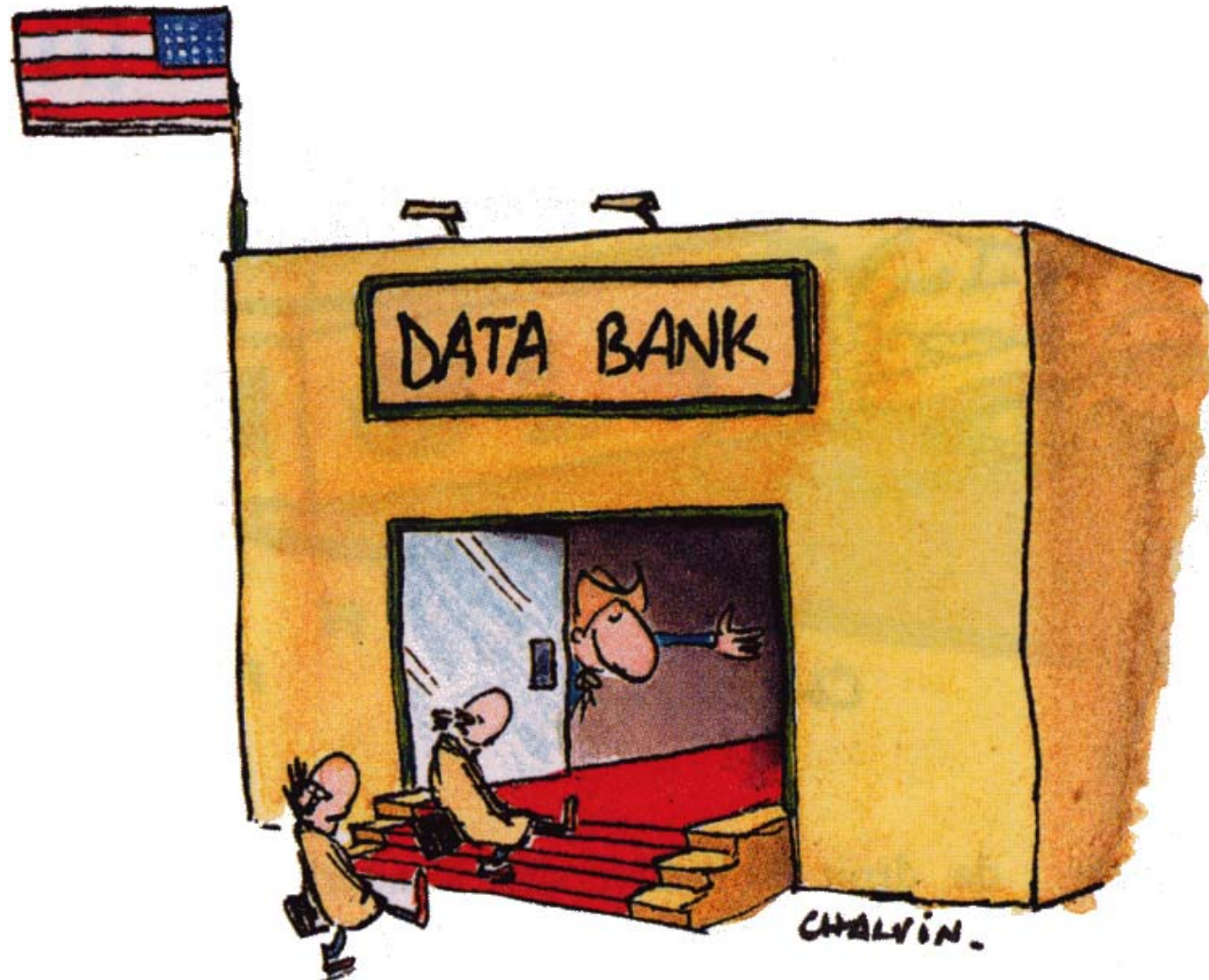


Ch2 Introduction to biological databases



Overview

- ❖ **What is a database?**
- ❖ **What is basis of database?**
- ❖ **What is the purpose of database?**
- ❖ **How many types of databases?**

What is database?

Database: a computerized archive to store and organize data in such a way that information can be retrieved easily according to various criteria.

What is the basis of databases?

Basis: computer hardware and computer software for data management.

What is the purpose of database?

1. Data retrieval
2. Knowledge discovery

Terminology

Entry

Field

Query

Entry, Field, and Query in PubMed

- <http://www.ncbi.nlm.nih.gov/>

Types of Databases

Flat files

Relational Database

Object oriented Databases

Flat files

A text file that can be edited by a text editor, such as Microsoft “notepad”.
Often, it contains many entries separated by a special delimiter such as tab or comma.
Each field within an entry can be separated by other delimiters, such as “|”.

Example: FASTA files

Flat files

Pros: easy for human to read and edit.

Easy for programming using SHELL, PERL and Python.

(Demo of SHELL and PERL script on flat files under Linux.)

Cons: very inefficient for computer to search for specific field or record because it would have to read through the entire file, i.e. very slow.

Relational Databases

Relational databases use a set of tables to organize data. Each table is called a “relation” and is made up of columns and rows.

Columns represent individual fields in a table.

Rows represent records of data in a table.

Columns in a table are indexed and linked so that they can be cross-referenced in other tables.

SQL and Relational DB

Relational databases use a set of tables to organize data. Each table is called a “relation” and is made up of columns and rows.

Columns represent individual fields in a table.

Rows represent records of data in a table.

Columns in a table are indexed and linked so that they can be cross-referenced in other tables.

Tables in Relational DB

Table A

Student #	Name	State
1	John Smith	Texas
2	Jane Doe	Kansas
3	William Brown	Illinois
4	Jennifer Taylor	New York
5	Howard Douglas	Texas

Table B

Student #	Course #
1	Biol 689
2	Bich 441
3	Chem 289
4	Hort 201
5	Math 172

Table C

Course #	Course name
Biol 689	Bioinformatics
Bich 441	Biochemistry
Chem 289	Organic chemistry
Hort 201	Horticulture
Math 172	Calculus



Select * from TableA and TableB where
TableA.Student# = TableB.Student#
And TableA.State = "Texas";

(See demonstration in Linux)

Object-oriented databases

Object-oriented databases store data as “objects” which are linked by pointers. In programming language, an object is like a module that combines data and mathematical routines that act on the data. Between objects are a set of pointers that define the predetermined relationships between objects. Searching the database is to navigate through the objects, which relies on predefined pointers.

Programming languages like C++, JAVA can be used to create object-oriented databases.

Biological Databases

Current biological databases use all three types of database, flat files, relational, and objected-oriented. Despite obviously drawback of flat files in database management, many biological databases are still use this format. A justification for that is that it is easy to understand by working biologists and easy to be handled by some sequence analysis tools.

Biological Databases

Primary databases: contain original biological data, raw sequence or structural data submitted by the scientific community. Ex. GenBank and protein data bank.

Secondary databases: contain computationally processed or manually curated information based on original information from primary databases. Ex. Swiss-Prot, InterPro, Ensembl.

Specialized databases: those that cater a particular research interest. Ex. Flybase, SGD, HIV sequence database, and Ribosomal Database Project.

Primary Sequence Databases

Three major public sequence databases: GenBank, EMBL, DDBJ. They store raw nucleic acid sequence data submitted by scientific community worldwide. There is only a minimal level of annotation.

The three databases collaborate and share new data on a daily basis

Nowadays, sequence submission to one of these three databases is a precondition for publication in most scientific journals.

SWISS-PROT

An example of secondary database.

(<http://us.expasy.org/sprot/>) An extensively curated protein sequence database with detailed annotation for confirmed protein sequences. The annotations include structure, function, and protein family assignment. It maintains a low level of redundancy by merging variants and fragments of a sequence into a single entry.

Specialized Databases

Specialized databases normally serve a specific research community or focus on a particular organism. The content of these databases may overlap with the primary database but may have new data directly submitted by authors. Since they are often curated by experts in the field, they may have unique organizations and additional annotations associated with the sequences.

Ex. genome databases (TIGR microbial genome database, Flybase, AceDB, TAIR, Ensembl).

Ex. functional databases. GenBank EST database and Microarray Gene Expression Database at EBI

Interconnection between databases

Except a few cases, poor coordination.

Often incompatible file formats.

Pitfalls of Databases

One of the problems associated biological databases is over-reliance sequence information and related annotations without understanding the reliability of the information.

Sequence errors.

Redundancy.

Errorneous annotations.

The errors are passed onto other databases causing propagation of errors.

Errors

Mostly experimental problems.

Sequencing errors which exhibit in the form of substitutions, insertions, deletions or contamination. Some of these errors cause frameshifts that make whole gene identification difficult. Sometimes, gene sequences are contaminated by sequences from cloning vectors.

Errors are more common for sequences produced before 1990s.

Therefore, excessive care should be taken when dealing with those dated sequences.

Redundancy

There is tremendous duplication of information in the databases due to :

- repeated submission of identical or overlapping sequences by the same or different authors;

- revision of annotations;

- dumping of EST data;

- poor database management that fails to detect the redundancy.

This makes databases excessively large and unwieldy for information retrieval.

Many sequences are duplicated either entirely or partially.

RefSeq

NCBI has now created a *non-redundant* database, called RefSeq.

Identical sequences from the same organism and associated fragments are merged into a single entry. That involves sequence similarity comparison of every entry.

RNA and proteins sequences derived from the same DNA sequences are explicitly linked as related entries.

Sequences variants from the same organism with very minor differences, which may well be due to sequencing errors, are treated distinctly related entries.

NCBI

National Center for Biotechnology Information: integrated databases and analysis tools.

Entrez: integrated gateway. It allows text-based searches for annotated genetic sequence information, structural information, and citations and abstracts, full papers, taxonomic data. BLAST searching allows one to connect one sequence with other sequences by underlying evolutionary relationships.











































Search across databases

GO

CLEAR

Help

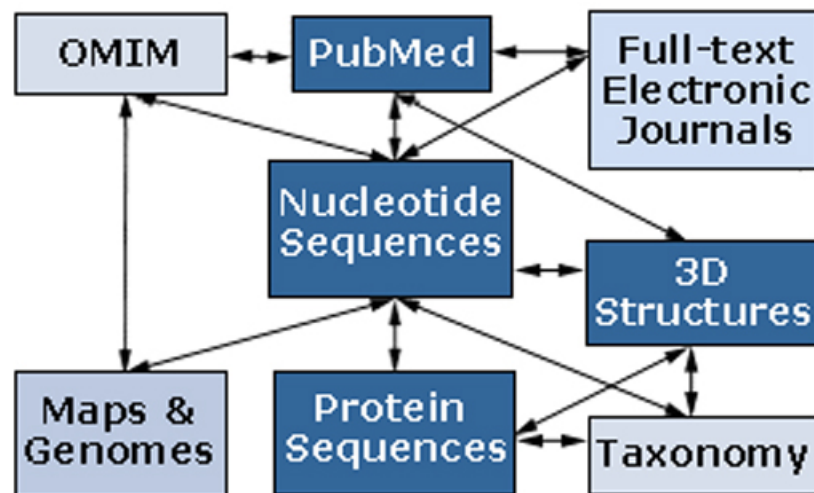
Welcome to the new Entrez cross-database search page

 PubMed: biomedical literature citations and abstracts		 Books: online books	
 PubMed Central: free, full text journal articles		 OMIM: Online Mendelian Inheritance in Man	
 Journals: detailed information about journals in Entrez		 Site Search: NCBI web and FTP sites	
 MeSH: detailed information about NLM's controlled vocabulary			
 Nucleotide: sequence database (GenBank)		 UniGene: gene-oriented clusters of transcript sequences	
 Protein: sequence database		 CDD: conserved protein domain database	
 Genome: whole genome sequences		 3D Domains: domains from Entrez Structure	
 Structure: three-dimensional macromolecular structures		 UniSTS: markers and mapping data	
 Taxonomy: organisms in GenBank		 PopSet: population study data sets	
 SNP: single nucleotide polymorphism		 GEO: expression and molecular abundance profiles	
 Gene: gene-centered information		 GEO DataSets: experimental sets of GEO data	

Enter terms and click 'GO' to run the search against ALL the databases, OR
 Click Database Name or Icon to go directly to the Search Page for that database, OR
 Click Question Mark for a short explanation of that database.

NCBI

Integrated Database: Interconnection between databases is the emphasis. Instead of having to visit multiple databases located in disparate places, the gateway allows access to various types of information by issuing only one query. Ex. nucleotide sequence linked to translated protein sequences to literature citations, and phylogeny of the organism and PubMed citations as well as full text articles whenever available.



Paper → Gene sequence → Protein structure
↓
Chromosome mapping

Unified Search Across Databases

<http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi>

PubMed

A biomedical literature database containing abstracts from nearly 4,000 journals.

Some abstracts have links to full text articles in journals' or publishers' websites or in a related database PubMed Central.

Major feature: MeSH terms.

PubMed Tutorials

<http://www.nlm.nih.gov/bsd/disted/pubmed.html>

PubMed Searches

- Search by Authors and key words
- Use Boolean operations, AND, OR, NOT
- Use tags, [AU] [TI] etc
- Use “Limits” (Journal and Dates)
- “Preview/Index” to help combine searches
- “History”,
- “Clipboard”, store temporary hits
- “Related links” to expand the searches
- PubMed services (Journal, Citation)
- Personalized search (MyNCBI)

PubMed Tags

Tag	Name
AB	Abstract
AD	Affiliation
AID	Article Identifier
AU	Author
DP	Publication Date
JID	Journal ID
LA	Language
OTO	Other Term Owner
PL	Place of Publication
PT	Publication Type
RN	EC/RN Number
SO	Source
TA	Journal Title Abbreviation
TI	Title
VI	Volume

GenBank

The most complete collection of annotated nucleic acid sequence data for almost every organism.

Contents: genomic DNA, mRNA, cDNA, ESTs, genome sequence raw data, sequence polymorphisms (SNPs).

Protein sequences (GenPept), the majority of which are conceptual translations from DNA sequences, but a small number of amino acid sequences are derived from peptide sequencing techniques.

Two ways to search GenBank

- Key words search
- BLAST by sequence

GenBank file format

Header

Features

Sequence entry

Header

LOCUS Q9ZGE9 440 aa linear BCT 15-JUN-2002
 DEFINITION Light-independent protochlorophyllide reductase subunit N (LI-POR subunit N) (DPOR subunit N).
 ACCESSION Q9ZGE9
 VERSION Q9ZGE9 GI:18203677
 DBSOURCE swissprot: locus BCHN_HELMO, accession Q9ZGE9;
 class: standard.
 created: Oct 16, 2001.
 sequence updated: Oct 16, 2001.
 annotation updated: Jun 15, 2002.
 xrefs: gi: [3820536](#), gi: [3820556](#)
 xrefs (non-sequence databases): InterProIPR000510, PfamPF00148
 KEYWORDS Photosynthesis; Bacteriochlorophyll biosynthesis; Oxidoreductase.
 SOURCE Heliobacillus mobilis
 ORGANISM [Heliobacillus mobilis](#)
 Bacteria; Firmicutes; Clostridia; Clostridiales; Heliobacteriaceae; Heliobacillus.
 REFERENCE 1 (residues 1 to 440)
 AUTHORS Xiong,J., Inoue,K. and Bauer,C.E.
 TITLE Tracking molecular evolution of photosynthesis by characterization of a major photosynthesis gene cluster from Heliobacillus mobilis
 JOURNAL Proc. Natl. Acad. Sci. U.S.A. 95 (25), 14851-14856 (1998)
 MEDLINE [99061957](#)
 PUBMED [9843979](#)
 REMARK SEQUENCE FROM N.A.
 COMMENT

 This SWISS-PROT entry is copyright. It is produced through a collaboration between the Swiss Institute of Bioinformatics and the EMBL outstation - the European Bioinformatics Institute. The original entry is available from <http://www.expasy.ch/sprot> and <http://www.ebi.ac.uk/sprot>

[FUNCTION] Uses Mg-ATP and reduced ferredoxin to reduce ring D of protochlorophyllide (Pchl) to form chlorophyllide a (Chl) (By similarity). This reaction is light-independent.
 [PATHWAY] Light-independent bacteriochlorophyll biosynthesis.
 [SUBUNIT] Protochlorophyllide reductase is thought to be composed of three subunits; bchL, bchN and bchB. Could form a heterotetramer of two bchB and two bchN subunits.
 [SIMILARITY] BELONGS TO THE BCHN / CHLN FAMILY.

Features

FEATURES Location/Qualifiers
 source 1..440
 /organism="Heliobacillus mobilis"
 /db_xref="taxon:28064"
 gene 1..440
 /gene="BCHN"
 Protein 1..440
 /gene="BCHN"
 /product="Light-independent protochlorophyllide reductase subunit N"
 /EC_number="1.18.-.-"

Sequence

ORIGIN
 1 merverengc fhtfcpiasv awlhrkikds fflivgthtc ahfiqtaldv mvyahsrfgf
 61 avleesdlvs aspteelgkv vqgvvdewhp kvifvlstcs vdilkmdlev sckdlstrfg
 121 fpvlpastsg idrsftgged avlhallpfv pkeapavepv eekkprwfsf gkesekae
 181 parnlvliga vtdstiqqlq welkqlglpk vdvfpdgdv kmpvinegtv vvplqpylnd
 241 tlatirrerr akvlstvfpi gpdgtarfle aiclegldt srikekeaga wrdleplqi
 301 lrgkkimflg dnllleplar fltscdvqv eagtpyihsk dlqgelellk erdvrvresp
 361 dftkqlqrmq eykpdvvag lgicnpleam gfttawsief tfaqihgfvn aidliklftk
 421 pllkrqalme hgwaagwle

//

Accession numbers and gi numbers

Accession number is unchanged, but the gi number are changeable.

So, always use accession number when searching.

FASTA

```
>gi|5106368|dbj|AB009351.1| Citrus sinensis mRNA for chalcone synthase, complete cds
AAACATATTCATTAAGGGTTCAACTTGAAATGGCAACCGTTCAAGAGATCAGAAACGCTCAGCGTGCCGA
CGGCCCCGGCCACCGTCCTCGCCATCGGTACGGCCACGCCTGCCCACAGTGTCAACCAGGCTGATTATCCC
GACTATTACTTCAGGATCACAAAGAGCGAGCATATGACGGAGCTTAAAGAGAAGTTCAAGCGCATGTGTG
ACAAGTCGATGATTAAGAAGCGTTACATGTACTTAACGGAAGAGATTTTGAAGGAAAACCCCAACATGTG
CGCTTACATGGCTCCATCACTCGACGCACGTCAAGACATTGTGGTGGTTCGAAGTGCCGAAGCTCGGGAAA
GAAGCTGCTACAAAGGCCATCAAAGAATGGGGCCAGCCCAAGTCTAAGATCACCCACCTCATCTTCTGCA
CCACCTCCGGCGTCGACATGCCCGGCGCCGACTACCAGCTCACCAAAGTATCGGGCCTGCGCCCCCTCCGT
CAAGCGCTTCATGATGTACCAACAAGGATGTTTCGCCGGCGGCACTGTTCTTCGCCTCGCTAAAGACTTG
GCTGAGAACAACAAGGGCGCTCGCGTTCTTGTTGTCTGTTTCGGAGATCACGGCAGTCACTTTCCGCGGCC
CTGCCGATACTCATCTTGATTCTCTTGTTGGGCCAGGCTTTGTTTCGGTGATGGTGCTGCTGTGATCGT
GGGTGCCGATCCTGACACGTTCGGTCGAGCGTCCGTTGTATCAGCTCGTGTCTGACTTCGCAGACGATCCTC
CCTGACTCTGACGGTGCAATTGACGGACACTTGCGCGAAGTCGGTCTCACTTTCCATTTGCTTAAAGACG
TCCCCGGCTTGATCTCAAAGAATATAGAGAAAAGCCTGTCTGAAGCGTTCGCCCCGCTTGGCATCAGCGA
CTGGAACCTCGATATTCTGGATCGCTCACCCCGGTGGGCCCCGCAATTCTGGACCAAGTGGAGTCTAAACTG
GGCCTCAAAGGGGAGAAGCTGAAGGCCACACGTCAAGTGTTGAGTGAGTACGGTAACATGTCAAGCGCTT
GTGTCTTGTTTCATCCTTGACGAGATGAGAAAGAAGTCTGTTGAGGAAGCGAAAGCCACCACCGGCGAAGG
GCTTGATTGGGGTGTGCTGTTTCGGGTTCGGGCGGGGCTCACCGTCGAGACCGTTGTGCTGCACAGTGTC
CCCATCAAAGCTTGAAGTGAGAACTCCATCAGTATGTTTAGTAGTTTCAGTAACTTTATGTTGTATGCTTT
CACAGTTGAGTTATTGGTTGATCGTGTGAAGGTTTAGTTTTGTCAATTGAGTTTAAGGCATCGTGCCTTT
TCTCTTATGACGTCACCAAACCTGGGCAACGCTTTGTGTTTATGCATAAATTCTTGGAATTTGAGAAAG
TAGTAAATTTGT
```

Summary

- Flat files can be parsed by PERL.
- Relational DB
 - Relational DB are made of table
 - In each entry (row) of a table, at least one field (column) must be unique.
 - Fields (columns) can be shared by different tables.
 - Different tables can be connected through shared fields.
- PubMed search
 - Use of tags
- GenBank sequence format
- FASTA sequence format

Reading assignment

–PubMed tutorial

<http://www.nlm.nih.gov/bsd/disted/pubmed.html>

–Chapter 2, Xiong Book

END