The Importance of Biological Databases in Biological Discovery

助理教授：張學偉

高雄醫學大學　生物醫學暨環境生物學系

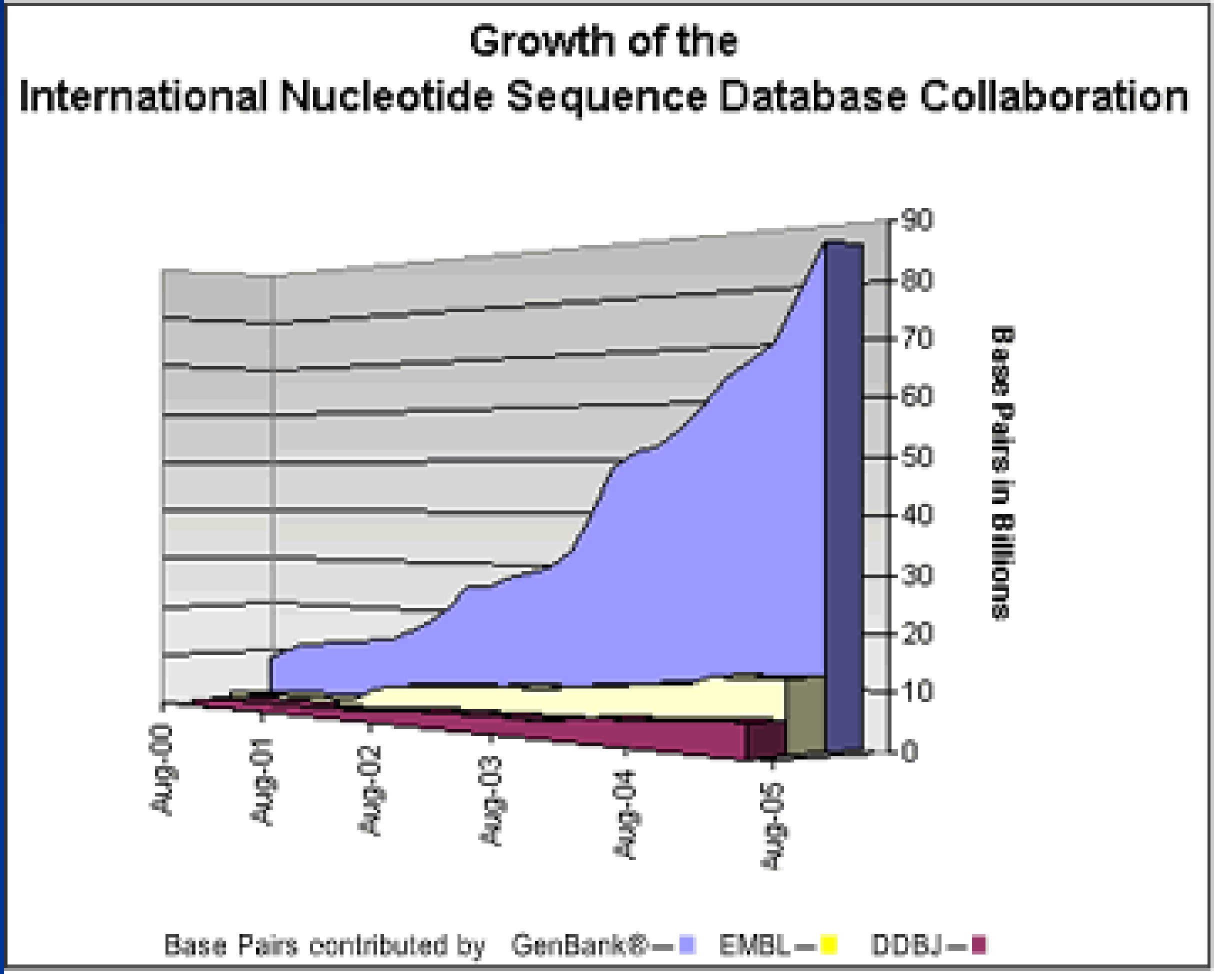
The database that most biologists are familiar with is GenBank.

- GenBank

-is the annotated collection of all publicly available ***DNA*** and ***protein*** sequences.

-is maintained by NCBI (National Center for Biotechnology) at NIH (National Institute of Health).

- -represents a collaborative effort between NCBI, EMBL (European Molecular Biology Laboratory) and DDBJ (DNA Data Bank of Japan)



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

- GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research 2007 Jan 1;35(Database issue):D16-20*).

- There are approximately 65,369,091,950 bases in 61,132,599 sequence records in the *traditional GenBank divisions* and 80,369,977,826 bases in 17,960,667 sequence records in the *WGS division* as of August 2006.

The growth of GenBank

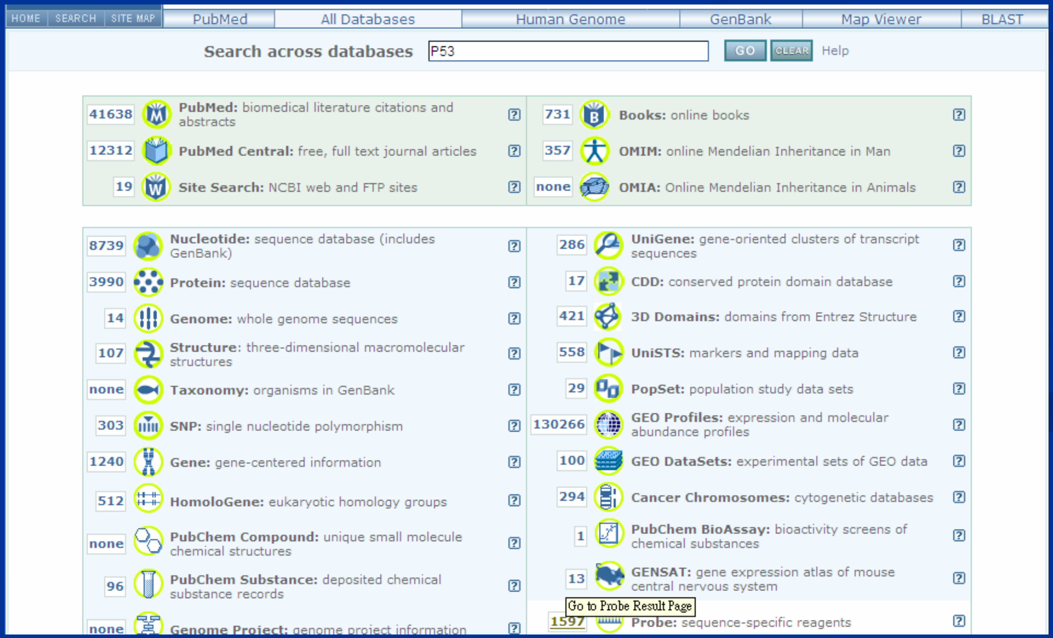
- Human Genome Project
- Other systematic sequencing projects

→ accumulation of sequence data

NCBI & Boolean Search

- Entrez homepage of NCBI

  is not a database itself but rather the interface through which all of its component databases can be accessed and traversed.



Search across databases P53

41638	**PubMed:** biomedical literature citations and abstracts
12312	**PubMed Central:** free, full text journal articles
19	**Site Search:** NCBI web and FTP sites
731	**Books:** online books
357	**OMIM:** online Mendelian Inheritance in Man
none	**OMIA:** Online Mendelian Inheritance in Animals

8739	**Nucleotide:** sequence database (includes GenBank)
3990	**Protein:** sequence database
14	**Genome:** whole genome sequences
107	**Structure**: three-dimensional macromolecular structures
none	**Taxonomy:** organisms in GenBank
303	**SNP:** single nucleotide polymorphism
1240	**Gene:** gene-centered information
512	**HomoloGene:** eukaryotic homology groups
none	**PubChem Compound:** unique small molecule chemical structures
96	**PubChem Substance:** deposited chemical substance records
none	**Genome Project:** genome project information
286	**UniGene:** gene-oriented clusters of transcript sequences
17	**CDD:** conserved protein domain database
421	**3D Domains:** domains from Entrez Structure
558	**UniSTS:** markers and mapping data
29	**PopSet:** population study data sets
130266	**GEO Profiles:** expression and molecular abundance profiles
100	**GEO DataSets:** experimental sets of GEO data
294	**Cancer Chromosomes:** cytogenetic databases
1	**PubChem BioAssay:** bioactivity screens of chemical substances
13	**GENSAT:** gene expression atlas of mouse central nervous system
1597	**Probe:** sequence-specific reagents

http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/Boolean.pdf

Boolean logic
From Wikipedia, the free encyclopedia

KEGG (Kyoto Encyclopedoa of Genes and Genomes)

- a widely-used compendium of biochemical pathways that has practical use in the modeling of expression data and in understanding higher-order cellular processes.

3. Useful links & tools- Glossary

- BioInformatics Glossary;
- Genome Glossary NCBI;
- -Omes and -omics glossary

Biological databases

- Like any other database
  - Data organization for optimal analysis

- Data is of different types
  - Raw data (DNA, RNA, protein sequences)
  - Curated data (DNA, RNA and protein annotated sequences and structures, expression data)

Characteristics of biological data

- Complex → Thoughtful data modeling

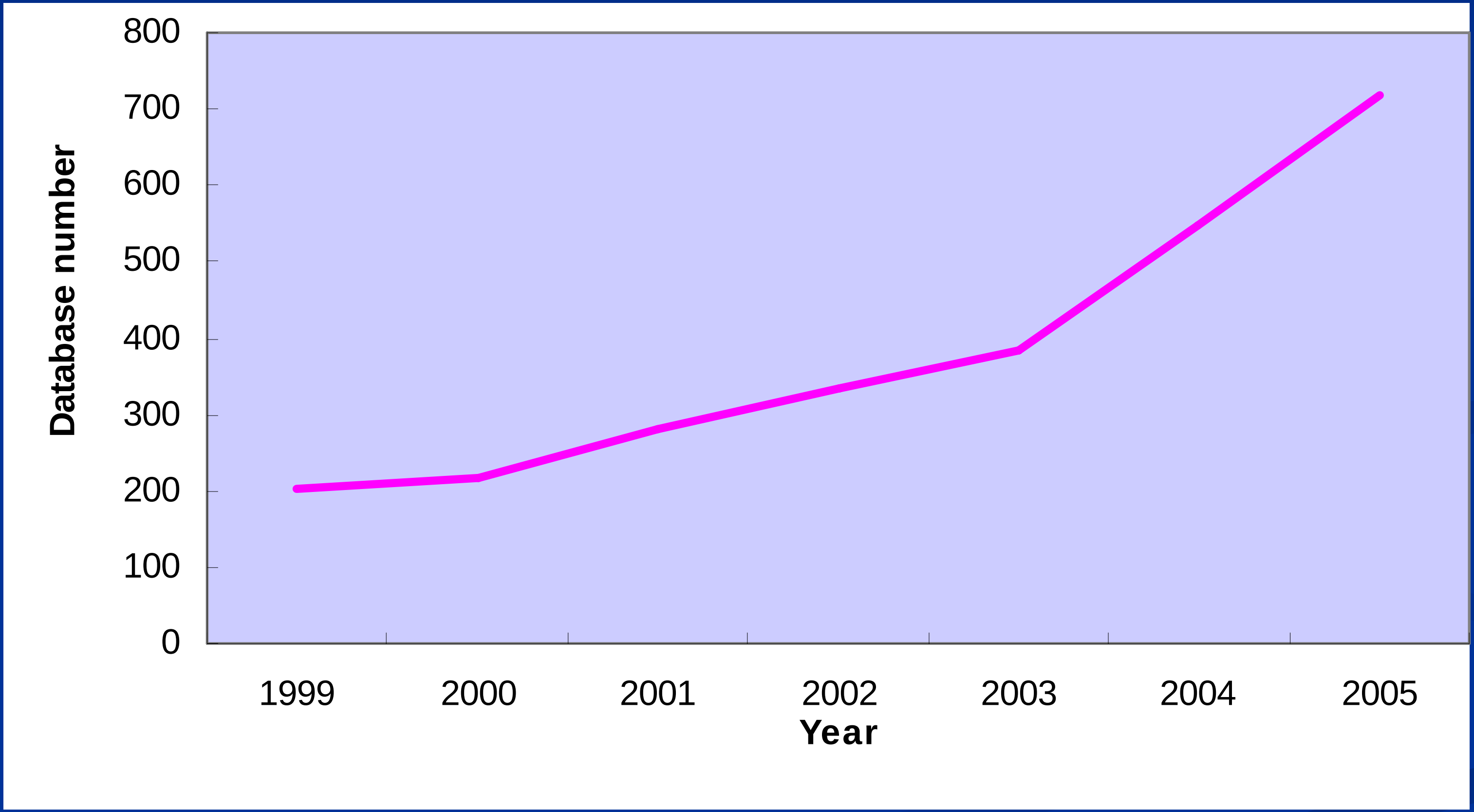
  Data types range from sequences, 3-dimensional structures, pathways, images, text, and a wide variety of annotation.
- Heterogeneous → Universal schema

  storage format, management, and access vary widely
- Dynamic → Flexible designing

  contents and schema change routinely and rapidly (twice/year)
- Inconsistent → Ontology

  lack standards at the ontology level
  - Controlled vocabulary for consistent naming for biomedical terms within and between databases
  - Data models for modeling or abstraction of biological system and processes

The growth of public domain bio-databases

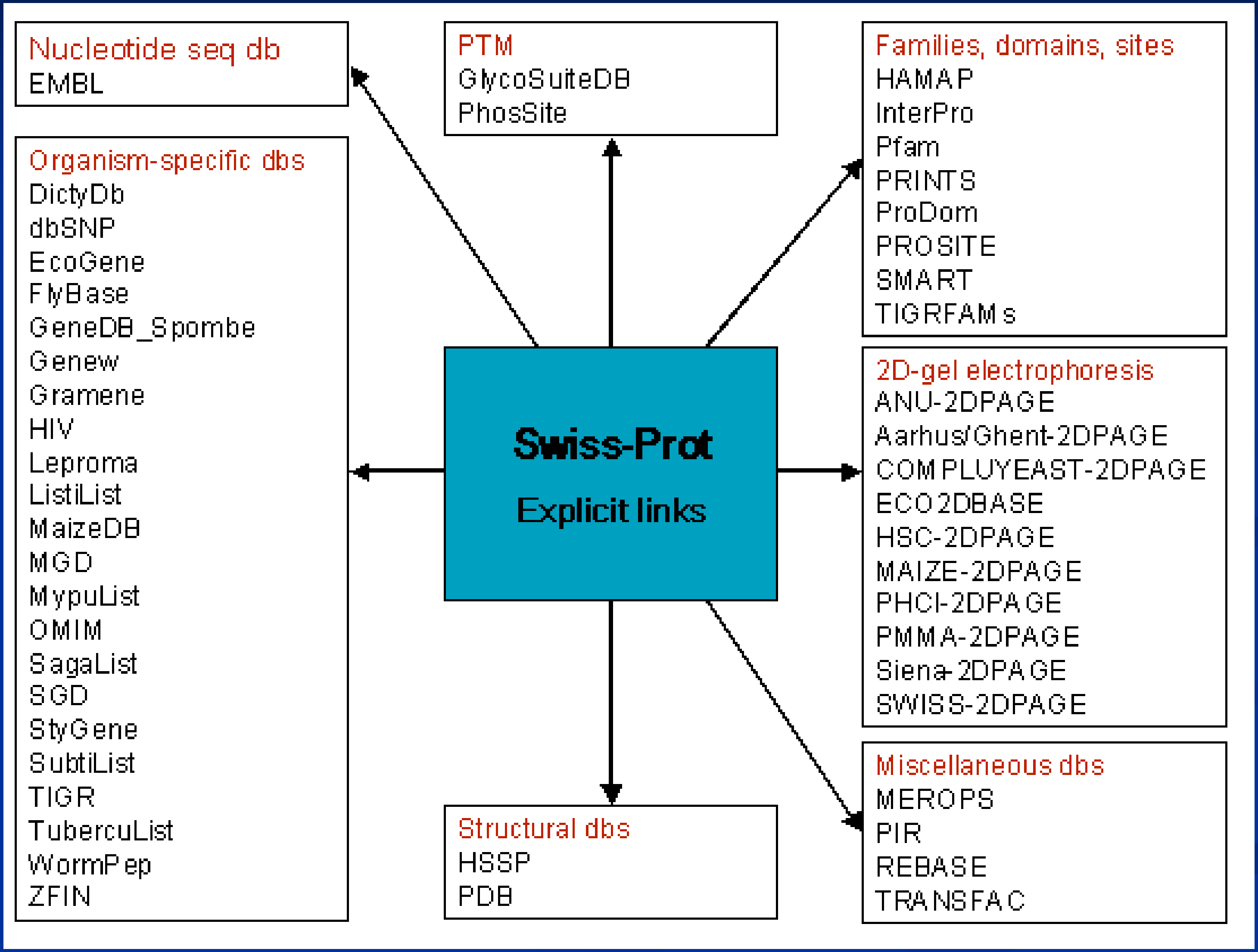


(The Molecular Biology Database Collection from *Nucleic Acids Research*)

Three main tendencies of bio-databases

- Database proliferation
  - Dozens to hundreds at the moment
- In the next 5 years biological data analysis will be trifurcated
  - Bio-webs : remote data analysis and mining
  - Bio-grids : transparent high-end computing
  - Bio-semantic webs : biological knowledge
- More and more scientific discoveries result from inter-database analysis and mining

Cross-references



Swiss-Prot

Explicit links

Nucleotide seq db
EMBL

PTM
GlycoSuiteDB
PhosSite

Families, domains, sites
HAMAP
InterPro
Pfam
PRINTS
ProDom
PROSITE
SMART
TIGRFAMs

Organism-specific dbs
DictyDb
dbSNP
EcoGene
FlyBase
GeneDB_Spombe
Genew
Gramene
HIV
Leproma
ListiList
MaizeDB
MGD
MypuList
OMIM
SagaList
SGD
StyGene
SubtiList
TIGR
TubercuList
WormPep
ZFIN

2D-gel electrophoresis
ANU-2DPAGE
Aarhus/Ghent-2DPAGE
COMPLUYEAST-2DPAGE
ECO2DBASE
HSC-2DPAGE
MAIZE-2DPAGE
PHCI-2DPAGE
PMMA-2DPAGE
Siena-2DPAGE
SWISS-2DPAGE

Structural dbs
HSSP
PDB

Miscellaneous dbs
MEROPS
PIR
REBASE
TRANSFAC

Gene Ontology database-1

"The Gene Ontology (GO) project seeks to provide a set of *structured vocabularies* for *specific biological domains* that can be used to describe gene products in any organism."

A few key points:

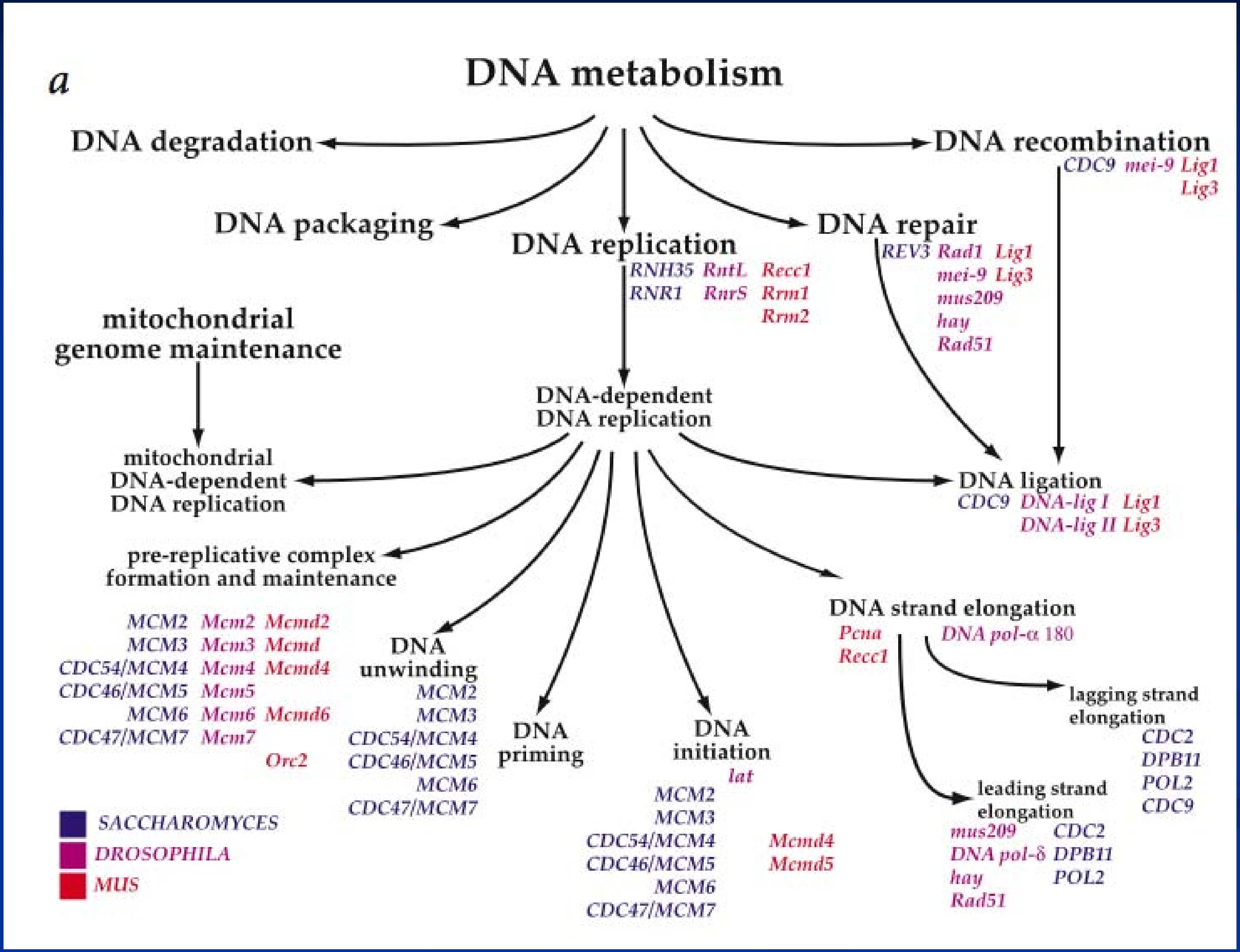
GO is a "structured" vocabulary, which is really a specialized type of a "controlled" vocabulary.

Gene Ontology database-2

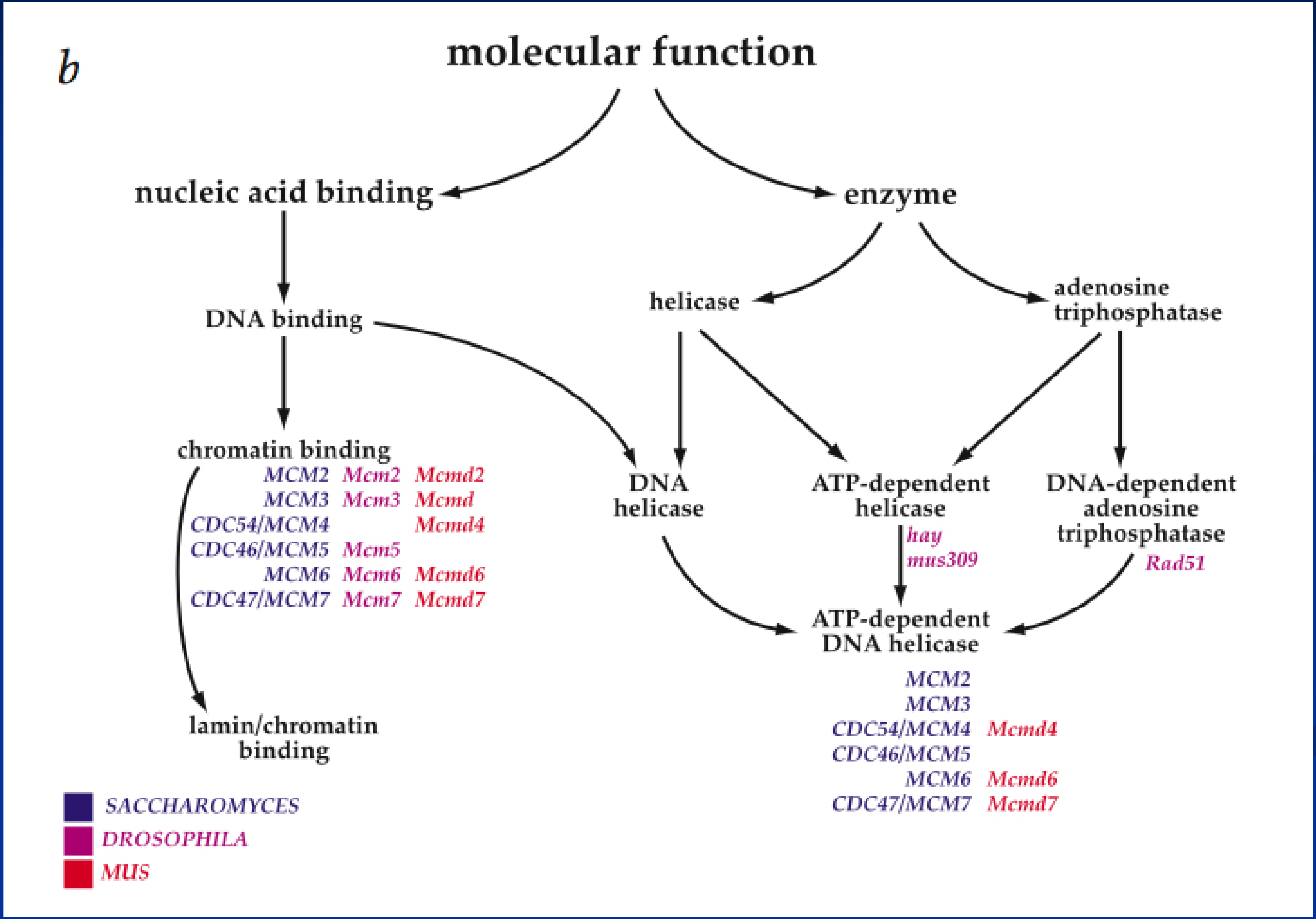
The ontologies in GO are intended to describe three biological areas, “molecular function”, “biological processes” and “cellular components”.

GO was *originally developed* through the collaboration of the members of *three model organism projects*: SGD, the *Saccharomyces* Genome database; FlyBase, the *Drosophila* genome database; and MGD/GXD, the Mouse Genome Informatics databases.

Biological Process Ontology



Molecular Function Ontology



b

molecular function

nucleic acid binding

enzyme

DNA binding

helicase

adenosine triphosphatase

chromatin binding

MCM2 Mcm2 Mcmd2
MCM3 Mcm3 Mcmd
CDC54/MCM4 Mcmd4
CDC46/MCM5 Mcm5
MCM6 Mcm6 Mcmd6
CDC47/MCM7 Mcm7 Mcmd7

DNA helicase

ATP-dependent helicase

hay
mus309

DNA-dependent adenosine triphosphatase

Rad51

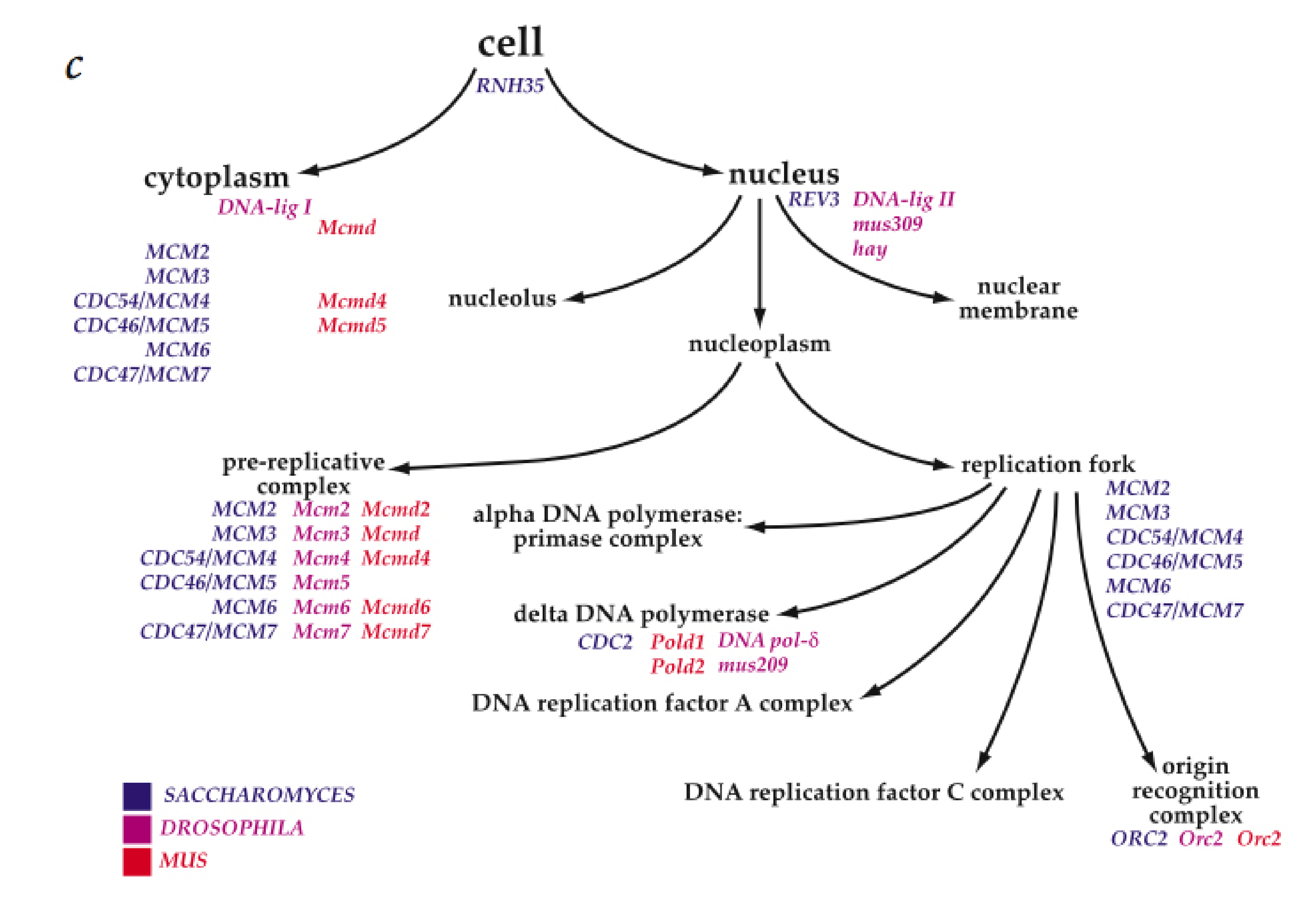
ATP-dependent DNA helicase

MCM2
MCM3
CDC54/MCM4 Mcmd4
CDC46/MCM5
MCM6 Mcmd6
CDC47/MCM7 Mcmd7

lamin/chromatin binding

🟪 *SACCHAROMYCES*
🟪 *DROSOPHILA*
🟥 *MUS*

Cellular Component Ontology



The CANCER GENOME ANATOMY PROJECT

- The Gene Ontology Browser (GO Browser) classifies human and mouse genes by molecular function, biological process, and cellular component.

What GO is Not

1. GO is ***not** a way* to unify biological databases. Sharing nomenclature is a step toward unification, but is not, in itself, sufficient.

2. GO is ***not*** a dictated standard, mandating nomenclature across databases. Groups participate because of self-interest and cooperate to arrive at a consensus.

3. GO does ***not*** define homologies between gene products from different organisms. The use of the GO results in shared annotations for gene products from different organisms, and this may reflect an evolutionary relationship, but the shared annotation is in itself not sufficient for such a determination.

Knowledge discovery in databases

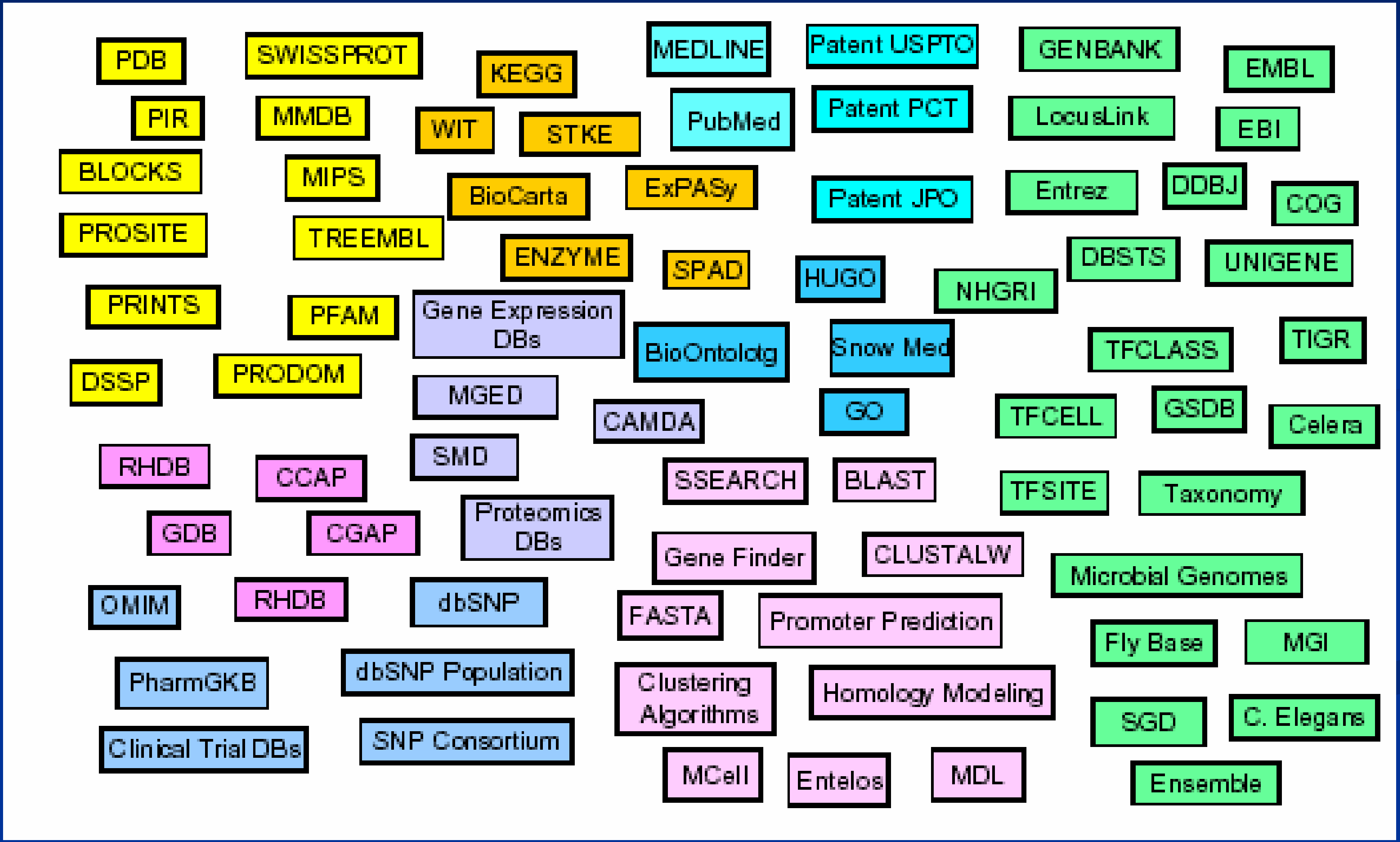
- **Data mining** is a technique to discover hidden information in large databases. This information, e.g. trends and patterns, can be used to build predictive models.

- Example: extracting predictive information of gene expression from genome sequence databases.

Bio-databases: a short word on problems

- Even today we face some key limitations
  - **There is no standard format**
    - Every database or program has its own format
  - **There is no standard nomenclature**
    - Every database has its own names
  - **Data is not fully optimized**
    - Some datasets have missing information without indications of it
  - **Data errors**
    - Data is sometimes of poor quality, erroneous, misspelled

Swimming in Data Sources



PDB

SWISSPROT

KEGG

MEDLINE

Patent USPTO

GENBANK

EMBL

PIR

MMDB

WIT

STKE

PubMed

Patent PCT

LocusLink

EBI

BLOCKS

MIPS

BioCarta

ExPASy

Patent JPO

Entrez

DDBJ

COG

PROSITE

TREEMBL

ENZYME

SPAD

HUGO

NHGRI

DBSTS

UNIGENE

PRINTS

PFAM

Gene Expression DBs

BioOntolog

Snow Med

TFCLASS

TIGR

DSSP

PRODOM

MGED

CAMDA

GO

TFCELL

GSDB

Celera

RHDB

CCAP

SMD

SSEARCH

BLAST

TFSITE

Taxonomy

GDB

CGAP

Proteomics DBs

Gene Finder

CLUSTALW

Microbial Genomes

OMIM

RHDB

dbSNP

FASTA

Promoter Prediction

Fly Base

MGI

PharmGKB

dbSNP Population

Clustering Algorithms

Homology Modeling

SGD

C. Elegans

Clinical Trial DBs

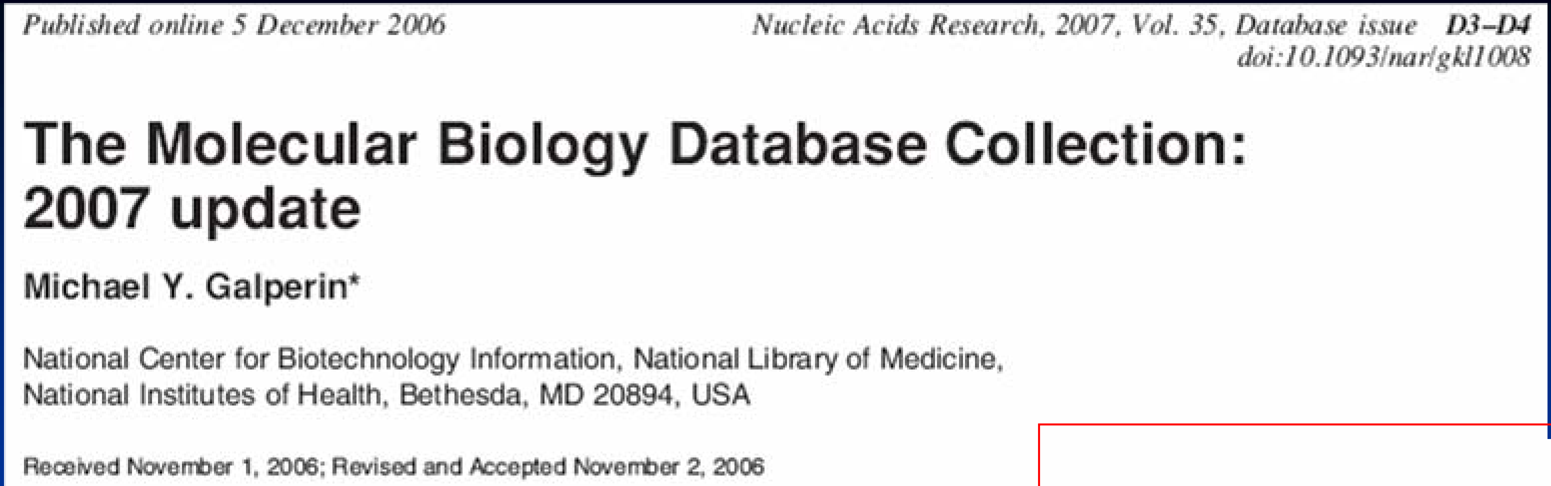
SNP Consortium

MCell

Entelos

MDL

Ensemble



Published online 5 December 2006

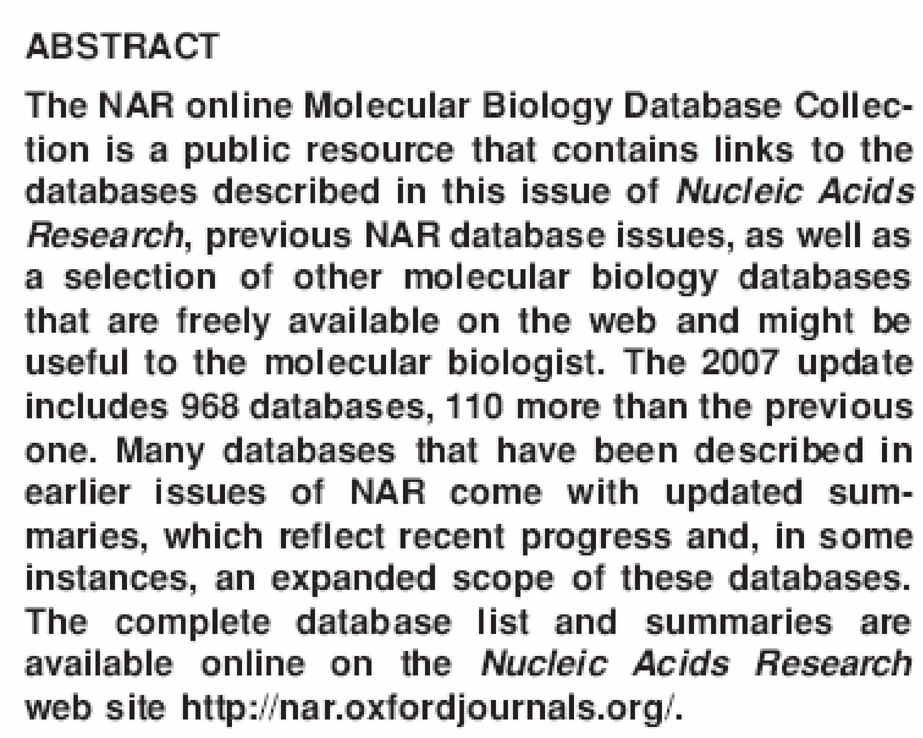
Nucleic Acids Research, 2007, Vol. 35, Database issue D3–D4
doi:10.1093/nar/gkl1008

The Molecular Biology Database Collection: 2007 update

Michael Y. Galperin*

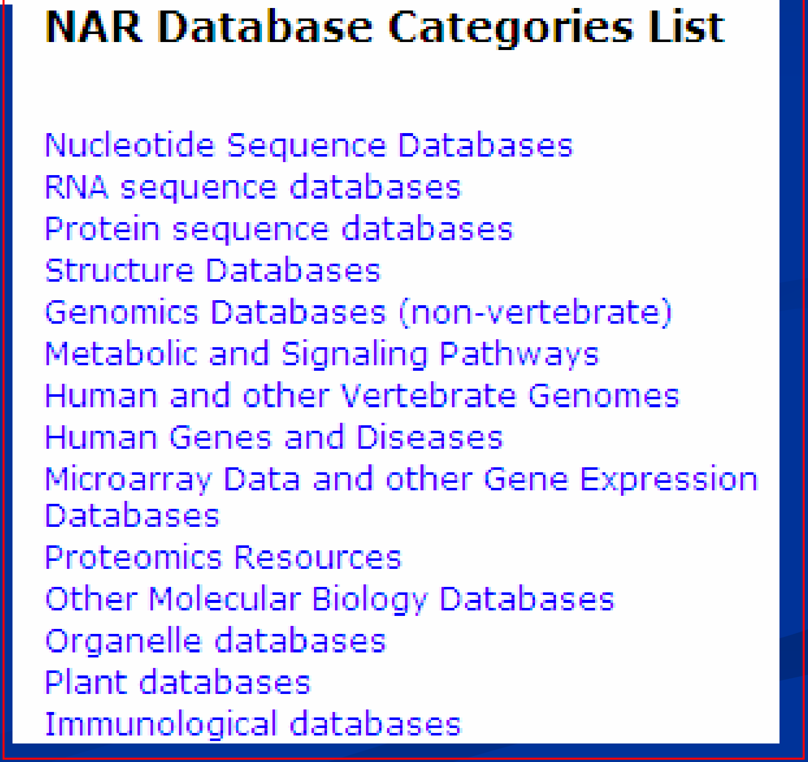
National Center for Biotechnology Information, National Library of Medicine,
National Institutes of Health, Bethesda, MD 20894, USA

Received November 1, 2006; Revised and Accepted November 2, 2006



ABSTRACT

The NAR online Molecular Biology Database Collection is a public resource that contains links to the databases described in this issue of *Nucleic Acids Research*, previous NAR database issues, as well as a selection of other molecular biology databases that are freely available on the web and might be useful to the molecular biologist. The 2007 update includes 968 databases, 110 more than the previous one. Many databases that have been described in earlier issues of NAR come with updated summaries, which reflect recent progress and, in some instances, an expanded scope of these databases. The complete database list and summaries are available online on the *Nucleic Acids Research* web site http://nar.oxfordjournals.org/.



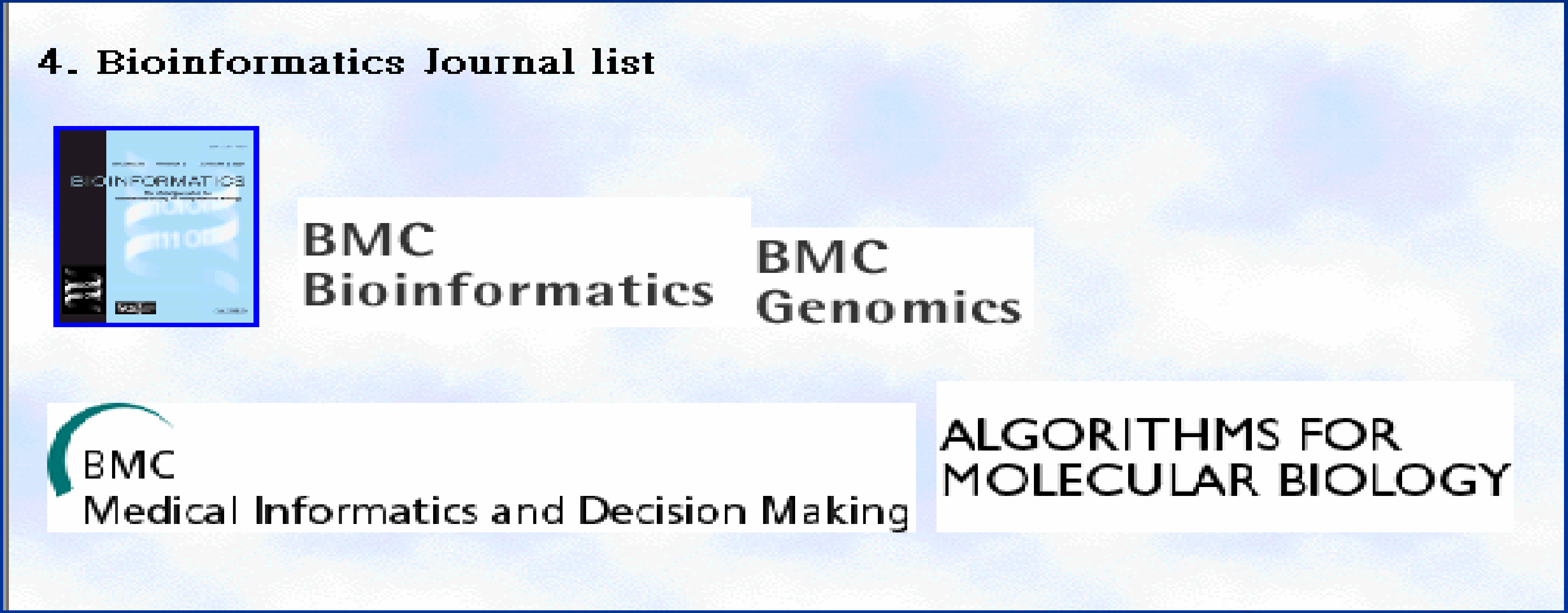
NAR Database Categories List

Nucleotide Sequence Databases
RNA sequence databases
Protein sequence databases
Structure Databases
Genomics Databases (non-vertebrate)
Metabolic and Signaling Pathways
Human and other Vertebrate Genomes
Human Genes and Diseases
Microarray Data and other Gene Expression Databases
Proteomics Resources
Other Molecular Biology Databases
Organelle databases
Plant databases
Immunological databases

http://nar.oxfordjournals.org/cgi/reprint/35/suppl_1/D3 Free download

Nucleic Acids Research

Database & server issues



4. Bioinformatics Journal list

BMC Bioinformatics

BMC Genomics

BMC Medical Informatics and Decision Making

ALGORITHMS FOR MOLECULAR BIOLOGY