCCTCCTCCAACCAC
ACCTCCTACGGCCAGGCTGATGACATCTACGTCAACCCAGACCCACACGGCACCGACCTGGTTCGTGCTG
CTTACGTCAAGCACCAGGATGATGACGACTTCATCCAGCCAGGCATCCTATACCGCGAGGTCCTGGATGA
GGGCGAGAAGGAGCGATTGGCAGACAACATCTCCAACGCAATGCAGGGCATCTCTGAGGCAACCGAGCCA
CGCGTCTACGACTACTGGAACAACGTTGATGAGAACCTCGGCGCTCGCGTCAAGGAGCTTTACCTCCAGA
AGAAGGCTTAA
```

Primer design is an art (and a bit voodoo). For an initial attempt, I design primers to have 4-6 codons (15 nt) of specificity on either end of the gene. Then, in order of importance, the primers should have nearly identical melting temperatures, end on G/C, and have 50% G/C content. You may break these “rules.”

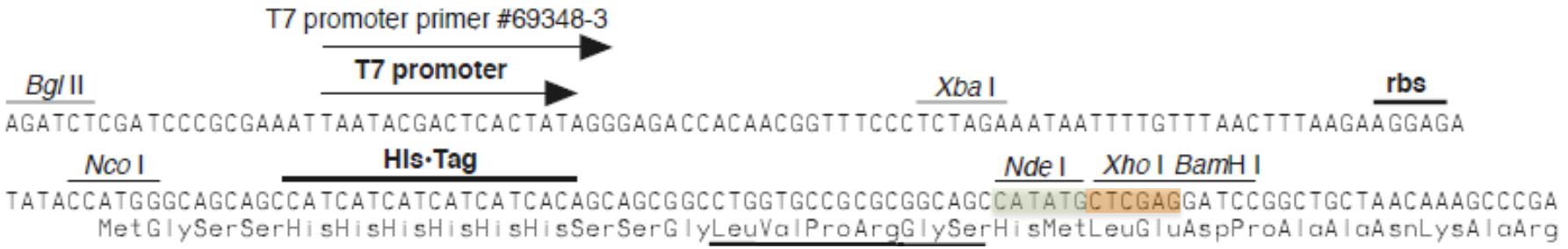
Design primers to clone gene into vector

```
5′ ATGTCTGAGAAGTCAGC……GENE……CCAGAAGAAGGCTTAA 3′
3′ TACAGACTCTTCAGTCG……GENE……GGTCTTCTTCCGAATT 5′
```

The gene sequence is above, and it is only showing the ~15 nt on the ends. For correct directionality, we want to clone the gene so the NdeI site is at the 5' end and the XhoI site is at the 3' end.

```
5′ CATATGTCTGAGAAGTCAGC……GENE……CCAGAAGAAGGCTTAACTCGAG 3′
3′ GTATACAGACTCTTCAGTCG……GENE……GGTCTTCTTCCGAATTGAGCTC 5′
```

The sequence above is what we want our product from PCR to look like. Notice that the start codon is included in the NdeI recognition sequence. There is, however, one last thing we want to consider.



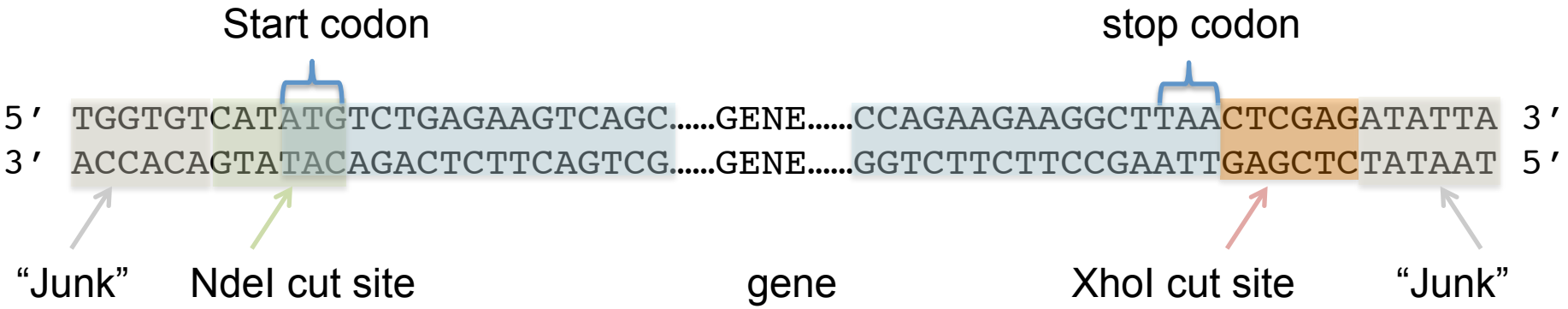
Design primers to clone gene into vector

Note: As a general rule and for enzymes not listed below, 6 base pairs should be added on on either side of the recognition site to cleave efficiently. The extra bases should be chosen so that palindromes and primer dimers are not formed. In most cases there is no requirement for specific bases.

```
5′ CATATGTCTGAGAAGTCAGC……GENE……CCAGAAGAAGGCTTAACTCGAG 3′
3′ GTATACAGACTCTTCAGTCG……GENE……GGTCTTCTTCCGAATTGAGCTC 5′
```

DNA digestion works more efficiently if the restriction enzyme has more DNA bases on either side of the restriction site. These are meaningless, “junk,” sequences only to provide space for the enzyme to bind.

Design primers to clone gene into vector



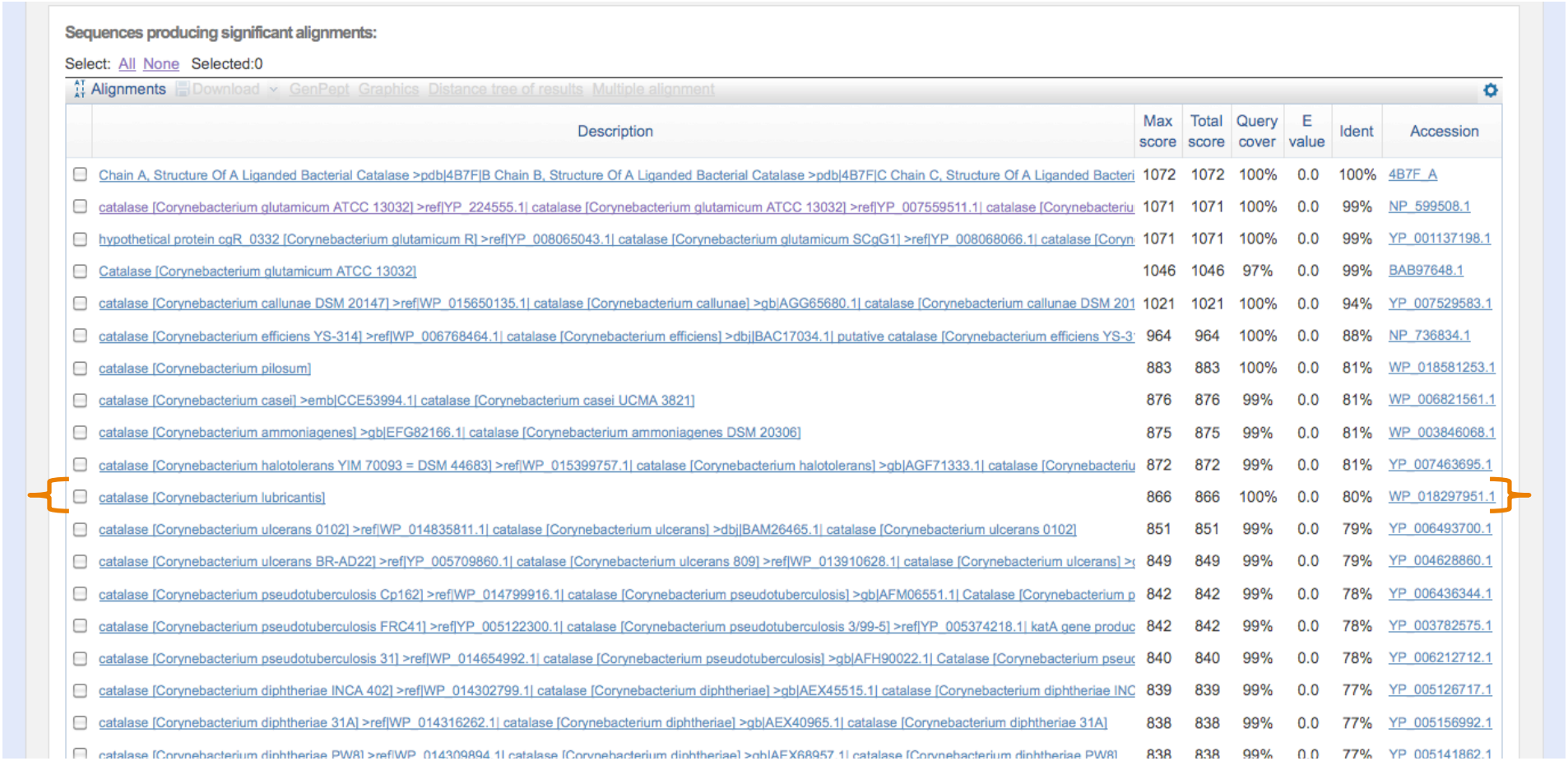
So this is what we’d actually like the PCR product to look like.

PCR amplifies DNA 5’ -> 3’, and we need two primers to synthesize the coding strand (top) and the complementary strand (bottom), and we commonly provide primers 5’ -> 3’

Our two primers, with a melting temperature of ~ 62 °C (http://www6.appliedbiosystems.com/support/techtools/calc/), are:

FWD: 5′ TGGTGTCATATGTCTGAGAAGTCAGC 3′
REV: 5′ TAATATCTCGAGTTAAGCCTTCTTCTTGG 3′

Make a multiple sequence alignment



Sequences producing significant alignments:

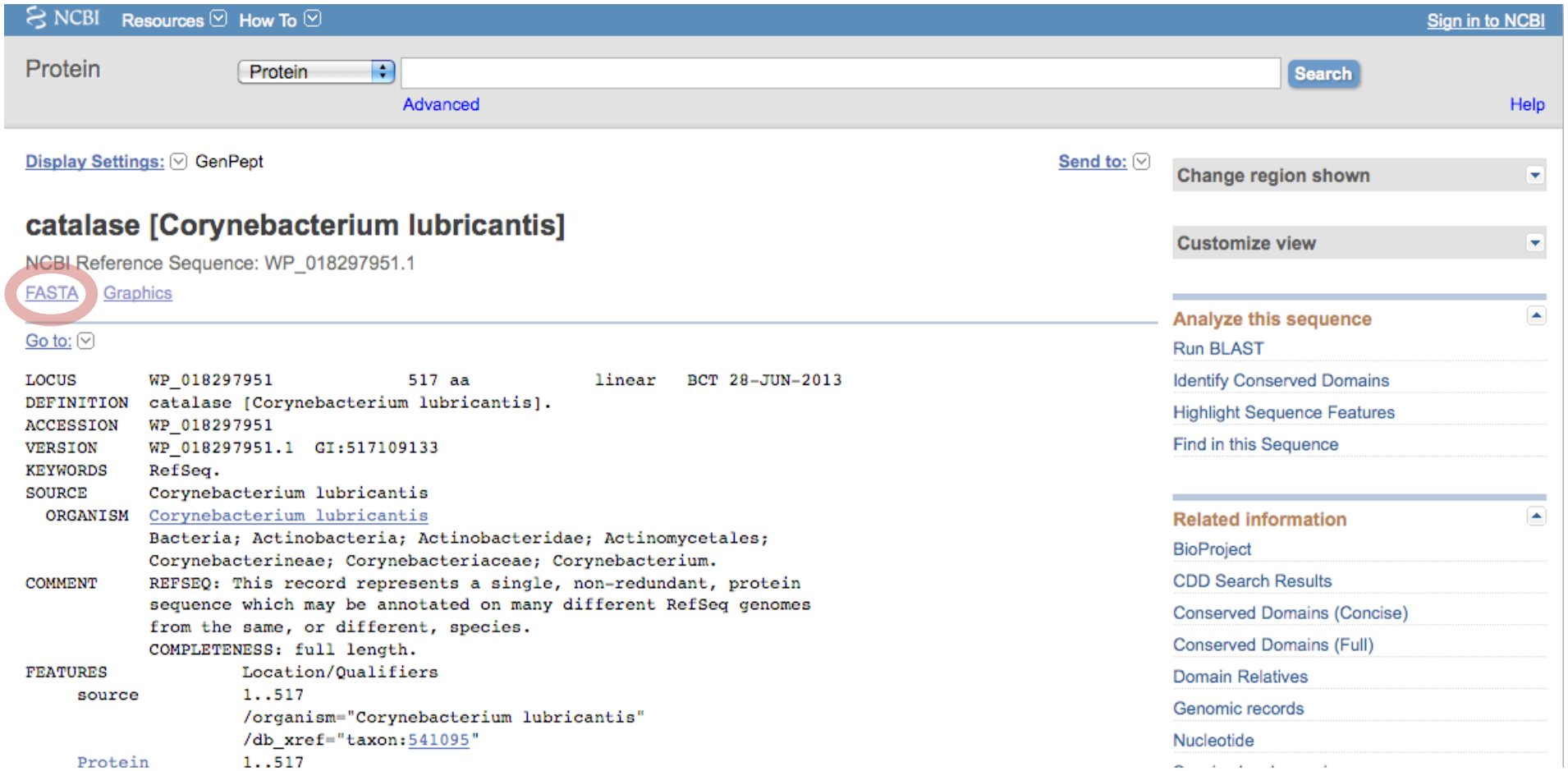
Select: All None Selected:0

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

Description	Max score	Total score	Query cover	E value	Ident	Accession
Chain A, Structure Of A Liganded Bacterial Catalase >pdb|4B7F|B Chain B, Structure Of A Liganded Bacterial Catalase >pdb|4B7F|C Chain C, Structure Of A Liganded Bacteri	1072	1072	100%	0.0	100%	4B7F_A
catalase [Corynebacterium glutamicum ATCC 13032] >ref|YP_224555.1| catalase [Corynebacterium glutamicum ATCC 13032] >ref|YP_007559511.1| catalase [Corynebacteriu	1071	1071	100%	0.0	99%	NP_599508.1
hypothetical protein cgR_0332 [Corynebacterium glutamicum R] >ref|YP_008065043.1| catalase [Corynebacterium glutamicum SCgG1] >ref|YP_008068066.1| catalase [Coryn	1071	1071	100%	0.0	99%	YP_001137198.1
Catalase [Corynebacterium glutamicum ATCC 13032]	1046	1046	97%	0.0	99%	BAB97648.1
catalase [Corynebacterium callunae DSM 20147] >ref|WP_015650135.1| catalase [Corynebacterium callunae] >gb|AGG65680.1| catalase [Corynebacterium callunae DSM 201	1021	1021	100%	0.0	94%	YP_007529583.1
catalase [Corynebacterium efficiens YS-314] >ref|WP_006768464.1| catalase [Corynebacterium efficiens] >dbj|BAC17034.1| putative catalase [Corynebacterium efficiens YS-3	964	964	100%	0.0	88%	NP_736834.1
catalase [Corynebacterium pilosum]	883	883	100%	0.0	81%	WP_018581253.1
catalase [Corynebacterium casei] >emb|CCE53994.1| catalase [Corynebacterium casei UCMA 3821]	876	876	99%	0.0	81%	WP_006821561.1
catalase [Corynebacterium ammoniagenes] >gb|EFG82166.1| catalase [Corynebacterium ammoniagenes DSM 20306]	875	875	99%	0.0	81%	WP_003846068.1
catalase [Corynebacterium halotolerans YIM 70093 = DSM 44683] >ref|WP_015399757.1| catalase [Corynebacterium halotolerans] >gb|AGF71333.1| catalase [Corynebacteriu	872	872	99%	0.0	81%	YP_007463695.1
catalase [Corynebacterium lubricantis]	866	866	100%	0.0	80%	WP_018297951.1
catalase [Corynebacterium ulcerans 0102] >ref|WP_014835811.1| catalase [Corynebacterium ulcerans] >dbj|BAM26465.1| catalase [Corynebacterium ulcerans 0102]	851	851	99%	0.0	79%	YP_006493700.1
catalase [Corynebacterium ulcerans BR-AD22] >ref|YP_005709860.1| catalase [Corynebacterium ulcerans 809] >ref|WP_013910628.1| catalase [Corynebacterium ulcerans] >(	849	849	99%	0.0	79%	YP_004628860.1
catalase [Corynebacterium pseudotuberculosis Cp162] >ref|WP_014799916.1| catalase [Corynebacterium pseudotuberculosis] >gb|AFM06551.1| Catalase [Corynebacterium p	842	842	99%	0.0	78%	YP_006436344.1
catalase [Corynebacterium pseudotuberculosis FRC41] >ref|YP_005122300.1| catalase [Corynebacterium pseudotuberculosis 3/99-5] >ref|YP_005374218.1| katA gene produc	842	842	99%	0.0	78%	YP_003782575.1
catalase [Corynebacterium pseudotuberculosis 31] >ref|WP_014654992.1| catalase [Corynebacterium pseudotuberculosis] >gb|AFH90022.1| Catalase [Corynebacterium pseuc	840	840	99%	0.0	78%	YP_006212712.1
catalase [Corynebacterium diphtheriae INCA 402] >ref|WP_014302799.1| catalase [Corynebacterium diphtheriae] >gb|AEX45515.1| catalase [Corynebacterium diphtheriae INC	839	839	99%	0.0	77%	YP_005126717.1
catalase [Corynebacterium diphtheriae 31A] >ref|WP_014316262.1| catalase [Corynebacterium diphtheriae] >gb|AEX40965.1| catalase [Corynebacterium diphtheriae 31A]	838	838	99%	0.0	77%	YP_005156992.1
catalase [Corynebacterium diphtheriae PW8] >ref|WP_014309894.1| catalase [Corynebacterium diphtheriae] >gb|AEX68957.1| catalase [Corynebacterium diphtheriae PW8]	838	838	99%	0.0	77%	YP_005141862.1

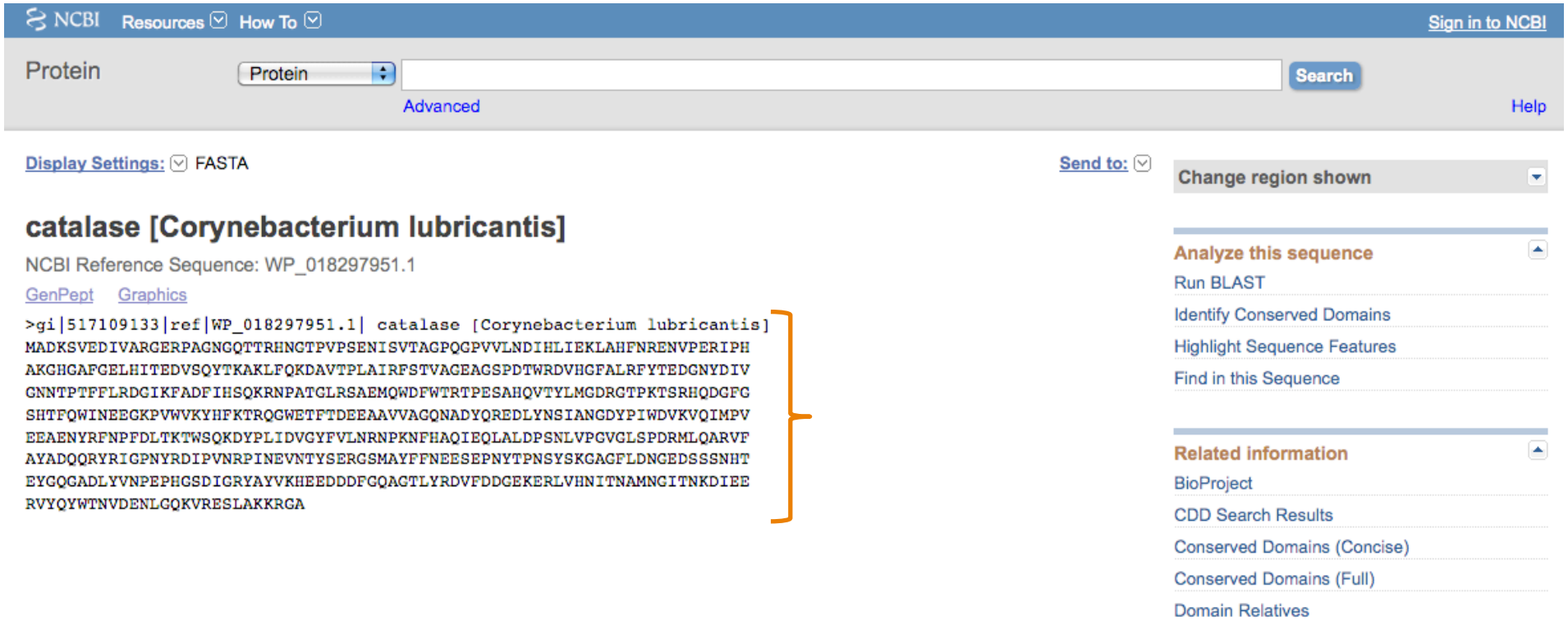
A multiple sequence alignment will allow us to determine conserved residues that are important for this enzyme to function. Go back to your BLAST results and pick proteins that are related to your own. Let's add this protein to our multiple sequence alignment.

Make a multiple sequence alignment



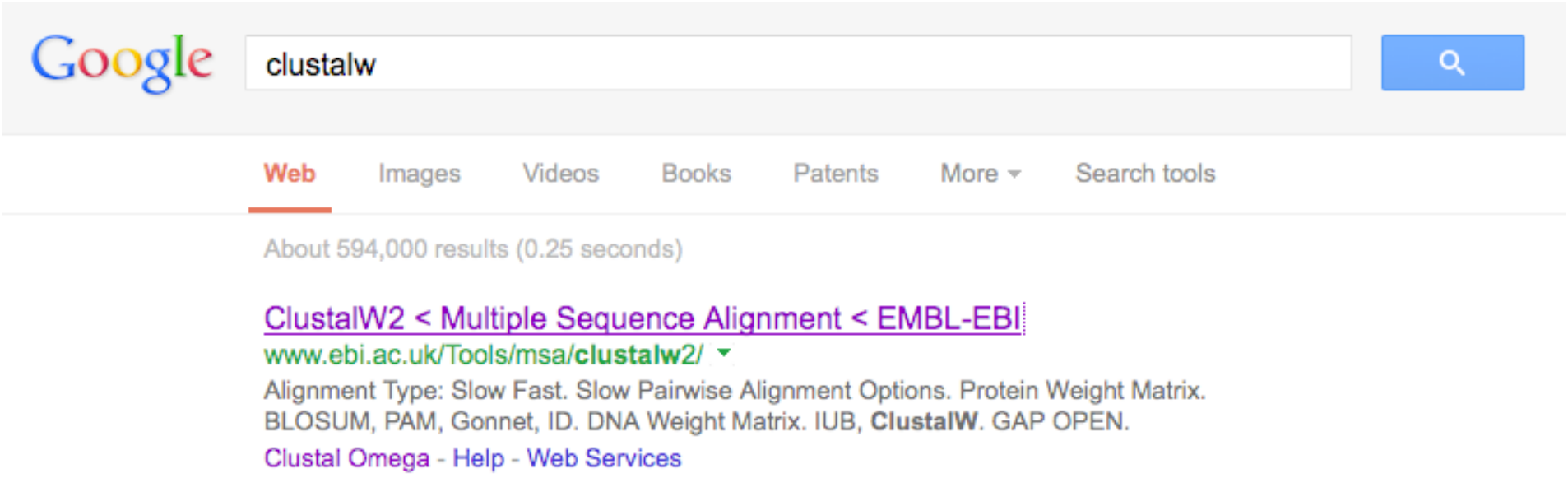
Click on the accession number and get the FASTA format for the sequence of this and the other proteins.

Make a multiple sequence alignment



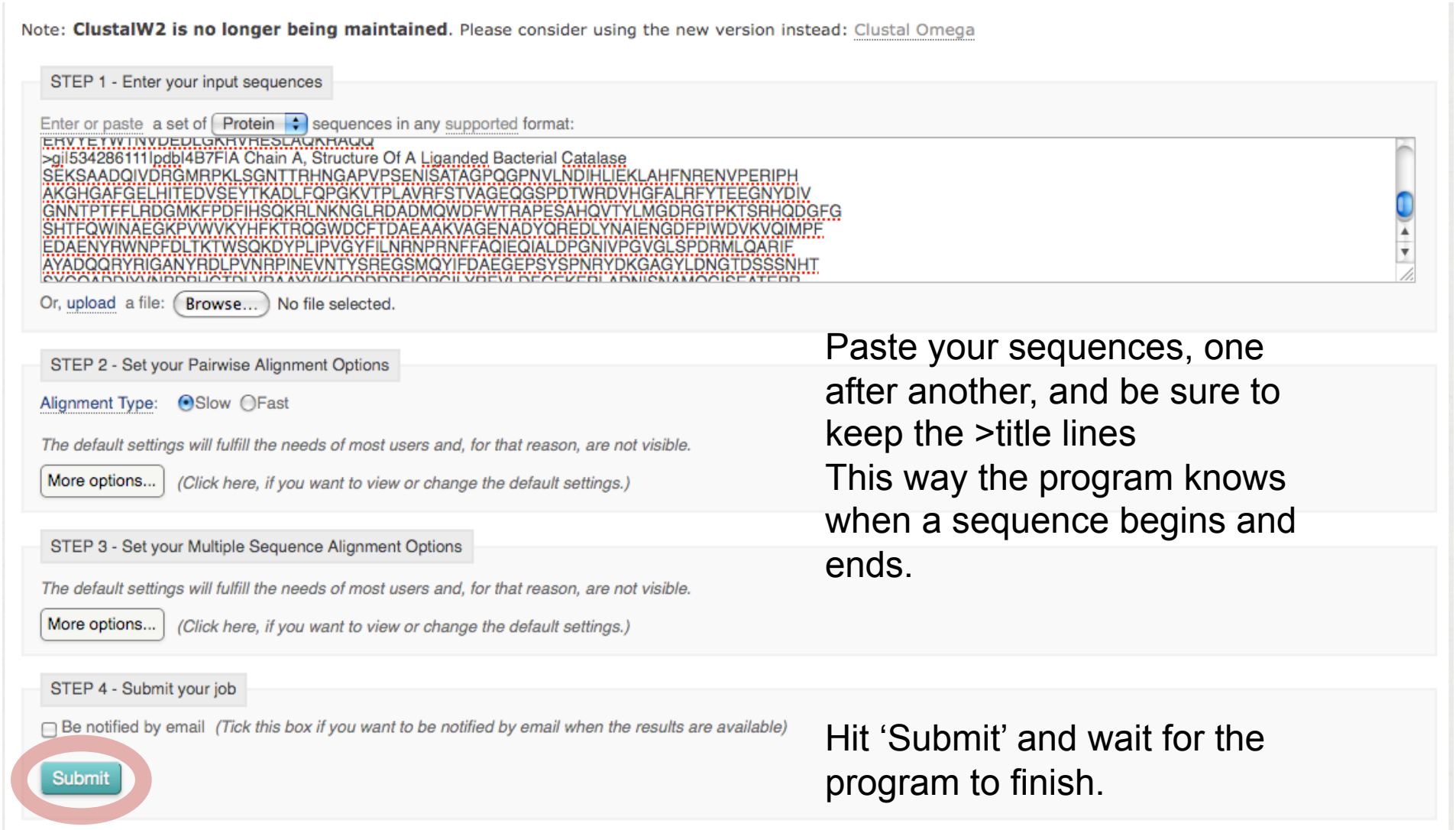
Copy the whole sequence, including the title line. Remember that everything that follows the carrot >IS A TITLE
After a return is the sequence.

Make a multiple sequence alignment



Use any program you like to make the multiple sequence alignment.
The ‘Clustal’ series are popular.

Make a multiple sequence alignment



Paste your sequences, one after another, and be sure to keep the >title lines
This way the program knows when a sequence begins and ends.

Hit ‘Submit’ and wait for the program to finish.

Make a multiple sequence alignment

```
gi|517109133|ref|WP_018297951.      -------MADKS-VEDIVARGERPAGNGQTTRHNGTPVPSENISVTAGPQ 42
gi|517408857|ref|WP_018581253.      -------MADKSAAENVVSRGERAAGDGTTTRLNGAPVPSENISVTAGPQ 43
gi|534286111|pdb|4B7F|A             --------SEKSAADQIVDRGMRPKLSGNTTRHNGAPVPSENISATAGPQ 42
gi|516831493|ref|WP_018121082.      --------MTKN-IDDKLDQGQREDAPGTTTRQGGQPIASENISITAGPQ 41
gi|497955492|ref|WP_010269648.      MTDNTTPAGNPAGSDDVAMRGVCP-VTGHSSNINGAPVRTEEHSVTVGAQ 49
                                              ::   :*      * ::. .* *: :*: * *.*.*

gi|517109133|ref|WP_018297951.      GPVVLNDIHLIEKLAHFNRENVPERIPHAKGHGAFGELHITEDVSQYTKA 92
gi|517408857|ref|WP_018581253.      GPNVLDDIHMIEKLAHFNRENVPERIPHAKGHGAFGELHITEDVSQYTKA 93
gi|534286111|pdb|4B7F|A             GPNVLNDIHLIEKLAHFNRENVPERIPHAKGHGAFGELHITEDVSEYTKA 92
gi|516831493|ref|WP_018121082.      GPNVLNDLQLIEKLQSFNRERVPERNPHAKGHGAFGEFHVTEDVSAYTKA 91
gi|497955492|ref|WP_010269648.      GPIALNDVHLIEKHAHFNRERIPERNVHAKGSGAFGELTVTEDVSKYTKA 99
                                    ** .*:*:::***   ****.:***  **** *****: :***** ****

gi|517109133|ref|WP_018297951.      KLFQKDAVTPLAIRFSTVAGEAGSPDTWRDVHGFALRFYTEDGNYDIVGN 142
gi|517408857|ref|WP_018581253.      KLFQKGTVTPMAGRFSTVAGEAGSPDTWRDVHGFALRFYTEDGNYDIVGN 143
gi|534286111|pdb|4B7F|A             DLFQPGKVTPLAVRFSTVAGEQGSPDTWRDVHGFALRFYTEEGNYDIVGN 142
gi|516831493|ref|WP_018121082.      DLFQPNKVTPMGIRFSTVAGEQGSPDTWRDVHGFALRFWTQEGNFDIVGN 141
gi|497955492|ref|WP_010269648.      DLFQPGRVTPMLARFSTVAGEQGYPDTVRDVRGFSLKFYTQEGNYDLVGN 149
                                    .*** . ***:  ******** * *** ***:**:*:*:*::**:*:***

gi|517109133|ref|WP_018297951.      NTPTFFLRDGIKFADFIHSQKRNPATGLRSAEMQWDFWTRTPESAHQVTY 192
gi|517408857|ref|WP_018581253.      NTPTFFLRDAIKFPDFIHSQKRNPASGLRDDEMQWDFWTRTPESAHQVTY 193
gi|534286111|pdb|4B7F|A             NTPTFFLRDGMKFPDFIHSQKRLNKNGLRDADMQWDFWTRAPESAHQVTY 192
gi|516831493|ref|WP_018121082.      NTPTFFLRDGIKFPDFIHSQKRTGASGLRDADMQWDFWTRTPESAHQVTY 191
gi|497955492|ref|WP_010269648.      NTPVFFLRDGIKFPDFIRSQKRLHGPGLQSADMQWDFWTRSPESAHQVTY 199
                                    ***.*****.:**.***:****    **:. :********:*********

gi|517109133|ref|WP_018297951.      LMGDRGTPKTSRHQDGFGSHTFQWINEEGKPVWVKYHFKTRQGWETFTDE 242
gi|517408857|ref|WP_018581253.      LMGDRGTPKTTRNQDGFGSHTFQWINEEGTPVWVKYHFKTRQGWDCFTDE 243
gi|534286111|pdb|4B7F|A             LMGDRGTPKTSRHQDGFGSHTFQWINAEGKPVWVKYHFKTRQGWDCFTDA 242
gi|516831493|ref|WP_018121082.      LMGDRGTPKTSRHQDGFGSHTFQWVNDKGEAFWVKYHFKTQQGWECFTDE 241
gi|497955492|ref|WP_010269648.      LMGDRGIPTDFRHMDGFSSHTYQWINADNERFWVKYHFKTRQGWKYFTDE 249
                                    ****** *.  *: ***.***:**:* ..  .********:***. ***

gi|517109133|ref|WP_018297951.      EAAVVAG-QNADYQREDLYNSIANGDYPIWDVKVQIMPVEEAENYRFNPF 291
gi|517408857|ref|WP_018581253.      EAEEMAG-KNADYHRQDLYEAIERGDYPIWDVKVQIMPFDEAENYRWNPF 292
gi|534286111|pdb|4B7F|A             EAAKVAG-ENADYQREDLYNAIENGDFPIWDVKVQIMPFEDAENYRWNPF 291
gi|516831493|ref|WP_018121082.      EAAEMAG-QNADYHREDLFKAIENGDYPRWDVYVQIMPFDEAENYKFNPF 290
gi|497955492|ref|WP_010269648.      EASQVLASQDFDHSRKDLWEAIEAGDYPTWDVKVQIMPLDEAEGYRWNPF 299
                                    **  : . :: *: *:**:::*  **:* *** *****.::**.*::***

gi|517109133|ref|WP_018297951.      DLTKTWSQKDYPLIDVGYFVLNRNPKNFHAQIEQLALDPSNLVPGVGLSP 341
gi|517408857|ref|WP_018581253.      DLTKTWSQKDYPLHDVGYFVLNRNPKNYHAQIEQMALDPSNLVPGVGLSP 342
gi|534286111|pdb|4B7F|A             DLTKTWSQKDYPLIPVGYFILNRNPRNFFAQIEQIALDPGNIVPGVGLSP 341
gi|516831493|ref|WP_018121082.      DLTKVWYTEDYPLQKVGYFELNRNPRNNFAQIEQIALDPSNIVPGTGLSP 340
gi|497955492|ref|WP_010269648.      DLTKTWSQKDYPLIPVGHFTLNRNPENFFAQIEQAAFAPSNIVPGIGFSP 349
                                    ****.*  :****  **:* *****.* .***** *: *.*:*** *:**
```

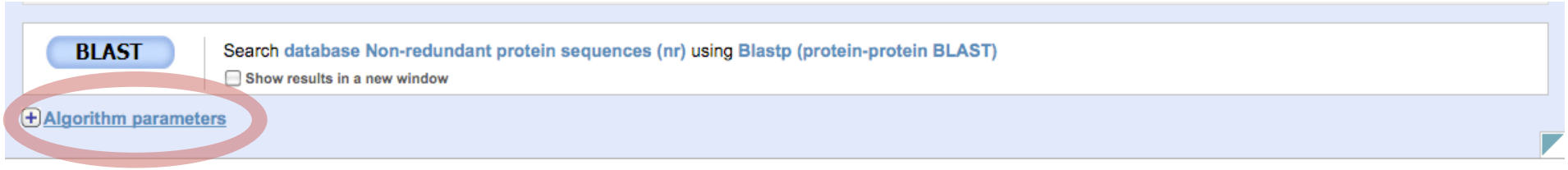
This is an example of a bad sequence alignment.

What's wrong?

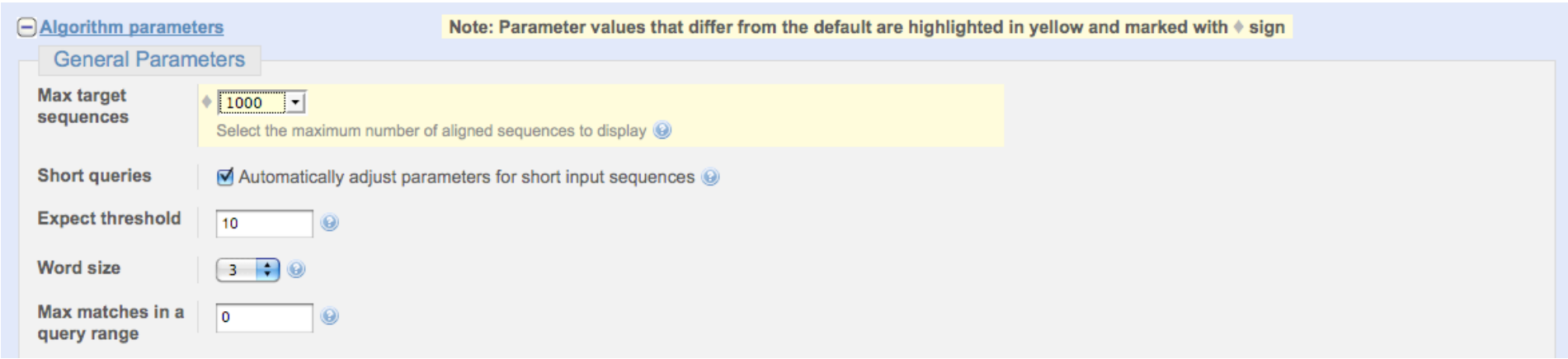
1) These sequences are far too similar. You can see that about 80% of residues here are identical. We need more diversity to discover the truly important residues.

2) The titles are meaningless. When you copy/paste, the title begins with the accession number, so you'll need to alter the title.

Make a multiple sequence alignment



There wasn't enough diversity in the results given the default settings. We'll have to change them. Go into 'Algorithm parameters.'



We'll have to get more results, so first try increase the number of result sequences given. BLAST it again.

Make a multiple sequence alignment

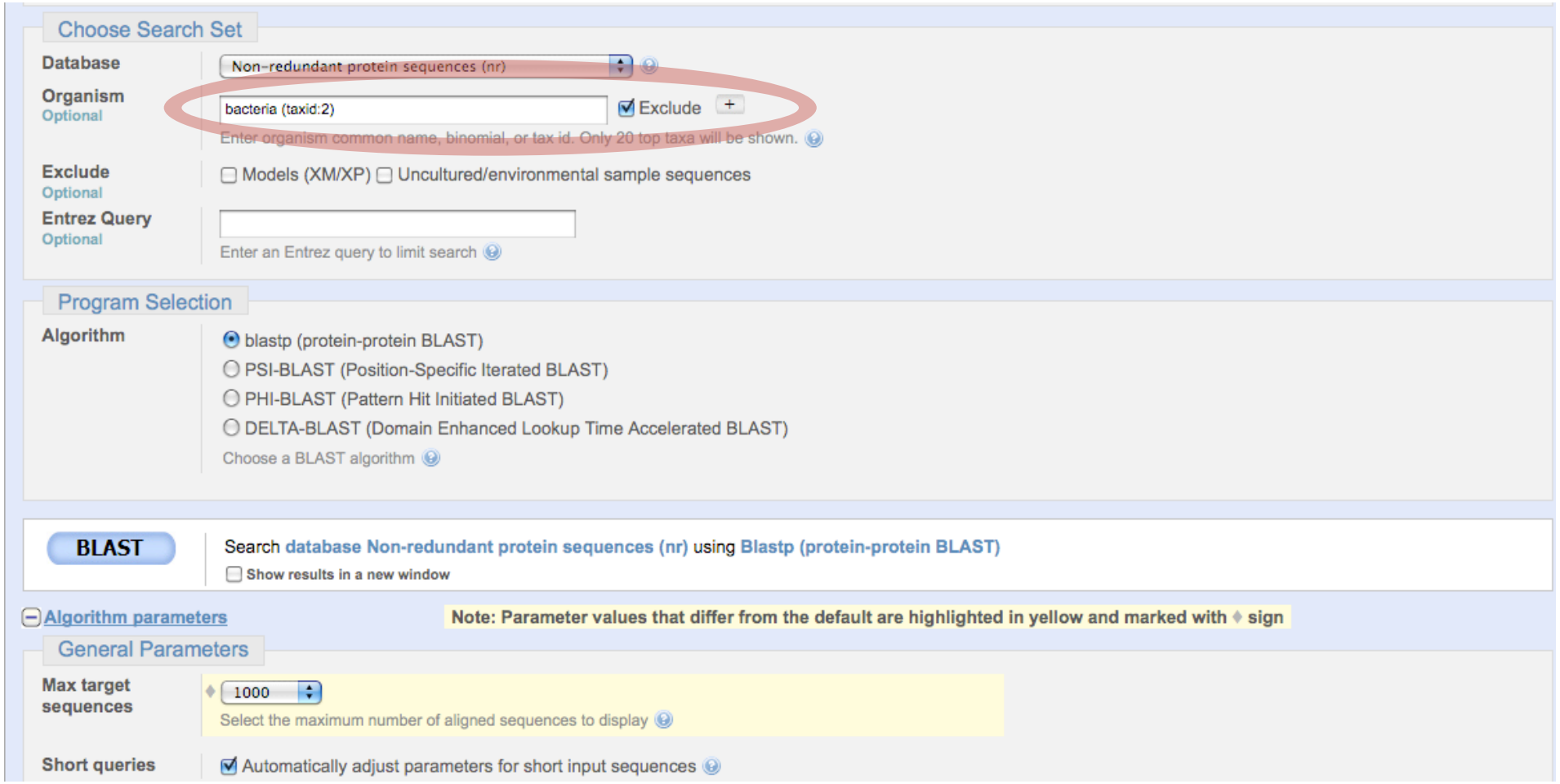
catalase [Pseudomonas syringae]	482	482	93%	1e-162	51%	WP_010438211.1
catalase [Pseudomonas fluorescens] >gb|EJK58295.1| catalase KatA [Pseudomonas fluorescens Q8r1-96]	482	482	93%	1e-162	51%	WP_003206086.1

catalase [Daphnia magna]	454	454	97%	3e-151	44%	ACU81116.1
catalase [Bordetella pertussis Tohama I] >ref|NP_886532.1| catalase [Bordetella parapertussis 12822] >ref|YP_005591739.1| ca	453	453	96%	3e-151	46%	NP_882347.1

Given 1000 results, there still isn't enough diversity! Remember that my protein was from *Corynebacterium glutamicum*, a Gram positive bacterium, so from the 1000 results, I've included *Pseudomonas syringae*, a Gram negative bacterium, and *Daphnia magna*, a water flea.

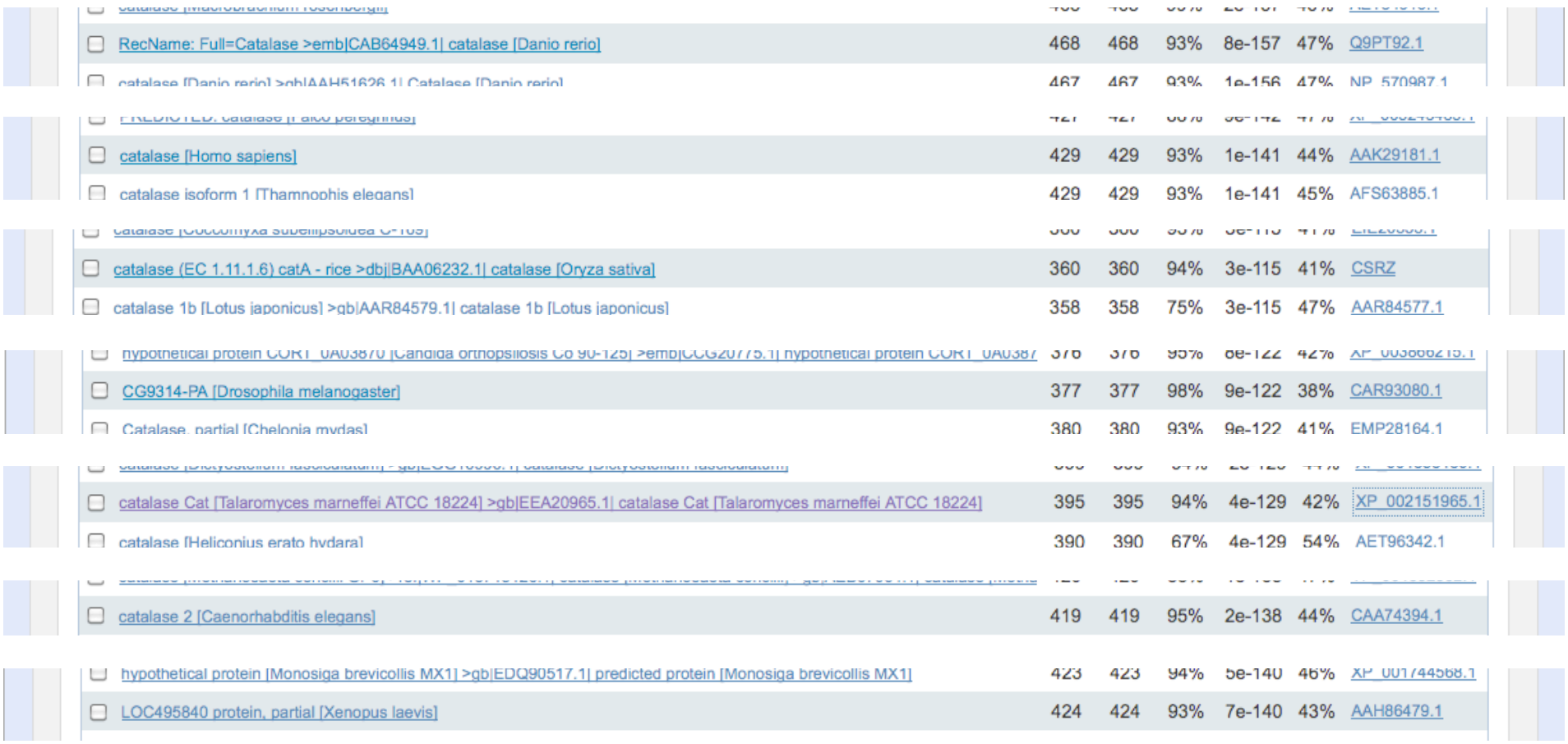
I have to go back and change more settings.

Make a multiple sequence alignment



I'm still asking for 1000 sequences, but I want to exclude all bacteria from my results in order to get greater diversity. You can also force the program to only give results from whatever taxon you desire.

Make a multiple sequence alignment



RecName: Full=Catalase >emb|CAB64949.1| catalase [Danio rerio]	468	468	93%	8e-157	47%	Q9PT92.1
catalase [Homo sapiens]	429	429	93%	1e-141	44%	AAK29181.1
catalase isoform 1 [Thamnophis elegans]	429	429	93%	1e-141	45%	AFS63885.1
catalase (EC 1.11.1.6) catA - rice >dbj|BAA06232.1| catalase [Oryza sativa]	360	360	94%	3e-115	41%	CSRZ
catalase 1b [Lotus japonicus] >gb|AAR84579.1| catalase 1b [Lotus japonicus]	358	358	75%	3e-115	47%	AAR84577.1
CG9314-PA [Drosophila melanogaster]	377	377	98%	9e-122	38%	CAR93080.1
Catalase, partial [Chelonia mydas]	380	380	93%	9e-122	41%	EMP28164.1
catalase Cat [Talaromyces marneffei ATCC 18224] >gb|EEA20965.1| catalase Cat [Talaromyces marneffei ATCC 18224]	395	395	94%	4e-129	42%	XP_002151965.1
catalase [Heliconius erato hydara]	390	390	67%	4e-129	54%	AET96342.1
catalase 2 [Caenorhabditis elegans]	419	419	95%	2e-138	44%	CAA74394.1
hypothetical protein [Monosiga brevicollis MX1] >gb|EDQ90517.1| predicted protein [Monosiga brevicollis MX1]	423	423	94%	5e-140	46%	XP_001744568.1
LOC495840 protein, partial [Xenopus laevis]	424	424	93%	7e-140	43%	AAH86479.1

This is better. I've got sequences from Zebra fish, human, rice, fruit fly, a fungus, a worm, and a frog.

Make a multiple sequence alignment

```
Danio.rerio                          MADDREKSTDQMKLWKEGRGSQRPDVLTTGAGVPIGDKLNAMTAGPRGPL 50
Homo.sapiens                         MADSRDPASDQMQHWKEQRAAQKADVLTTGAGNPVGDKLNVITVGPRGPL 50
Daphnia.magna                        -MAERDAATNQLKNFGESKKNEAKASVTTSAGAPIGFKTASLTAGPRGPM 49
Caenorhabditis.elegans               --MPNDPSDNQLKTYKETYP--KPQVITTSNGAPIYSKTAVLTAGRRGPM 46
Drosophila                           -MCSRDTASNQLIDYKNNDS-EVQREITTSSGTPVGVKDAIQTVGPRGPA 48
Talaromyces.marneffei                --MGKD---DEPKTYRYNET----PTYTTSNGCPVFDPESSQRIGKNGPL 41
Pseudomonas.syringae                 ------------------MSQKT---LTTASGAPVADNQNSRSAGPRGPL 29
BLASTEDproteinCorynebacterium.       ---SEKSAADQIVDR--GMRPKLSGNTTRHNGAPVPSENISATAGPQGPN 45
Oryza.sativa                         ----MDPCKFRPSSSFDTKT------TTTNAGAPVWNDNEALTVGPRGPI 40
                                                           *   * *:         * .**

Danio.rerio                          LVQDVVFTDEMAHFDRERIPERVVHAKGAGAFGYFEVTHDITRYSKAKVF 100
Homo.sapiens                         LVQNVVFTDEMAHFDRERIPERVVHAKGAGAFGYFEVTHDITKYSKAKVF 100
Daphnia.magna                        LLQDHVYIDETAHFDRERISERVVHAKGSGAFGYFEVTHDISKYCKAAIF 99
Caenorhabditis.elegans               LMQDVVYMDEMAHFDRERIPERVVHAKGAGAHGYFEVTHDITKYCKADMF 96
Drosophila                           LLQDFQFLDEVMHFDSERIPERVAYAKGAGAFGYFECTHDISKFCAASIF 98
Talaromyces.marneffei                LLQDFHLIDLLAHFDRERIPERVVHAKGAGAYGEFEVTDDISDITTIDML 91
Pseudomonas.syringae                 LLDDFHLIEKLAHFNRENIPERRVHAKGSGAHGTFTVTRDISQYTSAKLF 79
BLASTEDproteinCorynebacterium.       VLNDIHLIEKLAHFNRENVPERIPHAKGHGAFGELHITEDVSEYTKADLF 95
Oryza.sativa                         LLEDYHLIEKVAHFARERIPERVVHARGASAKGFFECTHDVTDITCADFL 90
                                     ::::    :   **  *.:.**  :*:* .* * :  * *::      .:

Danio.rerio                          EHVGKTTPIVVRFSTVAGEAGSPDTVRDPRGFAVKFYTDEGNWDLTGNNT 150
Homo.sapiens                         EHIGKKTPIAVRFSTVAGESGSADTVRDPRGFAVKFYTEDGNWDLVGNNT 150
Daphnia.magna                        SQVGKKTPVAVRFSTVGGESGSADTARDPRGFAVKFYTEEGNWDLVGNNT 149
Caenorhabditis.elegans               NKVGKQTPLLVRFSTVAGESGSADTVRDPRGFSLKFYTEEGNWDLVGNNT 146
Drosophila                           DKVRKRTAVAMRFSVACGEQGSADTVREQRGFAVKFYTDDGIWDIVGCNM 148
Talaromyces.marneffei                KGVGKKTKLVTRFSTVGGEKGSADSARDPRGFSVKFYTEQGNWDWVFNNT 141
Pseudomonas.syringae                 DTVGKQTPIFLRFSTVGGERGSADTERDPRGFAIKFYTEEGNWDIVGNNT 129
BLASTEDproteinCorynebacterium.       Q-PGKVTPLAVRFSTVAGEQGSPDTWRDVHGFALRFYTEEGNYDIVGNNT 144
Oryza.sativa                         RSPGAQTPVIVRFSTVIHERGSPETIRDPRGFAVKFYTREGNWDLLGNNF 140
                                           * :  ***..  * **.:: *: :**:::*** :* :*    *

Danio.rerio                          PTFFIRDTLLSPSFIHSQKRNPQTHLKDPDMVWDFWSLRPESLHQVSFLF 200
Homo.sapiens                         PIFFIRDPILFPSFIHSQKRNPQTHLKDPDMVWDFWSLRPESLHQVSFLF 200
Daphnia.magna                        PIFFIRDPILFPSFIHTQKRNPVTHLKDPDMFWDFISLRPETTHQVCFLF 199
Caenorhabditis.elegans               PIFFIRDAIHFPNFIHALKRNPQTHMRDPNALFDFWMNRPESIHQVMFLY 196
Drosophila                           PVHYVRDPMLFPSLVHAQKRNPQTHLKDPDMFWDFMTLRPETLHALLMYF 198
Talaromyces.marneffei                PVFFLRDPSKFPVFIHTQKRNPQTNLKDANMFWDYLSTHQESAHQVMHLF 191
Pseudomonas.syringae                 PVFFIRDPMKFPDFIHTQKRLPQTNLKSPQMMWDFWSHSPEALHQVTILF 179
BLASTEDproteinCorynebacterium.       PTFFLRDGMKFPDFIHSQKRLNKNGLRDADMQWDFWTRAPESAHQVTYLM 194
Oryza.sativa                         PVFFIRDGIKFPDVIHAFKPNPRSHVQEYWRVFDFLSHHPESLHTFFFLF 190
                                     * .::**    * .:*: *    . ::.    :*:     *: * .
```

Much better. Fewer conserved residues and some moderately mutated positions.

Catalase is a difficult protein to make a multiple sequence alignment because it is a ‘perfect’ protein, ubiquitous in nearly all living organisms.

For some proteins, a good set of relatives may only require sequences from different bacteria.

Make an informative image of protein

See the PyMol Demo on how to make an image.

The most important thing with this is to have the image tell us something useful. Don’t make an image just because we told you to do so.

Write 0.5 page summary on protein

Tell us something interesting about the protein. Say something about the function, and look through the primary literature to discuss something unique. The Pubmed reference for the protein that has a crystal structure is a good place to start looking for information. Wikipedia is a good place to start, but don’t stick to that alone.