Introduction to Bioinformatics

Monday, November 15, 2010
Jonathan Pevsner
pevsner@kennedykrieger.org
Bioinformatics
M.E.:800.707

What are the goals of the course?

- To provide an introduction to bioinformatics with a focus on the National Center for Biotechnology Information (NCBI), UCSC, and EBI
- To focus on the analysis of DNA, RNA and proteins
- To introduce you to the analysis of genomes
- To combine theory and practice to help you solve research problems

Who is taking this course?

- People with very diverse backgrounds in biology
- Some people with backgrounds in computer science and biostatistics
- Most people (will) have a favorite gene, protein, or disease

Textbook


I will make pdfs of the chapters available to everyone.

You can also purchase a copy at the bookstore, at amazon.com (now $60), or at Wiley with a 20% discount through the book’s website www.bioinfbook.org.

Web sites

The course website is reached via moodle:
http://pevsnerlab.kennedykrieger.org/moodle
(or Google “moodle bioinformatics”)
--This site contains the powerpoints for each lecture, including black & white versions for printing
--The weekly quizzes are here
--You can ask questions via the forum
--Audio files of each lecture will be posted here

The textbook website is:
http://www.bioinfbook.org
This has powerpoints, URLs, etc. organized by chapter. This is most useful to find “web documents” corresponding to each chapter.

Literature references

You are encouraged to read original source articles (posted on moodle). They will enhance your understanding of the material. Readings are optional but recommended.
Themes throughout the course: the beta globin gene/protein family

We will use beta globin as a model gene/protein throughout the course. Globins including hemoglobin and myoglobin carry oxygen. We will study globins in a variety of contexts including:

- sequence alignment
- gene expression
- protein structure
- phylogeny
- homologs in various species

Computer labs

There are no computer labs, but the seven weekly quizzes function as a computer lab. To solve the questions, you will need to go to websites, use databases, and use software.

Grading

60% moodle quizzes (your top 6 out of 7 quizzes). Quizzes are taken at the moodle website, and are due one week after the relevant lecture. Special extended due date for quizzes due immediately after Thanksgiving and the New Year.

40% final exam Monday, January 10 (in class). Closed book, cumulative, no computer, short answer / multiple choice. Past exams will be made available ahead of time.

Google "moodle bioinformatics" to get here;
Click "Bioinformatics" to sign in;
The enrollment key you need is...

The password to get the book chapter pdf is...
Outline for the course (all on Mondays)

1. Accessing information about DNA and proteins  Nov. 15
2. Pairwise alignment  Nov. 22
3. BLAST  Nov. 29
4. Multiple sequence alignment  Dec. 6
5. Molecular phylogeny and evolution  Dec. 13
6. Microarrays  Dec. 20
7. Genomes  Jan. 3
Final exam  Jan. 10

Outline for today

Definition of bioinformatics

Overview of the NCBI website
Accessing information: accession numbers and RefSeq
Entrez Gene (and UniGene, HomoloGene)
Protein Databases: UniProt, ExPASy
Three genome browsers: NCBI, UCSC, Ensembl
Access to biomedical literature

What is bioinformatics?

• Interface of biology and computers
• Analysis of proteins, genes and genomes using computer algorithms and computer databases
• Genomics is the analysis of genomes.
  The tools of bioinformatics are used to make sense of the billions of base pairs of DNA that are sequenced by genomics projects.

On bioinformatics

“Science is about building causal relations between natural phenomena (for instance, between a mutation in a gene and a disease). The development of instruments to increase our capacity to observe natural phenomena has, therefore, played a crucial role in the development of science - the microscope being the paradigmatic example in biology. With the human genome, the natural world takes an unprecedented turn: it is better described as a sequence of symbols. Besides high-throughput machines such as sequencers and DNA chip readers, the computer and the associated software becomes the instrument to observe it, and the discipline of bioinformatics flourishes.”

Martin Reese and Roderic Guigó, Genome Biology 2006 7(Suppl I):S1, introducing EGASP, the Encyclopedia of DNA Elements (ENCODE) Genome Annotation Assessment Project
Three perspectives on bioinformatics

The cell

The organism

The tree of life

Page 4

Page 5

After Pace NR (1997) Science 276:734
**Growth of GenBank**

![Growth of GenBank chart](image)

**Growth of GenBank + Whole Genome Shotgun (1982-November 2008): we reached 0.2 terabases**

![Growth of GenBank + WGS chart](image)

**Arrival of next-generation sequencing:**

In two years we have gone from 0.2 terabases to 71 terabases (71,000 gigabases) (November 2010)

![Arrival of next-generation sequencing chart](image)

**Central dogma of molecular biology**

![Central dogma of molecular biology diagram](image)

**Central dogma of bioinformatics and genomics**

![Central dogma of bioinformatics and genomics diagram](image)

**There are three major public DNA databases**

![DNA databases diagram](image)

EMBL ↔ GenBank ↔ DDBJ

The underlying raw DNA sequences are identical

![Underlying raw DNA sequences are identical diagram](image)
There are three major public DNA databases

**EMBL**
Housed at EBI
European Bioinformatics Institute

**GenBank**
Housed at NCBI
National Center for Biotechnology Information

**DDBJ**
Housed in Japan

---

The most sequenced organisms in GenBank

<table>
<thead>
<tr>
<th>Organism</th>
<th>Total Bases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens</td>
<td>14.9 billion bases</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>8.9b</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>6.5b</td>
</tr>
<tr>
<td>Bos taurus</td>
<td>5.4b</td>
</tr>
<tr>
<td>Zea mays</td>
<td>5.0b</td>
</tr>
<tr>
<td>Sus scrofa</td>
<td>4.8b</td>
</tr>
<tr>
<td>Danio rerio</td>
<td>3.1b</td>
</tr>
<tr>
<td>Strongylocentrotus purpurata</td>
<td>1.4b</td>
</tr>
<tr>
<td>Oryza sativa (japonica)</td>
<td>1.2b</td>
</tr>
<tr>
<td>Nicotiana tabacum</td>
<td>1.2b</td>
</tr>
</tbody>
</table>

Updated Oct. 2010
GenBank release 180.0
Excluding WGS, organelles, metagenomics

---

NCBI homepage

---

National Center for Biotechnology Information (NCBI)

NCBI key features: PubMed

• National Library of Medicine's search service
• 20 million citations in MEDLINE (as of 2010)
• links to participating online journals
• PubMed tutorial on the site or visit NLM: http://www.nlm.nih.gov/bsd/disted/pubmed.html

NCBI key features: Entrez search and retrieval system

Entrez integrates...

• the scientific literature;
• DNA and protein sequence databases;
• 3D protein structure data;
• population study data sets;
• assemblies of complete genomes

NCBI key features: BLAST

BLAST is...

• Basic Local Alignment Search Tool
• NCBI's sequence similarity search tool
• supports analysis of DNA and protein databases
• 100,000 searches per day

NCBI key features: OMIM

OMIM is...

• Online Mendelian Inheritance in Man
• catalog of human genes and genetic disorders
• created by Dr. Victor McKusick; led by Dr. Ada Hamosh at JHMI

NCBI key features: TaxBrowser

TaxBrowser is...

• browser for the major divisions of living organisms (archaea, bacteria, eukaryota, viruses)
• taxonomy information such as genetic codes
• molecular data on extinct organisms
• practically useful to find a protein or gene from a species
Structure site includes...

- Molecular Modelling Database (MMDB)
- biopolymer structures obtained from the Protein Data Bank (PDB)
- Cn3D (a 3D-structure viewer)
- vector alignment search tool (VAST)

Accession numbers are labels for sequences

NCBI includes databases (such as GenBank) that contain information on DNA, RNA, or protein sequences. You may want to acquire information beginning with a query such as the name of a protein of interest, or the raw nucleotides comprising a DNA sequence of interest.

DNA sequences and other molecular data are tagged with accession numbers that are used to identify a sequence or other record relevant to molecular data.

What is an accession number?

An accession number is a label that is used to identify a sequence. It is a string of letters and/or numbers that corresponds to a molecular sequence.

Examples (all for retinol-binding protein, RBP4):

- X02775: GenBank genomic DNA sequence
- Rs706946: dtSINP (single nucleotide polymorphism)
- NM_006744: RefSeq DNA sequence (from a transcript)
- NP_006744: RefSeq protein
- Q28369: SwissProt protein
- 1KT7: Protein Data Bank structure record

NCBI’s important RefSeq project: best representative sequences

RefSeq (accessible via the main page of NCBI) provides an expertly curated accession number that corresponds to the most stable, agreed-upon “reference” version of a sequence.

RefSeq identifiers include the following formats:

- Complete genome: NC_#####
- Complete chromosome: NC_#####
- Genomic contig: NT_#####
- mRNA (DNA format): NM_##### e.g. NM_006744
- Protein: NP_##### e.g. NP_006735

NCBI’s RefSeq project: many accession number formats for genomic, mRNA, protein sequences

<table>
<thead>
<tr>
<th>Accession</th>
<th>Molecule</th>
<th>Method</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC_123456</td>
<td>Genomic</td>
<td>Mixed</td>
<td>Alternate complete genomic</td>
</tr>
<tr>
<td>AP_123456</td>
<td>Protein</td>
<td>Mixed</td>
<td>Protein products; alternate</td>
</tr>
<tr>
<td>NC_123456</td>
<td>Genomic</td>
<td>Mixed</td>
<td>Complete genomic molecules</td>
</tr>
<tr>
<td>NG_123456</td>
<td>Genomic</td>
<td>Mixed</td>
<td>Incomplete genomic regions</td>
</tr>
<tr>
<td>NM_123456</td>
<td>mRNA</td>
<td>Mixed</td>
<td>Transcript products; mRNA</td>
</tr>
<tr>
<td>NM_123456789</td>
<td>mRNA</td>
<td>Mixed</td>
<td>Transcript products; 9-digit</td>
</tr>
<tr>
<td>NP_123456</td>
<td>Protein</td>
<td>Mixed</td>
<td>Protein products;</td>
</tr>
<tr>
<td>NP_123456789</td>
<td>Protein</td>
<td>Curation</td>
<td>Protein products; 9-digit</td>
</tr>
<tr>
<td>NR_123456</td>
<td>RNA</td>
<td>Mixed</td>
<td>Non-coding transcripts</td>
</tr>
<tr>
<td>NT_123456</td>
<td>Genomic</td>
<td>Automated</td>
<td>Genomic assemblies</td>
</tr>
<tr>
<td>NW_123456</td>
<td>Genomic</td>
<td>Automated</td>
<td>Genomic assemblies</td>
</tr>
<tr>
<td>NZ_ABCD12345678</td>
<td>Genomic</td>
<td>Automated</td>
<td>Whole genome shotgun data</td>
</tr>
<tr>
<td>XM_123456</td>
<td>mRNA</td>
<td>Automated</td>
<td>Transcript products</td>
</tr>
<tr>
<td>XP_123456</td>
<td>Protein</td>
<td>Automated</td>
<td>Protein products</td>
</tr>
<tr>
<td>XR_123456</td>
<td>RNA</td>
<td>Automated</td>
<td>Transcript products</td>
</tr>
<tr>
<td>YP_123456</td>
<td>Protein</td>
<td>Auto. &amp; Curated</td>
<td>Protein products</td>
</tr>
<tr>
<td>ZP_12345678</td>
<td>Protein</td>
<td>Automated</td>
<td>Protein products</td>
</tr>
</tbody>
</table>
### Outline for today

- Definition of bioinformatics
- Overview of the NCBI website
- Accessing information: accession numbers and RefSeq
  - Entrez Gene (and UniGene, HomoloGene)
- Protein Databases: UniProt, ExPASy
- Three genome browsers: NCBI, UCSC, Ensembl
- Access to biomedical literature

### Access to sequences: Entrez Gene at NCBI

Entrez Gene is a great starting point: it collects key information on each gene/protein from major databases. It covers all major organisms.

RefSeq provides a curated, optimal accession number for each DNA (NM_000518 for beta globin DNA corresponding to mRNA) or protein (NP_000509)

---

![From the NCBI home page, type “beta globin” and hit “Search”](image1)

![Follow the link to “Gene”](image2)

![Entrez Gene is in the header](image3)

![Using “limits” you can restrict your search to human (or any other organism)](image4)
By applying limits, there are now far fewer entries

Entrez Gene (top of page)

Note that links to many other HBB database entries are available

Entrez Gene (middle of page): genomic region, bibliography

Entrez Gene (middle of page, continued): phenotypes, function

Entrez Gene (bottom of page): RefSeq accession numbers

Entrez Gene (bottom of page): non-RefSeq accessions (it’s unclear what these are, highlighting usefulness of RefSeq)
You should learn the one-letter amino acid code!

<table>
<thead>
<tr>
<th></th>
<th>3-Letter</th>
<th>1-Letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>Ala</td>
<td>A</td>
</tr>
<tr>
<td>Asparagine</td>
<td>Asn</td>
<td>N</td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>Asp</td>
<td>D</td>
</tr>
<tr>
<td>Cysteine</td>
<td>Cys</td>
<td>C</td>
</tr>
<tr>
<td>Glutamic Acid</td>
<td>Gnu</td>
<td>Q</td>
</tr>
<tr>
<td>Glycine</td>
<td>Gly</td>
<td>G</td>
</tr>
<tr>
<td>Histidine</td>
<td>His</td>
<td>H</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Ile</td>
<td>I</td>
</tr>
<tr>
<td>Leucine</td>
<td>Leu</td>
<td>L</td>
</tr>
<tr>
<td>Lysine</td>
<td>Lys</td>
<td>K</td>
</tr>
<tr>
<td>Methionine</td>
<td>Met</td>
<td>M</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>Phe</td>
<td>F</td>
</tr>
<tr>
<td>Proline</td>
<td>Pro</td>
<td>P</td>
</tr>
<tr>
<td>Serine</td>
<td>Ser</td>
<td>S</td>
</tr>
<tr>
<td>Threonine</td>
<td>Thr</td>
<td>T</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>Trp</td>
<td>W</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Tyr</td>
<td>Y</td>
</tr>
<tr>
<td>Valine</td>
<td>Val</td>
<td>V</td>
</tr>
</tbody>
</table>

FASTA format:
versatile, compact with one header line
followed by a string of nucleotides or amino acids
in the single letter code
Comparison of Entrez Gene to other resources

Entrez Gene, Entrez Nucleotide, Entrez Protein: closely inter-related

Entrez Gene versus UniGene:
UniGene is a database with information on where in a body, when in development, and how abundantly a transcript is expressed

Entrez Gene versus HomoloGene:
HomoloGene conveniently gathers information on sets of related proteins

HomoloGene: an NCBI resource organized by organism to describe where genes are expressed (i.e. from which library) and how abundantly

Outline for today

Definition of bioinformatics
Overview of the NCBI website
Accessing information: accession numbers and RefSeq
Entrez Gene (and UniGene, HomoloGene)

Protein Databases: UniProt, ExPASy

Three genome browsers: NCBI, UCSC, Ensembl

Access to biomedical literature

ExPASy to access protein and DNA sequences

ExPASy sequence retrieval system (ExPASy = Expert Protein Analysis System)
Visit http://www.expasy.ch/

UniProt: a centralized protein database (uniprot.org)
This is separate from NCBI, and interlinked.
ExPASy: vast proteomics resources (www.expasy.ch)

Outline for today

Definition of bioinformatics
Overview of the NCBI website
Accessing information: accession numbers and RefSeq
Entrez Gene (and UniGene, HomoloGene)
Protein Databases: UniProt, ExPASy

Three genome browsers: NCBI, UCSC, Ensembl

Access to biomedical literature

Genome Browsers: increasingly important resources

Genomic DNA is organized in chromosomes. Genome browsers display ideograms (pictures) of chromosomes, with user-selected "annotation tracks" that display many kinds of information.

The two most essential human genome browsers are at Ensembl and UCSC. We will focus on UCSC (but the two are equally important). The browser at NCBI is not commonly used.

Ensembl genome browser (www.ensembl.org)

Ensembl output for beta globin includes views of chromosome 11 (top), the region (middle), and a detailed view (bottom).

There are various horizontal annotation tracks.
The UCSC Genome Browser: an increasingly important resource

- This browser’s focus is on humans and other eukaryotes
- you can select which tracks to display (and how much information for each track)
- tracks are based on data generated by the UCSC team and by the broad research community
- you can create “custom tracks” of your own data! Just format a spreadsheet properly and upload it
- The Table Browser is equally important as the more visual Genome Browser, and you can move between the two

[1] Visit http://genome.ucsc.edu/, click Genome Browser

[2] Choose organisms, enter query (beta globin), hit submit

[3] Choose the RefSeq beta globin gene

Example of how to access sequence data: HIV-1 pol

There are many possible approaches. Begin at the main page of NCBI, and type an Entrez query: hiv-1 pol

[4] On the UCSC Genome Browser:
- choose which tracks to display
- add custom tracks
- the Table Browser is complementary

Searching for HIV-1 pol: 150,000 nucleotide, protein hits
Example of how to access sequence data: HIV-1 pol

For the Entrez query: hiv-1 pol
there are about 150,000 nucleotide or protein records
(and >350,000 records for a search for “hiv-1”),
but these can easily be reduced in two easy steps:

--specify the organism, e.g. hiv-1[organism]
--limit the output to RefSeq!

Searching for HIV-1 pol:
using the command hiv-1[organism] limits the
output to just one entry

Example of how to access sequence data: histone

query for “histone”  # results
protein records 104,000
RefSeq entries 39,000
RefSeq (limit to human) 1171
NOT deacetylase 911

At this point, select a reasonable candidate (e.g.
histone 2, H4) and follow its link to Entrez Gene.
There, you can confirm you have the right protein.

Outline for today

Definition of bioinformatics
Overview of the NCBI website
Accessing information: accession numbers and RefSeq
Entrez Gene (and UniGene, HomoloGene)
Protein Databases: UniProt, ExPASy
Three genome browsers: NCBI, UCSC, Ensembl

Access to biomedical literature
PubMed is the NCBI gateway to MEDLINE. MEDLINE contains bibliographic citations and author abstracts from over 4,600 journals published in the United States and in 70 foreign countries. It has >20 million records dating back to 1950s.

MeSH is the acronym for "Medical Subject Headings." MeSH is the list of the vocabulary terms used for subject analysis of biomedical literature at NLM. MeSH vocabulary is used for indexing journal articles for MEDLINE. The MeSH controlled vocabulary imposes uniformity and consistency to the indexing of biomedical literature.

Use the pull-down menu to access related resources such as Medical Subject Headings (MeSH).

A "how to" pull-down menu links to tutorials.
Use “Advanced search” to limit by author, year, language, etc.

PubMed search strategies

Try the tutorial
Use boolean queries (capitalize AND, OR, NOT)
lipocalin AND disease

Try using limits (see Advanced search)
There are links to find Entrez entries and external resources

Obtain articles on-line via Welch Medical Library
(and download pdf files): http://www.welch.jhu.edu/

1 AND 2
lipocalin AND disease
(584 results)

1 OR 2
lipocalin OR disease
(2,500,000 results)

1 NOT 2
lipocalin NOT disease
(2,370 results)

WelchWeb is available at http://www.welch.jhu.edu

Reminder: Please enroll! Google “moodle bioinformatics” to get here; click “Bioinformatics” to sign in; The enrollment key is…