Bioinformatics Public Resources

CMSC423
Héctor Corrada Bravo
Center for Bioinformatics and Computational Biology
University of Maryland

Bioinformatics Databases

- General
  - GenBank - US
  - EMBL - Europe
- Specialized by data type
  - NCBI Trace Archive – raw sequencing data
  - SwissProt – curated protein information
  - KEGG – biological pathways
  - Gene Expression Omnibus – microarray data
- Specialized by organism
  - ZFIN – zebrafish
  - SGD – yeast
  - WormBase - worms

What data gets stored?

- DNA
  - string of letters
  - quality information, maybe chromatograms/intensities
  - location of genes/transcripts (ranges along a chromosome)

Proteins

  - string of letters
  - protein domains
  - 3D coordinates of each atom

Pathways

- graph of interactions between genes

For all – often store link to scientific articles related to data

How is data accessed?

- Gene by gene/object by object – targeted at manual inspection of data
  - usually lots of clicking involved
  - simple search capability
  - similarity searches in addition to text queries
- Bulk – targeted at computational analyses
  - often programmatic access through web server
  - most frequently – just bulk download (ftp)

NCBI

- National Center for Biotechnology Informatics
- Virtually all biological data generated in the US gets stored here!
- One-stop-shop for biological data
- Primarily focused on gene-by-gene analyses
- Provides simple scripts for programmatic access
- Provides ftp access for bulk downloads

- http://ncbi.nlm.nih.gov

EMBL

- European Molecular Biology Lab
- European version of NCBI
- Biomart query builder - really nice!
- http://www.embl.de

Expasy proteomics server

- Home of SwissProt and other useful information on proteins

- http://www.expasy.org

KEGG

- Kyoto Encyclopedia of Genes and Genomes
- Central repository of pathway information
  - No longer freely available...
- http://www.genome.jp/kegg/

Interaction Data

- General Gene/Protein interaction data

- http://string-db.org/

Ontologies

- Gene Ontology. http://www.geneontology.org

- The Gene Ontology project provides a controlled vocabulary to describe gene and gene product attributes in any organism. (text from GO homepage)
  - Note: similar to semantic web

- GO not the only one: http://www.obofoundry.org

- BioPortal: https://bioportal.bioontology.org/

One stop resource

- Everything you want to know about your favorite (human) gene:

- http://www.genecards.org/

Human genetic variation

- International hapmap project
  - Catalog of discovered common variants in human population
  - Includes single nucleotide polymorphisms, insertions, deletions
  - Initial phases included a small number of populations (caucasians, yoruba, chinese, etc.), but is rapidly expanding
- http://hapmap.ncbi.nlm.nih.gov/
- Data repository: dbSNP, http://www.ncbi.nlm.nih.gov/projects/SNP/
- Variation associated with disease: https://www.omim.org/
- Wiki: https://snpedia.com/
-

Biomedical Literature

- Pubmed: online index and search engine specific to biomedical literature: http://www.ncbi.nlm.nih.gov/pubmed
- Not all papers are open access! Pubmed is not a repository, it's an index. Some papers require journal subscription (personal or library), or one-time payment.
- This is changing to a degree. Big drive for open access publishing:
  - Most research sponsored by federal funds, why do we pay again to see results?
  - Pubmed Central: http://www.ncbi.nlm.nih.gov/pmc/ This is a repository, you can access papers openly here.
  - Open access journals:
    - BMC: http://www.biomedcentral.com/
    - PLoS: http://www.plos.org/
- Preprint servers: http://biorxiv.org/ https://arxiv.org/archive/q-bio

Experimental Data

- Short Read Archive
  - Experiments using next-generation sequencing
  - Stores reads and qualities (millions per sample per experiment)
    - Rarely stores raw data (more next lecture)
- http://www.ncbi.nlm.nih.gov/Traces/sra/

Experimental Data

- Gene Expression Omnibus
- Repository of microarray data
  - mainly gene expression
  - also *some* genotyping array data
  - some experiments with sequencing data (processed data, not raw reads like SRA)
- *This is an extremely valuable resource!*
- http://www.ncbi.nlm.nih.gov/geo/

Experimental Data

- dbGAP: Database of genotypes and phenotypes.
- Genotype experiment repository
- Large scale human genotyping studies along with some trait (phenotype) information about subjects
- Closed access: you must request access by submitting proposal and description of what you plan to do with the data. Why?
- http://www.ncbi.nlm.nih.gov/gap

Large-scale data producing projects

- Second-generation sequencing has allowed the explosion of large-scale projects beyond genome sequencing
- Functional Genomics
  - Encode (Encylopedia of DNA elements): expression, protein/DNA binding data, histone modification data, on reference human *cell lines.* http://www.genome.gov/10005107
  - Similar goal but using *primary human tissue.* (very new as of Fall 2013, not much data, but keep an eye). http://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/
  - ModEncode: Similar goal to Encode, but for model organisms (fly and worm). http://www.modencode.org/

Large-scale data generating projects

- 1000 genomes project: lots and lots of sequence data for a large number of human genomes. Data released early and often http://www.1000genomes.org/
- Genome 10k: Sequence data for 10,000 vertebrate species. https://genome10k.soe.ucsc.edu/
- Personal Genome Project: sequence data for human genomes along with *health* and *trait* data. http://www.personalgenomes.org/
  - Subjects opt-in, thus not restricted access
- The Cancer Genome Atlas: Sequence, expression, methylation, other for a number of different cancer types and corresponding normal tissue: http://cancergenome.nih.gov/
  - Some is restricted access

Genome Browsers

- UCSC Genome Browser: http://genome.ucsc.edu
- Ensembl Genome Browser: http://ensembl.org
- GBrowse: http://www.gmod.org

- Not just for browsing, their data can be accessed programmatically
- They have database backends that can be queried...

NCBI Programmatic Access

- http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html
  - must write your own HTTP client (LWP Perl module helps)
  - queries go directly to web server
  - data returned in XML
- http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?cmd=show&f=doc&m=obtain&s=stips
  - stub script provided (query_tracedb)
  - queries still go through web server
  - data returned in a variety of user selected formats
- For both, limits are set on the amount of data retrieved, e.g. less than 40,000 records at a time
- Download procedure:
  - figure out # of records to be retrieved ("count" query)
  - read data in allowable chunks

One more plug for R/Bioconductor

- An API for almost everything discussed here can be found in Bioconductor.

- Examples:

  - biomaRt: http://bioconductor.org/packages/release/bioc/html/biomaRt.html

  - UCSC genome browser track data: http://bioconductor.org/packages/release/bioc/html/rtracklayer.html

  - GEO: http://bioconductor.org/packages/release/bioc/html/GEOquery.html