Intro to Bioinformatics

Esti Yeger-Lotem
Gideon Greenspan

Lecture I: Introduction & Text Based Search

prepared with some help from friends...

Metsada Pasmanik-Chor, Benny Chor, Dan Geiger,
Hanah Margalit, Ron Pinter, Zohar Yakhini
and numerous web resources.
Course requirements:

1. Attend all lectures.
2. Submit all written assignments.
   - There will be 6-8 assignments.
   - Each assignment is to be done and submitted in pairs.
   - Assignments are due at the beginning of
     the next lecture after they are given.
   - Each assignment will be typed within two page limit
     (no e-mail or disketts).
3. Final project:
   Critically review a topic and analyze data.
   Propose and implement new approaches using tools taught in class.
   Submit a written report and make an oral presentation in class.
   Each presentation will be about 15 minutes long.
   Presentations will be scheduled for the last 2 meetings.
4. The course web site:
   http://dogbert.cs.technion.ac.il/courses/236806/
Course outline:

- **General information**: Introduction to bioInformatics.
- **BioInformatics tools**, NCBI.
- **Databases search**: ENTREZ, PubMed, OMIM:
  - choose a gene to work with throughout the course.
- **Nucleotides**: Pairwise sequence alignment (BLAST, FASTA).
- **Proteins**: Pairwise and multiple sequence alignment
  (BLASTP, PSI-BLAST, FASTA, CLASTALW).
- **Protein structure analysis**: secondary and tertiary structure.
- **Proteins families** - motifs, domains, clustering.
- **Projects' List**.
- **Phylogeny**: Tree reconstruction methods, gene duplication.
- **Genomics**: Procaryotes and eucaryotes, gene hunting, splicing.
- **The Human Genome Project**: drug design, other genomes.
- **Gene expression analysis**: DNA micro arrays (chips), clustering tools.
- **Presentations** of students projects.
- **Conclusions**: Further directions, additional courses, the future, etc.
LITERATURE:

References on the web: Please refer to class notes, and to the list of "important sites".
A Few Basic Concepts of Molecular Biology:

- Genetic material - DNA & RNA.
  - DNA as a sequence of bases (A,C,T,G).
  - Watson-Crick complementation.
- Proteins.
- The central dogma of molecular biology.
Central Dogma

Gene (DNA) \(\xrightarrow{\text{Transcription}}\) mRNA \(\xrightarrow{\text{Translation}}\) Protein

Cells express different subset of the genes in different tissues and under different conditions
The Central Dogma of Molecular Biology

Replication-
DNA duplication

Transcription-
RNA synthesis

Translation-
Protein synthesis

Central Paradigm of Molecular Biology

DNA → RNA → Protein → Symptoms (Phenotype)
Central Paradigm of Bioinformatics

Genetic information

SRAAINKHIVA
VSYQTVSRVVN
VSTATVSRALA
GVTTTVSIVV
SGVSAVSAILN
GVSEMTRDRDLN
TAYATIHVTVE
GSPQTVSRELA
MSIATITGRSN
ISRETVGRLK
FDISRSLHLFR
LRPSRLAHLFR
MTVETISRLLC
TLEFHHLRLFK
Central Paradigm of Bioinformatics

Genetic Information → Molecular Structure
Central Paradigm of Bioinformatics

Genetic Information → Molecular Structure → Biochemical Function

[Diagram showing protein structure with amino acid sequences and molecular structures of L-Phenylalanine and L-Tyrosine]
Central Paradigm of Bioinformatics

Genetic Information ➔ Molecular Structure ➔ Biochemical Function ➔ Symptoms
Central Paradigm of Bioinformatics

Genetic Information → Molecular Structure → Biochemical Function → Symptoms
Computer Science Tools are \textit{Crucial}

- New bio-technologies create \textit{huge} amounts of data
- It is impossible to analyze data by \textit{manual} inspection.

- \textbf{Novel} mathematical, statistical, algorithmic and computational tools are \textit{necessary}!
Three Specific Examples:

- Molecular evolution and the **TREE OF LIFE**. (a classical, basic science problem, since Darwin’s 1859 ""Origin of Species"").
- The Human Genome Project (HGP):
  - Write down all of human DNA on a single CD (“completed” 2001).
  - Identify all genes, their locations and function (far from completion).
- DNA Chips and **personalized medicine** (leading edge, future technologies).
TREE OF LIFE: Searching Protein Sequence Databases - How far can we see back?

-4.0

-3.0

-2.0

-1.0

-0.1

Time (billion years)

Formation of the solar system

First self replicating systems

Prokaryotes/ eukaryotes

Plant/ animals

Invertebrates/ vertebrates

Mammalian radiation

Adapted from Dayhoff et al., 1978.

Origin of the universe?
Microarrays ("DNA Chips")

- New technological breakthrough:
  - Measure, in one experiment:
    RNA expression levels of thousands of genes.
A Big Goal

“The greatest challenge, however, is analytical. ... Deeper biological insight is likely to emerge from examining datasets with scores of samples.”


BIOINFORMATICS:
Provide methodologies for elucidating biological knowledge from biological data.
What is BIOINFORMATICS?

A field of science in which Biology, Computer Science and Information Technology merge into a single discipline.

Goal: To enable the discovery of new biological insights and create a global perspective for biologists.
Disciplines:

- Development of new algorithms and statistics to assess relationships among members of large data sets.

- Analysis and interpretation of various types of data.

- Development and implementation of tools to efficiently access and manage different types of information.
Why use BIOINFORMATICS?

- An explosive **growth** in the amount of biological information necessitates the **use of computers** for cataloguing and retrieval.

- A more **global perspective** in experimental design (from “one scientist = one gene/protein/disease” paradigm to whole organism consideration).

- **Data mining** - functional/structural information is important for studying the molecular basis of diseases (and evolutionary patterns).
Genomics, Bioinformatics & Medicine

- Genomics
- Molecular Diagnostics
- Molecular Epidemiology
- Bioinformatics
- Identify Drug Targets
- Rational Drug Design
- Genetic Therapy

- Machine Learning
- Robotics
- Artificial Intelligence
- Statistics & Probability
- Databases
- Information Theory
- Algorithms
- Graph Theory

© Doug Brutlag, 1999
Listed below are some of the major events in bioinformatics over the last several decades. Most of the events in the list occurred long before the term, "bioinformatics", was coined. In most cases, links take the user to outside sites that provide further explanation of each event and access to related resources.

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1962</td>
<td>Pauling’s theory of molecular evolution</td>
</tr>
<tr>
<td>1967</td>
<td>Margaret Dayhoff’s Atlas of Protein Sequences</td>
</tr>
<tr>
<td>1970</td>
<td>Needleman-Wunsch algorithm</td>
</tr>
<tr>
<td>1977</td>
<td>DNA sequencing and software to analyze it (Staden)</td>
</tr>
<tr>
<td>1981</td>
<td>Smith-Waterman algorithm developed</td>
</tr>
<tr>
<td>1981</td>
<td>The concept of a sequence motif (Doolittle)</td>
</tr>
<tr>
<td>1982</td>
<td>GenBank Release 3 made public</td>
</tr>
<tr>
<td>1982</td>
<td>Phage lambda genome sequenced</td>
</tr>
<tr>
<td>1983</td>
<td>Sequence database searching algorithm (Wilbur-Lipman)</td>
</tr>
<tr>
<td>1985</td>
<td>FASTP/FASTN: fast sequence similarity searching</td>
</tr>
<tr>
<td>1988</td>
<td>National Center for Biotechnology Information (NCBI) created at NIH/NLM</td>
</tr>
</tbody>
</table>
1988 **EMBnet network** for database distribution

1990 **BLAST**: fast sequence similarity searching

1991 EST: expressed sequence tag sequencing

1993 **Sanger Centre**, Hinxton, UK

1994 **EMBL European Bioinformatics Institute**, Hinxton, UK

1995 First **bacterial genomes** completely sequenced—Haemophilus influenzae (2 Mb).

1996 **Yeast genome** completely sequenced—First Eukaryote genome (Saccharomyces cerevisiae (12 Mb)).

1997 **PSI-BLAST**

1998 **Worm (multicellular) genome** completely sequenced—First multi-cellular Eukaryote (Caenorhabditis elegans (100Mb)).

1999 **Fly genome** completely sequenced—A model organism for animal kingdom (Drosophila melanogaster).

- A model organism for plant kingdom—(Arabidopsis thaliana).

**Disclaimer**  **Privacy statement**

**Revised March 13, 2000**
Central Paradigm of Bioinformatics

Genetic Information → Molecular Structure → Biochemical Function → Phenotype

Genetic Information

Molecular Structure

Biochemical Function

Phenotype
Why is it Hard to Elucidate from Sequence?

- Genetic information is redundant
  - Genetic code
  - Accepted amino acid replacements
  - Intron-Exon variation
  - Strain variation
- Structural information is redundant
  - Conformational changes
  - Different structures may result in similar functions
  - Different sequences result in the same structure
- Single genes have multiple functions.
  - May act as an metabolic enzyme and as a regulator.
- Genes are 1-dimensional but function depends on 3-dimensional structure.
Biological Revolution Necessitates Bioinformatics:

New bio-technologies (automatic sequencing, DNA chips, protein identification, mass specs., etc.) produce large quantities of biological data.

Bioinformatics: Development of algorithms that enable the analysis of the data (from experiments or from databases).

Data produced by biologists and stored in database → Bioinformatics Algorithms and Tools → New information for biological and medical use
ENTREZ

A search and retrieval system for information integration.
Entrez: Genotype to Phenotype (1996)
Entrez Increases Discovery Space 1998

- PubMed abstracts
- Full text journals online
- Phylogeny (Taxonomy)
- 3-D Structure (MMDB)
- Genomes
- Nucleotide sequences
- Protein sequences
**Entrez** is a retrieval system for searching several linked databases.

It provides access to:

- **PubMed**: The biomedical literature (PubMed)
- **Nucleotide**: Sequence databases (GenBank)
- **Protein**: Sequence databases
- **Structure**: Three-dimensional macromolecular structures
- **Genome**: Complete genome assemblies
- **PopSet**: Population study data sets
- **Taxonomy**: Organisms in GenBank
- **OMIM**: Online Mendelian Inheritance in Man

- MedLine, abstracts and links to journals
- GenBank, EMBL, DDBJ, RefSeq
- GenBank, DDBJ, EMBL, PDB, PIR, Swiss-prot, RefSeq
- NCBI’s MMDB-derived from PDB
- Graphical view, mapping data
- Population and phylogenetic studies

<table>
<thead>
<tr>
<th>Database</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PubMed</td>
<td>National Library of Medicine</td>
</tr>
<tr>
<td>GenBank</td>
<td>National Institutes of Health</td>
</tr>
<tr>
<td>Genomes</td>
<td>GenBank for Genomes</td>
</tr>
<tr>
<td>LocusLink</td>
<td>GenBank for LocusLink</td>
</tr>
<tr>
<td>OMIM</td>
<td>GenBank for OMIM</td>
</tr>
<tr>
<td>Proteins</td>
<td>GenBank for Proteins</td>
</tr>
<tr>
<td>Structures</td>
<td>GenBank for Structures</td>
</tr>
</tbody>
</table>

**NCBI Search**

- **Search:** GenBank
- **For:** [Input Field]
- **Submit:** Go
<table>
<thead>
<tr>
<th>GenBank sections:</th>
<th>PLN - Plant sequences.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PRI - Primates</td>
</tr>
<tr>
<td></td>
<td>ROD - Rodents</td>
</tr>
<tr>
<td></td>
<td>MAM - Other mammals</td>
</tr>
<tr>
<td></td>
<td>VRT - Other vertebrates</td>
</tr>
<tr>
<td></td>
<td>INV - Invertebrates</td>
</tr>
<tr>
<td></td>
<td>BCT - Bacterial</td>
</tr>
<tr>
<td></td>
<td>PHG - Phage</td>
</tr>
<tr>
<td></td>
<td>VRL - Viral</td>
</tr>
<tr>
<td></td>
<td>SYN - Synthetic</td>
</tr>
<tr>
<td></td>
<td>UNA - Unannotated</td>
</tr>
<tr>
<td></td>
<td>PAT - Patent</td>
</tr>
<tr>
<td></td>
<td>NEW - New</td>
</tr>
</tbody>
</table>
GenBank Growth

<table>
<thead>
<tr>
<th>Year</th>
<th>bp</th>
<th>Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>384,939,485</td>
<td>555,694</td>
</tr>
<tr>
<td>1996</td>
<td>651,972,984</td>
<td>1,021,211</td>
</tr>
<tr>
<td>1997</td>
<td>1,160,300,687</td>
<td>1,765,847</td>
</tr>
<tr>
<td>1998</td>
<td>2,008,761,784</td>
<td>2,837,897</td>
</tr>
<tr>
<td>1999</td>
<td>3,841,163,011</td>
<td>4,864,570</td>
</tr>
<tr>
<td>2000</td>
<td>11,101,066,288</td>
<td>10,106,023</td>
</tr>
</tbody>
</table>
GenBank on FTP site

Current directory is /genbank

ftp> open ncbi.nlm.nih.gov
ftp> cd genbank

Release 124: 233 files; 51.7 Gigabytes uncompressed
PubMed:

- A Web-based retrieval system developed by NCBI at the NLM. It is part of NCBI's retrieval system, known as Entrez.

- PubMed is a database of bibliographic information drawn primarily from the life sciences literature.

- Links to full-text articles as well as links to other third party sites such as libraries and sequencing centers.

- PubMed provides access and links to the integrated molecular biology databases maintained by NCBI.
**PubMed:**

- The largest, most used and best known of NLM databases (90% of all searches are done in MEDLINE).
  - > 40 databases online.
  - > 20 million records.
  - External databases.
  - > 9 million searches per month.
Searching the literature

MedLine Indexing:

MESH (Medical Subject Heading):
Use a term to limit retrieval.
(Human, animal, male, female, age group, organism, etc.).

Publication Type:
Review, clinical trial, letter, journal article, etc.

Search Terms By:
Author name, title word, text word, journal title, publication date, phrase, or any combination of these.

- Words are automatically added, but Boolean operators (AND, OR, NOT, in UPPER CASE) are welcome.
Primary databases - Databases consisting of data derived experimentally.
Secondary databases - Data that are derived from analysis of primary databases.
(Example: A protein database consisting of the translation of nucleotide sequences is also a secondary database).
How much information is there?

- Nucleotide records
  - 12,243,766
- Nucleotides
  - 12,973,707,065
- Protein sequences
  - 708,648
- 3D structures
  - 15,557
- Expression data points
  - >20,000,000
- Human Unigene clusters
  - 97,521
- Maps and complete genomes
  - 1,161
- Different organisms
  - 101,357
- Human SNPs in dbSNP
  - 2,972,052
- Human LocusLink records
  - 20,582
- Finished human sequence
  - 1,140,365 kb (33%)
- PubMed records
  - >11,000,000
- OMIM records
  - 12,774
• Exponential growth of biological information: growth of sequences, structures, and literature.

• Efficient storage and management tools were most important.
Sequence families or neighborhoods can be defined.
Common sequences can be identified in multiple alignment.
These motifs can provide clues for biochemical function.
Clustering sequences into trees reflect the degree of similarity between species and evolutionary relationships.
NCBI bioinformatics tools

**BLAST**

The Basic Local Alignment Search Tool (BLAST), for comparing gene and protein sequences against others in public databases, now comes in several flavors including PSI-BLAST, PHI-BLAST, and BLAST 2 sequences. Specialized BLASTs are also available for human, microbial, and malaria genomes, as well as for vector contamination, immunoglobulins, and tentative human consensus sequences.

**Clusters of Orthologous Groups (COGs)** currently covers 21 complete genomes from 17 major phylogenetic lineages. A COG is a cluster of very similar proteins found in at least three species. The presence or absence of a protein in different genomes can tell us about the evolution of the organisms, as well as point to new drug targets.

**Map Viewer**

Map Viewer shows integrated views of chromosome maps currently for human, mouse, and Drosophila. Used to view the NCBI assembly of the complete human genome, Map Viewer is a valuable tool for the identification and localization of genes that contribute to human disease.

**LocusLink**

LocusLink combines descriptive and sequence information on human genes through a single query interface. LocusLink covers information on official nomenclature, aliases, sequence accession numbers, phenotypes, EC numbers, OMIM numbers, UniGene clusters, map information, and relevant web sites.
A UniGene cluster is a non-redundant set of sequences that represents a unique human, mouse, or rat gene. Well-characterized genes, as well as thousands of expressed sequence tag (EST) sequences have been included. Each cluster record also contains information such as the tissue types in which the gene has been expressed and map location. UniGene can assist in gene discovery, gene mapping projects, and large-scale expression analysis.

ORF finder identifies all possible ORFs in a DNA sequence by locating the standard and alternative stop and start codons. The deduced amino acid sequences can then be used to BLAST against GenBank. ORF finder is also packaged in the sequence submission software Sequin.

Electronic PCR allows you to search your DNA sequence for sequence tagged sites (STSs), which have been used as landmarks in various types of genomic maps. It compares the query sequence against data in NCBI's UniSTS, a unified, non-redundant view of STSs from a wide range of sources.
**VAST search** is a structure-structure similarity search service. It compares 3D coordinates of a newly determined protein structure to those in the MMDB/PDB database. VAST Search computes a list of similar structures that can be browsed interactively, using molecular graphics to view superimpositions and alignments.

The Human-Mouse Homology Maps compare genes in homologous segments of DNA from human and mouse sources, sorted by position in each genome. A total of 1793 loci are presented, most of which are genes. This map should be interpreted as a reflection of probable, not confirmed, homology relationships due to the lack of further information available for about half the loci.

**VecScreen** is a tool for identifying segments of a nucleic acid sequence that may be of vector, linker or adapter origin prior to sequence analysis or submission. VecScreen was developed to combat the problem of vector contamination in public sequence databases.

The Cancer Chromosome Aberration Project (CCAP) compiles information on the distinct chromosome aberrations that are associated with different cancers. The identification of chromosomal abnormalities by clinicians can enable the diagnosis of, classification of, and treatment selection for a given cancer.

**CGAP** aims to decipher the molecular anatomy of cancer cells. CGAP develops profiles of cancer cells by comparing gene expression in normal, precancerous, and malignant cells from a wide variety of tissues.
OTHER TEXT BASED SEARCHES:

- **SRS** (sequence retrieval system) at EBI, England.  
  [http://srs.ebi.ac.uk/](http://srs.ebi.ac.uk/)

- **STAG** at DDBJ, Japan.  
  [http://stag.genome.ad.jp/](http://stag.genome.ad.jp/)

- **Expasy** at SIB (Swiss Institute of Bioinformatics), Switzerland.  

Search the ExPASy site
GenBank is part of international collaboration of NCBI, DDBJ, EMBL.
Information puzzles end abruptly?
Database Searching with Entrez
Entrez is a retrieval system for searching several linked databases. It provides access to:

- PubMed: The biomedical literature (PubMed)
- Nucleotide sequence database (Genbank)
- Protein sequence database
- Structure: three-dimensional macromolecular structures
- Genome: complete genome assemblies
- PopSet: population study data sets
- OMIM: Online Mendelian Inheritance in Man
- Taxonomy: organisms in GenBank
- Books: online books
- ProbeSet: Gene Expression Omnibus (GEO)
- 3D Domains: domains from Entrez Structure

Pre-computed similarity searches are available for most database records producing a list of related sequences, structure neighbors, as well as related articles.
Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish.

NCBI

Search Nucleotide for sickle cell anemia

Display Summary, Save, Text, Clip Add

Show: 20

1: NM_000518
Homo sapiens hemoglobin, beta (HBB), mRNA
gi|13788565|ref|NM_000518.3|[13788565]

2: AF352795
Homo sapiens bilirubin UDP-glucuronosyltransferase 1-1 (UGT1A1) gene, UGT1A1*1 allele, partial cds
gi|13569708|gb|AF352795.1|AF352795[13569708]

3: U01317
Human beta globin region on chromosome 11
gi|455025|gb|U01317.1|HUMHBB[455025]

4: AF180372
Homo sapiens bilirubin UDP-glucuronosyltransferase 1-1 (UGT1) gene, UGT1*1 allele, partial cds
gi|6010649|gb|AF180372.1|AF180372[6010649]

Related Sequences, OMIM, Protein, PubMed, Taxonomy
LOCUS  NM_000518  626 bp  mRNA  PRI  04-MAY-2001
DEFINITION Homo sapiens hemoglobin, beta (HBB), mRNA.
ACCESSION  NM_000518
VERSION  NM_000518.3  GI:13788565
KEYWORDS .
SOURCE  human.
ORGANISM Homo sapiens
  Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
  Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE  1 (bases 1 to 626)
AUTHORS  Marotta CA, Forget BG, Cohen-Solal M and Weissman SM.
TITLE  Nucleotide sequence analysis of coding and noncoding regions of
  human beta-globin mRNA
JOURNAL  Prog. Nucleic Acid Res. Mol. Biol. 19, 165-175 (1976)
MEDLINE  77126403
PUBMED  1019344
REFERENCE  2 (bases 1 to 626)
AUTHORS  Proudfoot NJ and Brownlee GG.
Gaucher Disease:

- A rare inherited disease, caused by enzyme deficiency. Highest prevalence in Ashkenazi Jewish population.

- Characterization: remarkable degree of variability in clinical signs and symptoms.

- Clinical signs: anemia, bone damage, enlarged liver and spleen (most patients have Type 1 disease), some develop additional severe central nervous system damage (Type 2 & 3, neuronopathic).

- Cause: genetic disorder in the function of glucocerebrosidase, which result in accumulation of the lipid glucocerebrosides in cells lysosomes.

- Traditional therapy: total or partial removal of the spleen, blood transfusions, orthopedic procedures and bone marrow transplantation. Enzyme replacement therapy, very costly: 100,000 - 400,000 US$/patient/year.
A Simple GenBank Record

LOCUS XM_034490 1782 bp mRNA PRI 27-AUG-2001
DEFINITION Homo sapiens glucosidase, beta; acid (includes glucosylceramidase) (GBA), mRNA.
ACCESSION XM_034490
VERSION XM_034490.1 GI:14780919
KEYWORDS .
SOURCE human.
ORGANISM Homo sapiens
  Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE 1 (bases 1 to 1782)
AUTHORS NCBI Annotation Project.
TITLE Direct Submission
JOURNAL Submitted (23-AUG-2001) National Center for Biotechnology Information, NIH, Bethesda, MD 20894, USA
FEATURES Location/Qualifiers
  source 1..1782
  /organism="Homo sapiens"
  /db_xref="taxon:9606"
  /chromosome="1"
  gene 1..1782
  /gene="GBA"
  /note="GLUC"
  /db_xref="LocusID:2629"
  /db_xref="MIM:230800"
Keywords.  
Source:  human

Organism:  Homo Sapiens  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.

Accepted common name

Scientific name

Taxonomic lineage according to GenBank
Sequences and Databases

Locus, accession, gi, version

- **Locus Name**: LOCUS NM_000518
  - 2001
- **Sequence Length**: 626 bp
- **Mol. Type**: mRNA
- **GB Division**: PRI
- **Modification Date**: 04-MAY-

**DEFINITION**: Homo Sapiens ..., mRNA

**ACCESSION**: NM_000518

**VERSION**: NM_000518.3

**Accession Num.**: NM_000518

**Gi Num.**: GI:13788565

The number a sequence gets when published, unique to entry.

The number a sequence gets when entering database.
What is OMIM?

Online Mendelian Inheritance in Man.

A catalog of human genes and genetic disorders.

**USING OMIM:**

- Start with a syndrome search or a gene using OMIM.
- Identify the gene for the syndrome (if not known).
- Locate the gene on the genome (& other genes nearby).
- Information about the gene in LocusLink.
- Links to sequences and structure information.
- View structurally similar proteins (DART & CDD).
OMIM™ - Online Mendelian Inheritance in Man™

NEW OMIM is now incorporated into NCBI’s Entrez system and can be queried using the same approach as the other Entrez databases such as PubMed and GenBank. The previous OMIM pages are still available here.

Welcome to OMIM, Online Mendelian Inheritance in Man. This database is a catalog of human genes and genetic disorders authored and edited by Dr. Victor A. McKusick and his colleagues at Johns Hopkins and elsewhere, and developed for the World Wide Web by NCBI, the National Center for Biotechnology Information. The database contains textual information and references. It also contains copious links to MEDLINE and sequence records in the Entrez system, and links to additional related resources at NCBI and elsewhere.

You can do a search by entering one or more terms in the text box above. Advanced search options are accessible via the Limits, Preview/Index, History, and Clipboard options in the grey bar beneath the text box. The OMIM help document provides additional information and examples of basic and advanced searches.

4. What numbering system is used in the OMIM database?

Each OMIM entry is given a unique six-digit number whose first digit indicates the mode of inheritance of the gene involved:

1----- (100000- ) Autosomal dominant (entries created before May 15, 1994)
2----- (200000- ) Autosomal recessive (entries created before May 15, 1994)
3----- (300000- ) X-linked loci or phenotypes
4----- (400000- ) Y-linked loci or phenotypes
5----- (500000- ) Mitochondrial loci or phenotypes
6----- (600000- ) Autosomal loci or phenotypes (entries created after May 15, 1994)

An allelic variant is designated by the MIM number followed by a decimal point and 4 more unique digits.
### OMIM Statistics for September 8, 2001

#### Number of Entries

<table>
<thead>
<tr>
<th></th>
<th>Autosomal</th>
<th>X-Linked</th>
<th>Y-Linked</th>
<th>Mitochondrial</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Established genes or phenotype loci (*)</td>
<td>8949</td>
<td>503</td>
<td>37</td>
<td>37</td>
<td>9526</td>
</tr>
<tr>
<td>Phenotype descriptions (#)</td>
<td>836</td>
<td>68</td>
<td>0</td>
<td>22</td>
<td>926</td>
</tr>
<tr>
<td>Other loci or phenotypes (no prefix)</td>
<td>2322</td>
<td>164</td>
<td>2</td>
<td>0</td>
<td>2488</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>12107</strong></td>
<td><strong>735</strong></td>
<td><strong>39</strong></td>
<td><strong>59</strong></td>
<td><strong>12940</strong></td>
</tr>
</tbody>
</table>
OMIM search - sickle cell anemia

1: #603903
SICKLE CELL ANEMIA

2: #143020
HPA I RECOGNITION POLYMORPHISM, BETA-GLOBIN-RELATED; HPA1

3: *141900
HEMOGLOBIN--BETA LOCUS; HBB
BETA-THALASSEMIAS, INCLUDED
Gene map locus 11p15.5

4: *142470
HETEROCELLULAR HEREDITARY PERSISTENCE OF FETAL HEMOGLOBIN
Gene map locus 6q22.3-q23.1

5: *305435
F-CELL PRODUCTION 1, FCP1
HETEROCELLULAR HEREDITARY PERSISTENCE OF FETAL HEMOGLOBIN, SWISS TYPE, INCLUDED
Gene map locus Xp22.2
SICKLE CELL ANEMIA

TEXT

A number sign (#) is used with this entry because sickle cell anemia is the result of mutant beta globin (HBB [141900]) in which the mutation causes sickling of hemoglobin rather than reduced amount of beta globin which causes beta-thalassemia.

The most common cause of sickle cell anemia is Hb S [141900.0243], with SS disease being most prevalent in Africans.

CLINICAL FEATURES

In many children with sickle cell anemia, functional asplenia develops during the first year of life and septicemia is the leading cause of death in childhood. The risk of septicemia in sickle cell anemia is greatest during the first 3 years of life and is reduced markedly by prophylactic penicillin therapy. Less is known about splenic dysfunction and the risk of overwhelming sepsis in children with sickle cell-hemoglobin C disease (see Hb C; 141900.0038), although functional asplenia has been documented by radionuclide liver-spleen scans in some adult patients (Ballas et al., 1982) and an elevated erythrocyte pit count, a finding that indicates functional asplenia in children with sickle cell anemia, also has been found in some children with SC disease (Pearson et al., 1985). Lane et al. (1994) reported 7 fatal cases of pneumococcal septicemia in children with SC disease. The earliest death occurred in a 1-year-old child who had cyanotic congenital heart, the other children were aged 3.5 to 15 years. Only 1 child had received pneumococcal vaccine or prophylactic penicillin therapy. All 7
Hemoglobin-Beta Locus; HBB

Alternative titles; symbols

Beta-thalassemias, included
Methemoglobinemia, beta-globin type, included
Erythremia, beta-globin type, included

Gene map locus 11p15.5

Text

The alpha and beta loci determine the structure of the 2 types of polypeptide chains in adult hemoglobin, Hb A. Mutant beta globin that sickles causes sickle cell anemia (603903). Absence of beta chain causes beta-zero-thalassemia. Reduced amounts of detectable beta globin causes beta-plus-thalassemia. For clinical purposes, beta-thalassemia is divided into thalassemia major (transfusion dependent), thalassemia intermedia (of intermediate severity), and thalassemia minor (asymptomatic).

Patients with thalassemia major present in the first year of life with severe anemia, they are unable to maintain a hemoglobin level about 5 gm/dl. Clinical details of this disorder have been detailed extensively in numerous monographs and are summarized by Weatherall et al.
The OMIM Gene map presents the cytogenetic map location of disease genes and other expressed genes described in OMIM. See the OMIM Morbid Map for a list of disease genes organized by disease. For more refined maps of genes and DNA segments click on the Location to invoke NCBI Entrez Map Viewer.

Search for: [ ] Find  Find Next  (from the current location)

- Enter gene symbol, chromosomal location, or disorder keyword to search for, e.g. "CYP1", "5", "1pter", "Xq", or "alzheimer".
- You must capitalize X and Y to search for those chromosomes.

### 11p15.5, HBB to 11p15.5, SIRT3

<table>
<thead>
<tr>
<th>Location</th>
<th>Symbol</th>
<th>Title</th>
<th>MIM #</th>
<th>Disorder</th>
<th>Comments</th>
<th>Method</th>
<th>Mouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>11p15.5</td>
<td>Hbb</td>
<td>Hemoglobin beta</td>
<td>141900</td>
<td>Sickle cell anemia (3); Thalassemias, beta-(3); Methemoglobinemias, beta-(3); Erythremias, beta-(3); Heinz body anemias, beta-(3); HPFH, deletion type (3)</td>
<td>pseudogene HBBP1 between HBG and HBD loci</td>
<td>LD, AAS, F, Fd</td>
<td>7(Hbb)</td>
</tr>
<tr>
<td>11p15.5</td>
<td>HBD</td>
<td>Hemoglobin delta</td>
<td>142000</td>
<td>Thalassemia, delta- (3); Thalassemia due to Hb Lepore (3)</td>
<td></td>
<td>AAS</td>
<td></td>
</tr>
<tr>
<td>11p15.5</td>
<td>HBG1</td>
<td>Hemoglobin, gamma A</td>
<td>142200</td>
<td>HPFH, nondeletion type A (3)</td>
<td></td>
<td>RE</td>
<td></td>
</tr>
<tr>
<td>FEATURES</td>
<td>Location/Qualifiers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>---------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>source</td>
<td>1..1782</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/organism</td>
<td>&quot;Homo sapiens&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/db_xref</td>
<td>taxon:9606</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/chromosome</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene</td>
<td>1..1782</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/gene</td>
<td>&quot;GBA&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/note</td>
<td>&quot;GLUC&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/db_xref</td>
<td>&quot;LocusID:2629&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/db_xref</td>
<td>&quot;MIM:230800&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDS</td>
<td>36..1586</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/gene</td>
<td>&quot;GBA&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/codon_start</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/product</td>
<td>&quot;glucosidase, beta; acid (includes glucosylceramidase)&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/protein_id</td>
<td>&quot;XP_034490.1&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/db_xref</td>
<td>&quot;GI:148708902&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/translation</td>
<td>&quot;MAGSLTGLLLLQA...HTGTLGLLTLOPEQKFQKVKGFGG...VVLTDPEAAKYVHGIHAVH YLDFLAPAKATLGE...IYYLHVVGVWTDDWN...HGYS...HTYLNWRQQ&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>variation</td>
<td>56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/allele</td>
<td>&quot;A&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/allele</td>
<td>&quot;T&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/db_xref</td>
<td>&quot;dbSNP:1059731&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feature</td>
<td>Value</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Biosource</strong></td>
<td>/organism=&quot;Homo_sapiens&quot; /db_xref=&quot;taxon:9606&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>/chromosome=&quot;1&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Reading frame</strong></td>
<td>/gene=&quot;GBA&quot; /note=&quot;GLUC&quot; /db_xref=&quot;LocusID:2629&quot; /db_xref=&quot;MIM:230800&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Protein id</strong></td>
<td>/product=&quot;glucosidase, beta; acid (includes glucosylceramidase)&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>/protein_id=&quot;XP_034490.1&quot; /db_xref=&quot;GI:14780920&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
|              | /translation="MAGSLTGLLLQAVSWAGARPSCPFSFYGSSVVCVCNATYCDSDFPPTFPALGF
                TFSLYPESTGLGSMPLMIRQANHTGTGGLLTLQPQKFQKVKGFGG
                AMTDAALNLIASPPAQNLKKYSYFSEEYGNIIARPMASCDFSISRTYTYADTPDD
                FQLHNFLPSEEDTKLKIPLHIALQLAQRPVSSLASPWTSPTWKLKTNGAVNGKGS
                LKQFGDIYHQMWARYFVKFLDAYAEHKLQFWAVTAEENPSAGLLSGYFQCLGLFTP
                EQRDFIARDLGFTLANNTHHNLVRLNMLDDQRLLPHWAKWVLTDPEAAKYVHGI
                AVHWYLDFLAPAKATLGETHRLEPNTMLFASEACVGSKFWEQSVRLGSSWDRG
                MQYSHSITNLLYHVVGWTDWNLALNPQGPNWWRNFVDSPIPIVDTDFYKQPMF
                FYHLGHFSKIFPEGSSQRVGLVASQKNLDAVAMHPDGSAVVVVNLRSSKDV
                VPLTIDPAPWETFISPQYSIHTVLWRQ" |
<p>| <strong>Coding Sequence (proteins)</strong> | /variation=&quot;56&quot; /allele=&quot;A&quot; /allele=&quot;T&quot; /db_xref=&quot;dbSNP:1059731&quot; |</p>
<table>
<thead>
<tr>
<th>OMIM Number</th>
<th>Description</th>
</tr>
</thead>
</table>
| #230800     | Gaucher Disease, Type I  
GLUCOSIDASE, ACID BETA, INCLUDED; GBA, INCLUDED  
Gene map locus 1q21 |
| #231000     | Gaucher Disease, Type III  
GAUCHER DISEASE, NORRBOTTNIAN TYPE, INCLUDED |
| #230900     | Gaucher Disease, Type II |
| #231005     | Gaucher-Like Disease |
| #176801     | Prosaosin; PSAP  
Saposin A, INCLUDED  
Gene map locus 10q22.1 |

Related Entries, Genome, Nucleotide, Protein, PubMed
*230800
GAUCHER DISEASE, TYPE I

Alternative titles; symbols

GD I
GAUCHER DISEASE, NONCEREBRAL JUVENILE
GLUCOCEREBROSIDASE DEFICIENCY
ACID BETA-GLUCOSIDASE DEFICIENCY
GBA DEFICIENCY
GLUCOSIDASE, ACID BETA, INCLUDED; GBA, INCLUDED
GLUCOCEREBROSIDASE PSEUDOGENE, INCLUDED; GBAP, INCLUDED

Gene map locus 1q21

TEXT

The cardinal features of type I Gaucher disease are hematologic abnormalities with hypersplenism, bone lesions, skin pigmentation, and pingueculae (brown spots of Gaucher cells at corneoscleral limbus). The disorder is particularly frequent in Ashkenazi Jews. The several forms of Gaucher disease are cerebrosides lipidoses. The disease has been diagnosed as early as the first week of life and as late as 86 years. Although the disorder is clearly autosomal recessive in most cases, a dominant form was suggested by some observations.
The cardinal features of type I Gaucher disease are hematologic abnormalities with hypersplenism, bone lesions, skin pigmentation, and pingueculae (brown spots of Gaucher cells at corneoscleral limbus). The disorder is particularly frequent in Ashkenazi Jews. The several forms of Gaucher disease are cerabrosode lipidoses. The disease has been diagnosed as early as the first week of life and as late as 86 years. Although the disorder is clearly autosomal recessive in most cases, a dominant form was suggested by Hsia et al. (1959) on the basis of affected father and son. The father was German-Jewish and the mother Swedish-English. Even here, the mother may have been a carrier and this quasi-dominant mechanism is even more likely in reports of presumed dominant inheritance in Jewish groups where the frequency of the Gaucher gene is relatively high. Classification into types I, II (230900), and III (231000) was proposed by Knudson and Kaplan (1962). An instructive pedigree was reported by Herrlin and Hillborg (1962). Serum acid phosphatase (which, unlike the prostatic enzyme, is not inhibited by L-tartrate) is elevated. Choy (1985) found that serum acid phosphatase was elevated only in those patients with bone involvement. The elevation was due to isozyme type 5 of osteoclastic origin. Acid phosphatase was normal in the lymphocytes and cultured fibroblasts and was normal in the serum of all heterozygotes. Thus, it is apparently a secondary feature of the disease and unreliable for diagnosis. Wiedemann et al. (1965) found typical Gaucher cells in the bone marrow of 2 clinically normal parents and a normal sister of 2 affected children and in the 2 clinically normal parents and 2 normal sisters of