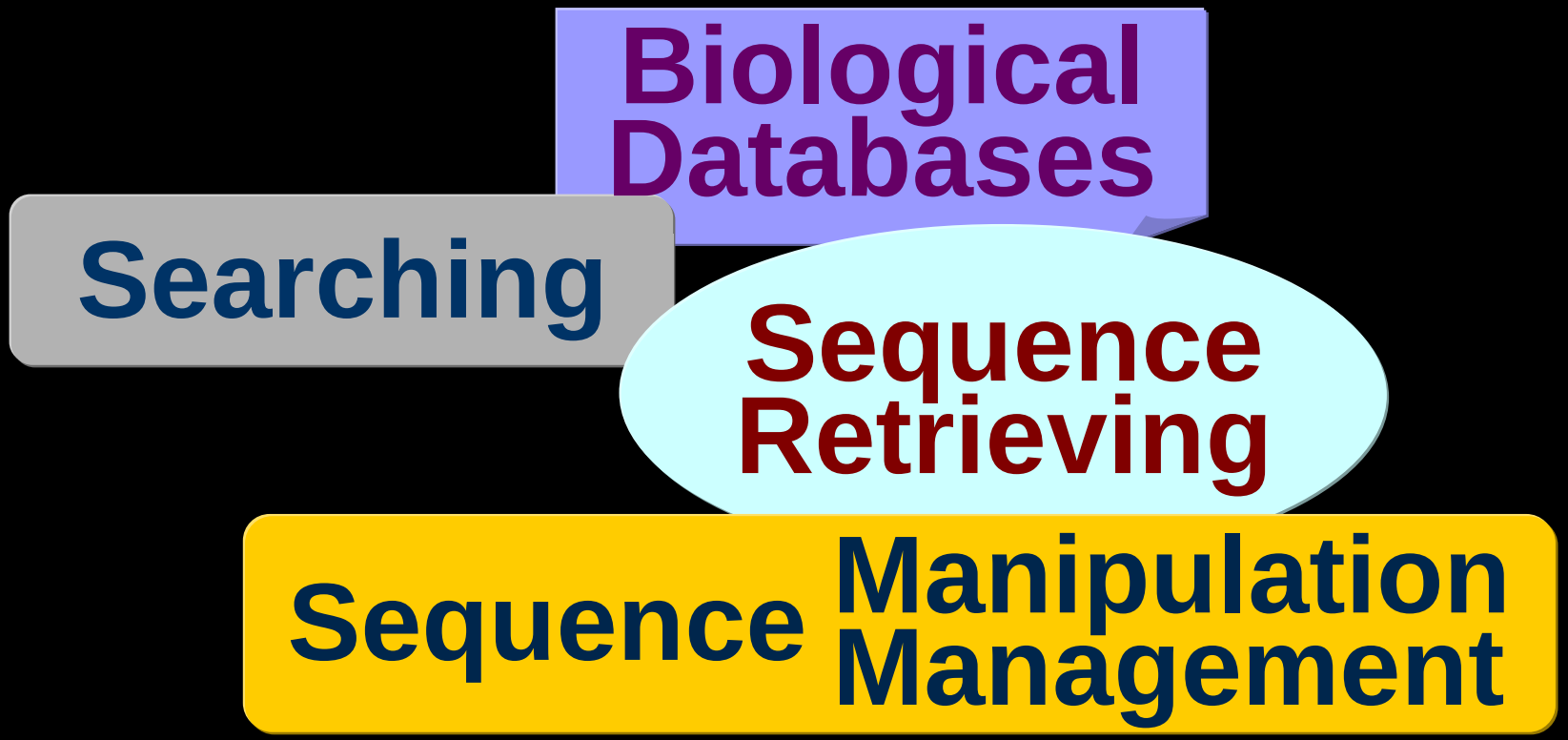
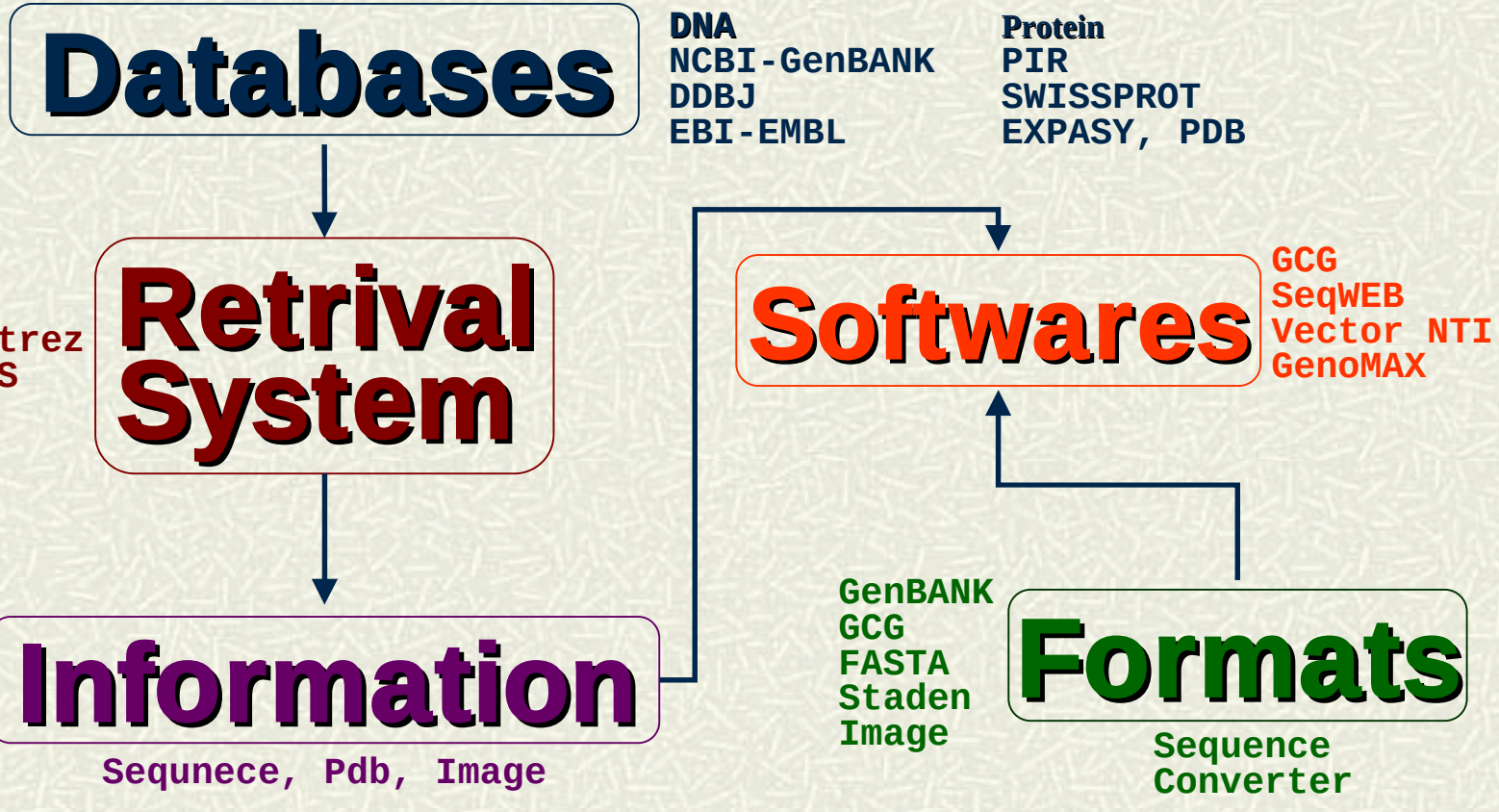
Bioinformatics 91-2

lecture 03+04



A Sequence Retrieving and Manipulation Network



Why Sequence Databases?

Available to scientists.

As much as possible of a particular type of information should be available in one single place.

Published data may be difficult to find or access, and collecting it from the literature is very time-consuming.

And not all data is actually published explicitly in an article (genome sequences!).

Available in computer-readable form.

Since analysis of biological data almost always involves computers, having the data in computer-readable form (rather than printed on paper) is a necessary first step.

The Protein Information Resource (PIR) is a division of the National Biomedical Research Foundation (NBRF) which is affiliated with Georgetown University Medical Center. The Resource was established in 1984.

http://pir.georgetown.edu/

PIR Protein Information Resource

About PIR | Databases | Search and Retrieval | Download | Support

An Integrated Public Resource of Protein Informatics to Support Genomic and Proteomic Research and Scientific Discovery

PIR produces the **Protein Sequence Database (PSD)** of functionally annotated protein sequences, which grew out of the *Atlas of Protein Sequence and Structure* (1965-1978) edited by Margaret Dayhoff and has been incorporated into an integrated knowledge base system of value-added databases and analytical tools.

***i*ProClass**, a central point for exploration of protein information, provides summary descriptions of protein family, function and structure for PIR-PSD, Swiss-Prot, and TrEMBL sequences, with links to over 50 biological databases . Release 2.17, 3-Mar-2003 contains 876,174 entries

PIR-NREF, a comprehensive database for sequence searching and protein identification, contains non-redundant protein sequences from PIR-PSD, Swiss-Prot, TrEMBL, RefSeq, GenPept, and PDB. Release 1.17, 3-Mar-2003 contains 1,159,203 entries.



PIR News Flash

View the PIR UNIPROT Press Release

Text Search Protein Databases: GO!

Find an Exact Peptide Match: GO!

Type in a string of single letter amino acid code (at least 3 letters)

PIR has recently joined forces with EBI (European Bioinformatics Institute) and SIB (Swiss Institute of Bioinformatics) to establish the UniProt (United Protein Databases), the central resource of protein sequence and function.

Primary & Secondary Databases

Primary databases experimental results directly into database

Nucleotide sequences
Protein sequences
Macromolecular 3D structure

Proteins sequence patterns or motifs
Gene expression data
Metabolic pathways
Regulatory pathways

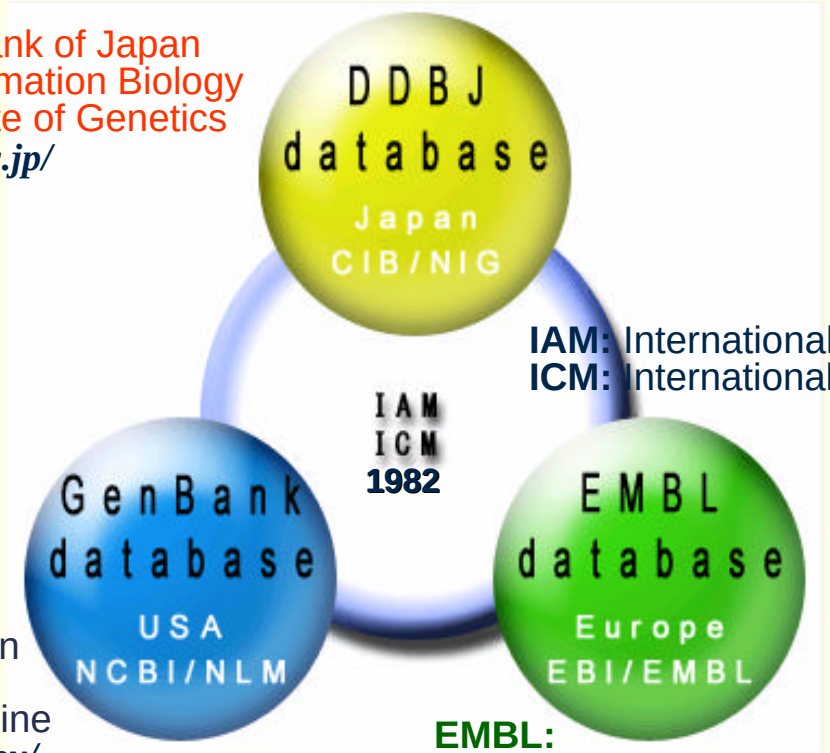
Secondary databases results of analysis of primary databases

Aggregate of many databases
Links to other data items
Combination of data
Consolidation of data

The International Nucleotide Sequence Database Collaboration

The entries in the EMBL, GenBank and DDBJ databases are **synchronized on a daily basis**, and the accession numbers are managed in a consistent manner between these three centers.

DDBJ: DNA Data Bank of Japan
CIB: Center for Information Biology
NIG: National Institute of Genetics
http://www.ddbj.nig.ac.jp/



IAM: International Advisory Meeting
ICM: International Collaborative Meeting

NCBI:
National Center for Biotechnology Information
NLM:
National Library of Medicine
http://www.ncbi.nlm.nih.gov/

EMBL:
European Molecular Biology Laboratory
EBI:
European Bioinformatics Institute
http://www.ebi.ac.uk

NCBI : GenBANK *http://www.ncbi.nlm.nih.gov*

GenBank:
An annotated collection of all publicly available nucleotide and amino acid sequences.

EST database:
A collection of expressed sequence tags, or short, single-pass sequence reads from mRNA (cDNA).

GSS database:
A database of genome survey sequences, or short, single pass genomic sequences.

HTG database:
A collection of high throughput genome sequences from large-scale genome sequencing centers; including unfinished and finished sequences.

SNPs database:
A central repository for both single base nucleotide substitutions and short deletion and insertion polymorphisms.

RefSeq:
A database of non-redundant reference sequences standards, including genomic DNA contigs, mRNAs and proteins for known genes. Multiple collaborations, both within NCBI and with external groups, support our data-gathering efforts.

STS database:
A database of sequence tagged sites; or short sequences that are operationally unique in the genome.

UniSTS:
A unified, non-redundant view of sequence tagged sites (STSs).

UniGene:
A collection of ESTs and full-length mRNA sequences organized into clusters, each representing a unique known or putative human gene annotated with mapping and expression information and cross-references to other sources.

EMBL
European Bioinformatics Institute



EBI:EMBL http://www.ebi.ac.uk

Nucleotide Sequence Databases

EMBL Information
EMBL Nucleotide Sequence Database information.
EMBL-Align database
EMBL-Align multiple sequence alignment database
Ensembl
Automatic annotation of eukaryotic genomes
dbEST and dbSTS Queries
Query dbEST and dbSTS.
EMEST
EMEST is a database of EST sequences.
EuroGeneIndexes
A database of EST alignments and clusters
MitBase Server
Mitochondrial DNA database server
IMGT
ImMunoGeneTics database.
EDGP
European Drosophila Genome Project server.
Parasites
Parasite Genome Databases
Mutations
Sequence variation database project.
Genomes Server
An overview of Completed Genomes at the EBI
Genome MOT
Genome Monitoring Table.

Protein Sequence Databases

SWISS-PROT
TrEMBL
InterPro

Sequence Structure Classification Databases

DSSP
Database of Secondary Structure Assignments.
HSSP
Homology Derived Secondary Structure Assignments.
FSSP
Fold Classification based on Structure-Structure Assignments.
DALI
Protein Structure Domain Dictionary
3Dee
Database of protein domain definitions.

Macromolecular Structure Databases

EBI-MSD
The EBI-Macromolecular Structure Database.

Sequence Mapping Databases

RHdb Server
Radiation Hybrid Database server.
GenomeMaps 98
Human Genome Maps 98.



DDBJ *http://www.ddbj.nig.ac.jp*

DDBJ (DNA Data Bank of Japan) began DNA data bank activities in earnest in 1986 at the National Institute of Genetics (NIG) with the endorsement of the Ministry of Education, Science, Sport and Culture. From the beginning, DDBJ has been functioning as one of the International DNA Databases, including EBI (European Bioinformatics Institute; responsible for the EMBL database) in Europe and NCBI (National Center for Biotechnology Information; responsible for GenBank database) in the USA as the two other members. Consequently, we have been collaborating with the two data banks through exchanging data and information on Internet and by regularly holding two meetings, the International DNA Data Banks Advisory Meeting and the International DNA Data Banks Collaborative Meeting.

DDBJ	15016100	22/1/02	DDBJNEW	1490104	14/3/02
DAD	945852	28/1/02	DADNEW	97212	14/3/02
SWISSPROT	105586	2/3/02	PIR	262528	11/12/01
PROSITE	1517	14/3/02	PROSITEDOC	1122	14/3/02
BLOCKS	4034	6/3/01	PRINTS	1050	6/3/01
PFAMA	2008	6/3/01	PFAMB	39228	6/3/01
SWISSPFAM	223208	6/3/01	PFAMHMM	2008	6/3/01
PFAMSEED	2008	6/3/01	PRODOM	149606	6/3/01
ENZYME	3869	29/10/01	PDB	17568	14/3/02
HSSP	15508	12/2/02	FSSP	2860	5/11/01
PATHWAY	7473	14/3/02	LENZYME	3829	13/3/02
LCOMPOUND	10158	13/3/02	SRSFAQ	10	6/3/01



Protein Databases

Protein Information Resources (PIR) *http://pir.georgetown.edu/*



In 1988, The Protein Information Resource (PIR), established a cooperative effort with the Munich Information Center for Protein Sequences (MIPS) and the Japan International Protein Information Database (JIPID) , produces the PIR-International . Protein Sequence Database (PIR-PSD) -- a comprehensive, non-redundant, expertly annotated, fully classified and extensively cross-referenced protein sequence database in the public domain. The PIR-PSD, PIR-NREF, iProClass and other PIR auxiliary databases provide an integration of sequences, functional, and structural information to support genomics and proteomics research
The PIR-PSD, Current Release 71.04, March 01, 2002, Contains 283153 Entries

SWISSPROT *http://www.ebi.ac.uk/swissprot/*



The SWISS-PROT Protein Knowledgebase is an annotated protein sequence database established in 1986. It is maintained collaboratively by the Swiss Institute for Bioinformatics (SIB) and the European Bioinformatics Institute (EBI).

Protein Databases

ExPASY Molecular Biology Server *http://tw.expasy.org*



The ExPASy (**Ex**pert **P**rotein **A**nalysis **Sy**stem) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE

Protein Data Bank *http://www.rcsb.org*



The Protein Data Bank (PDB) is operated by Rutgers, The State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; and the National Institute of Standards and Technology -- three members of the Research Collaboratory for Structural Bioinformatics (RCSB). The PDB is supported by funds from the National Science Foundation, the Department of Energy, and two units of the National Institutes of Health: the National Institute of General Medical Sciences and the National Library of Medicine.

Entrez
search and retrieval system

NCBI

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

Entrez

The Entrez system developed and accessible at the NCBI Entrez site. Similar to the SRS system, it provides search facilities for a large number of databases, and provides links between them. It provides a well-defined web interface which allows programs or web sites to define links that will query Entrez.
However, it appears that the Entrez system is not available to set up at one's own server. It is purely a system for accessing and searching the databases at NCBI.

Entrez
search and retrieval system

NCBI

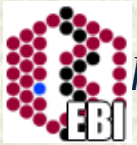
http://www.ncbi.nlm.nih.gov/Entrez/

Entrez is a retrieval system for searching several linked databases. It provides access to:

PubMed: The biomedical literature (PubMed)
Nucleotide sequence database (Genbank)
Protein sequence database
Structure: three-dimensional macromolecular structures
Genome: complete genome assemblies
PopSet: population study data sets
OMIM: Online Mendelian Inheritance in Man
Taxonomy: organisms in GenBank
Books: online books
ProbeSet: gene expression and microarray datasets
3D Domains: domains from Entrez Structure
UniSTS: markers and mapping data
SNP: single nucleotide polymorphisms
CDD: conserved domains

Database Interlinking

SRS



http://srs.ebi.ac.uk/

DDBJ *http://srs.ddbj.nig.ac.jp/*

The Sequence Retrieval System (SRS) developed by Thure Etzold is a system for integrating heterogenous databases. It is based on premade indexes of the items (words, entries, data fields, text,...) found in a set of documents (database files). Apart from the database files themselves, the indexing procedure requires a grammar (Icarus) that describes what different words in the data files mean, how they are to be indexed, and how they cross-reference to other items in other databases. SRS is a web-oriented system located on a server which is accessed through HTML pages and CGI scripts.

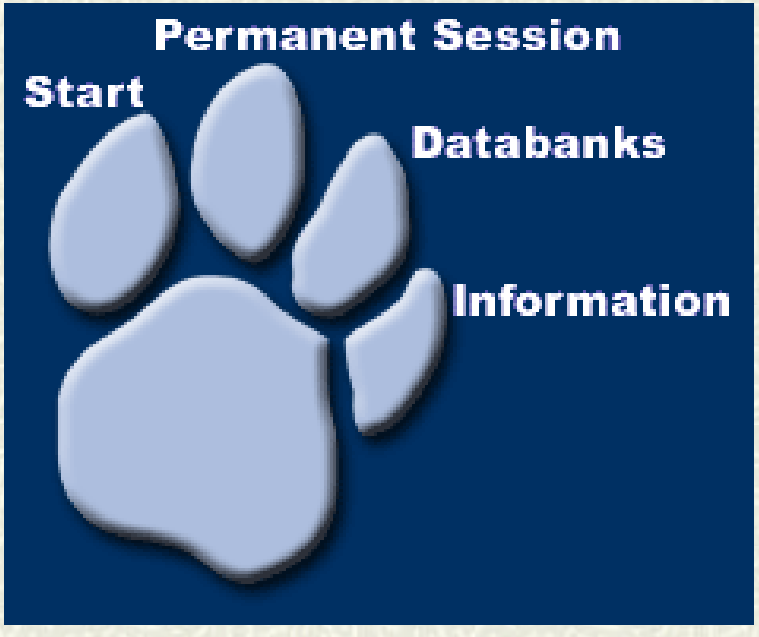
SRS started as an academic project, but it is now a commercial system developed and marketed by LION Bioscience AG. However, academic groups can license the SRS system free of charge and set it up at a server in their own lab.

EBI runs an SRS service which can be used by anyone. It indexes a large number of databases, and it also provides a well-defined web interface which allows programs or web sites to create links that query SRS at EBI.

SRS

http://srs.ebi.ac.uk/

DDBJ *http://srs.ddbj.nig.ac.jp/*



http://www.lionbioscience.com/

EMBL Nucleotide Database – Europe's primary collection of nucleotide sequences is maintained in collaboration with Genbank (USA) and DDBJ (Japan)

SWISS-PROT – A complete annotated protein sequence database

Macromolecular Structure Database - European Project for the management and distribution of data on macromolecular structures

ArrayExpress - for gene expression data

ENSEMBL - Metazoic genomes and the best possible automatic annotation.

What are you looking for?

You know NOTHING

Just Search

You know SOMETHING

Search and Compare

You've been scooped.
You've discovered a new member of a gene family.
Your sequence shares just some domains/motifs with known genes.
Your sequence is completely novel.

Bibliographic

NCBI PubMed/Medline

Sequence

NCBI-Entrez/EBI-SRS/DDBJ-SRS

Integrated

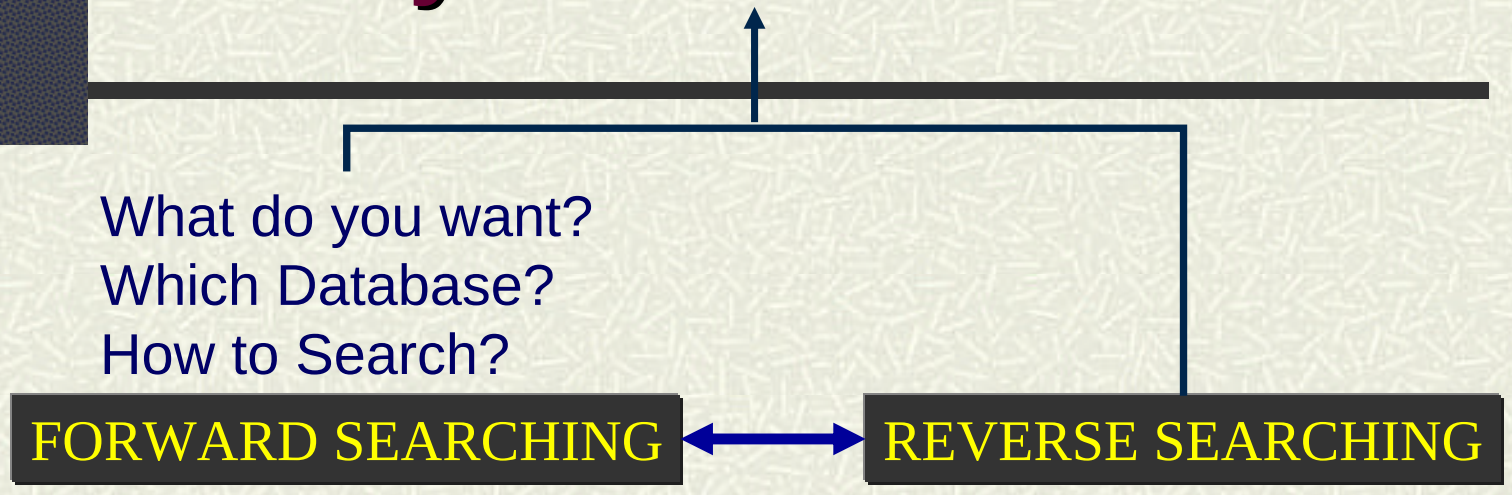
Entrez/PubMed

Entrez is a set of tightly linked databases, including nucleic sequences, protein sequences and MEDLINE. It has a very nice interface and is a very powerful and useful system. It has links to itself! In other words when you find an interesting nucleic sequence entry, you can quickly find others like it, the corresponding protein entry and abstracts of papers describing it .

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

Major Sequence Repositories
-Specilaized Sequence Databases
-Species Specific Databases
-Genome Sequencing Centers
-Human Chromosome Information
-RNA Databases
-Comparative, Phylogenetic and Evolutionary Databases
-Protein Sequence Databases
-Protein Structure Databases
-Protein Domains, Motifs and Signatures
-Other Databases

Built your own Database



How many sequences in the databases is homologous or identical to Yours

BLAST

(Basic Local Alignment Search Tool)

FastA

Entrez
search and retrieval system

NCBI

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

Entrez is a retrieval system for searching several linked databases. It provides access to:

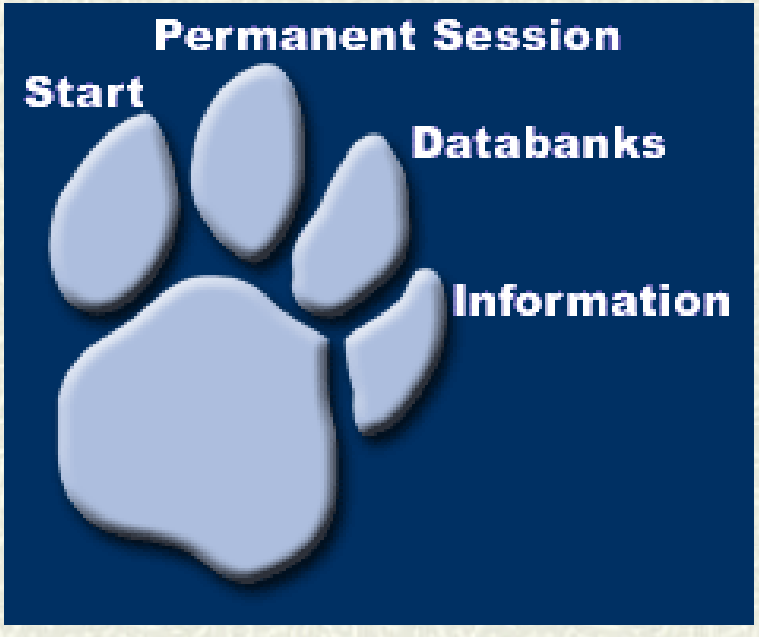
PubMed: The biomedical literature (PubMed)
Nucleotide sequence database (Genbank)
Protein sequence database
Structure: three-dimensional macromolecular structures
Genome: complete genome assemblies
PopSet: population study data sets
OMIM: Online Mendelian Inheritance in Man
Taxonomy: organisms in GenBank
Books: online books
ProbeSet: gene expression and microarray datasets
3D Domains: domains from Entrez Structure
UniSTS: markers and mapping data
SNP: single nucleotide polymorphisms
CDD: conserved domains

Database Interlinking

SRS

http://srs.ebi.ac.uk/

DDBJ *http://srs.ddbj.nig.ac.jp/*



http://www.lionbioscience.com/

EMBL Nucleotide Database – Europe's primary collection of nucleotide sequences is maintained in collaboration with Genbank (USA) and DDBJ (Japan)

SWISS-PROT – A complete annotated protein sequence database

Macromolecular Structure Database - European Project for the management and distribution of data on macromolecular structures

ArrayExpress - for gene expression data

ENSEMBL - Metazoic genomes and the best possible automatic annotation.

Exercise04-01

Use the database searching techniques you learned today to retrieve the reference sequence and the amino acid sequence of

Homo sapiens LEGUMAIN

From NCBI and EMBL
And then transfer the sequence(s) to
1. SeqWEB and
2. GCG Unix (in GCG format)

There are many different ways to DO it.
You can have your lunch now if you can make it.

Softwares & Sequence Formats

Program	Formats Default	Accept	Multiple sequence
WWW	text file	paste & Copy	
SeqWEB	text file	paste & copy	
GCG	GCG file	FASTA GenBANK EMBL Staden SwissProt	Multiple sequence file (msf) Rich sequence file (rsf) List files (lst)
VectorNTI	*.gb *.gp	FASTA GenBANK SwissProt	FASTA GenBank SwissProt



SEQWeb™
Web-based Sequence Analysis

The Sequence Manager in SeqWEB

http://gcg.nhri.org.tw:8003

SeqWeb Version 2

What is Sequence Manager?

The Sequence Manager lets you load and manage sequences in SeqWeb.

From the Sequence Manager you can load new sequences into SeqWeb as well as

retrieve,
create,
edit and document,
copy,
view,
delete, and
save sequences

Source of Sequences

Personal Sequences - Create, Edit and Add

You can add personal sequences to SeqWeb in three ways:
(1)You can specify a local file on your personal computer and upload it to the SeqWeb server,
(2) You can copy and paste a sequence into SeqWeb, or
(3) You can create a new sequence in SeqWeb.

Database Sequences - Retrieve and Loading

SeqWeb provides DNA and protein databases. All DNA databases are a combination of sequences in GenBank and the EMBL Data Library. Due to the large duplication between GenBank and EMBL, GCG has eliminated EMBL sequence entries sharing the same primary accession number as sequences in GenBank.

Reference & Sequence Searching

SeqWeb http:/gcg.nhri.org.tw:8003/
http://saturn.nchc.gov.tw:8001/

LookUp

LookUp identifies sequence database entries by name, accession number, author, organism, keyword, title, reference, feature, definition, length, or date. The output is a list of sequences.

StringSearch

StringSearch identifies sequences by searching for character patterns such as "globin" or "human" in the sequence documentation.

Pubmed & Entrez http://www.ncbi.nlm.nih.gov/(Pubmed/Entrez)

Entrez provides integrated access to biomedical literature, nucleotide and protein sequences, along with 3-D protein structures, complete genomes and population study data sets. Pre-computed similarity searches are available for most database records producing a list of related sequences, structure neighbors, as well as related articles.

Exercise03-1

(A) Adding a local sequence file
(B) Copying and pasting a sequence from the clipboard
(C) Adding database sequencing
(D) Editing sequences

Go to the Bioinformatics Teaching WEB
Download **naq.txt and psq.txt**
for this exercise.

Sequence Management in GCG Command Mode

Retrieve Sequences in GCG

Fetch

Copies GCG sequences or data files from the GCG database Into your directory or displays them on your terminal screen.
Syntax: % fetch [-Infile=]database:acession number
Example: fetch gb:l10131

SeqEd

An interactive editor for entering and modifying sequences and for assembling parts of existing sequences into new genetic constructs

Importing and Exporting

You need a FTP program to transfer files between your PC and GCG.
The sequence file must be in “plain text” format.

Chopup: converts a non-GCG format sequence file containing lines longer than 511 characters and as long as 32,000 characterters into a new file containing no longer than 50 characters.

Breakup: reads a non-GCG format sequence file containing more than 350,000 sequence characterters and writes it as a set of separate, shorter, overlapping sequence files than can be analyzed by GCG.

Reformat: rewrites sequence files, scoring matrix files, or enzyme data files so than they can be read by GCG programs.

FromStaden/EMBL/GenBank/PIR/IG/Fasta
T0Staden/PIR/IG/FastA

Exercise 03-2

(A) Transfer sequence files from your PC to GCG
(B) Chopup the sequence
(C) Reformat the sequence
(D) Edit the sequence

Create a folder “BIO” in your hard disk
Start WsFTP (ftp://gcg.nhri.org.tw)
Upload “naq.txt” & “psq.txt” to GCG
Start Netterm
Start GCG
Chopup “naq.txt” & “psq.txt”
Reformat “naq.dat” or “psq.dat”
Cat “naq.txt” or “psq.txt”

Exercise 04-1

Sequence Manipulation in GCG UNIX

Use the database searching techniques you learned today to retrieve the reference sequence and the amino acid sequence of

***Homo sapiens* LEGUMAIN**

From NCBI and EMBL
And then transfer the sequence(s) to
1. SeqWEB and
2. GCG Unix (in GCG format)

There are many different ways to DO it.
You can have your lunch now if you can make it.

ASSIGNMENT 1.

Use the Entrez searching techniques you learned today to retrieve the **Reference sequence and the corresponding amino acid sequences of**

All the subclasses of *Homo sapiens* cyclophilin

Transfer the sequences to GCG Unix,
Transform the sequences to GCG format

E-mail
1. The steps (including URL of WWW sites) you used and
2. The sequences in GCG format as attached file to
petang@mail.cgu.edu.tw before 20 March 2003
****郵件主旨： ASS1 bioinfo – (學號)