Database lecture notes

Learning outcomes

Where data comes from (experiments)
How it is stored and distributed (records, data structures, relational databases)
How it is tracked (identifiers)
How it is updated (data pipelines)
How it is retrieved (web interface, FTP, web-services)
Major molecular databases (NCBI and UniProt)
Where and how to find other databases (http://bioinformatics.ca/links_directory/index.php)
Distinguishing between good and bad databases.

Finding your gene of interest.

Find a record for your gene of interest.
Start here: http://www.ncbi.nlm.nih.gov/gene
Search for a gene name. Rpb1
Note ambiguity. Why?
Limit to a taxon (yeast).
Are the results still ambiguous? Why?
Note that the official gene name. (Rpb1)
What are Limits?
Point out link to help: http://www.ncbi.nlm.nih.gov/books/NBK3837/#EntrezHelp.Using_Limits
Point out advanced search: http://www.ncbi.nlm.nih.gov/gene/advanced
Go to Rpo21 record:http://www.ncbi.nlm.nih.gov/gene/851415
Redo the search using the gene id: 851415[uid].
Point out major fields. Gene ID, Gene Symbol
Go through display settings. The same record can be displayed multiple ways.

A **record** is an assembly of one or more pieces of data in an ordered manner.
Individual pieces of data fill “**fields**” of the record.
Fields have names and expected formats.
The set of fields and there expected formats are called a **data structure**.
Records can be displayed in variety of **formats**.
A record will have a **non-ambiguous, distinct identifier**.

Excercise:

1. Use Entrez Gene to find your gene of interest.
2. Give your neighbour the gene name (alias or official name) and see if they can find it.
3. Give your neighbour the Gene ID (when and if they get tired).
4. Try to compose a search that returns your gene record and only your gene record without using the Gene ID.

Secondary databases and primary databases.

Entrez Gene is a database of Gene Records.
The Gene concept has many associated attributes (like coding DNA and mRNA and protein products).
The Gene record collects many of these attributes in one place and links out to other databases

where additional information can be found that supports these attributes with experimental evidence.
Entrez Gene could be called a secondary database because it collects information from multiple sources and assembles them into a related concept.

No one has ever observed a “gene”. It’s a concept.
However, someone has observed the DNA sequence (contained within this gene) that encodes a protein that has been called Rpo21. Records in a primary database will contain references to primary experimental observations. Many other observations have been made that are collected under the Rpo21 Gene concept.

Excercise:

1. Look through the different sections of the Entrez Gene record. What kinds of information are available? Can you tell where the information is coming from?
2. Can you find a paper that describes the sequencing of your gene of interest?
3. What are GeneRIFs and where do they come from?
4. What is the difference between EntrezGene and SGD?

http://www.ncbi.nlm.nih.gov/gene/851415
http://www.yeastgenome.org/cgi-bin/locus.fpl?sgdid=S000002299
SGD is a model organism database. They annotate and curate many of the records in RefSeq and Entrez Gene.

5. What is the difference between RefSeq and GenBank?

http://www.ncbi.nlm.nih.gov/RefSeq/
http://www.ncbi.nlm.nih.gov/genbank/
http://www.ncbi.nlm.nih.gov/genbank/submit.html
GenBank is a **primary submission database** that accepts sequence submissions from researchers.
RefSeq is a **curated primary database** that is a subset of GenBank that may combine information from multiple records of the same sequence submitted by various researchers into one curated record.

6. What is the difference between NCBI and GenBank

NCBI is an organization. GenBank is just one of their databases/resources.
http://www.ncbi.nlm.nih.gov/
You can search across all their resources using the Entrez interface.
http://www.ncbi.nlm.nih.gov/gquery

7. Discuss data structures: http://www.ncbi.nlm.nih.gov/IEB/ToolBox/MainPage/index.html

Looking at identifiers in a protein record

Follow the link under protein product:
http://www.ncbi.nlm.nih.gov/protein/NP_010141.1
Point out: date, accession, version and GI and db_xref. 1733 AA.
Date: time the record was created or last updated (annotation only)
Accession: common to all versions of the sequence record.
Version and GI are specific to a version of the record.
Each time the sequence is updated a new record is made with the same Accession but a different version and GI.
Compare to this record that has multiple versions:
http://www.ncbi.nlm.nih.gov/protein/NP_012376.3
Try

http://www.ncbi.nlm.nih.gov/protein/NP_012376.2
You can find out what has changed by looking at the Revision History.
Select “Revision history” under “Display Settings”.
http://www.ncbi.nlm.nih.gov/protein/NP_012376.3?report=girevhist
Note that a Gene ID has been added to the latest version of the record.

Go back to the Entrez Gene page for Rpo21.
Look at the link to UniProt P04050. External link.
Look at the the links under related sequences:
http://www.ncbi.nlm.nih.gov/protein/2507347
This record is borrowed (copied) from UniProtKB but NCBI assigns their own GI.
Both sequences are identical. But this may not always be the case. There can be asynchrony between databases (when UniProtKB updates their record, there may be a lag-time until NCBI updates their version of the same record.

Then look at the original record in UniProtKB
http://www.uniprot.org/uniprot/P04050

Look at the bottom of the page under Entry information.
The entry name (RPB1_YEAST) is a combination of the gene name and the common organism name.
The accession (P04050) is common to all versions of the record.
Sometimes records are deprecated and used to make a new record. For example two records can be merged into one. In this case, the accessions of the two old records will be added to the new record as secondary accessions.
Note how many times the P04050 record has been modified. How many times the sequence has changed. See link to “Complete History”.

Excercise

1. UniProtKB (like RefSeq) is a curated database. Related records will almost always have reciprocal links to one another. Check your favourite protein for links between protein records in RefSeq and UniProtKB and also for reciprocal links between Entrez Gene and UniProtKB.
2. Use either history tool in RefSeq or UniProtKB to look at how the record for your favourite protein has chnaged over time.
3. A record may have one or more cross-references to other records in the same or in a different database.
Cross-references may refer to the identical entity or concept (but in another database).
Cross references also may refer to related information - like Gene Ontology annotation.
4. On your own: Google “curation policies refseq” and “curation policies uniprot”. Curation policies may or may not be public and they may not always be easy to find. Start with the database site itself and look for something like a manual, help or FAQ.
http://www.ncbi.nlm.nih.gov/RefSeq/RSfaq.html
http://www.uniprot.org/help/biocuration
http://www.uniprot.org/help/uniprotkb
http://www.uniprot.org/manual/entry_name

Retrieving data

Databases may often make a snap-shot of all their data and make it available via download via FTP (usually as tab-delimited files - but also in other formats like XML or FASTA). These snap-shots will usually have a release number and an associated date when the release was made.
In some cases, you can use your browser to retrieve files
try ftp://ftp.uniprot.org/pub/databases/uniprot/relnotes.txt or

ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.dat.gz
or
you can use an FTP application (from the command line or a graphical interface to FTP that you download and install).
Large data sets will often be compressed as .gz files or .zip files.
Windows users can use the application 7-zip to uncompress files (http://www.7-zip.org/)

Downloaded files (especially in tab-delimited format) can be quite useful for further processing (using Perl, R or a local instance of the database in MySQL for example).

UniProtKB FTP site: http://www.uniprot.org/downloads
NCBI FTP site: http://www.ncbi.nlm.nih.gov/Ftp/

Excercise

Use the Firefox browser to go here:
and then click on the Gene FTP link to go here
ftp://ftp.ncbi.nlm.nih.gov/gene/
Browse through the README here
ftp://ftp.ncbi.nlm.nih.gov/gene/README

Click on the DATA directory.
There is another README in this directory.
Try downloading the file “gene_info.gz”.
Try uncompressing it and viewing it with a text editor.
Find the description of this file in the README and read it.

Excercise

Demo
On the command line, try:
ftp ftp.uniprot.org
enter “anonymous” as your user name
enter “anonymous” as the password
use these commands to navigate the directory structure and retrieve data

dir	returns a listing of the files and subdirectories in the current directory
cd directory name	move to the “directory name”
cd ..	move up one level in the directory hierarchy

get filename	retrieves the file called “filename”
quit	quits the ftp application

Read more about FTP by googling “man FTP”.

Excercise

You can also retrieve data from your search results at some databases and customize them as a tab-delimited file.

Try a search at www.uniprot.org
Click on “Results Customize” (Upper left-hand corner).
Select the fields you want from each record.
Click on the orange download button (Upper right-hand corner).
Select the format you want (Tab delimited say) and click on the corresponding “Download” link.

Where and how to find other databases

Time permitting, we will talk about this site:
http://bioinformatics.ca/links_directory/index.php

This is an excellent starting point for finding bioinformatics databases, tools and other resources.

Working with cross-reference tools
Being able to convert between different types of identifiers is another important bioinformatics skill. Quite often you will have one identifier type but need another because the tool, website or database you want to use only understands a limited number of accession types.
There are a few tools that can help
We will cover working with two cross-reference tools in a separate excercise.
1. The UniProt ID mapping service. http://www.uniprot.org
2. BioMart. http://www.biomart.org

Summary - What we have learned.

All primary data has an origin (some experiment).
Data about these experiments are organized in records.
Data about concepts are also organized into records.
Records have distinct identifiers.
Identifiers are made up of two parts (a database name - usually implicit and an accession or identifier that can either be alphanumeric or an integer).
All data has a location (some one who takes care of it).
Databases can share information (make their own local copies of something).
Databases can also link to one another.
Cross-references can either be to related information or to identical concepts in another database.
Data is updated and modified constantly (usually by curators but also by original submitters).
Databases will usually have a method to track history of a record.
Databases will usually have curation policies explaining how data structures are used.
Databases will commonly have periodic releases of a snap-shot of their data that can be downloaded from an FTP site.
Search results can also be downloaded via a web-interface in some cases.
NCBI and EBI are two major bioinformatics centres.
There are many other centre and databases that may serve more specialized needs.
Not all databases will be maintained or updated according to the same standards.

Suggestions
You should become familiar with at least the NCBI and UniprotKB sites.
Familiarize yourself with major identifier types and how they are related.
Know where and how to download, uncompress and view data in a tab-delimited file.
Learn to use a database id conversion utility.