Marine Genomics

Bioinformatic Protocol

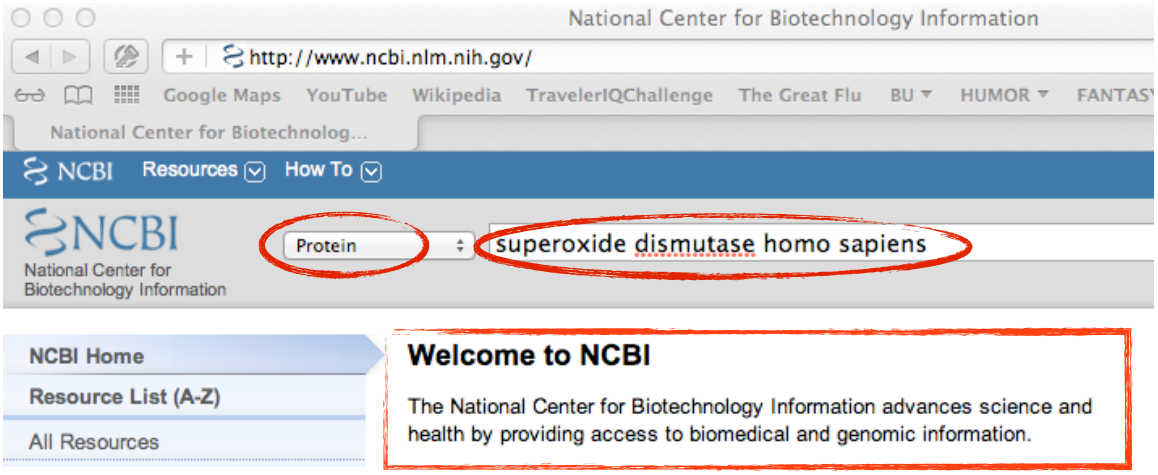
Identify a gene of interest in a "non-model" system

(e.g., *Nematostella*)

by John R. Finnerty

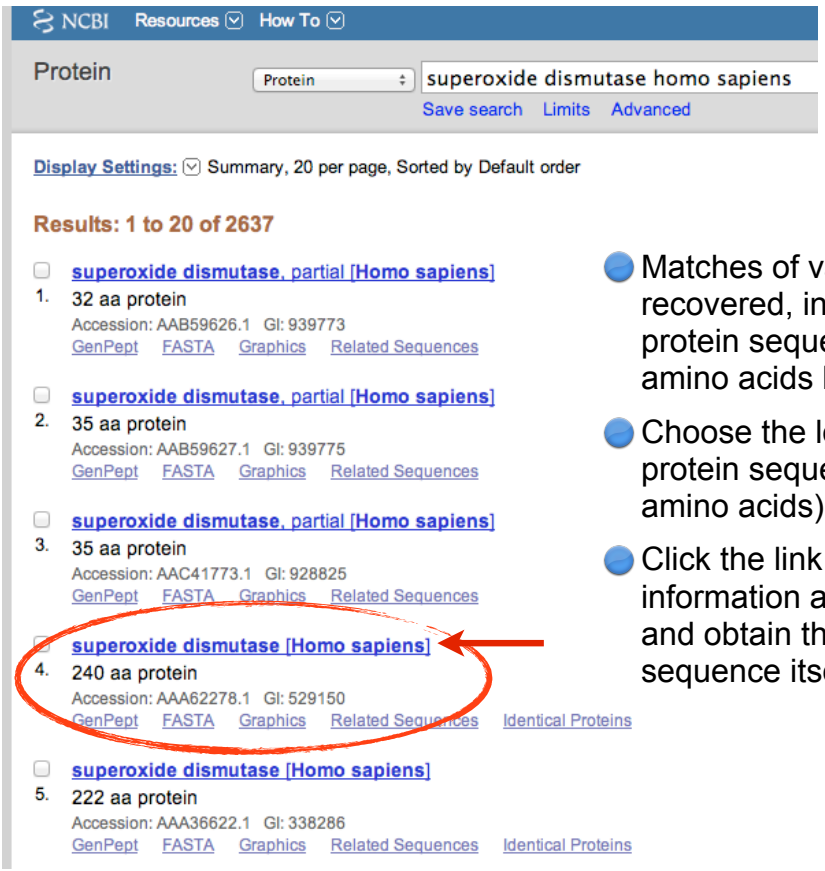
- The recent profusion of genome and transcriptome sequencing projects for marine organisms has generated an enormous amounts of DNA/RNA sequence data.
- However, these DNA/RNA sequences are generally not well "annotated." In other words, the individual genes have generally not been subjected to sufficient analysis to identify them by function or even to give them a name.
- If you want to identify a particular gene in an organism for which a well-annotated genome does not exist, it is often necessary (1) to first obtain the comparable (i.e., homologous) gene from an organism that has a well-annotated genome and (2) to search through the DNA/RNA sequences of the less well-characterized species to find a gene with a similar sequence that might be its homolog.
- In the example that follows: A researcher is seeking to determine how oxidative stress alters the gene expression of the starlet sea anemone *Nematostella vectensis*. In other well-studied animals, it is known that the anti-oxidant enzyme superoxide dismutase is up-regulated under conditions of oxidative stress.
- Does *Nematostella* have a gene for superoxide dismutase?
- To answer this question, we first obtain the gene from a well-annotated genome, such as the human genome.

- On any web browser, connect to the homepage of the **National Center for Biotechnology Information** (or NCBI; http://www.ncbi.nlm.nih.gov/).



What a wonderful use for our tax dollars!

- Select the "Protein" database, and type the terms "superoxide dismutase homo sapiens" into the search window.
- Hit Search



- Matches of various lengths are recovered, including "partial" protein sequences only 32 amino acids long.
- Choose the longest predicted protein sequence available (240 amino acids).
- Click the link to see additional information about this protein and obtain the amino acid sequence itself.

- The top of the page provides a locus ID number, also called an “accession number,” as well as information on any publications that are associated with the sequence.

superoxide dismutase [Homo sapiens]

GenBank: AAA62278.1

FASTA Graphics

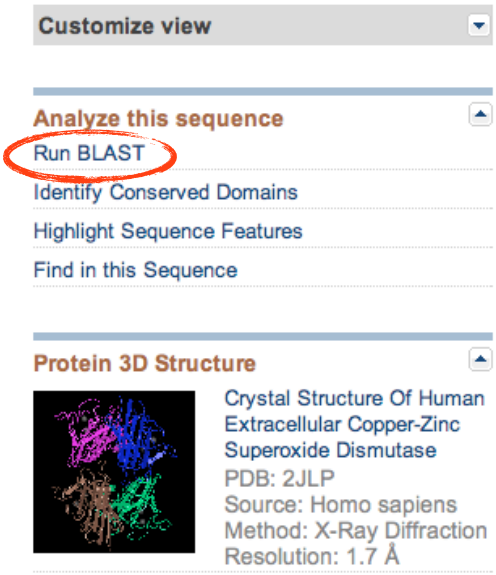
Go to:

```
LOCUS       AAA62278                 240 aa            linear   PRI 18-FEB-1995
DEFINITION  superoxide dismutase [Homo sapiens].
ACCESSION   AAA62278
VERSION     AAA62278.1  GI:529150
DBSOURCE    locus HSU10116 accession U10116.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 240)
  AUTHORS   Folz,R.J. and Crapo,J.D.
  TITLE     Extracellular superoxide dismutase (SOD3): tissue-specific
            expression, genomic characterization, and computer-assisted
            sequence analysis of the human EC SOD gene
  JOURNAL   Genomics 22 (1), 162-171 (1994)
```

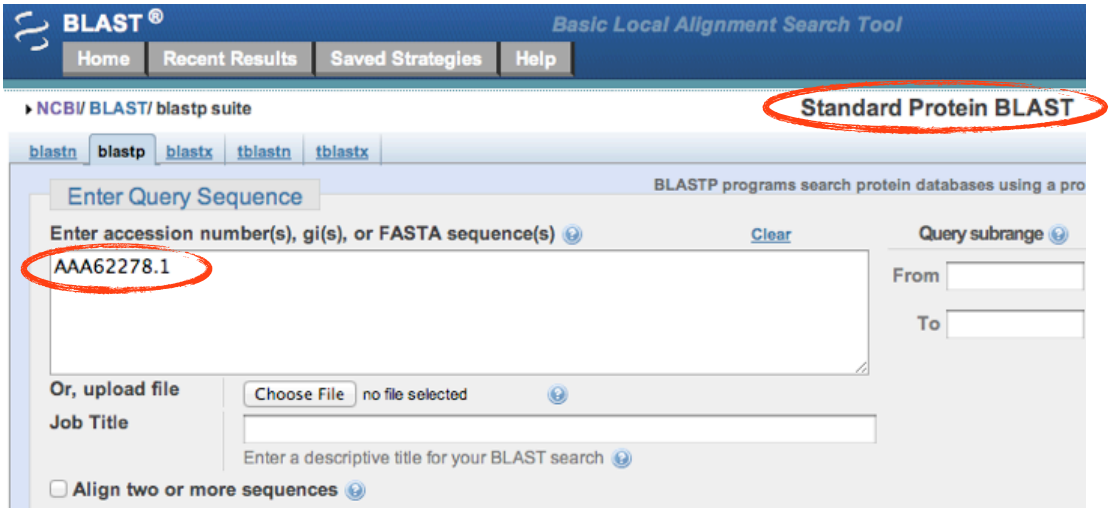
```
     Region          69..207
                     /region_name="Cu-Zn_Superoxide_Dismutase"
                     /note="Copper/zinc superoxide dismutase (SOD). superoxide
                     dismutases catalyse the conversion of superoxide radicals
                     to molecular oxygen. Three evolutionarily distinct
                     families of SODs are known, of which the
                     copper/zinc-binding family is one. Defects in the...;
                     cd00305"
                     /db_xref="CDD:48338"
     Site            order(70,81,83,118..120,174..175)
                     /site_type="other"
                     /note="E-class dimer interface [polypeptide binding]"
                     /db_xref="CDD:48338"
     Site            order(92,150)
                     /site_type="other"
                     /note="P-class dimer interface [polypeptide binding]"
                     /db_xref="CDD:48338"
     Site            order(114,116,131,142,145,181)
                     /site_type="active"
                     /db_xref="CDD:48338"
     Site            order(114,116,131,181)
                     /site_type="other"
                     /note="Cu2+ binding site [ion binding]"
                     /db_xref="CDD:48338"
     Site            order(131,139,142,145)
                     /site_type="other"
                     /note="Zn2+ binding site [ion binding]"
                     /db_xref="CDD:48338"
     CDS             1..240
                     /gene="SOD3"
                     /coded_by="U10116.1:5085..5807"
ORIGIN
        1 mlallcscll laagasdawt gedsaepnsd saewirdmya kvteiwqevm qrrdddgtlh
       61 aacqvqpsat ldaaqprvtg vvlfrqlapr akldaffale gfptepnsss raihvhqfgd
      121 lsqgcestgp hynplavphp qhpgdfgnfa vrdgslwryr aglaaslagp hsivgravvv
      181 hageddlgrg gnqasvengn agrrlaccvv gvcgpglwer qarehserkk rrreseckaa
//
```

- The bottom of the page lists conserved regions or sites within the protein, and characterizes their known function (e.g., “polypeptide binding”, “ion binding”).
- The full protein sequence is also given.

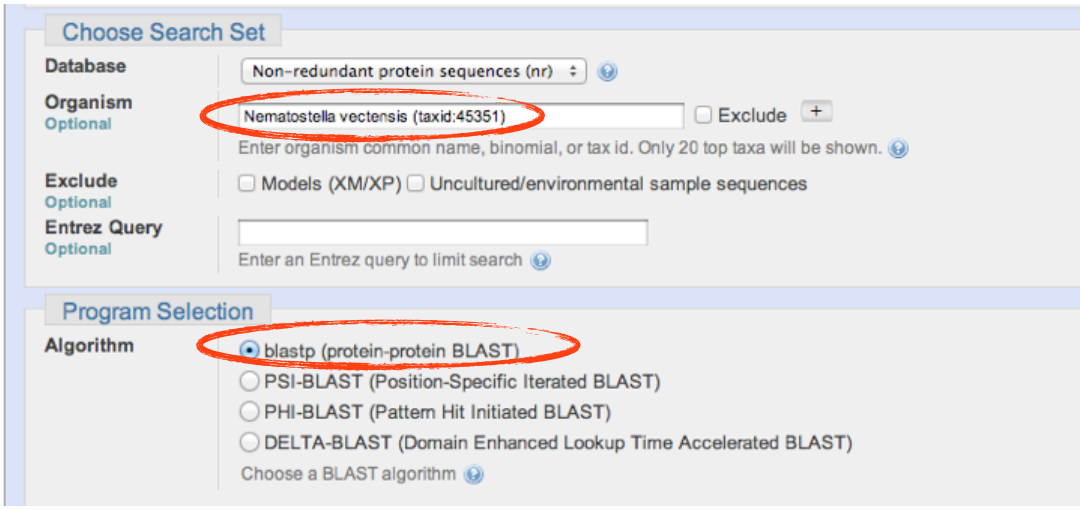
- At the top of the page, on the far right, there is a link to Run **BLAST** (Basic Local Alignment Search Tool). This will search the database for similar sequences, including sequences in other species.



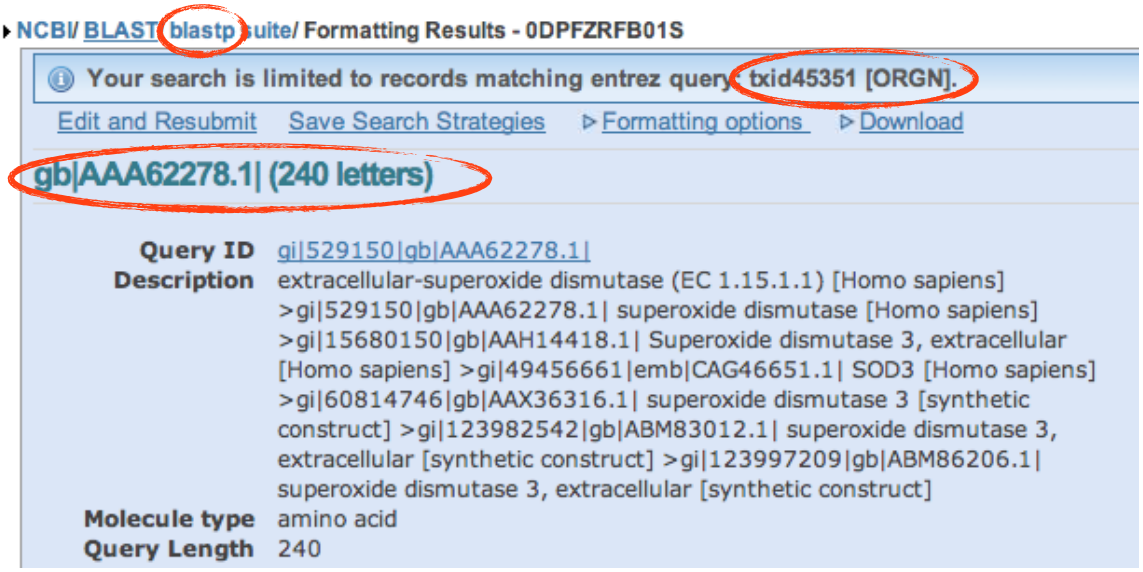
- Clicking this link takes you to a “Standard Protein BLAST.”
- The amino acid sequence of the human superoxide dismutase protein is automatically specified as the “Query Sequence” (notice that its accession number, AAA62278.1, is already pasted into the “Enter Query Sequence” box).



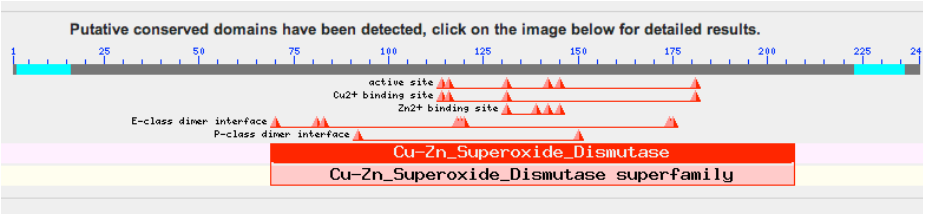
- To look for the closest match to this sequence in the starlet sea anemone, *Nematostella vectensis*, you need to restrict the "Search Set" by Organism by entering Nematostella vectensis. The database will automatically supply the taxonomic identification number for this species (taxid:45351).
- You can use the default search program, "blastp." This performs a standard protein BLAST search.
- Hit **BLAST**



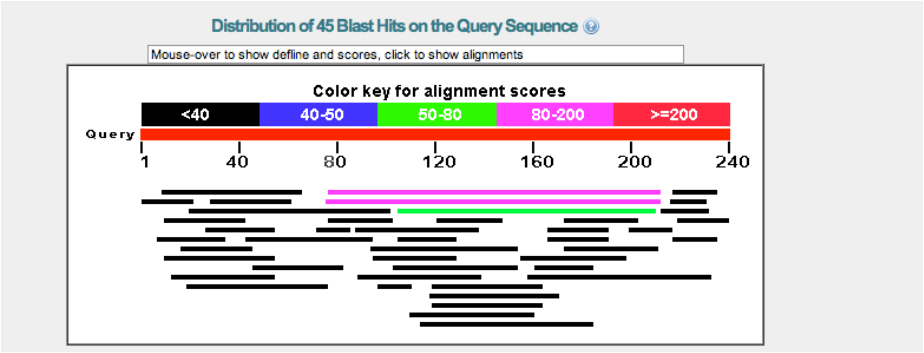
- The top of the **BLAST output page** describes:
  - the type of blast (blastp)
  - the query sequence you used (AAA62278.1), and
  - any restrictions you placed on the search (txid43451) .



- The **Graphic Summary** reveals where functionally characterized domains reside in the 240 amino acid query sequence (e.g., Cu2+ binding site).



- The **Distribution of 45 Blast Hits on the Query Sequence** reveals graphically how various *Nematostella* sequences exhibit similarity to different regions of the query sequence. A higher alignment score indicates a better match.



- The **Descriptions** section list all BLAST hits beginning with those that are most similar to the query sequence (those with the highest "Total score.")
- The **Expect value (E value)** expresses the probability that a degree of similarity as great as was observed could have occurred by chance.

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
XP_001634103.1	predicted protein [Nematostella vectensis] >gb|EDO42040.1| pr	99.0	99.0	56%	1e-27	41%	UG
XP_001634104.1	predicted protein [Nematostella vectensis] >gb|EDO42041.1| pr	95.1	95.1	56%	3e-26	41%	UG
XP_001635982.1	predicted protein [Nematostella vectensis] >gb|EDO43919.1| pr	75.5	75.5	43%	1e-19	39%	UG

- The **Alignment** section reveals how the amino acid sequence of the "Sbjct" from Nematostella aligns to the "Query" from human .
- Click the link to view more information on the *Nematostella* predicted protein.

```
>ref|XP_001634104.1| UG predicted protein [Nematostella vectensis]
 gb|EDO42041.1| G predicted protein [Nematostella vectensis]
Length=154

 GENE ID: 5513890 NEMVEDRAFT_v1g234825 | hypothetical protein
[Nematostella vectensis] (10 or fewer PubMed links)

 Score = 95.1 bits (235),  Expect = 3e-26, Method: Compositional matrix adjust.
 Identities = 59/143 (41%), Positives = 74/143 (52%), Gaps = 11/143 (8%)

Query  77   RVTGVVLFRQLAPRAKLDAFFALEGFPTEPNSSSRAIHVHQFGDLSQGCESTGPHYNPLA  136
             V GV+ F Q AP    D    L G  T         H+H+FGD + GC S G HYNP
Sbjct  14   EVKGVIHFTQQAP----DGPCTLRGRITGLTEGKHGFHIHEFGDNTNGCTSAGAHYNPHG  69

Query  137  VPH------PQHPGDFGNF-AVRDGSLWRYRAGLAASLAGPHSIVGRAVVVHAGEDDLGR  189
              H       +H GD GN  A  +G            SL G  SI+GR++VVH G DDLG
Sbjct  70   KMHGAPEDKDRHLGDLGNIEADANGIADVSITDCLVSLTGQCSIIGRSLVVHEGMDDLGA  129

Query  190  GGNQASVENGNAGRRLACCVVGV  212
            GG++ S+  GNAG R+AC V+G+
Sbjct  130  GGHELSLTTGNAGGRVACGVIGI  152
```

- This predicted protein was deposited in NCBI at the time the sequenced genome of *Nematostella* was published in the journal *Science* in 2007.
- Click the **Run BLAST** link to search the human genome using this *Nematostella* sequence as a query. This is a "reciprocal BLAST."

predicted protein [Nematostella vectensis]

NCBI Reference Sequence: XP_001634104.1

FASTA Graphics

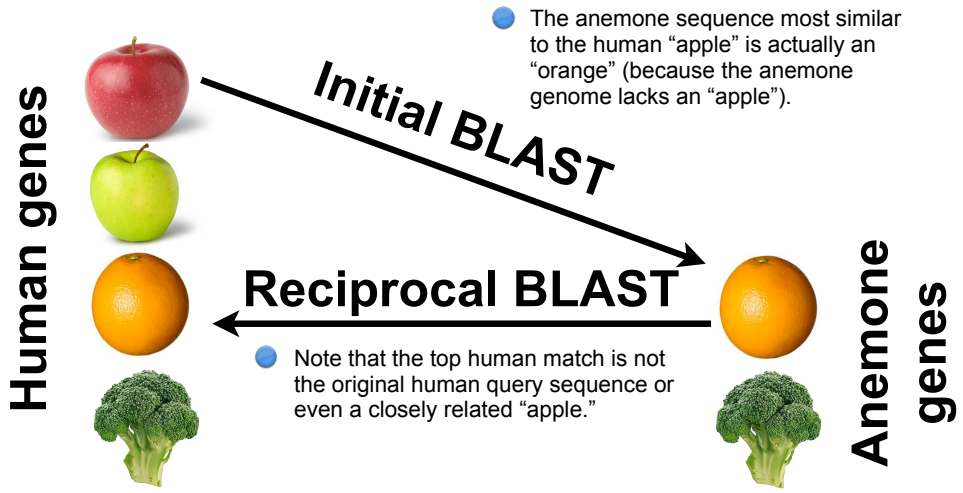
Go to:

```
LOCUS       XP_001634104             154 aa            linear   INV 27-AUG-2009
DEFINITION  predicted protein [Nematostella vectensis].
ACCESSION   XP_001634104
VERSION     XP_001634104.1  GI:156386810
DBSOURCE    REFSEQ: accession XM_001634054.1
KEYWORDS    .
SOURCE      Nematostella vectensis (starlet sea anemone)
  ORGANISM  Nematostella vectensis
            Eukaryota; Metazoa; Cnidaria; Anthozoa; Hexacorallia; Actiniaria;
            Edwardsiidae; Nematostella.
REFERENCE   1  (residues 1 to 154)
  AUTHORS   Putnam,N.H., Srivastava,M., Hellsten,U., Dirks,B., Chapman,J.,
            Salamov,A., Terry,A., Shapiro,H., Lindquist,E., Kapitonov,V.V.,
            Jurka,J., Genikhovich,G., Grigoriev,I.V., Lucas,S.M., Steele,R.E.,
            Finnerty,J.R., Technau,U., Martindale,M.Q. and Rokhsar,D.S.
  TITLE     Sea anemone genome reveals ancestral eumetazoan gene repertoire and
            genomic organization
  JOURNAL   Science 317 (5834), 86-94 (2007)
```

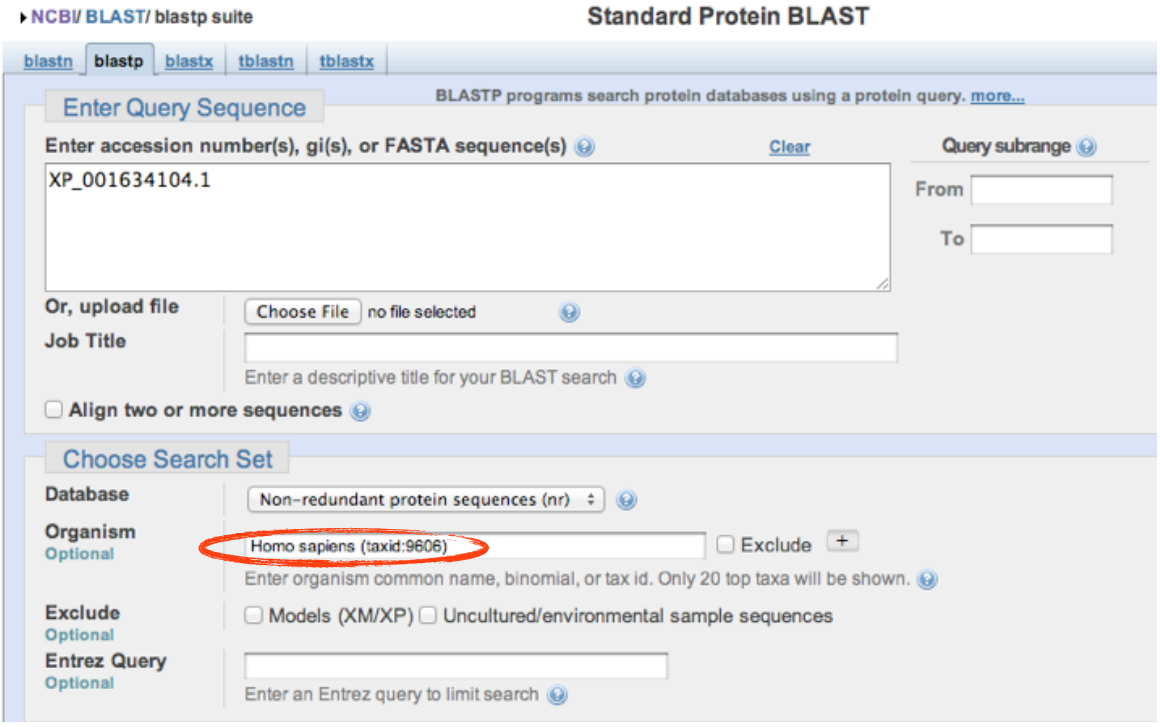
Why perform a reciprocal BLAST?

Haven't I already found the gene I'm looking for in *Nematostella*?

- When you BLASTed a human sequence against all the *Nematostella* sequences, you found the best *Nematostella* match to the human sequence, but it may not be the comparable gene if the sea anemone genome lacks that gene.
- If *Nematostella* lacks that gene, the reciprocal BLAST will reveal that we're "comparing apples and oranges."



- To look for the closest match to this sequence in the human, you need to restrict the “Search Set” by Organism by entering Homo sapiens. The database will supply the taxonomic identification number (taxid:9606).

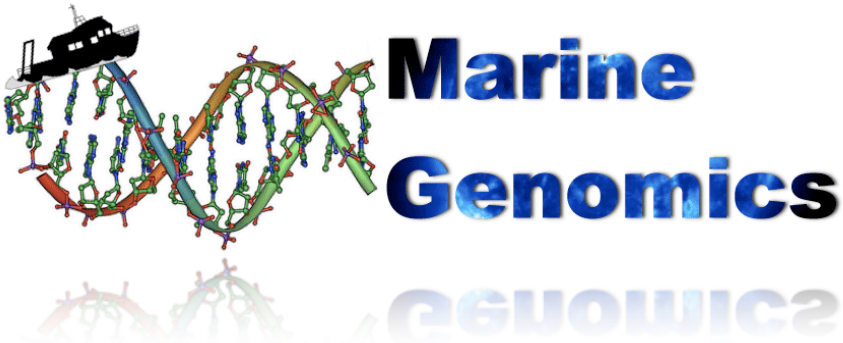


- The top match is a mutant form of human superoxide dismutase.

```
>pdb|1N18|A  S  Chain A, Thermostable Mutant Of Human Superoxide Dismutase, C6a,
C111s
 pdb|1N18|B  S  Chain B, Thermostable Mutant Of Human Superoxide Dismutase, C6a,
C111s
 pdb|1N18|C  S  Chain C, Thermostable Mutant Of Human Superoxide Dismutase, C6a,
C111s
 ▷7 more sequence titles
 Length=154

 Score =  197 bits (501),  Expect = 9e-67, Method: Compositional matrix adjust.
 Identities = 94/153 (61%), Positives = 117/153 (76%), Gaps = 0/153 (0%)

Query  1    MVIRGVCCLVGDNEVKGVIHFTQQAPDGPCTLRGRITGLTEGKHGFHIHEFGDNTNGCTS  60
            M  + V  L GD  V+G+I+F Q+  +GP  + G I GLTEG HGFH+HEFGDNT GCTS
Sbjct  1    MATKAVAVLKGDGPVQGIINFEQKESNGPVKVWGSIKGLTEGLHGFHVHEFGDNTAGCTS  60

Query  61   AGAHYNPHGKMHGAPEDKDRHLGDLGNIEADANGIADVSITDCLVSLTGQCSIIGRSLVV  120
            AG H+NP  + HG P+D++RH+GDLGN+ AD +G+ADVSI D ++SL+G  SIIGR+LVV
Sbjct  61   AGPHFNPLSRKHGGPKDEERHVGDLGNVTADKDGVADVSIEDSVISLSGDHSIIGRTLVV  120

Query  121  HEGMDDLGAGGHELSLTTGNAGGRVACGVIGIA  153
            HE  DDLG GG+E S  TGNAG R+ACGVIGIA
Sbjct  121  HEKADDLGKGGNEESTKTGNAGSRLACGVIGIA  153
```

- The second match, almost equally good as the top match, is a “wild-type” human superoxide dismutase.
- As the original query was “superoxide dismutase” and the reciprocal blast recovered superoxide dismutase, we can consider the *Nematostella* sequence to be superoxide dismutase.

```
>ref|NP_000445.1|  superoxide dismutase [Cu-Zn] [Homo sapiens]
 sp|P00441.2|SODC_HUMAN  RecName: Full=Superoxide dismutase [Cu-Zn]; AltName:
dismutase 1; Short=hSod1
 emb|CAA26182.1|  unnamed protein product [Homo sapiens]
 ▶12 more sequence titles
 Length=154

 GENE ID: 6647 SOD1 | superoxide dismutase 1, soluble [Homo sapiens]
(Over 100 PubMed links)

 Score =  196 bits (497),  Expect = 3e-66, Method: Compositional matrix adjust.
 Identities = 94/153 (61%), Positives = 117/153 (76%), Gaps = 0/153 (0%)

Query  1    MVIRGVCCLVGDNEVKGVIHFTQQAPDGPCTLRGRITGLTEGKHGFHIHEFGDNTNGCTS  60
            M  + VC L GD  V+G+I+F Q+  +GP  + G I GLTEG HGFH+HEFGDNT GCTS
Sbjct  1    MATKAVCVLKGDGPVQGIINFEQKESNGPVKVWGSIKGLTEGLHGFHVHEFGDNTAGCTS  60

Query  61   AGAHYNPHGKMHGAPEDKDRHLGDLGNIEADANGIADVSITDCLVSLTGQCSIIGRSLVV  120
            AG H+NP  + HG P+D++RH+GDLGN+ AD +G+ADVSI D ++SL+G   IIGR+LVV
Sbjct  61   AGPHFNPLSRKHGGPKDEERHVGDLGNVTADKDGVADVSIEDSVISLSGDHCIIGRTLVV  120

Query  121  HEGMDDLGAGGHELSLTTGNAGGRVACGVIGIA  153
            HE  DDLG GG+E S  TGNAG R+ACGVIGIA
Sbjct  121  HEKADDLGKGGNEESTKTGNAGSRLACGVIGIA  153
```



Unpublished Cnidarian Sequence Databases

Database	Species	Sequences	Biological Source Material	Sequencing methodology	Database URL / Reference
StellaBase v. 1	*Nematostella vectensis* (starlet sea anemone)	genomic DNA raw reads; assembled genome; Expressed Sequence Tags; assembled transcripts;	larva from the cross of two Maryland individiulas	Sanger sequencing (Putnam et al. 2007)	stellabase.org (Sullivan et al. 2006)
StellaBase v. 2 (encompasses v. 1 in addition to new data sources)	*Nematostella vectensis* (starlet sea anemone)	+ mRNA raw reads & assembled transcripts	+ intact adults & regenerating adults from New Jersey	+ next generation sequencing (Illumina platform)	Finnerty Lab server (Lubinski et al., unpublished)
PocilloporaBase	*Pocillopora damicornis* (cauliflower coral)	mRNA raw reads & assembled transcripts	adult colonies from 3 different sites in Oahu, Hawaii exposed to various stressors	next generation sequencing (454 platform)	pocilloporabase.org (Traylor-Knowles et al. 2011)
EdwardsiellaBase	*Edwardsiella lineata* (lined sea anemone)	mRNA raw reads & assembled transcripts	adult polyps, larvae, parasites, larva-to-adult transition; parasite-to-larva transition	next generation sequencing (Illumina platform)	Edwardsiellabase.org Stefanik et al., in Review
CorallimorphBase (doesn't exist as a free-standing site yet; sequences are searchable via internal version of EdBase)	*Actinodiscus* species	mRNA raw reads & assembled transcripts	adults exposed to hyposaline, hypersaline, or control conditions for 3 or 27 hours	next generation sequencing (Illumina platform)	Finnerty Lab server (Granger et al., unpublished)