Gaeun Jun
A16814573
gajun@ucsd.edu

The find-a-gene project assignment

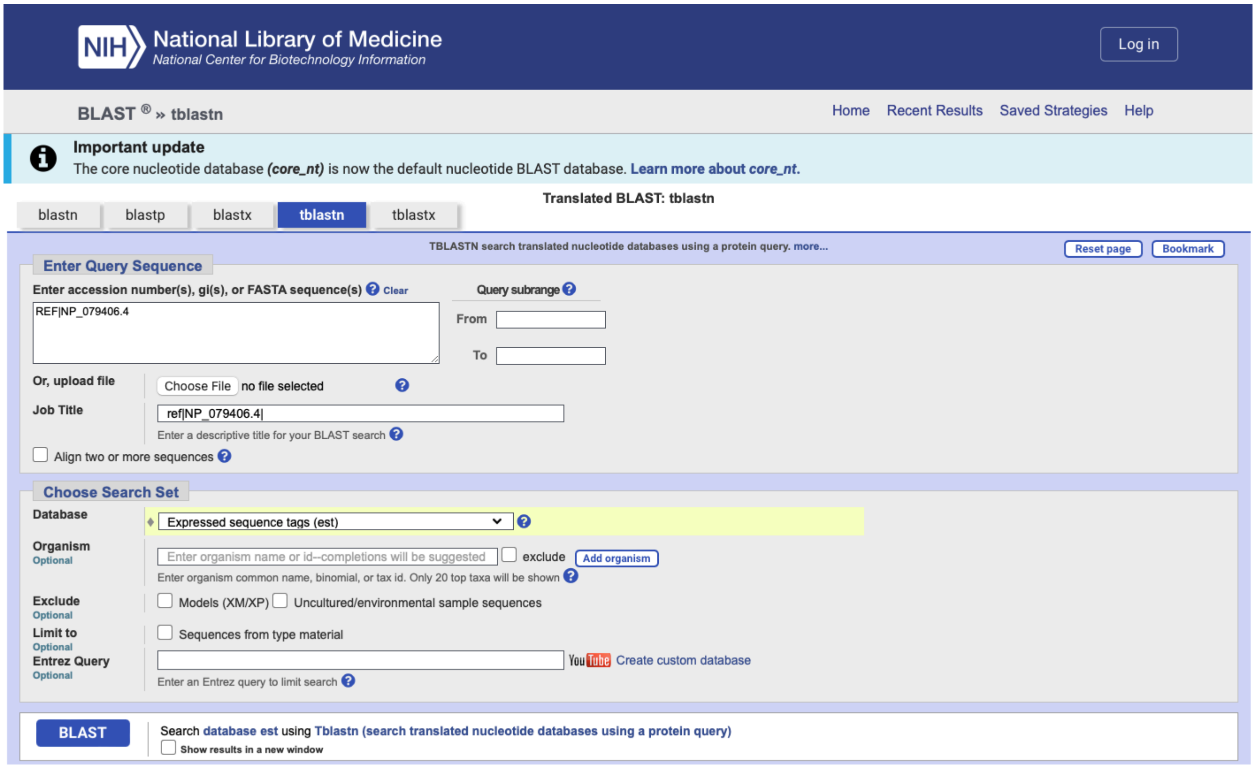
[Q1] Tell me the name of a protein you are interested in. Include the species and the accession number. This can be a human protein or a protein from any other species as long as its function is known.

Name: Hexokinase HKDC1
Accession: NP_079406.4
Species: Homo Sapiens

[Q2] Perform a BLAST search against a DNA database, such as a database consisting of genomic DNA or ESTs. The BLAST server can be at NCBI or elsewhere. Include details of the BLAST method used, database searched and any limits applied (e.g. Organism).

Method: TBLASTN
Database: Expressed Sequence Tags (est)

Also include the output of that BLAST search in your document. If appropriate, change the font to Courier size 10 so that the results are displayed neatly. You can also screen capture a BLAST output (e.g. alt print screen on a PC or on a MAC press ⌘-shift-4. The pointer becomes a bulls eye. Select the area you wish to capture and release. The image is saved as a file called Screen Shot [].png in your Desktop directory). It is not necessary to print out all of the blast results if there are many pages.



Important update
The core nucleotide database *(core_nt)* is now the default nucleotide BLAST database. **Learn more about *core_nt*.**

Translated BLAST: tblastn

TBLASTN search translated nucleotide databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear: REF|NP_079406.4

Query subrange: From ___ To ___

Or, upload file: no file selected

Job Title: ref|NP_079406.4|
Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Expressed sequence tags (est)

Organism (Optional): Enter organism name or id--completions will be suggested □ exclude
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude (Optional): □ Models (XM/XP) □ Uncultured/environmental sample sequences

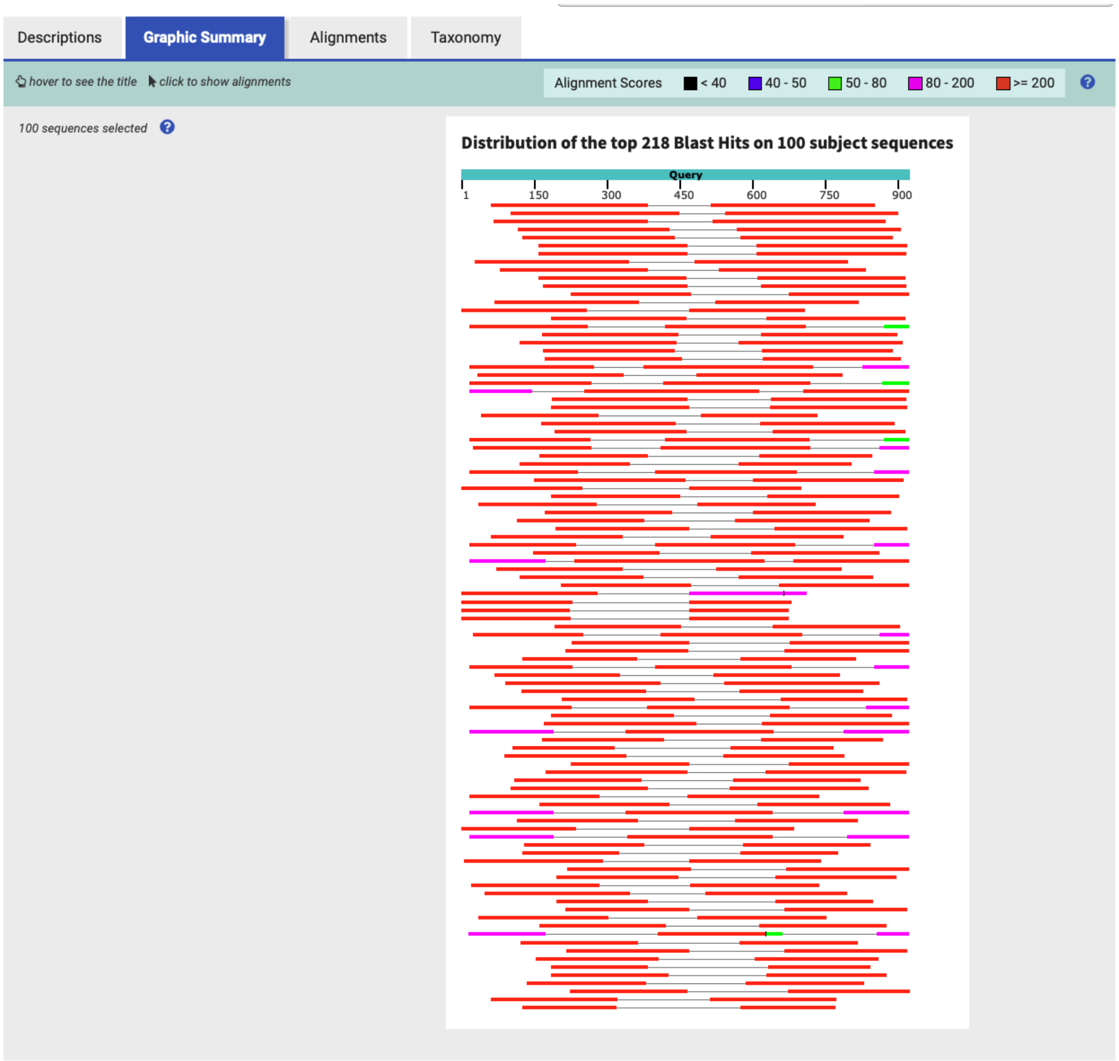
Limit to (Optional): □ Sequences from type material

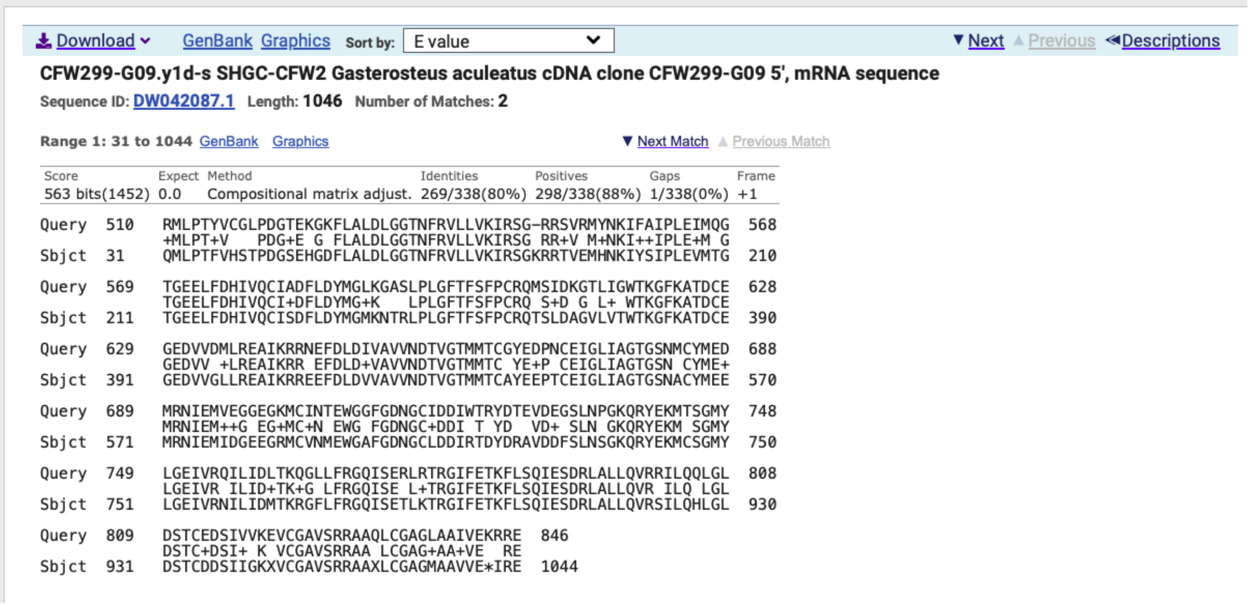
Entrez Query (Optional): ___ Create custom database
Enter an Entrez query to limit search

Search database est using Tblastn (search translated nucleotide databases using a protein query)
□ Show results in a new window

On the BLAST results, clearly indicate a match that represents a protein sequence, encoded from some DNA sequence, that is homologous to your query protein. I need to be able to inspect the pairwise alignment you have selected, including the E value and score. It should be labeled a "genomic clone" or "mRNA sequence", etc. - but include no functional annotation.

Chosen match: Accession DW042087.1, a 1046 base pair clone from *Gasterosteus aculeatus.* See below for alignment details.





CFW299-G09.y1d-s SHGC-CFW2 Gasterosteus aculeatus cDNA clone CFW299-G09 5', mRNA sequence

Sequence ID: DW042087.1 Length: 1046 Number of Matches: 2

Range 1: 31 to 1044 GenBank Graphics ▼ Next Match ▲ Previous Match

```
Score           Expect Method                        Identities     Positives      Gaps        Frame
563 bits(1452)  0.0    Compositional matrix adjust.  269/338(80%)   298/338(88%)   1/338(0%)   +1

Query  510   RMLPTYVCGLPDGTEKGKFLALDLGGTNFRVLLVKIRSG-RRSVRMYNKIFAIPLEIMQG  568
             +MLPT+V   PDG+E G FLALDLGGTNFRVLLVKIRSG RR+V M+NKI++IPLE+M G
Sbjct  31    QMLPTFVHSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKIYSIPLEVMTG  210

Query  569   TGEELFDHIVQCIADFLDYMGLKGASLPLGFTFSFPCRQMSIDKGTLIGWTKGFKATDCE  628
             TGEELFDHIVQCI+DFLDYMG+K   LPLGFTFSFPCRQ S+D G L+ WTKGFKATDCE
Sbjct  211   TGEELFDHIVQCISDFLDYMGMKNTRLPLGFTFSFPCRQTSLDAGVLVTWTKGFKATDCE  390

Query  629   GEDVVDMLREAIKRRNEFDLDIVAVVNDTVGTMMTCGYEDPNCEIGLIAGTGSNMCYMED  688
             GEDVV +LREAIKRR EFDLD+VAVVNDTVGTMMTC YE+P CEIGLIAGTGSN CYME+
Sbjct  391   GEDVVGLLREAIKRREEFDLDVVAVVNDTVGTMMTCAYEEPTCEIGLIAGTGSNACYMEE  570

Query  689   MRNIEMVEGGEGKMCINTEWGGFGDNGCIDDIWTRYDTEVDEGSLNPGKQRYEKMTSGMY  748
             MRNIEM++G EG+MC+N EWG FGDNGC+DDI T YD  VD+ SLN GKQRYEKM SGMY
Sbjct  571   MRNIEMIDGEEGRMCVNMEWGAFGDNGCLDDIRTDYDRAVDDFSLNSGKQRYEKMCSGMY  750

Query  749   LGEIVRQILIDLTKQGLLFRGQISERLRTRGIFETKFLSQIESDRLALLQVRRILQQLGL  808
             LGEIVR ILID+TK+G LFRGQISE L+TRGIFETKFLSQIESDRLALLQVR ILQ LGL
Sbjct  751   LGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSQIESDRLALLQVRSILQHLGL  930

Query  809   DSTCEDSIVVKEVCGAVSRRAAQLCGAGLAAIVEKRRE  846
             DSTC+DSI+ K VCGAVSRRAA LCGAG+AA+VE  RE
Sbjct  931   DSTCDDSIIGKXVCGAVSRRAAXLCGAGMAAVVE*IRE  1044
```

Alignment details:

>gb|DW042087.1| CFW299-G09.y1d-s SHGC-CFW2 Gasterosteus aculeatus
cDNA clone CFW299-G09 5', mRNA sequence
Length: 1046

Score = 563 bits (1452), Expect = 0.0, Method: compositional matrix adjust. Identities = 269/338 (80%), Positives = 298/338 (88%), Gaps = 1/338 (0%)
Frame = +1

```
Query  510  RMLPTYVCGLPDGTEKGKFLALDLGGTNFRVLLVKIRSG-RRSVRMYNKIFAIPLEIMQG  568
            +MLPT+V   PDG+E G FLALDLGGTNFRVLLVKIRSG RR+V M+NKI++IPLE+M G
Sbjct  31   QMLPTFVHSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKIYSIPLEVMTG  210

Query  569  TGEELFDHIVQCIADFLDYMGLKGASLPLGFTFSFPCRQMSIDKGTLIGWTKGFKATDCE  628
            TGEELFDHIVQCI+DFLDYMG+K   LPLGFTFSFPCRQ S+D G L+ WTKGFKATDCE
Sbjct  211  TGEELFDHIVQCISDFLDYMGMKNTRLPLGFTFSFPCRQTSLDAGVLVTWTKGFKATDCE  390

Query 629   GEDVVDMLREAIKRRNEFDLDIVAVVNDTVGTMMTCGYEDPNCEIGLIAGTGSNMCYMED  688
            GEDVV + LREAIKRR EFDLD+VAVVNDTVGTMMTC YE+P CEIGLIAGTGSN CYME+
Sbjct  391  GEDVVGLLREAIKRREEFDLDVVAVVNDTVGTMMTCAYEEPTCEIGLIAGTGSNACYMEE  570

Query 689   MRNIEMVEGGEGKMCINTEWGGFGDNGCIDDIWTRYDTEVDEGSLNPGKQRYEKMTSGMY  748
            MRNIEM++G EG+MC+N EWG FGDNGC+DDI T YD  VD+ SLN GKQRYEKM SGMY
Sbjct 571   MRNIEMIDGEEGRMCVNMEWGAFGDNGCLDDIRTDYDRAVDDFSLNSGKQRYEKMCSGMY  750

Query  749  LGEIVRQILIDLTKQGLLFRGQISERLRTRGIFETKFLSQIESDRLALLQVRRILQQLGL  808
            LGEIVR ILID+TK+G LFRGQISE L+TRGIFETKFLSQIESDRLALLQVR IL Q LGL
Sbjct  751  LGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSQIESDRLALLQVRSILQHLGL  930

Query  809  DSTCEDSIVVKEVCGAVSRRAAQLCGAGLAAIVEKRRE  846
            DSTC+DSI+ K VCGAVSRRAA LCGAG+AA+VE  RE
Sbjct  931  DSTCDDSIIGKXVCGAVSRRAAXLCGAGMAAVVE*IRE  1044
```

In general, [Q2] is the most difficult for students because it requires you to have a “feel” for how to interpret BLAST results. You need to distinguish between a perfect match to your query (i.e. a sequence that is not “novel”), a near match (something that might be “novel”, depending on the results of [Q4]), and a non-homologous result. If you are having trouble finding a novel gene try restricting your search to an organism that is poorly annotated.

[Q3] Gather information about this “novel” protein. At a minimum, show me the protein sequence of the “novel” protein as displayed in your BLAST results from [Q2] as FASTA format (you can copy and paste the aligned sequence subject lines from your BLAST result page if necessary) or translate your novel DNA sequence using a tool called EMBOSS Transeq at the EBI. Don’t forget to translate all six reading frames; the ORF (open reading frame) is likely to be the longest sequence without a stop codon. It may not start with a methionine if you don’t have the complete coding region. Make sure the sequence you provide includes a header/subject line and is in traditional FASTA format.

Chosen sequence: >EMBOSS_001_1 (sequence taken from EMBOSS Transeq at the EBI)

```
RSGFPG*LPCQMLPTFVHSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKI
YSIPLEVMTGTGEELFDHIVQCISDFLDYMGMKNTRLPLGFTFSFPCRQTSLDAGVLVTW
TKGFKATDCEGEDVVGLLREAIKRREEFDLDVVAVVNDTVGTMMTCAYEEPTCEIGLIAG
TGSNACYMEEMRNIEMIDGEEGRMCVNMEWGAFGDNGCLDDIRTDYDRAVDDFSLNSGKQ
RYEKMCSGMYLGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSQIESDRLALLQ
VRSILQHLGLDSTCDDSIIGKXVCGAVSRRAAXLCGAGMAAVVE*IRET
```

Here, tell me the name of the novel protein, and the species from which it derives. It is very unlikely (but still definitely possible) that you will find a novel gene from an organism such as S. cerevisiae, human or mouse, because those genomes have already been thoroughly annotated. It is more likely that you will discover a new gene in a genome that is currently being sequenced, such as bacteria or plants or protozoa.

Name: Three-spined stickleback
Species: Gasterosteus aculeatus

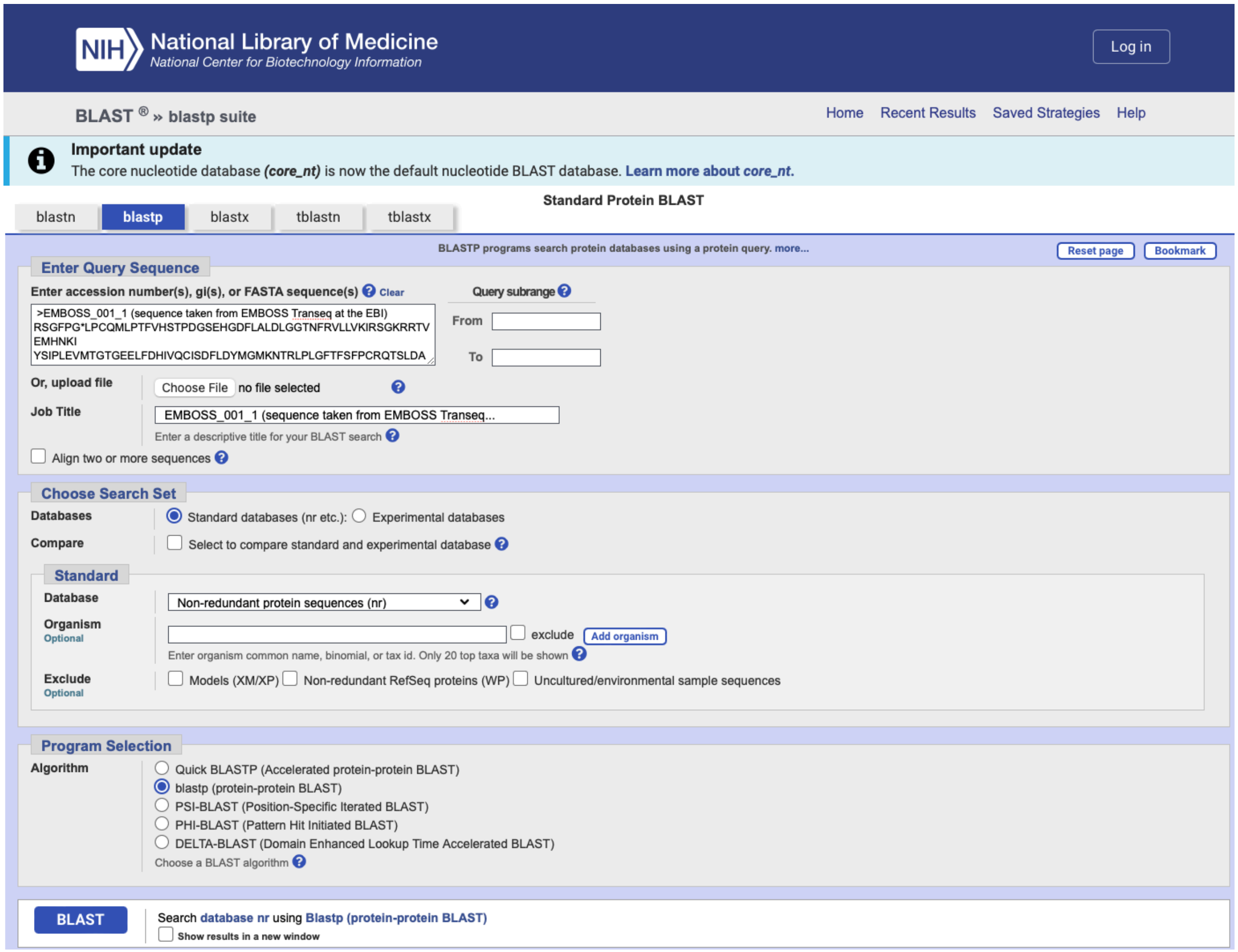
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Actinopterygii; Neopterygii; Teleostei; Neoteleostei;
Acanthomorphata; Eupercaria; Perciformes; Cottioidei;
Gasterosteales; Gasterosteidae; Gasterosteus.

[Q4] Prove that this gene, and its corresponding protein, are novel. For the purposes of this project, "novel" is defined as follows. Take the protein sequence (your answer to [Q3]), and use it as a query in a blastp search of the nr database at NCBI.

- If there is a match with 100% amino acid identity to a protein in the database, from the same species, then your protein is NOT novel (even if the match is to a protein with a name such as "unknown"). Someone has already found and annotated this sequence, and assigned it an accession number.
- If the top match reported has less than 100% identity, then it is likely that your protein is novel, and you have succeeded.
- If there is a match with 100% identity, but to a different species than the one you started with, then you have likely succeeded in finding a novel gene.
- If there are no database matches to the original query from [Q1], this indicates that you have partially succeeded: yes, you may have found a new gene, but no, it is not actually homologous to the original query. You should probably start over.

Details: A BLASTP search against NR database yielded a top hit result to a protein from *Gasterosteus aculeatus aculeatus* (three-spined stickleback).

See additional screen shots below for top hits and selected alignment details:



NIH National Library of Medicine
National Center for Biotechnology Information
Log in

BLAST ® » blastp suite

Important update
The core nucleotide database *(core_nt)* is now the default nucleotide BLAST database. **Learn more about *core_nt*.**

Standard Protein BLAST

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
>EMBOSS_001_1 (sequence taken from EMBOSS Transeq at the EBI)
RSGFPG*LPCQMLPTFVHSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTV
EMHNKI
YSIPLEVMTGTGEELFDHIVQCISDFLDYMGMKNTRLPLGFTFSFPCRQTSLDA
```

Query subrange
From
To

Or, upload file: Choose File no file selected

Job Title: EMBOSS_001_1 (sequence taken from EMBOSS Transeq...
Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Databases: ● Standard databases (nr etc.): ○ Experimental databases

Compare: ☐ Select to compare standard and experimental database

Standard

Database: Non-redundant protein sequences (nr)

Organism (Optional): ☐ exclude Add organism
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

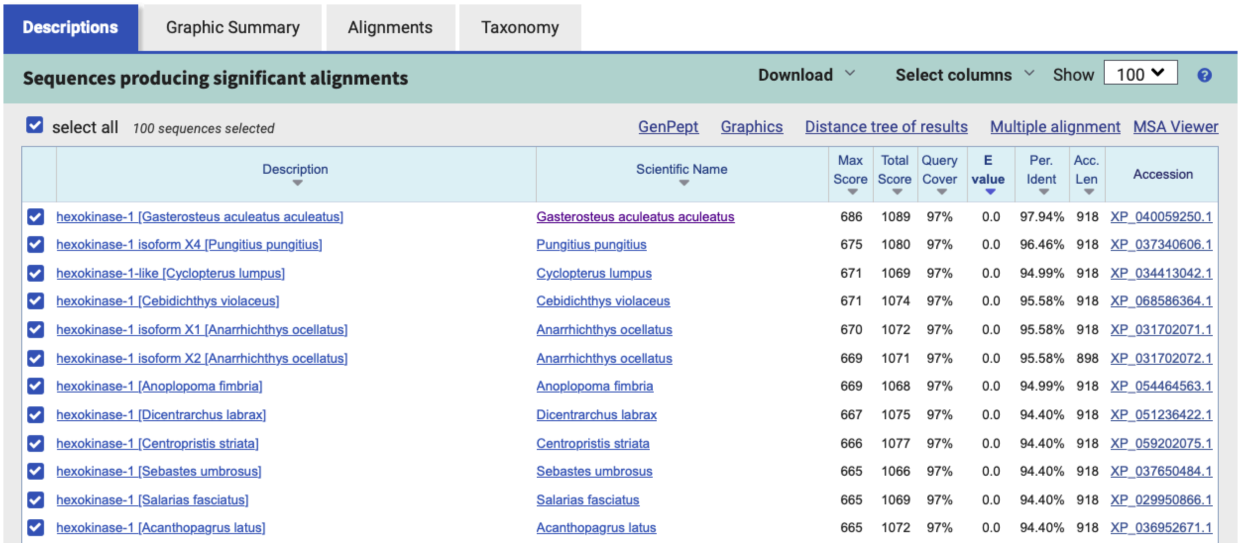
Exclude (Optional): ☐ Models (XM/XP) ☐ Non-redundant RefSeq proteins (WP) ☐ Uncultured/environmental sample sequences

Program Selection

Algorithm:
○ Quick BLASTP (Accelerated protein-protein BLAST)
● 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

Search **database nr** using **Blastp (protein-protein BLAST)**
☐ Show results in a new window

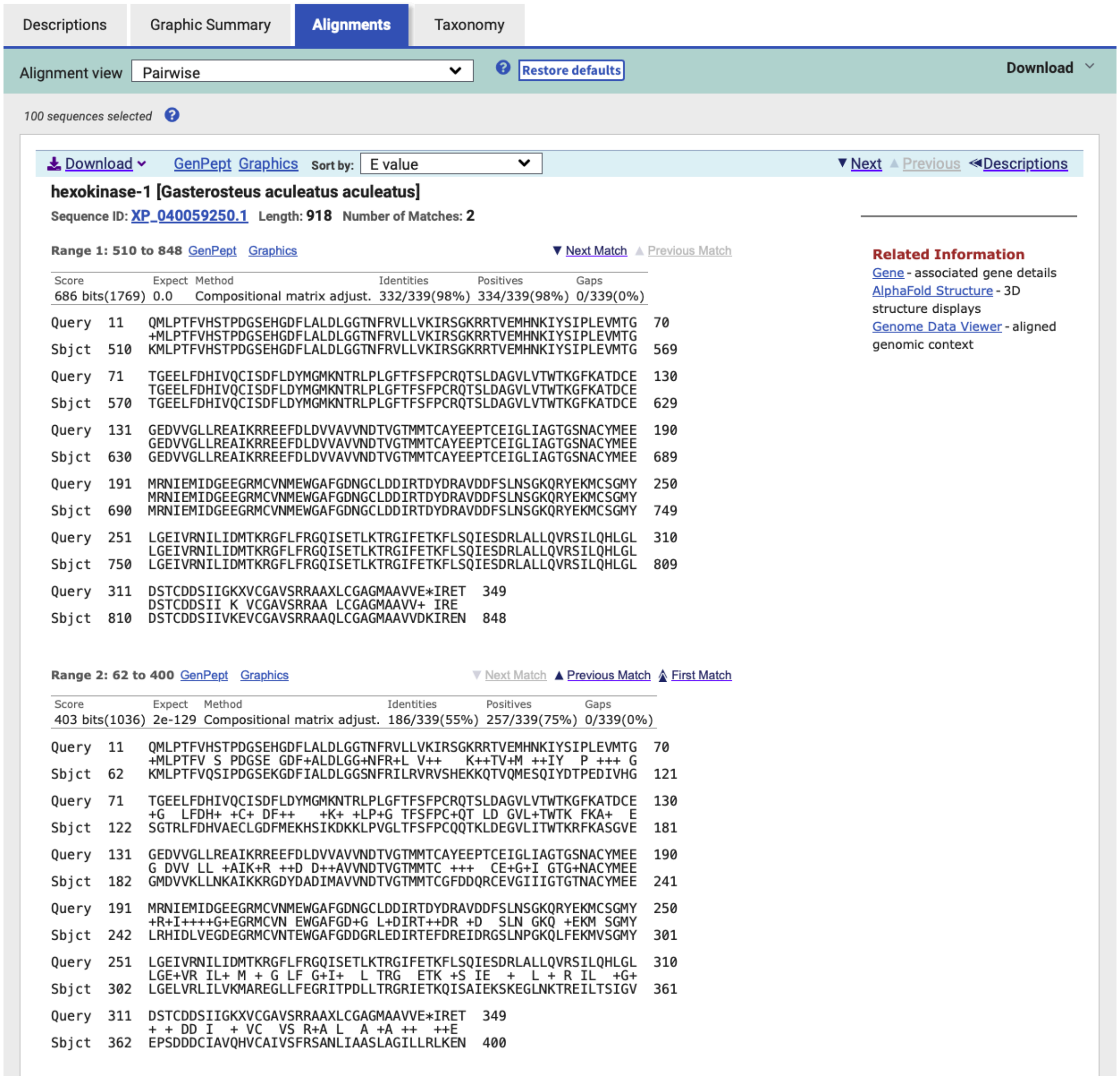
The top result is to a protein from *Gasterosteus aculeatus aculeatus* (three-spined stickleback), see second screen shot below for alignment details:



Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	hexokinase-1 [Gasterosteus aculeatus aculeatus]	Gasterosteus aculeatus aculeatus	686	1089	97%	0.0	97.94%	918	XP_040059250.1
☑	hexokinase-1 isoform X4 [Pungitius pungitius]	Pungitius pungitius	675	1080	97%	0.0	96.46%	918	XP_037340606.1
☑	hexokinase-1-like [Cyclopterus lumpus]	Cyclopterus lumpus	671	1069	97%	0.0	94.99%	918	XP_034413042.1
☑	hexokinase-1 [Cebidichthys violaceus]	Cebidichthys violaceus	671	1074	97%	0.0	95.58%	918	XP_068586364.1
☑	hexokinase-1 isoform X1 [Anarrhichthys ocellatus]	Anarrhichthys ocellatus	670	1072	97%	0.0	95.58%	918	XP_031702071.1
☑	hexokinase-1 isoform X2 [Anarrhichthys ocellatus]	Anarrhichthys ocellatus	669	1071	97%	0.0	95.58%	898	XP_031702072.1
☑	hexokinase-1 [Anoplopoma fimbria]	Anoplopoma fimbria	669	1068	97%	0.0	94.99%	918	XP_054464563.1
☑	hexokinase-1 [Dicentrarchus labrax]	Dicentrarchus labrax	667	1075	97%	0.0	94.40%	918	XP_051236422.1
☑	hexokinase-1 [Centropristis striata]	Centropristis striata	666	1077	97%	0.0	94.40%	918	XP_059202075.1
☑	hexokinase-1 [Sebastes umbrosus]	Sebastes umbrosus	665	1066	97%	0.0	94.40%	918	XP_037650484.1
☑	hexokinase-1 [Salarias fasciatus]	Salarias fasciatus	665	1069	97%	0.0	94.40%	918	XP_029950866.1
☑	hexokinase-1 [Acanthopagrus latus]	Acanthopagrus latus	665	1072	97%	0.0	94.40%	918	XP_036952671.1



100 sequences selected

hexokinase-1 [Gasterosteus aculeatus aculeatus]

Sequence ID: XP_040059250.1 Length: 918 Number of Matches: 2

Range 1: 510 to 848 GenPept Graphics ▼ Next Match ▲ Previous Match

```
Score           Expect Method                      Identities     Positives      Gaps
686 bits(1769)  0.0    Compositional matrix adjust. 332/339(98%)  334/339(98%)  0/339(0%)

Query  11   QMLPTFVHSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKIYSIPLEVMTG  70
            +MLPTFVHSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKIYSIPLEVMTG
Sbjct  510  KMLPTFVHSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKIYSIPLEVMTG  569

Query  71   TGEELFDHIVQCISDFLDYMGMKNTRLPLGFTFSFPCRQTSLDAGVLVTWTKGFKATDCE  130
            TGEELFDHIVQCISDFLDYMGMKNTRLPLGFTFSFPCRQTSLDAGVLVTWTKGFKATDCE
Sbjct  570  TGEELFDHIVQCISDFLDYMGMKNTRLPLGFTFSFPCRQTSLDAGVLVTWTKGFKATDCE  629

Query  131  GEDVVGLLREAIKRREEFDLDVVAVVNDTVGTMMTCAYEEPTCEIGLIAGTGSNACYMEE  190
            GEDVVGLLREAIKRREEFDLDVVAVVNDTVGTMMTCAYEEPTCEIGLIAGTGSNACYMEE
Sbjct  630  GEDVVGLLREAIKRREEFDLDVVAVVNDTVGTMMTCAYEEPTCEIGLIAGTGSNACYMEE  689

Query  191  MRNIEMIDGEEGRMCVNMEWGAFGDNGCLDDIRTDYDRAVDDFSLNSGKQRYEKMCSGMY  250
            MRNIEMIDGEEGRMCVNMEWGAFGDNGCLDDIRTDYDRAVDDFSLNSGKQRYEKMCSGMY
Sbjct  690  MRNIEMIDGEEGRMCVNMEWGAFGDNGCLDDIRTDYDRAVDDFSLNSGKQRYEKMCSGMY  749

Query  251  LGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSQIESDRLALLQVRSILQHLGL  310
            LGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSQIESDRLALLQVRSILQHLGL
Sbjct  750  LGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSQIESDRLALLQVRSILQHLGL  809

Query  311  DSTCDDSIIGKXVCGAVSRRAAXLCGAGMAAVVE*IRET  349
            DSTCDDSII K VCGAVSRRAA LCGAGMAAVV+ IRE
Sbjct  810  DSTCDDSIIVKEVCGAVSRRAAQLCGAGMAAVVDKIREN  848
```

Range 2: 62 to 400 GenPept Graphics ▼ Next Match ▲ Previous Match ⩓ First Match

```
Score           Expect Method                       Identities    Positives     Gaps
403 bits(1036)  2e-129 Compositional matrix adjust. 186/339(55%)  257/339(75%)  0/339(0%)

Query  11   QMLPTFVHSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKIYSIPLEVMTG  70
            +MLPTFV S PDGSE GDF+ALDLGG+NFR+L V++   K++TV+M ++IY  P +++ G
Sbjct  62   KMLPTFVQSIPDGSEKGDFIALDLGGSNFRILRVRVSHEKKQTVQMESQIYDTPEDIVHG  121

Query  71   TGEELFDHIVQCISDFLDYMGMKNTRLPLGFTFSFPCRQTSLDAGVLVTWTKGFKATDCE  130
            +G  LFDH+ +C+ DF++   +K+ +LP+G TFSFPC+QT LD GVL+TWTK FKA+  E
Sbjct  122  SGTRLFDHVAECLGDFMEKHSIKDKKLPVGLTFSFPCQQTKLDEGVLITWTKRFKASGVE  181

Query  131  GEDVVGLLREAIKRREEFDLDVVAVVNDTVGTMMTCAYEEPTCEIGLIAGTGSNACYMEE  190
            G DVV LL +AIK+R ++D D++AVVNDTVGTMMTC +++  CE+G+I GTG+NACYMEE
Sbjct  182  GMDVVKLLNKAIKKRGDYDADIMAVVNDTVGTMMTCGFDDQRCEVGIIIGTGTNACYMEE  241

Query  191  MRNIEMIDGEEGRMCVNMEWGAFGDNGCLDDIRTDYDRAVDDFSLNSGKQRYEKMCSGMY  250
            +R+I++++G+EGRMCVN EWGAFGD+G L+DIRT++DR +D  SLN GKQ +EKM SGMY
Sbjct  242  LRHIDLVEGDEGRMCVNTEWGAFGDDGRLEDIRTEFDREIDRGSLNPGKQLFEKMVSGMY  301

Query  251  LGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSQIESDRLALLQVRSILQHLGL  310
            LGE+VR IL+ M + G LF G+I+  L TRG  ETK +S IE  +  L + R IL  +G+
Sbjct  302  LGELVRLILVKMAREGLLFEGRITPDLLTRGRIETKQISAIEKSKEGLNKTREILTSIGV  361

Query  311  DSTCDDSIIGKXVCGAVSRRAAXLCGAGMAAVVE*IRET  349
            + + DD I  + VC  VS R+A L  A +A ++  ++E
Sbjct  362  EPSDDDCIAVQHVCAIVSFRSANLIAASLAGILLRLKEN  400
```

Related Information

Gene - associated gene details

AlphaFold Structure - 3D structure displays

Genome Data Viewer - aligned genomic context