

---

This is a guided tour. The homework is separate. In fact, this exercise is used for multiple classes.

The tour will take a few hours. Do it with a friend (or two or three)!

The NCBI site can be slow if it's very busy; this is hard to predict.

If you are using this exercise for a class, be sure to start well before the homework is due!

The details of how the NCBI site is configured are constantly changing. If specific steps in this guide don't work as expected, keep going, or try to recreate the result using the site as revised (e.g. if a button is not where I say it should be, look around).

The following are exercises to be carried out using a web browser (e.g. Firefox, Netscape, Safari or Microsoft Internet Explorer).

## GUIDED TOUR

---

### NCBI DATABASES

Go to <http://www.ncbi.nlm.nih.gov>. This is the National Center for Biotechnology web site. (To get to this URL: Depending on your browser and operating system you can open a new location with control-L, control-O, command-L or command-O, and then type in the address. You can also just enter the address on the top line, also known as the address bar).

You can also get to this address, and other useful bioinformatics addresses, from my bioinformatics links page, <http://www.clfs.umd.edu/labs/mount/Bioinformatics.html>.

Once at the NCBI site, notice the top line ("quicklinks bar"), which lists PubMed, All Databases, Blast, OMIM, Books, TaxBrowser and Structure. We will consider most of these in turn.

First, **click on All Databases**. This page presents all of the NCBI databases on a single page and allows you to search across databases. **Enter "hemoglobin" into the search window and click on "Go."** You should see numbers next to icons corresponding to each of the databases. At the very top are Pubmed, PubMed Central, Books, OMIM and site search. By clicking on one of these links you can see information about hemoglobin in PubMed, in books, on OMIM, etc.. More specialized databases are listed below. The following exercises will introduce you to these databases.

**PubMed** (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed>). This is a publicly available database of literature related to medicine, which includes all of molecular biology and genetics. To see how it works, navigate to that site using the top toolbar (**click on PubMed**) and **type in Dreyfuss Xing Matunis**

You should get a single article in Nucleic Acids Research on hnRNP F.

Note the box next to "Display" (near the top) with "Abstract Plus" written in it. This is a pull-down menu. If you put your mouse over the one labeled " Abstract Plus " and hold the mouse button down, you will see that there are other choices: Summary, Brief, Abstract, Citation, etc.. Most of these are different formats for presenting the citation. I encourage you to try them out. Because Nucleic Acids Research participates in PubMed Central, the full text of this article can be accessed directly by clicking on the link above the citation (on the right side). If the journal were not freely available to the public, it would still be possible to click a link here and connect to the article. In general, if the University of Maryland subscribes you should be able to connect to the journal from any campus computer.

From off-campus, you would need to access the journal article through Research Port (<http://researchport.umd.edu/>) and log in with your ID (use the 14 digit number on the back).

Try "**Links**" (upper right, next to the journal links) **and then select "CoreNucleotide (RefSeq)".** On the next page click the number next to "core nucleotide." You should now see:

The screenshot shows the NCBI Entrez Nucleotide search interface. The search term 'CoreNucleotide' is entered in the search box. The results are displayed in a table with 5 entries, each showing a GenBank accession number (NM\_001098206, NM\_001098204, NM\_001098205, NM\_001098207, NM\_004966) and a description of the Homo sapiens heterogeneous nuclear ribonucleoprotein F (HNRPF) transcript variant. The interface includes navigation tabs (Limits, Preview/Index, History, Clipboard, Details), a 'Display' menu set to 'Summary', and a 'Show' dropdown set to '20'. A sidebar on the left contains various navigation links like 'About Entrez', 'Entrez Nucleotide', and 'Entrez Tools'.

You can get to the GenBank entries for these sequences directly by clicking on the accession numbers. You can also get to these same entries via the Nucleotide database. Let's see how that works.

On the pull-down menu next to "Search," you will see the selections PubMed, Protein, Nucleotide, CoreNucleotide, Structure, etc..

**Select "CoreNucleotide," and then type in NM\_004966 and click "Go."** This takes you to the same entry you got through MedLine. Now, **click on NM\_004966,** and you will see the GenBank entry for the locus HUMHFP, which encodes the hnRNP F protein. ([ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucore&id=148470398](http://ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucore&id=148470398)) Please note how the sequence is provided with a great deal of associated information.

Note the pull-down menu labeled "Display" with "GenBank" written selected. Using your mouse, you can see that there are other choices: GenBank(Full), FASTA, ASN.1, Summary, etc.. **Select FASTA** (on the "Display" pull-down menu). You will see the sequence alone in a standard format known as "fasta format."

---

This is the fasta format for the nucleotide sequence. Protein and nucleotide sequences have distinct accessions, but both types of sequence can be represented in fasta format. To get to the associated protein sequence, look for the "**links**" **button** on the right top of the main part of the page. If you **click on it**, you will see a list: Gene, Genome Projects, etc..

**Click on 'Protein.'** You will see:

[NP\\_004957](#)

heterogeneous nuclear ribonucleoprotein F [Homo sapiens]  
gi|4826760|ref|NP\_004957.1|[4826760]

This is the refseq accession for the protein sequence.

Refseq accessions are standard versions of accessions for which there are multiple related entries.

Refseq accession can be identified by their standard formats.

See <http://www.ncbi.nlm.nih.gov/RefSeq/key.html> - [accessions](#) for more information

Follow the link for this protein (**click on NP\_004957**). You should now be at <http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=protein&val=4826760>

This time, **click on BLink**. You should see a list of proteins that are related to human hnRNP F by sequence similarity. In fact, this is the output of a pre-computed BLAST search. There are several very useful ways to sort this output. You may want to explore them now or come back to this later (after you do some BLAST searches of your own).

If you **go back** to the page for NP\_004957 and **click on "Links" again**, you will see that there was also a link to OMIM, the Online Mendelian Inheritance in Man. **Click on "OMIM."** This page provides links to OMIM pages that provide much of the same information as the protein accession, but list the position on the human genome and any associations with disease.

**Click on the link to 601037** (HNRPF).

You should now be at <http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=601037>

You will see a page with a short text description of the gene and a list of citations. If the gene were associated with human genetic disease that would be described here. **Click on 10q11.21-q11.22** to see the gene map region in OMIM. These pages are in OMIM. **Go back to the listing for NP\_004957 (click the back button your browser three times or just start over by entering NP\_004957 under a search for protein)**. This time, **click on Links and choose Map Viewer**. You will see a much more complex presentation of the same genetic region.

**Click on ug (next to HNRNPF, you will need to scroll right and down)**, and you will be taken to UniGene. HnRNP F has UniGene ID 808. **Click the Hs.808** to reach the Unigene page

<http://www.ncbi.nlm.nih.gov/UniGene/clust.cgi?ORG=Hs&CID=808>.

This page presents homologs in various species provides additional links based on homology. Please go back to the Map Viewer page and explore this on your own. Try zooming in and out and following various link.

## Ensembl

Another site with similar presentations of gene maps is Ensembl, at EMBL. Go to <http://www.ensembl.org/index.html> and use either NP\_004957 or HNRPF to get to

[http://www.ensembl.org/Homo\\_sapiens/geneview?gene=ENSG00000169813](http://www.ensembl.org/Homo_sapiens/geneview?gene=ENSG00000169813)

Explore this site on your own.

---

## UCSC

Another site with a focus on comparative information is the UCSC browser. Go to <http://genome.ucsc.edu/>

Click on Genomes, and then enter either NP\_004957 or HNRPF to get to

[http://genome.ucsc.edu/cgi-bin/hgTracks?position=chr10:43201070-43224620&hgSid=69090608&refGene=pack&hgFind.matches=NM\\_004966](http://genome.ucsc.edu/cgi-bin/hgTracks?position=chr10:43201070-43224620&hgSid=69090608&refGene=pack&hgFind.matches=NM_004966).

This browser attaches a variety of interesting features to the sequence and the user has a lot of choice about what to see. For example, you can reverse the polarity (which is useful if the gene goes right to left). The width of your view can be set to any value, including values greater than the width of your screen (in which case you will need to scroll right and left). The image and track options are selected under "configure."

I encourage you to explore it on your own.

---

## BLAST

Now, let's try to find a sequence by similarity. Go back to <http://www.ncbi.nlm.nih.gov>.

This time, **click on BLAST**. This takes you to <http://www.ncbi.nlm.nih.gov/BLAST/> and allows you to search for similar sequences using BLAST. You will see a number of selections in categories:

Nucleotide BLAST, Protein BLAST, blastx, tblastn and tblastx

There are many options. These are explained in help pages that you can access by clicking on the word HELP in the top toolbar.

The most commonly used options are blastp, blastn, tblastn and blastx.

**blastp** compares an amino acid query sequence against a protein sequence database.

**blastn** compares a nucleotide query sequence against a nucleotide sequence database.

**tblastn** compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames

**blastx** compares a nucleotide query sequence translated in all reading frames against a protein sequence database

For now, you want to see what is related to hnRNP F. **Select "protein BLAST."** Type the protein accession number, **NP\_004957, into the box**. When you click the blue oval labeled **BLAST** the search will begin. You may see a new page with some information including any conserved domains (what you see depends on how busy the site is). In this case, three RRM domains should have been found, along with Zf-RNPHF). There are links here to information about these domains. This page also presents the "Request ID" that you can use to obtain the results later (it will work for a day or so). **Save the request ID** (copy it to the clipboard with Control-C and paste it into a text or Word document or use later).

The request-ID should be something like F3NPHXDY014.

At some point, your search will finish and you will see alignments between the hnRNP F protein and other protein sequences in GenBank. If the search takes a long time, you can open a new window in your browser (this is usually in the upper left pull-down menu; control-N or command-N) and do other things while the search continues. You can also retrieve the results later using the "request ID" number that appears in this window. The time that the search takes has nothing to do with how fast your network connection is! The search takes place at NCBI, and the results are transmitted back to you as a fairly small file.

**Examine your results.**

You will first see a figure illustrating where each "hit" matches your query sequence. Below that, you will see something like this:

[Distance tree of results](#) **NEW**

Sequences producing significant alignments:	Score (Bits)	E Value	
<a href="#">ref NP_004957.1 </a> heterogeneous nuclear ribonucleoprotein F [H...	<u>825</u>	0.0	<b>G</b>
<a href="#">gb AAH16736.1 </a> HNRPF protein [Homo sapiens]	<u>824</u>	0.0	<b>G</b>
<a href="#">sp Q60HC3 HNRPF MACFA</a> Heterogeneous nuclear ribonucleoprotein...	<u>822</u>	0.0	
<a href="#">dbj BAC36361.1 </a> unnamed protein product [Mus musculus]	<u>816</u>	0.0	<b>G</b>
<a href="#">ref NP_598595.1 </a> heterogeneous nuclear ribonucleoprotein F [M...	<u>813</u>	0.0	<b>G</b>
<a href="#">ref XP_001490098.1 </a> PREDICTED: similar to heterogeneous nucle...	<u>805</u>	0.0	<b>G</b>
<a href="#">ref XP_534954.1 </a> PREDICTED: similar to Heterogeneous nuclear ...	<u>803</u>	0.0	<b>G</b>
<a href="#">ref NP_001014860.1 </a> heterogeneous nuclear ribonucleoprotein F...	<u>800</u>	0.0	<b>G</b>
<a href="#">gb EDL39909.1 </a> mCG129396 [Mus musculus]	<u>797</u>	0.0	<b>G</b>
<a href="#">gb AAH89313.1 </a> Hnrpf protein [Mus musculus]	<u>768</u>	0.0	<b>G</b>
<a href="#">ref XP_856932.1 </a> PREDICTED: similar to Heterogeneous nuclear ...	<u>727</u>	0.0	<b>G</b>
<a href="#">ref XP_001155889.1 </a> PREDICTED: heterogeneous nuclear ribonucl...	<u>687</u>	0.0	<b>G</b>
<a href="#">gb AAH29764.1 </a> Hnrpf protein [Mus musculus]	<u>663</u>	0.0	<b>G</b>
<a href="#">gb AAH27003.1 </a> Hnrpf protein [Mus musculus]	<u>661</u>	0.0	<b>G</b>
<a href="#">gb EDL02340.1 </a> mCG50680 [Mus musculus]	<u>660</u>	0.0	<b>G</b>
<a href="#">gb AAI29088.1 </a> Hnrphl protein [Rattus norvegicus]	<u>594</u>	4e-168	<b>G</b>
<a href="#">ref XP_856916.1 </a> PREDICTED: similar to Heterogeneous nuclear ...	<u>594</u>	4e-168	<b>G</b>
<a href="#">ref XP_856955.1 </a> PREDICTED: similar to Heterogeneous nuclear ...	<u>592</u>	1e-167	<b>G</b>
<a href="#">ref XP_001099754.1 </a> PREDICTED: heterogeneous nuclear ribonucl...	<u>590</u>	4e-167	<b>G</b>
<a href="#">ref XP_856589.1 </a> PREDICTED: similar to Heterogeneous nuclear ...	<u>590</u>	6e-167	<b>G</b>
<a href="#">ref XP_001490122.1 </a> PREDICTED: similar to heterogeneous nucle...	<u>590</u>	7e-167	<b>G</b>
<a href="#">ref XP_848777.1 </a> PREDICTED: similar to Heterogeneous nuclear ...	<u>584</u>	3e-165	<b>G</b>
<a href="#">ref NP_067485.1 </a> heterogeneous nuclear ribonucleoprotein H1 [...	<u>584</u>	4e-165	<b>G</b>
<a href="#">ref NP_005511.1 </a> heterogeneous nuclear ribonucleoprotein H1 [...	<u>584</u>	4e-165	<b>G</b>
<a href="#">emb CAG33059.1 </a> HNRPH1 [Homo sapiens]	<u>582</u>	2e-164	<b>G</b>


You can see the alignment for each "hit" by clicking on the score. If you **click on the bit score of the accession refNP\_0067485.1** you will see an alignment of the query (NP\_004957, human hnRNP F) with the hit (NP\_067485, mouse hnRNP H). "Query" refers to the sequence you searched with, while "Sbjct" (subject) refers to the matching database entry. The top scoring "hit" is a perfect match of the query to itself. The G on the right of this list takes you to the "[Entrez Gene](#)" page for this protein (Gene ID 3185 is human hnRNP F).

**Go back to the BLAST results.** A new feature that you might want to examine is a distance tree of results (just above the list of hits). Give it a try!

**Look at the alignments!!!!**

**Go back to the BLAST results** (if necessary, use your Request ID) and return to the alignments. If you **cruise through the list of hits**, you will see that most of the top scoring hits are similar proteins in more closely related species. In order to evaluate the meaning of a match, you should pay close attention to both the percent identity and E value.

For example, find [gi|25153681|ref|NP\\_741422.1|](#) and go to the alignment (by clicking on the score). You will see:

```
>[ref|XP_001100049.1|  PREDICTED: heterogeneous nuclear ribonucleoprotein H1 isoform
8 [Macaca mulatta]
Length=472

Score = 580 bits (1494), Expect = 7e-164, Method: Composition-based stats.
Identities = 309/413 (74%), Positives = 348/413 (84%), Gaps = 3/413 (0%)

Query 1      MMLGPEGGEGFVVKLRGLPWSCSVEDVQNFLSDCTIHDGAAGVHFIYTREGRQSGEAFVE 60
            MMLG EGGEFVVK+RGLPWSCS ++VQ F SDC I +GA G+ FIYTREGR SGEAFVE
Sbjct 1      MMLGTEGEGEFVVKVRGLPWSCSADEVQRFFSDCKIQNGAQGIRFIYTREGRPSGEAFVE 60
```

The E value of  $7e-164$  means that you would expect  $7 \times 10^{-164}$  hits this good or better if you searched a database of random sequences the size of the database that you actually searched. Better matches yield lower E values. Generally, E values less than 0.05 have **statistical** significance, but the **biological** significance may be meaningless unless E values are below  $1e-6$  or even lower. Matches with scores this good are not rare when related proteins are compared.

Many of the top hits in this search are extremely similar. You may want to find less similar proteins, in another species, for example. There are two ways to do this. First, **go to the top of the blast output window and click on "Taxonomy Reports."** This will generate a searchable window that presents the hits organized by the taxonomy of the species involved.

As an alternative, you can limit your search in the first place. To do this, go back to the main BLAST window, and repeat your search. This time, you want to restrict your search to the fruit fly *Drosophila melanogaster*. **Repeat the search after selecting *Drosophila melanogaster* under "Organism"** (Just start to type "Drosophila" and you will be given choices.) You may notice that the search takes less time this time around! That is because you are searching a smaller database. Look at your output, including the alignment, and you will see a set of related proteins. How many do you think are likely to have a similar function?

Many of the hits obtained with this search are proteins containing the RRM domain. You can see the location of RRM domains in hnRNP F by returning to the blast page and looking at the Conserved Domains hits, located just above the Request ID. Alternatively, return to the main NCBI page, <http://www.ncbi.nlm.nih.gov/> and **search Proteins using the accession NP\_004957**. Once you have the protein listing, you can click on "Conserved Domains," which is between BLink and Links on the upper right. You should see three RRM domains aligned with the hnRNP F sequence. If you **click on any one of these**, you will see an alignment of several specific RRM proteins with the query hnRNP F. You may want to explore links from this page. Note particularly the Pfam page for this domain (pfam00076).

In this exercise, I have taken you through just a few of the available programs in great detail. **Play around.** Explore the site yourself. You can also get help online by clicking on links within the BLAST pages. There are also help and tutorial links on the NCBI page themselves.

There are other sites that this tutorial did not cover, like Gene Ontology ([www.geneontology.org/](http://www.geneontology.org/)), that you may also want to look at; many of these are listed on my [links page](#).

Good luck and have fun!