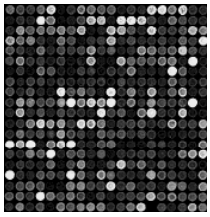
Introduction to Bioinformatics



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

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

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

Textbook

The course textbook has no required textbook. I wrote *Bioinformatics and Functional Genomics* (Wiley-Blackwell, 2nd edition 2009). The lectures in this course correspond closely to chapters.

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

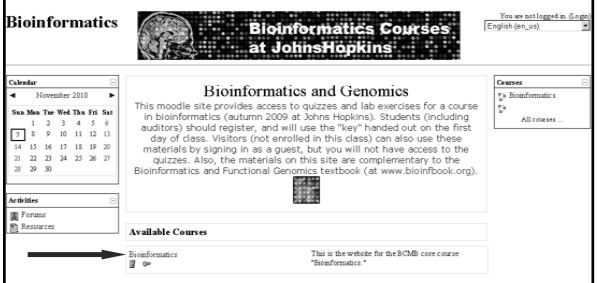
There are no computer labs, but the seven weekly quizzes function as a computer lab. To solve the questions, you will need to go to websites, use databases, and use software.

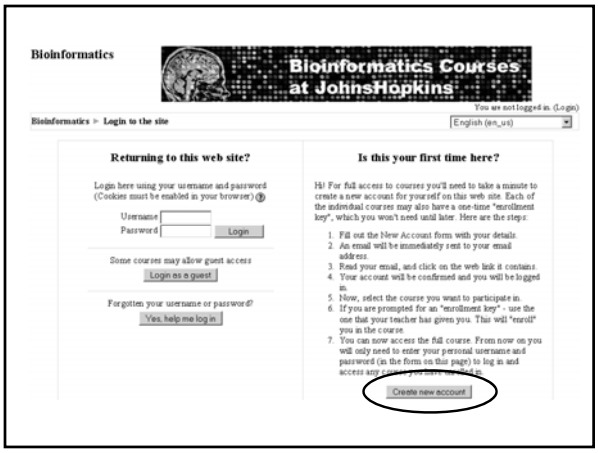
Grading

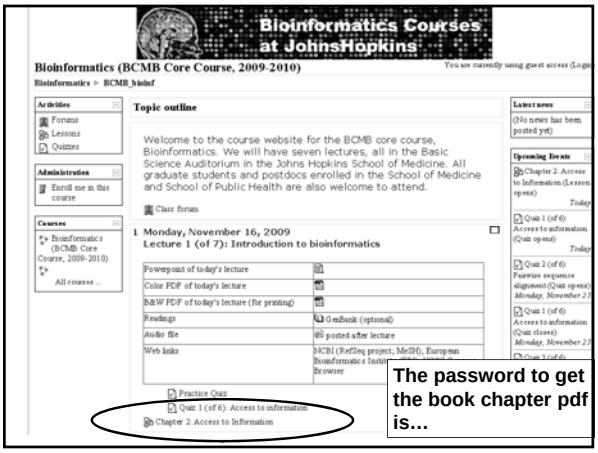
60% moodle quizzes (your top 6 out of 7 quizzes).
Quizzes are taken at the moodle website, and are due one week after the relevant lecture. Special extended due date for quizzes due immediately after Thanksgiving and the New Year.

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…







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.

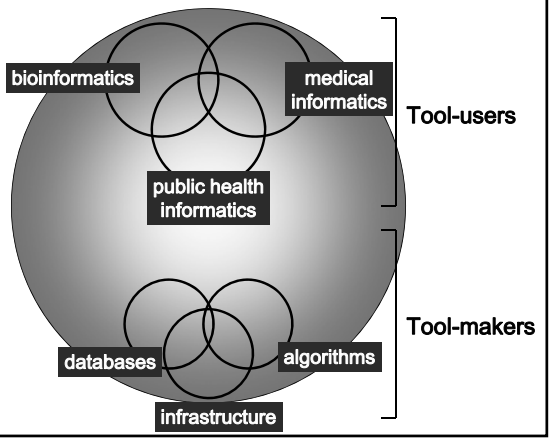
On bioinformatics

"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."

On bioinformatics

"However, as the separation between us (the observers) and the phenomena observed increases (from organism to cell to genome, for instance), instruments may capture phenomena only indirectly, through the footprints they leave. Instruments therefore need to be calibrated: the distance between the reality and the observation (through the instrument) needs to be accounted for. This issue of *Genome Biology* is about calibrating instruments to observe gene sequences; more specifically, computer programs to identify human genes in the sequence of the human genome."

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



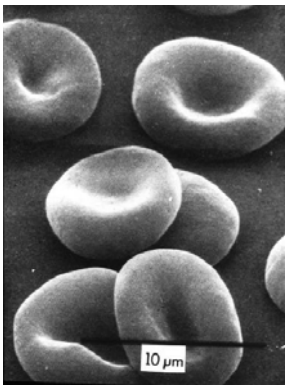
Three perspectives on bioinformatics

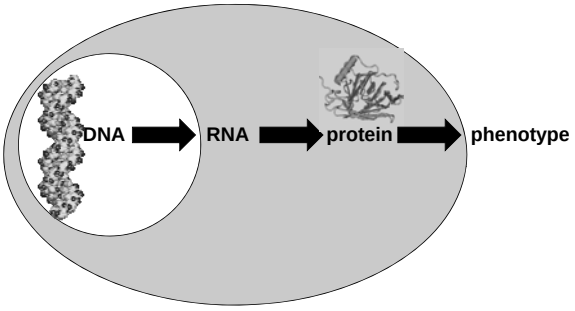
The cell

The organism

The tree of life

Page 4



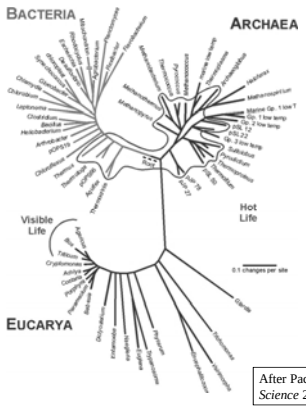


Page 5

Time of development

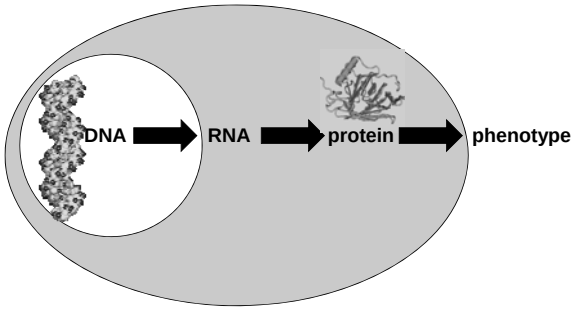
Body region, physiology, pharmacology, pathology

Page 5



After Pace NR (1997) *Science* 276:734

Page 6



Growth of GenBank

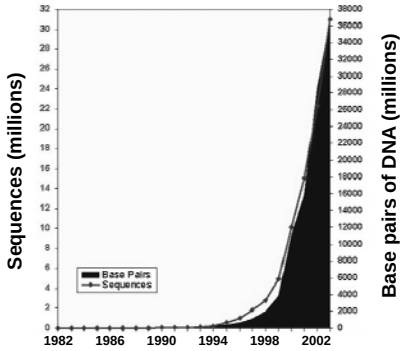


Fig. 2.1
Page 17

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

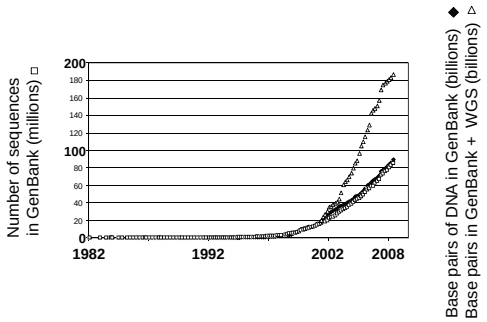


Fig. 2.1
Page 15

Arrival of next-generation sequencing: In two years we have gone from 0.2 terabases to 71 terabases (71,000 gigabases) (November 2010)

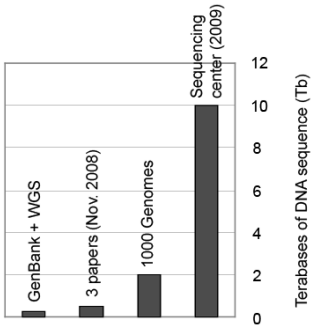
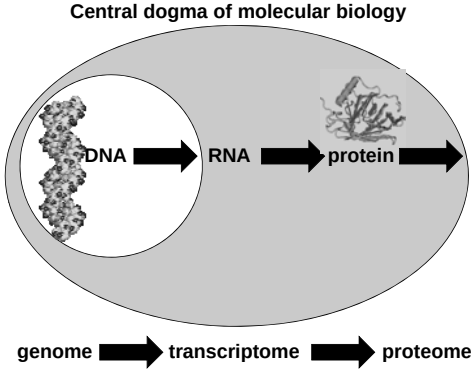


Fig. 2.1
Page 15

Central dogma of molecular biology



genome → transcriptome → proteome

Central dogma of bioinformatics and genomics

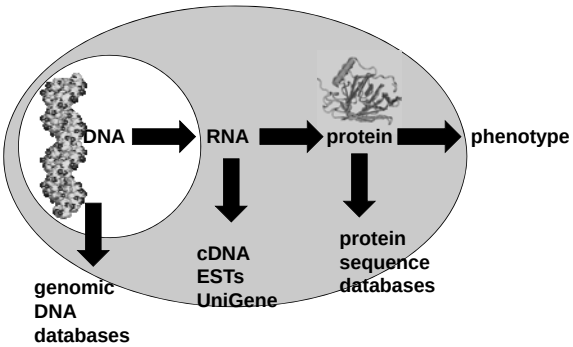


Fig. 2.2
Page 18

There are three major public DNA databases

EMBL ⟷ GenBank ⟷ DDBJ

The underlying raw DNA sequences are identical

Page 14

There are three major public DNA databases

EMBL ⟷ **GenBank** ⟷ **DDBJ**

Housed at EBI European Bioinformatics Institute

Housed at NCBI National Center for Biotechnology Information

Housed in Japan

Page 14

Taxonomy at NCBI:
>200,000 species are represented in GenBank

Ranks:	higher taxa	genus	species	lower taxa	total
Archaea	103	116	553	162	914
Bacteria	1126	2041	16236	8938	28341
Eukaryota	17050	52186	204884	17318	291435
Fungi	1211	3675	21980	1486	28352
Metazoa	12540	32954	92188	8278	145960
Viridiplantae	2012	13339	83523	6466	105340
Viruses	506	344	7020	56107	63977
All taxa	18810	54695	234123	82560	390188

10/10

Page 16
http://www.ncbi.nlm.nih.gov/Taxonomy/txstat.cgi

The most sequenced organisms in GenBank

Homo sapiens	14.9 billion bases
Mus musculus	8.9b
Rattus norvegicus	6.5b
Bos taurus	5.4b
Zea mays	5.0b
Sus scrofa	4.8b
Danio rerio	3.1b
Strongylocentrotus purpurata	1.4b
Oryza sativa (japonica)	1.2b
Nicotiana tabacum	1.2b

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

Table 2-2
Page 17

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

National Center for Biotechnology Information (NCBI)

www.ncbi.nlm.nih.gov

Page 23

NCBI homepage

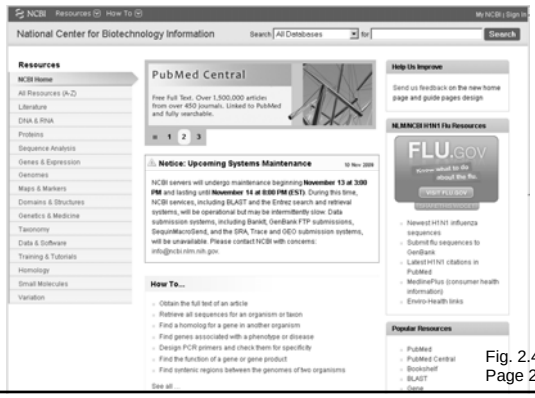


Fig. 2.4
Page 24

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

Page 23

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

Page 24

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

Page 25

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

Page 25

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

Page 26

NCBI key features: Structure

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)

Page 25

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

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.

Page 26

What is an accession number?

An accession number is label that 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	**DNA**
NT_030059	Genomic contig	
Rs7079946	dbSNP (single nucleotide polymorphism)	
N91759.1	An expressed sequence tag (1 of 170)	**RNA**
NM_006744	RefSeq DNA sequence (from a transcript)	
NP_007635	RefSeq protein	**protein**
AAC02945	GenBank protein	
Q28369	SwissProt protein	
1KT7	Protein Data Bank structure record	

Page 27

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

Page 27

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

Accession	Molecule	Method	Note
AC_123456	Genomic	Mixed	Alternate complete genomic
AP_123456	Protein	Mixed	Protein products; alternate
NC_123456	Genomic	Mixed	Complete genomic molecules
NG_123456	Genomic	Mixed	Incomplete genomic regions
NM_123456	mRNA	Mixed	Transcript products; mRNA
NM_123456789	mRNA	Mixed	Transcript products; 9-digit
NP_123456	Protein	Mixed	Protein products;
NP_123456789	Protein	Curation	Protein products; 9-digit
NR_123456	RNA	Mixed	Non-coding transcripts
NT_123456	Genomic	Automated	Genomic assemblies
NW_123456	Genomic	Automated	Genomic assemblies
NZ_ABCD12345678	Genomic	Automated	Whole genome shotgun data
XM_123456	mRNA	Automated	Transcript products
XP_123456	Protein	Automated	Protein products
XR_123456	RNA	Automated	Transcript products
YP_123456	Protein	Auto. & Curated	Protein products
ZP_12345678	Protein	Automated	Protein products

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"

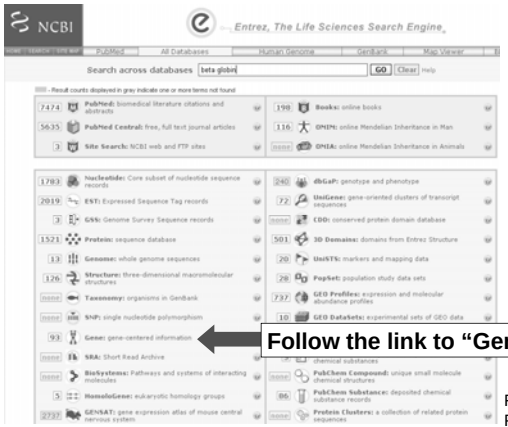
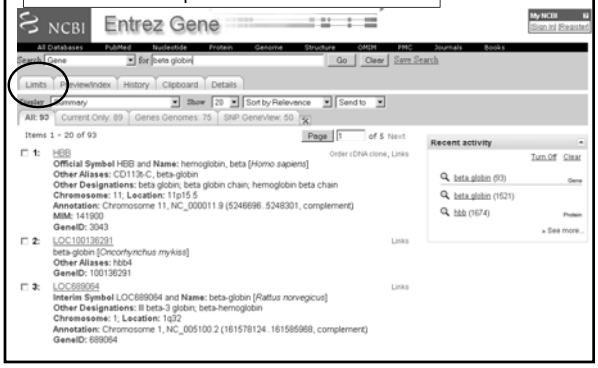
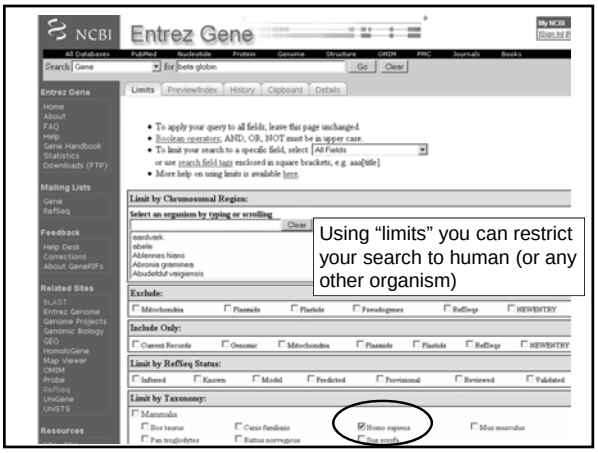


Fig. 2.5
Page 28

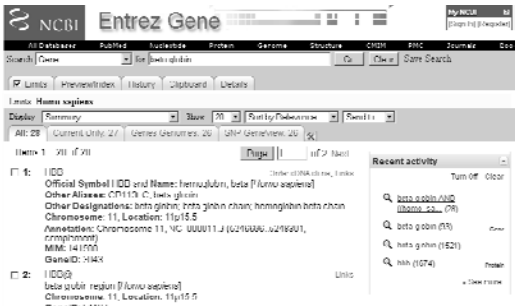
Entrez Gene is in the header
Note the "Official Symbol" HBB for beta globin
Note the "limits" option





Using "limits" you can restrict your search to human (or any other organism)

By applying limits, there are now far fewer entries

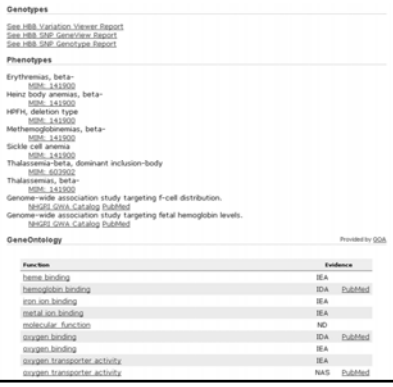


Entrez Gene (top of page)

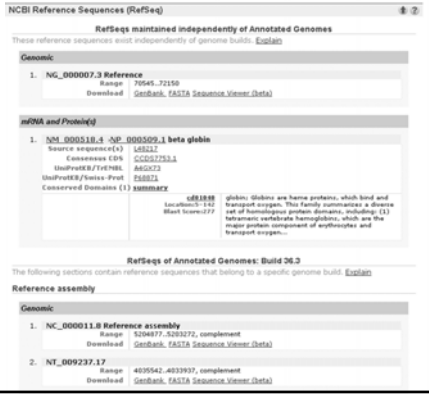
Note that links to many other HBB database entries are available

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

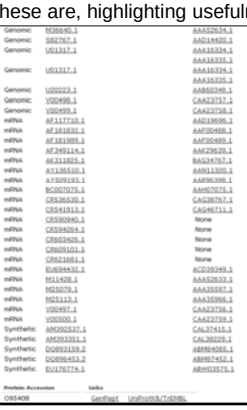
Entrez Gene (middle of page, continued): phenotypes, function

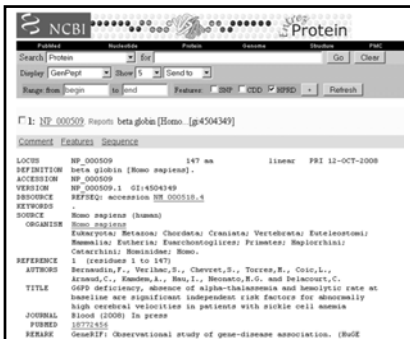


Entrez Gene (bottom of page): RefSeq accession numbers



Entrez Gene (bottom of page): non-RefSeq accessions
(it's unclear what these are, highlighting usefulness of RefSeq)





Entrez Protein: accession, organism, literature…

Fig. 2.8
Page 31

Entrez Protein:
…features of a protein, and its sequence in the one-letter amino acid code

```
Site            94
                /site_type="modified"
                /experiment="experimental evidence, no additional details
                recorded"
                /note="S-nitrosylation site"
                /citation=[7]
Site            121
                /site_type="glycosylation"
                /experiment="experimental evidence, no additional details
                recorded"
                /note="glycation site"
                /citation=[9]
CDS             1..147
                /gene="HBB"
                /gene_synonym="CD113t-C"
                /coded_by="NM_000518.4:51..494"
                /db_xref="CCDS:CCDS7753.1"
                /db_xref="GeneID:3043"
                /db_xref="HGNC:4827"
                /db_xref="HPRD:00786"
                /db_xref="MIM:141900"
ORIGIN
        1 mvhltpeeks avtalwgkvn vdevggealg rllvvypwtq rffesfgdls tpdavmgnpk
       61 vkahgkkvlg afsdglahld nlkgtfatls elhcdklhvd penfrllgnv ivcvlahhfg
      121 keftppvqaa yqkvvagvan alahkyh
//
```

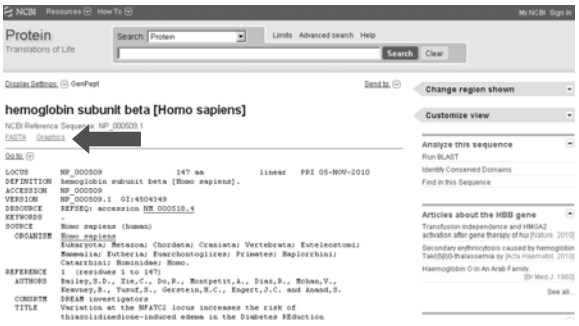
Fig. 2.8
Page 31

You should learn the one-letter amino acid code!

Name	3-Letter	1-Letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic Acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I

Name	3-Letter	1-Letter
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Entrez Protein:
You can change the display (as shown)…



Page 31

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

Fig. 2.9
Page 32

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

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

Page 32

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

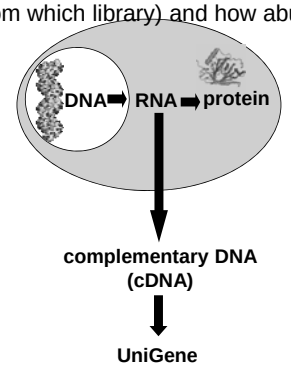
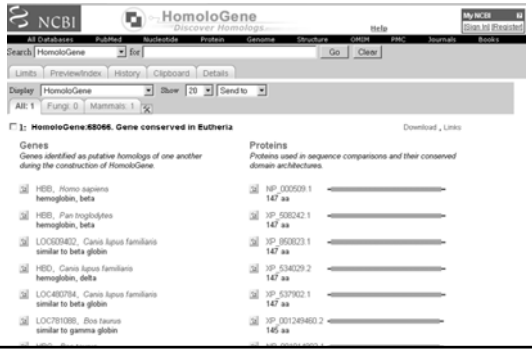


Fig. 2.3
Page 22

HomologoGene: an excellent NCBI resource that conveniently groups homologous eukaryotic genes (find links from Entrez search engine or Entrez gene)



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/

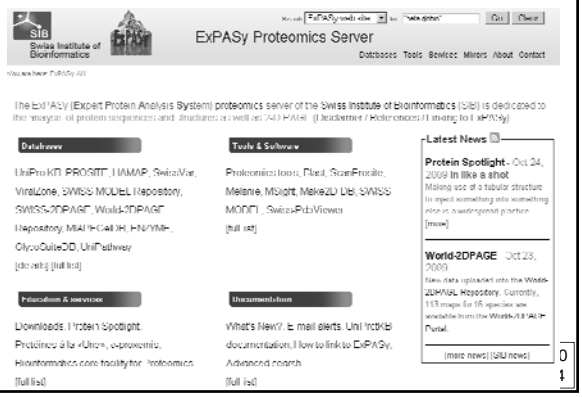
Page 33

UniProt:
a centralized protein database (uniprot.org)

This is separate from NCBI, and interlinked.

Page 33

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

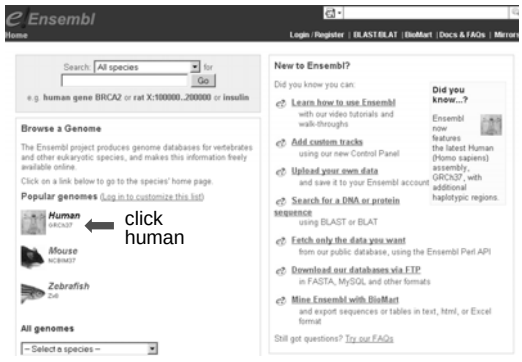
Access to biomedical literature

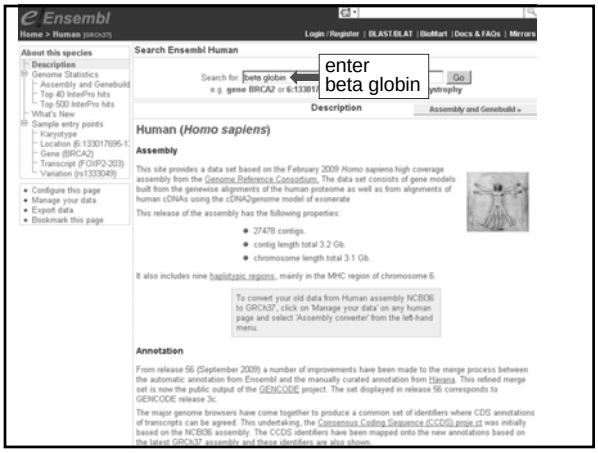
Genome Browsers: increasingly important resources

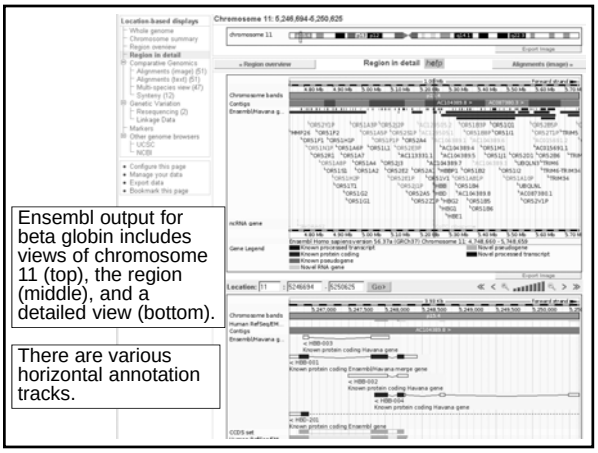
Genomic DNA is organized in chromosomes. Genome browsers display ideograms (pictures) of chromosomes, with user-selected "annotation tracks" that display many kinds of information.

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)



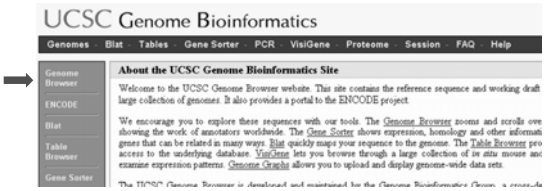




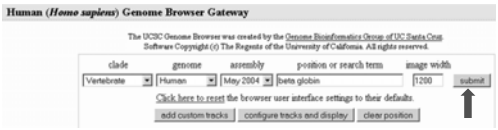
The UCSC Genome Browser: an increasingly important resource

- 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

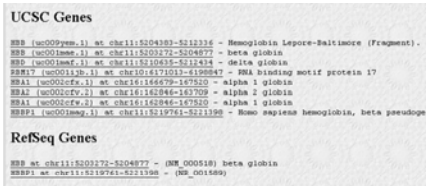


[2] Choose organisms, enter query (beta globin), hit submit

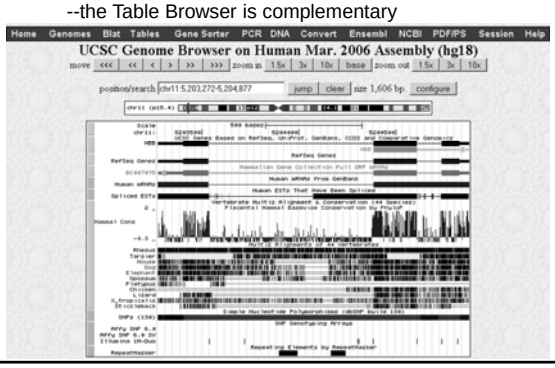


Page 36

[3] Choose the RefSeq beta globin gene



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

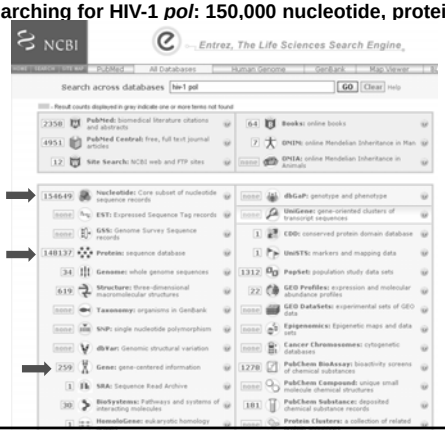


Example of how to access sequence data: HIV-1 *pol*

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

Page 36

Searching for HIV-1 *pol*: 150,000 nucleotide, protein hits



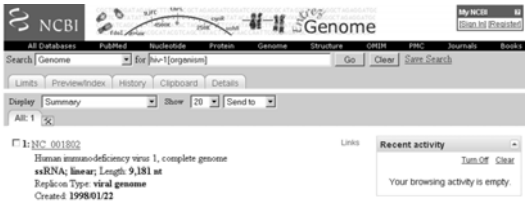
11/10

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:

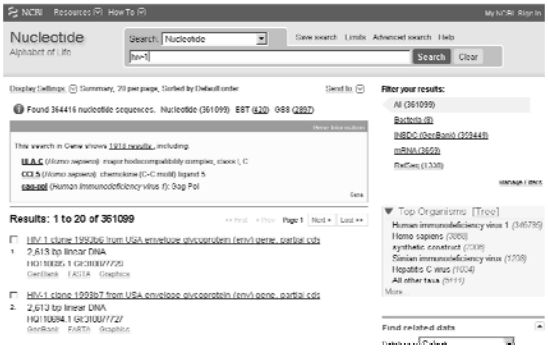
--specify the organism, e.g. hiv-1[organism]
--limit the output to RefSeq!

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



Try Taxonomy Browser to easily limit your query to your favorite organism(s). *Example*:
NCBI home → Taxonomy → Taxonomy browser → human → protein to find a human protein

Entrez Nucleotide features over 360,000 nucleotide entries for HIV-1 (but only one RefSeq for that virus)

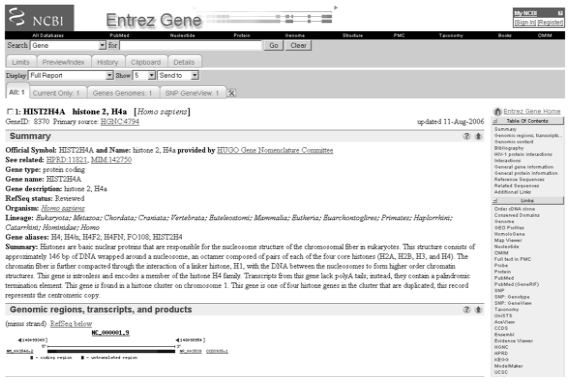


Example of how to access sequence data: histone

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.

Entrez Gene result for a histone



Outline for today

Definition of bioinformatics

Overview of the NCBI website

Accessing information: accession numbers and RefSeq

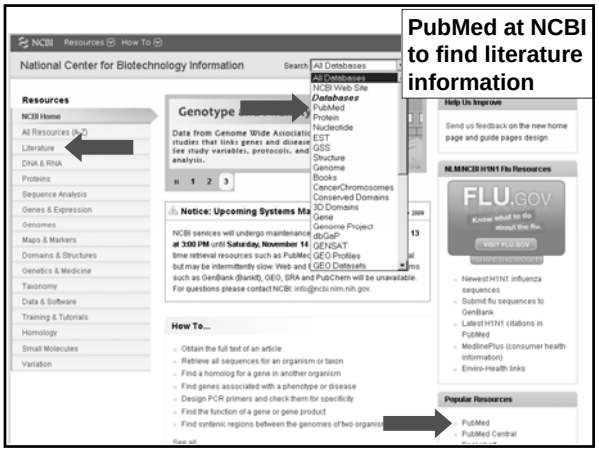
Entrez Gene (and UniGene, HomoloGene)

Protein Databases: UniProt, ExPASy

Three genome browsers: NCBI, UCSC, Ensembl

Access to biomedical literature

PubMed at NCBI to find literature information



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.

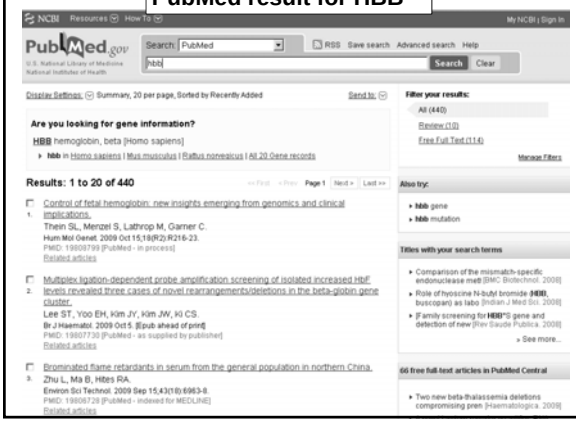
It has >20 million records dating back to 1950s.

MeSH is the acronym for "Medical Subject Headings."

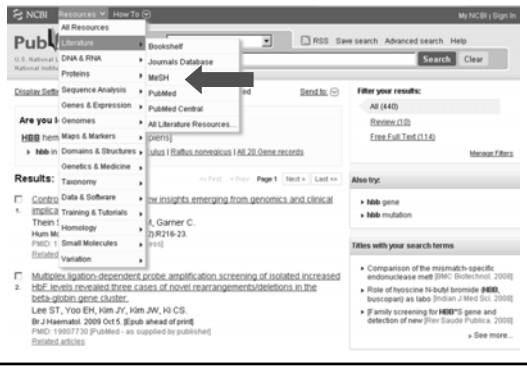
MeSH is the list of the vocabulary terms used for subject analysis of biomedical literature at NLM. MeSH vocabulary is used for indexing journal articles for MEDLINE.

The MeSH controlled vocabulary imposes uniformity and consistency to the indexing of biomedical literature.

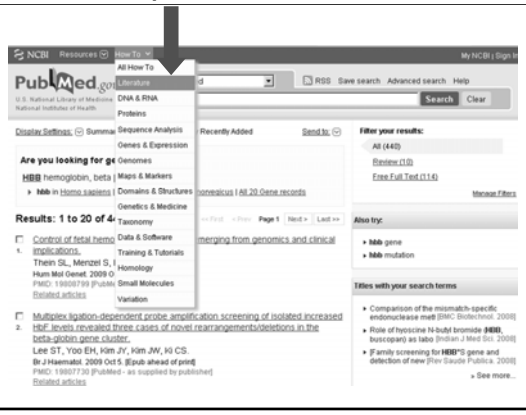
PubMed result for HBB



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

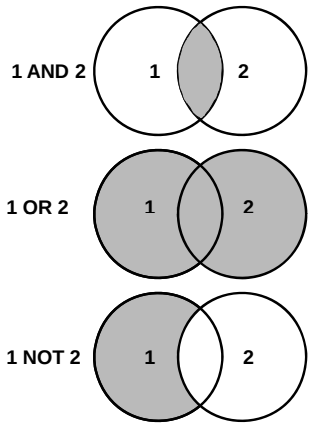
Try the tutorial

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/



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

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

Welch Medical Library liasons to the basic sciences

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