Biological Databases

Bioinformatics Workshop 2009

Chi-Cheng Lin, Ph.D.
Department of Computer Science
Winona State University
clin@winona.edu

MnSCU CTL Discipline Workshop in Bioinformatics 2009

Biological Databases

- Data Domains
- Types of Databases - By Scope
- Types of Databases - By Level of Curation
- GenBank
- RefSeq

Acknowledgement: The presentation includes adaptations from NCBI's Introduction to Molecular Biology Information Resources Modules

Data Domains

- **Types of data generated by molecular biology research:**
  - Nucleotide sequences (DNA and mRNA)
  - Protein sequences
  - 3-D protein structures
  - Complete genomes and maps
- **Also now have:**
  - Gene expression
  - Genetic variation (polymorphisms)

Types of Databases - By Scope

- **Comprehensive**
  - Contain data from many organisms and many different types of sequences. Examples:
  - Nucleotide
    - GenBank (overview)
    - EMBL: European Molecular Biology Laboratory
    - DDBJ: DNA Data Bank of Japan
      (The three databases above comprise the International Nucleotide Sequence Database Collaboration and currently include sequence data from >120,000 species.)
  - Protein, such as Swiss-Prot
  - Protein Structure, such as PDB: Protein Data Bank
  - Genomes and Maps, such as Entrez Genomes

Types of Databases - By Scope

- **Specialized**
  - Contain data from individual organisms, specific categories/functions of sequences, or data generated by specific sequencing technologies.
  - Example: Flybase, Wormbase, etc.

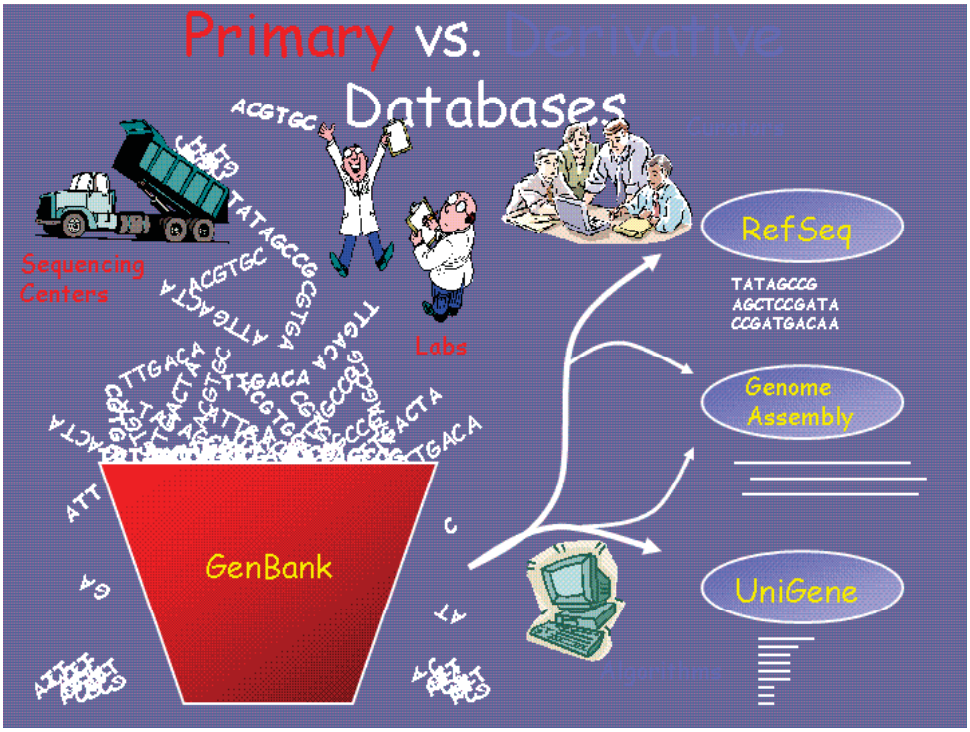
Types of Databases - By Level of Curation

- **Primary databases – Archival data**
  - Repository of information
  - Redundant; might have many sequence records for the same gene, each from a different lab
  - Submitters maintain editorial control over their records: what goes in is what comes out
  - No controlled vocabulary
  - Variation in annotation of biological features

Types of Databases - By Level of Curation

- **Secondary (derivative) databases – Curated data**
  - Non-redundant; one record for each gene, or each splice variant
  - Each record is intended to present an encapsulation of the current understanding of a gene or protein, similar to a review article
  - Records contain value-added information that have been added by an expert(s)

Primary vs. Derivative Databases



100's of Databases

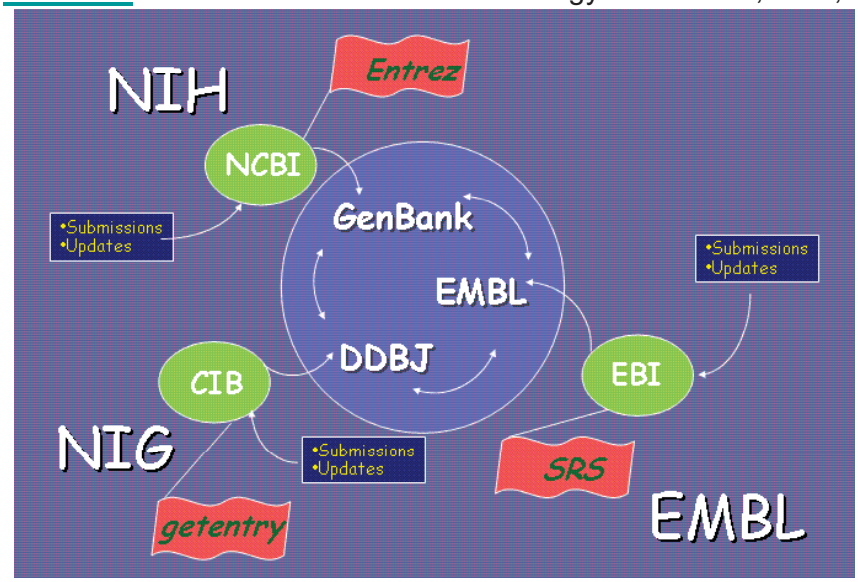
- 100's of databases available (http://www.expasy.ch/alinks.html). Which ones to use?
- Easiest to start with a single search system (such as Entrez) that combines data from the most commonly used comprehensive databases
- If user wants additional specialized databases, search the database and software directories

GenBank

- Archival database of nucleotide sequences from >130,000 organisms
- Records annotated with coding region (CDS) features also include amino acid translations
- Each record represents the work of a single lab
- Redundant; can have many sequence records for a single gene
- Part of the International Nucleotide Sequence Database Collaboration
- more information about GenBank... (http://www.ncbi.nlm.nih.gov/Sitemap/ResourceGuide.html#GenBank)

International Nucleotide Sequence Database Collaboration

- Collaboration among:
  - DDBJ - DNA Data Bank of Japan
  - EMBL - European Molecular Biology Laboratory, UK
  - GenBank - National Center for Biotechnology Information, NLM, NIH



RefSeq

- Database of reference sequences
- Curated
- Non-redundant; one record for each gene, or each splice variant, from each organism represented
- A representative GenBank record is used as the source for a RefSeq record
- Value-added information is added by an expert(s)
- Each record is intended to present an encapsulation of the current understanding of a gene or protein, similar to a review article
- Accessible through Entrez, BLAST, and FTP site
  - RefSeq records are available in various Entrez Databases such as Nucleotide, Protein, Genome, and are also accessible from Entrez Gene records
- More about RefSeq: http://www.ncbi.nlm.nih.gov/RefSeq/

Other Popular Databases

- OMIM (http://www.ncbi.nlm.nih.gov/omim/)
  - "Online Mendelian Inheritance in Man®, a comprehensive, authoritative, and timely compendium of human genes and genetic phenotypes."
- PubMed (http://www.ncbi.nlm.nih.gov/pubmed/)
  - "PubMed is a service of the U.S. National Library of Medicine that includes over 18 million citations from MEDLINE and other life science journals for biomedical articles back to the 1950s. PubMed includes links to full text articles and other related resources."
- NCBI databases (http://www.ncbi.nlm.nih.gov/Database/)