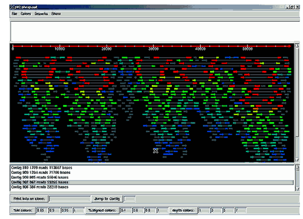
Bioinformatics: what is it?

- Fusion of biology & computer science
- Informatics: technologies for information management
- Uses information technology to store, curate, retrieve & analyze biological data
- Major areas of endeavor:
  – creation, storage & management of (large) biological data sets
  – development of tools (algorithms, statistical analysis) to determine relationships among members of these data sets
  – analysis & interpretation of biological data

Types of Biological Data

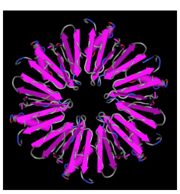
- Sequences: DNA, RNA, Protein



source: http://www.jgi.doe.gov/education/how/how_11.html

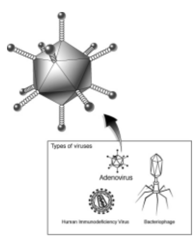
Types of Biological Data

- Structures of biological molecules



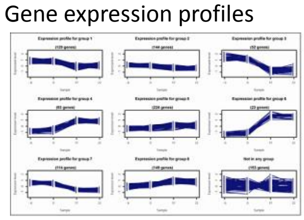
Tertiary structural model of HFQ protein from *E. coli*

source: Discovering Biology in a Digital World, http://scienceblogs.com/digitalbio/



Types of Biological Data

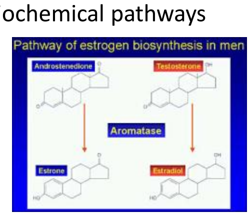
- Gene expression profiles



source: http://nai.arc.nasa.gov/team/index.cfm?page=projectreports&teamID=31&year=8&projectID=1658

Types of Biological Data

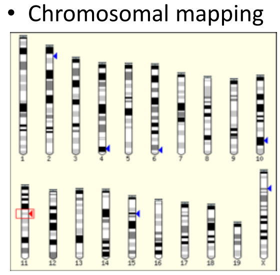
- Biochemical pathways



source: http://www.endotext.org/male/male17/male17.htm

Types of Biological Data

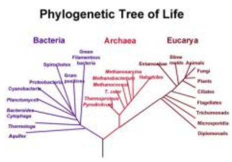
- Chromosomal mapping



source: http://animal.nibio.go.jp/research/genotyping/chromosome.png

Types of Biological Data

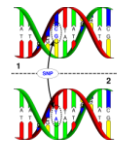
- Phylogenetic data



source: http://darwin.nmsu.edu/~molb470/fall2005/projects/pan/

Types of Biological Data

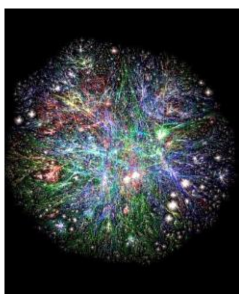
- Single Nucleotide Polymorphisms (SNPs)



source:
http://urgi.versailles.inra.fr/projects/GnpSNP/general_documentation.php

Bioinformatics & the Internet

- Advances in bioinformatics have paralleled advances in computer technology and the spread of knowledge via the World Wide Web
- Knowledge sharing via public databases has led to an explosion of additional knowledge



source: http://www.vlib.us/web/worldwideweb3d.html

Database

- Computerized archive for storage & organization of data
- Goal is ease of information retrieval
- Organization:
  - records (entries) consisting of
  - fields
    - each field holds a single piece of data
    - search mechanism often field-based

Historical perspective

- Major biological databases sprung from different sources, with different uses (and user communities) in mind
- Links between different types of information not always clear
- Major task in bioinformatics: reconciling different data sources (and formats), making them work cohesively

Sequence databases

- Primary sources: sites where new data are submitted by researchers
- GenBank
  (http://www.ncbi.nlm.nih.gov/Genbank/)
- EMBL
  (http://www.ebi.ac.uk/embl/)
- DDBJ
  (http://www.ddbj.nig.ac.jp/)

Features of primary sequence databases

- Predate the World Wide Web
- Relatively few in number
- Overlap one another (and themselves): sites now cooperate to mirror data
- Repositories of new information
  - may not be annotated
  - may not be complete
  - may not be good!

RefSeq

- http://www.ncbi.nlm.nih.gov/RefSeq/
- Aims to be one-stop shopping for sequence information
- Entries are:
  - non-redundant
  - annotated
  - consistent
- RefSeq is an example of a secondary, or "curated" database

More secondary sequence databases: a sampler

- Entrez Nucleotide: http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucleotide
- Unigene: http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene
- Homologene: http://www.ncbi.nlm.nih.gov/sites/entrez?db=homologene

A common thread

- You may have noticed that the last several databases had a major portion of their URLs in common:
  http://www.ncbi.nlm.nih.gov
- This is no accident; the National Center for Biotechnology Information (ncbi) is one of the premier sites for bioinformatics research, collecting several databases under a common umbrella
- A similar site, based in England, is the European Bioinformatics Institute:
  http://www.ebi.ac.uk/

Protein structure databases

- First major data collection effort: PIR (Protein Information Resource)
- Begun in 1970s by Margaret Dayhoff et. al. (http://pir.georgetown.edu/)
  - organized proteins into families based on sequence similarity
  - derived tables to reflect frequency of observed changes in closely related proteins
  - led to development of phylogenetic trees, PAM matrices

Protein structure databases

- Protein Data Bank (PDB): http://www.rcsb.org/pdb/home/home.do
- UniProt: http://www.uniprot.org/
- Entrez Protein: http://www.ncbi.nlm.nih.gov/sites/entrez?db=protein

Effective database searching: from the PubMed tutorial *

- **Developing a Search Strategy**
  - Before you can search for any information, you should first develop a search strategy.
- **What is a Search Strategy?**
  - A search strategy is a plan that helps you look for the information you need.

* http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/020_180.html

Continued from previous slide

- **Search Strategy Tips**
  - Identify the key concepts.
  - Determine alternative terms for these concepts, if needed.
  - Refine your search to dates, study groups, etc., as appropriate.
  - Practice helps. Strategies and styles will differ according to personal choice and professional discipline.

Information sources

- We have looked, and will continue to look, at many of the major biological databases
- There are other sources of information, however, and we don't want to ignore them
- Regardless of whether we want to search a biological database or a general-purpose database like Google, it is useful to know a thing or two about setting up queries

Forming a query

- A query is a request for information – a way of asking a question
- The proper structuring of a query can yield results that contain more pertinent information and less junk
- Query formulation is a critical research skill whether you are using general web search engines or specialized scientific databases

Relevance vs. recall

- No search query is perfect; we can only hope to strike a balance between:
  - relevance: the degree to which the results we get pertain to our information need
  - recall: the number of hits we get relative to the number of actual hits contained in the database

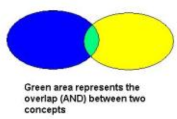
source: http://www.cdc.gov

Boolean logic

- System for stating how information should be divided or combined into compound sets
- All search engines and public biological databases use boolean logic in some form
- We form queries with Boolean logic by incorporating the operators AND, OR, and NOT

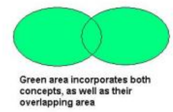
Boolean logic

- AND operation:
  - Narrows a search by requiring that both terms must appear in a query result
  - Example: breast AND cancer



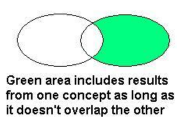
Boolean logic

- OR operation:
  - widens search by allowing query results that contain either term (or both)



Boolean logic

- NOT operation:
  - narrows search by excluding a term or terms



Boolean expressions & query formulation

- Boolean expressions are read left to right, just as arithmetic expressions are
- Operators, when combined, may not work left to right, however: consider this example from arithmetic:

  4 + 5 x 2
  - is the answer 14 or 18?
  - in arithmetic, we can clarify (and/or change) our intention by using parentheses: (4 + 5) x 2 = 18

Parentheses & Boolean expressions

- Can parenthesize Boolean expressions as well; part of expression in parentheses is evaluated first, working from inside out
- Good idea for particularly long search string
- Example:

  (breast AND cancer) AND (NOT inflammatory) AND (NOT estrogen) AND (NOT progesterone)

Quotation marks

- Many search engines and databases allow use of quotation marks to make keywords form whole phrases
- For example, “breast cancer” is actually more narrow than breast AND cancer because the quotes require the presence of the whole phrase, not just both terms somewhere in the results

Search Engines & Algorithms

- There are dozens of general-purpose search engines available for finding information on the world-wide web
- Each search engine has its own database, with a particular (and often proprietary) method for building the knowledge base and ranking pages

Google

- Consistently ranks among the best
  - Knowledge base is comprehensive
  - Pages are ranked by the number of links to them
  - Default combining term is AND
  - Can use quote marks and explicit boolean terms (and parenthesized phrases) or use Advanced Search to fill in form

Google

PubMed

- Web-based retrieval system for life science literature
- Part of Entrez
- Includes links to full-text articles, where available
- Provides point of access for MedLine, but more comprehensive than MedLine

PubMed and MeSH

- MeSH: Medical Subject Headings
  - controlled vocabulary for indexing of articles by subject
  - terms are assigned by human experts at the National Library of Medicine
  - hierarchical: broad terms are automatically expanded to include more specific terms (this is known as explosion)

Searching PubMed

- Enter query in box
  - can use parentheses for grouping
  - quotes and hyphens for phrase-based searching
  - Boolean operators must be in ALL CAPS
  - Can use qualified terms (tagged with one of the designators shown on next slide) to search specific fields

PubMed fields

Abstract	(AB)	ISSN	(IS)	Personal Name as Subject	(PS)
Copyright Information	(CI)	Issue	(IP)	Full Personal Name as Subject	(FPS)
Affiliation	(AD)	Journal Title Abbreviation	(TA)	Place of Publication	(PL)
Investigator Affiliation	(IRAD)	Journal Title	(JT)	Publication History Status	(PHST)
Article Identifier	(AID)	Language	(LA)	Publication Status	(PST)
Author	(AU)	Location Identifier	(LID)	Publication Type	(PT)
Full Author	(FAU)	Manuscript Identifier	(MID)	Publishing Model	(PUBM)
Corporate Author	(CN)	MeSH Date	(MHDA)	PubMed Central Identifier	(PMC)
Comments/Corrections		MeSH Terms	(MH)	PubMed Unique Identifier	(PMID)
Date Completed	(DCOM)	NLM Unique ID	(JID)	Registry Number/EC Number	(RN)
Date Created	(DA)	Number of References	(RF)	Substance Name	(NM)
Date Last Revised	(LR)	Other Abstract	(OAB)	Secondary Source ID	(SI)
Date of Electronic Publication	(DEP)	Other Copyright Information	(OCI)	Source	(SO)
Date of Publication	(DP)	Other ID	(OID)	Space Flight Mission	(SFM)
Entrez Date	(EDAT)	Other Term	(OT)	Status	(STAT)
Gene Symbol	(GS)	Other Term Owner	(OTO)	Subset	(SB)
General Note	(GN)	Owner	(OWN)	Title	(TI)
Grant Number	(GR)	Pagination	(PG)	Transliterated Title	(TT)
Investigator Name and Full	(IR)			Volume	(VI)
Investigator Name	(FIR)				

Scientific Literature & Authority

- Scientists trust their journals because contents are refereed
  - each paper vetted by group of experts prior to publication
  - author(s) may be required to make additions, corrections, even conduct further experiments before article is accepted

Reading Scientific Papers

- The next several slides are adapted from notes by Professor Gary Ritchison of Eastern Kentucky University
- His original notes may be found at:
  http://people.eku.edu/ritchisong/RITCHISO/801lecnotes.htm

Acquire background knowledge

- Papers are written for specialized audience; assume readers have subject matter experience and vocabulary
- You may need a dictionary, encyclopedia, and/or textbook(s) by your side as you read

Types of papers

- Review papers: provide historical perspective, summarize contributions of influential research, and may point out where additional work is needed.
- Reference section of these papers especially good source for primary literature
- Primary source work usually describes a single experiment or set of related experiments

Parts of papers

- Abstract: summary of paper and its findings
  - always read this first
  - will help you decide whether or not the rest of the paper is worth reading
- Introduction: describes study's objectives, often provides contextual information

Parts of papers

- Methods (sometimes Materials & Methods):
  - describes experimental design, use of controls, sampling techniques, etc.
  - probably the least likely section to contain useful information, unless you are conducting similar research

Parts of papers

- Results:
  - describes the findings of the experiment(s)
  - important to read this section carefully
  - pay attention to illustrations (figures and tables)

Parts of Papers

- Discussion (or Conclusion):
  - Describes the author's take on what his/her study or experiment shows
  - Should be read carefully and critically
    - do the data support the conclusions?
    - are your conclusions the same as the author's?
  - Go back to the Results section if you're not sure how to answer these questions

Parts of Papers

- References/Works Cited:
  - May lend additional authority to paper
  - Should not be empty!
  - Often useful for delving further into subject

Collecting bibliographic information

- Author(s)
- Date of publication (year is usually sufficient)
- Title
- Publisher or journal
- Volume & issue (if applicable)
- Page numbers
- Editor (if paper came from edited volume)