Ch1: General introduction

The National Center for Biotechnology Information (NCBI): Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information.
DNA sequences are stored in three major banks: GenBank (USA)
To get more info check http://www.ncbi.nlm.nih.gov/About/index.html

Exponential increase in DNA sequences:
http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

Explore:

Pubmed: literature searches-try Hartwell, L
Protein: Search "Keratin AND Human"-- get about 1000 entries
Blast:
OMIM (Online Mendelian Inheritance in Man)
Open statistic link under OMIM Facts on the left
Taxonomy: Click on Arabidopsis
Structure: 3D structure database for all nucleic acids and proteins whose shape has been determined by X-ray crystallography or nuclear magnetic resonance. 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.
Mapviewer (under hot spot): click on Homo sapiens
Click on Y chromasome-click on a unigene

Ch 2: How to use Entrez (All Databases)

Entrez' strength is that it provides links between related types of information. For instance, in storing a DNA sequence file, the file is associated with the protein translation of the sequence, with the literature reference, and with links to similar genes or proteins in other organisms. It also provides a characterization of any notable features, such as conserved regions, and ultimately chromosomal location and 3D structure of the gene product. The retrieval system moves relatively effortlessly among these various types of data. The links are updated as new data are added to the databases (or new databases are developed) and the resource becomes richer and richer over time.

Go to www.ncbi.nlm.nih.gov
Click on "All Databases" on top black bar
Search Dystrophin (underlying Duchenne muscular dystrophy) in Pubmed
Type "Dystrophin" in the search box and hit go
4393 entries-in reverse chronological order
Click on some of the entries to explore-get abstract
Let's call up a paper by "Hoffman, Brown and Kunkel (1987)"
Click "Limit" on light blue bar above display button
Enter "jan 1, 1987 to Dec 31, 1987" press GO
Now you see the "Hoffman Brown and Kunkel"
Click on the authors to get abstract, related articles, etc.
Click Link (upper right) drag down menu to click OMIM
Click on Xp21.2 -lead to a table

Click Xp21.2 again to get the human map (genomic map)
Click DMD to get to LocusLink (all the info about DMD)

Go back to 'All Databases"

Click Protein and then search for "Dystrophin" in the search box
Get 3359 entries
Click on P11532 (Accession number)
Click on the Blink on the upper right hand
The BLink program displays a graphic of an alignment of our gene with all related sequences in the database.
Click on any entry The display you see is the standard output format for GenBank. It has a specific list of sections, each with a particular type of information. It also includes highlighted links to other aspects of the databases.
Check different sections for information
What is the protein sequence?
Get the "FASTA" format by clicking the "DISPAY drag down menu" in upper left
Select FASTA from the dragdown menu

Restart Search by going back to "All Databases"
Click "Protein" and then search for
"Dystrophin and chicken" in the search box--get 83 entries
"Dystrophin not human" in the search box--get about 2081 entries

Ch 3: How to use Blast (Basic Local Alignment Search Tool)
Blast: the most important single software tool for searching sequence databases.
Query: The sequence used to initiate searches

Go to www.ncbi.nlm.nih.gov
Click BLAST
Click on 'Protein-protein BLAST [blastp]' under the Protein BLAST heading
1. Use following INSULIN sequence from zebra fish to search
Copy following sequence in FASTA format; paste the sequence into the search box
>gi|12053668|emb|CAC20109.1|insulin[Danio rerio]
MAVWIQAGALLVLLVVSSVSTNPGTPQHLCGSHLVDALYLVCGPTGFFYNP
KRDVEPLLGFLPPKSAQETEVADFAFKDHAELIRKRGIVEQCCHKPCSIFELQNYCN
Alternatively, you can also type the accession number in the window
Set subsequence allows you to search with a particular portion of the sequence. Leave it blank so that the entire sequence will be used in the search.
Choose database has a drop-down menu: choose nr for non-redundant-it includes one copy of each gene or protein in several of the main databases and excludes multiple copies of each record
Do CD-Search allows a comparison of the query sequence to a database of conserved domain patterns. This is a powerful tool for finding functional domains in genes. Leave it toggled on
Hit the submit buttom
On the top, you will see conserved domain (red bar), click on that to see CD

Click "format" to get the Blast result
You will see a detailed list of hits ordered by their alignment scores. They correspond to the ones displayed graphically. Note that each line gives the identification information for the protein followed by the alignment score and the E value. The entries are ranked from the lowest to the highest E value, which can be interpreted as from most similar to more distant. The top line, not surprisingly, is the record for Zebra fish insulin itself. If you click on the gene identifier link, it will call up the sequence from Entrez. When you click on the Score link, it will show you the particular alignment from lower down in the output file.
XXXX is a low complexity, or repetitive, region that is masked out in the query and ignored in the database search. Such regions may interfere with the alignment. You can see the actual masked sequence, since it is the same protein, in the subject line (LLVLLVVSSVS); it is a mostly hydrophobic, repetitive sequence.
Alignment score (S): indicating how strong the match was (higher is better).
E value: a statistical measure of the significance of the match; expectation that the match would have been found in the database by chance alone (lower is better).

Click on score 91.3: see imperfect matches
+ means similar but not identical
How do you calculate % identity vs % similarity
There are gaps (deletions and insertions) between alignment
Note the query vs subject
Note that the full length of the protein is not shown. BLAST is a local alignment tool and only displays the most strongly matching regions of the overall comparison.

2. Do another Blastp search using Dystrophin as the query
Click protein on top bar
Enter P11532