INTRODUCTION TO BIOINFORMATICS

Please take the initial BIOINF525 questionnaire:
<http://tinyurl.com/bioinf525-questions>

Barry Grant
University of Michigan
www.thegrantlab.org

BIOINF 525 http://bioboot.github.io/bioinf525_w16/ 12-Jan-2016

COURSE LOGISTICS

Lectures: Tuesdays 2:30-4:00 PM
Rm. 2062 Palmer Commons

Labs: Session I: Thursdays 2:30-4:00 PM
Session II: Fridays 10:30-12:00 PM
Rm. 2036 Palmer Commons

Website: http://tinyurl.com/bioinf525-w16
Lecture, lab and background reading material plus homework and course announcements

MODULE OVERVIEW

Objective: Provide an introduction to the practice of bioinformatics as well as a practical guide to using common bioinformatics databases and algorithms

1.1 • Introduction to Bioinformatics
1.2 • Sequence Alignment and Database Searching
1.3 • Structural Bioinformatics
1.4 • Genome Informatics: High Throughput Sequencing Applications and Analytical Methods

TODAYS MENU

Overview of bioinformatics
• The what, why and how of bioinformatics?
• Major bioinformatics research areas.
• Skepticism and common problems with bioinformatics.

Bioinformatics databases and associated tools
• Primary, secondary and composite databases.
  • Nucleotide sequence databases (GenBank & RefSeq).
  • Protein sequence database (UniProt).
  • Composite databases (PFAM & OMIM).

Database usage vignette
• Searching with ENTREZ and BLAST.
• Reference slides and handout on major databases.

HOMEWORK

☐ Complete the initial course questionnaire:
http://tinyurl.com/bioinf525-questions

☐ Check out the “Background Reading” material on Ctools:
http://tinyurl.com/bioinf525-w16

☐ Complete the lecture 1.1 homework questions:
http://tinyurl.com/bioinf525-quiz1
Q. What is Bioinformatics?

"Bioinformatics is the application of computers to the collection, archiving, organization, and analysis of biological data.”

[AFTER Orengo, 2003]

… Bioinformatics is a hybrid of biology and computer science…

Bioinformatics is computer aided biology!

Computer based management and analysis of biological and biomedical data with useful applications in many disciplines, particularly genomics, proteomics, metabolomics, etc...

MORE DEFINITIONS

‣ “Bioinformatics is conceptualizing biology in terms of macromolecules and then applying “informatics” techniques (derived from disciplines such as applied maths, computer science, and statistics) to understand and organize the information associated with these molecules, on a large-scale.


‣ “Bioinformatics is research, development, or application of computational approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize and analyze such data.”

National Institutes of Health (NIH) (http://tinyurl.com/l3gxr6b)

Key Point: Bioinformatics is Computer Aided Biology

Major types of Bioinformatics Data

Genomes DNA & RNA sequence DNA & RNA structure Protein families, motifs and domains Protein interactions Pathways Systems Gene expression Protein sequence Protein structure Chemical entities Pathways Systems Literature and ontologies DNA & RNA structure DNA & RNA sequence Gene expression Protein sequence Protein structure Chemical entities Pathways Systems Literature and ontologies Genomes DNA & RNA sequence DNA & RNA structure Protein families, motifs and domains Protein interactions Pathways Systems Gene expression Protein sequence Protein structure Chemical entities Pathways Systems Literature and ontologies

Goal: Integrate sequence, 3D structure, expression patterns, interaction and function of biomolecules to gain a deeper understanding of biological mechanisms, process and systems.
Major types of Bioinformatics Data

- Genomes
- Protein sequence
- DNA & RNA structure
- Protein interactions
- Pathways
- Systems

Goal: Bioinformatics aims to bridge the gap between data and knowledge.

BIOINFORMATICS RESEARCH AREAS

Include but are not limited to:
- Organization, classification, dissemination and analysis of biological and biomedical data (particularly ‘omics’ data).
- Biological sequence analysis and phylogenetics.
- Genome organization and evolution.
- Regulation of gene expression and epigenetics.
- Biological pathways and networks in healthy & disease states.
- Protein structure prediction from sequence.
- Modeling and prediction of the biophysical properties of biomolecules for binding prediction and drug design.
- Design of biomolecular structure and function.

With applications to Biology, Medicine, Agriculture and Industry

Where did bioinformatics come from?

Bioinformatics arose as molecular biology began to be transformed by the emergence of molecular sequence and structural data.

Recap: The key dogmas of molecular biology
- DNA sequence determines protein sequence.
- Protein sequence determines protein structure.
- Protein structure determines protein function.
- Regulatory mechanisms (e.g. gene expression) determine the amount of a particular function in space and time.

Why do we need Bioinformatics?

Bioinformatics is necessitated by the rapidly expanding quantities and complexity of biomolecular data.

- Bioinformatics provides methods for the efficient:
  - storage
  - annotation
  - search and retrieval
  - data integration
  - data mining and analysis

Why do we do Bioinformatics?

- A “bioinformatics approach” involves the application of computer algorithms, computer models and computer databases with the broad goal of understanding the action of both individual genes, transcripts, proteins and large collections of these entities.
How do we actually do Bioinformatics?

Pre-packaged tools and databases
- Many online
- New tools and time consuming methods frequently require downloading
- Most are free to use

Tool development
- Mostly on a UNIX environment
- Knowledge of programming languages frequently required (Python, Perl, R, C, Java, Fortran)
- May require specialized or high performance computing resources…

SIDE-NOTE: SUPERCOMPUTERS AND GPUS

Skepticism & Bioinformatics
We have to approach computational results the same way we do wet-lab results:
- Do they make sense?
- Is it what we expected?
- Do we have adequate controls, and how did they come out?
- Modeling is modeling, but biology is different…
  What does this model actually contribute?
- Avoid the miss-use of ‘black boxes’

Common problems with Bioinformatics
Confusing multitude of tools available
- Each with many options and settable parameters

Most tools and databases are written by and for nerds
- Same is true of documentation - if any exists!

Most are developed independently

Notable exceptions are found at the:
- EBI (European Bioinformatics Institute) and
- NCBI (National Center for Biotechnology Information)
Even Blast has many settable parameters

Related tools with different terminology

Key Online Bioinformatics Resources: NCBI & EBI

The NCBI and EBI are invaluable, publicly available resources for biomedical research

The NCBI and EBI are invaluable, publicly available resources for biomedical research

The NCBI and EBI are invaluable, publicly available resources for biomedical research

The NCBI and EBI are invaluable, publicly available resources for biomedical research

Key Online Bioinformatics Resources: NCBI & EBI

The NCBI and EBI are invaluable, publicly available resources for biomedical research

NCBI's mission includes:

• Establish public databases
• Develop software tools
• Education on and dissemination of biomedical information

We will cover a number of core NCBI databases and software tools in the lecture


http://www.ebi.ac.uk

http://www.ebi.ac.uk
Notable NCBI databases include: GenBank, RefSeq, PubMed, dbSNP and the search tools ENTREZ and BLAST

European Bioinformatics Institute (EBI)

- Created in 1997 as a part of the European Molecular Biology Laboratory (EMBL)
- EBI’s mission includes:
  - providing freely available data and bioinformatics services
  - and providing advanced bioinformatics training
- We will briefly cover several EBI databases and tools that have advantages over those offered at NCBI
The EBI makes available a wider variety of online tools than NCBI. These include multiple tools in the following areas:

- **Database sequence searches**
  - FASTA, BLAST, InterProScan

- **Sequence alignments**
  - Pairwise: Needle, LAlign
  - Multiple: ClustalW, T-Coffee, MUSCLE, Jalview

- **Protein structure search and alignment**
  - PDBeFold, DALI

- **Genome browsers**
- **Gene expression analysis**
- **Protein function analysis**

The EBI also provides a growing selection of online tutorials on EBI databases and tools.

Notable EBI databases include: ENA, UniProt, Ensembl and the tools FASTA, BLAST, InterProScan, ClustalW, T-Coffee, MUSCLE, DALI, HMMER.

**What is a database?**

A computerized store of data that is organized to provide efficient retrieval. Key database features allow for:
- Adding, changing, removing and merging of records
- User-defined queries and extraction of specified records

Desirable features include:
- Contains the data you are interested in
- Allows fast data access
- Provides annotation and curation of entries
- Provides links to additional information (possibly in other databases)
- Allows you to make discoveries
Bioinformatics Databases

The most popular bioinformatics databases focus on:

- Biomolecular sequence (e.g. GenBank, UniProt)
- Biomolecular structure (e.g. PDB)
- Vertebrate genomes (e.g. Ensemble)
- Small molecules (e.g. PubChem)
- Biomedical literature (e.g. PubMed)

There are also many "boutique" databases for:

- Classifying protein families, domains and motifs (e.g. Pfam, PROSITE)
- Specific organisms (e.g. WormBase, FlyBase)
- Specific proteins of biomedical importance (e.g. KinaseDB, GPCRDB)
- Specific diseases, mutations (e.g. OMIM, HGMD)
- Specific fields or methods of study (e.g. GOA, IEDB)

See Online: Handout_Major_Databases.pdf

Primary, secondary & composite databases

Bioinformatics databases can be usefully classified into primary, secondary and composite according to their data source.

- Primary databases (or archival databases) consist of data derived experimentally.
  - GenBank: NCBI's primary nucleotide sequence database.
  - PDB: Protein X-ray crystal and NMR structures.
- Secondary databases (or derived databases) contain information derived from a primary database.
  - RefSeq: non redundant set of curated reference sequences primarily from GenBank.
  - Pfam: protein sequence families primarily from UniProt and PDB
- Composite databases (or metadatabases) join a variety of different primary and secondary database sources.
  - OMIM: catalog of human genes, genetic disorders and related literature
  - GENE: molecular data and literature related to genes with extensive links to other databases.

Finding Bioinformatics Databases

The most popular bioinformatics databases focus on:

- Biomolecular sequence (e.g. GenBank, UniProt)
- Biomolecular structure (e.g. PDB)
- Vertebrate genomes (e.g. Ensemble)
- Small molecules (e.g. PubChem)
- Biomedical literature (e.g. PubMed)

The are also many popular "boutique" databases for:

- Classifying protein families, domains and motifs (e.g. Pfam, PROSITE)
- Specific organisms (e.g. WormBase, FlyBase)
- Specific proteins of biomedical importance (e.g. KinaseDB, GPCRDB)
- Specific diseases, mutations (e.g. OMIM, HGMD)
- Specific fields or methods of study (e.g. GOA, IEDB)

http://www.oxfordjournals.org/nar/database/c/
GENBANK & REFSEQ: NCBI’S NUCLEOTIDE SEQUENCE DATABASES

What is GenBank?
• GenBank is NCBI’s primary nucleotide only sequence database
  ‣ Archival in nature - reflects the state of knowledge at time of submission
  ‣ Subjective - reflects the submitter point of view
  ‣ Redundant - can have many copies of the same nucleotide sequence
• GenBank is actually three collaborating international databases from the US, Japan and Europe
  ‣ GenBank (US)
  ‣ DNA Database of Japan (DDBJ)
  ‣ European Nucleotide Archive (ENA)

GenBank, ENA and DDBJ
Share and synchronize data

GenBank sequence record
The GenBank flat file format has defined fields including unique identifiers such as the accession number. This same general format is used for other sequence database records too.

Side node: Database accession numbers
Database accession numbers are strings of letters and numbers used as identifying labels for sequences and other data within databases
• Examples (all for retinol-binding protein, RBP4):
  X02775   GenBank genomic DNA sequence  DNA
  NT_030059  Genomic contig
  NR1759.1   An expressed sequence tag (1 of 170)  RNA
  NM_006744  RefSeq DNA sequence (from a transcript)
  NP_007635  RefSeq protein
  AAC02945  GenBank protein
  Q28369  UniProtKB/SwissProt protein
  1KT7  Protein Data Bank structure record
  PMID: 12205585  PubMed IDs identify articles at NCBI/NIH
  Literature
FASTA sequence files consist of records where each record begins with a ‘>’ and header information on that same line. Each subsequent line of the record is sequence information. This format is commonly used by sequence analysis programs.

The FEATURES section contains annotations including a conceptual translation of the nucleotide sequence.

The actual sequence entry starts after the word ORIGIN.
RefSeq: NCBI’s Derivative Sequence Database

- RefSeq entries are hand curated best representation of a transcript or protein (in their judgement)
- Non-redundant for a given species although alternate transcript forms will be included if there is good evidence
  - Experimentally verified transcripts and proteins accession numbers begin with "NM_" or "NP_"
  - Model transcripts and proteins based on bioinformatics predictions with little experimental support accession numbers begin with "XM_" or "XP_"
  - RefSeq also contains contigs and chromosome records

UniProt: THE PREMIER PROTEIN SEQUENCE DATABASE

UniProt: Protein sequence database

UniProt is a comprehensive, high-quality resource of protein sequence and functional information
- UniProt comprises four databases:
  1. UniProtKB (Knowledgebase)
     - Containing Swiss-Prot and TrEMBL components (these correspond to hand curated and automatically annotated entries respectively)
  2. UniRef (Reference Clusters)
     - Filtered version of UniProtKB at various levels of sequence identity
     - e.g. UniRef90 contains sequences with a maximum of 90% sequence identity to each other
  3. UniParc (Archive) with database cross-references to source.
  4. UniMES (Metagenomic and Environmental Sequences)

The two sides of UniProtKB

UniProtKB/Swiss-Prot
- Non-redundant, high-quality manual annotation - reviewed

UniProtKB/TrEMBL
- Redundant, automatically annotated - unreviewed

Indicators of which part of UniProt an entry belongs to include the color of the stars and the ID

The main information added to a UniProt/Swiss-Prot entry
UniProt/Swiss-Prot vs UniProt/TrEMBL

- **UniProtKB/Swiss-Prot** is a non-redundant database with one entry per protein
- **UniProtKB/TrEMBL** is a redundant database with one entry per translated ENA entry (ENA is the EBI’s equivalent of GenBank)
  - Therefore TrEMBL can contain multiple entries for the same protein
  - Multiple UniProtKB/TrEMBL entries for the same protein can arise due to:
    - Erroneous gene model predictions
    - Sequence errors (Frame shifts)
    - Polymorphisms
    - Alternative start sites
    - Isoforms
    - OR because the same sequence was submitted by different people

Side note: Automatic Annotation (sharing the wealth)

Swiss-Prot manually annotated

![Graph showing trend from 1997 to 2015](image)

Same domain composition = same function = annotation transfer

InterPro is an EBI database that collates protein domain signatures

We will talk more about sequence similarity and annotation transfer next week.

DATABASE VIGNETTE

You have just come out a seminar about gastric cancer and one of your co-workers asks:

“**What do you know about that ‘Kras’ gene the speaker kept taking about?”**

You have some recollection about hearing of ‘Ras’ before. How would you find out more?

- Google?
- Library?
- Bioinformatics databases at NCBI and EBI!


Hands on demo (or see following slides)
Example Questions:

What chromosome location and what genes are in the vicinity?
Example Questions: What ‘molecular functions’, ‘biological processes’, and ‘cellular component’ information is available?

GO: Gene Ontology
GO provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data

Why do we need Ontologies?
• Annotation is essential for capturing the understanding and knowledge associated with a sequence or other molecular entity
• Annotation is traditionally recorded as “free text”, which is easy to read by humans, but has a number of disadvantages, including:
  ‣ Difficult for computers to parse
  ‣ Quality varies from database to database
  ‣ Terminology used varies from annotator to annotator
• Ontologies are annotations using standard vocabularies that try to address these issues
• GO is integrated with UniProt and many other databases including a number at NCBI
The 'Gene Ontology' or GO is actually maintained by the EBI so let's switch or link over to UniProt also from the EBI.

UniProt will detail much more information for protein coding genes such as this one.

Example Questions:
- What positions in the protein are responsible for GTP binding?
- What variants of this enzyme are involved in gastric cancer and other human diseases?
- Are high resolution protein structures available to examine the details of these mutations?
Example Questions:
What is known about the protein family, its species distribution, number in humans and residue-wise conservation, etc… ?
ENTREZ & BLAST: TOOLS FOR SEARCHING AND ACCESSING MOLECULAR DATA AT NCBI

Entrez: integrated search of NCBI databases

Entrez was setup to allow you to navigate to related data in different databases without having to run additional searches.

Relies on pre-computed and pre-compiled data links:
- Neighbor knowledge based on calculations
- Hard links based on things we know about

Global Entrez Query: All NCBI Databases

Entrez is available from the main NCBI homepage or from the homepage of individual databases


The Entrez system: 38 (and counting) integrated databases
Search Results

Discovery Column (sort, filter, link)

Advanced: Search Builder

Helps build complex fielded queries

Items from search history can be included / combined / modified

Complex Query Results

"Danio rerio"[Organism] AND "creatinine kinase"[Title] AND refseq[Filter] AND mrna[Filter]

Controlled Vocabularies

- Taxonomy primary controlled vocabulary / classification system for molecular databases at NCBI

- Medical Subject Headings (MeSH) primary controlled vocabulary / classification system (ontology) for molecular databases at NCBI
BLAST is a very important tool available from the NCBI Homepage

BLAST – Basic Local Alignment Search Tool

BLAST performs sequence similarity searches of query sequences vs sequence databases. We will cover this in detail in the next lecture.

SUMMARY
- Bioinformatics is computer aided biology.
- Bioinformatics deals with the collection, archiving, organization, and interpretation of a wide range of biological data.
- There are a large number of primary, secondary and tertiary bioinformatics databases.
- The NCBI and EBI are major online bioinformatics service providers.
- Introduced GenBank, RefSeq, UniProt, PDB databases as well as a number of ‘boutique’ databases including PFAM and OMIM.
- Introduced the notion of controlled vocabularies and ontologies.
- Described the use of ENTREZ and BLAST for searching databases.

HOMEWORK
- Complete the initial course questionnaire: http://tinyurl.com/bioinf525-questions
- Check out the “Background Reading” material on Ctools: http://tinyurl.com/bioinf525-w16
- Complete the lecture 1.1 homework questions: http://tinyurl.com/bioinf525-quiz1

THANK YOU

NCBI Metadatabases
- Gene
  ▪ molecular data and literature related to genes
- HomoloGene
  ▪ automated collection of homologous genes from selected eukaryotes
- Taxonomy
  ▪ access to NCBI data through source organism taxonomic classification
- PubChem
  ▪ small organic molecules and their biological activities
- BioSystems
  ▪ biochemical pathways and processes linked to NCBI genes, gene products, small molecules, and structures
PubMed

- Curated database of biomedical journal articles
- Data records are annotated with MeSH terms (Medical Subject Headings)
- Contract workers actually read all of the articles and classify them with the MeSH terms
- PubMed entries contain article abstracts
- PubMed Central contains full journal articles, but the majority are not freely re-distributable

PubMed results

Limits and Advanced search can be used to refine searches

Small molecule databases have been added at NCBI

HomoloGene - Homologous genes from different organisms

Online Mendelian Inheritance in Man – OMIM

The NCBI Bookshelf includes many well known molecular biology texts.

OMIM is essentially a set of reviews of human genes, gene function and phenotypes. Includes causative mutations where known.
GEO: Gene Expression Omnibus

- Gene expression data (mostly from microarrays but also RNA-seq data, 2 methods for measuring RNA levels)

Series - (GSExxx) is an original submitter-supplied record that summarizes a study. May contain multiple individual Samples (GSMxxx).

DataSets - (GDSxxx) are curated collections of selected Samples that are biologically and statistically comparable

GO Ontologies

- There are three ontologies in GO:
  - Biological Process
    A commonly recognized series of events e.g. cell division, mitosis,
  - Molecular Function
    An elemental activity, task or job e.g. kinase activity, insulin binding
  - Cellular Component
    Where a gene product is located e.g. mitochondrion, mitochondrial membrane

QuickGO is a fast web-based browser of the Gene Ontology and Gene Ontology annotation data

GO annotation in UniProt

An example UniProt entry for hemoglobin beta (HBB_human, P68871) with GO annotation displayed.
GO annotation in UniProt
An example UniProt entry for hemoglobin beta (HBB_human, P68871) with GO annotation displayed.

DAVID: a online tool for assessing GO term enrichment in gene lists
DAVID allows you to upload lists of genes and search for enriched GO and search for functionally related genes not in your list
http://david.abcc.ncifcrf.gov

Example output: enriched functions from GO