6 Database search

6.1 Biological databases

Biological databases contain biological information, mainly collected from molecular biology experiments, life science literature, and bioinformatics analyses.

Categories of databases

Annual Nucleic Acids Research database issue includes the following database categories.

- Nucleotide Sequence Databases
- RNA sequence databases
- Protein sequence databases
- Structure Databases
- Proteomics Resources
- Human and other Vertebrate Genomes
- Genomics Databases (non-vertebrate)
- Plant databases
- Human Genes and Diseases
- Metabolic and Signaling Pathways
- Immunological databases
- Microarray Data and other Gene Expression Databases
- Cell biology
- Organelle databases
- Other Molecular Biology Databases

NCBI (National Center for Biotechnology Information)

- It hosts a series of databases
- URL: http://www.ncbi.nlm.nih.gov

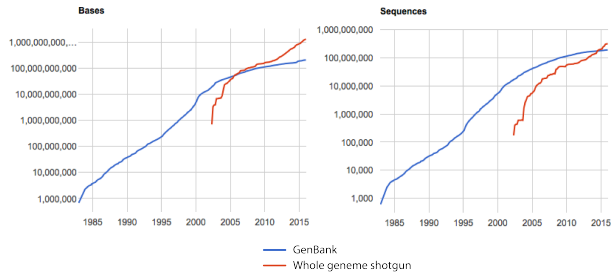

Figure 6.1: Growth of GenBank and WGS (source: NCBI)

Sequence data

- Identifier
- Sequence

Data format

FASTA is the most popular format for sequence data.

```
>gi|31563518|ref|NP_852610.1| microtubule-associated proteins 1A/1B light chain 3A isoform b
MKMRFFSSPCGKAAVDPADRCKEVQQIRDQHPSKIPVIIERYKGEKQLPVLDKTKFLVPDHVNMSELVKI
IRRRLQLNPTQAFFLLVNQHSMVSVSTPIADIYEQEKDEDGFLYMVYASQETFGFIRENE
```

Sequence search tools

- BLAST at NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi)
- BLAT/BLAST at Ensembl (http://www.ensembl.org/Multi/Tools/Blast)

Annotation data

Sequences databases usually contain annotations in addition to sequences.

- Notes and descriptions of important regions and components
- Meta data

Data format

Annotation data can be downloaded in many different formats. GFF is one of the popular formats of sequence annotations.

```
0  ##gff-version 3.2.1
1  ##sequence-region   ctg123 1 1497228
2  ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
3  ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001
4  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
5  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
6  ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
7  ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
```

Visualization tools

- UCSC Genome Browser (https://genome.ucsc.edu)
- Ensemble Genome Browser (http://www.ensembl.org)

Data download tools

- UCSC Table Browser (http://genome.ucsc.edu/cgi-bin/hgTables)
- Ensemble BioMart (http://www.ensembl.org/biomart)