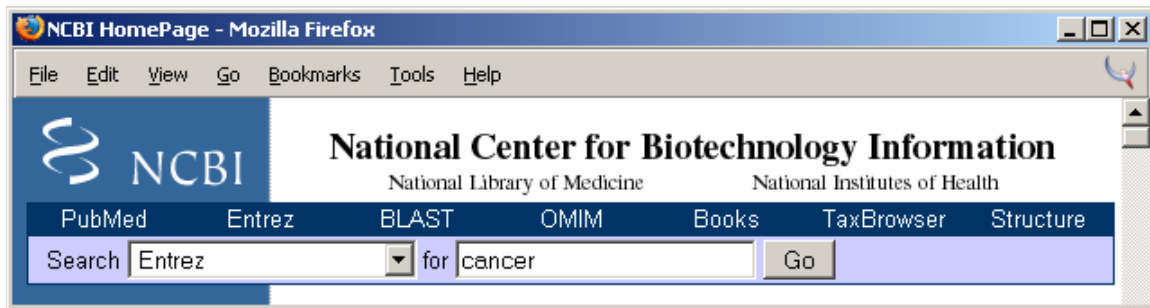


NCBI Exercises

Global Query: Controlled Vocabularies and Limits

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

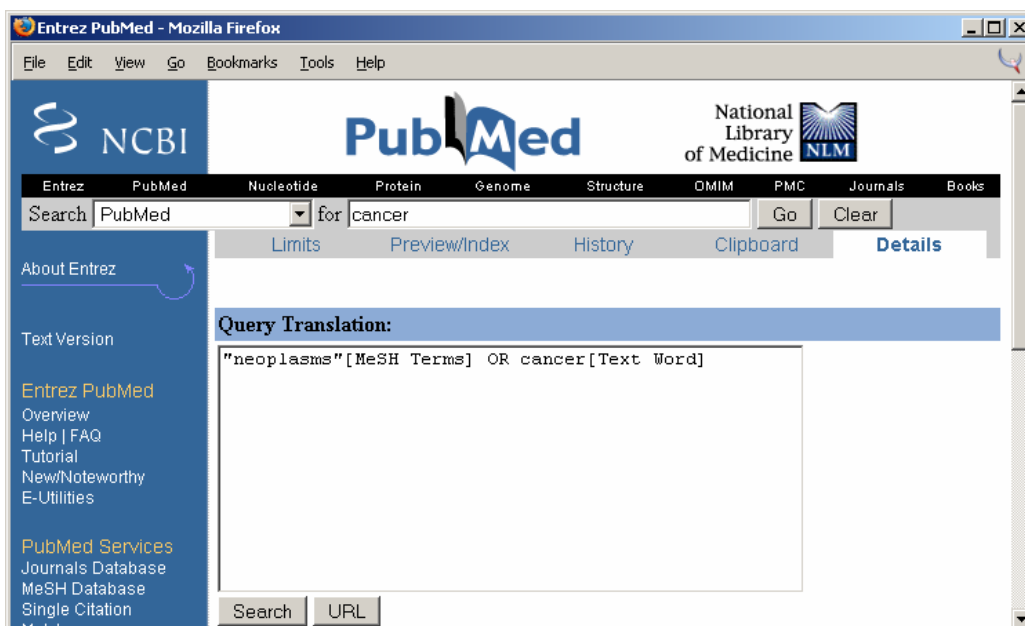


This query returns results in all of the Entrez databases. However the query is interpreted differently in different databases.

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 page. The search bar contains the text "cancer". The "Limits" tab is active. Below the search bar, there are several dropdown menus and checkboxes for filtering results. The "MeSH Terms" dropdown menu is set to "MeSH Terms". Other options include "only items with abstracts", "Publication Types", "Languages", "Subsets", "Ages", "Human or Animal", "Gender", and "Entrez Date".

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

Use the Web browser’s back button to return to the Global query page. Retrieve the results for the nucleotide database. 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.

[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 with the organism translation.

Taxonomy

[Finally, 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 for the genus *Cancer*. The search results are displayed as follows:

Lineage (full): [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](#)

◊ [Cancer](#) [140](#) [53](#) [12](#) [LinkOut](#) *Click on organism name to get more information.*

- [Cancer antennarius](#) (Pacific rock crab) [2](#) [2](#) [1](#) [LinkOut](#)
- [Cancer borealis](#) (Jonah crab) [1](#) [3](#) [2](#) [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) [121](#) [7](#) [6](#) [LinkOut](#)
- [Cancer novaezealandiae](#) [1](#) [1](#)
- [Cancer oregonensis](#) (pygmy rock crab) [1](#) [1](#) [LinkOut](#)
- [Cancer pagurus](#) (edible crab) [7](#) [34](#) [3](#) [LinkOut](#)
- [Cancer productus](#) (red rock crab) [1](#) [1](#) [1](#) [LinkOut](#)

Nucleotide: Zebrafish estrogen receptor

Zebrafish nucleotide sequences

Perform a nucleotide search to retrieve all zebrafish sequences. 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 EST sequences from your results. Check the box next to “exclude ESTs” and run the search.

Notice the large reductions in the number of results; the majority of zebrafish sequences in GenBank are EST records.

The screenshot shows the NCBI Entrez Nucleotide search interface. The search query is "zebrafish" and the "Limits" tab is selected. The "Limited to:" section is expanded, showing the following options:

- Organism: zebrafish
- exclude ESTs
- exclude STSs
- exclude GSS
- exclude TPA
- exclude working draft
- exclude patents
- exclude all of the above
- Molecule: [dropdown]
- Gene Location: [dropdown]
- Segmented Sequences: [dropdown]
- Only from: [dropdown]
- Modification Date: [dropdown]
- Modification Date: [dropdown] 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_est[PROP]”. PROP is the abbreviation for the Properties field.

The screenshot shows the NCBI Entrez Nucleotide search page. The search query is 'zebrafish'. The 'Limits' tab is active, and the 'Query Translation' section displays the following criteria: `('Danio rerio'[Organism] AND ((1900[MDAT] : 3000[MDAT]) NOT gbdiv_est[PROP]))`. The interface includes a search bar, navigation tabs (Entrez, PubMed, Nucleotide, Protein, Genome, Structure, PMC, Taxonomy, Books), and a sidebar with links to 'About Entrez', 'Entrez Nucleotide', 'Entrez Tools', 'Check sequence revision history', and 'LinkOut'.

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 nucleotide 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 “estrogen receptor” 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 “estrogen receptor” in their titles. A list containing term estrogen receptor and its expansion is now displayed with the number of records for each term.

Title Preview Index

Click to add terms selected from Index to the query box.

estrogen receptor(704)	Up
estrogen receptor 1(19)	
estrogen receptor 1 alpha(1)	
estrogen receptor 2(8)	
estrogen receptor 2 beta(2)	
estrogen receptor 2 er beta(3)	
estrogen receptor 2a(2)	
estrogen receptor 2b(1)	
estrogen receptor a(2)	
estrogen receptor alpha(63)	Down

Select “estrogen receptor” 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 estrogen receptor 2a mRNA (NM_180966). (Zebrafish have two distinct *esr2* genes; mammals apparently have only one.) 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.

Finding the genomic sequence

Click on the “Links” pop up menu in the upper right of the record.

A number of links are displayed. For mouse, rat and human Reference Sequences you could link directly to the assembled and annotated genome in the Map Viewer. However, NCBI does not yet have an assembly of the zebrafish genome. There are a 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.

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-274C14, accession BX255911, 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 eighth 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 by searching with a earlier version number, for example, BX255911.2. The revision history for this record shows all the forms it has taken in the Entrez system.

The screenshot shows the NCBI Sequence Revision History page for accession BX255911. The page is displayed in a Mozilla Firefox browser window. The title is "Sequence Revision History". The search bar contains "BX255911" and the "Show" button is selected. The "GenBank/GenPept" dropdown menu is visible. The table below shows the revision history for BX255911.

GI	Version	Update Date	Status	I	II
33695235	8	Aug 16 2003 11:11 PM	Live	<input checked="" type="radio"/>	<input type="radio"/>
31074709	7	May 28 2003 11:09 PM	Dead	<input type="radio"/>	<input checked="" type="radio"/>
31074709	7	May 24 2003 11:13 PM	Dead	<input type="radio"/>	<input type="radio"/>
30962419	6	May 20 2003 11:12 PM	Dead	<input type="radio"/>	<input type="radio"/>
30524820	5	May 10 2003 11:26 PM	Dead	<input type="radio"/>	<input type="radio"/>
30387122	4	May 5 2003 11:11 PM	Dead	<input type="radio"/>	<input type="radio"/>
30348707	3	May 2 2003 11:11 PM	Dead	<input type="radio"/>	<input type="radio"/>
28460261	2	Apr 11 2003 11:21 PM	Dead	<input type="radio"/>	<input type="radio"/>
28460261	2	Feb 21 2003 12:48 AM	Dead	<input type="radio"/>	<input type="radio"/>

Accession [BX255911](#) was first seen at NCBI on Feb 21 2003 12:48 AM

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 BX255911 is no longer a draft sequence but is in the traditional vertebrate (VRT) division of GenBank. In contrast to typical traditional GenBank records, BX255911 has almost no biological annotation.

Examine the feature table of the record and verify that the estrogen receptor gene is not annotated there.

Clearly, you could not have found this record using a text search for estrogen receptor.

Making a gene model

You can use the mRNA sequence and the genomic clone sequence to produce a gene model for the estrogen receptor. The NCBI utility Spidey will align mRNA to genomic sequence using consensus splice sites to constrain the alignment. Spidey is available at the following URL:

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

Load the Spidey page in your Web browser. Type or paste the genomic accession (BX255911) in the upper text area on the form, and type or paste the mRNA accession (NM_180966) in the lower text area. Press the “Align” button to run the program.

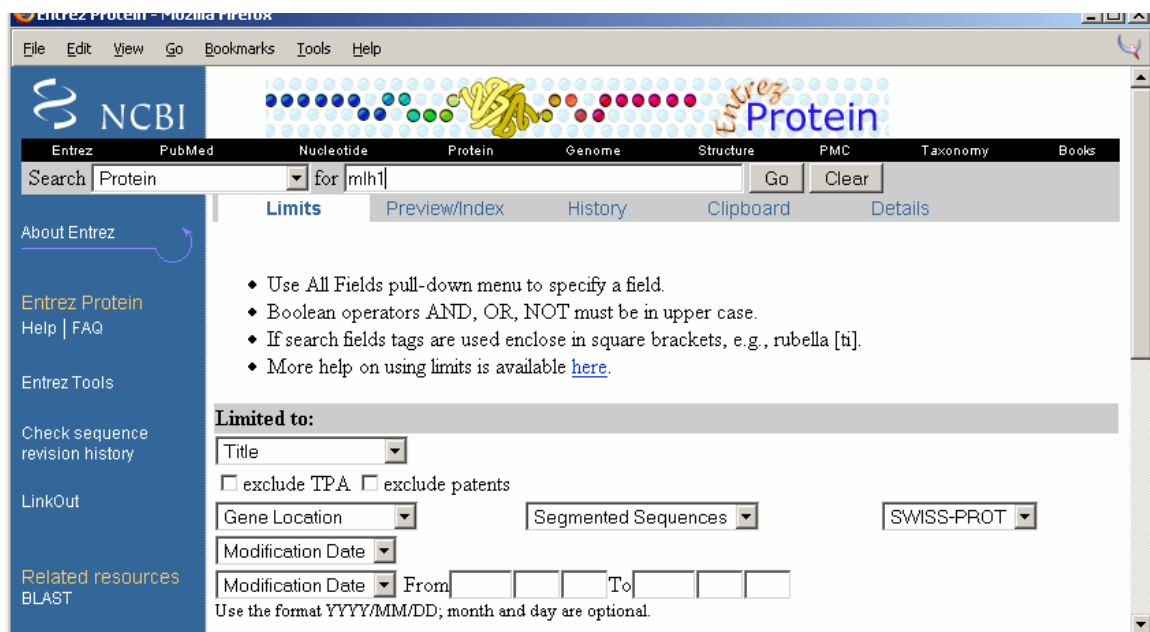
The entire *esr2b* gene is contained on this BAC clone.

Protein

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.



The screenshot shows the NCBI Entrez Protein search interface. The search bar contains "Protein" and "for mlh1". The search results are displayed under the "Limits" tab. The "Limited to:" section includes a dropdown menu for "Title", checkboxes for "exclude TPA" and "exclude patents", a dropdown for "Gene Location", a dropdown for "Segmented Sequences", and a dropdown for "SWISS-PROT". There is also a "Modification Date" dropdown and a date range selector (From [] [] [] To [] [] []). The interface includes a sidebar with "About Entrez", "Entrez Protein Help | FAQ", "Entrez Tools", "Check sequence revision history", "LinkOut", and "Related resources BLAST".

P40692 is a record imported from the Swiss-Prot database. Swiss-Prot is a small 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.

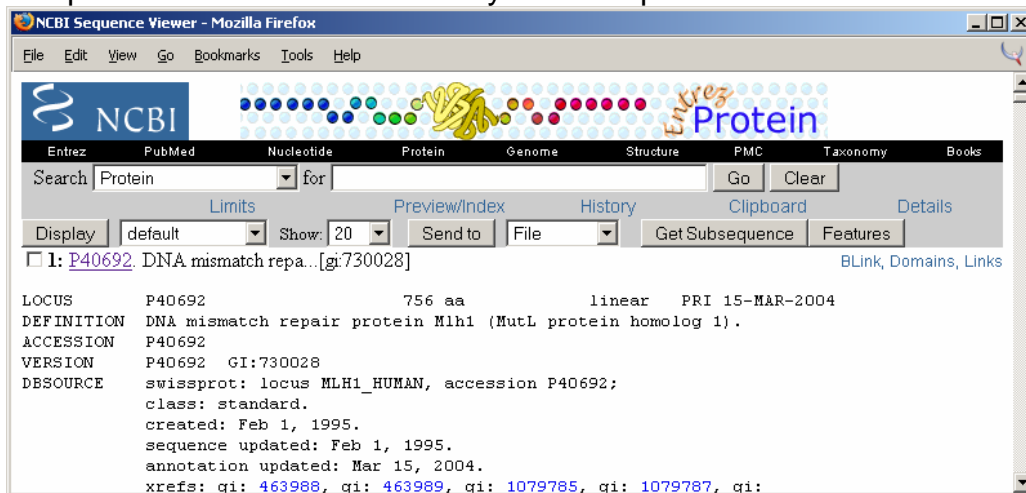
```

FEATURES             Location/Qualifiers
     source           1..756
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
     gene            1..756
                     /gene="MLH1"
                     /note="synonym: COCA2"
     Protein         1..756
                     /gene="MLH1"
                     /product="DNA mismatch repair protein Mlh1"
     Region          28
                     /gene="MLH1"
                     /region_name="Variant"
                     /note="P -> L (in HNPCC2). /FTId=VAR_004433."
     Region          32
                     /gene="MLH1"
                     /region_name="Variant"
                     /note="I -> V (in dbSNP:2020872). /FTId=VAR_014876."
     Region          35
                     /gene="MLH1"
                     /region_name="Variant"
                     /note="M -> R (in HNPCC2). /FTId=VAR_004434."
     Region          37
                     /gene="MLH1"
                     /region_name="Variant"
                     /note="E -> ELNH (in endometrial cancer; somatic).
                     /FTId=VAR_004435."
    
```

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:



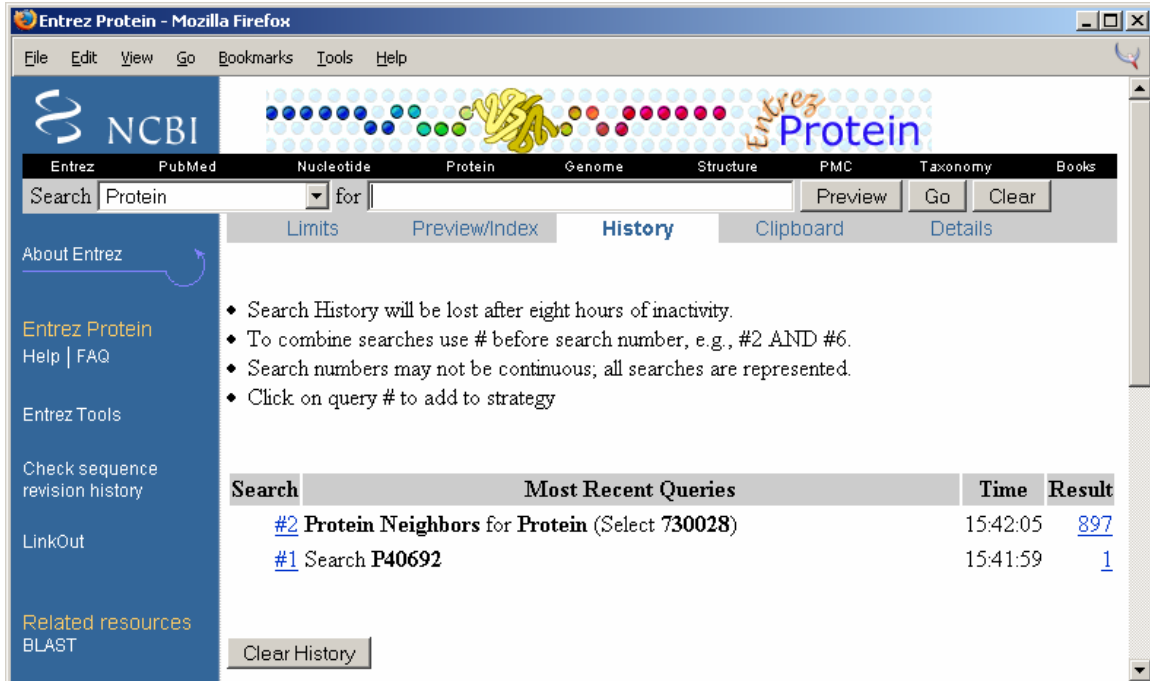
The screenshot shows the NCBI Sequence Viewer interface. The search bar contains 'Protein' and the search results for 'P40692: DNA mismatch repair protein Mlh1 (MutL protein homolog 1)'. The 'Links' pop-up menu is open, showing options like 'Gene', 'Full text in PMC', 'Related Sequences', 'Domain Relatives', 'Map Viewer', 'OMIM', 'PubMed', 'Taxonomy', and 'LinkOut'.

[Display the “Links” pop-up menu and follow the “Related Sequences” link.](#)

The resulting display is a list of similar sequences 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. Included in that set are corresponding records representing this human protein from five different sources; Swiss-Prot, PIR, 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 “Show” drop-down menu to select 500 records. Press the “Display” button to update the page. 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 can find all of the proteins from a particular taxon in the list through the “History” tab.

[Click on the "History" tab.](#)


The screenshot shows the Entrez Protein search history page. The browser window title is "Entrez Protein - Mozilla Firefox". The page has a navigation bar with tabs: Entrez, PubMed, Nucleotide, Protein, Genome, Structure, PMC, Taxonomy, and Books. The search bar contains "Protein" and "for". Below the search bar are buttons for "Limits", "Preview/Index", "History" (selected), "Clipboard", and "Details".

On the left side, there is a sidebar with "About Entrez", "Entrez Protein Help | FAQ", "Entrez Tools", "Check sequence revision history", "LinkOut", and "Related resources BLAST".

The main content area contains a list of search history entries:

- Search History will be lost after eight hours of inactivity.
- To combine searches use # before search number, e.g., #2 AND #6.
- Search numbers may not be continuous; all searches are represented.
- Click on query # to add to strategy

Search	Most Recent Queries	Time	Result
#2	Protein Neighbors for Protein (Select 730028)	15:42:05	897
#1	Search P40692	15:41:59	1

At the bottom of the history list, there is a "Clear History" button.

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](#)[#2 AND mouse\[Organism\]](#)[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 makes it much easier to find homologs in other species and 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” and press the “Display” button to refresh the page.

The screenshot shows the Entrez Protein database interface. The search results are displayed in a table format. The 'Display' dropdown menu is open, showing 'Structure Links' selected. The results list includes entries for P40692, AAC, AAA, and AAO22994, each with associated links for GenPept, GI list, Graphics, etc.

Display	Structure Links	Show: 20	Send to: Text
□ 1: P40692	GenPept GI list Graphics TinySeq XML LinkOut Protein Neighbors Domain Links 3D Domain Links Gene Links	g7	BLink, Domains, Links
□ 2: AAC	Genome Links DN LinkOut HomoloGene Links Nucleotide Links	g46	BLink, Domains, Links
□ 3: AAA	OMIM Links PMC Links PopSet Links PubMed Links Mut SNP Links Structure Links Taxonomy Links UniGene Links	g13	BLink, Domains, Links
□ 4: AAO22994			BLink, Domains, Links

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.

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 “View 3D Structure” button.

In order to display the structure you will need to have the NCBI structure viewer, Cn3D installed. Follow the hyperlink labeled “Get Cn3D” and follow the instructions to install the viewer.

This 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.

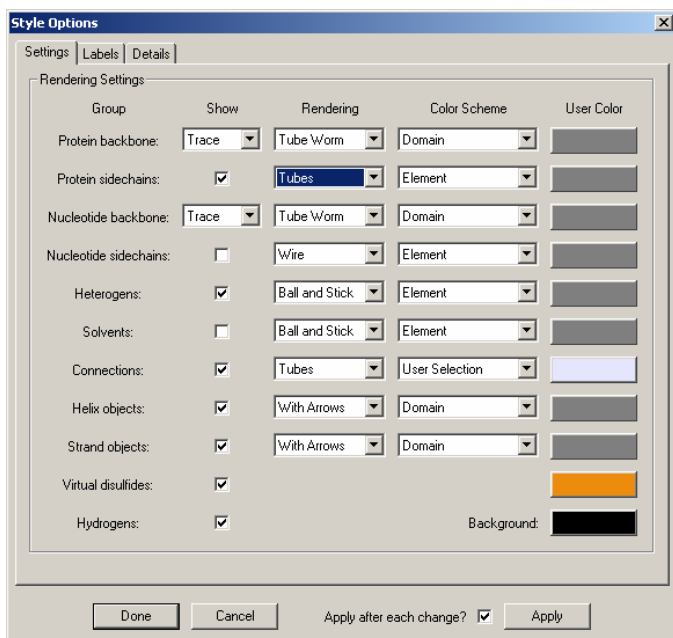
[Use the “Coloring shortcuts” on the “Style” menu to color by Domain.](#)

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 the magnesium ion. One 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 checkbox on the corresponding 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 protein set with the biologically uninteresting patent sequences removed. At the top of the page, is a list of the gi numbers 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 of differences in the sequences from different sources for the same protein. You can easily identify the best protein match from the plant *Arabidopsis thaliana*.

[Click on the BLAST score on the line containing the best *Arabidopsis* protein.](#)

The new window shows the BLAST 2 Sequences alignment between the human MLH1 and the best match in *Arabidopsis thaliana*. 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 “3D Structures” button.

[Click on the “3D Structures” button.](#)

This displays the equivalent of a protein BLAST search against PDB. The protein from 1B63 that you looked at earlier is one of the aligned proteins. As before, the linked BLAST score will show the alignment in BLAST 2 Sequences. The blue dot next to the BLAST score will load the alignment into Cn3D.

[Click on the blue dot on the line with 1B63 and then click the view 3D structure button on the “Related Structures” page.](#)

The 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) a 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 the record from the nucleotide database by typing accession BG618460 in the search box on the NCBI homepage and display the record.

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.

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. NCBI also has linked tag counts from quantitative SAGE libraries to the UniGene clusters.

Follow the "Gene to Tag" mapping link to see a "virtual Northern" display of the counts of reliable tags from this cluster in SAGE libraries.

What library 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. You may want to remove the cytogenetic map from the display before continuing. Do this by clicking the "X" at the top of the map. Uncheck the "Compress Maps" option on the left-hand-side to see the full marker labels. 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?

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 a point-to-point synteny of the three genomes. Removing the UniGene map may make this easier to view. Notice that the structure of the albumin gene family is conserved in these three mammalian genomes. In the

current mouse build (32), the structure surrounding afm, alb and afp is distorted because of a low quality assembly in that region resulting from the inclusion of early phase BAC clone sequences.

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.

- 1. Follow the link from the BLAST home page (<http://www.ncbi.nlm.nih.gov/BLAST/>) to the rat genome BLAST page.**
- 2. Type the accession number for the human albumin precursor, NM_000477, into the search box on the BLAST form.**
- 3. 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.

- 4. Format your results.**

Were you able to find the rat homolog?

- 5. Repeat the search. This time uncheck the “Use Megablast” option.**

You should have found some hits this time. The graphical overview shows that some 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.

- 6. 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

7. **Go back to the rat genome BLAST page and the human RefSeq protein accession, NP_000468, in the search box.**
8. **Select the translating BLAST search, tblastn, from the program selection drop down menu.**

9. **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.

1. **Follow the link from the BLAST home page (<http://www.ncbi.nlm.nih.gov/BLAST/>) to the chicken genome BLAST page.**
2. **Type the accession number for the human albumin precursor, P02768, into the search box on the BLAST form.**
3. **Select the translating BLAST search, tblastn, from the program selection drop down menu.**
4. **Leave the database menu set on “genome” and click the BLAST button.**

The translating searches take much longer than the standard BLAST searches.

5. **Format your results.**

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

Using NCBI BLAST

Identifying sequences

Michael Crichton's fantasy about cloning dinosaurs, *Jurassic Park*, contains a putative dinosaur DNA sequence. Use nucleotide-nucleotide BLAST against the default 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.

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 Advanced Options

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 default nucleotide database.

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 in human genome BLAST to get a clearer view of the product predicted from genomic DNA in the Map Viewer. You'll need to uncheck the “Use megablast” box to get this to work.

Now try these modified primers in standard nucleotide-nucleotide BLAST. There is one mismatch in each near the middle. You will need to use the advanced options to increase the expect value to 100 to compensate for the mismatches.

```
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.

Use the advanced options to adjust the Word Size from 11 to 7 and run the search again.

Do you find the original hits again? Are they still among the best hits? Can you devise a modification in the search strategy that will make them the best hits again?

There are special BLAST forms available from the BLAST homepage that has the expect value and word size preset to find matches using short sequences such as these primers. In the next exercise, we will use the protein-protein “Search for short nearly exact matches” page to count occurrences of a short peptide in the database.

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.

1. **Follow the link from the BLAST homepage to the protein “Search for short, nearly exact, matches” BLAST form.**

Examine the advanced options on the form and notice that the expect value and word size are preset to find short matches. In addition low complexity filtering is off and the scoring matrix is changed from BLOSUM62 to PAM30. PAM30 promotes the significance of nearly identical alignments.

2. **Type ELVIS in the search box.**
3. **Adjust the number of descriptions in the formatting options to 1000 to include all Elvises.**
4. **Run the search.**

What is the expect value for an exact match to ELVIS? How many Elvises are there in the database?

The number of Elvises increases in a linear fashion with the size of the database in accordance with the random behavior of protein sequences.

Protein-protein BLAST and advanced options

The *Caenorhabditis elegans* gene SMA-4 is a member of the dwarfins gene family, also called the MAD (mothers against decapentaplegic) family. SMA/MAD gene products play a role in transforming growth factor beta-mediated signal transduction. In this example we will attempt to find homologs for the SMA-4 protein (SMA4_CAEEL, Accession P45897) in vertebrate species.

Of course, this protein already is in the Entrez protein and BLAST databases. Remember that if the goal is to find a homolog in another species for a protein that is already present in the Entrez system, it is not necessary to perform a BLAST search; the pre-calculated similarities are already available through the related sequences link or through the BLink link.

Find homologs for SMA-4 in chicken by following the BLink link from P45897 in Entrez Protein. Click the best hits button and find the best protein hit to chicken (*Gallus gallus*).

The alignment between SMA-4 and the best chicken match is available by clicking on the linked BLAST score.

Return to the NCBI homepage and link to the protein-protein blast page and enter the SMA-4 accession number (P45897) in the Search text area.

We will search the default BLAST protein database, called “nr” on the database drop down menu.

To simulate performing a BLAST search with a novel protein, we will use an Entrez query to remove all *Caenorhabditis* proteins from the BLAST database. In fact, we can modify the default database so that it contains only vertebrate sequences. The ability to modify the BLAST database using an Entrez query is one of the “Options for advanced blasting.” You can type any valid Entrez query in the box labeled “Limit by entrez query.” Only those queries that give clearly defined sets of proteins would be useful here though. As we have seen previously, the organism query in Entrez uses the taxonomy controlled vocabulary for molecular databases. Notice that we have a drop down list that allows you to select from a limited list of pre-formulated organism searches. However, you can formulate queries for any taxon that NCBI recognizes. You can also formulate complex Boolean queries. For this search, we will use a complex query to select vertebrate proteins and remove the human proteins

Enter the following Entrez search in the "Limit by Entrez query" box under the "Options" section of the form:

vertebrates[Organism] NOT human[Organism]

Because there are a large number of related proteins in the BLAST database, we also need to increase the number of descriptions or BLAST hits that will be shown.

Use the drop down list to increase the number of descriptions to 500 in the "Format" section of the BLAST form. Run the search by clicking the BLAST button.

On the formatting page, you can see that the CD-search has identified conserved domains in this protein. You can click on the graphic to see what these domains are and what their function is.

Click the format button to retrieve your BLAST results. Look at your BLAST graphical output and verify that the Entrez query eliminated the query protein from the database; you should see no full-length matches.

Follow the link to the “Taxonomy reports” to verify that there are only vertebrate proteins in your output without any human sequences.

You can follow the link to the number of hits to chickens in the Taxonomy reports to see the results for only the chicken proteins in the output.

Close the “Taxonomy reports” window and look at the list of sequences found by BLAST. The most significant protein match is the SMAD8 protein from mouse.

Scroll down in the list of sequences in the descriptions. In the non-significant e-values (> 1), there are two proteins from the horse (*Equus caballus*) labeled as MAD proteins (Smad8 and Smad7). These protein fragments are homologs of SMA-4, but we did not demonstrate that with this particular search because these fragments -- 89 aa and 77 aa -- don't overlap well with the conserved portions of the *C. elegans* protein. (It is easy to find these as significant matches to the more closely related mouse SMAD8.) We can also use PSI-BLAST to show that these horse proteins are significant matches to the *C. elegans* SMA-4. Be sure to retain your formatting page for these results or copy your request ID so you can format them for PSI-BLAST for the next exercise.

Open a new browser window so you don't lose your results against the nr and run the search again. This time restrict the search to chicken proteins using the Entrez query option:

Chicken[Organism]

Are the same proteins found? Compare the expectation values of these hits to the same hits found against nr with no organism restriction. Why are the e-values different for the same scores and alignments?

PSI-BLAST

Any protein-protein BLAST search on the NCBI web pages can be extended to a PSI-BLAST search simply by re-formatting the results.

Check the "Format for PSI-BLAST" box on the formatting page for the SMA-4 protein search that you saved from the exercise above and click format.

The results are the same except that they are formatted differently. There is a line across the descriptions section of the results corresponding to the PSI-BLAST inclusion threshold of 0.005. Position-specific information from a multiple sequence alignment of the sequences above this line are used to generate a position-specific score matrix (PSSM) in the next iteration. Notice that Smad8

and Smad7 from the horse are below this line. Make a note the e-values of these two hits.

Now click the "Run PSI-BLAST iteration 2" button on the results page.

The Formatting page is refreshed in its separate window, generating a new Request ID number.

Click the "Format" button and the results of iteration 2 will load. Click on the "Skip to the first new sequence" link on the Iteration 2 results page. What is this sequence? What is its new expect value? Notice that there are now several new sequences above threshold. Check the e-values of the horse Smad8 and Smad7 proteins and compare them to the e-values from the standard protein-protein BLAST search. These new sequences will be used to construct a new PSSM for iteration 3 and so on. After a few more iterations no more sequences will be found; at this point the search is said to have converged.

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 their may have been balancing selection for heterozygotes at this position during human evolution. The EST data for human represent 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:

- 1. From the BLAST homepage, link to the “Protein query vs. translated database.”**
- 2. Type the prion protein accession number, P04156, in the search text area.**
- 3. Use the select subsequence boxes to use only residues 100 to 160.**
- 4. Choose the “est_human” database.**

Since we’re interested in looking at changes at a particular residue, it is more useful to have the BLAST output as a stacked master-slave pairwise alignment (query-anchored).

- 5. Use the “Format” options section of the BLAST form to set the “Alignment view” to “query-anchored with identities.”**
- 6. Set the number of alignments and descriptions to 500.**
- 7. Finally click the BLAST button to run the search.**
- 8. Click the format button on the formatting page to retrieve your results.**

Look at the alignments to see the 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?