Bioinformatics: Deduction of knowledge by computer analysis of biological data.

biological data + computers + algorithms

Organizing knowledge: databanks (UK) and databases (US and elsewhere)

Protein sequence analysis

Sequence alignment
Multiple alignment and sequence profiles
Phylogenetic trees
Bioinformatics key areas

Sequence analysis

- Organization of knowledge (sequences, structures, functional data)
- e.g. homology searches

Structural bioinformatics

Databases
Introduction to Databases

What Is a Database?

• A *database* is a computerized archive used to store and organize data in such a way that information can be retrieved easily via a variety of search criteria.

• *Databases* are composed of computer hardware and software for data management.

• The chief objective of the development of a database is to organize data in a set of structured records to enable easy retrieval of information.

• A computer program (*database management system, DBMS*) can manage and query the database to get answers to questions. It is a set of tools that stores, extracts and modifies data.

• *Schema* is the description of data in terms of a data model.
Types of Databases

- **Hierarchical**: organizes data in a tree structure. There is a hierarchy of parent and child data segments. It has only one root record. Each root record may participate in relationship with many child records. Each child may itself have many child records. A parent record “owns” its child records.

- **Network**: collection of record types. Record types are associated together by links and there is no constraint on the number and direction of the links that can be established. There is no root record and each record can participate in any number of “owns” relationships.

- **Object-Oriented**: add database functionality to object programming languages.

- **Relational**
  - Data is presented as a collection of relations
  - Each relation is depicted as a table
  - Columns are attributes
  - Rows ("tuples") represent entities
  - Every table has a set of attributes that taken together as a "key"
Why Relations?
• Very simple model.
• Often matches how we think about data.
• Abstract model that underlies SQL, the most important database language today.

Relational Data Model
• Composed of tables
• Row
  - Number column
  - Primary key
    Reference data in the table
    A column or set of columns in table contains unique data
<table>
<thead>
<tr>
<th>number</th>
<th>name</th>
<th>department</th>
<th>salary</th>
<th>location</th>
</tr>
</thead>
<tbody>
<tr>
<td>23603</td>
<td>Jones</td>
<td>413</td>
<td>1100</td>
<td>New Jersey</td>
</tr>
<tr>
<td>24568</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Primary key

Column
• Primary key uniquely identifies each row
  – Rule of Entity Integrity
• Composite primary key
• Lines connecting tables
  – Relationships
    • One-to-many relationship
• Foreign key
  – Join multiple tables
  – Rule of Referential Integrity
Examples

One Patient can take many Medications

Patient ID
First
Last
Address
City
State
Zip

Medications
Med ID
Medication
Description

Patient Meds
PM ID
Patient ID
Med ID
Dosage
Directions

One Kind of Medication can be taken by Many Patients
Hierarchical Database Model

Disadvantages:
- difficult to handle child entries that are initially unrelated to parent
- example:
  - Want to add a new entertainer, but cannot add entertainer until it is associated with an Agent
  - Have to insert a “dummy” record
- does not support complex relationships
  - redundant data: many to many relationship between Clients and Entertainers, so Schedule (date, time) will have same entries as Engagements (date, time)
Relational Database Model

Advantages:
- built in multilevel data integrity:
  - table: records are not duplicated, and detect missing primary keys
  - relationship: ensure connection between tables is valid
- logical and physical data independence from database application:
  - changes in physical implementation of database would not (in theory) adversely affect applications built to use DB
- easy data retrieval: data may be retrieved from any table or from any number of related tables
- Relationships
  - may be: one-to-one, one-to-many, and many-to-many
  - Established through matching values of a shared field
    - Example: Agents are associated to Clients by an AgentsID field in Clients table

- Data may be accessed in an almost unlimited combination of ways
The theory of relational database operations are based on set theory.

The **Relational Algebra** provides a collection of operations to manipulate relations. It supports the notion of a *query*, or request to retrieve information from a database. There are *set operations*: Union, Intersection, Difference, and Cartesian Product.

The **Relational Calculus** is a formal query language. Instead of having to write a sequence of relational algebra operations, we simply write a single declarative expression, describing the results that we want. It is somewhat like writing a program in C or Java instead of assembler.
Structured Query Language (SQL)

It came from an IBM Research project entitled "SEQUEL" where the intent was to create a structured English-like query language to interface to the early System R database system. Along with QUEL, SQL was the first high level declarative database language.

Data Definition Language (DDL): allows a database administrator or database designer to define tables, create views, etc.

Data Manipulation Language (DML): allows an end user to retrieve information from tables.
```
CREATE DATABASE (NAME)

The CREATE TABLE statement is used to define a new table. CREATE TABLE Students (sid CHAR(20),
name CHAR(30),
login CHAR(20),
age INTEGER
gpa REAL)
```

<table>
<thead>
<tr>
<th>sid</th>
<th>name</th>
<th>login</th>
<th>age</th>
<th>gpa</th>
</tr>
</thead>
<tbody>
<tr>
<td>50000</td>
<td>Dave</td>
<td>dave@cs</td>
<td>19</td>
<td>3.3</td>
</tr>
<tr>
<td>53666</td>
<td>Jones</td>
<td>jones@cs</td>
<td>18</td>
<td>3.4</td>
</tr>
<tr>
<td>53688</td>
<td>Smith</td>
<td>smith@chem</td>
<td>18</td>
<td>3.2</td>
</tr>
<tr>
<td>53650</td>
<td>Smith</td>
<td>smith@ee</td>
<td>19</td>
<td>3.8</td>
</tr>
<tr>
<td>53831</td>
<td>Madayan</td>
<td>madayan@music</td>
<td>11</td>
<td>1.8</td>
</tr>
<tr>
<td>53832</td>
<td>Guldu</td>
<td>guldu@math</td>
<td>12</td>
<td>2.0</td>
</tr>
</tbody>
</table>

An instance S1 if the Students Relation

An instance of a relation is a set of tuples, also called records
Tuples are inserted using the `INSERT` command. We can insert a single tuple into the Students table as follows:

```
INSERT INTO Students (sid, name, login, age, gpa) VALUES (53688, 'Smith', 'smith@chem', 18, 3.2)
```

We can delete tuples using the `DELETE` command. We can delete all Students tuples with name equal to Smith using the command.

```
DELETE FROM Students S WHERE S.name = 'Smith'
```

We can modify the column values in an existing row using the `UPDATE` command. For example we can increment the age and decrement the gpa of the student with sid 53688.

```
UPDATE Students S
SET S.age = S.age + 1, S.gpa = S.gpa - 1
WHERE S.id = 53688
```
The **WHERE** clause is applied first and determines which rows are to be modified. The **SET** clause then determines how these rows are to be modified. If the column that is being modified is also used to determine the new value, the value used on this expression on the right side of equals ( =) is the old value, that is before the modification. To illustrate these points further, consider,

```
UPDATE Students S
SET S.gpa = S.gpa – 0.1
WHERE S.gpa >= 3.3
```
A relational database query is a question about the data and the answer consists of a new relation containing the result. A query language is a specialized language for writing queries.
Consider the following three tables

<table>
<thead>
<tr>
<th>sid</th>
<th>sname</th>
<th>rating</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>Dustin</td>
<td>7</td>
<td>45.0</td>
</tr>
<tr>
<td>29</td>
<td>Brutus</td>
<td>1</td>
<td>33.0</td>
</tr>
<tr>
<td>031</td>
<td>Lubber</td>
<td>8</td>
<td>55.5</td>
</tr>
<tr>
<td>32</td>
<td>Andy</td>
<td>8</td>
<td>25.5</td>
</tr>
<tr>
<td>58</td>
<td>Rusty</td>
<td>10</td>
<td>35.0</td>
</tr>
<tr>
<td>64</td>
<td>Horatio</td>
<td>7</td>
<td>35.0</td>
</tr>
<tr>
<td>71</td>
<td>Zorba</td>
<td>10</td>
<td>16.0</td>
</tr>
<tr>
<td>74</td>
<td>Horatio</td>
<td>9</td>
<td>35.0</td>
</tr>
<tr>
<td>85</td>
<td>Art</td>
<td>3</td>
<td>25.5</td>
</tr>
<tr>
<td>95</td>
<td>Bob</td>
<td>3</td>
<td>63.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sid</th>
<th>bid</th>
<th>day</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>101</td>
<td>10/10/98</td>
</tr>
<tr>
<td>22</td>
<td>102</td>
<td>10/10/98</td>
</tr>
<tr>
<td>22</td>
<td>103</td>
<td>10/8/98</td>
</tr>
<tr>
<td>22</td>
<td>104</td>
<td>10/7/98</td>
</tr>
<tr>
<td>31</td>
<td>102</td>
<td>11/10/98</td>
</tr>
<tr>
<td>31</td>
<td>103</td>
<td>11/6/98</td>
</tr>
<tr>
<td>31</td>
<td>104</td>
<td>11/12/98</td>
</tr>
<tr>
<td>64</td>
<td>101</td>
<td>9/5/98</td>
</tr>
<tr>
<td>64</td>
<td>102</td>
<td>9/8/98</td>
</tr>
<tr>
<td>74</td>
<td>103</td>
<td>9/8/98</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>bid</th>
<th>bname</th>
<th>color</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>Interlake</td>
<td>blue</td>
</tr>
<tr>
<td>102</td>
<td>Interlake</td>
<td>red</td>
</tr>
<tr>
<td>103</td>
<td>Clipper</td>
<td>green</td>
</tr>
<tr>
<td>104</td>
<td>Marine</td>
<td>red</td>
</tr>
</tbody>
</table>

**More Queries**

```
SELECT * 
FROM Students S 
WHERE S.age < 18 
SELECT S.name, S.login 
FROM Students S, Enrolled E 
WHERE S.sid = E.sid AND E.grade = 'A'
```
1) Find the names and ages of all sailors

```
SELECT DISTINCT S.sname, S.age
FROM Sailors. S
```

Without DISTINCT

<table>
<thead>
<tr>
<th>sname</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dustin</td>
<td>45.0</td>
</tr>
<tr>
<td>Brutus</td>
<td>33.0</td>
</tr>
<tr>
<td>Lubber</td>
<td>55.5</td>
</tr>
<tr>
<td>Andy</td>
<td>25.5</td>
</tr>
<tr>
<td>Rusty</td>
<td>35.0</td>
</tr>
<tr>
<td>Horatio</td>
<td>35.0</td>
</tr>
<tr>
<td>Zorba</td>
<td>16.0</td>
</tr>
<tr>
<td>Horatio</td>
<td>35.0</td>
</tr>
<tr>
<td>Art</td>
<td>25.5</td>
</tr>
<tr>
<td>Bob</td>
<td>63.5</td>
</tr>
</tbody>
</table>

With DISTINCT

<table>
<thead>
<tr>
<th>sname</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dustin</td>
<td>45.0</td>
</tr>
<tr>
<td>Brutus</td>
<td>33.0</td>
</tr>
<tr>
<td>Lubber</td>
<td>55.5</td>
</tr>
<tr>
<td>Andy</td>
<td>25.5</td>
</tr>
<tr>
<td>Rusty</td>
<td>35.0</td>
</tr>
<tr>
<td>Horatio</td>
<td>35.0</td>
</tr>
<tr>
<td>Zorba</td>
<td>16.0</td>
</tr>
<tr>
<td>Art</td>
<td>25.5</td>
</tr>
<tr>
<td>Bob</td>
<td>63.5</td>
</tr>
</tbody>
</table>
2) Find all sailors with rating above 7
   SELECT   S.sid, S.sname, S.age
   FROM     Sailors. AS S
   WHERE    S.rating > 7

3) Find the names of sailors who have reserved boat number 103
   SELECT   S.sname
   FROM     Sailors.S, Reserves.R
   WHERE    S.sid = R.sid AND R.bid = 103

4) Find the sids of sailors who have reserved a red boat
   SELECT   R.sid
   FROM     Boats B, Reserves R
   WHERE    B.sid = R.sid AND B.color = 'red'

5) Find the names of sailors who have reserved a red boat
   SELECT   S.sname
   FROM     Sailors.S, Reserves.R, Boats B
   WHERE    S.sid = R.sid AND R.bid = B.bid AND B.color = 'red'

6) Find the colors of boats reserved by Lubber
   SELECT   B.color
   FROM     Sailors.S, Reserves.R, Boats B
   WHERE    S.sid = R.sid AND R.bid = B.bid AND S.sname = 'Lubber'

7) Find the names of boats who have reserved at least one boat
   SELECT   S.sname
   FROM     Sailors.S, Reserves.R
   WHERE    S.sid = R.sid
Nested Queries

Query 3 above, *find the names of sailors who have reserved boat number 103* can be written also as:

```
SELECT   S.sname
FROM     Sailors.S
WHERE    S.sid  IN ( SELECT  R.sid
                      FROM    Reserves R
                      WHERE   R.bid = 103 )
```

Query 5 above, *find the names of sailors who have reserved a red boat* can be written also as:

```
SELECT   S.sname
FROM     Sailors.S
WHERE    S.sid  IN ( SELECT  R.sid
                      FROM    Reserves R
                      WHERE   R.bid  IN  ( SELECT  B.bid
                                            FROM      Boats B
                                            WHERE   B.color = 'red' )
                      )
```

Find the names of sailors who have not reserved a red boat

```
SELECT   S.sname
FROM     Sailors.S
WHERE    S.sid  NOT IN ( SELECT  R.sid
                          FROM    Reserves R
                          WHERE   R.bid  IN  ( SELECT  B.bid
                                                FROM      Boats B
                                                WHERE   B.color = 'red' )
                          )
```
MySQL

- Multi-user and multi-threaded RDBMS server
- Uses SQL to interact with and manipulate data
- Supports various programming languages
- Access tables from different databases
- Handle large databases

Perl Database Interface

- Access relational databases from Perl programs
- Database independent
- Handles
  - Driver handles
  - Database handles
  - Statement handles
Web Resources

- www.sql.org
- www.mysql.com
- www.microsoft.com/sql
- www.microsoft.com/sql/downloads/default.asp
- www.postgresql.org
- www.interbase.com
- www.maverick-dbms.org
- www.devshed.com
- www.cql.com
- leap.sourceforge.net
- www.voicenet.com/~gray/Home.html
- www.w3school.com/sql
- www.sqlmag.com
The Data

- Information stored in the genetic code (DNA)
- Protein sequences
- 3D structures
- Experimental results from various sources
- Patient statistics
- Scientific literature
The challenge of the information space:

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleotides</td>
<td>44,575,745,176</td>
</tr>
<tr>
<td>Nucleotide records</td>
<td>49,127,925</td>
</tr>
<tr>
<td>Protein sequences</td>
<td>5,785,962</td>
</tr>
<tr>
<td>3D structures in PDB</td>
<td>28,905</td>
</tr>
<tr>
<td>BIND Interactions</td>
<td>134,886</td>
</tr>
<tr>
<td>Human Unigene Cluster</td>
<td>52,888</td>
</tr>
<tr>
<td>Completed Genome project</td>
<td>238</td>
</tr>
<tr>
<td>Different taxonomy Nodes</td>
<td>249,219</td>
</tr>
<tr>
<td>dbSNP records</td>
<td>18,883,945</td>
</tr>
<tr>
<td>RefSeq Genomic records:</td>
<td>180,770</td>
</tr>
<tr>
<td>RefSeq RNA Records:</td>
<td>352,275</td>
</tr>
<tr>
<td>RefSeq Protein Records:</td>
<td>1,310,899</td>
</tr>
<tr>
<td>GenSAT images</td>
<td>98,680</td>
</tr>
<tr>
<td>GEO profiles</td>
<td>11,288,275</td>
</tr>
<tr>
<td>Homologene gene</td>
<td>38,137</td>
</tr>
<tr>
<td>PubChem compounds</td>
<td>897,246</td>
</tr>
<tr>
<td>PubMed records</td>
<td>15,382,675</td>
</tr>
<tr>
<td>PubMed Central records</td>
<td>341,602</td>
</tr>
<tr>
<td>OMIM records</td>
<td>16,521</td>
</tr>
</tbody>
</table>

Feb 11th 2005
Algorithmic Developments

Important part of research in bioinformatics:
methods for
data storage
data retrieval
data analysis
Interdisciplinary Research

- Rapidly developing branch of biology
- Highly interdisciplinary: using techniques and concepts from informatics, statistics, mathematics, chemistry, biochemistry, physics, and linguistics.
- Many practical applications in biology and medicine.
Computations in Biology...

Similar to other sciences:

- computational physics, computational chemistry
- derivation of physics laws from astronomical data

Biologists always wanted to derive knowledge by induction

Reasons for recent development:

- development of computers and networks
- availability of data (sequences, 3D structures)
- amount of data
Why?

An avalanche of data:

- Sequences
- Function related
- Structures

requires computational approaches
Structural Bioinformatics

Sequence

Structure

Function

?
Structural Bioinformatics

• Prediction of structure from sequence
  – secondary structure
  – homology modelling, threading
  – ab initio 3D prediction
• Analysis of 3D structure
  – structure comparison/ alignment
  – prediction of function from structure
  – molecular mechanics/ molecular dynamics
  – prediction of molecular interactions, docking
• Structure databases (RCSB)
Organizing knowledge in databases

- Sequence databanks and databases
  - EMBL, SwissProt, TREMBL

- SRS: Sequence Retrieval system

- 3D structure database: the RCSB - PDB

- Domain databases
Biological Databases

• Very fast growth of biological data
• Diversity of biological data:
  – primary sequences
  – 3D structures
  – functional data
• Database entry usually required for publication
  – Sequences
  – Structures
• Database entry may replace primary publication
  – genomic approaches
DNA Sequence Databases

- Three databanks exchange data on a daily basis
- Data can be submitted and accessed at either location

- Genebank

- EMBL
  - www.ebi.ac.uk/embl/index.html

- DNA DataBank of Japan (DDBJ)
  - www.nig.ac.jp/home.html
There are three major public DNA databases:

- **EMBL**: Housed at EBI, European Bioinformatics Institute
- **GenBank**: Housed at NCBI, National Center for Biotechnology Information
- **DDBJ**: Housed in Japan
The storage of next-generation sequence data is an emerging challenge.

NCBI offers a sequence read archive (SRA), but the best storage strategies are uncertain. http://www.ncbi.nlm.nih.gov/Traces/sra/

The European Bioinformatics Institute (EBI) offers the European Nucleotide Archive (ENA), http://www.ebi.ac.uk/ena/

For individual labs, tens of terabytes of storage are routinely needed.
A project to sequence 10,000 human genomes requires about 3 petabytes of storage (3,000 terabytes). To buy a server with 100 Tb now costs $10,000.
EMBL Database Growth
Taxonomy at NCBI:
>250,000 species are represented in GenBank

<table>
<thead>
<tr>
<th>Ranks:</th>
<th>higher taxa</th>
<th>genus</th>
<th>species</th>
<th>lower taxa</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archaea</td>
<td>115</td>
<td>128</td>
<td>457</td>
<td>0</td>
<td>700</td>
</tr>
<tr>
<td>Bacteria</td>
<td>1209</td>
<td>2267</td>
<td>11207</td>
<td>768</td>
<td>15451</td>
</tr>
<tr>
<td>Eukaryota</td>
<td>18758</td>
<td>59473</td>
<td>248454</td>
<td>18774</td>
<td>345459</td>
</tr>
<tr>
<td>Fungi</td>
<td>1354</td>
<td>4093</td>
<td>25262</td>
<td>970</td>
<td>31679</td>
</tr>
<tr>
<td>Metazoa</td>
<td>13631</td>
<td>38819</td>
<td>116258</td>
<td>9466</td>
<td>178174</td>
</tr>
<tr>
<td>Viridiplantae</td>
<td>2290</td>
<td>14014</td>
<td>98767</td>
<td>8106</td>
<td>123177</td>
</tr>
<tr>
<td>Viruses</td>
<td>571</td>
<td>378</td>
<td>1973</td>
<td>0</td>
<td>2922</td>
</tr>
<tr>
<td>All taxa</td>
<td>20679</td>
<td>62254</td>
<td>262124</td>
<td>19542</td>
<td>364599</td>
</tr>
</tbody>
</table>
Distribution of entries

- Homo sapiens: 51%
- All Other Organisms: 15%
- Mus musculus: 4%
- Drosophila melanogaster: 2%
- Arabidopsis thaliana: 2%
- Caenorhabditis elegans: 2%
- Tetraodon nigroviridis: 1%
- Oryza sativa: 1%
- Rattus norvegicus: 1%
- Pan troglodytes: 1%
- Danio rerio: 1%
Notice: Upcoming Systems Maintenance

NCBI servers will undergo maintenance beginning **November 13 at 3:00 PM** and lasting until **November 14 at 8:00 PM (EST)**. During this time, NCBI services, including BLAST and the Entrez search and retrieval systems, will be operational but may be intermittently slow. Data submission systems, including BankIt, GenBank FTP submissions, SequinMacroSend, and the SRA, Trace and GEO submission systems, will be unavailable. Please contact NCBI with concerns: info@ncbi.nlm.nih.gov.
NCBI key features d PubMed

• National Library of Medicine's search service
• 22 million citations in MEDLINE (as of 2012)
• links to participating online journals
• PubMed tutorial on the site
  or visit NLM:
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
Accession numbers are labels for sequences

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

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

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

Examples (all for beta globin, \textit{HBB}):  

- X02775: GenBank genomic DNA sequence  
- NG\_000007.3: RefSeqGene  
- rs192792910: dbSNP (single nucleotide polymorphism)  
- AA970968.1: An expressed sequence tag (1 of 2,345)  
- NM\_000518.4: RefSeq DNA sequence (from a transcript)  
- NP\_000509.1: RefSeq protein  
- CAA00182.1: GenBank protein  
- Q14473: SwissProt protein  
- 1YE0|B: Protein Data Bank structure record
NCBI’s important RefSeq project: best representative sequences

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

RefSeq identifiers include the following formats:

- Complete genome: NC_########
- Complete chromosome: NC_########
- Genomic contig: NT_########
- mRNA (DNA format): NM_######## e.g. NM_000518
- Protein: NP_######## e.g. NP_000509
NCBI’s RefSeq project: many accession number formats for genomic, mRNA, protein sequences

<table>
<thead>
<tr>
<th>Accession</th>
<th>Molecule</th>
<th>Method</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC_123456</td>
<td>Genomic</td>
<td>Mixed</td>
<td>Alternate complete genomic</td>
</tr>
<tr>
<td>AP_123456</td>
<td>Protein</td>
<td>Mixed</td>
<td>Protein products; alternate</td>
</tr>
<tr>
<td>NC_123456</td>
<td>Genomic</td>
<td>Mixed</td>
<td>Complete genomic molecules</td>
</tr>
<tr>
<td>NG_123456</td>
<td>Genomic</td>
<td>Mixed</td>
<td>Incomplete genomic regions</td>
</tr>
<tr>
<td>NM_123456</td>
<td>mRNA</td>
<td>Mixed</td>
<td>Transcript products; mRNA</td>
</tr>
<tr>
<td>NM_123456789</td>
<td>mRNA</td>
<td>Mixed</td>
<td>Transcript products; 9-digit</td>
</tr>
<tr>
<td>NP_123456</td>
<td>Protein</td>
<td>Mixed</td>
<td>Protein products;</td>
</tr>
<tr>
<td>NP_123456789</td>
<td>Protein</td>
<td>Curation</td>
<td>Protein products; 9-digit</td>
</tr>
<tr>
<td>NR_123456</td>
<td>RNA</td>
<td>Mixed</td>
<td>Non-coding transcripts</td>
</tr>
<tr>
<td>NT_123456</td>
<td>Genomic</td>
<td>Automated</td>
<td>Genomic assemblies</td>
</tr>
<tr>
<td>NW_123456</td>
<td>Genomic</td>
<td>Automated</td>
<td>Genomic assemblies</td>
</tr>
<tr>
<td>NZ_ABCD12345678</td>
<td>Genomic</td>
<td>Automated</td>
<td>Whole genome shotgun data</td>
</tr>
<tr>
<td>XM_123456</td>
<td>mRNA</td>
<td>Automated</td>
<td>Transcript products</td>
</tr>
<tr>
<td>XP_123456</td>
<td>Protein</td>
<td>Automated</td>
<td>Protein products</td>
</tr>
<tr>
<td>XR_123456</td>
<td>RNA</td>
<td>Automated</td>
<td>Transcript products</td>
</tr>
<tr>
<td>YP_123456</td>
<td>Protein</td>
<td>Auto. &amp; Curated</td>
<td>Protein products</td>
</tr>
<tr>
<td>ZP_123456</td>
<td>Protein</td>
<td>Automated</td>
<td>Protein products</td>
</tr>
</tbody>
</table>
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_000518.4 for beta globin DNA corresponding to mRNA) or protein (NP_000509.1)
From the NCBI home page, type “beta globin”
<table>
<thead>
<tr>
<th>Database Type</th>
<th>Description</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>PubMed</td>
<td>biomedical literature citations and abstracts</td>
<td>8403</td>
</tr>
<tr>
<td>PubMed Central</td>
<td>free, full text journal articles</td>
<td>15047</td>
</tr>
<tr>
<td>Site Search</td>
<td>NCBI web and FTP sites</td>
<td>31</td>
</tr>
<tr>
<td>Nucleotide</td>
<td>core subset of nucleotide sequence records</td>
<td>2805</td>
</tr>
<tr>
<td>EST</td>
<td>Expressed Sequence Tag records</td>
<td>2033</td>
</tr>
<tr>
<td>GSS</td>
<td>Genome Survey Sequence records</td>
<td>3</td>
</tr>
<tr>
<td>Protein</td>
<td>sequence database</td>
<td>2078</td>
</tr>
<tr>
<td>Genome</td>
<td>whole genome sequences</td>
<td>621</td>
</tr>
<tr>
<td>Structure</td>
<td>three-dimensional macromolecular structures</td>
<td>353</td>
</tr>
<tr>
<td>Taxonomy</td>
<td>organisms in GenBank</td>
<td>697</td>
</tr>
<tr>
<td>SNP</td>
<td>short genetic variations</td>
<td>125</td>
</tr>
<tr>
<td>dbVar</td>
<td>Genomic structural variation</td>
<td>108</td>
</tr>
<tr>
<td>Gene</td>
<td>gene-centered information</td>
<td>12</td>
</tr>
<tr>
<td>GEO Profiles</td>
<td>expression and molecular abundance profiles</td>
<td>3160</td>
</tr>
<tr>
<td>GEO DataSets</td>
<td>experimental sets of GEO data</td>
<td>93</td>
</tr>
<tr>
<td>Epigenomics</td>
<td>Epigenetic maps and data</td>
<td>23</td>
</tr>
</tbody>
</table>

Follow the link to “Gene”
Entrez Gene is in the header
Note the “Official Symbol” HBB for beta globin
Note the “limits” option
HBB hemoglobin, beta [Homo sapiens]

Gene ID: 3043, updated on 6-Nov-2011

Summary

- **Official Symbol**: HBB provided by HGNC
- **Official Full Name**: hemoglobin, beta provided by HGNC
- **Primary source**: HGNC:4827
- **See related**: Ensembl:ENSG00000244734; HPRD:00786; MIM:141900
- **Gene type**: protein coding
- **RefSeq status**: REVIEWED
- **Organism**: Homo sapiens
- **Lineage**: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
- **Also known as**: CD113t-C; beta-globin

**Summary**
The alpha (HBA) and beta (HBB) loci determine the structure of the 2 types of polypeptide chains in adult hemoglobin, Hb A. The normal adult hemoglobin tetramer consists of two alpha chains and two beta chains. Mutant beta globin causes sickle cell anemia. Absence of beta chain causes beta-zero-thalassemia. Reduced amounts of detectable beta globin causes beta-plus-thalassemia. The order of the genes in the beta-globin cluster is 5'-epsilon--gamma-G--gamma-A--delta--beta--3'. [provided by RefSeq, Jul 2008]
“Gene” page at NCBI offers a wealth of information

- Genomic context
- Bibliography
- Phenotypes
- Gene Ontology (organizing principles of biological process, molecular function, cellular component)
- Reference sequences
- Additional (non-RefSeq sequences)
- Many, many links to NCBI resources (e.g. HomoloGene)
- Many, many links to external resources
Entrez Gene (bottom of page): non-RefSeq accessions (it’s unclear what these are, highlighting usefulness of RefSeq)

<table>
<thead>
<tr>
<th>Genomic</th>
<th>Accession 1</th>
<th>Accession 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M56640.1</td>
<td>AAA52634.1</td>
</tr>
<tr>
<td>Genomic</td>
<td>SB2767.1</td>
<td>AAD14420.1</td>
</tr>
<tr>
<td></td>
<td>U01317.1</td>
<td>AAA16334.1</td>
</tr>
<tr>
<td></td>
<td>U01317.1</td>
<td>AAA16335.1</td>
</tr>
<tr>
<td>Genomic</td>
<td>U20223.1</td>
<td>AAB60348.1</td>
</tr>
<tr>
<td>Genomic</td>
<td>U20498.1</td>
<td>CAA23757.1</td>
</tr>
<tr>
<td>Genomic</td>
<td>U20498.1</td>
<td>CAA23753.1</td>
</tr>
<tr>
<td>mRNA</td>
<td>AF117710.1</td>
<td>AAD19695.1</td>
</tr>
<tr>
<td>mRNA</td>
<td>AF181832.1</td>
<td>AAF00488.1</td>
</tr>
<tr>
<td>mRNA</td>
<td>AF181989.1</td>
<td>AAF00489.1</td>
</tr>
<tr>
<td>mRNA</td>
<td>AF349114.1</td>
<td>AAK29639.1</td>
</tr>
<tr>
<td>mRNA</td>
<td>AK311625.1</td>
<td>BAG34767.1</td>
</tr>
<tr>
<td>mRNA</td>
<td>AV138510.1</td>
<td>AAN11320.1</td>
</tr>
<tr>
<td>mRNA</td>
<td>AV502993.1</td>
<td>AAF96398.1</td>
</tr>
<tr>
<td>mRNA</td>
<td>BCO07075.1</td>
<td>AAR07075.1</td>
</tr>
<tr>
<td>mRNA</td>
<td>CR536530.1</td>
<td>CAC38757.1</td>
</tr>
<tr>
<td>mRNA</td>
<td>CR541013.1</td>
<td>CAG46711.1</td>
</tr>
<tr>
<td>mRNA</td>
<td>CR509430.1</td>
<td>None</td>
</tr>
<tr>
<td>mRNA</td>
<td>CR509430.1</td>
<td>None</td>
</tr>
<tr>
<td>mRNA</td>
<td>CR509430.1</td>
<td>None</td>
</tr>
<tr>
<td>mRNA</td>
<td>CR509430.1</td>
<td>None</td>
</tr>
<tr>
<td>mRNA</td>
<td>CR509430.1</td>
<td>None</td>
</tr>
<tr>
<td>mRNA</td>
<td>EU594342.1</td>
<td>ACD39343.1</td>
</tr>
<tr>
<td>mRNA</td>
<td>M11420.1</td>
<td>AAA52633.1</td>
</tr>
<tr>
<td>mRNA</td>
<td>MC5079.1</td>
<td>AAA35597.1</td>
</tr>
<tr>
<td>mRNA</td>
<td>MC5079.1</td>
<td>AAA35595.1</td>
</tr>
<tr>
<td>mRNA</td>
<td>Y00497.1</td>
<td>CAA23755.1</td>
</tr>
<tr>
<td>mRNA</td>
<td>Y00497.1</td>
<td>CAA23754.1</td>
</tr>
<tr>
<td>Synthetic</td>
<td>AM392537.1</td>
<td>CAL37415.1</td>
</tr>
<tr>
<td>Synthetic</td>
<td>AM393951.1</td>
<td>CAL38229.1</td>
</tr>
<tr>
<td>Synthetic</td>
<td>DDB93159.2</td>
<td>APN84095.1</td>
</tr>
<tr>
<td>Synthetic</td>
<td>DDB96453.2</td>
<td>APN87452.1</td>
</tr>
<tr>
<td>Synthetic</td>
<td>EU176774.1</td>
<td>ABW03575.1</td>
</tr>
</tbody>
</table>
## hemoglobin subunit beta [Homo sapiens]

**NCBI Reference Sequence:** NP_000509.1

### Goto

<table>
<thead>
<tr>
<th>LOCUS</th>
<th>NP_000509</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEFINITION</td>
<td>hemoglobin subunit beta [Homo sapiens].</td>
</tr>
<tr>
<td>ACCESSION</td>
<td>NP_000509</td>
</tr>
<tr>
<td>VERSION</td>
<td>NP_000509.1 GI:4504349</td>
</tr>
<tr>
<td>DESCSEQ</td>
<td>accession NM_000518.4</td>
</tr>
<tr>
<td>KEYWORDS</td>
<td>.</td>
</tr>
<tr>
<td>SOURCE</td>
<td>Homo sapiens (human)</td>
</tr>
<tr>
<td>ORGANISM</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>LTR</td>
<td>1 aa</td>
</tr>
<tr>
<td>TITLE</td>
<td>T-quaternary structure of oxy human adult hemoglobin in the presence of two allosteric effectors, L35 and IHP</td>
</tr>
<tr>
<td>PUBLMED</td>
<td>21703224</td>
</tr>
<tr>
<td>REMARK</td>
<td>GeneRIF: Data indicate that the fluctuation of the tertiary structure of Hb seems to be caused by both the structural perturbation of alpha(1)beta(1) (or alpha(2)beta(2)) intra-dimeric interface.</td>
</tr>
</tbody>
</table>

### Reference

| AUTHORS    | Kanori,K., Tajiri,Y., Tsuneshige,A., Ishigami,I., Ogura,T., Tajima,K., Neya,S. and Yonetsui,T. |
| PUBLMED    | 21703224 |
| REMARK     | GeneRIF: Data indicate that the fluctuation of the tertiary structure of Hb seems to be caused by both the structural perturbation of alpha(1)beta(1) (or alpha(2)beta(2)) intra-dimeric interface. |

### Additional Information

- **Articles about the HBB gene**
  - Molecular characteristics of three hemoglobin variants observed in a Chir [Mol Med Rep. 2]
  - Mechanism of escape from nonsense-mediated mRNA decay of human beta-globin tra [RNA. 2]
  - The association between intragenic SNP haplotypes and mutant [Blood Cells Mol Dis. 2]

- **Identical proteins for NP_000509.1**
  - Sequence 1352 from patent US 799869d [AEN15]
  - Sequence 7 from patent US 7989993 [JAE184]
  - PREDICTED: hemoglobin subunit [KP_003312]

- **Pathways for the HBB gene**
  - African trypanosomiasis
  - Malaria
  - Folate Metabolism
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 mwhltpeeks avtalwggkvn vdevggealg rllvvyputq rffesfgdls tpdavmgnpk
61 vkahgkkvgl afsglahld nlkgftacs elhcdklhvd penfr11gnv lvcvlahhfg
121 keftppvqaa yqkvavgyan alahkyh
```
//
You should learn the one-letter amino acid code!

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

NCBI Reference Sequence: NP_000509.1

FASTA  Graphics

Go to:

LOCUS NP_000509  147 aa  linear  PRI 05-NOV-2010
DEFINITION hemoglobin subunit beta [Homo sapiens].
ACCESSION NP_000509
VERSION NP_000509.1  GI:4504349
DBSOURCE REFSEQ: accession NM_000518.4
KEYWORDS
SOURCE Homo sapiens (human)
ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchothoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.
REFERENCE 1  (residues 1 to 147)
CONSRTM DREAM investigators
TITLE Variation at the NFATC2 locus increases the risk of thiazolidinedione-induced edema in the Diabetes REDuction
FASTA format:
versatile, compact with one header line
followed by a string of nucleotides or amino acids
in the single letter code

hemoglobin subunit beta [Homo sapiens]
NCBI Reference Sequence: NP_000509.1

>gi|4504349|ref|NP_000509.1|  hemoglobin subunit beta [Homo sapiens]
MVHLPPEEKSATLUGKVNVDEVGGEALGRLLLIVVPWTQRFŞESFGDSLSTPDAVNGPKVKAHGKKVLG
AFSGDLAHLYNKLGRFTATLSELHCDKLVHPENFRLLGNVLCVLAHHFGKEFTTPVQAAAYQKVVAGVAN
ALAHKYH
While FASTA is one file format, there are many others

<table>
<thead>
<tr>
<th>Format</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>FASTA</td>
<td>Sequences in one letter DNA or protein code</td>
</tr>
<tr>
<td>FASTQ</td>
<td>DNA sequences with quality scores for each base</td>
</tr>
<tr>
<td>BAM</td>
<td>compressed binary version of SAM</td>
</tr>
<tr>
<td>SAM</td>
<td>Sequence Alignment/Map file (tab-delimited)</td>
</tr>
<tr>
<td>VCF</td>
<td>variant call format (genomic variants; indels)</td>
</tr>
</tbody>
</table>

(See genome.ucsc.edu/FAQ/FAQformat.html for the following:)

<table>
<thead>
<tr>
<th>Format</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BED</td>
<td>a table including chromosome, start, end</td>
</tr>
<tr>
<td>WIG</td>
<td>wiggle format (displays dense, continuous data)</td>
</tr>
<tr>
<td>GFF</td>
<td>General Feature Format (tab separated)</td>
</tr>
</tbody>
</table>

Also, besides Excel (.xls, .xlsx) spreadsheets can also be:

<table>
<thead>
<tr>
<th>Format</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>.txt</td>
<td>tab-delimited text file (or space delimited)</td>
</tr>
<tr>
<td>.csv</td>
<td>comma separated text file</td>
</tr>
</tbody>
</table>
**FASTA format**

one line header, starting with >
some programs require several characters without space after >

```
>sp|043057|MSS4_SCHPO Guanine nucleotide exchange factor MSS4 homolog - Schizo
  MSNLRIVCQHCPSVVFNNKRPDVVKRPTMSAMLHSETQEDLETDDFFLLKDPMFAFDNVS
  SKPLANNYKLLACADCEKGPLGYDKNNEYLLLLCSLEKN
```
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 less commonly used.
Human (Homo sapiens)

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 genewise alignments of the human proteome as well as from alignments of human cDNAs using the cDNA2genome model of exonerate

This release of the assembly has the following properties:

- 27478 contigs.
- contig length total 3.2 Gb.
- chromosome length total 3.1 Gb.

It also includes nine haplotypic regions, mainly in the MHC region of chromosome 6.

To convert your old data from Human assembly NCBI36 to GRCh37, click on 'Manage your data' on any human page and select 'Assembly converter' from the left-hand menu.

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 Havana. This refined merge set is now the public output of the GENCODE project. The set displayed in release 56 corresponds to GENCODE release 3c.

The major genome browsers have come together to produce a common set of identifiers where CDS annotations of transcripts can be agreed. This undertaking, the Consensus Coding Sequence (CCDS) project was initially based on the NCBI36 assembly. The CCDS identifiers have been mapped onto the new annotations based on the latest GRCh37 assembly and these identifiers are also shown.
Ensembl output for beta globin includes views of chromosome 11 (top), the region (middle), and a detailed view (bottom).

There are various horizontal annotation tracks.
The UCSC Genome Browser: 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
Note that there are choices of assemblies such as hg19

An assembly (or “build”) is a fixed version of a genome. Builds are released every several years. In practice, you should always be aware whether you are using hg18 or hg19 for the human genome. They are annotated with different types of information such as experimental data sets.

To learn more visit http://www.ncbi.nlm.nih.gov/assembly/basics/
Choose the RefSeq beta globin gene (HBB)

**UCSC Genes**

- **HBB (uc001mce.1) at chr11:5246696-5248301** - Homo sapiens hemoglobin, beta (HBB), mRNA.
- **HBD (uc001maf.1) at chr11:5254059-5255858** - Homo sapiens hemoglobin, delta (HBD), mRNA.
- **RB17 (uc010gav.2) at chr10:6131309-6159422** - Homo sapiens RNA binding motif protein 17 (RB17), transcript variant 2, mRNA.
- **RB17 (uc0011j1b.3) at chr10:6130949-6159422** - Homo sapiens RNA binding motif protein 17 (RB17), transcript variant 1, mRNA.
- **HBA1 (uc002cfk.1) at chr16:226679-227520** - Homo sapiens hemoglobin, alpha 1 (HBA1), mRNA.
- **HBA2 (uc002cfv.4) at chr16:222846-223709** - Homo sapiens hemoglobin, alpha 2 (HBA2), mRNA.
- **HBBP1 (uc001mag.3) at chr11:5263185-5264822** - Homo sapiens hemoglobin, beta pseudogene 1 (HBBP1), non-coding RNA.
- **TMEM158 (uc011baf.2) at chr3:45265956-45267814** - Homo sapiens transmembrane protein 158 (gene/pseudogene) (TMEM158), mRNA.

**RefSeq Genes**

- **HBB at chr11:5246696-5248301** - (NM_000518) hemoglobin subunit beta
- **HBBP1 at chr11:5263185-5264822** - (NR_001589)

**Non-Human RefSeq Genes**

- **hbb at chr11:5247810-5275746** - (NM_001201019) hemoglobin subunit beta
- **HBB at chr11:5246810-5248260** - (NM_001164428) hemoglobin subunit beta
- **HBD at chr11:5246820-5248251** - (NM_001164047) hemoglobin subunit beta
- **HBB at chr11:5246828-5248251** - (NM_001164018) hemoglobin subunit beta
- **HBB at chr11:5254194-5255663** - (NM_001164018) hemoglobin subunit beta
- **HBB at chr11:5246828-5248301** - (NM_001144841) hemoglobin subunit beta
[4] On the UCSC Genome Browser:
--choose which tracks to display
--add custom tracks
--the Table Browser is complementary
Exploring the UCSC Genome Browser

The human genome can be viewed with different “assemblies” (hg18, hg19). These contain different data sets.

- You can get information about a track by clicking its header (e.g. “RefSeq Genes”).

- You can choose the density to display each track (e.g. hide, dense, squish, pack).
Types of files in GenBank

• From one-gene investigators
  – Often a very well annotated cDNA
  – A genomic segment from a new invertebrate
  – A mitochondria or virus

• From population/phylogenetic analysis

• From Genome Centers:
  – Gene expression:
    • Expressed Sequence Tags
    • Full Length Insert cDNA
  – Genome sequencing projects
## Organismal Divisions

**Used in which database?**

<table>
<thead>
<tr>
<th>Organism</th>
<th>Description</th>
<th>Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCT</td>
<td>Bacterial</td>
<td>DDBJ - GenBank</td>
</tr>
<tr>
<td>FUN</td>
<td>Fungal</td>
<td>EMBL</td>
</tr>
<tr>
<td>HUM</td>
<td>Homo sapiens</td>
<td>DDBJ - EMBL</td>
</tr>
<tr>
<td>INV</td>
<td>Invertebrate</td>
<td>all</td>
</tr>
<tr>
<td>MAM</td>
<td>Other mammalian</td>
<td>all</td>
</tr>
<tr>
<td>ORG</td>
<td>Organelle</td>
<td>EMBL</td>
</tr>
<tr>
<td>PHG</td>
<td>Phage</td>
<td>all</td>
</tr>
<tr>
<td>PLN</td>
<td>Plant</td>
<td>all</td>
</tr>
<tr>
<td>PRI</td>
<td>Primate (also see HUM)</td>
<td>all (not same data in all)</td>
</tr>
<tr>
<td>PRO</td>
<td>Prokaryotic</td>
<td>EMBL</td>
</tr>
<tr>
<td>ROD</td>
<td>Rodent</td>
<td>all</td>
</tr>
<tr>
<td>SYN</td>
<td>Synthetic and chimeric</td>
<td>all</td>
</tr>
<tr>
<td>VRL</td>
<td>Viral</td>
<td>all</td>
</tr>
<tr>
<td>VRT</td>
<td>Other vertebrate</td>
<td>all</td>
</tr>
</tbody>
</table>
Functional Divisions

**PAT**  Patent
**EST**  Expressed Sequence Tags
**STS**  Sequence Tagged Site
**GSS**  Genome Survey Sequence
**HTG**  High Throughput Genome (unfinished)
**HTC**  High throughput cDNA (unfinished)
**CON**  Contig assembly instructions
**ENV**  Environmental sampling methods

Organismal divisions:

**BCT**  **FUN**  **INV**  **MAM**  **PHG**  **PLN**
**PRI**  **ROD**  **SYN**  **VRL**  **VRT**
Guiding Principals

In GenBank, records are grouped for various reasons: understand this is key to using and fully taking advantage of this database.
Identifiers

• You need identifiers which are stable through time
• Need identifiers which will always refer to specific sequences
• Need these identifiers to track history of sequence updates
• Also need feature and annotation identifiers
LOCUS, Accession, NID and protein_id

**LOCUS**: Unique string of 10 letters and numbers in the database. Not maintained amongst databases, and is therefore a poor sequence identifier.

**ACCESSION**: A unique identifier to that record, citable entity; does not change when record is updated. A good record identifier, ideal for citation in publication.

**VERSION**: New system where the accession and version play the same function as the accession and gi number.

**Nucleotide gi**: Geninfo identifier (gi), a unique integer which will change every time the sequence changes.

**Protein gi**: Geninfo identifier (gi), a unique integer which will change every time the sequence changes.

**protein_id**: Identifier which has the same structure and function as the nucleotide Accession.version numbers, but slightly different format.
<table>
<thead>
<tr>
<th>LOCUS</th>
<th>HSU40282</th>
<th>1789 bp</th>
<th>mRNA</th>
<th>PRI</th>
<th>21-MAY-1998</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEFINITION</td>
<td>Homo sapiens integrin-linked kinase (ILK) mRNA, complete cds.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACCESSION</td>
<td>U40282</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VERSION</td>
<td>U40282.1</td>
<td>GI:3150001</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**CDS**

- **gene**: "ILK"
- **note**: "protein serine/threonine kinase"
- **codon_start**: 1
- **product**: "integrin-linked kinase"
- **protein_id**: "AAC16892.1"
- **db_xref**: "GI:3150002"
EMBL entry for insulin receptor

ID    HSINSR24    standard; DNA; PRI; 873 BP.
AC    M32972;
DT    10-JUL-1990 (Rel. 24, Created)
DT    05-SEP-1992 (Rel. 33, Last updated, Version 4)
DE    Human insulin receptor (hINSR) gene, exon 22.
KW    insulin receptor.
OS    Homo sapiens (human)
OC    Eukaryota; Animalia; Metazoa; Chordata; Vertebrata; Mammalia;
OC    Theria; Eutheria; Primates; Haplorhini; Catarrhini; Hominidae.
RN    [1]
RP    1-873
RA    Seino S., Seino M., Bell G.I.;
RT    "Human insulin-receptor gene";
CC    Draft entry and computer-readable sequence for [1] kindly submitted
CC    by G.I.Bell, 14-MAR-1990.
FH    Key Location/Qualifiers
FH
<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>source</td>
<td>/note=&quot;human insulin receptor precursor&quot;</td>
</tr>
<tr>
<td>prim_transcript</td>
<td>/note=&quot;human insulin receptor&quot;</td>
</tr>
<tr>
<td>intron</td>
<td>/note=&quot;hINSR mRNA and introns&quot;</td>
</tr>
<tr>
<td></td>
<td>/note=&quot;hINSR intron U&quot;</td>
</tr>
</tbody>
</table>
EMBL entry 3: sequence

SQ  Sequence 873 BP; 199 A; 217 C; 234 G; 223 T; 0 other;
gacgtgtcct tctgccccgc agcactgacc tcatgcgcct catgtggcaa ttcaacccca  120
agatgaggcc aaccttcctg gagattgctca acctgctcaa ggacgacctg cacccccagct  180
tttcagaggt gtctttccttc cacagcgagg agaacaaggg tccccgagagt gaggagctgg  240
...
ccccccacccg cccccagcag atggaaagaa agcacctgtt tttacaaatt ctttttttttt  780
ttttttttttt ttttttttttg cttgtgtcttg agcttcagta taaaagacaa aacttcctgt  840
tgtgagaaca aatattcggaa agaaaaaacc aaa  873
//
What is UniProt?

UniProt is a new protein sequence database that is the result of a merge from SWISS-PROT and PIR (Protein Information Resource) and is in great part funded by the NIH. It is the main distributed, annotated, and curated protein sequence database. Data in UniProt is primarily derived from coding sequence annotations in EMBL (GenBank/DDBJ) nucleic acid sequence data, but also from sequences in PIR and SP. UniProt is a Flat-File database just like EMBL and SwissProt

- http://www.pir.uniprot.org/
- Bairoch et al., The Universal Protein Resource (UniProt) Nucl. Acids Res. 2005 33: D154-D159
<table>
<thead>
<tr>
<th>Your Query Result Sets (Page - 1) [Data Set Manager]</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP31373</td>
</tr>
<tr>
<td>1 entry</td>
</tr>
</tbody>
</table>

Protein CYS3_YEAST

New Query | Download Protein | Bookmark Protein (Ctrl+D)

<table>
<thead>
<tr>
<th>Basic</th>
<th>Extended</th>
<th>Viewers: Flat File</th>
<th>XML</th>
<th>ExPASy</th>
<th>SRS</th>
<th>PIR</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;UniProt/Swiss-Prot</td>
<td>P31373</td>
<td>CYS3_YEAST Cystathionine gamma-lyase</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TLQESDKFATKAIHAGEHVDVHGSVIEPSLSTTFKQSSPAHPIGTYEYSRSQHPNRENL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ERAVALENQAQYGLAFSGSATTATILQSLPQ6SHAVSIGDYGGTHRYFTKVANNHGVE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TSFTNDLLNLPLQIMKENTKLVWIETPNPKTVTDIQKVAIDLIKQGAIQDVILVVDNT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLSPISNPLNFGDIVVSATKINGHDVVLGLATANPKLYEQLQLQNAIGAIPSP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FDAMLTHRGLKTLHRLVRQAALSANKTAEFLAADDKVAVNYPLKTPNYDVLKQHR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DALGGGMTSFRIKGGAARSKFASSTRFTLAESSLQGIESLLEVPRAVMTHGIPKEAREA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SGVFDDLVRISVQIETDDILEDIKQLKQATN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Pop-Up Fasta View

<table>
<thead>
<tr>
<th>Uniparc</th>
<th>UniRef100</th>
<th>UniRef90</th>
<th>UniRef50</th>
</tr>
</thead>
</table>
**Flat File Protein Viewer**

<table>
<thead>
<tr>
<th>Your Query Result Sets (Page - 1) [DataSet Manager]</th>
</tr>
</thead>
<tbody>
<tr>
<td>~ IHP31373</td>
</tr>
<tr>
<td>lenmy</td>
</tr>
</tbody>
</table>

**Protein CYS3_YEAST**

New Query | Download Protein | Bookmark Protein (Ctrl+D)

<table>
<thead>
<tr>
<th>Basic</th>
<th>Extended</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>CYS3_YEAST</td>
</tr>
<tr>
<td>AC</td>
<td>P31373</td>
</tr>
<tr>
<td>DT</td>
<td>01-JUL-1993 (Rev. 26, Created)</td>
</tr>
<tr>
<td>DT</td>
<td>01-JUL-1993 (Rev. 26, Last sequence update)</td>
</tr>
<tr>
<td>DT</td>
<td>25-JAN-2005 (Rev. 46, Last annotation update)</td>
</tr>
<tr>
<td>DE</td>
<td>Cystathionine gamma-lyase (EC 4.4.1.1) (Gamma-cystathionase).</td>
</tr>
<tr>
<td>QT</td>
<td>Name: CYS3; Synonyms: CY31, STR1; OrderedLocusNames=XNL012W;</td>
</tr>
<tr>
<td>GN</td>
<td>ORFNames=FUBF35;</td>
</tr>
<tr>
<td>OS</td>
<td>Saccharomyces cerevisiae (Baker's yeast).</td>
</tr>
<tr>
<td>UC</td>
<td>Eukaryota; Fungi; Ascomycota: Saccharomycotina; Saccharomyces;</td>
</tr>
<tr>
<td>DC</td>
<td>Saccharomycetales: Saccharomycetaceae: Saccharomyces.</td>
</tr>
<tr>
<td>NCBI TaxID</td>
<td>38292;</td>
</tr>
<tr>
<td>PN</td>
<td>14</td>
</tr>
<tr>
<td>RP</td>
<td>NUCLEOTIDE SEQUENCE, AND PARTIAL PROTEIN SEQUENCE.</td>
</tr>
<tr>
<td>RX</td>
<td>MEDLINE: 92950430; PubMed: 1577698;</td>
</tr>
<tr>
<td>RA</td>
<td>Uno B.-I., Tanaka K., Naito K., H histo C., Shimoda S., Yamamoto S.;</td>
</tr>
<tr>
<td>RA</td>
<td>Ohmori S., Ushina I., Toh E A.;</td>
</tr>
<tr>
<td>RT</td>
<td>&quot;Cloning and characterization of the CYS3 (CY31) gene of Saccharomyces cerevisiae.&quot;</td>
</tr>
<tr>
<td>RL</td>
<td>J. Bacteriol. 174:3339-3347(1992);</td>
</tr>
<tr>
<td>RN</td>
<td>2;</td>
</tr>
<tr>
<td>RP</td>
<td>NUCLEOTIDE SEQUENCE, AND CHARACTERIZATION.</td>
</tr>
<tr>
<td>DC</td>
<td>STAND-009139;</td>
</tr>
<tr>
<td>RX</td>
<td>MEDLINE: 93328680; PubMed: 8135636;</td>
</tr>
</tbody>
</table>
**Swiss-Prot**

**ID** CY3_YEAST STANDARD; PRT; 393 AA.

**AC** P31373;

**DT** 01-JUL-1993 (REL. 26, CREATED)

**DE** CYSTATHIONINE GAMMA-LYASE (EC 4.4.1.1) (GAMMA-CYSTATHIONASE).

**GN** CY3 OR CYI1 OR STR1 OR YAL012W OR FUN35.

**OS** SACCHAROMYCETACEAE; SACCHAROMYCIES.

**RX** CATALOGIC ACTIVITY: L-CYSTATHIONINE + H(2)O = L-CYSTEINE + NH(3) + 2-OXOBUTANOATE.

**CC** -!- COFACTOR: PYRIDOXAL PHOSPHATE.

**CC** -!- PATHWAY: FINAL STEP IN THE TRANS-SULFURATION PATHWAY SYNTHESIZING L-CYSTEINE FROM L-METHIONINE.

**CC** -!- SUBUNIT: HOMOTETRAMER.

**CC** -!- SUBCELLULAR LOCATION: CYTOPLASMIC.

**CC** -!- SIMILARITY: BELongs TO THE TRANS-SULFURATION ENZYMES FAMILY.

**CC** DISCLAMOR

**DR** DATABASE cross-reference

**KW** CYSTEINE BIOSYNTHESIS; HYDROLYASE; PYRIDOXAL PHOSPHATE.

**FT** INIT_MET 0 0

**FT** BINDING 203 203 PYRIDOXAL PHOSPHATE (BY SIMILARITY).

**SQ** SEQUENCE 393 AA; 42411 MW; 55BA2771 CRC32; TLQESDDKFA KAIHAGEHVD VHGHSIEPIS LSTTFKQSSP ANPITYEY KSNQPNRENL ERAWALENA QYGLFSSSQA ATTTAILQSL PGHSVSAIGV DYGGHTRYF TKVANAHGVE TSFTNDLND LPQIKENTK LWIETPTNP TLKVTIDQKV ADLKKHAG QDVILVVDNT FLPSIYNSPL NFGADIVHS ATKYINGHS VVGLVLATN KFLYERQQL QNAIGAIPSP FDALWTHRG KLTLHRVQA ALSANKIAEF LAADKENVA VNYQGLKTHP NYDVVLQHR DLAGGGMSF RIKGGGAES KFASSTRFL LAESLGGIES LLEVPAVMTH GGIPKEAREA SGVFDLVRI SVGIEDTDDL LEDIKQALKQ ATN

**RA** OHMORI S., OHMORI S., OSHIMA T., TOH-E A.;

**RP** SEQUENCE FROM N.A.; PARTIAL SEQUENCE.

**RN** [1]

**RT** “Cloning and bacterial expression of the CYS3 gene encoding cystathionine gamma-lyase of Saccharomyces cerevisiae and the rat gamma-cystathionase and Escherichia coli cystathionine gamma-synthase enzymes.”


**RA** YAMAGATA S., D'ANDREA R.J., FUJISAKI S., ISAJI M., NAKAMURA K.;

**RP** SEQUENCE FROM N.A., AND PARTIAL SEQUENCE.

**RN** [2]

**RT** “Cystathionine gamma-lyase of Saccharomyces cerevisiae and the rat gamma-cystathionase and Escherichia coli cystathionine gamma-synthase enzymes.”


**RA** ONO B.-I., TANAKA K., NAITO K., HEIKE C., SHINODA S., YAMAMOTO S.;

**RP** SEQUENCE FROM N.A., AND PARTIAL SEQUENCE.

**RN** [3]

**RT** “Physical localization of yeast CYS3, a gene whose product resembles the rat gamma-cystathionase and Escherichia coli cystathionine gamma-synthase activity.”


**RA** OUELLETTE B.F.F., CLARK M.W., KENG T., STORMS R.K., ZHONG W.W.,

**RP** SEQUENCE FROM N.A., AND PARTIAL SEQUENCE.

**RN** [4]

**RT** “Catalytic and kinetic properties of Cys3p, the yeast cystathionine gamma-lyase.”


**RA** RALES R.H., OSHIMA T., TOH-E A.,

**RP** SEQUENCE FROM N.A., AND PARTIAL SEQUENCE.

**RN** [5]

**RT** “Catalytic and kinetic properties of Cys3p, the yeast cystathionine gamma-lyase.”


**RA** OHMORI S., OSHIMA T., TOH-E A.,

**RP** SEQUENCE FROM N.A., AND PARTIAL SEQUENCE.

**RN** [6]

**RT** “Physical localization of yeast CYS3, a gene whose product resembles the rat gamma-cystathionase and Escherichia coli cystathionine gamma-synthase enzymes.”


**RA** RALES R.H., OSHIMA T., TOH-E A.,

**RP** SEQUENCE FROM N.A., AND PARTIAL SEQUENCE.

**RN** [7]

**RT** “Catalytic and kinetic properties of Cys3p, the yeast cystathionine gamma-lyase.”


**RA** CHRISTENSEN J.,

**RP** SEQUENCE FROM N.A., AND PARTIAL SEQUENCE.

**RN** [8]

**RT** “Physical localization of yeast CYS3, a gene whose product resembles the rat gamma-cystathionase and Escherichia coli cystathionine gamma-synthase enzymes.”


**RA** FUXA D., OSHIMA T., TOH-E A.,

**RP** SEQUENCE FROM N.A., AND PARTIAL SEQUENCE.

**RN** [9]

**RT** “Catalytic and kinetic properties of Cys3p, the yeast cystathionine gamma-lyase.”


**RA** OHMORI S., OSHIMA T., TOH-E A.,

**RP** SEQUENCE FROM N.A., AND PARTIAL SEQUENCE.

**RN** [10]

**RT** “Physical localization of yeast CYS3, a gene whose product resembles the rat gamma-cystathionase and Escherichia coli cystathionine gamma-synthase enzymes.”


**RA** OHMORI S., OSHIMA T., TOH-E A.,

**RP** SEQUENCE FROM N.A., AND PARTIAL SEQUENCE.

**RN** [11]

**RT** “Physical localization of yeast CYS3, a gene whose product resembles the rat gamma-cystathionase and Escherichia coli cystathionine gamma-synthase enzymes.”


**RA** OHMORI S., OSHIMA T., TOH-E A.,

**RP** SEQUENCE FROM N.A., AND PARTIAL SEQUENCE.

**RN** [12]

**RT** “Physical localization of yeast CYS3, a gene whose product resembles the rat gamma-cystathionase and Escherichia coli cystathionine gamma-synthase enzymes.”


**RA** OHMORI S., OSHIMA T., TOH-E A.,

**RP** SEQUENCE FROM N.A., AND PARTIAL SEQUENCE.

**RN** [13]

**RT** “Physical localization of yeast CYS3, a gene whose product resembles the rat gamma-cystathionase and Escherichia coli cystathionine gamma-synthase enzymes.”


**RA** OHMORI S., OSHIMA T., TOH-E A.,

**RP** SEQUENCE FROM N.A., AND PARTIAL SEQUENCE.
UniProt

• UniProt incorporates:
  • Function of the protein
  • Post-translational modification
  • Domains and sites.
  • Secondary structure.
  • Quaternary structure.
  • Similarities to other proteins;
  • Diseases associated with deficiencies in the protein
  • Sequence conflicts, variants, etc.
The RCSB-PDB (www.rcsb.org/pdb)

- Data bases for 3D structures of biological macromolecules (proteins, nucleic acids)

- RCSB (Research Collaboratory for Structural Bioinformatics) maintains and develops the PDB (Protein Data Bank)

- others:
  - MMDB (EBI): msd.ebi.ac.uk
Results of a simple query

Your query found 2 structures in the current PDB release and you have selected 0 structures so far. You can select specific structures by clicking on the checkbox next to their id. If you do not select any structures, certain options will default to all structures. To examine an individual structure select the Explore link!

Pull down to select option: New Search

- **1FWQ**
  - Title: Solution Structure Of Human Mss4, A Guanine Nucleotide Exchange Factor For Rab Proteins
  - Classification: Metal Binding Protein
  - Compound: Mol_Id: 1; Molecule: Guanine Nucleotide Exchange Factor; Chain: A; Engineered: Yes
  - Deposited: 24-Sep-2000
  - Exp. Method: NMR, Minimized Average Structure

- **1HXR**
  - Title: Crystal Structure Of Mss4 At 1.65 Angstroms
  - Classification: Metal Binding Protein
  - Compound: Mol_Id: 1; Molecule: Guanine Nucleotide Exchange Factor Mss4; Chain: A, B; Engineered: Yes
  - Deposited: 16-Jan-2001
  - Exp. Method: X-ray Diffraction
  - Resolution: 1.65 Å

© RCSB
Summary Information

Title: Solution Structure Of Human Mss4, A Guanine Nucleotide Exchange Factor For Rab Proteins

Compound: Mol_Id: 1; Molecule: Guanine Nucleotide Exchange Factor; Chain: A; Engineered: Yes

Authors: H. Yu, S. L. Schreiber

Exp. Method: NMR, Minimized Average Structure

Classification: Metal Binding Protein

Source: Homo Sapiens


Deposition Date: 24-Sep-2000

Release Date: 04-Oct-2000

Polymer Chains: A

Residues: 123

Atoms: 911

HET groups:

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZN</td>
<td>ZINC ION</td>
<td>ZN&lt;sub&gt;1&lt;/sub&gt;</td>
</tr>
</tbody>
</table>

© RCSB
View structures

Structure Explorer - 1FWQ

Interactive 3D Display:
Choose from the following display options:

- VRML (default options): Interactive immersive ribbon diagram
- VRML (custom options, full screen display): Interactive immersive ribbon or cylinder diagram with ligands
- Rasmol
- MICE - Molecular Interactive Collaborative Environment (requires Java Plugin)
- FirstGlance (needs Chime)
- Protein Explorer (needs Chime)
- Sting Millennium (needs Chime and Java)

- Java (simple interactive sequence/structure/property backbone diagram):

Still Images:

- Ribbons (250x250)
- Cylinders (250x250)
- Ribbons (500x500)
- Cylinders (500x500)

Custom Size Images:
Links to Databases

at Institut Pasteur
  www.pasteur.fr/recherche/banques

at Infobiogen (Evry)
  www.infobiogen.fr/services/deambulum/fr

European bioinformatics institute (ebi)
  www.ebi.ac.uk/Databases/index.html

at the swiss institute for bioinformatics (SIB)
  www.expasy.org
  www.expasy.org/alinks.html#Proteins