

Bioinformatic Protocol

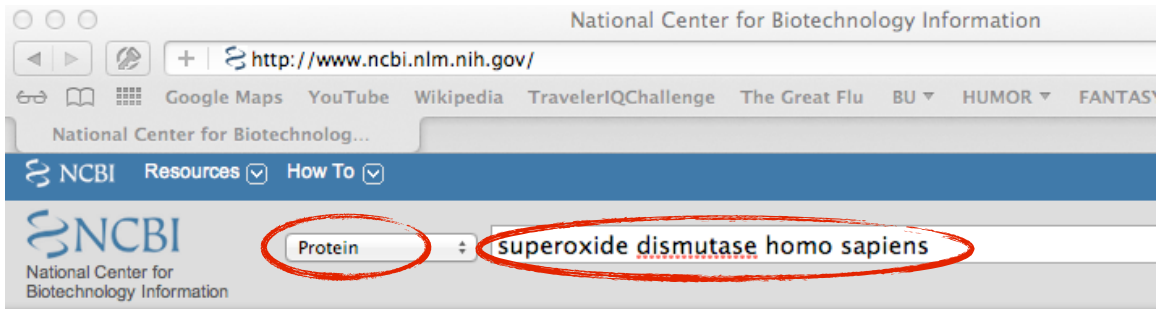
# Identify a gene of interest in a “non-model” system

(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/>).



NCBI Home  
Resource List (A-Z)  
All Resources

**Welcome to NCBI**  
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

**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 

Protein Protein superoxide dismutase homo sapiens  
Save search Limits Advanced

Display Settings: Summary, 20 per page, Sorted by Default order

Results: 1 to 20 of 2637

- [superoxide dismutase, partial \[Homo sapiens\]](#)  
1. 32 aa protein  
Accession: AAB59626.1 GI: 939773  
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#)
- [superoxide dismutase, partial \[Homo sapiens\]](#)  
2. 35 aa protein  
Accession: AAB59627.1 GI: 939775  
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#)
- [superoxide dismutase, partial \[Homo sapiens\]](#)  
3. 35 aa protein  
Accession: AAC41773.1 GI: 928825  
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#)
- [superoxide dismutase \[Homo sapiens\]](#)  
4. 240 aa protein  
Accession: AAA62278.1 GI: 529150  
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)
- [superoxide dismutase \[Homo sapiens\]](#)  
5. 222 aa protein  
Accession: AAA36622.1 GI: 338286  
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)

- 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:](#)

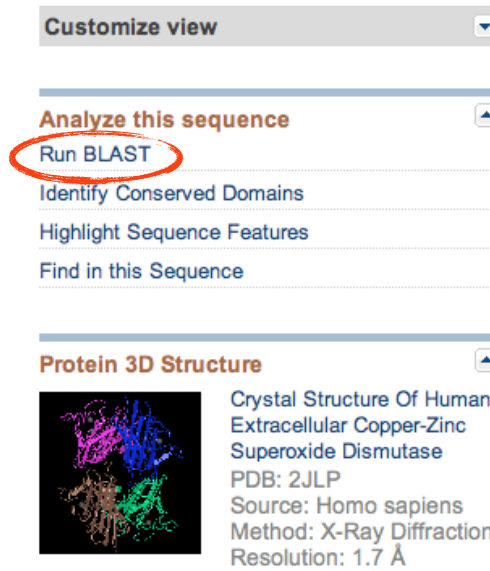
LOCUS **AAA62278** 240 aa linear PRI 18-FEB-1995  
 DEFINITION superoxide dismutase [Homo sapiens].  
 ACCESSION AAA62278  
 VERSION AAA62278.1 GI:529150  
 DBSOURCE locus HSU10116 accession [U10116.1](#)  
 KEYWORDS .  
 SOURCE Homo sapiens (human)  
 ORGANISM [Homo sapiens](#)  
 Eukaryota; Metazoa; 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="Cu<sup>2+</sup> binding site [ion binding]"  
 /db\_xref="CDD:[48338](#)"  
[Site](#) order(131,139,142,145)  
 /site\_type="other"  
 /note="Zn<sup>2+</sup> binding site [ion binding]"  
 /db\_xref="CDD:[48338](#)"  
[CDS](#) 1..240  
 /gene="SOD3"  
 /coded\_by="U10116.1:5085..5807"

ORIGIN  
 1 mlallcsc1l laagasdwt gedsaepnsd saewirdmya kvteiwqevm qrrddgqtlh  
 61 aacqvqpsat ldaaqrvtg vvlfrqlapr akldaffale gfptepnsss raihvhqfgd  
 121 lsqgcestgp hynplavphp qhpgdfgnfa vrdgslwryr aglaaslagp hsivgravvv  
 181 hageddligrq gnqasvengn agrrlaccvv gvcgpglwer qarehserkk rreseckaa  
 //

- 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.

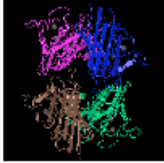


Customize view

Analyze this sequence

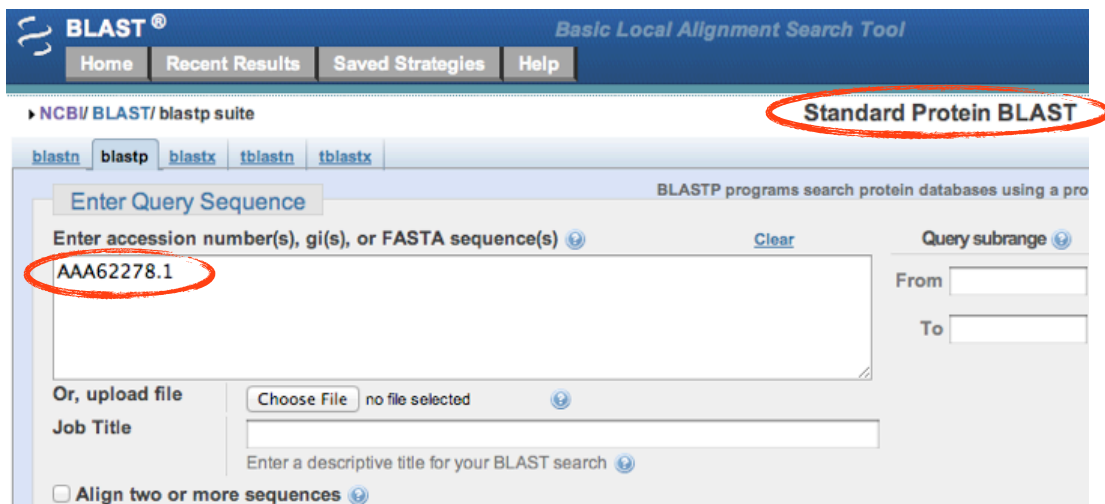
- Run BLAST
- Identify Conserved Domains
- Highlight Sequence Features
- Find in this Sequence

Protein 3D Structure



Crystal Structure Of Human Extracellular Copper-Zinc Superoxide Dismutase  
PDB: 2JLP  
Source: Homo sapiens  
Method: X-Ray Diffraction  
Resolution: 1.7 Å

- 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).



BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/BLAST/blastp suite

blastn blastp blastx tblastn tblastx

Standard Protein BLAST

Enter Query Sequence

BLASTP programs search protein databases using a pro

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

AAA62278.1 From To

Or, upload file Choose File no file selected

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

- 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**

**Choose Search Set**

Database: Non-redundant protein sequences (nr)

Organism: Nematostella vectensis (taxid:45351)  Exclude

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

Entrez Query:

**Program Selection**

Algorithm:  blastp (protein-protein BLAST)

PSI-BLAST (Position-Specific Iterated BLAST)

PHI-BLAST (Pattern Hit Initiated BLAST)

DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

- 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) .

▶ [NCBI/BLAST](#) **blastp** suite/ Formatting Results - 0DPFZRFB01S

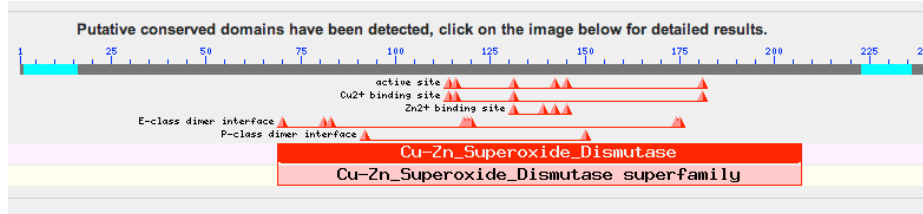
Your search is limited to records matching **entrez query** txid45351 [ORGN].

[Edit and Resubmit](#) [Save Search Strategies](#) ▶ [Formatting options](#) ▶ [Download](#)

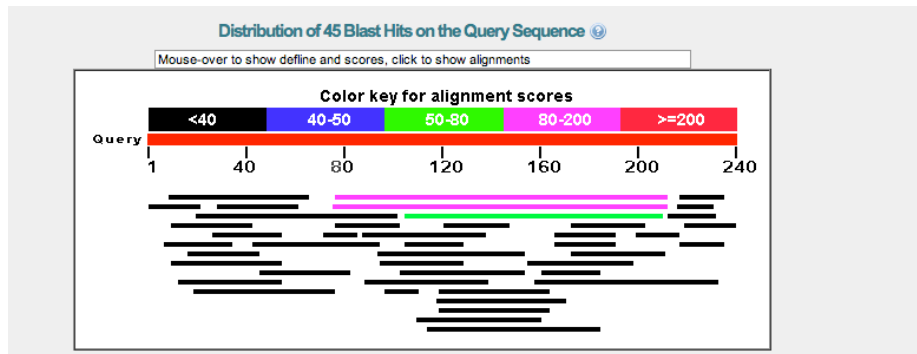
gb|AAA62278.1| (240 letters)

<b>Query ID</b>	<a href="#">gi 529150 gb AAA62278.1 </a>
<b>Description</b>	extracellular-superoxide dismutase (EC 1.15.1.1) [Homo sapiens] >gi 529150 gb AAA62278.1  superoxide dismutase [Homo sapiens] >gi 15680150 gb AAH14418.1  Superoxide dismutase 3, extracellular [Homo sapiens] >gi 49456661 emb CAG46651.1  SOD3 [Homo sapiens] >gi 60814746 gb AAX36316.1  superoxide dismutase 3 [synthetic construct] >gi 123982542 gb ABM83012.1  superoxide dismutase 3, extracellular [synthetic construct] >gi 123997209 gb ABM86206.1  superoxide dismutase 3, extracellular [synthetic construct]
<b>Molecule type</b>	amino acid
<b>Query Length</b>	240

- The **Graphic Summary** reveals where functionally characterized domains reside in the 240 amino acid query sequence (e.g., Cu<sup>2+</sup> 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
<a href="#">XP_001634103.1</a>	predicted protein [Nematostella vectensis] >gb EDO42040.1  pi	99.0	99.0	56%	1e-27	41%	<a href="#">UG</a>
<a href="#">XP_001634104.1</a>	predicted protein [Nematostella vectensis] >gb EDO42041.1  pi	95.1	95.1	56%	3e-26	41%	<a href="#">UG</a>
<a href="#">XP_001635982.1</a>	predicted protein [Nematostella vectensis] >gb EDO43919.1  pi	75.5	75.5	43%	1e-19	39%	<a href="#">UG</a>

- 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| UG predicted protein [Nematostella vectensis]
  Length=154

  GENE ID: 5513890 NEMVEDRAFT_vlq234825 | 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  RVTGVVLFRLAPRAKLDFALEGFPTEPNSSSRRAIHVHQFGDLSQCCESTGPHYNPLA  136
          V  GV+ F Q AP      D   L G T          H+H+FGD + GC S G HYNP
  Sbjct  14  EVKGVIHFTQQAP----DGPCTLRGRITGLTEGKHGFHIEFGDNTNGCTSAGAHYNPHG  69

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

  Query  190  GGNQASVENGNAGRRLACCVGV  212
          GG++ S+  GNAG R+AC V+G+
  Sbjct  130  GGHELSTTGNAGGRVACCVIGI  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](#)

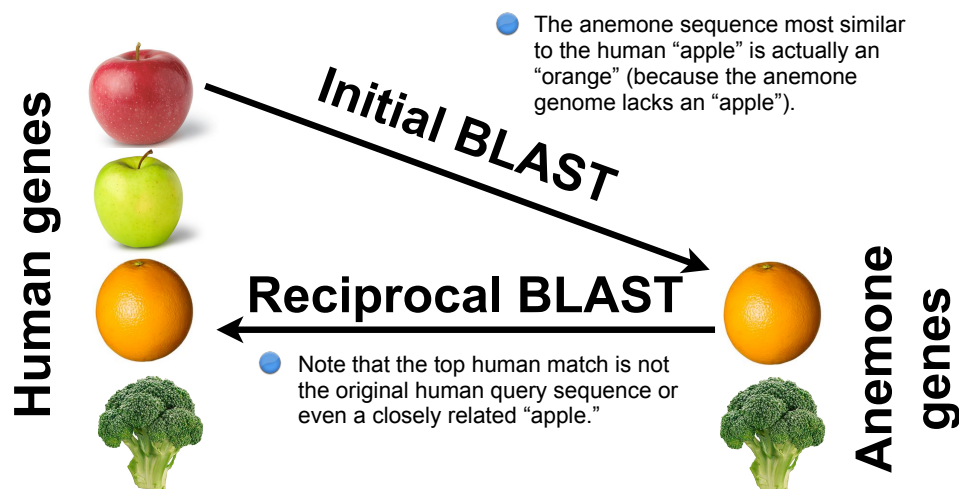
[Go to:](#)

LOCUS	XP_001634104	154 aa	linear	INV 27-AUG-2009
DEFINITION	predicted protein [ <i>Nematostella vectensis</i> ].			
ACCESSION	XP_001634104			
VERSION	XP_001634104.1 GI:156386810			
DBSOURCE	REFSEQ: accession <a href="#">XM_001634054.1</a>			
KEYWORDS	.			
SOURCE	<i>Nematostella vectensis</i> (starlet sea anemone)			
ORGANISM	<a href="#">Nematostella vectensis</a> Eukaryota; Metazoa; Cnidaria; Anthozoa; Hexacorallia; Actiniaria; Edwardsiidae; <i>Nematostella</i> .			
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).

NCBI/BLAST/blastp suite Standard Protein BLAST

blastn **blastp** blastx tblastn tblastx

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

**Enter Query Sequence**

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

XP\_001634104.1

Query subrange

From

To

Or, upload file  no file selected

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

**Choose Search Set**

Database

Organism   Exclude

Optional Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude  Models (XM/XP)  Uncultured/environmental sample sequences

Optional

Entrez Query

Optional Enter an Entrez query to limit search

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

```
> pdb|1N18|A S Chain A, Thermostable Mutant Of Human Superoxide Dismutase, C6a,
C11s
pdb|1N18|B S Chain B, Thermostable Mutant Of Human Superoxide Dismutase, C6a,
C11s
pdb|1N18|C S Chain C, Thermostable Mutant Of Human Superoxide Dismutase, C6a,
C11s
>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 MVIRGVCCLVGDNEVKGVIHFTQQAPDGPCTLRGRITGLTEGKHFHIEFGDNTNGCTS 60
M + V L GD V+G+I+F Q+ +GP + G I GLTEG HGFH+HEFGDNT GCTS
Sbjct 1 MATKAVAVLKGDGFPVQGIINFEQKESNGPVKVVWGSIKGLTEGLHGFHVHEFGDNTAGCTS 60

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

Query 121 HEGMDDL GAGGHELSLTGNAGGRVACGVIGIA 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| USGM superoxide dismutase [Cu-Zn] [Homo sapiens]
sp|P00441.2|SODC\_HUMAN SGM RecName: Full=Superoxide dismutase [Cu-Zn]; AltName:
dismutase 1; Short=hSod1
emb|CAA26182.1| SGM unnamed protein product [Homo sapiens]
  >12 more sequence titles
  Length=154

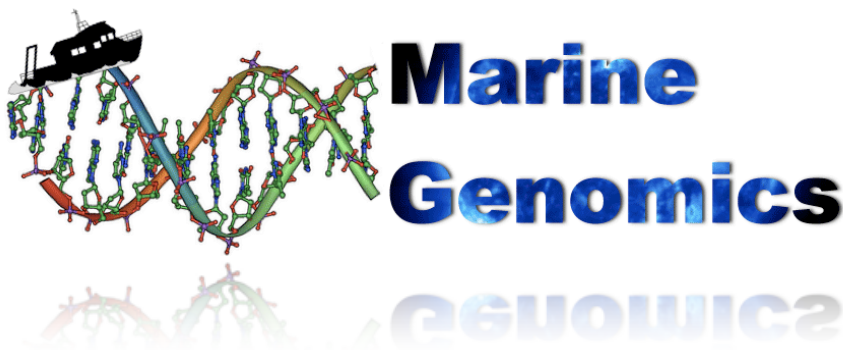
  GENE ID: 6647 SOD1 | superoxide dismutase 1, soluble [Homo sapiens]
  (Over 100 PubMed links)

  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  MVIRGVCCLVGDNEVKGVIHFTQQAPDGPCTLRGRITGLTEGKHGFHIEFGDNTNGCTS 60
           M + VC L GD V+G+I+F Q+ +GP + G I GLTEG HGFH+HEFGDNT GCTS
  Sbjct 1  MATKAVCVLKGDPVQGIINFEQKESNGPVKVGSIKGLTEGLHGFHVHEFGDNTAGCTS 60

  Query 61 AGAHYNPHGKMHGAPEDKDRHLGDLGNIEADANGIADVSI TDCLVSLTGQCSIIGRSLVV 120
           AG H+NP + HG P+D++RH+GDLGN+ AD +G+ADVSI D ++SL+G IIGR+LVV
  Sbjct 61 AGPHFNPLSRKHGGPKDEERHVGD LGNVTADKDG VADVSIEDSVISLSGDHCCIIGRTL VV 120

  Query 121 HEGMDDL GAGGHEL SLTTGNAGGRVACGVIGIA 153
            HE DDLG GG+E S TGNAG R+ACGVIGIA
  Sbjct 121 HEKADDL GKGNEESTKTGNAGSRLACGVIGIA 153
  
```



**Unpublished Cnidarian Sequence Databases**

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