NCBI Databases

Nociones Básicas de Bioinformática y Genómica

(Máster en Bioinformática, Universidad de Valencia)

David Montaner

2015-02-16

NCBI

- **National Center for Biotechnology Information** (NCBI)
  http://www.ncbi.nlm.nih.gov/
- part of the United States National Library of Medicine (NLM)
  https://www.nlm.nih.gov/
- a division of the National Institutes of Health (NIH)
  http://www.nih.gov/

Resources

Databases of biomedical and genomic information for *all* organisms:
(and related tools)

- Submission
  - GenBank
- Databases
  - GenBank
  - **RefSeq**
  - ...
- Downloads
  - FTP sites
  - PubMed
- Tools
  - Basic Local Alignment Search Tool (BLAST)
  - ...

See: http://www.ncbi.nlm.nih.gov/guide/all/

Documents

NCBI Help Manual:

- quick overview about topics
- usually just FAQs
- online (HTML) book

NCBI Handbook:

- nice introduction to each tool or database
- online (HTML) book
- but Chapters may be downloaded in PDF format
- see a chapter example describing GenBank

Many other:

- Glossary
- NCBI Educational Resources

Entrez: Search and Retrieval System

- the indexing and retrieval system used at the NCBI
- used for all major NCBI databases:
  - PubMed
  - Nucleotide and Protein Sequences
  - Protein Structures
  - Complete Genomes,
  - Taxonomy
  - OMIM
  - ...
- **text-based** searches over several record fields
- In practical terms, the web interface

NCBI Databases

Main

- GenBank: collection of **all** publicly available DNA sequences
- RefSeq: **non-redundant** set of reference standards

Other

- Sequence: Gene, Genomes, Protein
- Variation: dbSNP, dbVar
- Experimental Data: GEO, SRA, BioProject, BioSample
- Health: OMIM, ClinVar,dbGaP,
- Literature: PubMed, PubMed Central, Bookshelf, MeSH,
- Species: Taxonomy, HomoloGene

See Entrez Databases at EntrezHelp

GenBank

- collection of publicly available *annotated* nucleotide sequences and their **protein** translations
  - mRNA sequences with coding regions
  - segments of genomic DNA with single or multiple genes
  - ribosomal RNA gene clusters
  - genome shotgun reads
  - isolated genes
  - complete genomes
  - ...
- primary sequence data; **not curated**; minor checks done by the NCBI
- just authors submit and revise
- may have multiple records for same loci
- records can contradict each other
- no limit to species included

INSDC: International Nucleotide Sequence Database Collaboration

INSDC members:

- **GenBank**
- ENA: European Nucleotide Archive
- DDBJ: DNA Data Bank of Japan

http://www.insdc.org/

GenBank Access

GenBank Home Page (not very informative)
https://www.ncbi.nlm.nih.gov/genbank/

Primarily access via the NCBI **Nucleotide** database which is divided into three divisions:

- CoreNucleotide: the main collection (same as Nucleotide)
- EST: short single-read transcript sequences (Expressed Sequence Tags)
- GSS: unannotated short single-read primarily genomic sequences

But some other ways are available:

- BLAST: align against GenBank sequences
- FPT: ftp://ftp.ncbi.nlm.nih.gov/genbank/

GenBank ID System

Two *parallel* id nomenclatures for DNA, RNA and protein:

GI: GenInfo Identifier

- introduced in GenBank Release 81.0 (February, 1994)
- integer number
- new GI is assigned to every sequence version
- internal database keys

ACCESSION

- Character and numbers
- Accession.Version. Eg: AB000349.2, AB000349.3
- The accession portion of these identifiers is stable and will not change
- the version portion is incremented whenever the underlying sequence changes.

GenBank Record Format

See an Example of GenBank Record

RefSeq: The Reference Sequence database

http://www.ncbi.nlm.nih.gov/refseq/

- a **curated** collection of DNA, RNA, and protein sequences
- created by the NCBI from existing data (GeneBank)
- unique example of each natural biological molecule (for each major organisms)
- not all organisms available
- for each model organism, RefSeq aims to provide separate and linked records for:
  - the genomic DNA
  - the gene transcripts
  - and the proteins arising from those transcripts

RefSeq

- non-redundant set of reference standards (**NR**)
- includes:
  - chromosomes
  - complete genomic molecules (organelle genomes, viruses, plasmids)
  - intermediate assembled genomic contigs
  - curated genomic regions, mRNAs, RNAs
  - proteins
  - alternatively spliced transcripts
- generated to provide reference standards for multiple purposes
- facilitates database inquiries based on:
  - genomic location
  - sequence
  - text annotation

RefSeq Access

- Entrez: `http://www.ncbi.nlm.nih.gov/refseq/`
- NCBI Gene: include nomenclature, maps, pathways . . .
- NCBI Genome: information on genomes including sequences, maps, chromosomes, assemblies, and annotations
- NCBI Assembly: Genome assembly
- NCBI UniGene: A Unified View of the Transcriptome

Example Homo sapiens (human) genome.

RefSeq Accession Format

Accession format: accession number that begins with two characters followed by an underscore.
There are several RefSeq accession prefixes

- NM_: mRNA
- NR_: RNA (non coding)
- NC_: Complete genomic molecule, usually a reference assembly.

Curation **VERSION** is indicated after a dot:

- NM_000014.4
- NM_000014.5

Usual fasta id for a sequence:

```
>gi|262118207|ref|NM_000202.5| Homo sapiens iduronate ...
```

RefSeq Curation Levels

- There are several RefSeq curation levels.
- See status codes here
- RefSeq records with a status of **VALIDATED** or **REVIEWED** are intended to represent the current state of genomic knowledge.

FTP Downloads

See: `ftp://ftp.ncbi.nlm.nih.gov/refseq/`

Use shell commands `wget` or `curl`

EXERCISE: Search in "All" databases

Go to the NCBI web and search for gene **SMN1** in *All* databases

- What is in the output you get?
- How are organized the different queried databases?
- Why is there no link to the *RefSeq* or *GenBank* databases?
- How many *genomes* are associated to this gene ID?
- How many homologous genes are registered at the NCBI?
- Why are there two links to Human genes at the *HomoloGene*?
- If *D.rerio* has an homologous for gene SMN1 ... Why is there no link to the *D.rerio* genome in the *Genome* results?

Links:

http://www.ncbi.nlm.nih.gov/gquery/?term=SMN1

http://www.ncbi.nlm.nih.gov/genome/?term=SMN1

http://www.ncbi.nlm.nih.gov/homologene/?term=SMN1

EXERCISE: Assembly vs. Genomes

- How are NCBI *Assembly* and *Genome* databases different? Hint: Search for "Homo Sapiens" in both databases.
- Can you find the reference genome through the *Genome* output page?
- In which format is it available?
- Where is the file stored?
- Use `wget` to download its corresponding GFF file and `gunzip` to uncompress it. Explore the file. Read here about GFF formats.

Links:
www.ncbi.nlm.nih.gov/assembly/?term=Homo+sapiens
www.ncbi.nlm.nih.gov/genome/?term=Homo+sapien
www.ncbi.nlm.nih.gov/genome/?term=zebra+fish

EXERCISE: Gene version

- How many *versions* of the **NM_002020** gene have been submitted to the NCBI? Hint: find the *Nucleotid* information about the gene.