

## Introduction to Genomic Databases

Starting link: <http://www.ncbi.nlm.nih.gov>

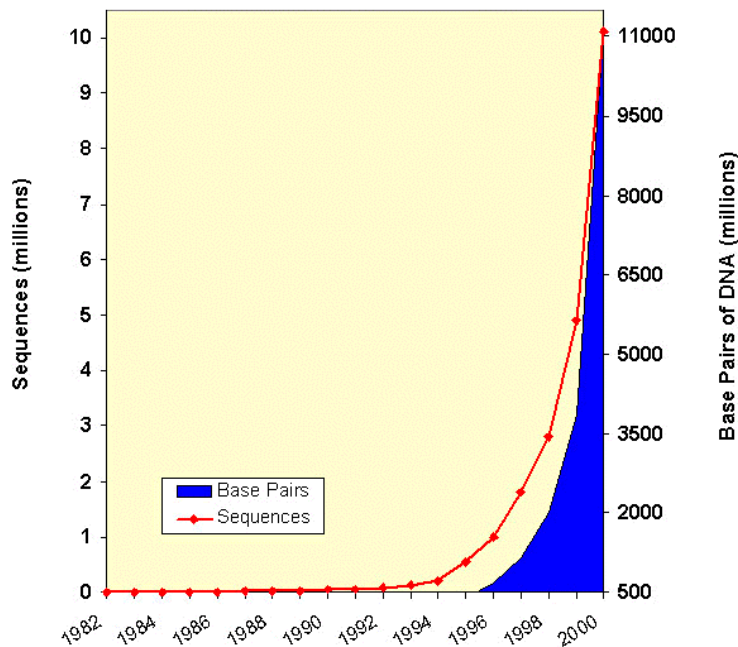
Everyone has a difficult time keeping up with the flow of new information. This is particularly true in biology now as the pace of discovery accelerates. Databases have become an essential tool for accumulating and archiving raw data. They also play a major role in analyzing and presenting information to researchers and the public in an easily accessible form. In this *Exploring Genomes* tutorial we will survey one of these resources: The National Center for Biotechnology Information (NCBI) located in Washington, D.C.

---

One of the roles of NCBI is to archive raw DNA sequence data. The sequence information comes from research efforts in laboratories around the world as well as from large-scale, dedicated genome sequencing centers. The resulting database is referred to as GenBank. GenBank shares its resources with the European and Japanese equivalents so that there are three primary public repositories of such information in the world. Because of the automation of DNA sequencing over the past decade, these databases are increasing in size exponentially. GenBank includes the sequences of the *E. coli*, *Drosophila*, and human genome, as well as data from thousands of other species. At present it comprises some 10 million DNA sequences with a cumulative length of 11 billion base pairs. You can see the rate of growth by clicking [here](#).

[<http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html>]

### Growth of GenBank



Now let's go back to the NCBI home page, the public access point to its many resources (<http://www.ncbi.nlm.nih.gov/>). A quick glance at this page shows that NCBI contains far more than just the sequence repository. It is a rich source of information on all aspects of genetics and genomics. All of the divisions are searchable and information ranging from gene sequences, to the position of a locus on a human chromosome, to direct access to the scientific literature dealing with a particular gene, is immediately accessible.

---

There are six major categories of service listed across the top of the NCBI homepage. These are PubMed, Entrez, BLAST, OMIM, Taxonomy, and Structure. In addition there are specialized projects and databases listed on the right hand side of the page. In this *Exploring Genomes* tutorial we will take a quick look at some of these resources. Subsequent tutorials will explore a particular resource in much greater depth. Let's click on each of the six resources at the top in turn. Click on '**PubMed**' first.

---

PubMed is the NCBI gateway to the biomedical research literature. It is a searchable database and information can be retrieved based on combinations of parameters such as author, subject key words, or organism. A complex query can be entered and a list of publications matching it will be returned. For instance, we could enter a simple search by author. If we entered **Hartwell LH** (one of the winners of the 2001 Nobel Prize in

Medicine) and pressed **'Go'**, PubMed would return a list of his current publications. Give it a try.

---

The list of Hartwell's publications is 5 pages long. Only the top and therefore most recent papers are displayed on the first page. Each paper is linked to its abstract and sometimes the full text of the articles. We will explore PubMed much more fully in a later tutorial. Let's go back and try another NCBI division by clicking the **NCBI icon** on the top left to get to the NCBI homepage again, and then clicking the button for **'Entrez'**.

---

The Entrez button opens a search engine that links all of the NCBI databases together. This allows access to everything from sequence to structure. The options are in the drop-down window at the left. The PubMed link that we looked at initially was only one of these options. Let's try searching for the human keratin protein sequence in the Protein database. Choose **'Protein'** from the drop-down menu, and then type **'keratin AND human'** in the text box. Now press **'Go'**.

---

The search returns a list of database entries including keratin-associated proteins and keratins themselves. Note that only the first 20 entries of a total of 464 are displayed on the first page! For each entry, clicking on the associated links will display the sequence, related information, and links to other parts of the databases. We can explore these later. Let's go back to the **NCBI** homepage and try the **'BLAST'** button.

---

BLAST is a powerful nucleic acid or protein alignment tool. It allows us to dynamically search the sequence databases (all 11 billion base pairs!) to find similar sequences in different organisms. It is extremely versatile and comes in many different forms for doing different types of searches. The underlying method is the same in each case, however. This is our standard software tool for doing such searches. It is very important and we will devote an entire tutorial to its use later. For now, let's move back to the **NCBI** homepage and click on to **'OMIM'**.

---

OMIM is the Online Mendelian Inheritance in Man database. Note from the overview on the OMIM homepage below that it integrates the known Mendelian genetics of human disease with the resources made available through Entrez at NCBI. We will explore this in detail later but for now let's open the **OMIM Statistics** link, under OMIM Facts on the left-hand menu.

---

The OMIM statistics page gives an overview of the breadth of the resource. Note the categorization by Mendelian inheritance pattern. Note also that over 13,000 entries are available with at least some information. Keep in mind that we believe there are more

than 35,000 human genes! All genes will not necessarily be associated with a disease or visible phenotype and therefore for many genes little information is yet available. The Human Genome Project has provided a glimpse at large numbers of new genes of unknown function, reminding us of how much remains to be done.

Next, click back to the **NCBI** homepage and on to the **Taxonomy** database.

---

The Taxonomy section groups all data by taxonomic classification. You may type in a species name to find out if any sequence information is available; for the major genetic systems, simply click on the links on the Taxonomy Browser home page. Try *Caenorhabditis elegans*, a nematode worm, one of our most powerful model genetic systems whose full genome sequence was recently determined.

---

This organism-specific page gives important information regarding phylogenetic lineage as well as the number of sequences of various types that have been deposited. These groupings may be called up at will and each of the genes listed are linked to the Entrez system. One important use of the groupings is to restrict other types of searches. For instance a BLAST search can be launched from the BLAST page and the database searched restricted to *Caenorhabditis elegans* only (or any other species).

Now, let's go to the last major subdivision, '**Structure**'.

---

The structure database contains the 3D structure for all nucleic acids and proteins whose shape has been determined by X-ray crystallography or nuclear magnetic resonance. We can call up these 3D models at will. The structure database is associated with the VAST program that allows for 3D structural comparisons among different proteins. It also will search the database on the basis of structure. These are very powerful tools and we will give them a try later in the term.

Now, let's go back to **NCBI** home page to see some of its other resources.

---

Apart from the basic databases and access software, NCBI has a wide variety of highly specialized databases and analyses. These focus on particular problems or interest groups. Some are listed on the right hand side of the NCBI homepage below. We will just touch on a couple of them for now to get a sense of their capabilities.

The first is the Human map viewer. This allows us to visualize the various human chromosomes and the genetic loci on them. Let's take a look by clicking on '**Human map viewer**' under the Hot Spots list.

---

The human genome view is a visualization of the full chromosome set. Clicking on any chromosome number will expand the view of that particular chromosome. Let's try the **Y** chromosome.

---

The graphic of the Y chromosome is expanded so that we can see its banding pattern as seen by a cytologist. Aligned along it are the blocks of DNA that have been sequenced and represented here (contigs) followed by the genes located on these pieces of DNA. The genes are represented in the first column as the compiled summary data in Unigene (another of the NCBI databases). Subsequent columns provide links to the various genes that have been annotated on the DNA sequence. Let's look more closely at **Hs 56336**, near the top of the Unigene list below.

---

The Unigene summary is an annotated view. It is compiled by a curator and takes into account all known information regarding the sequence. Unigene also includes links to OMIM as well as a comparison to the most similar genes in other organisms. It is a very rich source of information. All of the data is cross-referenced by links into Entrez and various other databases that provide the gene sequence and research literature references. Ultimately the entire human genome will be available in finished form from this database.

Now let's go back to the **NCBI** homepage and click on the first resource under Hot Spots, '**Cancer genome anatomy project**'.

---

The Cancer Genome Anatomy Project is the last topic we'll briefly explore. This database focuses only on tumor tissue and strives to provide cross-referenced information for all genes thought to be involved in cancer. Each of the subsections enters the database with a different type of approach. Ultimately you can explore the database to great depth, searching out the genes involved, chromosomal locations and aberrations, and biochemical pathways. All of this information is related to the tissues and cells where the tumor that you are interested in originates.

---

Over the course of these *Exploring Genomes* tutorials, we will look at parts of these databases in much greater detail. A facility for handling them is an essential skill for a modern biologist. Ultimately, the only way to familiarize yourself with a resource of this type is to go to the web site (<http://www.ncbi.nlm.nih.gov/>) and start exploring some of the links. You might start by searching for the answer to the following simple questions:

- What is the publication history of your Biology and in particular your Genetics instructors?
- What organisms do they work with?
- What types of question are they trying to answer?