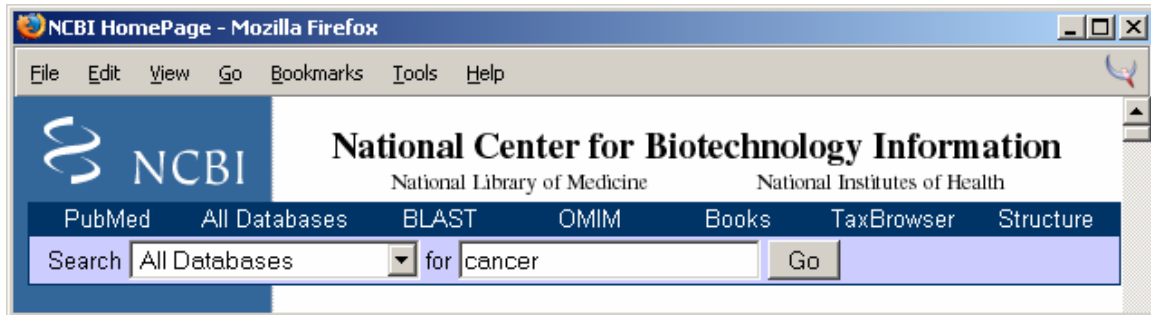


## Entrez Exercises

### *Global Query: Controlled Vocabularies and Limits*

Type the word “cancer” in the search box on the NCBI homepage and run the search.

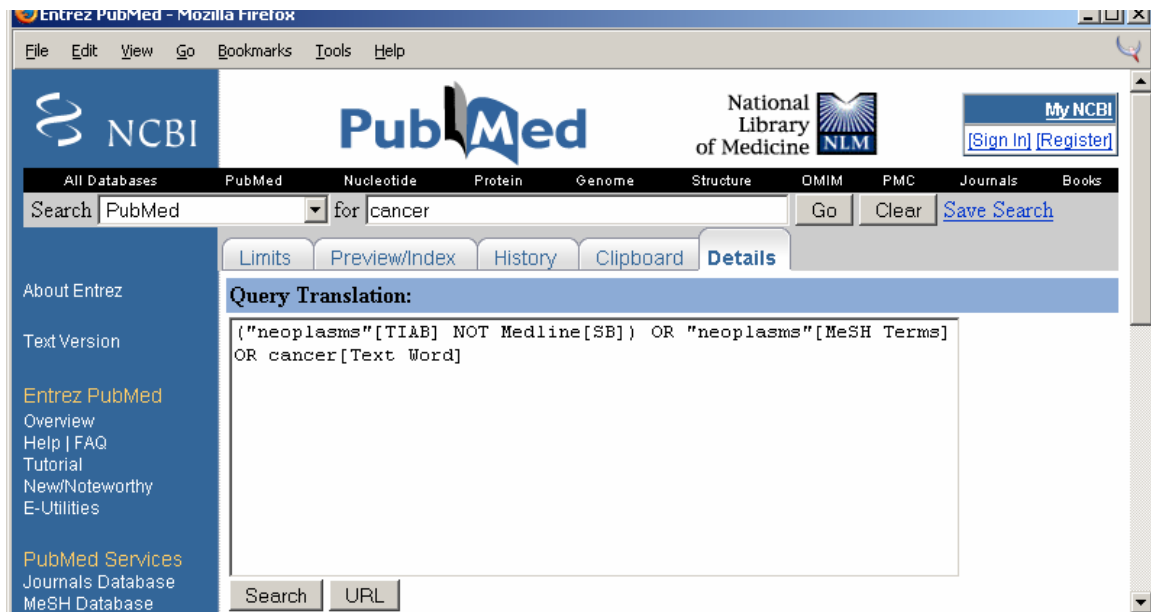


This query returns results in all of the Entrez databases. However the query is interpreted differently in different databases.

## PubMed

Retrieve the result for the PubMed database. Click the “Details” tab to see how the query was interpreted in this database.

Notice that the term cancer was translated to the Medical Subject Heading (MeSH) term “neoplasms” (“neoplasms”[MeSH Terms]).



MeSH is a controlled vocabulary that is used to index all articles in PubMed. In the details box, edit the query to remove the portion that searched for cancer as a text word and run the search.

Notice that the number of articles retrieved has changed. These will be a more relevant set of results.

You can force the PubMed engine to only search the MeSH vocabulary or specify any other indexed field through the "Limits" tab.

**Use the Web browser's back button to return to the Global query page and retrieve the PubMed results again. Now click on the "Limits" tab. Select "MeSH terms" from the first drop-down menu, the one headed by "All Fields".**

The screenshot shows the Entrez PubMed search interface. The search bar contains the text "cancer". The "Limits" tab is selected, and the "Tag Terms" section is expanded, showing a dropdown menu with "MeSH Terms" selected. The interface includes a navigation bar with "All Databases", "PubMed", "Nucleotide", "Protein", "Genome", "Structure", "OMIM", "PMC", "Journals", and "Books". The "Limits" section includes options for "Search by Author", "Search by Journal", and "Full Text, Free Full Text, and Abstracts". Below these are two columns of checkboxes for "More Publication Types" and "Age Groups". The "Tag Terms" section has a dropdown menu set to "MeSH Terms". A "GO" button and a "Clear All Limits" button are at the bottom.

**Now run the search with the limit in place and check the "Details" tab to verify that only the MeSH term translation was used.**

## Nucleotide

The Nucleotide database in the Global query is now three separate databases. The two large bulk sequence divisions, the expressed sequence tags (EST) and the genome survey sequences (GSS) are in their own separate Entrez databases. The remaining sequence records are in the database.

**Use the Web browser's back button to return to the Global query page. Retrieve the results for the CoreNucleotide database.**

The screenshot shows the NCBI Entrez Nucleotide search interface. The search query is 'cancer' in the Nucleotide database. The results show 3899750 items, with a breakdown by database: bacteria (1609), mRNA (3489304), and RefSeq (29294). The first result is NM\_013154, Rattus norvegicus CCAAT/enhancer binding protein (C/EBP), delta (Cebpd), mRNA. The 'Details' tab is selected, showing the sequence and organism information.

**Click the “Details” tab to see how the query was interpreted for this molecular database.**

In this database, the term cancer was translated into the crustacean genus name *Cancer* ("Cancer"[Organism]). The organism field stores NCBI's taxonomic classification for the source organism for the record. This is the most important controlled vocabulary for the bio-molecular Entrez databases. In this case, this translation has an unintended consequence of retrieving unrelated records: those from the crustacean genus *Cancer*, and those containing the term cancer most often in the context of the disease.

**In the details box, edit the query to remove the portion that searched for cancer in all fields so that you are just performing a search with "Cancer"[Organism] and run the search.**

This retrieves all of the nucleotide sequences for the genus *Cancer*. As you did with PubMed and the MeSH terms, you can use the “Limits” tab in the bio-molecular databases to restrict your search to the Organism field and obtain only the records from the crab genus.

## Taxonomy

**Go back to the global query results or run the search again for cancer on the NCBI homepage and retrieve the single result for the taxonomy database and click on the linked name.**

This takes you into the taxonomy browser and allows you to see all entries for the genus *Cancer*. You can check the boxes at the top to see the number of records from this genus in the various bio-molecular databases. (You must click the “Display” button to see the numbers.) These numbers are hyperlinks that will retrieve the records from the databases. The taxonomy database and browser are very useful as a global query for organism names in the bio-molecular databases.

The screenshot shows the NCBI Taxonomy Browser interface in a Mozilla Firefox browser window. The page title is "Taxonomy browser (Cancer) - Mozilla Firefox". The browser's menu bar includes File, Edit, View, History, Bookmarks, Tools, and Help. The NCBI logo is on the left, and the "Taxonomy Browser" title is on the right. Below the logo is a navigation bar with tabs for Entrez, PubMed, Nucleotide, Protein, Genome, Structure, PMC, Taxonomy, and Books. The "Taxonomy" tab is selected.

The search bar contains "Cancer" and "as complete name" is selected. There are "Go" and "Clear" buttons. Below the search bar, "Display 3 levels using filter: none" is shown. A grid of checkboxes allows filtering results by database type, with many options checked, including Nucleotide, Protein, Structure, Genome Sequences, Genome Projects, Popset, SNP, 3D Domains, Domains, GEO Datasets, GEO Expressions, UniGene, UniSTS, PubMed Central, Gene, HomoloGene, MapView, LinkOut, BLAST, and TRACE.

The "Lineage (full)" section lists a taxonomic path: [root](#); [cellular organisms](#); [Eukaryota](#); [Fungi/Metazoa group](#); [Metazoa](#); [Eumetazoa](#); [Bilateria](#); [Coelomata](#); [Protostomia](#); [Panarthropoda](#); [Arthropoda](#); [Mandibulata](#); [Pancrustacea](#); [Crustacea](#); [Malacostraca](#); [Eumalacostraca](#); [Eucarida](#); [Decapoda](#); [Pleocyemata](#); [Brachyura](#); [Eubrachyura](#); [Heterotremata/Thoracotremata group](#); [Heterotremata](#); [Cancroidea](#); [Cancridae](#).

The "Cancer" section is expanded, showing a list of species with their respective counts and "LinkOut" links:

- ◊ [Cancer](#) 183 70 6 27 [LinkOut](#) *Click on organism name to get more information.*
  - [Cancer antennarius](#) (Pacific rock crab) 2 2 1 [LinkOut](#)
  - [Cancer borealis](#) (Jonah crab) 9 10 4 [LinkOut](#)
  - [Cancer branneri](#) (furrowed rock crab) 1 1 [LinkOut](#)
  - [Cancer gracilis](#) (graceful rock crab) 1 1 [LinkOut](#)
  - [Cancer irroratus](#) (Atlantic rock crab) 4 2 1 [LinkOut](#)
  - [Cancer magister](#) (Dungeness crab) 144 18 2 10 [LinkOut](#)
  - [Cancer novaezealandiae](#) 1 1 [LinkOut](#)
  - [Cancer oregonensis](#) (pygmy rock crab) 1 1 [LinkOut](#)
  - [Cancer pagurus](#) (edible crab) 17 33 4 12 [LinkOut](#)
  - [Cancer productus](#) (red rock crab) 3 1 1 1 [LinkOut](#)

## ***Nucleotide: Zebrafish prolactin***

### **Zebrafish nucleotide sequences**

Perform a search to retrieve all zebrafish sequences in the CoreNucleotide database. Use the "Limits" tab to select the "Organism" field to force the translation to an organism search as in the first exercise.

### **Limits: the Properties field**

You can now use the "Limits" to eliminate certain types of sequences from your results.

Click on "Limits" and use the checkboxes to remove the high throughput genomic (HTG or Working Draft) sequences from your results. Check the box next to "exclude working draft" and run the search.

Entrez Nucleotide - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

NCBI

My NCBI [Sign In] [Register]

All Databases PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search CoreNucleotide for zebrafish Go Clear

Limits Preview/Index History Clipboard Details

- Use All Fields pull-down menu to specify a field.
- Boolean operators AND, OR, NOT must be in upper case.
- If search fields tags are used enclose in square brackets, e.g., rubella [ti].
- More help on using limits is available [here](#).

Limited to:

All Fields

exclude STSs  exclude TPA  exclude working draft  exclude patents  exclude all of the above

Molecule Gene Location Segmented Sequences

Only from Modification Date

Modification Date From To

Use the format YYYY/MM/DD; month and day are optional.

Click on the “Details” tab to see how Entrez managed this query.

Notice the term “NOT gbdiv\_htg[PROP]”. PROP is the abbreviation for the Properties field

Limit Details

Field: Organism, Limits: exclude working draft

Query Translation:

```
"Danio rerio"[Organism] AND (1900[MDAT] : 3000[MDAT] NOT gbdiv_htg[PROP])
```

Search URL

The Properties field terms are a controlled vocabulary for classifying sequence records. These terms are somewhat cryptic, but they are very helpful. Three useful types are the `biomol`, `gbdiv` and `srcdb` sets. The `biomol` terms classify records based on the type and origin of the molecule, for example `biomol mrna` or `biomol genomic`. The `gbdiv` sets of terms index records by the GenBank division code; `gbdiv est`, `gbdiv pri`, `gbdiv htg` and so on. The `srcdb` terms classify records based upon their database of origin. For nucleotide records these could be GenBank, EMBL, DDBJ, RefSeq or PDB (`srcdb genbank`, `srcdb embl`, `srcdb ddbj`, `srcdb refseq`). Many of the available filters on the “Limits” tab are managed through the Properties field terms.

**Preview/Index: adding terms to query**

**Return to the CoreNucleotide search results. Go into “Limits” again and use the “Molecule” drop-down menu to select mRNA and run the search.**

The results now contain all non-EST zebrafish mRNA sequences from the primary databases and the RefSeq database.

**Click on the “Preview/Index” tab.**

At the bottom of the “Preview/Index” page, is a search box with a drop-down menu that allows you to add terms to your search and restrict to certain fields if you like.

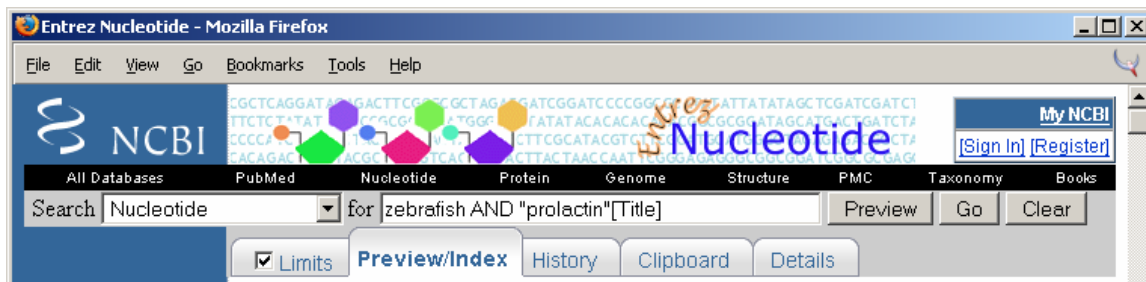
**Now, type “prolactin” in the search box.**

Although the vocabulary used is not strictly controlled, the name of a gene or gene product is generally in the title of a record. The title is displayed in the “Summary” view in Entrez and is identical to the DEFINITION line in a record in GenBank format. Select “Title” from the drop-down menu to the left of the search box and click the “Index” button. This checks the index for the “Title” field for records having “prolactin” in their titles. A list containing term prolactin and its expansion is now displayed with the number of records for each term.

The screenshot shows the search interface with the following elements:

- A search box containing the text "prolactin".
- A dropdown menu set to "Title".
- Buttons for "Preview" and "Index".
- A section with buttons for "AND", "OR", and "NOT" to add terms to the query box.
- A list of search terms with their respective record counts:
  - prolactin (922)
  - prolactin 1 (3)
  - prolactin 2 (6)
  - prolactin a (1)
  - prolactin and (3)
  - prolactin b (2)
  - prolactin family (93)
  - prolactin gene (194)
  - prolactin gene promoter (2)
  - prolactin gene, promoter (2)
- Vertical scroll bars and "Up" and "Down" buttons for navigating the list.

**Select “prolactin” from the list and add it to the search by clicking the “AND” button. Then run the search.**



The results contain records from GenBank / EMBL / DDBJ and NCBI’s RefSeq database. The RefSeq records are easily identified by their characteristic style of accession numbers. Retrieve the RefSeq record for the zebrafish prolactin mRNA (NM\_181437). This RefSeq contains sequence data derived from a traditional GenBank record, but also has additional annotations and cross references added by the NCBI RefSeq staff. Unlike many primary database records, this RefSeq record will be updated and maintained as the state of knowledge about the biology of

this gene and organism advances. The search also retrieves a gene model RefSeq for the prolactin receptor (XM\_685247). This sequence has been predicted from analysis of the assembled zebrafish genome using the NCBI gene prediction program called Gnomon.

## Finding the genomic BAC clone sequence

Click on the “Links” menu in the upper right of the record (NM\_181437).

A number of links are displayed. You can now link directly to the assembled and annotated whole genome shotgun assembly of the zebrafish genome in the Map Viewer. There are also a growing number of finished BAC clone sequences from the zebrafish genome project that are available in Entrez. You can use the “Related Sequences” feature of Entrez to find a BAC clone that contains the exons of this gene.

Follow the “Related Sequences” link.

The screenshot shows the NCBI Sequence Viewer interface for record NM\_181437. The top navigation bar includes links for PubMed, Nucleotide, Protein, Genome, Structure, PMC, Taxonomy, OMIM, and Books. The search bar contains 'CoreNucleotide' and 'for'. The display settings are set to 'GenBank' with 'Show 5' items. The 'Links' menu is open, showing a list of related sequences and links.

**Links**

- Gene
- Genome Project
- Master
- PubMed (RefSeq)
- Related Sequences
- Map Viewer
- GEO Profiles
- Protein
- PubMed
- Taxonomy
- UniGene

**Record Information:**

LOCUS NM\_181437 1396 bp mRNA linear VRT 21-NOV-2006  
 DEFINITION Danio rerio prolactin (prl), mRNA.  
 ACCESSION NM\_181437  
 VERSION NM\_181437.2 GI:31340660  
 KEYWORDS .  
 SOURCE Danio rerio (zebrafish)  
 ORGANISM [Danio rerio](#)  
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;  
 Actinopterygii; Neopterygii; Teleostei; Ostariophysi;  
 Cypriniformes; Cyprinidae; Danio.

This provides a list of nucleotide sequences that are related by BLAST similarity. Similarity scores are precomputed between all sequences in the database. The related sequences list is ranked in order of decreasing BLAST score. For the nucleotide database, the significance threshold is very stringent, so that it is unusual to see nucleotide sequences from other species in the list. Therefore, the nucleotide related sequences link is often a useful as a way of collecting all sequences for a particular gene and its products from one species. Often you can't easily collect all of them using a text search because of inconsistencies or errors in the annotation.

**You should find the sequence from BAC clone DKEY-16P21, accession BX511021, in the list of related sequences. Retrieve this record through the linked identifier.**

This is a typical finished BAC clone from a genome project. Notice that this is the ninth version of this record. In previous versions, this was a draft sequence in the high throughput genomic (HTG) GenBank division. You can see all versions of the record in the revision history available through the reports link.

NCBI Sequence Viewer v2.0 - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

1: [BX511021](#) Reports: Zebrafish DNA seq... [gi:46200452] Links

**Reports**

- ▶ ASN.1
- ▶ XML
- ▶ Summary
- ▶ FASTA
- ▶ TinySeq XML
- ▶ GenBank
- ▶ GBSeq XML
- ▶ INSDSeq XML
- ▶ GenBank(Full)
- ▶ GI list
- ▶ Graphic
- ▶ Revision History

LOCUS BX5110 235632 bp DNA linear VRT 03-APR-2004

DEFINITION Zebrafish DNA sequence from clone DKEY-16P21 in linkage group 3,

ACCESSION BX5110

VERSION BX5110 452

KEYWORDS HTG.

SOURCE Danio

ORGANISM [Danio](#) (Chordata; Craniata; Vertebrata; Euteleostomi; Actinopterygii; Teleostei; Ostariophysi; Cyprinidae; Danio.)

REFERENCE 1 (base pairs)

AUTHORS Hunter et al.

The revision history for this record shows all the forms it has taken in the Entrez system.

NCBI

Sequence Revision History

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books

Find (Accessions, GI numbers or Fasta style Seq/ids)  Go Clear

About Entrez

Show difference between I and II as

Revision history for [BX511021](#)

GI	Version	Update Date	Status	I	II
46200452	9	<a href="#">Oct 20 2006 2:13 PM</a>	Live	<input checked="" type="radio"/>	<input type="radio"/>
46200452	9	<a href="#">Apr 5 2004 12:43 PM</a>	Dead	<input type="radio"/>	<input checked="" type="radio"/>
44844493	8	<a href="#">Mar 1 2004 11:12 PM</a>	Dead	<input type="radio"/>	<input checked="" type="radio"/>
42820886	7	<a href="#">Feb 25 2004 11:07 PM</a>	Dead	<input type="radio"/>	<input checked="" type="radio"/>
42733255	6	<a href="#">Feb 22 2004 11:18 PM</a>	Dead	<input type="radio"/>	<input checked="" type="radio"/>
42733255	6	<a href="#">Feb 20 2004 11:16 PM</a>	Dead	<input type="radio"/>	<input checked="" type="radio"/>
42592613	5	<a href="#">Feb 18 2004 11:07 PM</a>	Dead	<input type="radio"/>	<input checked="" type="radio"/>
42592613	5	<a href="#">Feb 17 2004 11:05 PM</a>	Dead	<input type="radio"/>	<input checked="" type="radio"/>
42538790	4	<a href="#">Feb 11 2004 11:13 PM</a>	Dead	<input type="radio"/>	<input checked="" type="radio"/>
32959715	3	<a href="#">Jul 17 2003 11:57 PM</a>	Dead	<input type="radio"/>	<input checked="" type="radio"/>
31071334	2	<a href="#">Jul 1 2003 11:29 PM</a>	Dead	<input type="radio"/>	<input checked="" type="radio"/>
31071334	2	<a href="#">May 23 2003 11:13 PM</a>	Dead	<input type="radio"/>	<input checked="" type="radio"/>
30910895	1	<a href="#">May 19 2003 11:07 PM</a>	Dead	<input type="radio"/>	<input checked="" type="radio"/>

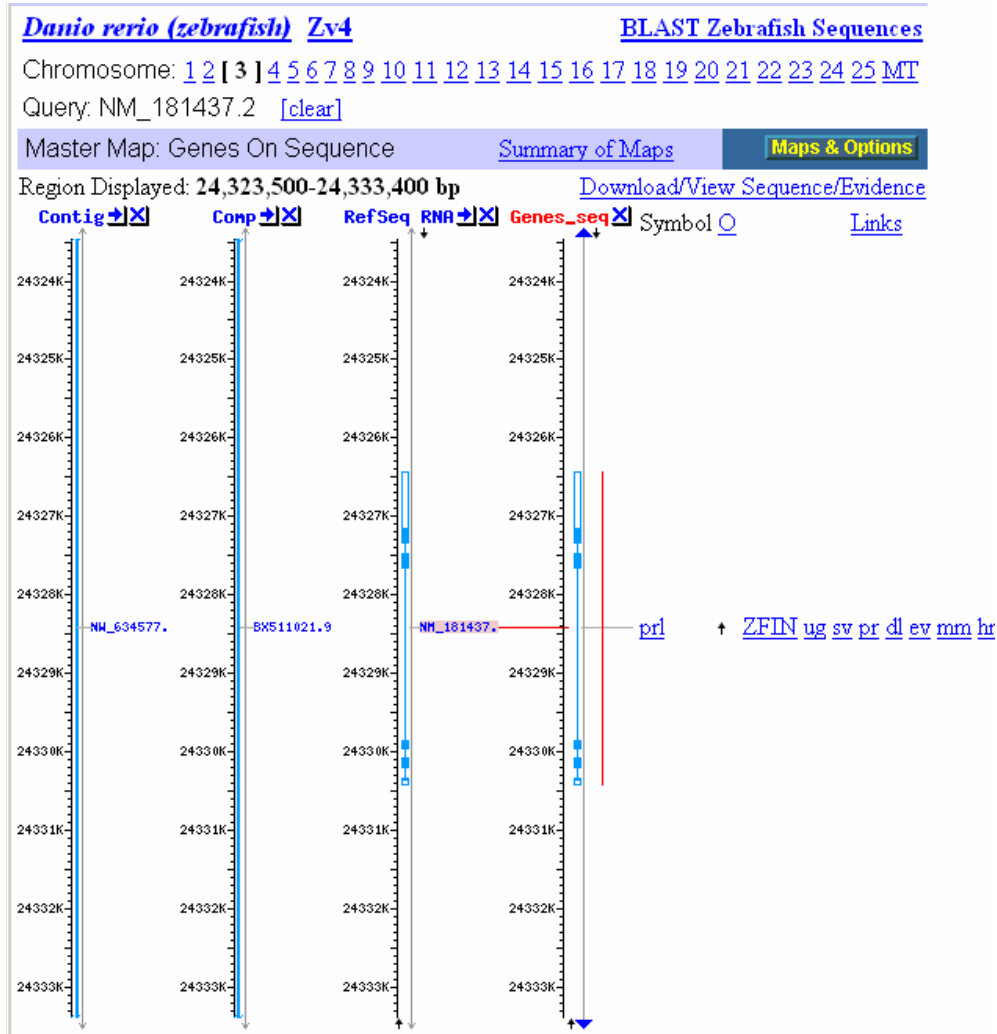
Accession [BX511021](#) was first seen at NCBI on May 19 2003 11:07 PM

The revision history also shows the gi number and accession.version number changes. These identifiers change together and only when the sequence itself changes. Other non-sequence changes can be made that do not affect these identifiers but are reflected in changes in the "Update Date".

The current version of BX511021 is no longer a draft sequence but is in the traditional vertebrate (VRT) division of GenBank. In contrast to typical traditional GenBank records, BX511021 has almost no biological annotation.

**Examine the feature table of the record and verify that the prolactin gene is not annotated there.**

Clearly, you could not have found this record using a text search for prolactin. However, in this case, now that there is an assembled zebrafish genome you could easily have found this BAC clone as a part of the assembly. The "Master" link on the links menu leads to the contig that contains this BAC. You could also find the assembly by following the link to Map viewer and adjusting the "Maps & Options" so that the "contig" and "component" maps are displayed. BX511021 appears as one of the components.



## Making a gene model

You can use the mRNA sequence and the genomic clone sequence to produce a gene model for the prolactin gene. The NCBI utilities Spidey and Splign will align mRNA to genomic sequence using consensus splice sites to constrain the alignment. Spidey is a fairly simple spliced alignment tool that produces good results in uncomplicated cases. Splign is a more sophisticated tool capable of producing alternative models. Splign is the tool currently used at the NCBI to aid in genome annotation. The Spidey and Splign pages are available at the following URLs:

<http://www.ncbi.nlm.nih.gov/spidey/>

<http://www.ncbi.nlm.nih.gov/sutils/splign/splign.cgi>

- Load the Spidey page in your Web browser.
- Type or paste the genomic accession (BX511021) in the upper text area on the form
- Type or paste the mRNA accession (NM\_181437) in the lower text area.
- Press the “Align” button to run the program.

The entire prolactin gene is contained on this BAC clone.

Genomic: [g|46200452|emb|BX511021.9](#) Zebrafish DNA sequence from clone DKEY-16P21 in linkage group 3

mRNA: [g|31340660|ref|NM\\_181437.2](#) Danio rerio prolactin (prl), mRNA

Alignment is on minus strand of genomic sequence and on plus strand of mRNA sequence  
 mRNA coverage: 100%  
 Overall percent identity: 97.6%  
 Non-aligning poly(A) tail: 10

	Genomic coordinates	mRNA coordinates	length	identity	mismatches	gaps	Donor site	Acc. site
<a href="#">Exon 1</a>	16204-16285	1-82	82	98.8%	1	0	d	
<a href="#">Exon 2</a>	15949-16070	83-204	122	99.2%	1	0	d	a
<a href="#">Exon 3</a>	15728-15835	205-312	108	100.0%	0	0	d	a
<a href="#">Exon 4</a>	13338-13520	313-495	183	99.5%	1	0	d	a
<a href="#">Exon 5</a>	12301-13205	496-1396	901	96.6%	31	6		a

Use Spign to perform the same operations.

Spign is the tool currently used at NCBI to make the mRNA to genomic alignments.

## Protein and Structures

### Example 1: Zebrafish prolactin

Display the Links menu from the zebrafish prolactin mRNA (NM\_181437) from the Nucleotides example and follow the link to the protein database.

From the protein record (NP\_852102) we can easily find homologs in other species and a structure model for the zebrafish protein.

The screenshot shows the NCBI Entrez Protein database interface. The search bar contains "Protein" and the search results list one entry: "NP\_852102. Reports prolactin [Danio ...[gi:31088862]". The entry details are as follows:

```

LOCUS      NP_852102                210 aa          linear   VRT 21-NOV-2006
DEFINITION prolactin [Danio rerio].
ACCESSION  NP_852102
VERSION    NP_852102.1  GI:31088862
DBSOURCE   REFSEQ: accession NM\_181437.2
KEYWORDS   .
SOURCE     Danio rerio (zebrafish)
  ORGANISM Danio rerio
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Actinopterygii; Neopterygii; Teleostei; Ostariophysi;
            Cypriniformes; Cyprinidae; Danio.
  
```

Navigation links include "Comment", "Features", "Sequence", "BLink", "Conserved Domains", and "Links".

## Using BLink to find homologs

The BLink output provides direct access to sequence similarity results that are equivalent to BLAST search against the default (nr) protein database.

**Click on the BLink link from the zebrafish prolactin protein record (NP\_852102).**

File Edit View History Bookmarks Tools Help

NCBI

BLAST Protein Structure PubMed Taxonomy  
Genome Nucleotide 3D-Domains Books Help

Query: gi|31088862 prolactin [Danio rerio]  
Matching gi: 28848616

Show identical Best hits Common Tree Taxonomy Report 3D structures CDD-Search GI list Run BLAST

200 BLAST hits to 100 unique species [Sort by taxonomy proximity](#)

0 Archaea 0 Bacteria 197 Metazoa 0 Fungi 0 Plants 0 Viruses 0 Other Eukaryotae

Keep only  Cut-Off 100  New search by GI: 31088862

210 aa

SCORE	P	ACCESSION	GI	PROTEIN DESCRIPTION
<a href="#">Conserved Domain Database hits</a>				
1077	31	AAH92358	62185655	Prl protein [Danio rerio]
1066	31	XP_001...	125812645	PREDICTED: hypothetical protein [Danio rerio]
999	28	P29235	130931	Prolactin precursor (PRL)
992	28	ABJ90338	116248046	prolactin [Tinca tinca]
980	28	P09585	130928	Prolactin precursor (PRL)
978	28	AAT74865	50345847	prolactin [Carassius auratus]
972	28	P35395	548596	Prolactin precursor (PRL)
809	24	AAK53436	14040042	prolactin hormone [Heteropneustes fossilis]
790	24	P51904	17380514	Prolactin precursor (PRL)
766	24	AA828018	425491	prolactin, PRL [Ictalurus punctatus=catfish,
733	21	AAA49611	532239	prolactin
733	21	P48096	1346801	Prolactin precursor (PRL)
731	21	CAA45407	64157	preprolactin [Oncorhynchus keta]
730	21	1406188A	226011	prolactin
730	21	P34181	464464	Prolactin precursor (PRL)
728	21	Q91364	17368546	Prolactin-2 precursor (Prolactin II) (PRL-II)
713	20	Q722V3	66773806	Prolactin precursor (PRL)
710	21	P69131	59800152	Prolactin-1 precursor (Prolactin I) (PRL-I)
707	21	S02304	85540	prolactin I - chum salmon
700	21	S06677	85541	prolactin II - chum salmon

This output shows the top 200 non-redundant hits from a protein-protein BLAST search against nr. Notice that there is often more than one protein in the list from the same species. This can occur because of multiple entries with different sequences for the same protein or because the protein belongs to a family of related proteins— in this case the growth hormone family with several members in each organism.

**To make it easier to find the one protein for each organism that has the best BLAST score, click the “Best hits” button.**

The output now shows one protein from each organism in the list identified by the species name.

File Edit View History Bookmarks Tools Help

Show identical All hits Common Tree Taxonomy Report 3D structures CDD-Search GI list Run BLAST

200 BLAST hits to 100 unique species [Sort by taxonomy proximity](#)

0 Archaea 0 Bacteria 197 Metazoa 0 Fungi 0 Plants 0 Viruses 0 Other Eukaryotae

Keep only  Cut-Off 100 Select

Reset

New search by GI: 31088862 Go

210 aa

SCORE	P	ACCESSION	GI	N	ORGANISM
<u>Conserved Domain Database hits</u>					
1077	31	AAH92358	62185655	6	Danio rerio
999	28	P29235	130931	2	Hypophthalmichthys nobilis
992	28	ABJ90338	116248046	1	Tinca tinca
980	28	P09585	130928	2	Cyprinus carpio
978	28	AAT74865	50345847	1	Carassius auratus
972	28	P35395	548596	1	Hypophthalmichthys molitrix
809	24	AAK53436	14040042	1	Heteropneustes fossilis
790	24	P51904	17380514	2	Ictalurus punctatus
733	21	AAA49611	532239	1	Oncorhynchus mykiss
733	21	P48096	1346801	1	Salmo salar
731	21	CAA45407	64157	4	Oncorhynchus keta
730	21	P34181	464464	1	Coregonus autumnalis
728	21	Q91364	17368546	2	Oncorhynchus tshawytscha
713	20	Q72ZV3	66773806	4	Anguilla japonica
700	20	Z014393A	744477	3	Anguilla anguilla
668	21	Q9YGV6	17367857	2	Paralichthys olivaceus
642	21	AAO11695	37727307	2	Epinephelus coioides
641	21	BAF45226	72669218	1	Mullus barbatus
276	15	P29234	62512134	3	Mustela vison
275	15	AAB20171	238002	1	Papio
273	15	Q28632	2500852	1	Oryctolagus cuniculus
273	15	P33089	62906851	2	Balaenoptera borealis
270	15	Q8HXS1	62510727	1	Ailuropoda melanoleuca
269	15	NP_001...	114052168	1	Macaca mulatta
269	15	XP_001...	114605661	1	Pan troglodytes
269	15	AAV17320	62866815	1	Nomascus leucogenys
269	15	XP_545363	74004132	1	Canis familiaris
267	1	ABM87406	123999795	3	synthetic construct
267	15	1RW5A	61679820	8	Homo sapiens
266	15	P01238	585731	2	Sus scrofa
252	15	1105254A	224452	1	Balaenoptera physalus
242	15	NP_776378	46810277	5	Bos taurus
241	15	Q28318	2500849	2	Capra hircus
241	15	Q6UC74	62510720	1	Cervus elaphus
240	15	P01240	130939	15	Ovis aries
237	15	CAA24561	1619610	11	Rattus norvegicus
237	15	ABK54367	117650786	1	Bubalus bubalis
236	15	1004240A	223892	17	Mus musculus

The most similar sequences are other fish prolactins, but further down the list are hits to mammalian prolactins including human, mouse and rat. For the mouse human and rat sequences in this output, the protein sequence that is shown is an arbitrarily chosen member of a non-redundant set. The NCBI reference sequence (RefSeq) is usually the most useful record from one of these redundant sets. The "Keep only" pull-down can be adjusted so that only the search results against the RefSeq database are shown.

Set the “Keep only” pull-down list on the Best hits BLink output to “REFSEQ” and click the “Select” button.

The screenshot shows the NCBI BLAST interface. The query is 'gi|31088862 prolactin [Danio rerio]' and the matching GI is '28848616'. The 'Keep only' dropdown menu is open, showing 'REFSEQ' selected. The 'Cut-Off' is set to 100. The 'Select' button is visible. The 'New search by GI' field contains '31088862'. The 'Conserved Domain Database hits' table is partially visible below the dropdown.

SCORE	P	ACCESSION	GI	N	ORGANISM
1077	31	AAH92358	62185655	-	6 <i>Danio rerio</i>
999	28	P29235	130931	-	2 <i>Hypophthalmichthys nobilis</i>

The screenshot shows the NCBI BLAST results page for the query 'gi|31088862 prolactin [Danio rerio]'. The 'Keep only' dropdown is set to 'REFSEQ'. The 'Conserved Domain Database hits' table is displayed, showing a list of proteins from various organisms with their scores and accessions.

SCORE	P	ACCESSION	GI	N	ORGANISM
1066	31	XP_001...	125812645	-	5 <i>Danio rerio</i>
560	21	NP_001...	118344638	-	1 <i>Takifugu rubripes</i>
306	15	NP_990797	49169789	-	2 <i>Gallus gallus</i>
290	15	NP_001...	112807242	-	2 <i>Felis catus</i>
276	15	NP_001...	74136543	-	2 <i>Monodelphis domestica</i>
269	15	NP_001...	114052168	-	11 <i>Macaca mulatta</i>
269	15	XP_001...	114605661	-	46 <i>Pan troglodytes</i>
269	15	XP_545363	74004132	-	2 <i>Canis familiaris</i>
267	15	NP_000939	4506105	-	9 <i>Homo sapiens</i>
263	15	NP_999091	47522634	-	2 <i>Sus scrofa</i>
242	15	NP_776378	46810277	-	12 <i>Bos taurus</i>
240	15	NP_001...	57164329	-	3 <i>Ovis aries</i>
236	15	NP_036761	6981404	-	18 <i>Rattus norvegicus</i>
227	15	NP_035294	6755164	-	24 <i>Mus musculus</i>

The resulting output provides a list of the best matching RefSeq proteins from each organism. The human, mouse and rat prolactin sequences are easily identified. The linked accessions (NP\_000939, NP\_035294 and NP\_036761) will retrieve those records. The linked SCORE will launch BLAST 2 Sequences to compare the zebrafish sequence with the listed one from the other species.

## Using Related Structure to find a structure model

The Related Structure shortcut on the Links menu of protein records provides access to BLAST results against the protein sequences from the Structure database and is the fastest way of finding a potential structure for a protein in the database.

**Display the Links menu from the zebrafish prolactin protein (NP\_852102) and follow the link to Related Structure.**

NCBI Sequence Viewer v2.0 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

gi|31088862 view NCBI Sequence Viewer v2.0

NCBI Entrez Protein

My NCBI [Sign In] [Register]

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books

Search Protein for [GI:31088862] Go Clear

Limits Preview/Index History Clipboard Details

Display GenPept Show 5 Send to

Range: from begin to end Features:  CDD + Refresh

1: [NP\\_852102](#). Reports prolactin [Danio ...[gi:31088862] BLink, Conserved

[Comment](#) [Features](#) [Sequence](#)

LOCUS NP\_852102 210 aa linear VRT 21-NOV-2006

DEFINITION prolactin [Danio rerio].

ACCESSION NP\_852102

VERSION NP\_852102.1 GI:31088862

DBSOURCE REFSEQ: accession [NM\\_181437.2](#)

KEYWORDS .

SOURCE Danio rerio (zebrafish)

ORGANISM [Danio rerio](#)  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;  
Actinopterygii; Neopterygii; Teleostei; Ostariophysi;  
Cypriniformes; Cyprinidae; Danio.

Links

- ▶ Gene
- ▶ Genome Project
- ▶ PubMed (RefSeq)
- ▶ Related Structure
- ▶ UniGene
- ▶ Related Sequences
- ▶ Domain Relatives
- ▶ Nucleotide
- ▶ PubMed
- ▶ Taxonomy
- ▶ LinkOut

NCBI

HOME SEARCH SITE MAP PubMed Blast Entrez Structure Help

Query: prolactin [Danio rerio]  
[gi: 31088862]

List All MMDB sequences, sort by Blast E value and display as Graphic

Show Page 1 at 50 structures per page

14 hits with known structures found

Page 1 of 1

Structure	E Value
<a href="#">1RH5_A</a>	1e-27
<a href="#">1N9D_A</a>	1e-27
<a href="#">1F6F_A</a>	1e-17
<a href="#">1HGU</a>	1e-15
<a href="#">3HHR_A</a>	1e-14
<a href="#">1HMG_A</a>	1e-14
<a href="#">1A22_A</a>	1e-14
<a href="#">1HHH_A</a>	1e-14
<a href="#">1BP3_A</a>	1e-14
<a href="#">1KF9_A</a>	1e-13
<a href="#">1KF9_D</a>	1e-13
<a href="#">1Z7C_A</a>	1e-12
<a href="#">1AXI_A</a>	1e-12
<a href="#">1HUM</a>	click for ...

The Related Structures shows the BLAST alignments of proteins with solved structures. In this case the first two are human prolactin NMR structures. These are the most similar proteins and would be the best structure models. These are lower resolution structures than the X-ray crystal structures for the growth hormones that are listed below these. Notice the drop in E-value (significance) from prolactin (1e-27) to the growth hormone entries (1e-15 to 1e-12).

The structure 1BP3 is an X-ray crystal structure of the human growth hormone in a complex with the extra-cellular portion of the prolactin receptor.

**Follow the linked identifier [1BP3\\_A](#) to the structure summary for 1BP3.**

The structure contains two chains; the A chain, the growth hormone, and the B chain, the extracellular domains of the prolactin receptor.

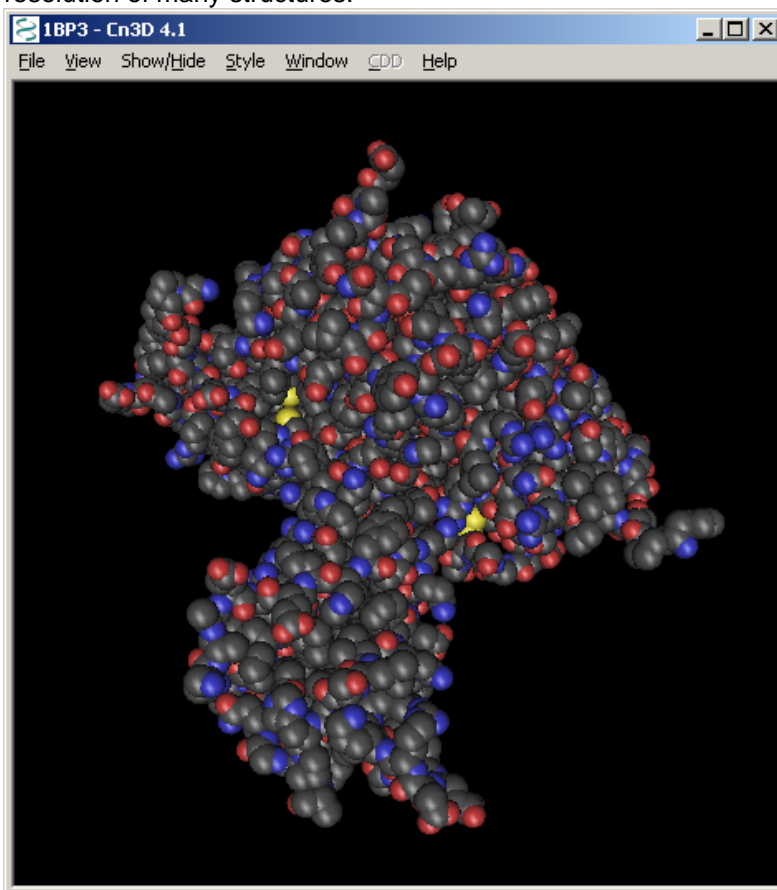
**Click the “View 3D structure” button to display the structure in Cn3D.**

(Cn3D must be installed on your computer first. If you are attending an NCBI workshop, Cn3D should already have been installed. You can install Cn3D on your own computer by following the instructions linked to "Download Cn3D".)

The structure is displayed showing only the alpha carbon backbone. It is colored by secondary structure, and the secondary structure regions are indicated by special objects. The alpha helices are indicated with green cylindrical arrows pointing in the C-terminal direction, the beta strands are indicated by flat tan arrows. You change the rendering of the structure through the Style menu of the viewer.

**Use the Style menu and the Rendering Shortcuts to change to Space Fill. Then use the Style Coloring Shortcuts to color by Element.**

This now more closely resembles a molecular model of the entire complex including the amino acid side chains. Notice that all of the elements are represented except for hydrogen (carbon = black, oxygen = red, nitrogen = blue, sulfur = yellow). Hydrogen atoms are below the limit of resolution of many structures.



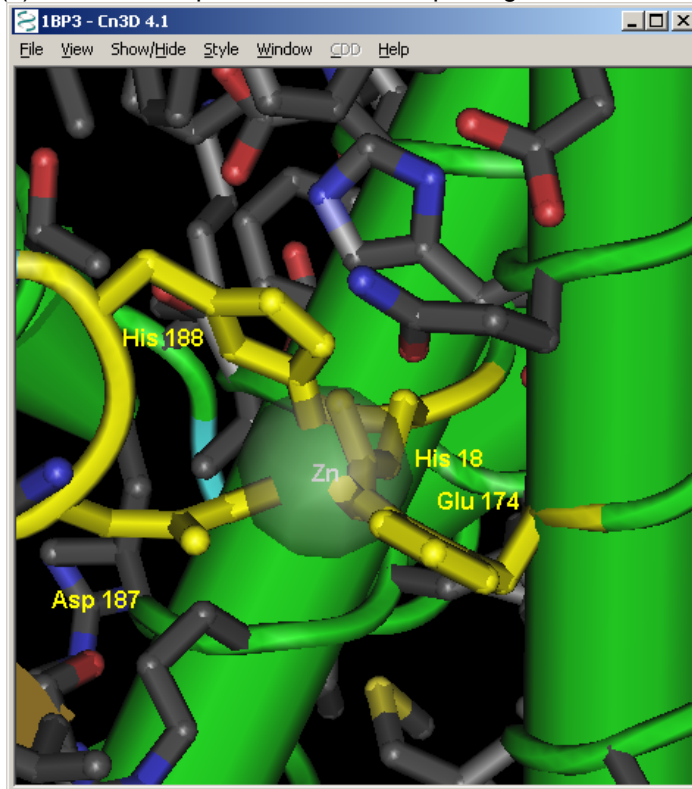
**Restore the earlier rendering and coloring by setting Style:Rendering Shortcuts:Worms and Style:Coloring Shortcuts:Secondary Structure.**

**Hold the mouse button down and rotate the structure so that the zinc ion is visible. Use the View menu and Zoom In to get a closer view of the region surrounding the zinc ion.**

If the zinc ion gets off-center you can hold the shift key down and drag the structure with mouse while holding button down. You can turn on the side chains to see which ones are coordinating the zinc ion.

- Use the **Style:Rendering Shortcuts:Toggle Sidechains** to turn on the amino acid side chains.
- Use **Style: Edit Global Style** to display the **Global Style** menu and change the rendering of the Protein side chains to **Tubes** to make them easier to see.
- You can highlight the four amino acids that are making contact with the zinc by double clicking on the residues in the structure viewer.
- Notice that the residue highlighted in the structure are also highlighted in the **Sequence/Alignment Viewer** window

There is a histidine (h) and a glutamate (e) from the hormone and an aspartate (d) and a histidine (h) from the receptor involved in complexing the zinc ion.



## Example 2: Human MutL Homolog 1

MLH1 is the product of a well-known human disease gene that is mutated in some heritable cancer syndromes.

**Use the global query to retrieve the Swiss-Prot record for human DNA mismatch repair protein MLH1. To save time, you can retrieve it directly using the accession, P40692.**

Of course, you could perform a global text search for mlh1, retrieve the protein results, then use the "Limits" tab as you did with the nucleotide searches to get more precise results.

Entrez Protein - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

NCBI Entrez Protein My NCBI [Sign In] [Register]

All Databases PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search Protein for mlh1 Go Clear

Limits Preview/Index History Clipboard Details

- Use All Fields pull-down menu to specify a field.
- Boolean operators AND, OR, NOT must be in upper case.
- If search fields tags are used enclose in square brackets, e.g., rubella [ti].
- More help on using limits is available [here](#).

Limited to:

Title

exclude TPA  exclude patents

Gene Location Segmented Sequences SWISS-PROT

Modification Date

Modification Date From To

Use the format YYYY/MM/DD; month and day are optional.

P40692 is a record imported from the Swiss-Prot database. Swiss-Prot is a smaller database of highly informative protein records. Many of them are equivalent to review articles on a particular protein. The present record has a large amount of information on the biology of MLH1 including a large list of polymorphisms.

**Examine the FEATURES table of the record and locate several of the polymorphisms in the first 50 residues of the protein.**

FEATURES	Location/Qualifiers
<a href="#">source</a>	1..756 /organism="Homo sapiens" /db_xref="taxon:9606"
<a href="#">gene</a>	1..756 /gene="MLH1" /note="synonym: COCA2"
<a href="#">Protein</a>	1..756 /gene="MLH1" /product="DNA mismatch repair protein Mlh1"
<a href="#">Region</a>	1..756 /gene="MLH1" /region_name="Mature chain" /experiment="experimental evidence, no additional details recorded" /note="DNA mismatch repair protein Mlh1." /FTId=PRO_0000178000."
<a href="#">Region</a>	8..>575 /gene="MLH1" /region_name="MutL" /note="DNA mismatch repair enzyme (predicted ATPase) [DNA replication, recombination, and repair]; COG0323" /db_xref="CDD:30671"

Region 18  
 /gene="MLH1"  
 /region\_name="Variant"  
 /experiment="experimental evidence, no additional details recorded"  
 /note="R -> C (in HNPCC2). /FTId=VAR\_022663."

Region 28  
 /gene="MLH1"  
 /region\_name="Variant"  
 /experiment="experimental evidence, no additional details recorded"  
 /note="P -> L (in HNPCC2). /FTId=VAR\_004433."

Region 31..122  
 /gene="MLH1"  
 /region\_name="HATPase\_c"  
 /note="Histidine kinase-like ATPases; This family includes several ATP-binding proteins for example: histidine kinase, DNA gyrase B, topoisomerases, heat shock protein HSP90, phytochrome-like ATPases and DNA mismatch repair proteins; cd00075"  
 /db\_xref="CDD: [28956](#)"

Region 32  
 /gene="MLH1"  
 /region\_name="Variant"  
 /experiment="experimental evidence, no additional details recorded"  
 /note="I -> V (in dbSNP:rs2020872). /FTId=VAR\_014876." order(34,38,41,61,63,65,67..68,101..104,115,117,122)

Site  
 /gene="MLH1"  
 /site\_type="other"  
 /note="ATP binding site"  
 /db\_xref="CDD: [28956](#)"

Region 35  
 /gene="MLH1"  
 /region\_name="Variant"  
 /experiment="experimental evidence, no additional details recorded"  
 /note="M -> R (in HNPCC2). /FTId=VAR\_004434."

Region 37  
 /gene="MLH1"  
 /region\_name="Variant"  
 /experiment="experimental evidence, no additional details recorded"  
 /note="E -> ELNH (in endometrial cancer; somatic). /FTId=VAR\_004435."

Site 38  
 /gene="MLH1"  
 /site\_type="other"  
 /note="Mg2+ binding site"  
 /db\_xref="CDD: [28956](#)"

Region 44  
 /gene="MLH1"  
 /region\_name="Variant"  
 /experiment="experimental evidence, no

```

additional details
recorded"
/note="S -> F (in HNPCC2; the equivalent
substitution in yeast causes loss of function in
a mismatch repair assay).
/FTId=VAR_004436."

```

Several of these polymorphisms are annotated with the name of a disease or syndrome, for example, hereditary non-polyposis colorectal cancer type 2 (HNPCC2). There is also a polymorphism at position 32 that is cross-referenced to NCBI's dbSNP. In the following sections, you will use some of the pre-computed Entrez relationships to map these polymorphisms onto a 3D structure.

## Links: Related Sequences

The protein record has a "Links" pop-up menu similar to that on the nucleotide record you saw previously. There are also two other hyperlinks; BLink, providing a pre-computed protein BLAST search against nr, and Domains, providing a pre-computed conserved domain analysis of the protein:

Display the "Links" pop-up menu and follow the "Related Sequences" link.

The screenshot shows the NCBI protein record for P40692. The record title is "P40692. Reports DNA mismatch repa...[gi:730028]". The record is displayed in a table format with columns for "Comment", "Features", and "Sequence". The "Links" pop-up menu is open, showing a list of links including "Gene", "Full text in PMC", "Protein (RefSeq)", "Gene Genotype", "GeneView in dbSNP", "Related Structure", "Related Sequences", "Domain Relatives", "OMIM", "PubMed", "Taxonomy", and "LinkOut". The "Related Sequences" link is highlighted.

LOCUS P40692 756 aa linear PRI 13-NOV-2007

DEFINITION DNA mismatch repair protein Mlh1 (MutL protein homolog 1).

ACCESSION P40692

VERSION P40692.1 GI:730028

DBSOURCE swissprot: locus MLH1\_HUMAN, accession [P40692.1 release reviewed](#); class: standard. created: Feb 1, 1995. sequence updated: Feb 1, 1995. annotation updated: Nov 13, 2007.

xrefs: [U07343.1](#), [AAC50285.1](#), [U07418.1](#), [AAA17374.1](#), [U40978.1](#), [AAA82079.1](#), [U40960.1](#), [U40961.1](#), [U40962.1](#), [U40963.1](#), [U40964.1](#), [U40965.1](#), [U40966.1](#), [U40967.1](#), [U40968.1](#), [U40969.1](#), [U40970.1](#), [U40971.1](#), [U40972.1](#), [U40973.1](#), [U40974.1](#), [U40975.1](#), [U40976.1](#), [U40977.1](#), [U17857.1](#), [AAA85687.1](#), [U17839.1](#), [U17840.1](#), [U17841.1](#), [U17842.1](#), [U17843.1](#), [U17844.1](#), [U17845.1](#), [U17846.1](#), [U17847.1](#), [U17848.1](#), [U17849.1](#), [U17851.1](#), [U17852.1](#), [U17853.1](#), [U17854.1](#), [U17855.1](#), [U17856.1](#), [AY217549.1](#), [AA022994.1](#), [BC006850.1](#), [AAH06850.1](#), [S43085](#)

The resulting display is a list of similar sequences in arranged in descending order by BLAST score as with the nucleotide "Related Sequences". Unlike the nucleotide "Related Sequences", the protein similarities typically do find sequences from other species. How exactly the protein sequences in this list are related to the sequence in P40692 is not easily seen. Some of these proteins are identical to P40692; some are very similar over the entire length, some share only a domain in common. All that the list tells you is that the sequences are significantly related. Although it isn't obvious, the first several proteins are, in fact, identical sequences. That set includes corresponding records representing this human protein from at least four different sources; Swiss-Prot, PRF, RefSeq and more than one translation of a GenBank/EMBL/DDBJ sequence. The inclusion of records from outside protein databases plus our own RefSeq database results in a high degree of redundancy at the sequence level in the protein data. The records themselves are not redundant, however, since the annotation on the records is different. When creating a BLAST database and for BLink, identical sequences are represented as a single sequence. The non-redundant database is about 50% smaller than the entire Entrez protein database.

**Change the “Display” drop-down menu to show 500 records. Scroll through the list to see records from other species.**

There are proteins in the list from a wide range of taxa: bacteria, green plants, protozoa, multicellular animals. Although the distance of a particular protein from the top of the list appears to approximate the evolutionary distance from human, keep in mind that some proteins in the list are fragments and may have low scores simply because they are short. You modify the search to find all of the proteins from a particular taxon through the “History” tab.

**Click on the “History” tab.**

The screenshot shows the Entrez Protein search interface. The search history is displayed as follows:

Search	Most Recent Queries	Time	Result
#16	Related Sequences for Protein (Select 730028)	12:33:41	<a href="#">1230</a>
#15	Search mlh1 Field: Title, Limits: SWISS-PROT	12:33:29	<a href="#">5</a>

This is the protein search history that is maintained on our Web server. You can combine the entries in your history with other searches. For example, you can combine the entry for related proteins, called Protein Neighbors, with an organism query.

**Type the number of the entry in your history for the Protein Neighbors in the search box followed by an organism search for mouse. For example**

#16 AND mouse[Organism]

You will need to turn Limits off if you used them previously.

**Then run the search.**

There are several proteins from mouse in the related sequences that are now displayed. Since the related sequences search is combined with another Entrez search, the sorting order is lost. The mouse proteins are listed in arbitrary order, not by their BLAST score with the human MLH1. The BLink option that you will use later makes it much easier to find homologs in other species. It also allows you to see alignments themselves.

## Links: Finding a related structure

Previously you saw that there are a number of sources that contribute to the protein database. One source is the Protein Databank (PDB). PDB is a database of 3D biomolecular structures.

NCBI imports these structures and makes them available in the Entrez system as the Structure database. In addition, protein sequences are extracted from the structures and entries are created in the protein database. This makes it easy to find a structure for a particular protein or a homolog if one exists. Several related proteins in the MLH1 example are PDB entries and have links to the structure database.

**Use the browser “Back” button or the “History” tab to return to the list of related sequences to P40692. Use the “Display” drop down menu to select “Structure Links.” The page will automatically refresh.**

The screenshot shows the Entrez Protein database interface. The search results for 'Protein' are displayed. The 'Display' dropdown menu is open, showing 'Structure Links' selected. The results list includes entries like '1B63' and '1B64' with links to structure records. The interface includes a search bar, navigation tabs (Limits, Preview/Index, History, Clipboard, Details), and a list of related resources on the left.

The new set of results that is displayed contains structure records. Notice that the graphic at the top of the page has changed, and you are now in the Entrez structure database. As with the previous example, the sorting order is lost. Several of these are structures of bacterial DNA mismatch repair proteins.

**Retrieve the structure summary for 1B63 by clicking on the linked identifier.**

Structure Summary, 1B63, 10447 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

NCBI

Structure Summary  
MMDB

PubMed BLAST Structure Taxonomy OMIM Help? Cn3d

**Reference:** Ban C, Junop M, Yang W [Transformation of MutL by ATP binding and hydrolysis: a switch in DNA mismatch repair](#) *Cell* v97, p.85-97

**Description:** MutL Complexed With Adpnp.

**Deposition:** 1999/1/20

**Taxonomy:** [Escherichia coli](#)

**MMDB:** [10447](#) **PDB:** [1B63](#) **Related Structures:** [VAST](#)

View options (Click image to view 3D structure)  
[Download Cn3D!](#)

Molecular components in the MMDB structure are listed below. The icons indicate macromolecular chains, 3D domains, protein classifications and ligands. Please hold the mouse over each icon for more information on the component. You may also click the thumbnails below to view corresponding chains and domains in Cn3D.

Protein  
3d Domains  
Domain Family

Sequence #

HATPase\_c

MutL\_Trans\_MutL

The structure summary page shows a graphic representing the biomolecular chains in the record with the 3D domains and conserved domains mapped onto the chain.

### Display the structure by clicking the image of the structure in the summary.

In order to display the structure, you will need to have the NCBI structure viewer, Cn3D installed. If the viewer is not already installed, follow the hyperlink labeled “*Get Cn3D*” and follow the instructions to install Cn3D.

The 1B63 record is the X-ray crystal structure of the N-terminal portion of the MutL DNA mismatch repair protein from *E. coli*. The default display in Cn3D shows the alpha carbon backbone of the protein colored by the type of secondary structure; alpha helices are green, beta strands are tan, and random coil is blue. There are also 3D objects representing the helices and strands. You can rotate the structure by dragging it with the mouse pointer while holding down the left mouse button. Holding the Shift key down will allow you to move the entire structure by dragging it with the mouse pointer.

You can modify the way the structure is rendered through the “Style” menu of the viewer.

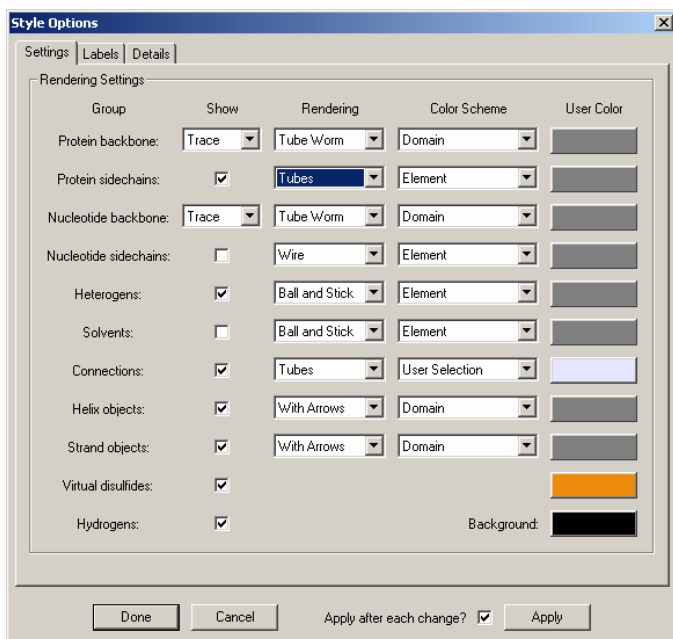
### Use the “Coloring shortcuts” on the “Style” menu to color by Domain.

The color scheme matches that on the structure summary Web page. The purple domain corresponds to the 3D domain also identified as a histidine kinase-like ATPase domain. This domain contains many of the protein polymorphisms associated with disease.

Use the “View” menu to zoom in to the ATPase domain.

An ATP analog is co-crystallized in this domain. Oxygen atoms on the three phosphates of the ATP analog make close contact with a magnesium ion. An amino acid side chain completes the coordination sphere of this metal ion. You can turn on the protein side chains to identify this residue.

Now, use the “Style” menu on the viewer to open the “Global” style dialog box.



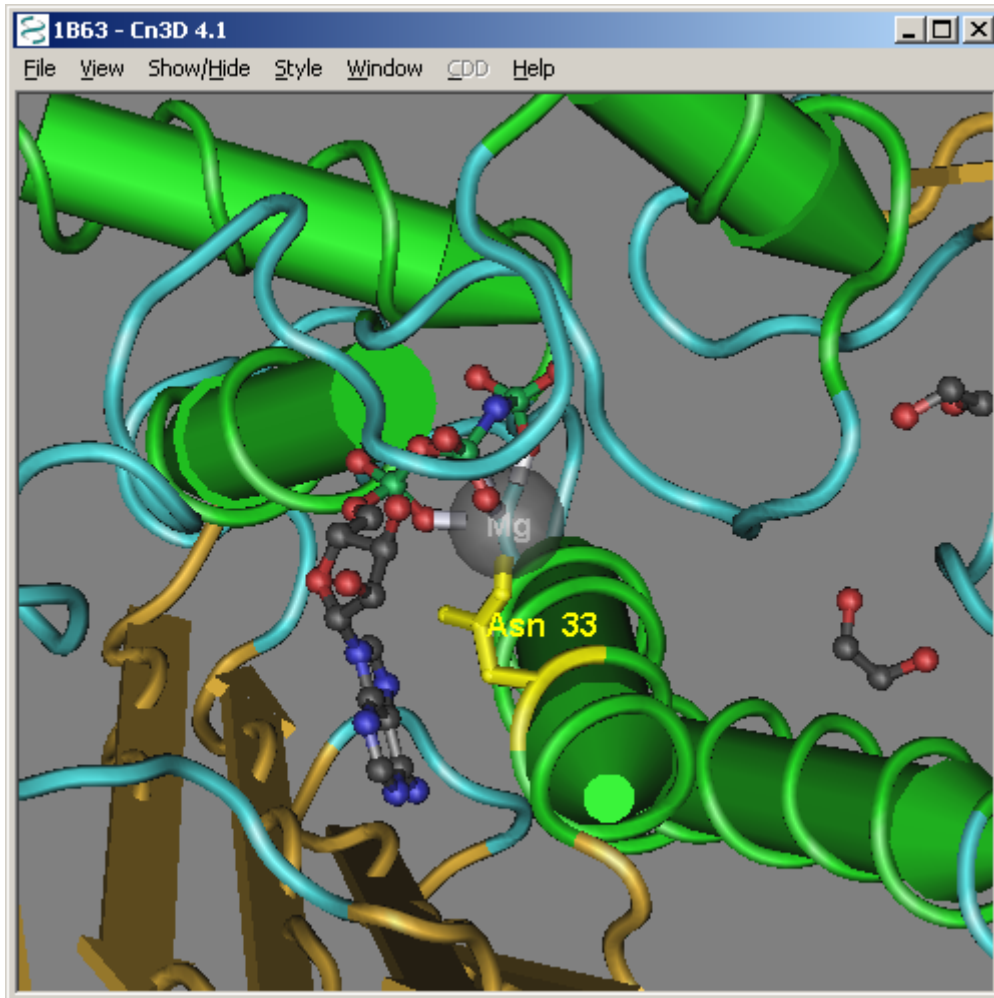
Turn on the protein side chains by checking the box on that line.

You can alter the way the side chains are rendered using the corresponding drop down menu.

Press the “Done” button to close the “Style Options” dialog box.

Zoom in to the region of the protein near the magnesium ion and find the side chain that makes a contact with the metal ion. Double click on this residue to highlight it.

You can identify the residue and its position by looking at the residue now highlighted in yellow in the sequence viewer.



## BLink: non-redundant protein neighbors

Use the global query on the NCBI homepage to retrieve P40692 again and follow the BLink hypertext link.

BLink provides a way of viewing related sequences that is more like a standard protein BLAST output. The top 200 non-redundant related sequences are shown. The source database for BLink is essentially the BLAST nonredundant protein database. This is the Entrez set with the biologically uninteresting patent sequences removed. At the top of the page, is a list of the gi number of sequences in the protein database that are identical to P40692. The graphic alignment shows the regions of the proteins that align to the query. The hyperlinked BLAST score shows the detailed alignment between the two proteins.

Click the “Best Hits” button to limit the display to the best protein match from each species in the list.

In many cases there may be more than one protein in the display from the same species. Sometimes this is because of the presence of paralogous proteins or because there may be differences in the sequences from different sources for the same protein. You can easily identify the best protein match from the tomato, *Solanum lycopersicum*.

Click on the **BLAST score** on the line containing the best tomato (*Solanum lycopersicum*) protein.

The new window shows the BLAST 2 Sequences alignment between the human MLH1 and the best match in tomato. This is a highly significant local alignment that extends nearly the entire length of both proteins.

You can use the “Keep only” drop down menu to limit to various subsets of the protein data, for example PDB to find structures as we did with the Entrez related proteins. Another way of finding structures for related proteins is through the “Related structure” shortcut on the protein links menu.

## Using Related Structures to find a structural model

The “Related structures” shortcut on the protein links menu provides a simple way to find to find related structures.

Retrieve the protein record **P40692** and display the links menu.

The screenshot shows the NCBI Sequence Viewer interface. The search bar contains 'Protein' and the search results list includes 'P40692 Reports DNA mismatch repa...[gi:730028]'. The 'Links' menu is open, displaying a list of links for the protein record.

**Links**

- Gene
- Full text in PMC
- Protein (RefSeq)
- Gene Genotype
- GeneView in dbSNP
- Related Structure
- Related Sequences
- Domain Relatives
- OMIM
- PubMed
- Taxonomy
- LinkOut

**Protein Record Details:**

LOCUS P40692 756 aa linear PRI 13-NOV-2007

DEFINITION DNA mismatch repair protein Mlh1 (MutL protein homolog 1).

ACCESSION P40692

VERSION P40692.1 GI:730028

DBSOURCE swissprot: locus MLH1\_HUMAN, accession [P40692.1 release reviewed](#); class: standard.

created: Feb 1, 1995.

sequence updated: Feb 1, 1995.

annotation updated: Nov 13, 2007.

xrefs: [U07343.1](#), [AAC50285.1](#), [U07418.1](#), [AAA17374.1](#), [U40978.1](#), [AAA82079.1](#), [U40960.1](#), [U40961.1](#), [U40962.1](#), [U40963.1](#), [U40964.1](#), [U40965.1](#), [U40966.1](#), [U40967.1](#), [U40968.1](#), [U40969.1](#), [U40970.1](#)

Follow the “Related Structure” link.

NCBI Sequence Alignment Visualization Service - Graphic Summary - Mozilla Firefox

File Edit View History Bookmarks Tools Help

NCBI

HOME SEARCH SITE MAP PubMed Blast Entrez Structure Help

Query: DNA mismatch repair protein Mlh1 (MutL protein homolog 1)  
[gi: 730028]

List All MMDB sequences, sort by Blast E value and display as Graphic

Show Page 1 at 50 structures per page

13 hits with known structures found

Page 1 of 1

Query

CDs

HATPase\_c MutL\_Trans\_MLH1 MutL

Structure	E Value
<a href="#">1B63 A</a>	1e-57
<a href="#">1NHH A</a>	1e-57
<a href="#">1NHI A</a>	1e-57
<a href="#">1BKN A</a>	1e-57
<a href="#">1BKN B</a>	1e-57
<a href="#">1B62 A</a>	1e-57
<a href="#">1NHJ A</a>	1e-56
<a href="#">1H7S A</a>	1e-45
<a href="#">1H7S B</a>	1e-45
<a href="#">1E66 A</a>	1e-45
<a href="#">1E66 B</a>	1e-45
<a href="#">1H7U A</a>	1e-45
<a href="#">1H7U B</a>	1e-45

Page 1 of 1

The related structures display provides an output similar to BLink, providing a BLAST output sorted by similarity with access to the alignments. The identifiers for the protein chains from structures (1B63 A etc.) on the left hand side link directly to the structure summary. The pink alignment graphic links to a page with the sequence alignment that allows loading the alignment display in Cn3D.

**Click on the alignment graphic for 1B63\_A in the Related Structures display.**

NCBI Sequence-Structure Alignment Visualization Service - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

NCBI

Related Structures

r SVFVGNIPYEATEEQLEKDI FSEVGPVVSFRLVVDRETGKPKGY  
mNMYIGNLSYRVKKEADLRQVMEYVGTVD SVKLIIDRET RKS KGF

PubMed BLAST OMIM Taxonomy Structure Help?

Query: DNA mismatch repair protein Mlh1 (MutL protein homolog 1) [gi: 730028]  
Structure: 1B63Chain A, MutL Complexed With Adpnp.  
MMDB: [1B63 A](#) Reference: [PubMed](#)

View 3D Structure with  (To display structure, download [Cn3D](#))

	10	20	30	40	50	60
gi 730028	8	IRRLDET	VVNR	IAAGEV	IQRPA	NAIKEMIENCLDAKSTS
1B63 A	5	IQVLP	PPQLANQ	IAAGEV	VERPAS	VVKELVENS
	70	80	90	100	110	120
gi 730028	68	IRKEDLD	IVCERF	TTSK	LQSFED	LASISTYGRGEALAS
1B63 A	65	IKKDEL	LALALAR	HATSK	IASLDD	LEAIIISLGRGEALAS
	130	140	150	160	170	180
gi 730028	128	AsYSDGK	-LKAPP	KPCAG	NQGTQ	ITVEDLFYNIATR
1B63 A	125	A-YAEG	RmMNV	TKPA	AHPV	GTTLEVLDLFY
	190	200	210	220	230	240
gi 730028	187	NAGISF	SVKKQ	GETV	ADV	RTLPNASTVD-N
1B63 A	184	RFDVT	INLSH	NGKIV	RQYRA	VPEGGQKEr
	250	260	270	280	290	300
gi 730028	246	ISNAN	YSVKK	C--IF	LLFIN	HRLVESTSLR
1B63 A	242	VADPN	HHTP	ALae	IQCY	VNGRMMRD
	310	320	330			
gi 730028	304	DVMV	HPTK	HEV	HFL	HEESILERV
1B63 A	302	DVMV	HPAK	HEV	RFH	QSRLVHDF

Score(bits) = 204, E\_value = 1e-57  
Aligned Length = 324 , Sequence Identity = 35 %

This display shows the BLAST 2 Sequences alignment between the human protein sequence and the *E. coli* sequence from the 1B63 A chain.

**Click on the “View 3D Structure” button.**

The resulting Cn3D display now shows the 1B63 structure colored by sequence conservation from the alignment of the human MLH1 (bottom sequence) and the N-terminal sequence region of MutL (top sequence). You can use the sequence alignment to map the human residues onto the *E. coli* protein structure. In other words, the human protein is assumed to fold up into a very similar structure; the sequence alignment is used as a proxy for the structural alignment. This is reasonable as long as the proteins are similar at the sequence level. You can confirm the validity of this to some extent by verifying that structurally and functionally significant residues in the structure line up with corresponding residues in the aligned protein sequences.

**Manipulate the structure in the viewer and use the view menu on the viewer to zoom in to the ATP binding site residue; the asparagine (n) at position 33 of the structure. Verify that this residue is aligned with an asparagine in the human sequence.**

You can now look at some of the polymorphisms reported in the FEATURES table of P40692 in the context of the structure of the protein. Notice that the isoleucine to valine change at position 32 of the human protein, which is not reported as associated with human disease, occurs on the side of the helix containing the ATP binding site residue that is away from ATP. In fact, the residue in that position in the *E. coli* protein is a valine. A disease causing polymorphism in the human protein replaces the proline at position 28 of the human protein with a leucine. The proline in this position, which is conserved in *E. coli*, may be important in constraining the turn at the end of the helix.

## NCBI Exercises Set 2

### ***NCBI Genomic Resources***

Albumins constitute a small family of genes in mammals. The human, mouse and rat genomes, and probably all mammals contain at least four members: albumin, alpha-fetoprotein, afamin (alpha albumin) and the vitamin D binding protein. We will look at various aspects of this gene family in the NCBI genome resources.

### **UniGene and Gene**

UniGene is the best NCBI resource to identify the gene (or suspected gene) that corresponds to a particular database sequence. This is especially true for ESTs where there may be no annotations on the sequence, but may also be important for other sequences where the annotation may be incomplete or obsolete. Database identifiers for UniGene searches may come from BLAST output or from microarray (hybridization) data. For example, an mRNA that hybridized to the EST sequence with accession number BG618460 was highly expressed in a human liver tumor sample. We can identify this gene using UniGene.

**Retrieve BG618460 from the EST database. You can use the search box on the NCBI homepage and retrieve the link to EST on the global query page.**

Is there any information indicating what gene this is?

**Now link to UniGene from the "Links" menu in the upper right.**

What is the name of this gene?

**Link to "Gene" from the UniGene cluster links menu.**

What is the function of this protein?

**Go back to UniGene.**

**Look at the ESTs in this cluster. How many are there? A pair of ESTs (a 5' and 3' read) that come from the same clone ID are T58928 and T58869. You'll need to display all ESTs and scroll down to see these. Also, identify the RefSeq mRNA in the cluster. You should be able to recognize the RefSeq by the characteristic accession.**

**Link to the BLAST homepage and use BLAST 2 Sequences to align the 5' and 3' reads to the RefSeq mRNA.**

Notice the mismatches that are most likely due to sequencing errors in the ESTs. Expression information is implied by the sources of the cDNA libraries in a particular cluster.

Follow the "Expression profile" mapping link to see a "virtual Northern" display of the counts from this cluster in UniGene libraries.

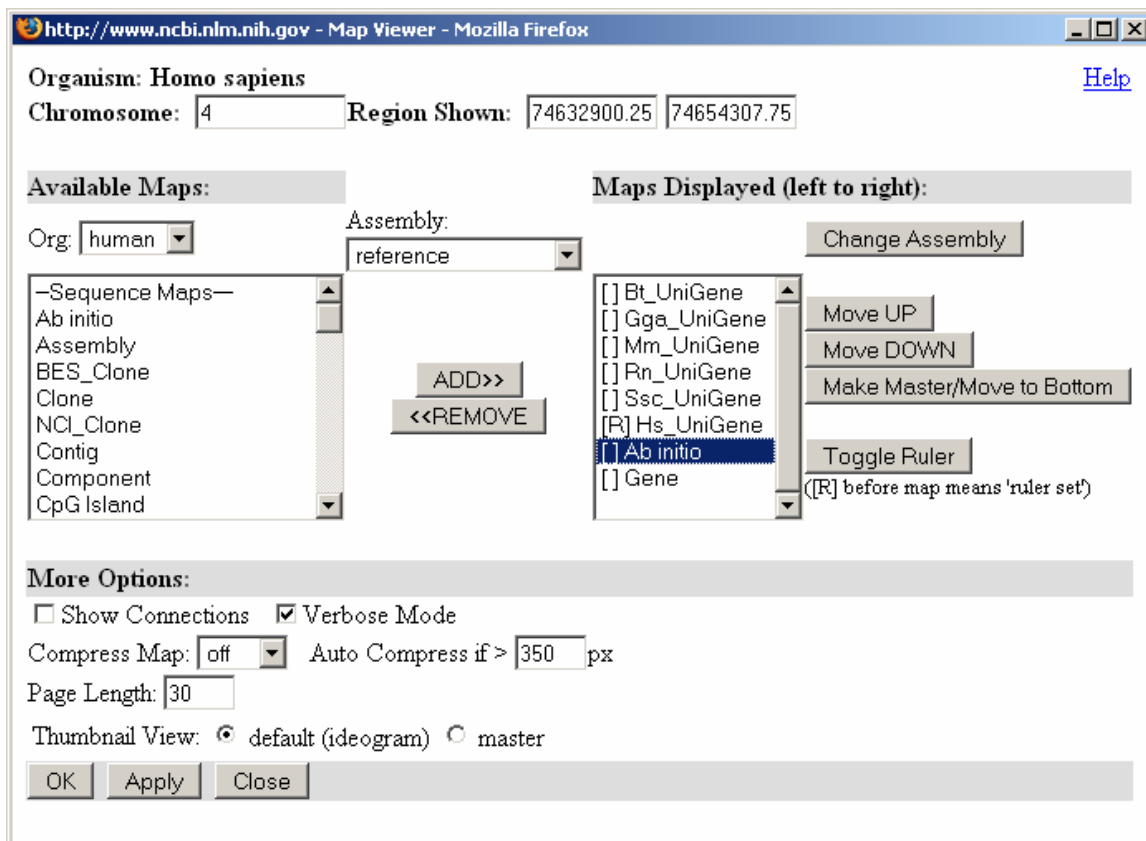
What library pool shows the highest relative expression of this gene?

## Map Viewer

From the "Gene" page for human Alb, use the "Links" menu to display this gene in the Map Viewer.

What chromosomal region is this? What maps are displayed? You can click on the map name at the top to learn more about the information displayed for each map. The UniGene map shows the density of EST hits on the genome. Generally the peaks in this histogram highlight the exons of expressed genes. Notice that there are some hits that don't correspond to the exons shown in the gene model on the Genes map. What could these represent?

You may want to use the "Maps and Options" dialog box to remove all except the "Gene" map from the display for easier viewing. The "Maps and Options" can be accessed through the button on the upper right of the maps. Click "OK" once you adjust the maps.



Use the zoom graphic on the left hand side of the map viewer to zoom out and display two other members of the albumin gene family, AFP and AFM.

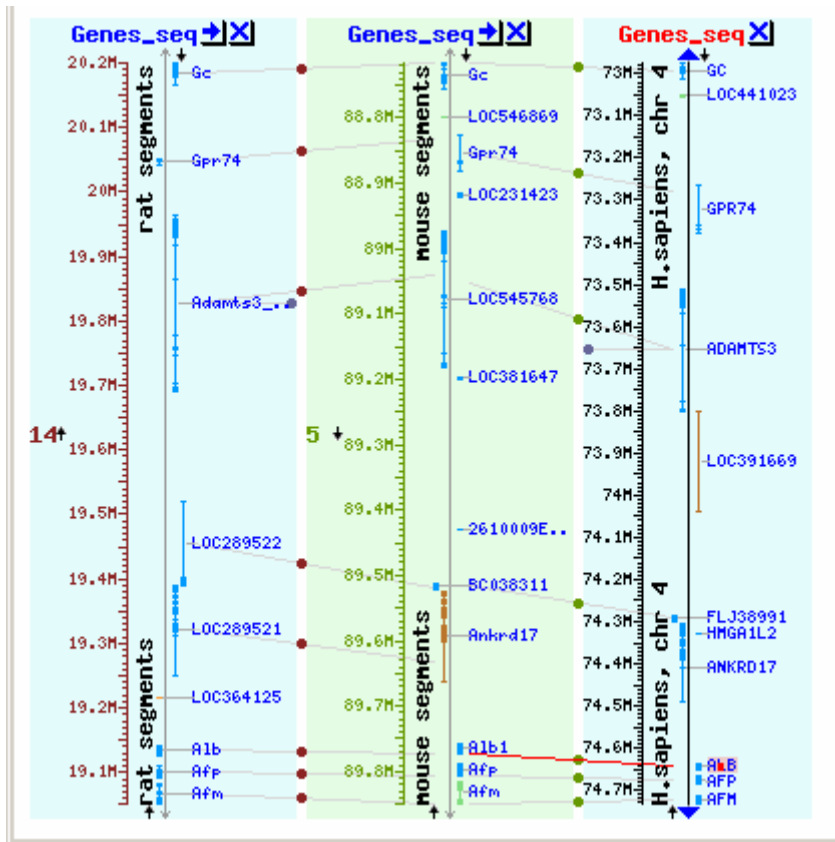
Are these in the same orientation?

The fourth member of this small family, the vitamin D binding protein, also called group-specific component (GC), is somewhat removed from these on chromosome 4.

Display the entire region between GC and AFM by typing these symbols in the "Region Shown" boxes on the left-hand-side and pressing the "Go" button.

Use the "Maps and Options" link to add the mouse and rat gene maps to the display.

This display shows gene-to-gene connections of the three genomes. Removing the UniGene map may make this easier to view. These connections are created through the HomoloGene database. Notice that the structure of the albumin gene family is conserved in these three mammalian genomes.



## Genomic BLAST pages

Some of the higher genome BLAST pages are helpful because they allow the genomic context of the BLAST search to be displayed in the Map Viewer. We can use the human albumin RefSeq transcript to identify the homolog in the rat genome.

Follow the link from the BLAST home page (<http://www.ncbi.nlm.nih.gov/blast/>) to the rat genome BLAST page.

Type the accession number for the human albumin precursor, NM\_000477, into the search box on the BLAST form.

### Run the search without changing the default settings.

This will use megablast against the assembly. This is faster but less sensitive than ordinary blastn when run in contiguous word-hit mode (word size =28, exact match required) as it is here.

### Format your results.

Were you able to find the rat homolog?

### Repeat the search. This time choose the cross-species megablast option.

You should have found some hits this time. The graphical overview shows that some parts of the human albumin transcript did not find any significant matches in the rat. Albumins are not highly conserved genes. Notice that the alignments shown in the output are some of the exons of the rat albumin gene. The exon matches are ordered by significance; the longest and best conserved exons are shown first. Another more interesting way to display these is by the position in the genome.

**Display your results in the rat Map Viewer by linking through the linked RefSeq identifier to the contig on rat chromosome 14.**

Notice that not all exons were found and that none of the other gene family members were identified. You can potentially identify the other members of the gene family in rat by searching with the human protein using the translation of the genome

**Go back to the rat genome BLAST page and type the human RefSeq protein accession, NP\_000468, in the search box.**

**Select the translating BLAST search, tblastn, from the program selection drop down menu.**

**Display these results in the Map Viewer as with the nucleotide results.**

You may need to adjust the zoom level to see your results clearly. What albumin gene family members did you find? If you compare the corresponding region in the human genome and mouse genomes you may notice that the rat and the mouse have an additional member close to afamin.

Albumins are also present in other vertebrates. We can use the specialized genomic BLAST pages to try to find homologs in other organisms. Links to some of the genome specific BLAST pages are available from the BLAST home page (<http://www.ncbi.nlm.nih.gov/blast/>). For some genomes, it's necessary to follow the BLAST link from the Map Viewer homepage. This is the link on the BLAST page labeled "[list all genomic BLAST databases](#)".

## BLAST Assembled Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

- |   |  |  |
|---|--|--|
| <input type="checkbox"/> <a href="#">Human</a>                | <input type="checkbox"/> <a href="#">Oryza sativa</a>            | <input type="checkbox"/> <a href="#">Gallus gallus</a>   |
| <input type="checkbox"/> <a href="#">Mouse</a>                | <input type="checkbox"/> <a href="#">Bos taurus</a>              | <input type="checkbox"/> <a href="#">Pan troglodytes</a> |
| <input type="checkbox"/> <a href="#">Rat</a>                  | <input type="checkbox"/> <a href="#">Danio rerio</a>             | <input type="checkbox"/> <a href="#">Microbes</a>        |
| <input type="checkbox"/> <a href="#">Arabidopsis thaliana</a> | <input type="checkbox"/> <a href="#">Drosophila melanogaster</a> | <input type="checkbox"/> <a href="#">Apis mellifera</a>  |

This links to the Map Viewer Homepage (<http://www.ncbi.nlm.nih.gov/mapview/>).

**Follow the link from the BLAST or Map Viewer home page to the chicken (Gallus gallus) genome BLAST page.**

**Type the accession number for the human albumin precursor, NP\_000468, into the search box on the BLAST form.**

**Select the translating BLAST search, tblastn, from the program selection drop down menu.**

**Leave the database menus set on "genome" and click the BLAST button.**

**Format your results.**

Were you able to find matches in the chicken genome? How many potential homologs did you find?

Repeat the search for the albumin family using the horse genome BLAST page. The horse genome BLAST page is linked through the “B” icon on the Map Viewer homepage.

## Using NCBI BLAST

### Identifying sequences

Michael Crichton's fantasy about cloning dinosaurs, *Jurassic Park*, contains a putative dinosaur DNA sequence. Use basic nucleotide BLAST against the nucleotide database, nr, to identify the real source of the following sequence from the novel. You can retrieve the sequence from the NCBI ftp site:

<ftp.ncbi.nih.gov/pub/FieldGuide/jurassic.txt>

**Select, copy and paste the sequence into the BLAST form window and run the search against the nr(nt) database. Use the default Megablast algorithm.**

What is the sequence that Michael Crichton used?

This search is an example of the most common use of nucleotide-nucleotide BLAST: sequence identification, establishing whether an exact match for a sequence is already present in the database.

Mark Boguski, who was at the NCBI at the time, noticed this obvious contaminant and supplied Crichton with a better sequence for the sequel, *The Lost World*. You can also retrieve this sequence from the NCBI ftp site:

<ftp.ncbi.nih.gov/pub/FieldGuide/lostworld.txt>

**Select, copy and paste the sequence into the BLAST form window and run the search.**

Identify the most likely source of this sequence using nucleotide-nucleotide BLAST.

Mark imbedded his name in the sequence he provided. To see Mark's name, use the translating BLAST (blastx) page with the sequence. (Look for MARK WAS HERE NIH).

The most important use of the translating BLAST services is to look for similar proteins (identify potential homologs) in other species.

## Short Nucleotide Sequences and Algorithm Parameters

A frequent use of nucleotide-nucleotide BLAST is to check the specificity oligonucleotides for hybridization or PCR. The goal most people have when doing this is to make sure that the primer will give a unique product from the target genome or cDNA population. Because BLAST is local and searches both strands, one can simply concatenate a pair of +/- strand primers and use them in a single search.

**Combine the following pair of candidate PCR primers in a nucleotide-nucleotide search against the nr(nt) database. Be sure to choose blastn (Somewhat similar sequences) as the BLAST program under “Program selection.”**

F12 GTCAAGTGGCAACTCCGTCAG

R8 TTGAGAGATGGATTGTTGCTC

**To prevent false matches that overlap the forward and reverse primer sequences, type ten or more “n’s” between the sequences when using them as a query.**

GTCAAGTGGCAACTCCGTCAGnnnnnnnnnnTTGAGAGATGGATTGTTGCTC

**Retrieve the results and identify the gene amplified by these primers.**

What is the predicted size of the product that would be amplified by PCR from cDNA (RT-PCR)? How could you distinguish the products amplified from genomic DNA versus cDNA?

You can also try these primers against the human genomic plus transcript database to get a clearer view of the product predicted from genomic DNA in the Map Viewer.

**Now try these modified primers in standard nucleotide-nucleotide BLAST. There is one mismatch in each near the middle.**

F12\_mod GTCAAGTGGCgACTCCGTCAG

R8\_mod TTGAGAGATGtATTGTTGCTC

GTCAAGTGGCgACTCCGTCAGnnnnnnnnnnTTGAGAGATGtATTGTTGCTC

Notice that the previous hits are completely missing. This is because the default word size setting requires an exact match of 11 before extensions can occur. A mismatch in the middle of a 21-mer will prevent any initial word hits. There is an automatic adjustment for short sequences that will allow these hits with mismatches to be found. However the sequence with the linking “n’s” is too long to trigger the adjustment.

**Run the search again with the forward and reverse primers as separate sequences. Copy and paste the following FASTA formatted primers in the search box.**

```
>F12_mod
GTCAAGTGGCgACTCCGTCAG
>R8_mod
TTGAGAGATGtATTGTTGCTC
```

Your results should now display a message that your search parameters were adjusted to search for a short input sequence, and you should see results for both primers. Notice that although there are now hits, the original hits are still missing. This is because the expect value of the mismatch hits is above 10.

You can manually adjust search parameters to short sequence setting through the “Algorithm parameters” section of the nucleotide BLAST form. After adjusting these, the search with the concatenated mismatched primer will work.

Go back to the BLAST form. Click on the reset page link at the top to restore the default settings. Then select blastn under “Program Selection” and expand the “Algorithm parameters” section of the form. Make the following changes.

- Uncheck the box next to “Automatically adjust parameters for short input sequences.”
- Increase the expect threshold to 100.
- Set the Word size to 7
- Set the Match Mismatch Scores to 1, -3
- Uncheck any Filter options

Now run the search again with the concatenated mismatch primers.

GTCAAGTGCCgACTCCGTCAGnnnnnnnnnnTTGAGAGATGtATTGTTGCTC

The screenshot shows the 'Algorithm parameters' section of the NCBI BLAST form. It is divided into three sub-sections: General Parameters, Scoring Parameters, and Filters and Masking.

- General Parameters:**
  - Max target sequences: 100
  - Short queries:  Automatically adjust parameters for short input sequences
  - Expect threshold: 10
  - Word size: 11
- Scoring Parameters:**
  - Match/Mismatch Scores: 2,-3
  - Gap Costs: Existence: 5 Extension: 2
- Filters and Masking:**
  - Filter:  Low complexity regions,  Species-specific repeats for: Human
  - Mask:  Mask for lookup table only,  Mask lower case letters

The screenshot shows the 'Algorithm parameters' section of the NCBI BLAST form after modifications. A note at the top right states: "Note: Parameter values that differ from the default are highlighted in yellow".

- General Parameters:**
  - Max target sequences: 100
  - Short queries:  Automatically adjust parameters for short input sequences
  - Expect threshold: 100
  - Word size: 7
- Scoring Parameters:**
  - Match/Mismatch Scores: 1,-3
  - Gap Costs: Existence: 5 Extension: 2
- Filters and Masking:**
  - Filter:  Low complexity regions,  Species-specific repeats for: Human
  - Mask:  Mask for lookup table only,  Mask lower case letters

Do you find the original hits now?

## Protein-protein BLAST and Short Peptides: ELVIS lives

As the database grows, so does the number of chance occurrences of amino acid motifs that spell out words or people's names in single-letter amino acid codes. One such name motif is ELVIS. In this example we will count the number of occurrences of ELVIS in the default protein database. The automatic adjustment of search parameters will allow us to find matches with this short peptide

**Type ELVIS in the search box on the blastp form.**

**Expand the Algorithm parameters section and adjust the number of Max target sequences to 1000 or more to include all Elvises.**

**Run the search.**

What is the expect value for an exact match to ELVIS? The number of Elvises increases in a linear fashion with the size of the database in accordance with the random behavior of protein sequences.

**Click on the “Edit and Resubmit” link at the top of the BLAST form. Examine the Algorithm parameters section to see how the settings were adjusted to search with this short peptide.**

## PSI-BLAST and Conserved Domains

The histine kinase-like ATPase domain (HATPase\_c) is present in a wide variety of proteins with quite different functions. These include bacterial sensor histidine kinases, DNA mismatch repair proteins, topoisomerases, DNA gyrases and 90 KDa heat shock protein homologs. We can use PSI-BLAST to demonstrate the similarity among these proteins that is not apparent with ordinary BLAST.

**Use the human DNA mismatch repair protein MLH1 (NP\_000240) in an ordinary blastp search and examine the conserved domain results to verify the presence of the HATPase\_c domain.**

**From the results of the above search, click the “Edit and Resubmit” link and make the following changes to prepare to run a PSI-BLAST search with just the region of MLH1 that corresponds to the HATPase\_c domain**

- **Set the query subrange in the boxes on the right hand side of the form. Use 32 as the “From” coordinate and 122 as the “To” coordinate.**
- **Change the database to “swissprot.”**
- **Change the “Program Selection to PSI-BLAST.”**
- **Expand the “Algorithm parameters” section and set the “Max target sequences” to 5000.**

**Now click the BLAST button to run the first iteration of PSI-BLAST and examine the results.**

The results are just the blastp results that are formatted for PSI-BLAST. Notice that the descriptions sections of the results is divided into two sections. The upper section contains the sequence with alignments that will be used to generate the position specific score matrix in the next iteration of PSI-BLAST. These sequence alignments have e-values less than 0.005. This cut-off is empirically determined to give good results in PSI-BLAST searches. All of the proteins above this threshold in the first iteration are DNA mismatch repair proteins PMS, MutL and HexB homologs. Just below the PSI-BLAST threshold with e-values ranging from 0.008 to 6.0 are several bacterial signaling histidine kinases. Some of these have marginally significant e-values in ordinary BLAST but many are not distinguishable from chance matches.

**Now click the “Run PSI-BLAST iteration 2” button to run the second iteration of PSI-BLAST and examine the results.**

There are now new proteins less than the 0.005 threshold. Notice that these are now marked with a “New” graphic while the proteins found in the previous iteration are marked with a green ball. Many of the new proteins are topoisomerases or DNA gyrases. There are also many more gyrases and topoisomerases just above the 0.005 threshold.

**Retrieve a few of the new proteins in Entrez by clicking on the linked identifier and verify that that they contain the HATPase\_c domain.**

**Click the “Run PSI-BLAST iteration 3” button to run the third iteration of PSI-BLAST and examine the results.**

Again there are new proteins, not only more gyrases and topoisomerases, but also signaling histidine kinases and HSP90 chaperonins.

**Continue to run PSI-BLAST iterations until you have collected some plant phytochrome and ethylene receptor proteins below the 0.005 threshold.**

These are plant signaling proteins. As these results show, plant ethylene receptors and phytochromes are related by sequence similarity to the two component sensor kinase system of bacteria.

**Demonstrate the similarity between the HATPase\_c domain of the *E. coli* sensor protein PhoR (PHOR\_ECOLI, P08400) and plant ethylene receptors by performing a first iteration PSI search against swissprot. Use a query subrange on PhoR of 318 to 421.**

**Now, continue to run PSI-BLAST iterations until the plant phytochromes appear.**

The number of iterations should be fewer than when using the MLH1 protein as a query.

**Retrieve the protein record for an ethylene receptor (ETR1\_LYCES, ETR1\_ARATH) and a phytochrome (PHYA\_ARATH, PHY\_PICA) by following the link to Entrez. Compare their domain structures by following the links to the pre-computed Conserved Domains results.**

What three domains do they have in common?

**Retrieve a protein record for one of the bacterial sensor proteins (PHOR\_ECOLI) and examine its domain structure.**

Notice that the plant proteins and the bacterial protein all contain the histidine kinase domain (HATPase\_c) and the HisKA (phosphoacceptor) domain. In the classic two component bacterial system, the HisKA domain is phosphorylated on a conserved histidine residue by the HATPase\_c domain in response to an external signal. This phosphate is then transferred from the HisKA

domain of the sensor protein to a conserved receiver domain on a separate response regulator protein. In the case of PhoR the response regulator is PhoB.

**Retrieve the *E. coli* PhoB protein (PHOB\_ECOLI, P0AFJ5) and examine its domain structure as before.**

Notice the presence of the receiver domain (REC) and a DNA binding effector domain (trans\_reg\_C) in PhoB. In the plant ethylene receptors examined previously there is a receiver domain is on the receptor itself, but the effector domain present in PhoB is lacking. The plant ethylene receptors apparently mediate their effects through the MAP-kinase pathway. Unlike the ethylene receptors, the phytochromes function as serine/threonine kinases but also appear to share an ancestry with bacterial histidine kinases.

## Translating BLAST searches, mining polymorphisms

The prion protein is found in high concentrations in the brains of humans and other mammals. In certain degenerative neurological diseases, prion proteins aggregate into polymers. Several of these prion diseases seem to be transmissible. Perhaps the most remarkable aspect of these is that the infectious agent appears to be an aberrant form of the prion protein itself. Bovine spongiform encephalopathy (BSE) is one of the transmissible prion diseases that has received much recent notoriety. There are a number of polymorphisms that have been identified in the prion proteins for several mammals, notably human, mouse, and sheep. Some of these are associated with inherited prion diseases and some with susceptibility to transmissible forms.

**Retrieve the SWISS-PROT record for the human prion protein (PRIO\_HUMAN, P04156) and look at the FEATURE table to see the various polymorphisms.**

Notice the methionine / valine polymorphism at position 129. The amino acid at this position affects the particular disease phenotype when another disease causing mutation is present. People who are heterozygous at this position appear to be more resistant to *kuru*, one of the transmissible encephalopathies. There is population genetic evidence that they may have been balancing selection for heterozygotes at this position during human evolution. The EST data for human represents a large number of individuals and can be used as a resource for identifying nucleotide polymorphisms. In this case, we can investigate the prevalence of the two alleles at position 129 of the prion protein in the EST data for human. We will use one of the formatting options to make the different alleles easier to identify.

**Set up and run this search by following these steps:**

- **From the BLAST homepage, link to the tblastn form “Search translated nucleotide database using a protein query.”**
- **Type the prion protein accession number, P04156, in the search text area.**
- **Use the “Query subrange” boxes to use only residues 100 to 160.**
- **Choose the “Expressed sequence tags (est)” database.**
- **Type “human” in the Organism limit box and choose human (taxid:9606) from the resulting list to limit to human sequences.**
- **Open the Algorithm parameters section and set the Max target sequences to 1000**
- **Turn off the “Low complexity” filter option.**

- **Click the BLAST button to run the search.**
- **Immediately click the “Formatting options” link at the top of the intermediate page.**
- **Set the alignment view to “Query-anchored with dots for identities.” This is a stacked pairwise alignment format that makes it easy to see changes relative to the query sequence in all the database hits at once.**

Click the “View report” button to display the results.

Look at the alignments to see how the query-anchored format helps to investigate changes in sequences. Find position 129 in the query. Which amino acid is most prevalent at position 129?

## WGS and Trace Archive Data in Entrez and BLAST

Verify that nearly all of the rabbit DNA records in the NCBI database are whole genome shotgun. You can retrieve all nucleotide rabbit sequences by using the Limits tab and setting the field restriction in the pull-down list to organism. You can further limit to genomic DNA through the “molecule” pull-down list.

How many records are there?

**Follow the link to the “CoreNucleotide” results before continuing. Now restrict to whole genome shotgun records by adding the following query term to your search.**

wgs[Properties]

The overall search performed now is

rabbit [Organism] AND biomol\_genomic[Properties] AND wgs[Properties]

The first record is the master record for the project that gathers all of the contigs. You can get only this record by adding wgs\_master[Properties] to the search.

**Retrieve the first contig record in you list and verify that it is unannotated –no genes or other features.**

Using BLAST, Spidey and Salign to annotate wgs

You can find the genomic sequences corresponding to a rabbit (*Oryzolagus cuniculus*) mRNA sequence by using BLAST to search the wgs database. A sequence that demonstrates this is the rabbit apolipoprotein A-1 mRNA (NM\_001101687).

- **From the BLAST homepage select the blastn page**
- **Type NM\_001101687 in the “Search box” and select wgs as the database.**
- **Use the Organism limit feature to limit to rabbit (taxid:9986)**
- **Expand the Algorithm parameters section and set the e-value threshold to 1e-12.**

- **Run the search and re-format your results using the “CDS feature” option and “Pairwise with identities” Alignment view option.**

Your search should hit one wgs contig (AAGW01335306). How many exons did you identify in each?

**Use the sort by “Query start position” to put the exons in the genomic order on AAGW01335306.**

This is a rather primitive gene model because it does not constrain the alignment breaks to splice junctions.

**Use the same mRNA and genomic sequences as above, make gene models using the spliced alignment tools Spidey and Splign and compare them to the BLAST results.**

The spliced alignment tools place two of the exon-intron boundaries at slightly different points than BLAST alignments.

## Trace Archive

Some sequences are only available through the NCBI trace archive.

<http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?>

These data can be retrieved by species code or trace number. The most important way to search these is through the Trace archive megablast pages. Both standard (contiguous) megablast and cross-species megablast are available. These are linked through the BLAST tab on the main trace archive page or through the BLAST homepage in the Specialized BLAST section.



The screenshot shows the NCBI Trace Archive v4.1 interface in a Mozilla Firefox browser. The page has a navigation menu with tabs for Main, Obtaining Data, Statistics, Tracking, Documentation, Assembly Archive, and BLAST. The BLAST tab is active, showing options for Mega BLAST and Cross-Species Mega BLAST. Below the search bar, there is a section for "News, Events and Notifications" and a table titled "Last week Top 10 Arrivals (02/11/2007 - 02/17/2007)".

Organism	Count
CAVIA PORCELLUS	1,676,926
DELFTIA ACIDOVORANS SPH-1	1,409,615
TARSIVUS SYRICHTA	1,184,313
BACILLUS WEIHENSTEPHANENSIS KBAB4	1,034,406
CANDIDATUS DESULFOCOCCUS OLEOVORANS HXD3	979,232
STENOTROPHOMONAS MALTOPHILIA R551-3	780,912
MACROPUS EUGENII	741,027
GLYCINE MAX	505,978
THERMOSIPHO MELANESIENSIS BI429	419,225
FERVIDOBACTERIUM NODOSUM RT17-B1	382,741

We can use the cross-species page to find an HSP70 gene homolog in the sea lamprey (*Petromyzon marinus*) traces

- **Go to the BLAST homepage and choose the trace archive search from the Specialized BLAST section.**
- **Enter the accession number for the human HSP70 1A mRNA Reference Sequence (NM\_005345) in the search box on the BLAST form.**
- **Choose Petromyzon marinus-WGS as the database and set blastn as the program.**
- **Click the BLAST button to run the search.**

Because HSP70 is well conserved it is easy to find homologs in the sea lamprey at the nucleotide level using the human sequence. Many less well conserved genes may only be identified at the protein level. Unfortunately the large size of the trace databases makes translating searches impractical.

## New BLAST Displays

### TreeView

The treeview display in BLAST will not always produce reasonable phylogenetic species trees or gene trees because the alignments are not multiple sequence alignments and don't necessarily include all residues. Nevertheless searches with complete mitochondrial genomes often reproduce accepted phylogenetic groupings.

**From the BLAST homepage, choose the blastn page. . Select the RefSeq genomic database from the database pull-down list and put the accession for the wolf mitochondrial genome (NC\_008092) in the "Search" box as a query.**

The Refseq genomic database contains chromosome (NC\_) RefSeqs including plastid genomes, mitochondrial genomes and chromosomes for prokaryotic genomes.

**Use the following Entrez limit to restrict to the mammalian order carnivora (dogs, cats, seals, hyenas, weasels etc.).**

carnivores[organism] NOT gene in genomic[properties]

This last term, "NOT gene in genomic[properties]", eliminates hits to mitochondrial insertion sequences present in the dog genome.

**Run the search. Click on the "Distance Tree of Results" link under the BLAST graphic to display the tree. Compare the groupings to the classification of the carnivores in the NCBI Taxonomy database.**

The family groupings correspond to those in the tree. However, many families of carnivores are not represented because the mitochondrial genomic sequences are not available yet.

## New View of Results and Genome and Transcript Databases

The new human genome and transcript database provides direct access to the human genome through the main BLAST page. The new view options provide a more organized and sortable presentation of the results.

**Use nucleotide-nucleotide BLAST (blastn) to search the human genome plus transcript database with the human alcohol lactate dehydrogenase B (LDHB) transcript (NM\_002300).**

**Use the new sorting options and summary statistics to identify the functional multi-exon gene by sorting using the “Total Score” column.**

On which chromosome is the functional gene? How many exons does it have?

**Use the “Sort alignments” feature to sort by “Query start position” to get the exons in genomic order.**

On which chromosome is the longest retrocopy pseudogene?

**Follow the linked identifiers to the human Map Viewer to display the hits for both the functional gene and the retrocopy pseudo gene.**