Integrating Heterogeneous Biological Databases

Russ B. Altman

MIS 214/ CS 427

Copyright Russ B. Altman

An explosion of biological data

Biological data is usually stored in special purpose, focused collections of data:

- DNA sequence
- Protein sequence
- Protein structure
- Mutations leading to disease
- Sequence motifs associated with function
- Biological literature
- Genome organization information

In addition, there are “virtual databases” created by application programs that associated data with other data.

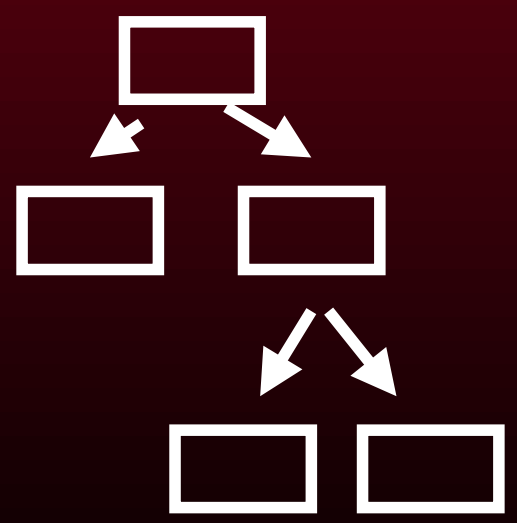
Copyright Russ B. Altman

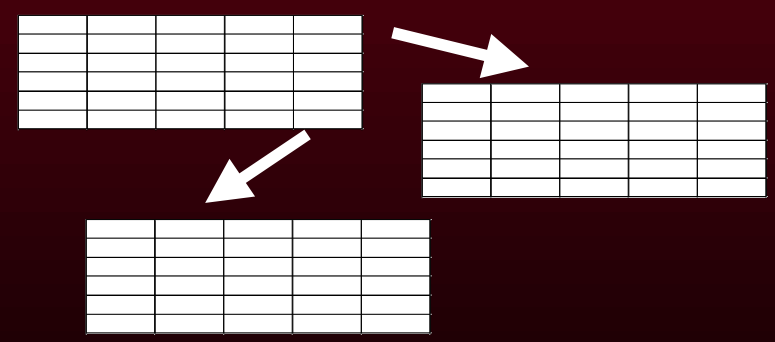
Methods for storing biological data

Biological data can be organized in many different manners:

```
LOCUS       SYNWHLMG     507 bp    DNA          SYN      15-MAR-1989
DEFINITION  Sperm whale synthetic myoglobin gene, complete cds.
ACCESSION   J03566
NID         g209563
KEYWORDS    myoglobin.
SOURCE      Synthetic DNA.
  ORGANISM  artificial sequence
            artificial sequence.
REFERENCE   1  (bases 1 to 507)
  AUTHORS   Springer,B.A. and Sligar,S.G.
  TITLE     High-level expression of sperm whale myoglobin in Escherichia coli
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 84, 8961-8965 (1987)
  MEDLINE   88097408
FEATURES             Location/Qualifiers
     source          1..507
                     /organism="artificial sequence"
                     /db_xref="taxon:29278"
     CDS             24..498
                     /note="synthetic myoglobin"
                     /codon_start=1
                     /db_xref="PID:g209564"
                     /transl_table=11
                     /translation="MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKF
                     DRFKHLKTEAEMKASEDLKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPI
                     KYLEFISEAIIHVLHSRHPGNFGADAQGAMNKALELFRKDIAAKYKELGYQG"
BASE COUNT     155 a    108 c    115 g    129 t
ORIGIN
        1 ctgcagataa ctaactaaag gagaacaaca acaatggttc tgictgaagg tgaatggcag
       61 ctggtictgc atgttgggc  taaagttgaa gctgacgtcg ctggtcatgg tcaggatatc
      121 ttgatitcgac tgttcaaatc tcatctggaa actctggaaa aattcgatcg ttitcaaatcat
      181 ctgaaaactg aagctgaaat gaaagcttct gaagatctga aaaaacatgg tgttaccgtg
      241 ttaactgccc taggtgctat ccttaagaaa aaagggcatc atgaagctga gctcaaaccg
      301 ctgcgcaat cgcatgctac taaacataag atccegatca aatacctgga attcatctct
      361 gaagcgatca tccatgtict gcatictaga catccaagta actcggtgc tgacgctcag
      421 ggtgctatga acaaagctct cgagctgttc cgtaaagata tcgctgctaa gtacaaagaa
      481 ctgggttacc agggttaatg aggtacc
```

1. Flat text files databases
2. Relational databases
3. Object oriented databases





Copyright Russ B. Altman

Flat Text File

- Entries are stored in text.
- Text fields/attributes are labelled with identifiers
- May be standard vocabulary used for values of attributes (or not)

- Search by searching for text
- Can be indexed for faster search
- Easy to import/export
- Not platform dependent
- Ubiquitous
- Hard to do complicated queries

Copyright Russ B. Altman

```
LOCUS       SYNWHLMG      507 bp    DNA             SYN       15-MAR-1989
DEFINITION  Sperm whale synthetic myoglobin gene, complete cds.
ACCESSION   J03566
NID         g209563
KEYWORDS    myoglobin.
SOURCE      Synthetic DNA.
  ORGANISM  artificial sequence
            artificial sequence.
REFERENCE   1  (bases 1 to 507)
  AUTHORS   Springer,B.A. and Sligar,S.G.
  TITLE     High-level expression of sperm whale myoglobin in Escherichia coli
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 84, 8961-8965 (1987)
  MEDLINE   88097408
FEATURES             Location/Qualifiers
     source          1..507
                     /organism="artificial sequence"
                     /db_xref="taxon:29278"
     CDS             34..498
                     /note="synthetic myoglobin"
                     /codon_start=1a
                     /transl_table=11
                     /db_xref="PID:g209564"
                     /translation="MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKF
                     DRFKHLKTEAEMKASEDLKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPI
                     KYLEFISEAIIHVLHSRHPGNFGADAQGAMNKALELFRKDIAAKYKELGYQG"
BASE COUNT      155 a    108 c    115 g    129 t
ORIGIN      5 bp upstream of PstI site.
        1 ctgcagataa ctaactaaag gagaacaaca acaatggttc tgtctgaagg tgaatggcag
       61 ctggttctgc atgtttgggc taaagttgaa gctgacgtcg ctggtcatgg tcaggacatc
      121 ttgattcgac tgttcaaatc tcatccggaa actctggaaa aattcgatcg tttcaaacat
      181 ctgaaaactg aagctgaaat gaaagcttct gaagatctga aaaaacatgg tgttaccgtg
      241 ttaactgccc taggtgctat ccttaagaaa aaagggcatc atgaagctga gctcaaaccg
      301 cttgcgcaat cgcatgctac taaacataag atcccgatca aatacctgga attcatctct
      361 gaagcgatca tccatgttct gcattctaga catccaggta acttcggtgc tgacgctcag
      421 ggtgctatga acaaagctct cgagctgttc cgtaaagata tcgctgctaa gtacaaagaa
      481 ctgggttacc agggttaatg aggtacc
//
```

Copyright Russ B. Altman

Structured Flat File

Use a standard syntax for facilitating automated reading of flat text files.

```
Publications =
{[title: string,
  authors: {‖[name: string,initial: string]‖},
  journal: <uncontrolled: string,
            controlled: <medline-jta: string,
                 % Medline journal title abbreviation
                        iso-jta: string,
                 % ISO journal title abbreviation
                        journal-title: string,
                 % Full journal title
                        issn: string>>
                 % ISSN number
  volume: string,
  issue: string,
  year: int,
  pages: string,
  abstract: string,
  keywd: {string}]}
```

Copyright Russ B. Altman

Relational Database

All data represented in tables, representing relations within the data.

R1(A|B,C,D)
R2(B|E,F)
etc...

The collection of tables represents the information represented in the database.

Easy to perform basic operations, since structure is very constrained.

sequences:

Column name, type, length, nulls?
seq_id , ddt_int_id , 4 , 0
seq_lab_symbol , varchar , 40 , 8
sequence , text , 16 , 0
sts_flag , ddt_flag_tiny , 1 , 0

Primary key(s) seq_id

seqd_dict:

Column name, type, length, nulls?
seqd_code , ddt_dict_char_code , 1 , 0
seqd_desc , ddt_dict_desc , 255 , 0

Primary key(s) seqd_code

Copyright Russ B. Altman

Object oriented databases

- Data organized into a hierarchy of concepts or *classes*.
- Each concept has a set of *attributes*, which can have typed values.
- Concepts can *inherit* values of attributes fromparents in the hierarchy.
- Can model richer set of relationships than the relational model…but queries are not as efficient for that reason.

Copyright Russ B. Altman

Class: Molecule
name:
molecular-weight:
type:

Class: Polymer
name:
molecular-weight:
type:
length:
sequence:
basic subunit:

Class: DNA
name:
molecular-weight:
type: nucleic-acid
length:
basic subunit: base
sequence:

Class: Protein
name:
molecular-weight:
type: polypeptide
length:
basic subunit: amino acid
sequence:

Copyright Russ B. Altman

Problem: performing queries across databases

Example: ***Find the DNA sequences associated with a 3D protein structure.***

Problem:

3D protein structure stored in PDB
DNA sequences stored in GENBANK

Example: ***Find the a mutation of a 3D protein structure known to cause disease.***

Problem:

3D protein structure stored in PDB
DNA sequences stored in GENBANK
Mutations in sequences stored in OMIM.

Copyright Russ B. Altman

Why is integration hard?

- vocabularies not shared different terms for same concept
- same term for different concepts different dependencies in the data
  - inheritance in object oriented systems implies relations
  - tables in relations imply logical connections
- queries are very different in the systems
  - text search
  - SQL
  - OOSQL
- different formats even when semantics are shared
- different semantics for basic concepts

Copyright Russ B. Altman

Creating integrated access to databases for performing queries.

Two strategies:

1. Consolidation
   - create single homogenous mega-DB
   - required DBs to use same tables, concepts

2. Federation
   - incorporate links from between DBs
   - couple the DBs loosely with common query language
   - construct a data warehouse

Copyright Russ B. Altman

Fundamental data integration issues

Integrating two databases fundamentally involves identifying information that is implicitly or explicitly shared, and can be used to create new relationships from existing ones.

Database X: R1(A|B,C)
Database Y: R2(B|D,E)

Allows new relation R3(A,B,C,D,E) to be created.

Copyright Russ B. Altman

Linking databases

To link two databases, need to

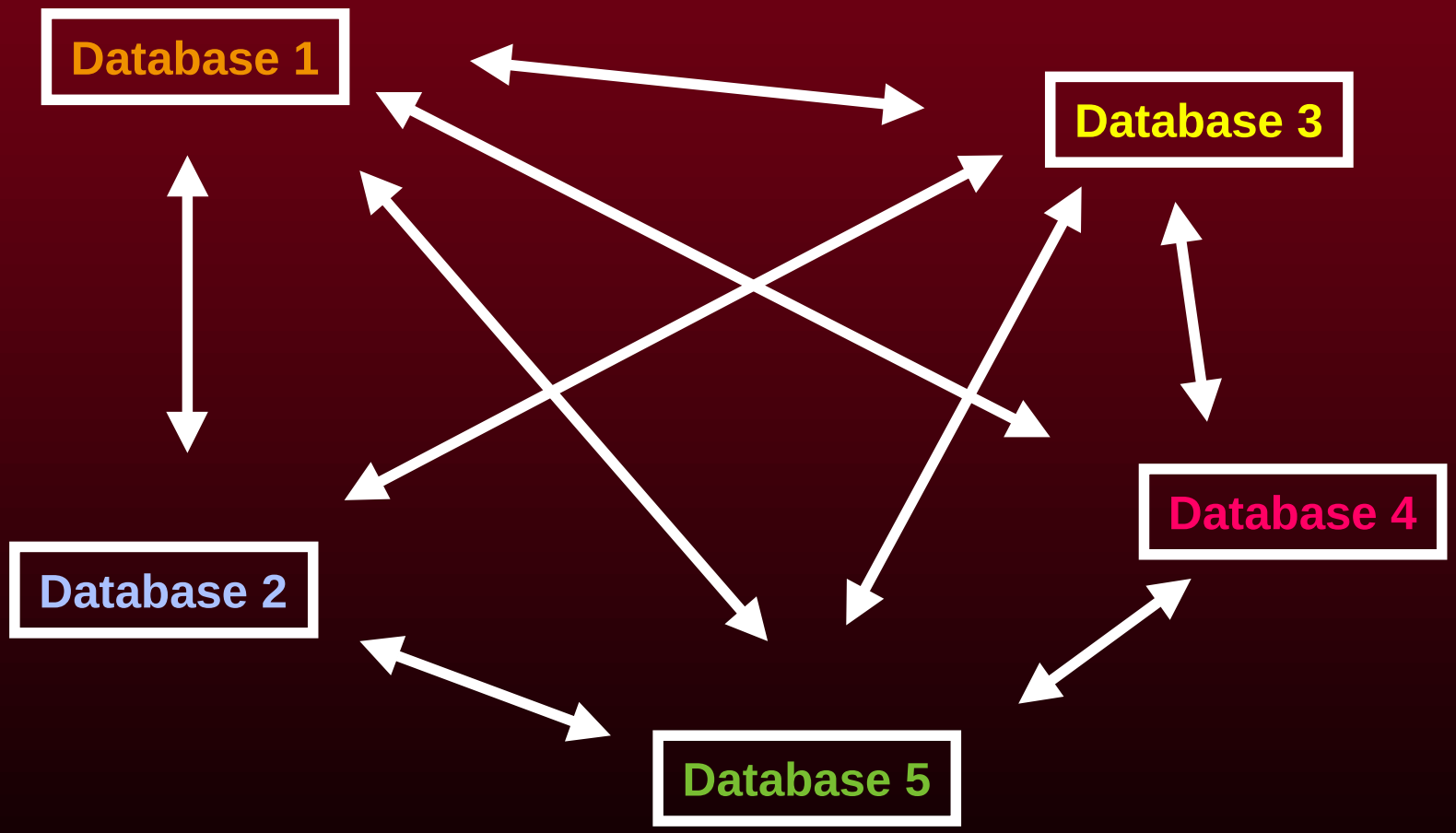
1. **Identify the critical shared data elements that allow relationships to be combined.**
2. **Need to be sure that semantics of the shared data is similar.**
3. **Need to create theasaurus for corresponding concepts.**

This creates a mapping from DB-1 to DB-2

For N databases $O(N^2)$ mappings needed?

Copyright Russ B. Altman

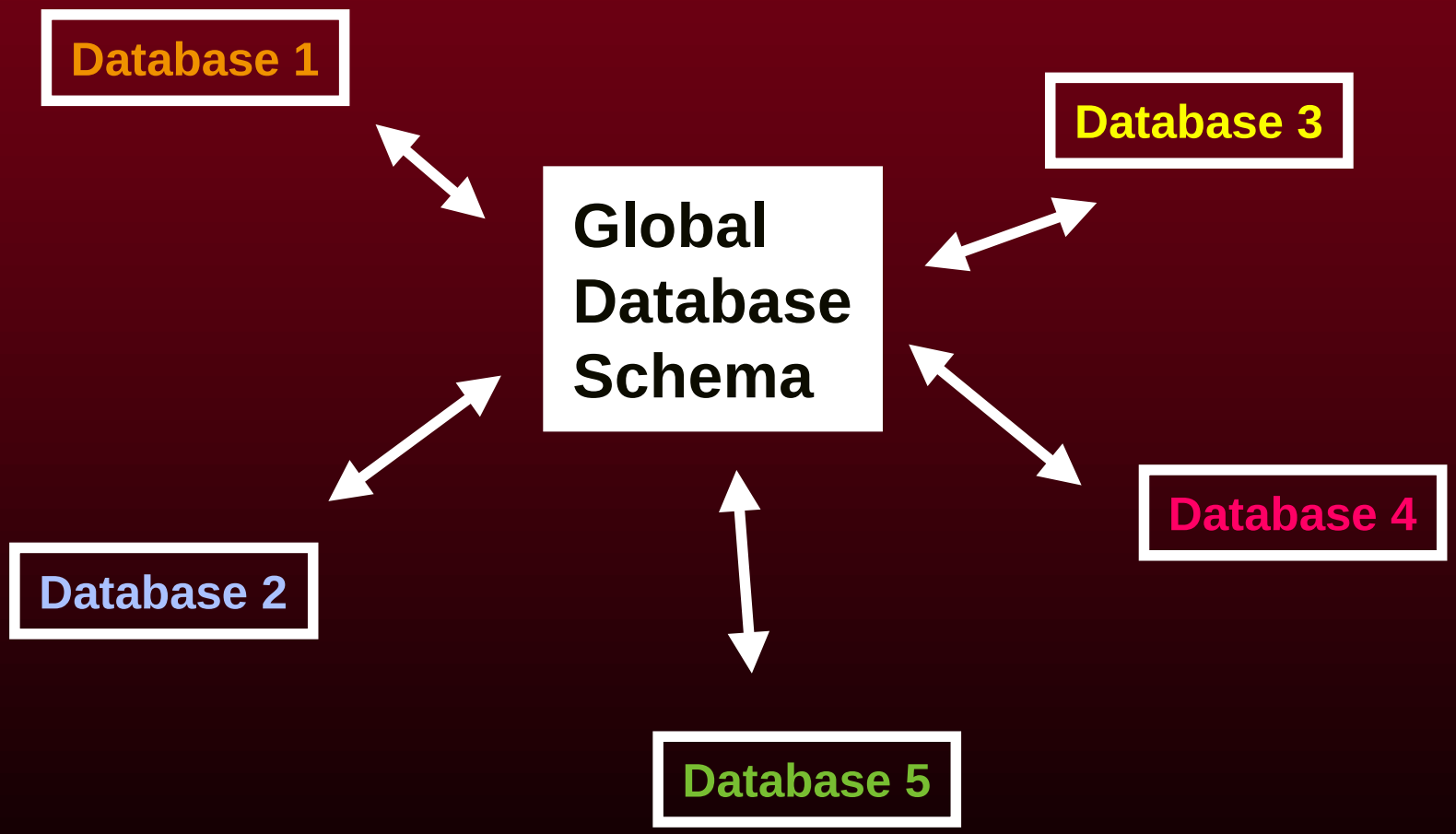
Integration requiring lots of mappings



Copyright Russ B. Altman

Global Schema

The $O(N^2)$ problem can be avoided by defining a single GLOBAL schema to which all databases can be mapped--only $O(N)$ mappings required.



Copyright Russ B. Altman

Creating Global Schema

Global schemas are very difficult to create, since they must be general enough to handle any type of data in the contributing databases.

Levels of "globality"

1. Global query model
--write queries in local language, but then combine queries as needed at global level (KLEISLI).

2. Global data model and query model
--write queries in global language, less knowledge of local structure required (OPM)

Copyright Russ B. Altman

Consolidation

- Gather together data of interest, and translate it once into common database structure, with all incompatibilities “scrubbed out.”

- Remove the contributing legacy DBs once and for all.

- Rarely done--more likely in industry
- Nontrivial to resolve semantic incompatibilities
- Nontrivial to scrub data to match resolved semantics

- With rapidly growing data, one time translation is not adequate.

Copyright Russ B. Altman

Federation 1: incorporate links within databases to one another

- Often hypertext links from item in A to item in B.
- Data retrieval by traversal of links
- No other compatibility imposed on DBs
- Prone to missing or inconsistent links, especially in the setting of rapidly growing databases (are links static or dynamically computed?)
- No general purpose query facility to retrieve multiple records that satisfy some set of general requirements.
- Most commonly used technique:
  DBGET and SRS

Copyright Russ B. Altman

```
HEADER    OXYGEN STORAGE                          05-APR-73   1MBN      1MBNH  1
COMPND    MYOGLOBIN  (FERRIC IRON - METMYOGLOBIN)                       1MBN   4
SOURCE    SPERM WHALE (PHYSETER CATODON)                               1MBNM  1
AUTHOR    H.C.WATSON,J.C.KENDREW                                       1MBNG  1
REVDAT  20   27-OCT-83 1MBNS   1       REMARK                          1MBNS  1
JRNL        AUTH   H.C.WATSON                                          1MBNG  2
JRNL        TITL   THE STEREOCHEMISTRY OF THE PROTEIN MYOGLOBIN        1MBNG  3
JRNL        REF    PROG.STEREOCHEM.              V.   4   299 1969     1MBNG  4
JRNL        REFN   ASTM PRSTAP  US ISSN 0079-6808                 419  1MBNG  5
REMARK   1                                                             1MBNG  6
REMARK   1 REFERENCE 1                                                 1MBNQ  1
REMARK   1  AUTH   J.C.KENDREW                                         1MBNQ  2
REMARK   1  TITL   MYOGLOBIN AND THE STRUCTURE OF PROTEINS (NOBEL      1MBNQ  3
REMARK   1  TITL 2 LECTURE, DECEMBER 11, 1962)                         1MBNQ  4
REMARK   1  REF    PRIX NOBEL                          103 1963        1MBNQ  5
REMARK   1  REFN   ASTM PRIXAL  SW ISSN 0546-8175                 945  1MBNS  2
REMARK  10 CORRECTION. REORDER THE ATOMS OF THE HEME GROUP AND CHANGE  1MBNF 35
REMARK  10  THE CONECT RECORDS CORRESPONDINGLY. 23-AUG-77.             1MBNF 36
SEQRES   1    153  VAL LEU SER GLU GLY GLU TRP GLN LEU VAL LEU HIS VAL 1MBN  39
SEQRES   2    153  TRP ALA LYS VAL GLU ALA ASP VAL ALA GLY HIS GLY GLN 1MBN  40
HET    HEM      1      44     PROTOPORPHYRIN IX WITH FE(OH), FERRIC    1MBND 10
FORMUL   2  HEM    C34 H32 N4 O4 FE1 +++ .                             1MBNG 25
FORMUL   2  HEM    H1 O1                                               1MBNG 26
HELIX    1   A SER      3  GLU     18  1 N=3.63,PHI=1.73,H=1.50       1MBN  52
HELIX    2   B ASP     20  SER     35  1 N=3.72,PHI=1.69,H=1.47       1MBN  53
HELIX    3   C HIS     36  LYS     42  1 SHORT IRREGULAR HELIX        1MBN  54
3
SITE     1 HMB  9 PHE    43  ARG    45  HIS    64  VAL    68           1MBNN  4
CRYST1   64.500   30.900   34.700  90.00 106.00  90.00 P 21         2 1MBN  65
ORIGX1       1.00000   0.00000   0.00000        0.00000                1MBN  66
ORIGX2       0.00000   1.00000   0.00000        0.00000                1MBN  67
ORIGX3       0.00000   0.00000   1.00000        0.00000                1MBN  68
SCALE1        .01550   0.00000    .00445        0.00000                1MBN  69
SCALE2       0.00000    .03236   0.00000        0.00000                1MBN  70
SCALE3       0.00000   0.00000    .02998        0.00000                1MBN  71
ATOM      1  N   VAL     1      -2.900  17.600  15.500  1.00  0.00   2 1MBN  72
ATOM      2  CA  VAL     1      -3.600  16.400  15.300  1.00  0.00   2 1MBN  73
```

P

D

B

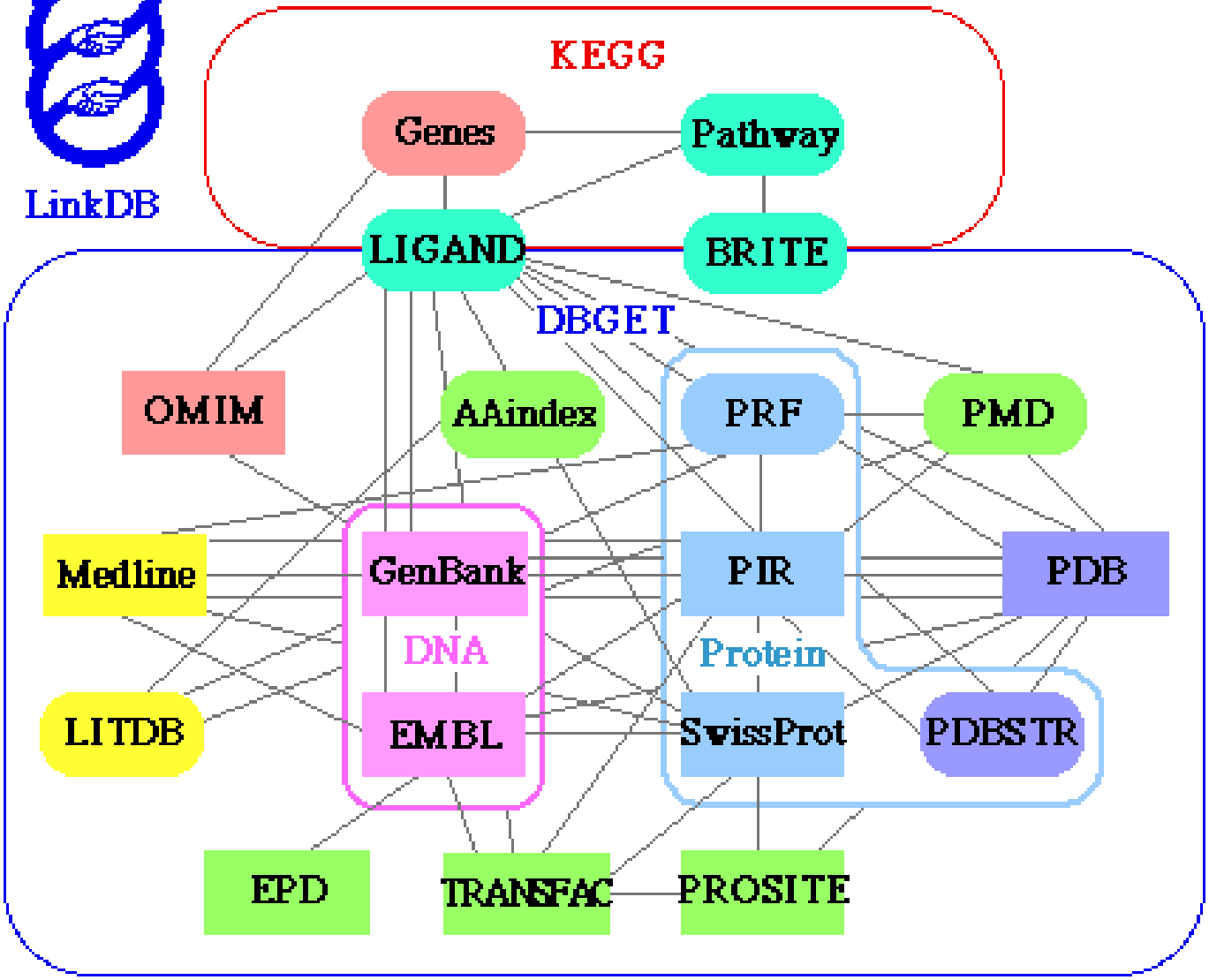
Copyright Russ B. Altman

```
ID   AGP1_YEAST     STANDARD;      PRT;   633 AA.
AC   P25376;
DT   01-MAY-1992 (REL. 22, CREATED)
DT   01-MAY-1992 (REL. 22, LAST SEQUENCE UPDATE)
DT   01-NOV-1997 (REL. 35, LAST ANNOTATION UPDATE)
DE   ASPARAGINE/GLUTAMINE PERMEASE.
GN   AGP1 OR YCL25C.
OS   SACCHAROMYCES CEREVISIAE (BAKER'S YEAST).
OC   EUKARYOTA; FUNGI; ASCOMYCOTINA; HEMIASCOMYCETES.
RN   [1]
RP   SEQUENCE FROM N.A.
RA   HOLLENBERG C.P., KLEINHANS U., LUETZENKIRCHEN K., RAD M.R., XU G.;
RL   SUBMITTED (MAR-1992) TO EMBL/GENBANK/DDBJ DATA BANKS.
RN   [2]
RP   CHARACTERIZATION.
RA   SCHREVE J.L., SIN J., GARRETT J.M.;
RL   UNPUBLISHED OBSERVATIONS (JUL-1997).
CC   -!- FUNCTION: BROAD SUBSTRATE RANGE PERMEASE WHICH TRANSPORTS
CC       ASPARAGINE AND GLUTAMINE WITH INTERMEDIATE SPECIFICITY.
CC   -!- SUBCELLULAR LOCATION: INTEGRAL MEMBRANE PROTEIN (PROBABLE).
CC   -!- SIMILARITY: BELONGS TO A FAMILY THAT GROUPS MANY AMINO ACID
CC       PERMEASES.
DR   EMBL; X59720; E264437; -.
DR   PIR; S19352; S19352.
DR   SGD; L0003271; AGP1.
DR   PROSITE; PS00218; AMINO_ACID_PERMEASE; 1.
KW   TRANSPORT; AMINO-ACID TRANSPORT; TRANSMEMBRANE.
FT   TRANSMEM    125    141       POTENTIAL.
FT   TRANSMEM    152    169       POTENTIAL.
FT   TRANSMEM    191    214       POTENTIAL.
```

Swiss-Prot

Copyright Russ B. Altman

DBGET Database Links



Copyright Russ B. Altman

LinkDB Search Result: PDB 1mbn - Netscape

File Edit View Go Communicator Help

Back Forward Reload Home Search Guide Print Security Stop

Bookmarks Location: http://www.genome.ad.jp/dbget-bin/www_bfind_sub?dbkey=linkdb&keywords=pdb%3A1mbn&organism=none&mode=blink

LinkDB Search Result

Database: LinkDB

Database of Link Information
Release 98-05-14, May 98
Institute for Chemical Research, Kyoto University
5,131,470 entries

PDB : 1mbn - related entries

Factual Links (9 hits from 5 databases)

1. SWISS-PROT (1)
2. PIR (2)
3. PDBSTR (1)
4. PROSITE (1)
5. MEDLINE (4)
6. All databases (9)

Link table for PDB

Document: Done

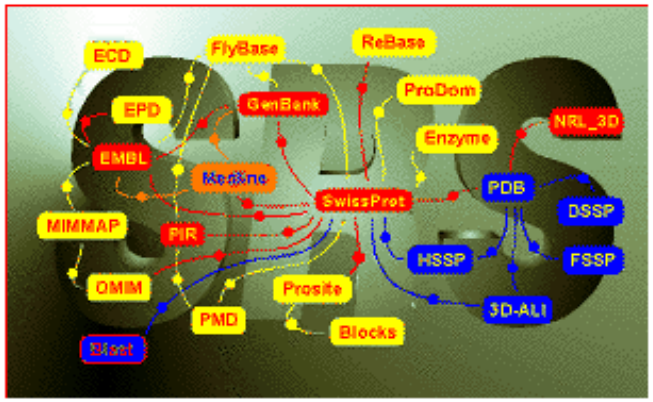
SRSWWW at EMBL-EBI, Hinxton, UK - Netscape

File Edit View Go Communicator Help

Bookmarks Location: http://srs.ebi.ac.uk:5000/

Sequence Retrieval System

Network Browser for Databanks in Molecular Biology



Start — Start a new SRS session

The SRS Manual

SRS World Wide

The SRS newsgroup

SRS Developers

Document: Done

Copyright Russ B. Altman

File Edit View Go Communicator Help

Bookmarks Location: http://srs.ebi.ac.uk:5000/srs5bin/cgi-bin/wgetz

Navigation Toolbar

Top Page | Query Form | Query Manager | View Manager | Databanks | Help

Select one or more databanks and continue (explode + or collapse − all groups)

Continue Reset

− **Sequence** all

SWISSPROT SWISSNEW PIR EMBL EMBLNEW NRL3D

SPTREMBL REMTREMBL IMGT TREMBLNEW

+ *SeqRelated*

+ *TransFac*

+ *Application Results*

− **Protein3DStruct** all

PDB DSSP HSSP FSSP PDBFINDER PDBREPORT

+ *Genome*

+ *Mapping*

+ *Mutations*

+ *Locus Specific Mutations*

+ *Others*

Bookmark this link to return to your session: resume

... tired of looking at all this data? Change their color! Try plain

If you find problems or have suggestions please mail the SRS administrator

Document: Done

Copyright Russ B. Altman

View Select Page - Netscape

File Edit View Go Communicator Help

Bookmarks Location: http://srs.ebi.ac.uk:5000/srs5bin/cgi-bin/wgetz?-id+5VggB194GFd+-fun+Pagelibinfo+-info+PDB

WWW

http://www.pdb.bnl.gov/

Data-fields in SRS

Name	Short Name	Type	No of Keys	No of Entry References	Indexing Date	Status
ID	id	ID	7596	7596	5/14/98	ok
Date	dat	int	2086	7596	5/14/98	ok
Supersedes	spr	string	299	304	5/14/98	ok
Compound	com	string	8107	126425	5/14/98	ok
Source	src	string	5192	86294	5/14/98	ok
EntryAuthors	eat	string	7293	26243	5/14/98	ok
Authors	aut	string	15149	66912	5/14/98	ok
Title	tit	string	8107	126425	5/14/98	ok
JournalName	jnl	string	757	19601	5/14/98	ok
Remark	rem	string	51338	1741048	5/14/98	ok
CRYST1	cr1	show				not indexed
ORIGXn	ori	show				not indexed
SCALEn	sca	show				not indexed
MTRIXn	mtr	show				not indexed
TVECT	tve	show				not indexed
MODEL	mod	show				not indexed

Links To

none

Document: Done

Copyright Russ B. Altman

Netscape

File Edit View Go Communicator Help

Bookmarks Location: http://srs.ebi.ac.uk:5000/srs5bin/cgi-bin/wgetz?-fun+PageIcarusFile+-l+PDB+-ifile+i

```
PDB_DB:$library:[PDB group:@PROTSTRUCT_LIBS
  format:@PDB_FORMAT maxNameLen:30 ifiles:{"pdb.i" "pdb.is"}
]

PDB_FORMAT:$libformat:[fileType:{@PDB_FILE} syntax:@PDB_SYNTAX
  fields:{
    $field:@DF_ALL
    $field:[@DF_HeaderField name:'Title Section']
    $field:[@DF_ID code:header index:id indexToken:id]
    $field:[@DF_Date code:header index:int indexToken:date]
    $field:[@DF_Supersedes code:sprsde index:str indexToken:sprsde]
    $field:[@DF_Compound code:compnd index:str indexToken:'wordX|compnd'
      tableToken:'t_fields|compnd']
    $field:[@DF_Source code:source index:str indexToken:'wordX|source'
      tableToken:'t_fields|source']
    $field:[@DF_EntryAuthors code:author index:str indexToken:author]
    $field:[@DF_Authors code:auth index:str indexToken:auth]
    $field:[@DF_Title code:titl index:str indexToken:'wordX|compnd']
    $field:[@DF_Journal code:ref index:str indexToken:ref]
    $field:[@DF_Remark code:remark index:str indexToken:'wordX|remark']
```

Document: Done

Copyright Russ B. Altman

Netscape

File Edit View Go Communicator Help

Back Forward Reload Home Search Guide Print Security Stop

Bookmarks Location: http://srs.ebi.ac.uk:5000/srs5bin/cgi-bin/wgetz?-fun+PagelcarusFile+-l+PDB+-ifile+is

```
$rules={
  # the entry
  entry:     ~ {$In:[file:text] $Out init{$Skip:0}}
               'HEADER    ' {pre {$entryFip=0} $Wrt} appLn
               ('END    ' {$Not} appLn)*  ~
  data:      ~ {$In:[file:data share:text] $Out}
               ln {$Wrt $dataFip=$Fip} appLn+  ~

  # fields from text and data
  fields:    ~ {$In:entry $Out}
               (tag { $t=$Ct
                      if:$fn.$t!="" {if:$fn.$t!=$fPrev $Wrt:$fn.$t else $App
                      $fPrev = $fn.$t}
                    } ln {$App})+ ~
  dfields:   ~ {$In:data $Out}
               (tag { $t=$Ct
                      if:$fn.$t!="" {if:$fn.$t!=$fPrev $Wrt:$fn.$t else $App
                      $fPrev = $fn.$t}
                    } ln {$App})+ ~
  tag:       ~ /JRNL +[A-Z]+/ | /REMARK +1 +[A-Z]+/ | /[A-Z0-9]+/ ~
```

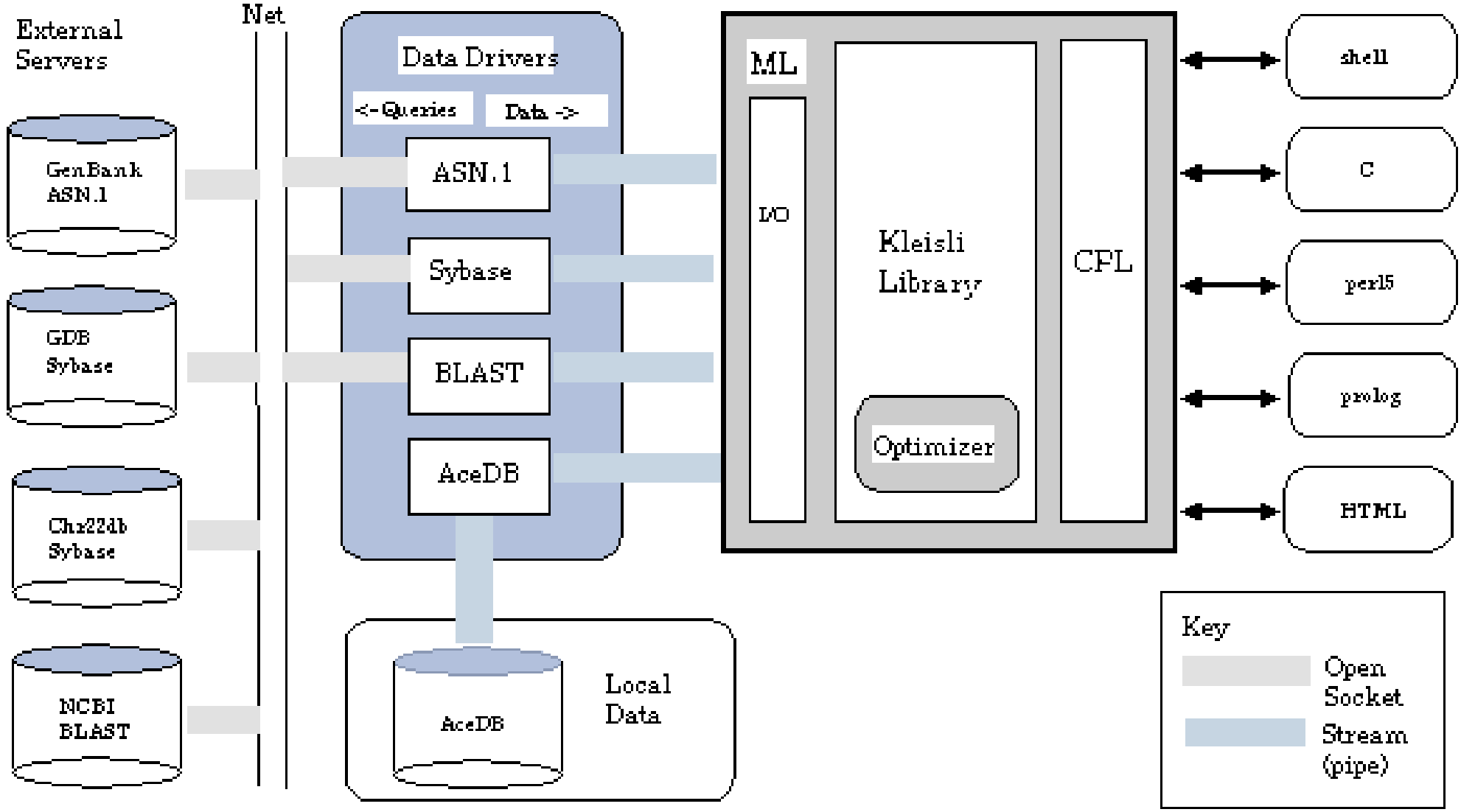
Document: Done

Copyright Russ B. Altman

Federation 2: couple DBs loosely

- Construct queries over multiple DBs without touching the DBs themselves
- Create query processor which can map query to individual search capabilities of DBs and integrate answer together
- Kleisli: uses common query language, with queries mapped directly into valid searches in the local DB.
- OPM: uses common data model, and maps local DBs onto the common model, and then queries the model (with translation into local DB)

Copyright Russ B. Altman



Copyright Russ B. Altman

Sample Kleisli Syntax

For relational database like GDB, define query to determine whether entry is part of chromosome 22:

```
define GDB == Open-Sybase([server="GDB",
  user="guest", password="smith@stanford"]);
define Loci22 == GDB([query=
      "select locus_symbol, genbank_ref
               from locus
               where loc_cyto_chrom_num = '22'"])
```

For flat text file database Genbank, find accession number:

```
define Genbank==Open-ASN([server="NCBI",
  user="guest", password = ""]);
define ASN-IDs==\accession =>
  Genbank([db="na", select="accession" ^ accession,
      path = "Seq-entry.seq.id..giim",args=[]]);
```

Query for all genbank entries that occur on chromosome 22:

```
{[locus=locus]| \locus <- Loci22, \uid<- ASN-
  IDs(locus.genbank_ref)}
```

Copyright Russ B. Altman

Difficult Kleisli Query

http://adenine.krdl.org.sg:8080/demos/biokleisli/subbiah/

Determine which sequences of unknown structure should be prioritized for structure determination.

Insight: Find sequences that are referred to in literature at high rate, but which don't have a structure.

Query plan:

1. **Extract all sequences of known structures from SCOP**
2. **Extract all sequences of proteins from Swiss-Prot**
3. **Remove from (2) all homologous to (1)**
4. **Cluster remaining sequences with BLAST**
5. **Rank clusters based on number of MEDLINE references to members of the cluster**

Copyright Russ B. Altman

Netscape

File Edit View Go Communicator Help

Back Forward Reload Home Search Guide Print Security Stop

Bookmarks Location: http://adenine.krdl.org.sg:8080/cgi-bin/examples/subbiah/by-size.pl

Proteins of Unknown Structure

Clusters that have at least 3 members and at least Medline references. Members of these clusters does not have any BLAST-relationship at a pscore better than 1.0E-4 to any protein of known structure. Members of the same cluster are BLAST-related at a pscore no more than 1.0E-4. The clusters are sorted by Medline count.

1. **cluster-rep** mus musculus (mouse). kappa-type opioid receptor (kor-1) (msl-1). 10/96
   cluster-size 8942
   medline-count 8058

2. **cluster-rep** bacillus subtilis. probable aldehyde dehydrogenase ycbd (ec 1.2.1.3). 10/96
   cluster-size 192
   medline-count 195

3. **cluster-rep** sulfolobus solfataricus. aspartate aminotransferase (ec 2.6.1.1) (transaminase a) (aspat). 11/95
   cluster-size 170
   medline-count 179

4. **cluster-rep** bos taurus (bovine). acetylcholine receptor protein, delta chain precursor. 10/96
   cluster-size 144
   medline-count 173

Sample OPM query

Find name and annotation in GSDB of Gene called ACHE in GDB.

```
SELECT Name = GSDB:gene.name,
  Annotation = GDB:Gene.annotation
FROM GSDB:Gene, GDB:Gene
WHERE GDB:Gene.accessionID =
  GSDB.Gene.gdb_xref AND
  GSDB:Gene.name = “ACHE”
```

Copyright Russ B. Altman

The Genome Database - Netscape

File Edit View Go Communicator Help

The Genome Database

An international collaboration in support of the Human Genome Project.

Hosted by The Johns Hopkins University School of Medicine, Baltimore, Maryland USA and available at mirror sites worldwide

What's New (9-Mar-98):

- **GDB 6.4 released** (includes Mapview 2.4)
- New Mirror Site in Taiwan
- Bioinformatics Seminar Series
- GDB Termination Announcement

GDB Termination Announcement

Simple Search

Search	Genomic Segments All Biological Data People Citations	by	Name/GDB ID Keyword DNA Sequence ID	pHM.17.E1 Submit Reset

Note: When doing Name/ID searches, adding * to the end of your search text may improve your results.

Other Search Options

Edit | **Help** | **Site Map** | **About GDB** | **Reports** | **Resources** | **Prefs** | **GDB Termination Feedback** | **5.6**

For help, contact help@gdb.org or 1-410-955-9705. For best viewing, use Netscape 3.0 and higher

Please note: The Johns Hopkins University Bioinformatics Web Server (formerly at www.gdb.org) has a new address:

Document: Done

Federation 3: Data warehouse

- Develop a global schema for all the data in the DBs
- Data are tranformed into this common schema and loaded in central repository on a regular basis
- Query facilities provided by central repository
- Need to update global schema if/when local DBs change their data formats/schemas

- Similar to consolidation strategy, but local DBs remain and are synchronized in the global DB.

Copyright Russ B. Altman

ENTREZ system from NCBI

http://www.ncbi.nlm.nih.gov/Entrez/

1. All data is translated into Abstract Syntax Notation (ASN.1) structured files.

2. Most links to other parts of database are determined at the time of translation into ASN.1

3. Some dditional links can be computed on the fly using keyword searches, sequence similarity searches, or any other comparison metrics over the entire DB.

4. Updates are performed at regular interval.

Copyright Russ B. Altman

Entrez Browser - Netscape

File Edit View Go Communicator Help

Bookmarks Location: http://www.ncbi.nlm.nih.gov/Entrez/

NCBI

Entrez Help

The Entrez Databases

Network Entrez

Retrieve large data sets

Making WWW Links to Entrez

Entrez

Search WWW Entrez at NCBI

- Nucleotides
- Proteins
- 3D structures
- Genomes
- Taxonomy
- Literature - PubMed

The Entrez Browser is provided by the National Center for Biotechnology Information. NCBI also builds, maintains, and distributes the GenBank Sequence Database.

Document: Done

PubMed protein query - Netscape

File Edit View Go Communicator Help

Bookmarks Location: http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=p_r

NCBI Entrez **Protein QUERY** BLAST Entrez ?

Details myoglobin[All Fields] Search Clear

Docs Per Page: 20 Mod. Date limit: No Limit

citations 1-20 displayed (out of 417 found), page 1 of 21

Display GenPept Report for the articles selected (default all).

☐ P02205
MYOGLOBIN
gi|127697|sp|P02205|MYG_THUAL [127697]
(View GenPept Report, FASTA report, ASN.1 report, Graphical view, 1 MEDLINE link, or 1528 protein neighbors)

☐ P02185
MYOGLOBIN
gi|127687|sp|P02185|MYG_PHYCA [127687]
(View GenPept Report, FASTA report, ASN.1 report, Graphical view, 4 MEDLINE links, or 987 protein neighbors)

Document: Done

Copyright Russ B. Altman

ASN.1 version of data record

```
Seq-entry ::= seq {
  id {
    pdb {
      mol "1MBN" ,
      rel
        std {
          year 1994 ,
          month 1 } } ,
    gi 230152 } ,
  descr {
    pdb {
      deposition
        std {
          year 1973 ,
          month 4 ,
          day 5 } ,
      class "Oxygen Storage" ,
      compound {
        "Myoglobin  (Ferric Iron - Metmyoglobin)" } ,
      source {
        "Sperm Whale (Physeter Catodon)" } ,
      exp-method "X-Ray Diffraction" },
AND MORE...
```

Copyright Russ B. Altman