Everything you ever wanted to know about DATA & SOFTWARE
(...but were afraid to ask)

Presented in the Introduction to Genomic Medicine series on 2 May 2011

Kristi Holmes, PhD
Bioinformaticist, Becker Medical Library
Local and National Outreach Coordinator, VIVO (http://vivoweb.org)
Information

- Consumption
  - How to find, assess, process, & apply
- What do these results mean and how did I get them?
- What do I do next?
bioinformatics
Outline

• General concepts to keep in mind when thinking about data
• Data sources
• Necessary software and tools
• Tutorials, classes, support resources
• On the horizon...
information overload
Top 10 List: THINGS TO KEEP IN MIND WHEN THINKING ABOUT DATA

10 - **Scope**: What are you looking for?

9 - **Data quality**: Depends heavily on data curation, validation can be difficult (huge range)

8 - **Freshness of data**: Well-maintained databases with current information can indicate a higher quality database

7 - **Data quantity**: Never assume completeness of a data set

6 - **Availability**: Download restrictions and IP restrictions

5 - **Technical architecture**: Interfaces which promote ease of use

4 – **Boolean Searching**: AND, OR, NOT

3 – “”,*,(),: write better searches!

2 – **Use limits to narrow your search**: again, what are you looking for?

1 – **Use information directories**: collections of information
Top 10 List: THINGS TO KEEP IN MIND WHEN THINKING ABOUT DATA

10 - Scope: What are you looking for?

9 - Data quality: Depends heavily on data curation, validation can be difficult (huge range)

8 - Freshness of data: Well-maintained databases with current information can indicate a higher quality database

7 - Data quantity: Never assume completeness of a data set

6 - Availability: Download restrictions and IP restrictions

5 - Technical architecture: Interfaces which promote ease of use

4 – Boolean Searching: AND, OR, NOT

3 – “”,*,(): write better searches!

2 – Use limits to narrow your search: again, what are you looking for?

1 – Use information directories: collections of information
Top 10 List: THINGS TO KEEP IN MIND WHEN THINKING ABOUT DATA

10 - Scope: What are you looking for?
9 - Data quality: Depends heavily on data curation, validation can be difficult (huge range)
8 - Freshness of data: Well-maintained databases with current information can indicate a higher quality database
7 - Data quantity: Never assume completeness of a data set
6 - Availability: Download restrictions and IP restrictions
5 - Technical architecture: Interfaces which promote ease of use

4 – Boolean Searching: AND, OR, NOT
3 – “”, *, (): write better searches!
2 – Use limits to narrow your search: again, what are you looking for?
1 – Use information directories: collections of information

4 – Boolean Searching
What is a Boolean Operator?

- **AND**: To ‘AND’ two search terms together instructs Entrez to find all documents that contain BOTH terms.

- **OR**: To ‘OR’ two search terms together instructs Entrez to find all documents that contain EITHER term.

- **NOT**: To ‘NOT’ two search terms together instructs Entrez to find all documents that contain search term 1 BUT NOT search term 2.

[Links to relevant resources]
The Entrez search rules and syntax for using Boolean operators are:

1. Boolean operators AND, OR, NOT must be entered in UPPERCASE (e.g., promoters OR response elements).

2. Entrez processes all Boolean operators in a left-to-right sequence. The order in which Entrez processes a search statement can be changed by enclosing individual concepts in parentheses. The terms inside the parentheses are processed first as a unit and then incorporated into the overall strategy. For example, the search statement: g1p3 AND (response element OR promoter) is processed by Entrez by ORing the terms response element OR promoter first and then ANDing the resulting set of documents with g1p3. g1p3 AND (response element OR promoter)

3. See Entrez Searching Options for more information on using Boolean Operators and Entrez Search Field Qualifiers. (this is a great resource!)

Write better searches!

- The use of quotation marks can change your search results significantly!
  - 1000 Genomes Project = 6,980,000 results
  - “1000 Genomes Project” = 125,000 results

- The use of wildcards can change your search results significantly!

- The use of parentheses can change your search results significantly!
  - g1p3 AND (response element OR promoter) retrieves 9 records
  - g1p3 AND response element OR promoter retrieves 148,553 records

[An unmatched left parenthesis creates an unresolved tension that will stay with you all day.](http://xkcd.com/859/)
2 – Use limits to narrow your search

Enter your search in command language, indicating field qualifiers in square brackets [].

- **Syntax:** term[field] BOOLEAN term[field] BOOLEAN term[field] etc.
- **Boolean operators** (AND, OR, NOT) must be written in **UPPER CASE**

Boolean operators are processed from **left to right** unless parentheses are used for nesting. If **parentheses** are used, the portions of the query in parentheses will be processed first, then the remaining Boolean operators will be processed from left to right.

- **LIST** of Search Fields and Qualifiers and search fields available by database
- Search results for **field AND handbook[book]**

**presenilin 1[titl] AND human[orgn] (Nucleotide database)**

Some search fields at NCBI

- Properties field
  - molecule type (e.g., biomol_genomic[prop])
  - GenBank division (e.g., gbdiv_EST[prop])
  - gene location (e.g., gene in mitochondrion[prop])
  - source database (e.g., srcdb_refseq[prop])
  - citation status (e.g., cit_inpress[prop])
  - HTGS phase (e.g., htgs_phase3[prop])
  - other features (e.g., partial[prop], complete[prop])
  - submission types (e.g., population study[prop])
- Feature Key (nucleotides only)
- Sequence Length
- Molecular Weight (proteins only)
- Organism (uses Taxonomy db)

Use field limits to see what you’re looking for...

1 – Take advantage of information directories

**Nucleic Acids Research Database Issue**
- Complete table of contents for the NAR database issue  *(Tip: to see the table of contents from the database issue for a previous year, just reduce the volume number in the URL (to the complete table of contents) by one.)*
- Searchable database of summary papers

**Nucleic Acids Research Web Server Issue**
- 2010 Web Server complete table of contents
- Searchable database of web server summaries

**Nucleic Acids Research Methods Index**

**Bioinformatics Links Directory** *(described in an NAR article, July 2007 web server issue)*

**ExPASY Life Science Directory**
- >1000 links on a single page, organized by category

**BioMed Central Databases** collection

**Biocatalog by EBI**
- Database providing summary and access information for a wide range of molecular biology databases and software; browse category of interest or search complete db

**Online Bioinformatics Resources Collection (OBRC)** from the HSLS, University of Pittsburgh

**Max Planck Society bioinfowiki**
- 250 tools and resources which have been developed within the institutes and research groups of the Max Planck Society.
Information sources...
Top 2 List: DATA SOURCES

2 – biological data
   (NCBI, EBI, many others)

1 – bibliographic data
   (Web of Science, Scopus, PubMed)
Top 10 List: SOME INTERESTING TOOLS

10 – Sequence Manipulation Suite
9 – BLAST
8 – Blink
7 – Flink
6 – Multiple sequence alignment software
5 – Partek
4 – JMP Genomics
3 – GeneGo Metacore
2 – Genome Browsers (UCSC, MapViewer, G-Browse)
1 – BioMart
Sequence Manipulation Suite

- A collection of JavaScript programs for generating, formatting, and analyzing short DNA and protein sequences. It is commonly used by molecular biologists, for teaching, and for program and algorithm testing.

- See the about the SMS [page](http://www.ualberta.ca/~stothard/javascript/index.html) for more information about individual Sequence Manipulation Suite programs.
9 – BLAST

Basic Local Alignment Search Tool

• BLAST finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

• BLAST assembled genomes

• Basic BLASTs (BLASTn, BLASTp, BLASTx, tBLASTn, tBLASTx)

• Specialized BLASTs can accomplish a host of tasks

9 – BLAST
Specialized BLAST

Make specific primers with Primer-BLAST
Search trace archives
Find conserved domains in your sequence (cds)
Find sequences with similar conserved domain architecture (cdart)
Search sequences that have gene expression profiles (GEO)
Search immunoglobulins (IgBLAST)
Search using SNP flanks
Screen sequence for vector contamination (vecscren)
Align two (or more) sequences using BLAST (bl2seq)
Search protein or nucleotide targets in PubChem BioAssay
Search SRA transcript and genomic libraries
Constraint Based Protein Multiple Alignment Tool
Needleman-Wunsch Global Sequence Alignment Tool
Search RefSeqGene
Search WGS sequences grouped by organism
• The output includes the positions of up to 200 BLAST hits on the query sequence, scores, and alignments. (View sample BLink output for human MLH1 protein.)

• BLink offers a variety of display options, including the distribution of hits by taxonomic grouping, the best hit to each organism, the protein domains in the query sequence, similar sequences that have known 3-D structures, and more.

• Additional options allow you to specify which taxa you would like to exclude, increase or decrease the BLAST cutoff score, or filter the BLAST hits to show only those from a specific source database, such as RefSeq or Swiss-Prot.

• [Link to BLink help page](http://www.ncbi.nlm.nih.gov/sutils/static/blinkhelp.html)
Multiple Sequence Alignment (MSA) is generally the alignment of three or more biological sequences (protein or nucleic acid) of similar length.

ClustalW2 MSA
Kalign MSA
MAFFT MSA
MUSCLE MSA
T-Coffee MSA

WebPRANK - phylogeny-aware multiple sequence alignment program

http://www.ebi.ac.uk/Tools/msa/
Partek Genomic Suite

Partek Genomics Suite is a comprehensive suite of advanced statistics and interactive data visualization specifically designed to reliably extract biological signals from noisy data. The software is unique in supporting all microarray and next generation sequencing technologies, allowing for analysis of multiple applications in one complete solution.
Integrated Genomics in Partek Genomics Suite
4 – JMP Genomics

- JMP Genomics is a SAS-based suite of tools to analyze Expression, Genetics, Copy Number, Exon and Proteomics data.
3 – GeneGo MetaCore

- Integrated knowledge database and software suite for pathway analysis of experimental data and gene lists
- Data types includes microarray and sequence-based gene expression, SNPs and CGH arrays, proteomics, metabolomics, Co-IP pull-out and other custom interactions
Software Managed by Becker Library

Becker Medical Library licenses and supports many different types commercially available software. For assistance with any of these packages or to submit requests for new software, please email BioInfo@wusm.wustl.edu.

Cambridge Structural Database

The Cambridge Structural Database (CSD) is the principal product of the CCDC. It is the central focus of the CSD System, which also comprises software for database access, structure visualization and data analysis, and structural knowledge bases derived from the CSD.

CSD software has been installed on one of the workstations located in the west bank of computers on the main floor of Becker Medical Library. This software is available for use by any WUSM faculty, student, or staff member.

DiscoveryGate

The Washington University community now has access to a new, innovative resource tool to find basic, molecular, biological, medicinal and other chemistry-related information. Medical School clinical, pre-clinical and research faculty, staff and students can use the web-based DiscoveryGate interface to access Crossfire Beilstein, and Gmelin.

All users will need to install some helper applications when they first access the site. Please contact the Becker Medical Library’s Information Services Desk at 314-362-7625 or email to Bernard@becker.wustl.edu for technical assistance.
Genome browsers organize tremendous volumes of genome sequence data, adding context to genomic sequence with many types of annotations. Several major genome web browsers are widely used to search, retrieve, and display genome information for human and numerous other species. 

Ensembl, Map Viewer, UCSC Genome Browser, and the Integrated Microbial Genomes (IMG) browser, Gbrowse

http://www.openhelix.com/cgi/tutorialInfo.cgi?id=65
BioMart is an open source query-oriented data management system developed jointly by the Ontario Institute for Cancer Research (OICR) and the European Bioinformatics Institute (EBI).
Top 10 List: INTERESTING TOOLS

10 – Sequence Manipulation Suite
9 – BLAST
8 – Blink
7 – Flink
6 – Multiple sequence alignment software
5 – Partek
4 – JMP Genomics
3 – GeneGo Metacore
2 – Genome Browsers (UCSC, MapViewer, G-Browse)
1 – BioMart
and this is just the tip of the iceberg!
Top 8 List: WHERE TO LEARN MORE & TRY THINGS OUT

• EBI Training: http://www.ebi.ac.uk/training/
• Random resources (wikis, slides, etc.)
• Training through the library (NCBI training in July)
• Consultations through the library
• Training through the CRTC and CBMI
• The Research Pod: http://becker.wustl.edu/services/researchpod.html
• OpenHelix: http://www.openhelix.com/
TUTORIALS

• **2Can Support Portal** The bioinformatics educational resource.
• NCBI Education site
• OpenHelix Tutorials – Becker Medical Library subscribes to the entire suite and it is available across both campuses. You will need to register on the OpenHelix site using your “wustl.edu” email address for access.
  • All tutorials
  • Free tutorials
On the horizon...
Top 5 List: ON THE HORIZON

• Visualization
• The semantic web
• Social networking
• Data (curation, sharing, preservation)
• Nanopublications
Visualization

http://www.computerworld.com/s/article/9215504/22_free_tools_for_data_visualization_and_analysis
The Semantic Web

- The Semantic Web is a "web of data" that enables machines to understand the semantics, or meaning, of information on the World Wide Web. It extends the network of hyperlinked human-readable web pages by inserting machine-readable metadata about pages and how they are related to each other, enabling automated agents to access the Web more intelligently and perform tasks on behalf of users. (Wikipedia)

- Information is stored using the Resource Description Framework (RDF).
- Data are structured in the form of “triples” as subject-predicate-object.
- Concepts and their relationships use a shared ontology to facilitate the exchange of data from multiple sources and for multiple purposes.

- By storing data in RDF and using standard ontologies, the information can either be displayed in a human readable web page or delivered directly to other systems as RDF. This allows data to be harvested, aggregated, and integrated into the Linked Open Data cloud.

- Tools that use and repurpose Linked Data are increasing in popularity and are likely to become a significant part of the research work flow in the years to come.

http://computer.howstuffworks.com/semantic-web1.htm
Linked Open Data

http://richard.cyganiak.de/2007/10/lod/lod-datasets_2010-09-22_colored.html
VIVO: a semantic web-based collaboration and profiling application at WU and other institutions
Data

Data management and dissemination policies:

http://datadryad.org/

http://figshare.com/

...etc.
nanopublications

Some reading:

Image credits

- http://upload.wikimedia.org/wikipedia/commons/e/e1/Blind.JPG
- http://www.flickr.com/photos/dirtysouth/2982720117/
- http://www.flickr.com/photos/thenvlue/5617535720/
- http://www.indiamike.com/photopost/data/500/IMG_1836.JPG
- http://www.flickr.com/photos/kyle_johnson/2887121154
- http://www.genego.com/metacore.php
- http://www.flickr.com/photos/jazzysteps/5556449519
- http://www.flickr.com/photos/ghb624/1356102993
- http://www.flickr.com/photos/goobi/4021009835/