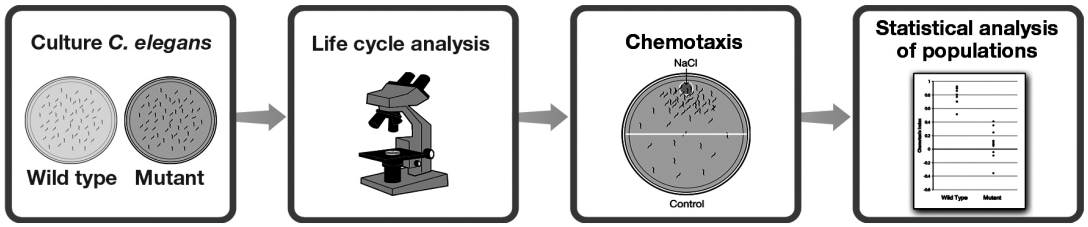
Biotechnology Explorer™

C. elegans Behavior Kit

Bioinformatics Supplement

explorer.bio-rad.com

Catalog #166-5120EDU



This kit contains temperature-sensitive reagents.
Open immediately and see individual components for storage temperature.

Please see redemption instructions on how to receive your *C. elegans*.

Please visit explorer.bio-rad.com to access our selection of language translations for Biotechnology Explorer kit curricula.

BIO-RAD

BLAST Extension Activity

At this point, you have performed experiments that show the impact of a genotypic change to the *daf-18* gene on a phenotypic response in *C. elegans*, the ability to learn to associate NaCl with food. *C. elegans* was used in these studies as a model organism since it is easy to work with and the entire genome and connectome have been determined. Model organisms are traditionally used to help us understand more complex organisms, such as humans, where there might be ethical or experimental issues in performing and studying the same types of genotype-phenotype links.

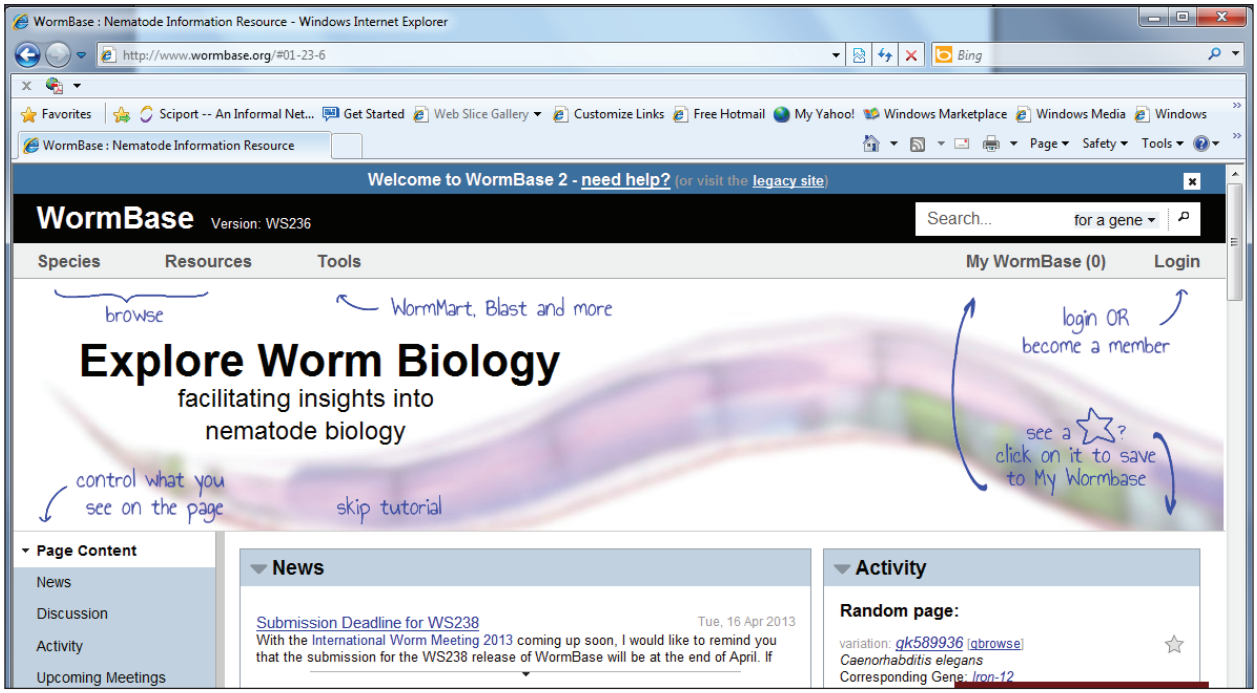
However, when using a model system, it is important to understand what similarities and differences exist between the model system and the more complex system. For example, what genes in a mouse or a human are similar to the *C. elegans daf-18* gene? Do the proteins produced from these mouse and human genes perform the same function as the protein produced by the *daf-18* gene? We can answer these questions by delving into data stored in large genetics databases like those maintained by the National Center for Biotechnology Information (NCBI), such as GenBank.

By performing a homology search — a search in a large genetic information database for all sequences that are close matches with a specific sequence of interest — you can begin to determine similarities and differences among various species for certain genes and proteins. Do higher species even have a gene similar to *daf-18*? If so, what protein does that gene code for in higher species? How does what is known about the function of that protein in higher species compare with what is known about the function of the protein produced by *daf-18* in *C. elegans*?

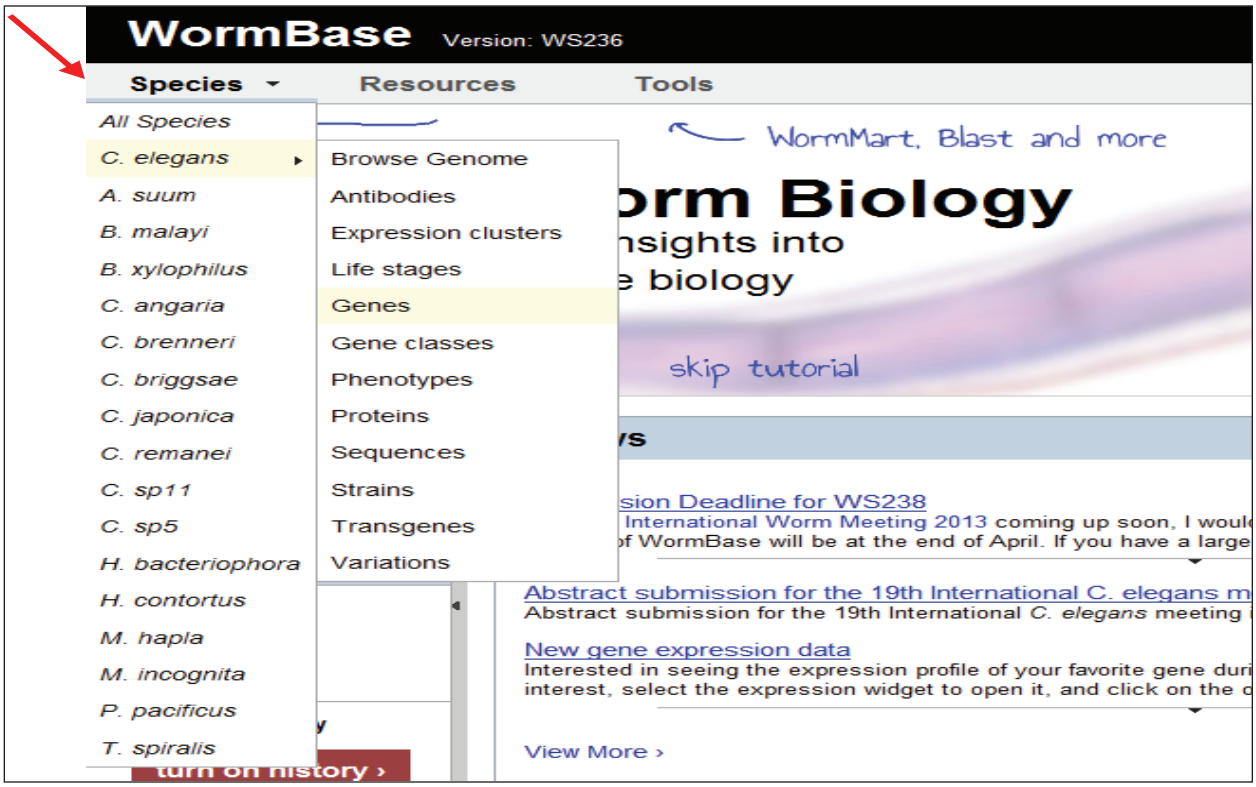
To answer the question of whether higher species have a gene similar to *daf-18*, a BLAST (Basic Local Alignment Search Tool) search can be performed. Programs in the BLAST suite look for regions of sequence that are similar within the huge databases maintained by NCBI and return results for the closest matches. The procedure listed below will help you to use BLAST to check the homology of the coding DNA sequence of *C. elegans daf-18* with all other nucleotide sequences that have been uploaded into GenBank.

Retrieve the *C. elegans* Coding DNA Sequence for *daf-18* from WormBase.org

Open a Web browser and go to WormBase.org.

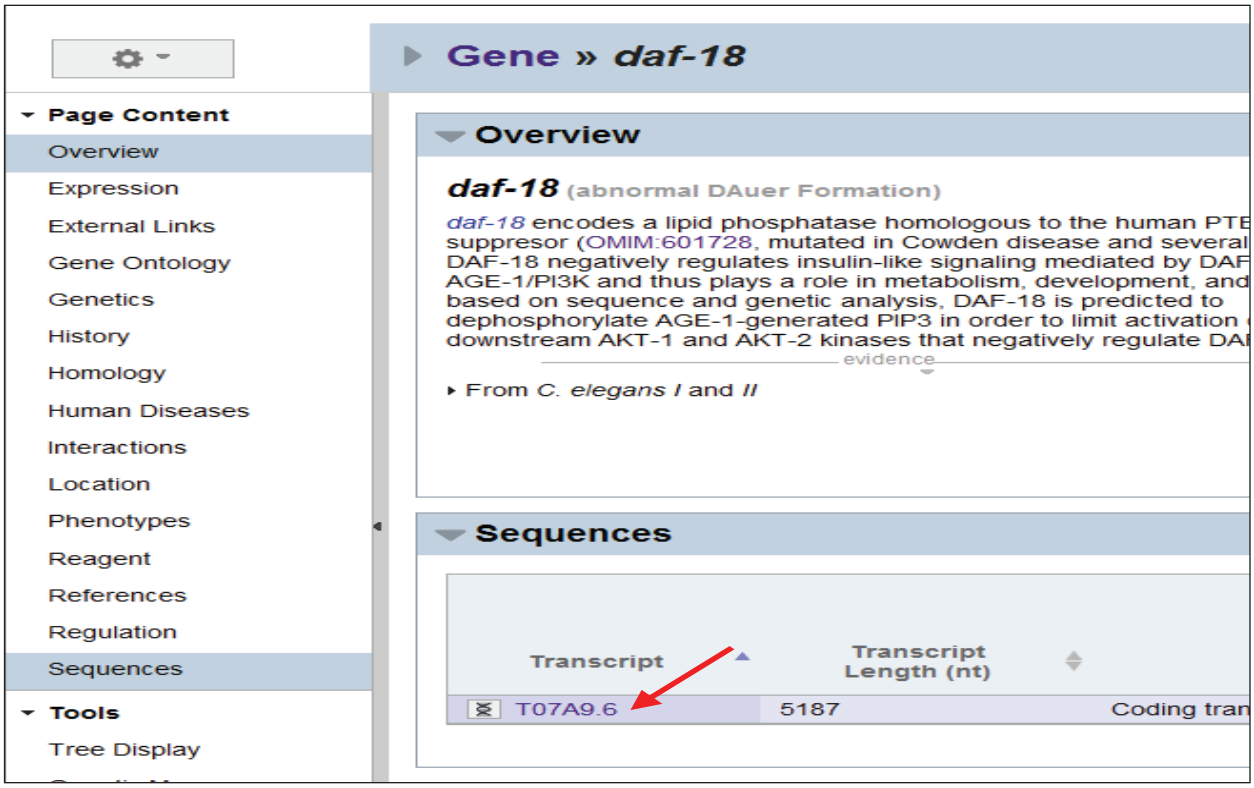


Hover the mouse over the Species menu, then over the *C. elegans* option (it should turn yellow), and finally over the Genes option. Click Genes.

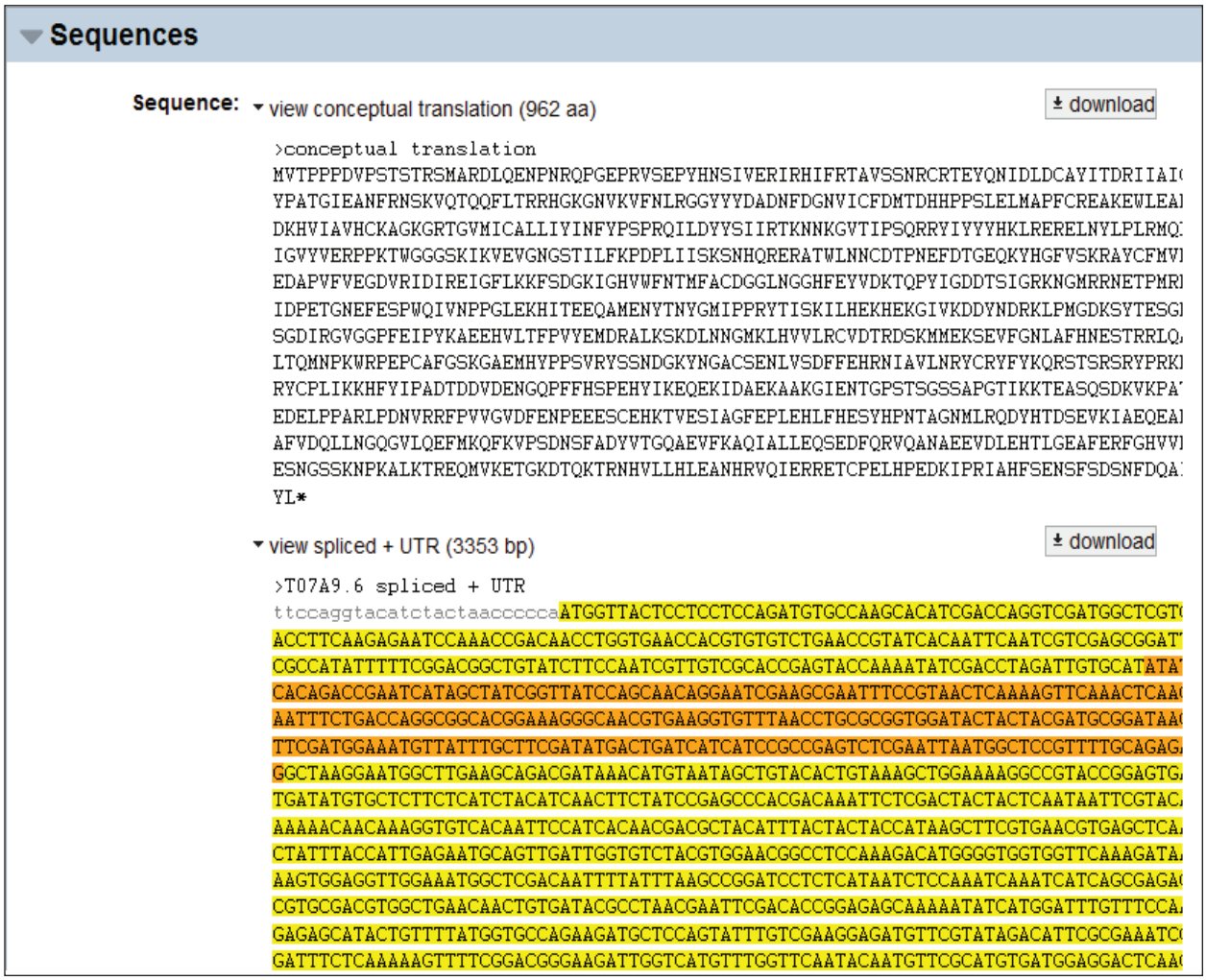


A new window will open which has a basic search option. Type daf-18 into the search box. If “for a gene” is not listed next to where you typed daf-18, use the down arrow toggle menu to choose “for a gene” and press the magnifying glass button to begin the search.

A new window with information about the *daf-18* gene will open. Look for the box marked Sequences. Under the heading Transcript, click **T07A9.6** to open up the sequence data.

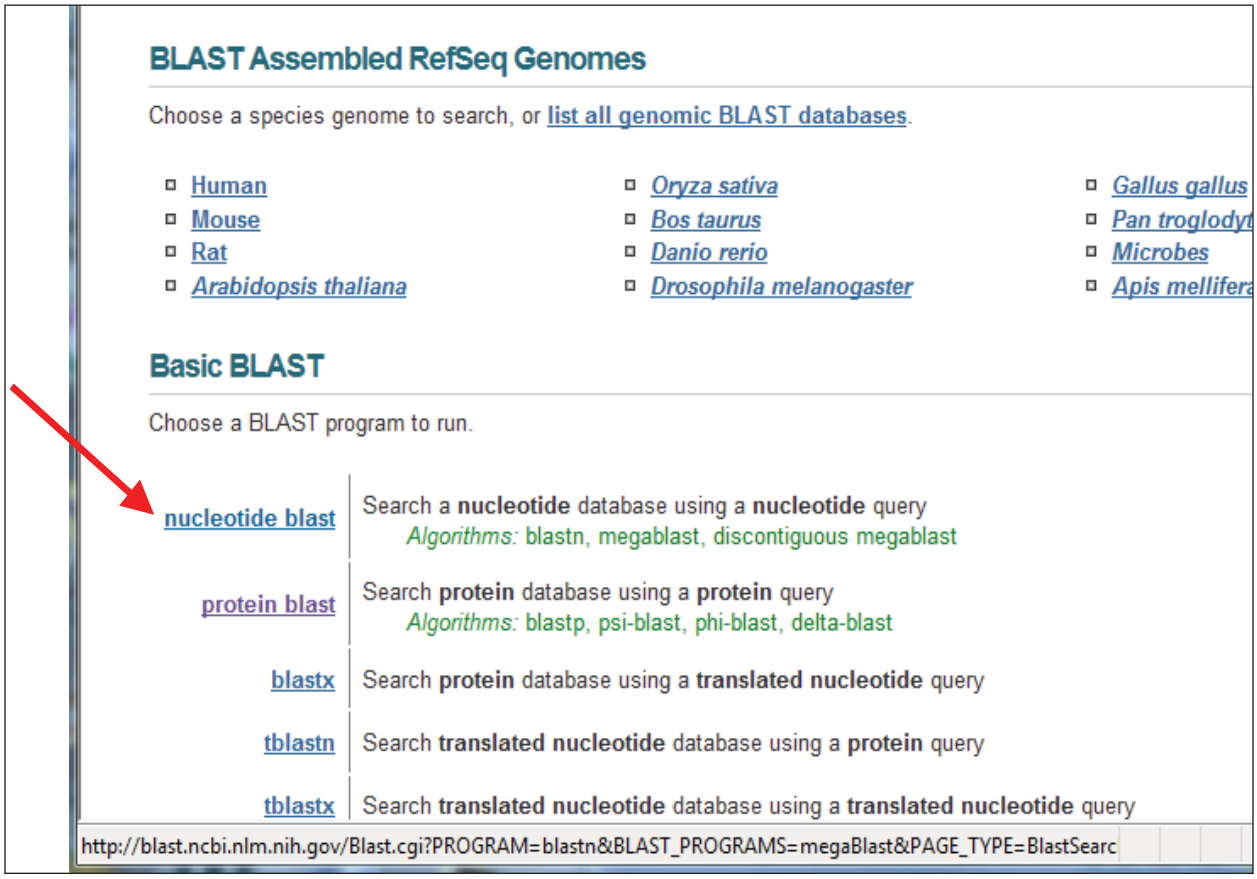


A window that contains both protein and DNA sequences will open. If all you see is a protein (amino acid) sequence listed as “view conceptual translation (962 aa),” scroll down the window until you see “view spliced + UTR (3353 bp)” and click the toggle arrow next to the words so that it points downward. A sequence labeled T07A9.6 spliced + UTR will now show. This sequence is derived by removing the intron sequence regions and linking the exon sequence regions. UTR stands for untranslated region and this is shown in gray lower case letters that have not been highlighted in yellow or orange. Each yellow and orange section is a separate exon coding region of the *daf-18* gene.



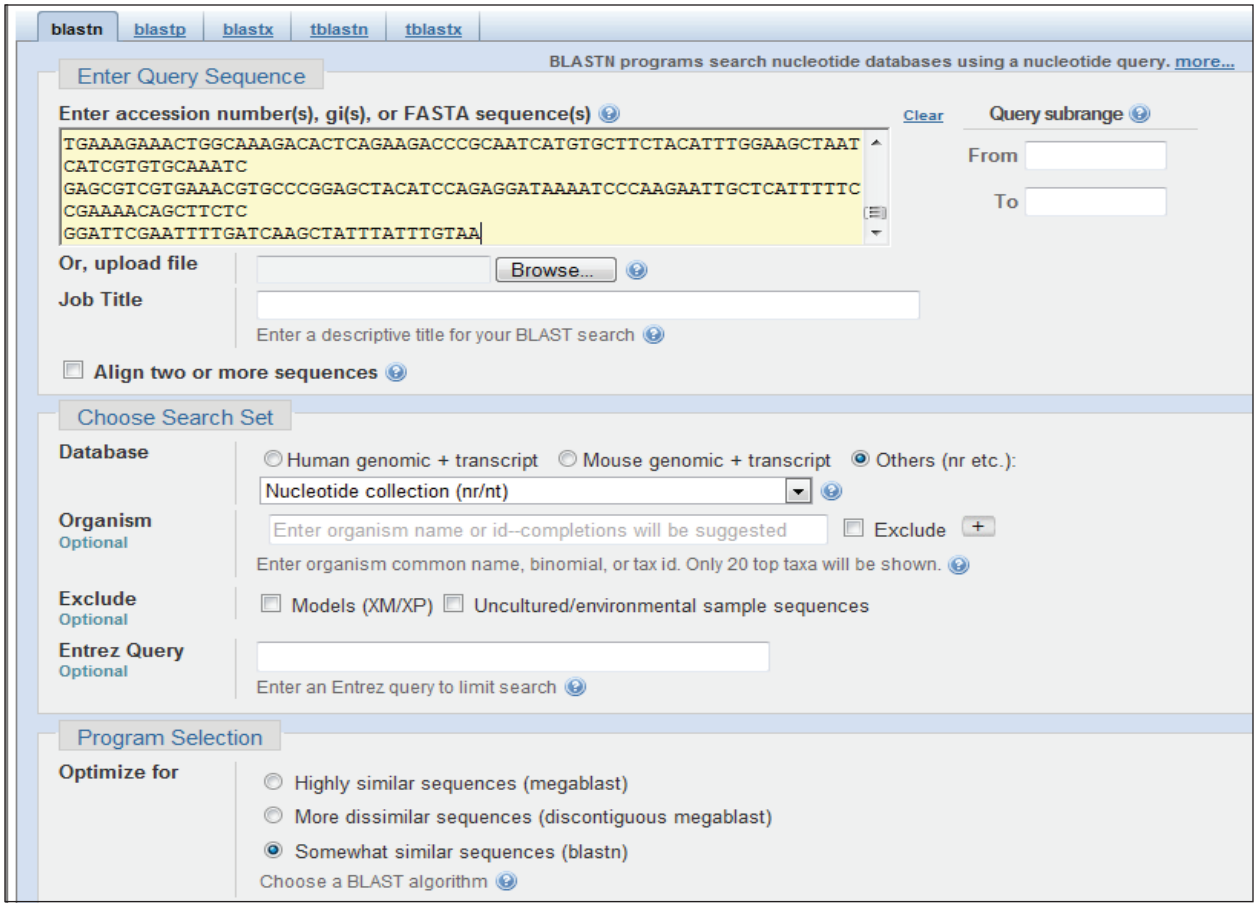
Using BLAST to Find Sequences Similar to the *C. elegans daf-18* Gene

Open a new browser window and go to the NCBI BLAST homepage at blast.ncbi.nlm.nih.gov/Blast.cgi. Choose nucleotide blast from the Basic BLAST options.

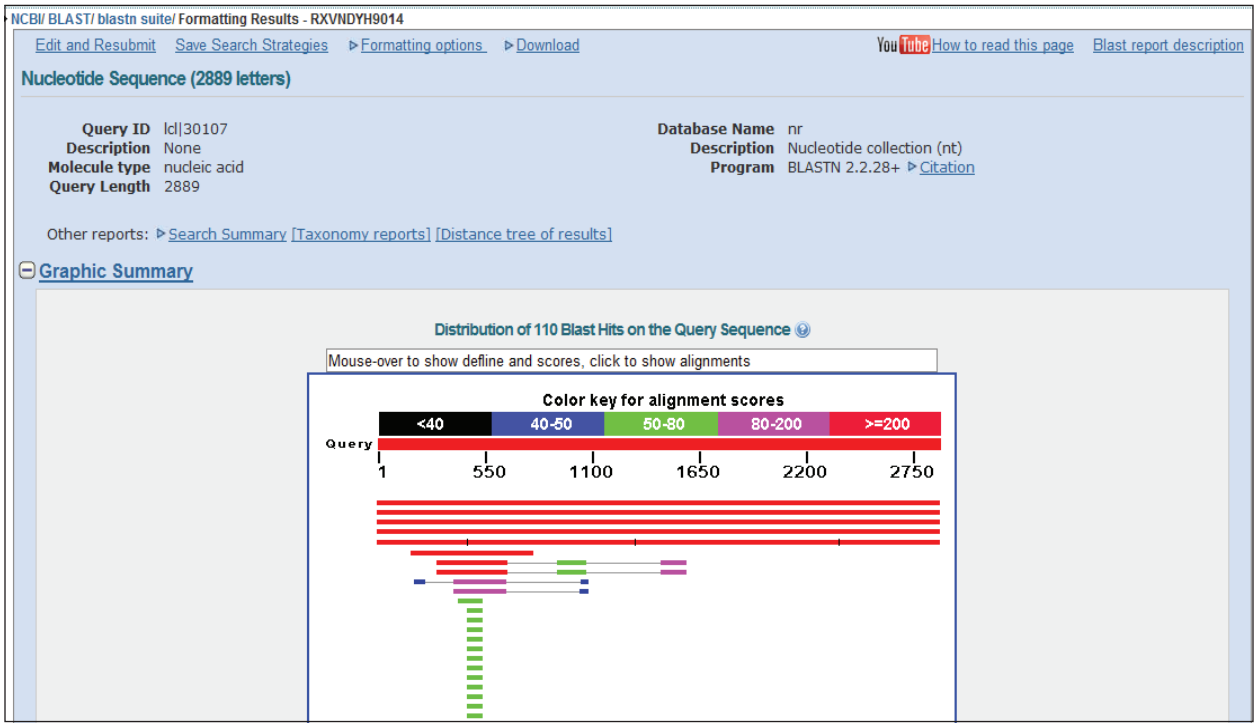


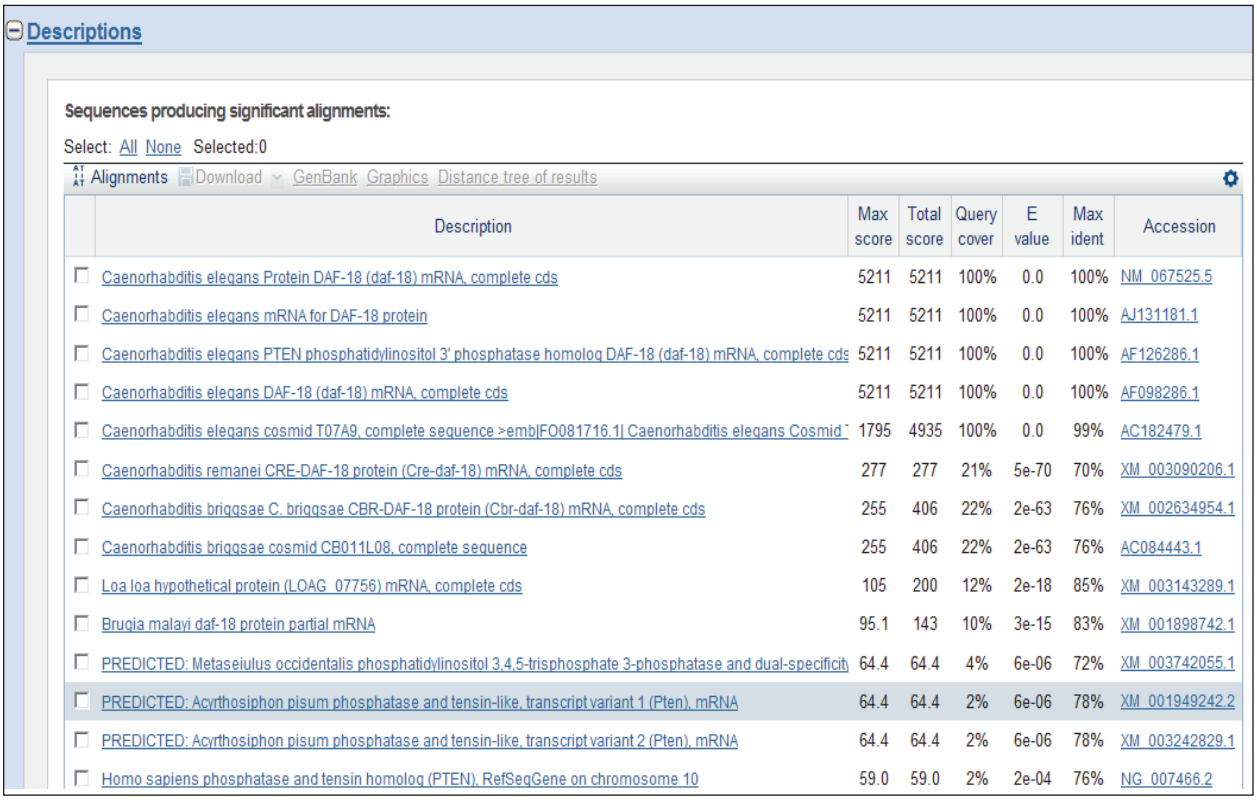
Return to the WormBase sequence page in your first browser window and copy all of the highlighted data from the “view spliced + UTR” sequence information. Do not copy the gray, lower case nucleotide data at the beginning and end of the yellow- and orange-highlighted data.

- Paste these data into the yellow box of the BLAST nucleotide search page in the Enter Query Sequence box labeled Enter accession number(s), gi(s), or FASTA sequence(s)
- In the Choose Search Set box, make sure that nucleotide collection (nr/nt) is selected
- Under Program Selection, choose Somewhat similar sequences (blastn)
- Press the blue BLAST button to begin the search



The BLAST program will perform a search and then display data for the results that are top matches for the *C. elegans daf-18* gene coding region.





Description	Max score	Total score	Query cover	E value	Max ident	Accession
Caenorhabditis elegans Protein DAF-18 (daf-18) mRNA, complete cds	5211	5211	100%	0.0	100%	NM_067525.5
Caenorhabditis elegans mRNA for DAF-18 protein	5211	5211	100%	0.0	100%	AJ131181.1
Caenorhabditis elegans PTEN phosphatidylinositol 3' phosphatase homolog DAF-18 (daf-18) mRNA, complete cds	5211	5211	100%	0.0	100%	AF126286.1
Caenorhabditis elegans DAF-18 (daf-18) mRNA, complete cds	5211	5211	100%	0.0	100%	AF098286.1
Caenorhabditis elegans cosmid T07A9, complete sequence >emb|FO081716.1| Caenorhabditis elegans Cosmid T	1795	4935	100%	0.0	99%	AC182479.1
Caenorhabditis remanei CRE-DAF-18 protein (Cre-daf-18) mRNA, complete cds	277	277	21%	5e-70	70%	XM_003090206.1
Caenorhabditis briggsae C. briggsae CBR-DAF-18 protein (Cbr-daf-18) mRNA, complete cds	255	406	22%	2e-63	76%	XM_002634954.1
Caenorhabditis briggsae cosmid CB011L08, complete sequence	255	406	22%	2e-63	76%	AC084443.1
Loa loa hypothetical protein (LOAG_07756) mRNA, complete cds	105	200	12%	2e-18	85%	XM_003143289.1
Brugia malayi daf-18 protein partial mRNA	95.1	143	10%	3e-15	83%	XM_001898742.1
PREDICTED: Metaseiulus occidentalis phosphatidylinositol 3,4,5-trisphosphate 3-phosphatase and dual-specificity	64.4	64.4	4%	6e-06	72%	XM_003742055.1
PREDICTED: Acyrthosiphon pisum phosphatase and tensin-like, transcript variant 1 (Pten), mRNA	64.4	64.4	2%	6e-06	78%	XM_001949242.2
PREDICTED: Acyrthosiphon pisum phosphatase and tensin-like, transcript variant 2 (Pten), mRNA	64.4	64.4	2%	6e-06	78%	XM_003242829.1
Homo sapiens phosphatase and tensin homolog (PTEN), RefSeqGene on chromosome 10	59.0	59.0	2%	2e-04	76%	NG_007466.2

Understanding blastn Results

The results from a blastn search include many different kinds of information and statistics. These bits of information include the size of the database, length of each query sequence, statistics that describe the number and percent of matching bases, a BLAST score, and the E value.

At the top of the blastn results page is a graphical representation of the results. A thick red bar represents the full length of the query sequence. In this case, the coding sequence for *C. elegans daf-18* is 2,889 bases long.

Below the thick red query bar are thinner colored bars that represent sequences from the database (subject sequences) that align with the query sequence. The accession number and description of the subject sequence will appear in the box at the top of the graph when you mouse over a bar. The colors of these bars show the degree of alignment based on the color key above the red query bar. The colors are based on the blastn max scores. The subject sequence with the highest max score is at the top.

If the subject sequences do not continuously match the query, the colored bars are connected by thin gray lines representing regions where there is no homology to the query.

Next in the blastn results page is a table that summarizes the statistics. Each row contains a matching sequence with the closest match sequence at the top of the table.

Description — the description refers to the source of the matching sequence. The description usually contains information on the species, possible gene name, and what type of sample was used to generate the sequence (that is, mRNA, complete genomic code, predicted genes based on computer algorithms, etc.).

Max score (maximum score) — each of the colored bars in the BLAST alignment graph (at the top of the BLAST search results page) has been assigned a score based on the extent of the match. The max score comes from the block of aligned sequence that had the highest score.

The top four matches to the *daf-18* gene queried are all submissions from different researchers of mRNA for *C. elegans daf-18*. The sixth match from the top, for *C. remanei* CRE-DAF-18 protein mRNA has a much lower max score (277) and hence is a much lower match. Even though *C. remanei* is also a nematode, the DNA sequence similarity between its daf-18 and *C. elegans' daf-18 gene* is not very high.

Total score — the total score is obtained by adding the max scores from any region of the query sequence that matches any region on the subject sequence. In this example, where we used only exon sequence data, there should not be long gaps of non-matching sequence followed by long stretches of matching sequence, so this score will be comparable to the max score. Total score is more important when trying to match genomic DNA sequences that include both exon sequences expected to have higher similarity and intron sequences that are not expected to have much similarity between different species.

Query coverage— the query coverage corresponds to the fraction of the entire query sequence that is matched by parts of the subject sequence. In this case, for the top match that is not a *C. elegans* sequence, only 21% of the query sequence matches the subject sequence (the sequence in GenBank). The query that was submitted was 2,889 bases long and 606 of these were found to align with the subject sequence in the database.

E value — the E value represents the number of sequences closely matched to the query sequence that would be expected to turn up by chance in a search of the database. The E value can have such a large range that it is reported as a power of 10 (for example, e-2 means 10^{-2}). E values below 1 can be translated to the probability that two sequences will match to a large extent. So an E value of 0.01 signifies a 1% chance of finding a good match in a database of random sequences. While low E values indicate that any matches will not be random, high E values suggest that it is possible to find an equally good match by chance. In the top row of this example, the E value is 0. This means that there is a 0% chance of finding this match in a database of random sequences. In other words, a match is statistically not likely to occur by chance.

Two additional factors have a strong influence on E values; the length of the sequence, because it is easier to find a perfect match to a shorter sequence than it is to a longer sequence, and the size of the database, because it is easier to find a match in a larger database than it is in a smaller one.

Max identity (maximum identity) — this column shows the block of a sequence that has the highest percentage of matching bases. In this example, the maximum identity of any matching block with the *C. remanei* CRE-DAF-18 protein mRNA is 70%. This sequence has 444 of 634 aligned bases that match; but it has 13 gaps as well. This can be seen if you scroll down the GenBank search page until you reach this match.

Accession number — an accession number is the unique identifier given to a DNA sequence when it is submitted to a database. (It can also refer to a submitted protein sequence.) The submitted data can be for mRNA, a synthetically generated cosmid construct, genes predicted from genomic sequences using computer algorithms that look for certain DNA patterns, an entire genome, a chromosome within a genome, or other forms.

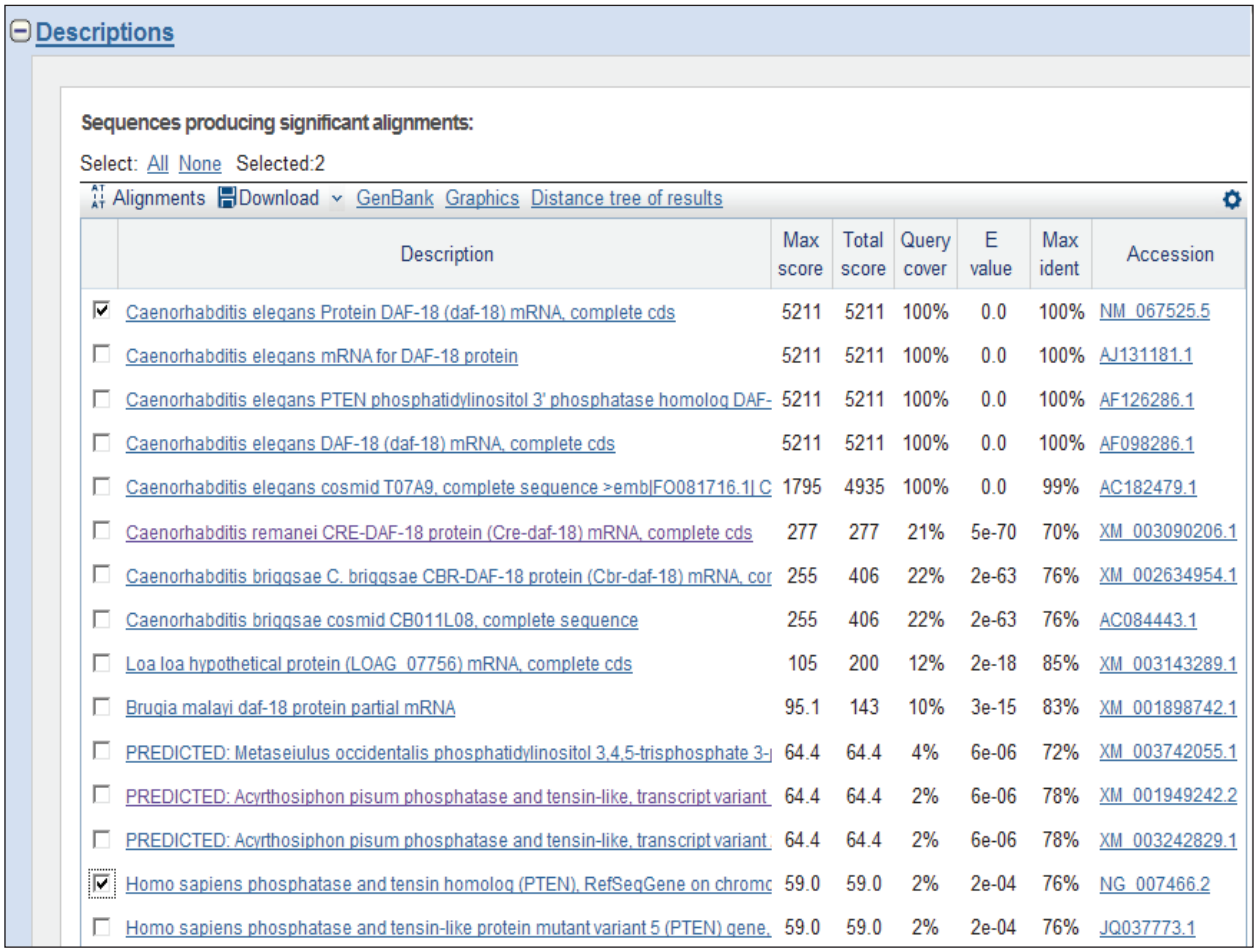
Links — the final column in the blastn alignment table contains links to other databases, which are identified in a key above the table on the BLAST results page. In this example, there are no links to other databases.

Examining Sequence Homology between *C. elegans*, Other Model Organisms, and Humans

Description	Max Score	Total Score	Query Coverage	E Value	Max Identity

Click in the box to the left of the *C. elegans* sequence with the top homology (the first sequence on the list) to choose it for analysis. Record the description, max score, total score, query coverage, E value, and max identity for this sequence in the table above.

Click in the box to the left of the top *Homo sapiens* phosphatase and tensin homolog (PTEN) to choose it for analysis. Record the description, max score, total score, query coverage, E value and max identity for this sequence in the table on page above.



Descriptions

Sequences producing significant alignments:

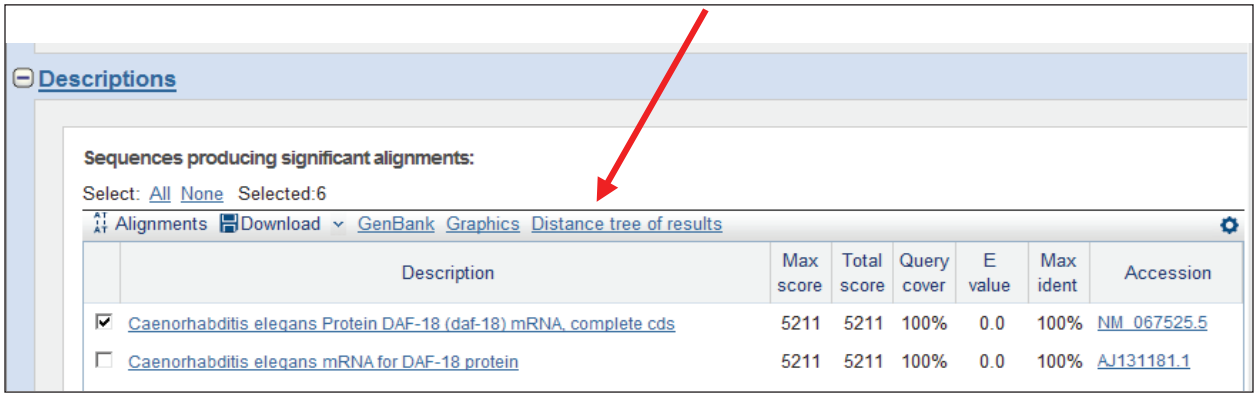
Select: All None Selected:2

Alignments Download GenBank Graphics Distance tree of results

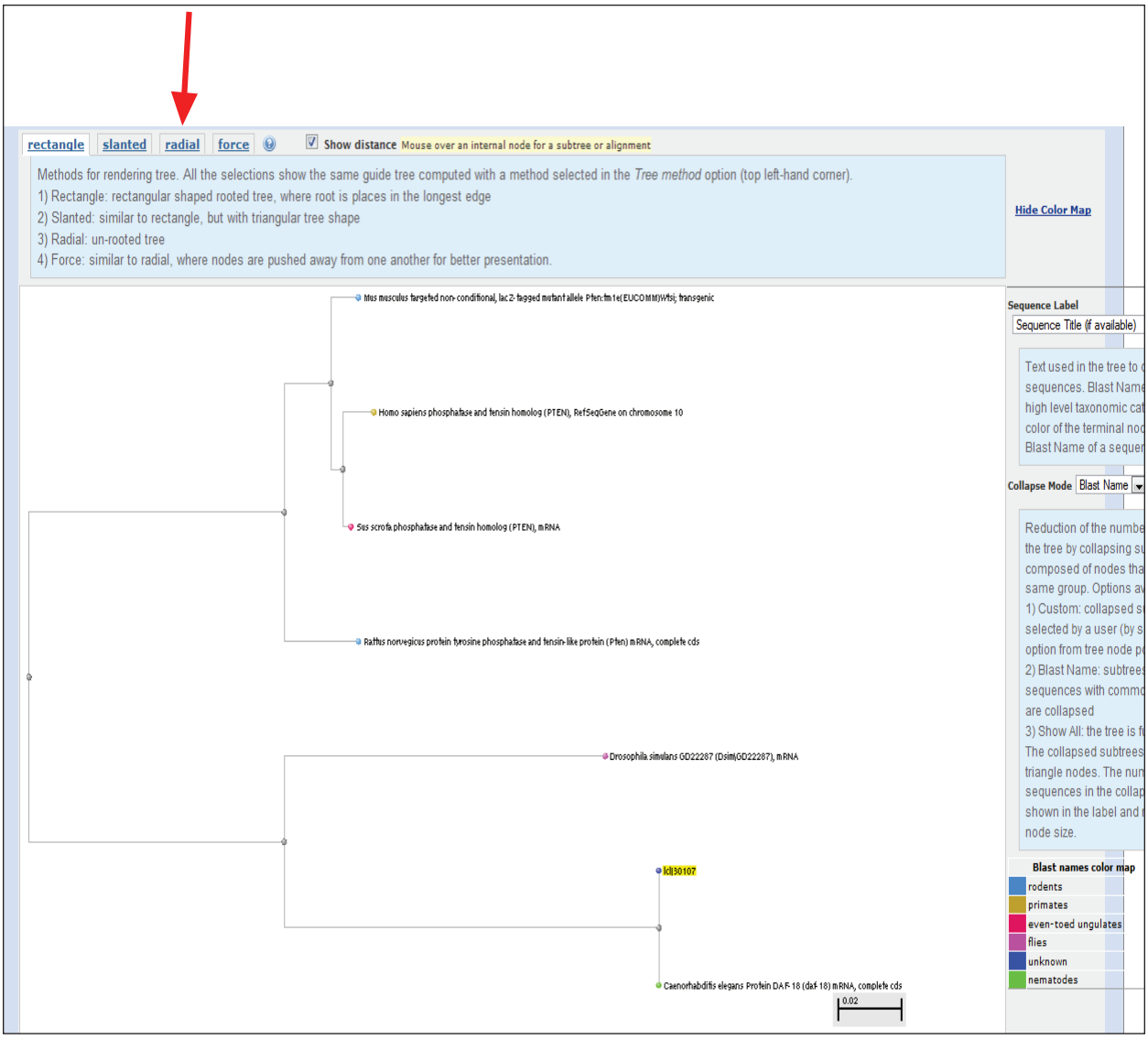
	Description	Max score	Total score	Query cover	E value	Max ident	Accession
☑	Caenorhabditis elegans Protein DAF-18 (daf-18) mRNA, complete cds	5211	5211	100%	0.0	100%	NM_067525.5
☐	Caenorhabditis elegans mRNA for DAF-18 protein	5211	5211	100%	0.0	100%	AJ131181.1
☐	Caenorhabditis elegans PTEN phosphatidylinositol 3' phosphatase homolog DAF-	5211	5211	100%	0.0	100%	AF126286.1
☐	Caenorhabditis elegans DAF-18 (daf-18) mRNA, complete cds	5211	5211	100%	0.0	100%	AF098286.1
☐	Caenorhabditis elegans cosmid T07A9, complete sequence >emb|FO081716.1| C	1795	4935	100%	0.0	99%	AC182479.1
☐	Caenorhabditis remanei CRE-DAF-18 protein (Cre-daf-18) mRNA, complete cds	277	277	21%	5e-70	70%	XM_003090206.1
☐	Caenorhabditis briggsae C. briggsae CBR-DAF-18 protein (Cbr-daf-18) mRNA, cor	255	406	22%	2e-63	76%	XM_002634954.1
☐	Caenorhabditis briggsae cosmid CB011L08, complete sequence	255	406	22%	2e-63	76%	AC084443.1
☐	Loa loa hypothetical protein (LOAG_07756) mRNA, complete cds	105	200	12%	2e-18	85%	XM_003143289.1
☐	Brugia malayi daf-18 protein partial mRNA	95.1	143	10%	3e-15	83%	XM_001898742.1
☐	PREDICTED: Metaseiulus occidentalis phosphatidylinositol 3,4,5-trisphosphate 3-i	64.4	64.4	4%	6e-06	72%	XM_003742055.1
☐	PREDICTED: Acyrthosiphon pisum phosphatase and tensin-like, transcript variant	64.4	64.4	2%	6e-06	78%	XM_001949242.2
☐	PREDICTED: Acyrthosiphon pisum phosphatase and tensin-like, transcript variant	64.4	64.4	2%	6e-06	78%	XM_003242829.1
☑	Homo sapiens phosphatase and tensin homolog (PTEN), RefSeqGene on chromc	59.0	59.0	2%	2e-04	76%	NG_007466.2
☐	Homo sapiens phosphatase and tensin-like protein mutant variant 5 (PTEN) gene,	59.0	59.0	2%	2e-04	76%	JQ037773.1

Scroll through the list and select 5-6 other organisms. We recommend selecting at least one other member of the *Caenorhabditis* genus as well as other species that you may be very familiar with such as *Rattus norvegicus* (rat), *Drosophila simulans* (fruit fly), *Mus musculus* (mouse) or *Sus scrofa* (pig). For each organism selected, click in the box to the left to choose it for analysis. Record the description, max score, total score, query coverage, E value, and max identity in the table on page 12.

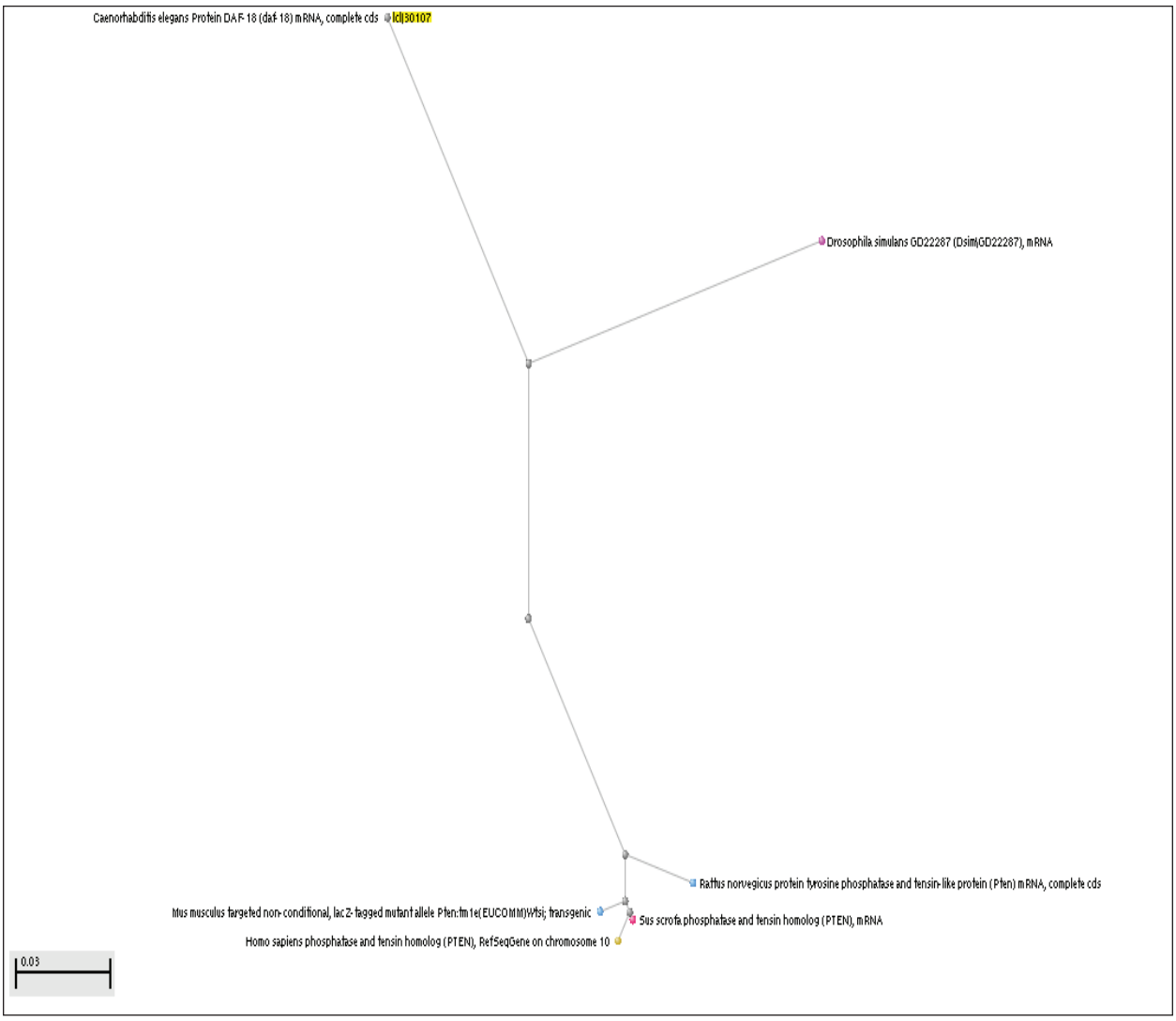
At the top of the Descriptions box, click **Distance tree of results** to generate a cladogram of your unknown sequence and the other model organism and human sequences.



A new window will open. Scroll down until you see the tree generated from the data as well as options for how to view the tree. Change the viewing option from the default rectangle to radial.



Rectangle view



Radial view

The distance tree generated shows which sequences are most similar (have the shortest distance between their branches) for the sequences picked. The query sequence (what you copied from WormBase.org) matches with *C. elegans*, which is good since this is a sequence taken from a worm database! The next closest match is fruit flies, followed by mice and rats, and then pigs and humans. If a distance tree were drawn using other genes or using physical traits, the tree would be similar to this one generated from the *daf-18* gene and its closest matches in GenBank.

What implications does this have for using *C. elegans* as a model system for humans? The *daf-18* gene is much less similar to the human PTEN gene than, say, the pig PTEN gene is to the human PTEN gene. However, generating mutations in pig DNA and raising and studying pigs, which have much more complicated connectomes than *C. elegans*, would be costly and difficult and would present some ethical issues. But even though there is little similarity at the DNA level between *C. elegans daf-18* and human PTEN, there is much similarity in function when the protein level is examined.

The protein produced by the *daf-18* gene and its human homolog PTEN both function as phosphatases, or enzymes, which remove phosphate groups from other proteins. A second functional domain found on both the protein product of the daf-18 gene and on its human homolog PTEN is a protein binding domain. This domain allows the protein to noncovalently (without sharing electron pairs) associate with other specific proteins in order to perform its enzymatic function of removing phosphate groups. While the DNA sequences vary a lot, the *C. elegans* protein product of *daf-18* and the human protein product of PTEN perform similar functions. So studying disruptions of the *C. elegans* protein can lead to new insights into how PTEN mutations might impact human phenotypes.

Bio-Rad Laboratories, Inc.

Life Science Group

Web site *www.bio-rad.com* ***USA*** *800 424 6723* ***Australia*** *61 2 9914 2800* ***Austria*** *01 877 89 01* ***Belgium*** *09 385 55 11* ***Brazil*** *55 11 5044 5699*
Canada *905 364 3435* ***China*** *86 21 6169 8500* ***Czech Republic*** *420 241 430 532* ***Denmark*** *44 52 10 00* ***Finland*** *09 804 22 00*
France *01 47 95 69 65* ***Germany*** *089 31 884 0* ***Greece*** *30 210 9532 220* ***Hong Kong*** *852 2789 3300* ***Hungary*** *36 1 459 6100* ***India*** *91 124 4029300*
Israel *03 963 6050* ***Italy*** *39 02 216091* ***Japan*** *03 6361 7000* ***Korea*** *82 2 3473 4460* ***Mexico*** *52 555 488 7670* ***The Netherlands*** *0318 540666*
New Zealand *64 9 415 2280* ***Norway*** *23 38 41 30* ***Poland*** *48 22 331 99 99* ***Portugal*** *351 21 472 7700* ***Russia*** *7 495 721 14 04*
Singapore *65 6415 3188* ***South Africa*** *27 861 246 723* ***Spain*** *34 91 590 5200* ***Sweden*** *08 555 12700* ***Switzerland*** *026 674 55 05*
Taiwan *886 2 2578 7189* ***Thailand*** *800 88 22 88* ***United Kingdom*** *020 8328 2000*

10040011 Rev A US/EG *Sig 1212*