

Biology 362. Principles of Genetics

NCBI: GenBank Tutorial

The goal of this lab is to provide a broad overview of the utility of the National Center for Biotechnology Information Website (NCBI). It was established in 1988 as a national resource for Molecular Biology information. NCBI creates databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information – all for the better understanding of molecular processes affecting human health and disease.

1. Exploring specific Databases on NCBI

The NCBI website offers many sources of information. The information provided is in a series of databases. Entrez allows you to search the linked databases available. The databases include:

PubMed: The biomedical literature (PubMed)

Nucleotide: DNA sequence database (Genbank)

Protein: Amino acid sequence database

Structure: Three-dimensional macromolecular structures

Genome: Complete genome assemblies

PopSet: Population study data sets

Taxonomy: Organisms in GenBank

OMIM: Online Mendelian Inheritance in Man, human database of genes and their related disorders.

Popset Database

Goto the Popset Tutorial at:

<http://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowSection&rid=coffeebrk.chapter.3> – On the left panel select the link under boxes and work through the tutorial DNA Barcoding to answer the questions below.

- a. Which gene is being used by scientists and with which objectives?
- b. What is the Popset Database?
- c. How many DNA barcoding sets were available as of April 2005?
- d. Why in the study Identification of birds through DNA barcodes says “this study does not contain any alignments”?
- e. In the Results page, where are the accession numbers display?
- f. How can you obtain interconnected data and where is the location of the link?

Search the SNPs database for HFE

<http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=coffeebrk.box.301>

- a. What is a single nucleotide polymorphisms (SNP) and which is its importance?
- b. Do SNPs in the HFE gene have functional consequences?
- c. Define hemochromatosis and explain who causes it.
- d. What is the function of HFE?
- e. Is this hereditary genetic disorder recessive or dominant?
- f. What does EV show us? What do the thick and thin bars indicate?
- g. How many SNP were found in one of the HFE transcript variants?
- h. Which is the importance of the cys260 mutation?
- i. What is the OMIM database (Online Mendelian Inheritance in Man) and what does its records list?

2. BLAST: Basic Local Alignment Search Tool

BLAST finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequences databases and calculates the statistical significance of matches. Can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Most of the time you will want to search the general nucleotide, protein, or genome databases. However, you can also search for more specific databases such as blastn for nucleotide-nucleotide searches, blatsx, tblastn, or tblastx for translated queries or blastp for protein queries.

NCBI explains how to select the appropriate BLAST program. It depends on the given search and is influenced by three factors: the nature of the query, the purpose of the search, and the database intended as the target of the search and its availability. [These NCBI page](#) provides recommendations on how to make a selection of the database to use.

Now, go to

http://www.DigitalWorldBiology.com/dwb/Tutorials/Entries/2009/1/26_BLAST_for_Beginners.html and click the green arrow to find the answer to the following questions:

- a. What is the gi number?

- b. How can you determine the entry point of a sequence into the database?
- c. What is the E value?
- d. What is the Score (bits) value?
- e. What does “CDS” stands for?

3. Using MapView – see where genes are on a chromosome

At the top of the page (blue menu bar) click on MapViewer. If you want to find hemoglobin in humans, you can either:

- click on “Homo sapiens” under “Vertebrates” and then type in hemoglobin,
- OR use the pull down menu at the top (i.e. Search Homo sapiens for hemoglobin)
- Be sure to click on the organism name, not BLAST which will take you elsewhere.
- In MapViewer you should see a representation of all the chromosomes. Genes with the name hemoglobin in the name are marked in red.
- Find hemoglobin, gamma G under the search results. Click on Genes_cyto under Maps.
- At the top of the page, the chromosome you are looking at is highlighted. The region displayed should be a code like 11p15.4. The 11 tells you which chromosome you are on. The p tells you the gene is on the short arm of the chromosome; regions on the long arm are labeled q. For a whole-chromosome view, look at the Ideogram on the lower left.
- You will notice that some loci have multiple genes. Multiple genes located in the same place are called gene clusters.
- Next to the gene symbol there are a bunch of links (under LinkOut). These are links to other databases in NCBI.

4. Using Gene – close-up of a gene

There are a few different ways of getting to the Gene view of hemoglobin gamma 2:

- from the page you were just on (genes_cyto), click the gene Symbol (HBG2)
- start from the Cross-Database Search page and click on Gene
- If you are lost now, search Gene for HBG2 and look for the entry with GeneID 3048

You should be looking at pages and pages of gene information. You can change the amount of information by the Display menu. For right now, set it to Full Report.

Under “Genomic regions, transcripts, and products” you’ll see a diagram of the gene with its introns and exons.

One important section is the General Gene Information area. The entries under Gene Ontology tell you the biological processes this gene is involved in.

5. Using Nucleotide – looking up gene sequences

From Gene you can look at a gene's DNA sequence, RNA sequence, or protein sequence. Under NCBI Reference Sequences, click on

- Reference NG_000007
- mRNA Sequence NM_000184
- or Product NP-000175

Notice that genomic DNA has a code starting with NG, mRNA with NM, and proteins with NP

For right now, go to the mRNA sequence in Nucleotide, some important things to note are:

Accession Number: this is the unique number by which this specific gene is referred (especially in literature)

Length: if you're trying to PCR this gene, it's useful to know how long your product is supposed to be (look next to Locus, there's a number like 538bp)

References: are you writing a paper about this gene? Clicking on the PubMed references takes you to specific papers.

Exercise

1. How many chromosomes does a tomato have?
2. Find the gene sonic hedgehog in chicken, and determine on which chromosome is this gene located.
3. Is this gene located on the long or short arm?
4. How many coding regions does this gene have?
5. What is the mRNA code for this gene?
6. Say you are interested in the human disease lamellar ichthyosis. How many free, full-text articles can you find in PubMed?
7. Say this disorder is caused by a defect in gene ABCA12 in humans. What are some biological processes this gene is involved in (list 2)?
8. What is ABCA12's mRNA accession number?

6. NCBI Protein Structure

We will get familiar with NCBI's Molecular Modeling Database (MMDB) and learn how to download a protein structure file, visualize, and annotate it. The MMDB database is a subset of three-dimensional structures from the Protein Data Bank (PDB). Most structures are obtained from X-ray crystallography and NMR spectroscopy. To visualize and work with these structures you will use the freely available CnD3 4.1 software. Cn3D operates on Windows, Mac, and Unix. It displays structure, sequence, and alignment, and has annotation and alignment editing features.

1) Download and install CnD3 4.1 [here](#).

2) Instructions for using CnD3 and its features are provided here:

<http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3dtut.shtml>

Before proceeding, carefully read through the sections titled:

[Retrieving individual structures \(MMDB\)](#)

[Viewing individual structures in Cn3D](#)

[Annotating a structure](#)

[Saving structures and images](#)

You also can use CnD3 to explore alignments of two or more protein structures, and CnD3 has several other advanced features that you may choose to explore:

[Retrieving structure alignments \(VAST\)](#)

[Viewing structure alignments in Cn3D](#)

[Alignment editing](#)

Now it is time to get started. Go to MMDB and find a particular protein structure you would like to analyze and learn more about. Type the name of your protein at the top of the page:

Search Entrez: Structure and include any taxonomic names you wish to use to restrict your search. Examples:

Hemoglobin anseriformes – 4 hemoglobin structures

DNA polymerase human – 196 human DNA polymerase structures

Myoglobin Physeter catodon – 174 sperm whale myoglobin structures

Hemocyanin – 13 hemocyanin structures

Click on the MMDB link for each structure to display the journal reference, description, and map of the molecular components of the structure.

Click on the “PDB” link to access the Protein Data Bank link to this structure (much more information is available here).

Click on the ‘VAST’ link to view the database of structural neighbors.

To download, the structure select: ‘All Atom Model’, ‘CnD3’, and ‘Save File’ and then click ‘View 3D Structure’.

Now you can open the structure file in CnD3 and start to work with it. **The rest is up to you.**

Next week, each one of you will give a 5 min oral presentation and tell the class what you learned about your structure using CnD3. Be prepared to tell the class about your protein, how it functions, and why you chose it.