

Biological Databases (Part 1)

Kam D. Dahlquist, Ph.D.
 Department of Biology
 Loyola Marymount University

BIOL 367
 September 17, 2019

Outline

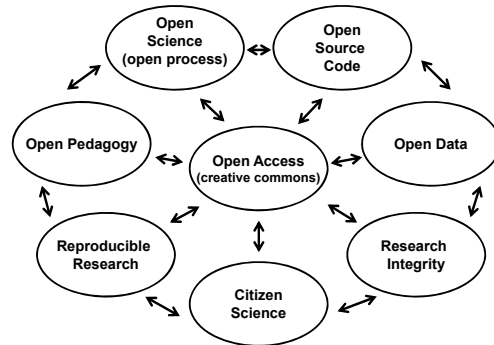
- What is bioinformatics?
- Open science ecosystem
- Life cycle of data
- What is a database anyway?
- Key vocabulary
- Classification and evaluation of biological databases
- “Gold Standard” databases

What is bioinformatics?

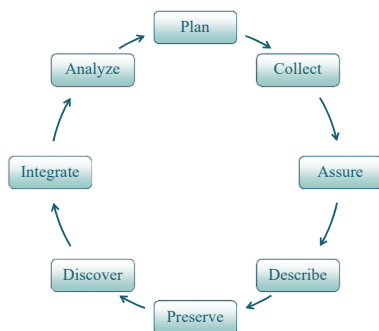
- Application of information technology (informatics) to biological data
- Informatics: representation, organization, manipulation, distribution, maintenance, use of digital information
- Interdisciplinary: biology, computer science, math, chemistry, physics, engineering . . .



Open Science Ecosystem

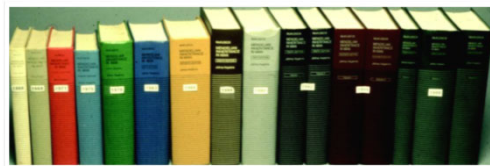


The Data Life Cycle



DataONE Education Module: Data Management. DataONE. Retrieved Nov12, 2012. From http://www.dataone.org/sites/all/documents/L01_DataManagement.pptx

A Database Is Simply a Collection of Information that Exists over a Long Period of Time



12 print editions of Mendelian Inheritance in Man 1966-1998; McKusick (2007)

- Most databases are stored using computers.
- Databases are managed by a database management system, a type of computer program.
- Examples of everyday databases...
- Increasingly, biological data is being stored in databases.

Database Management Systems Are Used by Programmers and System Administrators

- They perform the following functions:
 - create new databases with a particular *schema*
 - allow users to *query* the data
 - support *storage* of large amounts of data
 - enable *durability*: recovery of the database in the face of failures, errors, misuse
 - *control access* from many users

You Need to Understand Some Vocabulary When Working with Biological Databases

- **IDs** = identifiers (machine, not human readable)
 - unique identifiers = primary key
 - sometimes called accession numbers
- **Record** = entry in a database
- **Field** = type of data associated with a record
- When you search for data in a database you are executing a query.
- Different databases have different file formats in which you can download the data.
 - Some formats are standardized.

The Part of Biological Databases that You See Online Is a Graphical User Interface (GUI)

• It allows you to perform a subset of functions.

• It is usually an add-on to the actual database management system

Databases Can Be Classified by Type

- Type of data stored in them (think about the physical entity that is represented by the data)
 - Sequence (DNA, RNA, protein)
 - 3-D structure (DNA, RNA, protein)
 - Model Organism Databases (MODs)
 - Abstracts or scientific literature
 - Other types of high-throughput data or biological information
 - See January 1 issue of *Nucleic Acids Research*
- Type of data source
 - primary source vs. secondary source (*meta*)
 - curated vs. non-curated
 - electronic vs. human curation
 - staff vs. community curation
- Type of access/sponsoring organization
 - public vs. private (free vs. subscription)
 - large national or multinational entity or small lab group
- “Professional” or “hobby”