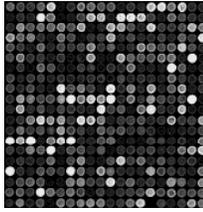


## Introduction to Bioinformatics



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, 2<sup>nd</sup> 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](http://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.

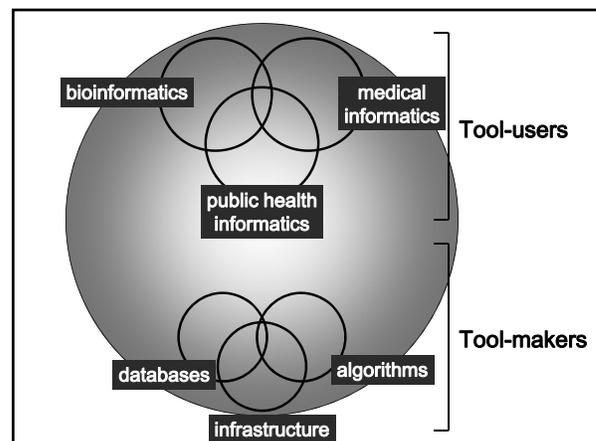
## On bioinformatics

"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 1):S1, introducing EGASP, the Encyclopedia of DNA Elements (ENCODE) Genome Annotation Assessment Project



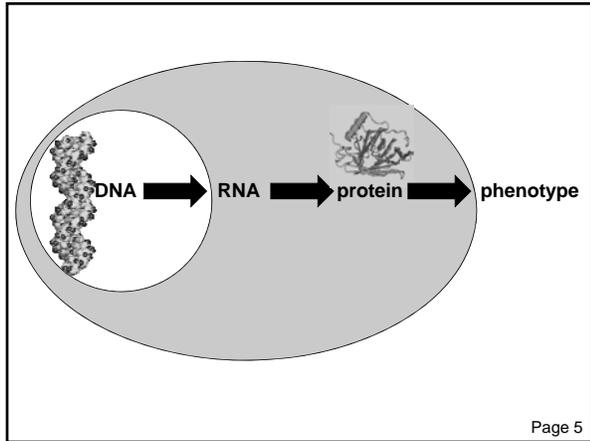
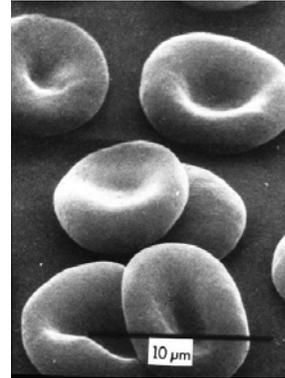
### Three perspectives on bioinformatics

The cell

The organism

The tree of life

Page 4



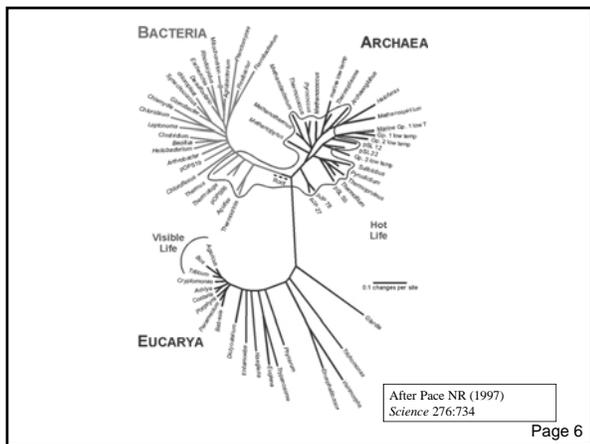
Page 5



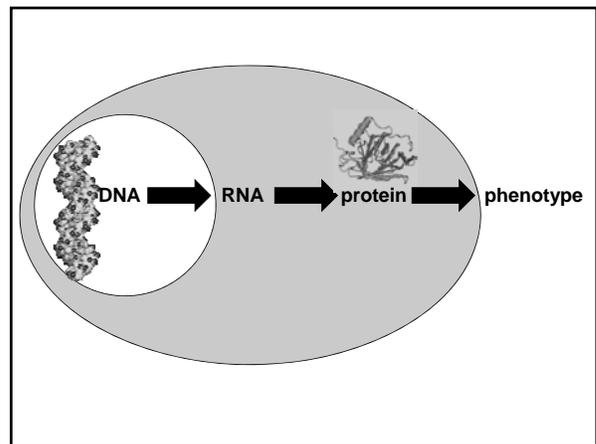
Time of development

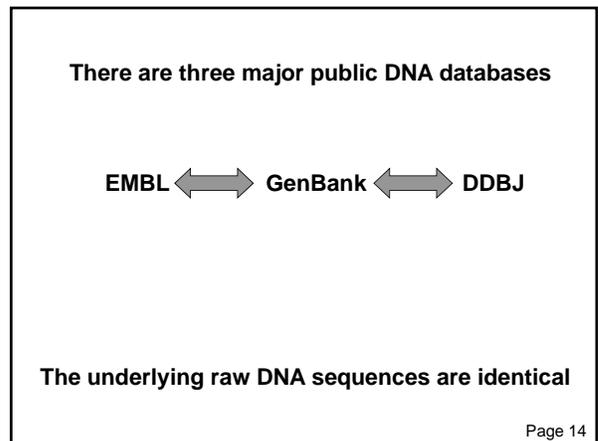
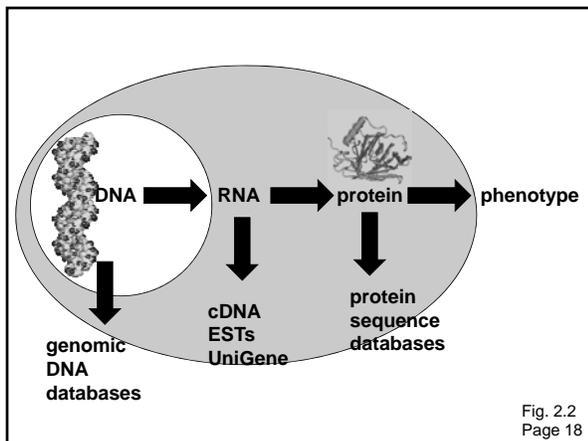
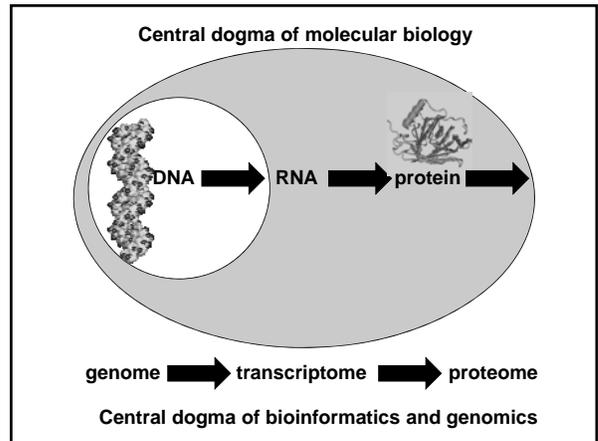
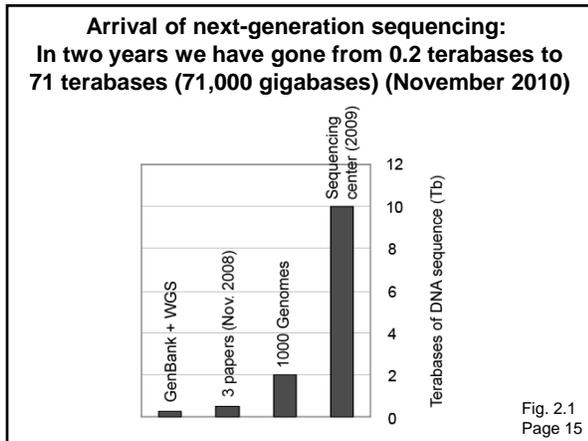
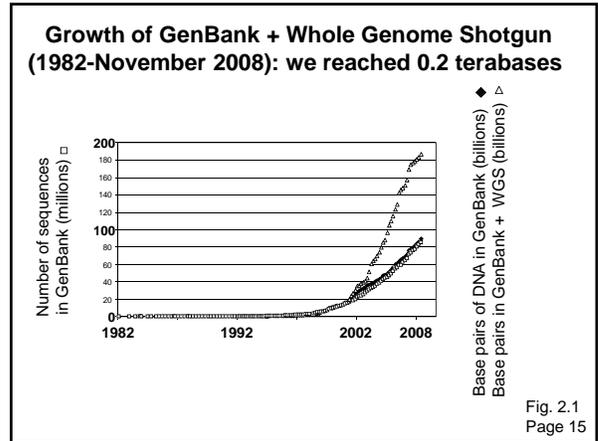
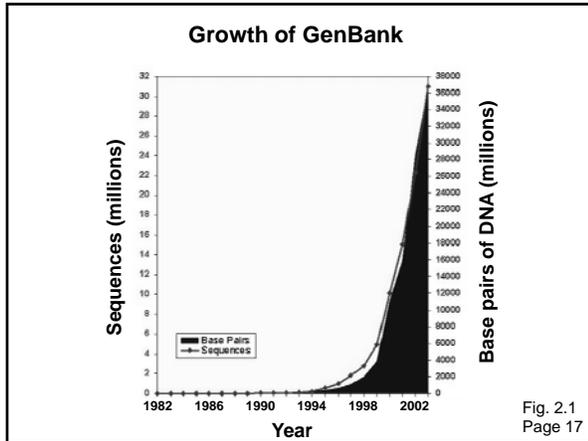
Body region, physiology, pharmacology, pathology

Page 5



Page 6







## 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>

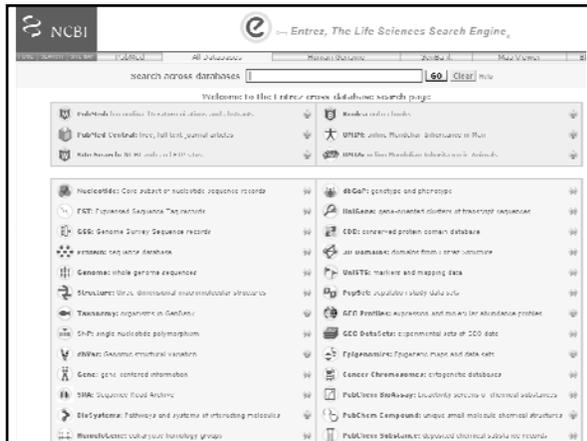
Page 23

## 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	<b>DNA</b>
NT_030059	Genomic contig	
Rs7079946	dbSNP (single nucleotide polymorphism)	
N91759.1	An expressed sequence tag (1 of 170)	<b>RNA</b>
NM_006744	RefSeq DNA sequence (from a transcript)	
NP_007635	RefSeq protein	<b>protein</b>
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)

Page 29

From the NCBI home page, type "beta globin" and hit "Search"

The screenshot shows the NCBI home page with the search bar containing the text 'beta globin'. A red arrow points to the search button. The page includes various navigation links and a 'Resources' sidebar.

The screenshot shows the search results for 'beta globin'. A red arrow points to the 'Gene' link in the 'Follow the link to Gene' box. The results list various databases like PubMed, GenBank, and RefSeq.

Fig. 2.5 Page 28

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

The screenshot shows the Entrez Gene page for HBB. The 'Limits' dropdown menu is open, showing options like 'Human', 'Mammalia', and 'All'. The 'Official Symbol' HBB is highlighted.

The screenshot shows the Entrez Gene page for HBB. The 'Limits' dropdown menu is open, showing options like 'Human', 'Mammalia', and 'All'. The 'Limits by Chromosomal Region' section is visible, with 'Human' selected.

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

**By applying limits, there are now far fewer entries**

The screenshot shows the Entrez Gene search interface. The search term 'HBB' is entered, and several filters are applied: 'Current Only', 'Genes Genomes', and 'SNP Genes/View'. The results list shows two entries for HBB, with the first entry selected. The 'Recent activity' sidebar shows search history for HBB and related terms like 'beta globin' and 'hemoglobin beta chain'.

**Entrez Gene (top of page)**

This screenshot shows the top of the Entrez Gene page for HBB. It includes the gene's official symbol, full name, primary source, and a detailed summary. A note on the right side of the page states: "Note that links to many other HBB database entries are available". The page number "Page 30" is visible in the bottom right corner.

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

This screenshot shows the middle of the Entrez Gene page for HBB. It features a genomic map of chromosome 11p15.5, a bibliography section with a table of references, and a section for related articles published in the field.

Year	Author	Journal	PMID
2007	Wang J, et al.	Nat Genet	1716
2007	Wang J, et al.	Nat Genet	1716
2007	Wang J, et al.	Nat Genet	1716
2007	Wang J, et al.	Nat Genet	1716

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

This screenshot shows the middle of the Entrez Gene page for HBB, focusing on phenotypes and Gene Ontology (GO) terms. It lists various clinical conditions like Erythremia, beta-thalassemia, and sickle cell anemia, along with GO terms such as 'hemoglobin binding' and 'iron ion binding'.

Phenotype	Gene Ontology (GO) Term
Erythremia, beta-	hemoglobin binding
Sickle cell anemia	iron ion binding
Thalassemia-beta, dominant inclusion-body	metal ion binding
Thalassemias, beta-	molecular function
	oxygen binding
	oxygen transporter activity

**Entrez Gene (bottom of page): RefSeq accession numbers**

This screenshot shows the bottom of the Entrez Gene page for HBB, detailing RefSeq accession numbers. It lists genomic reference sequences (e.g., NC\_000007.3), mRNA sequences (e.g., NM\_000518.4), and protein sequences (e.g., P02012). It also includes information about RefSeqs of Annotated Genomes and reference assembly.

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

This screenshot shows a list of non-RefSeq accessions for HBB. The list includes various identifiers such as AF032013.1, AF032014.1, AF032015.1, etc., along with their corresponding accession numbers and dates.

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="O-glycosylation site"
          /citation[1]
Site      121
          /site_type="glycosylation"
          /experiment="experimental evidence, no additional details
          recorded"
          /note="glycosylation site"
          /citation[2]
CD        1..147
          /gene="HBB"
          /gene_synonym="CD1131-C"
          /coded_by="NM_000518.4:51..494"
          /db_xref="CCDS:CCDS7933.1"
          /db_xref="GeneID:1041"
          /db_xref="MIM:10276"
          /db_xref="MIM:131200"

ORIGIN
1  MSHLPSKPK AVKALGKPK VDRVGGKGLK KLVVYRPTK CFFKFGKLG KGVKGGKPK
41  VSKHGKIVPK AFKQKLVKDK KIKPKKATK KLVKDKKIVKDK PAKFKVIGKPK IVKIVKHKPK
121  KETFKVQKPK VQKVVQVYKPK ALKHKPK
//
  
```

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

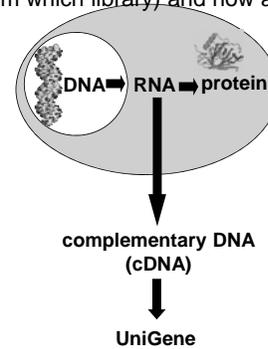


Fig. 2.3  
Page 22

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

The screenshot shows the NCBI HomoloGene search results for 'Hemoglobin, beta'. It lists several genes from different species, including Homo sapiens, Pan troglodytes, and Canis lupus familiaris, along with their accession numbers and protein lengths.

Gene	Protein
HBB, Homo sapiens hemoglobin, beta	NP_000029.1 147 aa
HBB, Pan troglodytes hemoglobin, beta	XP_508242.1 147 aa
LOC638462, Canis lupus familiaris similar to beta-globin	XP_559323.1 147 aa
HBD, Canis lupus familiaris hemoglobin, delta	XP_534029.2 147 aa
LOC480784, Canis lupus familiaris similar to beta-globin	XP_537902.1 147 aa
LOC781986, Bos taurus similar to gamma-globin	XP_001249460.2 145 aa

## 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.

The screenshot shows the UniProt homepage with a search bar, a 'WELCOME' message, and a 'NEWS' section. The 'WELCOME' message states: 'The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.'

The 'NEWS' section includes: 'UniProt release 15.10 - Nov 3, 2009', 'What are UniProt Complete proteome sets? How to retrieve them?', 'Change to cross-reference to OMA', 'Statistics for UniProt: Swiss-Prot, TrEMBL', 'Fortcoming changes', and 'New articles'.

The 'SITE TOUR' section includes: 'UniProtKB: Protein knowledgebase, consists of two sections: Swiss-Prot, which is manually annotated and reviewed; TrEMBL, which is automatically annotated and is not reviewed. Includes Complete Proteome Sets.', 'UniRef: Sequence clusters, used to speed up similarity searches.', 'UniParc: Sequence archive, used to keep track of sequences and their identifiers.', and 'Supporting data: Literature citations, taxonomy, keywords and more.'

The UniProt logo is visible at the bottom of the page.

Page 33

### ExpASY: vast proteomics resources (www.expasy.ch)

The ExpASY (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the storage of protein sequences and structures, as well as 3D-PDB (3) databases / References / Links to other XREFs.

**Databases**

- UniProt (UniProt, PROSITE, TrEMBL, SwissProt, Swiss-Prot)
- ViralZone, SWISS-MODEL Repository, SWISS-PROF, UniProt, UniProt
- Repository: MATH-Cat, H-10/10/10
- Cytochrome P450, UniProt

**Tools & Software**

- ProtParam, ProtParam2, ProtParam
- Melanie, MSight, Melkard UniProt, UniProt

**Latest News**

- Protein Spotlight** - Oct 24, 2009: In like a shot: Making use of a tubular structure to inject something into something else or something else into something else.
- World-2PAGE** - Oct 23, 2009: New data uploaded into the World-2PAGE Repository. Currently, 113 images for 16 species are available from the World-2PAGE Portal.

**Protein Annotations**

- Downloads: Protein Spotlight, Proteomics for UniProt, Proteomics, Bioinformatics core facility for Proteomics

**Documentation**

- What's New? / e-mail alerts / UniProt documentation, how to link to ExpASY, Advanced search

### 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 genome browser (www.ensembl.org)

Search: All species for [Go]

e.g. human gene BRCA2 or rat X:10000..20000 or insulin

**Browse a Genome**

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Click on a link below to go to the species' home page.

**Popular genomes** (Log in to customize this list)

- Human
- Mouse
- Zebrafish

All genomes: [Select a species]

**New to Ensembl?**

- Learn how to use Ensembl with our video tutorials and walkthroughs
- Add custom tracks using our new Control Panel and save it to your Ensembl account
- Search for a DNA or protein sequence using BLAST or BLAT
- Fetch only the data you want from our public database, using the Ensembl Perl API
- Download our databases via FTP in FASTA, MySQL and other formats
- Mine Ensembl with BiMart and export sequences or tables in text, html, or Excel format

Ensembl genome browser search results for 'beta globin'.

Search for: beta globin

Human (Homo sapiens)

**Description**

- Location (6:133017695-133017700)
- Gene (BRCA2)
- Transcript (FOXP2-203)
- Variation (rs1333049)

**Assembly**

This site provides a data set based on the February 2009 Homo sapiens high coverage assembly from the Genome Reference Consortium. The data set consists of gene models built from the genome alignments of the human proteome as well as from alignments of human cDNAs using the cDNA genome model of exons.

This release of the assembly has the following properties:

- 27,478 contigs
- contig length total 3.2 Gb
- chromosome length total 3.1 Gb

**Annotation**

From release 56 (September 2009) a number of improvements have been made to the merge process between the automatic annotation from Ensembl and the manually curated annotation from GENCODE. This refined merge set is now the public output of the GENCODE project. The set displayed in release 56 corresponds to GENCODE release 3c.

Ensembl genome browser detailed view of chromosome 11.

Chromosome 11: 5,246,694-6,250,655

Region overview: [Chromosome ideogram]

Region in detail: [Detailed view of the region]

Gene legend: [Legend for gene tracks]

CCDS set: [CCDS identifiers]



## 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!

Page 37

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

The screenshot shows the NCBI Entrez search interface. The search term is 'hiv-1[organism]'. The results are limited to one entry: 'Human immunodeficiency virus 1, complete genome'. The entry details include: 'ssRNA; linear; Length: 9,181 nt', 'Region Type: viral genome', and 'Created: 1998/01/22'.

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)

The screenshot shows the NCBI Entrez Nucleotide search results for 'hiv-1'. It displays 'Found 364416 nucleotide sequences. Nucleotide (361099) EST (420) 988 (282)'. The results are sorted by default order. A list of top organisms is shown, including 'Human immunodeficiency virus 1 (34679)', 'Homo sapiens (3303)', and 'Simian immunodeficiency virus (1709)'.

## 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.

11-10

The screenshot shows the NCBI Entrez search results for 'histone'. It displays 'Found 3800 nucleotide sequences. Nucleotide (3800) EST (420) 988 (282)'. The results are sorted by default order. A list of top organisms is shown, including 'Human immunodeficiency virus 1 (34679)', 'Homo sapiens (3303)', and 'Simian immunodeficiency virus (1709)'.

## Entrez Gene result for a histone

The screenshot shows the NCBI Entrez Gene result for 'HIST2H4A histone 2, H4a'. The gene is located on chromosome 1. The summary section includes: 'Official Symbol: HIST2H4A and Name: histone 2, H4a provided by HUGO Gene Nomenclature Committee', 'Gene name: HIST2H4A', 'Gene type: protein coding', 'Gene description: histone 2, H4a', and 'RefSeq status: Referred'. The genomic regions, transcripts, and products section shows the gene structure on chromosome 1.

## 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**

Resources  
 NCI Home  
 All Resources (9/2)  
 Literature  
 DNA & RNA  
 Proteins  
 Sequence Analysis  
 Genes & Expression  
 Genomes  
 Maps & Markers  
 Domains & Structures  
 Genetics & Medicine  
 Taxonomy  
 Data & Software  
 Training & Tutorials  
 Homology  
 Small Molecules  
 Variation

Genotype  
 Data from Genome Wide Association studies that links genes and disease. See study variables, protocols, and analysis.

Notice: Upcoming Systems Maintenance  
 NCBI services will undergo maintenance at 3:00 PM until Saturday, November 14. Some retrieval resources such as PubMed, GEO Profiles, and GEO Datasets may be intermittently slow. Web and GEO Datasets such as GeneBank (BankIt), GEO, SRA and PubChem will be unavailable. For questions please contact NCBI: info@ncbi.nlm.nih.gov

How To...  
 - Obtain the full text of an article  
 - Retrieve all sequences for an organism or taxon  
 - Find a homology for a gene in another organism  
 - Find genes associated with a phenotype or disease  
 - Design PCR primers and check them for specificity  
 - Find the function of a gene or gene product  
 - Find syntentic regions between the genomes of two organisms

Popular Resources  
 PubMed  
 PubMed Central

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.

Page 38

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.

Page 38

**PubMed result for HBB**

Search: PubMed  
 hbb

Display Settings: Summary, 20 per page, Sorted by Recently Added

Are you looking for gene information?  
 HBB hemoglobin, beta [Homo sapiens]  
 HBB in Homo sapiens | Mus musculus | Rattus norvegicus | All 20 Gene records

Results: 1 to 20 of 440

1. Control of fetal hemoglobin: new insights emerging from genomics and clinical implications.  
 Thien SL, Menzel S, Lathrop M, Garner C.  
 Hum Mol Genet. 2009 Oct 15;18(R2):R216-23.  
 PMID: 19807730 [PubMed - in process]  
 Related articles

2. Multiplex ligation-dependent probe amplification screening of isolated increased HbF levels revealed three cases of novel rearrangements/deletions in the beta-globin gene cluster.  
 Lee ST, Yoo EH, Kim JY, Kim JW, Ki CS, Br J Haematol. 2009 Oct 5. [Epub ahead of print].  
 PMID: 19807730 [PubMed - as supplied by publisher]  
 Related articles

3. Promoted frame rearrangements in serum from the general population in northern China.  
 Zhu L, Ma B, Hites RA.  
 Environ Sci Technol. 2009 Sep 15;43(18):6853-8.  
 PMID: 19807728 [PubMed - indexed for MEDLINE]  
 Related articles

Filter your results:  
 All (440)  
 Review (13)  
 Free Full Text (116)

Also try:  
 Hbb gene  
 Hbb mutation

Titles with your search terms:  
 Comparison of the mismatch-specific endonuclease meB [PMC: Bioelectron. 2008]  
 Role of tyrosine N-butyl bromide (NBB, N-butyl) as labo [Indian J Med Sci. 2008]  
 Family screening for HBB\* gene and detection of new [Iran Saudi Publica. 2008]  
 See more...

66 free full text articles in PubMed Central  
 Two new beta-thalassemia deletions comprising gene [Haematologica. 2008]

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

Resources  
 NCI Home  
 All Resources (9/2)  
 Literature  
 DNA & RNA  
 Proteins  
 Sequence Analysis  
 Genes & Expression  
 Genomes  
 Maps & Markers  
 Domains & Structures  
 Genetics & Medicine  
 Taxonomy  
 Data & Software  
 Training & Tutorials  
 Homology  
 Small Molecules  
 Variation

MeSH

MeSH

Popular Resources  
 PubMed  
 PubMed Central

**A "how to" pull-down menu links to tutorials**

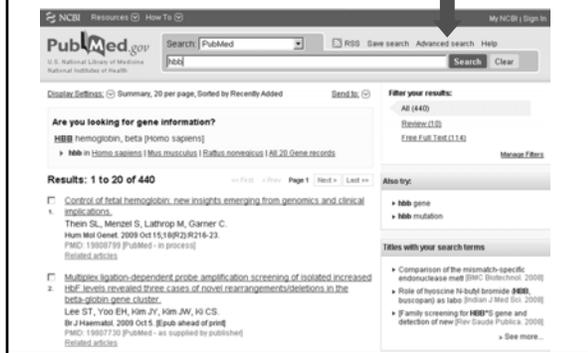
Resources  
 NCI Home  
 All Resources (9/2)  
 Literature  
 DNA & RNA  
 Proteins  
 Sequence Analysis  
 Genes & Expression  
 Genomes  
 Maps & Markers  
 Domains & Structures  
 Genetics & Medicine  
 Taxonomy  
 Data & Software  
 Training & Tutorials  
 Homology  
 Small Molecules  
 Variation

How To

How To

Popular Resources  
 PubMed  
 PubMed Central

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

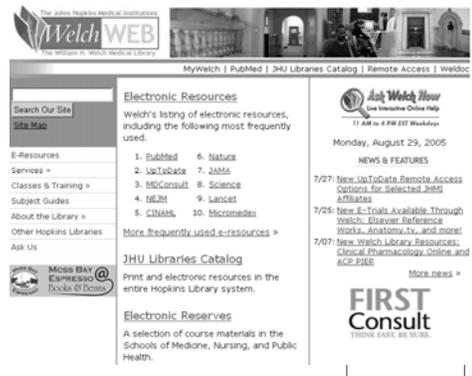
Obtain articles on-line via Welch Medical Library (and download pdf files): <http://www.welch.jhu.edu/>

1 AND 2 lipocalin AND disease (504 results)

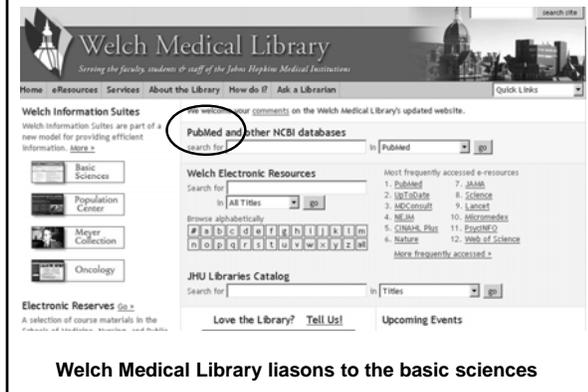
1 OR 2 lipocalin OR disease (2,500,000 results)

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

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



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



Welch Medical Library liaisons to the basic sciences

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

