Bioinformatics in *Monascus* Genome

食品工業發展研究所
生物資源保存及研究中心
副研究員 王俊霖

2005/12/05

Outline

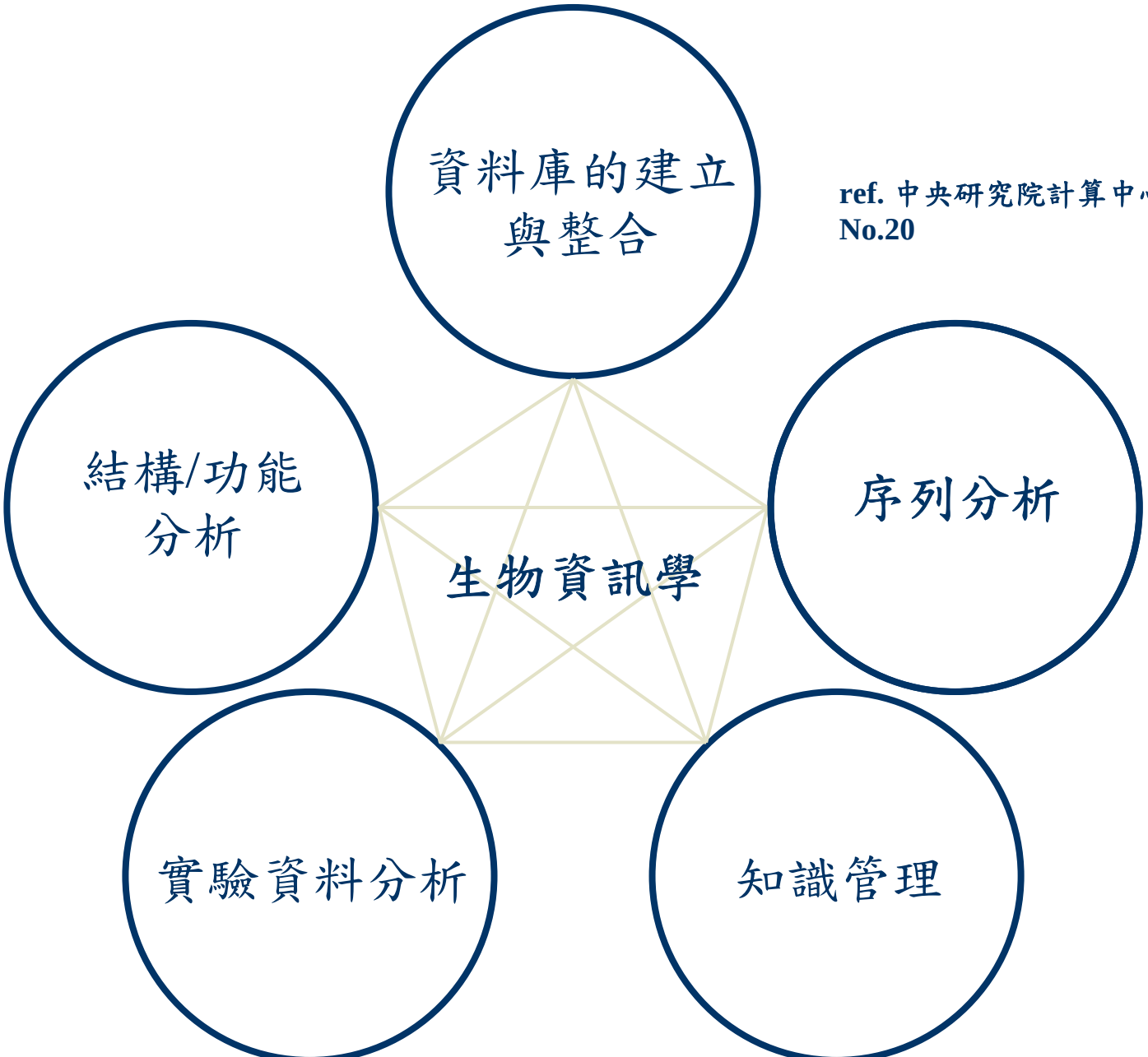
- Fields of Bioinformatics
- Genome Projects Today
- Introduction of Relational Database
- Bioinformatics in Monascus Genome

Bioinformatics

Bioinformatics

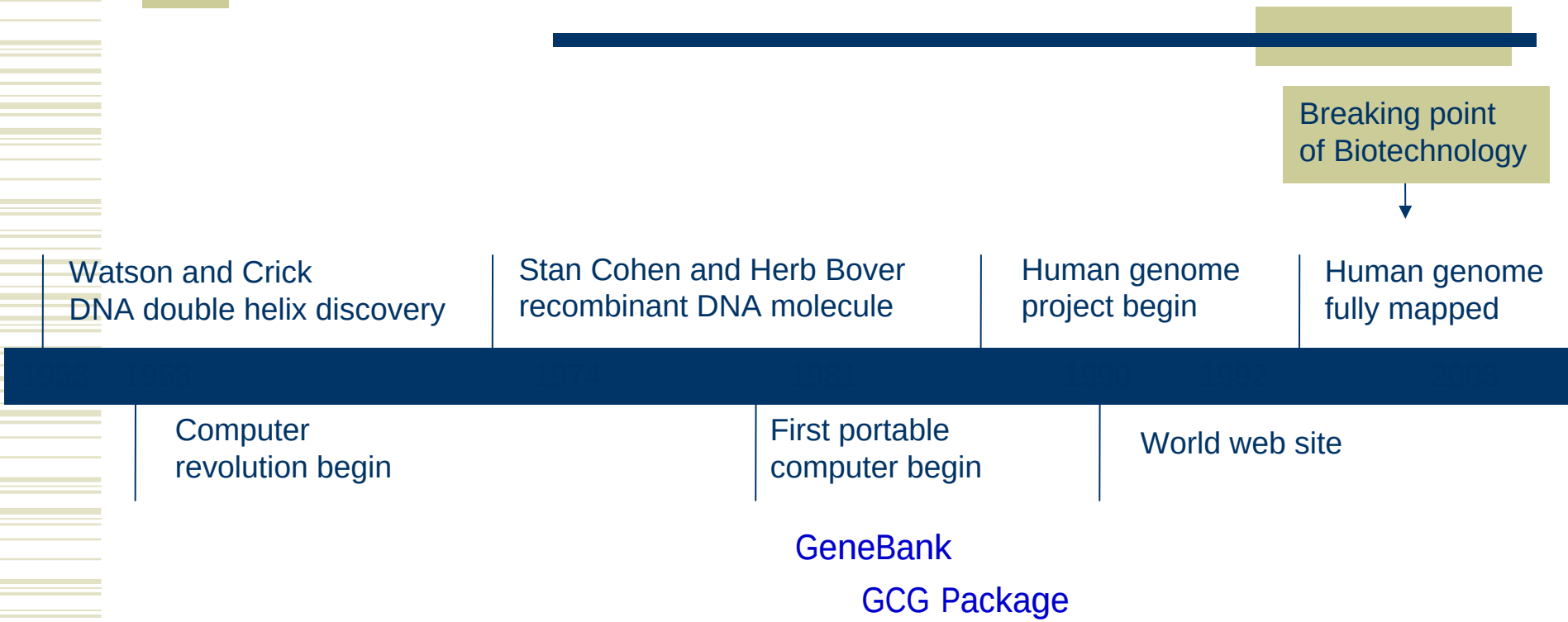
- Bioinformatics is the discipline of biology that has evolved to gather, store and manage in specialized databanks the vast amounts of biological data, which it then mines for knowledge

生物資訊的領域



ref. 中央研究院計算中心通訊 Vol.19 No.20

Biotech and Computer Science



The breaking point of Biotechnology is Human Genome Project

Bioinformatics- hot issues

- Genome Analysis
  - Pipeline Analysis
  - Genome Annotation
  - SNP
- Data warehouse/ Databases integration
- New Algorithm
- Literature Mining
- System Biology/ Microarray Analysis

Bioinformatics Applications in Genome Science

Raw Sequence data

- Aligned sequences
- Annotated sequences
- Repeat sequences
- Grouped sequences
- Isolated genes

- Genetic screening
- Expression profiling
- Novel gene discovery
- Electronic cloning
- Pharmacogenomics
- Structure biology
- Chromosomal localization
- Proteomics

- Drug discovery
- Disease diagnosis and prodiagnosis
- Clinical trial
- Gene mapping
- Gene therapy
- Customized medicine
- Population genomics
- DNA fingerprinting

Data → Information → Knowledge → Intelligence

Genome Projects Today

Genome Project

- Published Complete Genomes: 324(165)
  - prokayotic: 260(145)
  - eukaryotic: 40(20)
    - Human
    - *Mus musculus* **(**Mouse)
    - *Rattus norvegicus* (Rat)
    - Plant
    - Fish
    - Yeast
    - Fungi
    - Protozoa

Last Update: November 30, 2005(2003)

Genome Project

- Prokaryotic Ongoing Genomes: 802(415)
- Eukaryotic Ongoing Genomes: 548(360)
  - Fungi
  - Fish
  - Chicken
  - Animal: cow, dog, pig
  - Plant: barley, corn, cotton, rice, soybean, wheat

Source:GOLD, Genomes OnLine Database

Primary Online Genome Information Sources

- TIGR
- Sanger Center
- Broad Institute, MIT (Whitehead)
- NCBI / EBI COGENT
- NIH
- DOE / USDA

Model Organism Genome Projects

- Why Model Organism Genomes?
  - relatively well known
  - comparative between species
- Major Model Organisms
  - Ensembl – Human..etc.
  - RGD – Rat
  - FlyBase - Drosophia
  - WormBase – *C. elegan*

有夢最美 vs. 夢醒時分

- February 12, 2001-Celera Genomics (NYSE: CRA), an Applera Corporation business, today announced that its scientists have published an accurate assembly of the human genome and an initial interpretation of the sequence.
- 最實際的市場反應- 股價表現
- Why?
  - Genome decoding != understand the genome
  - Ex. The estimate number of genes in human genome

Relational Database

WHAT is a database?

- A collection of data that needs to be:
  - Structured
  - Searchable
  - Updated (periodically)
  - Cross referenced

- Challenge:
  - To change “meaningless” data into useful information that can be accessed and analysed the best way possible.

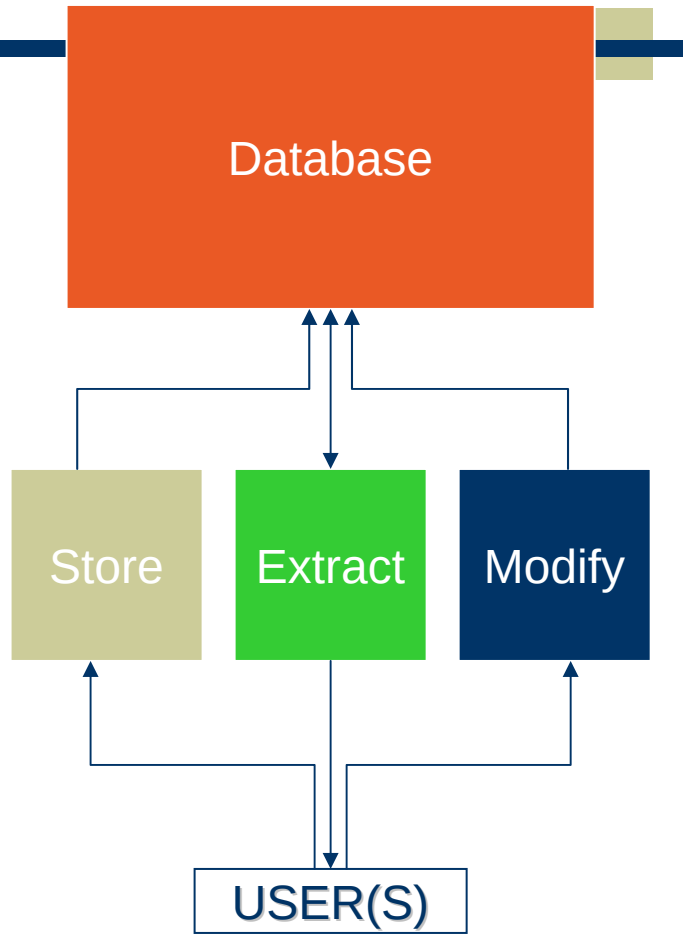
For example:

HOW would YOU organise all biological sequences so that the biological information is optimally accessible?

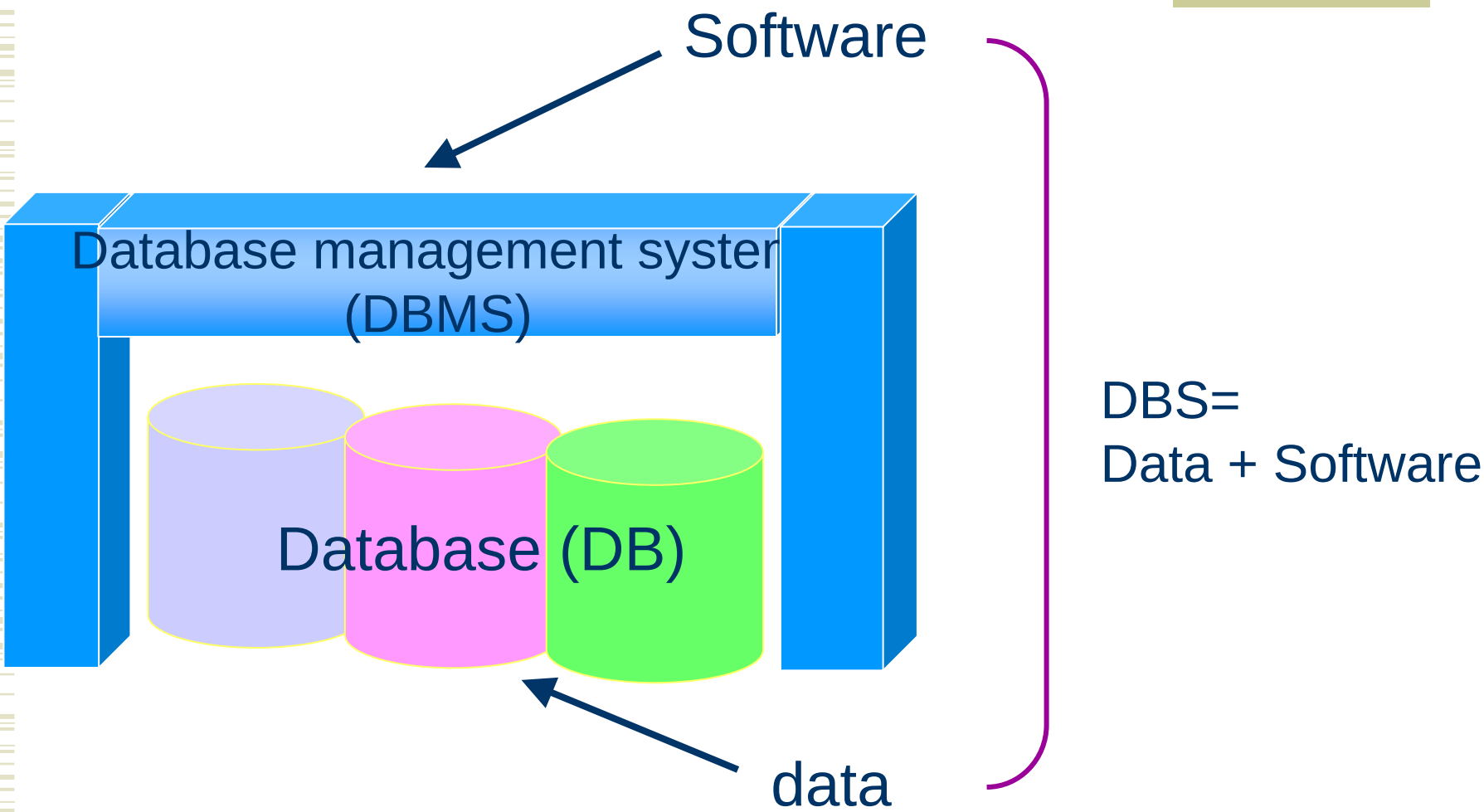
You need an appropriate data management system (DBMS)

DBMS

- Internal organization
  - Controls speed and flexibility
- A unity of programs that
  - Store
  - Extract
  - Modify



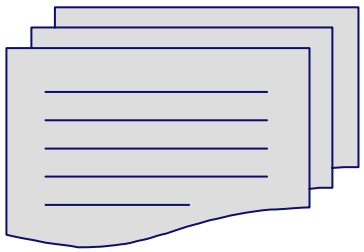
Database System



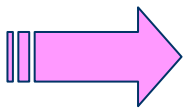
DBMS organisation types

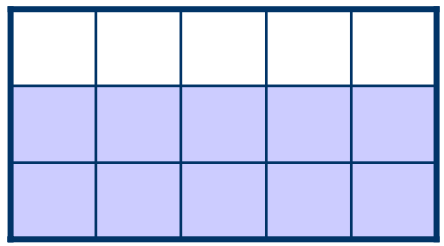
- Flat file databases (flat DBMS)
  - Simple, restrictive, table
- Relational databases (RDBMS)
  - Complex, versatile, tables
- Object-oriented databases (OODBMS)
  - Complex, versatile, objects
- Object-relational databases (ORDBMS)
  - Complex, versatile, objects

Construction process



files



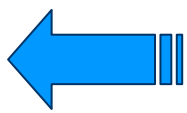


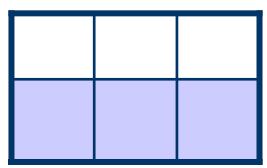
table

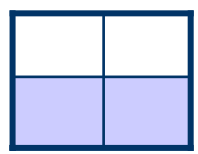
Depend on data attribute

Database software

database



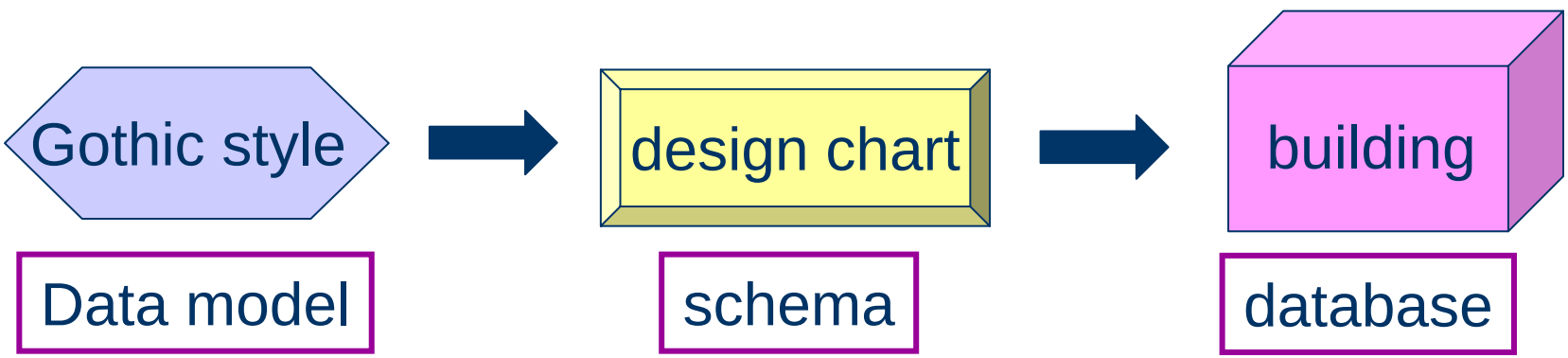




tables

Database construction

- A *data model* is a collection of concepts for describing data behavior.
  - Example: Relational data model: *relations*, *primary key*, *foreign key*, *view*, *constrain*, *etc.*….
- A *schema* is a description of a particular collection of data, using a given data model.



Our flat file database

FLAT DATABASE 2 Course details

Name	Deprtmnt	Course	E1	E2	E3	P1	P2	
Student 1	Chemistry	Biology	A	B	B	A	C	.....
Student 1	Chemistry	Maths	C	C	B	A	A	.....
Student 1	Chemistry	English	A	A	A	A	A	.....
.								
.								
.								
.								
Student 2	Ecology	Biology	A	B	A	A	A	.....
Student 2	Ecology	Maths	A	D	A	A	A	.....
.								
.								
.								
.								

Normalize (1NF) …

- We remove repeating records (rows)

sID	Name	dID
1	Student1	1
2	Student2	2

cID	Course
1	Biology
2	Maths
3	English

dID	Department
1	Chemistry
2	Ecology

Primary keys

sID	cID	E1	E2	E3	P1	P2	
1	1	A	B	B	A	C	.....
1	2	C	C	B	A	A	.....
1	3	A	A	A	A	A	.....
2	1	A	B	A	A	A	.....
2	2	A	D	A	A	A	.....

Normalize (2NF) …

- We remove redundant fields (columns)

sID	Name	dID
1	Student1	1
2	Student2	2

gID	Grade
1	A
2	B
3	C

dID	Department
1	Chemistry
2	Ecology

cID	Course
1	Biology
2	Maths
3	English

wID	Project
1	E1
2	E2
3	E3
4	P1
5	P2

sID	cID	gID	wID
1	1	1	1
1	1	2	2
1	1	2	3
1	1	1	4
1	1	3	5
2	1	1	1
2	1	1	2
2	1	2	3
2	1	1	4
2	1	1	5

Query Languages

- The standard
  - SQL (Structured Query Language) originally called SEQUEL (Structured English QUEry Language)
  - Developed by IBM in 1974
  - Introduced commercially in 1979 by Oracle Corp.
  - RDMS (SQL), ODBMS (Java, C++, OQL etc)

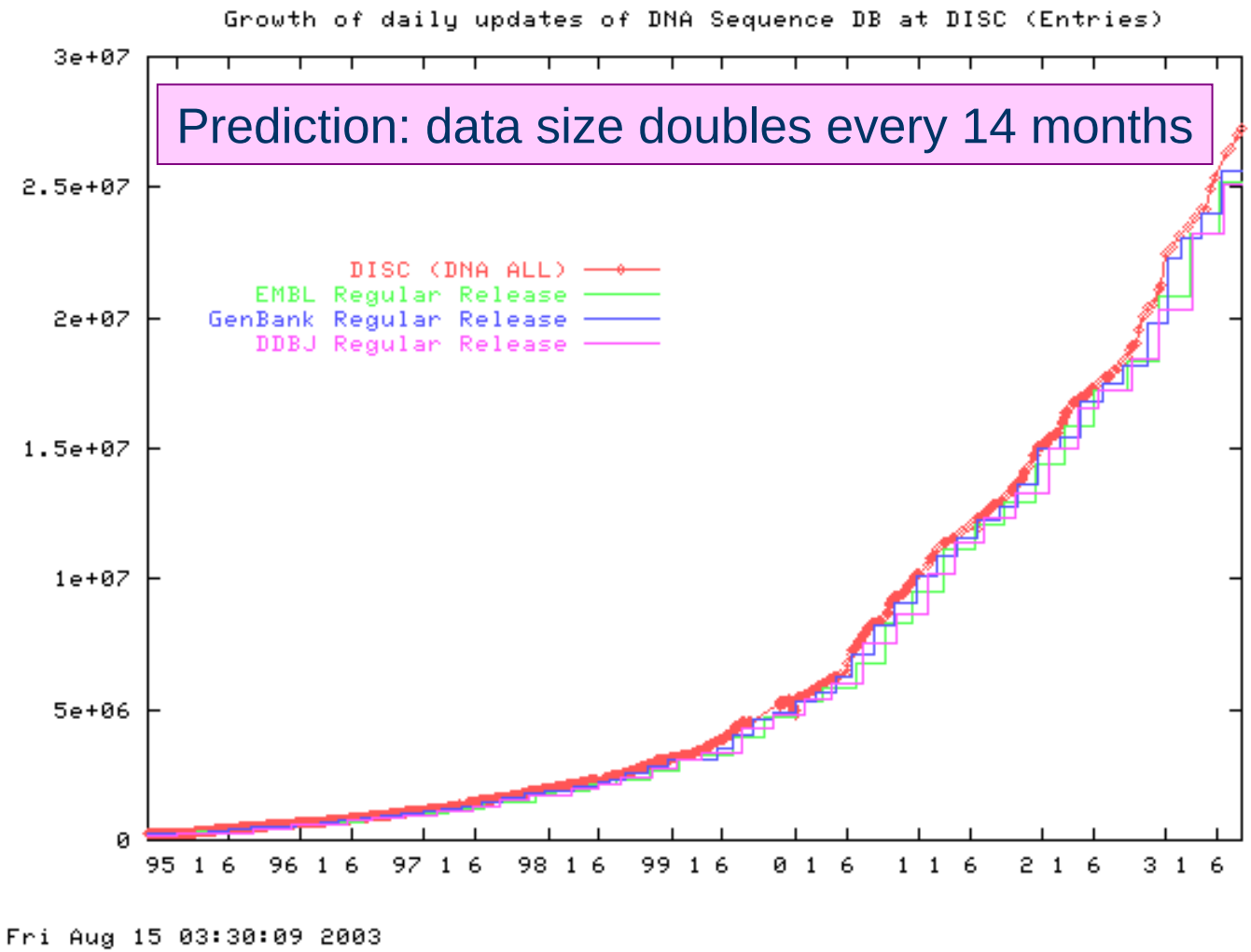
Relational Databases

- What have we achieved?
  - No repeating information
  - Less storage space
  - Better reality representation
  - Easy modification/management
  - Easy usage of any combination of records

Why need to use database system?

- There needs to be a way to
  - **manage** the massive amount of data effectively,
  - **search** data of interest fast and in independent manner,
  - **combine** and relate the data in new ways,
  - **analyze** the data (across different species), and
  - highly **consistent**, non-redundant data,
  - **integrate** the data from different labs

The growth of Genbank (updates)



44,575,745,176 bases, from 40,604,319 reported sequences (up to Dec.,15, 2004)

Databases is important to biology research

- Bioinformatics Lab ~ Database Lab
- Kinds of databases created in the lab

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

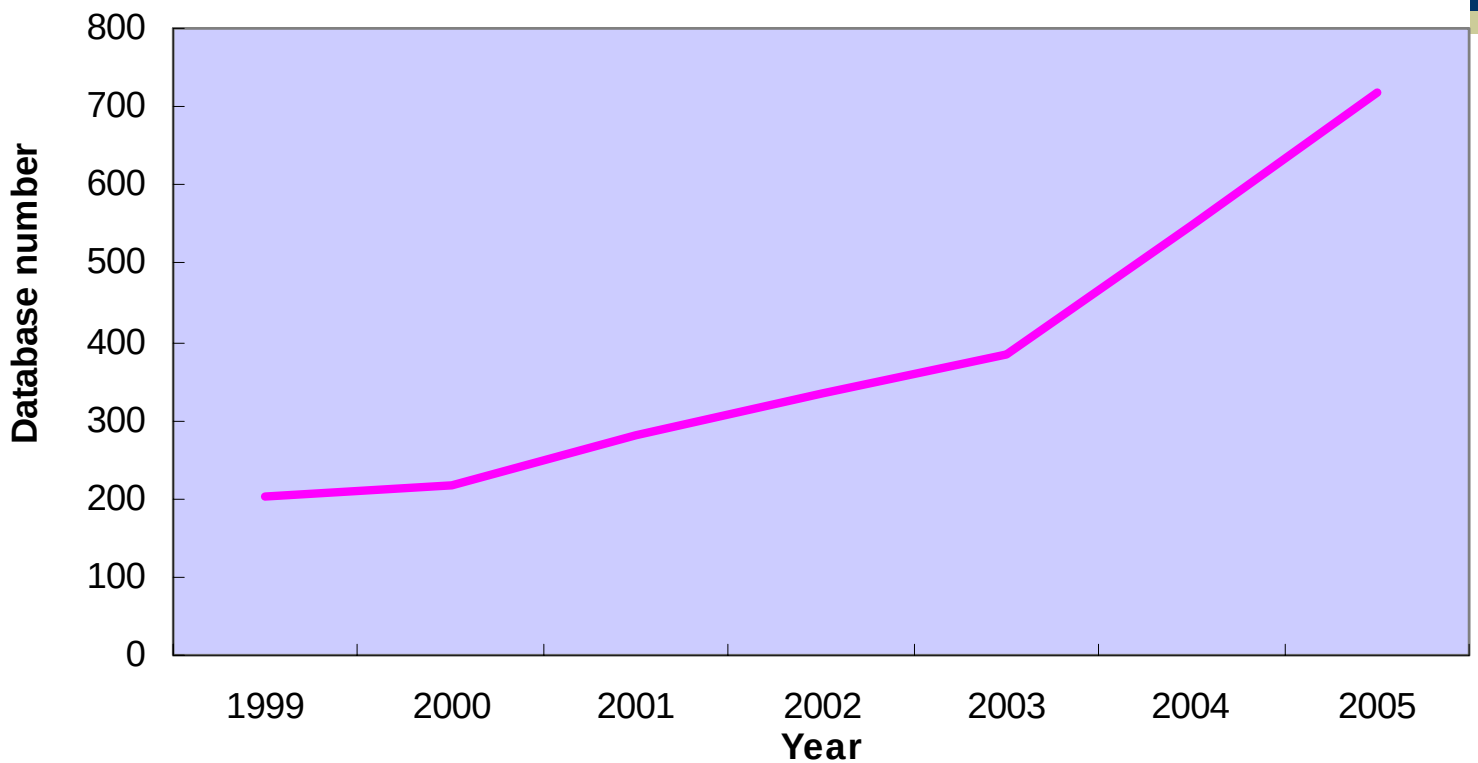
  Data types range from sequences, 3-dimensional structures, pathways, images, text, and a wide variety of annotation.
- Heterogeneous → Universal schema

  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

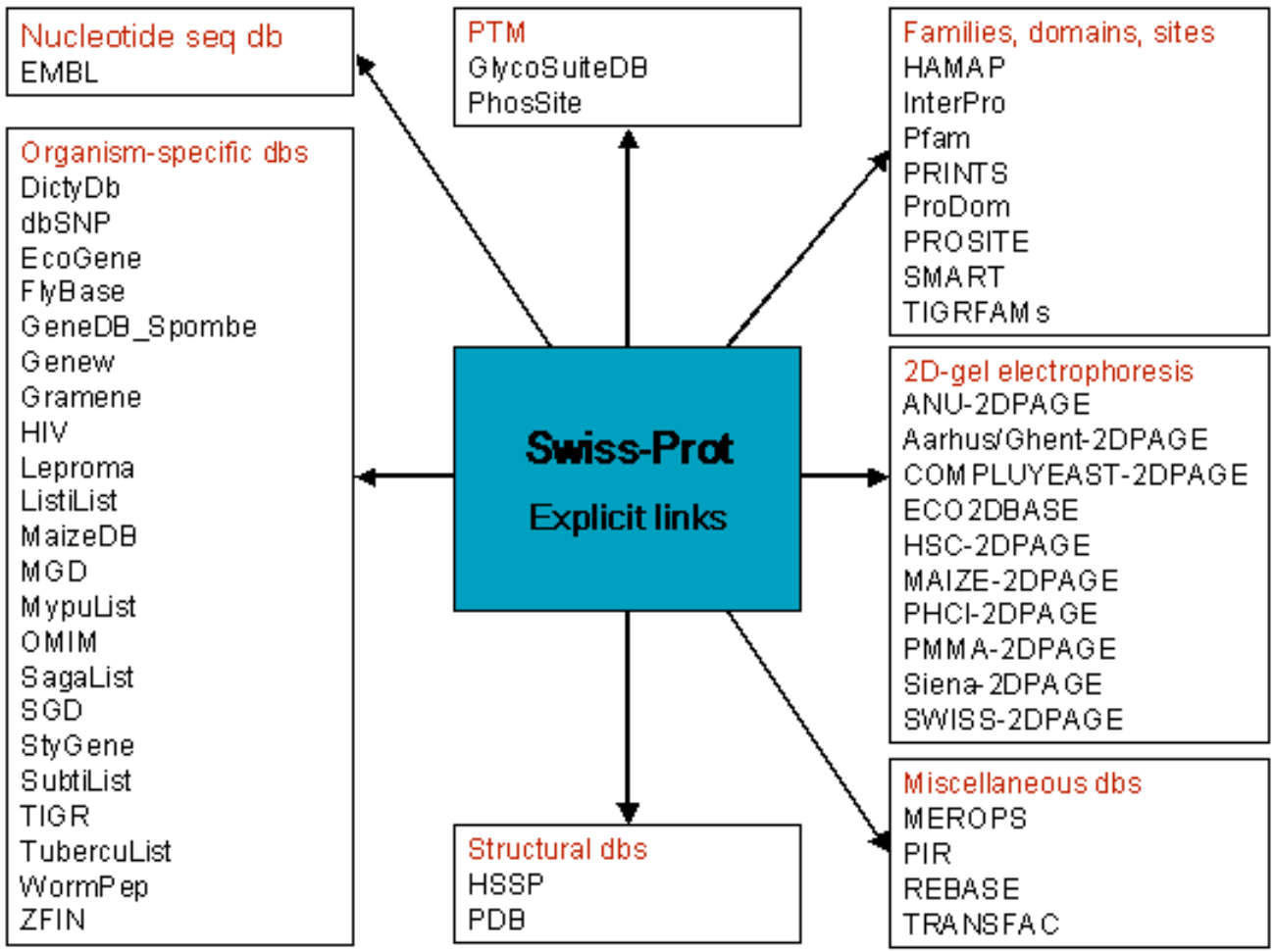


(The Molecular Biology Database Collection from *Nucleic Acids Research*)

Three main tendencies

- 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

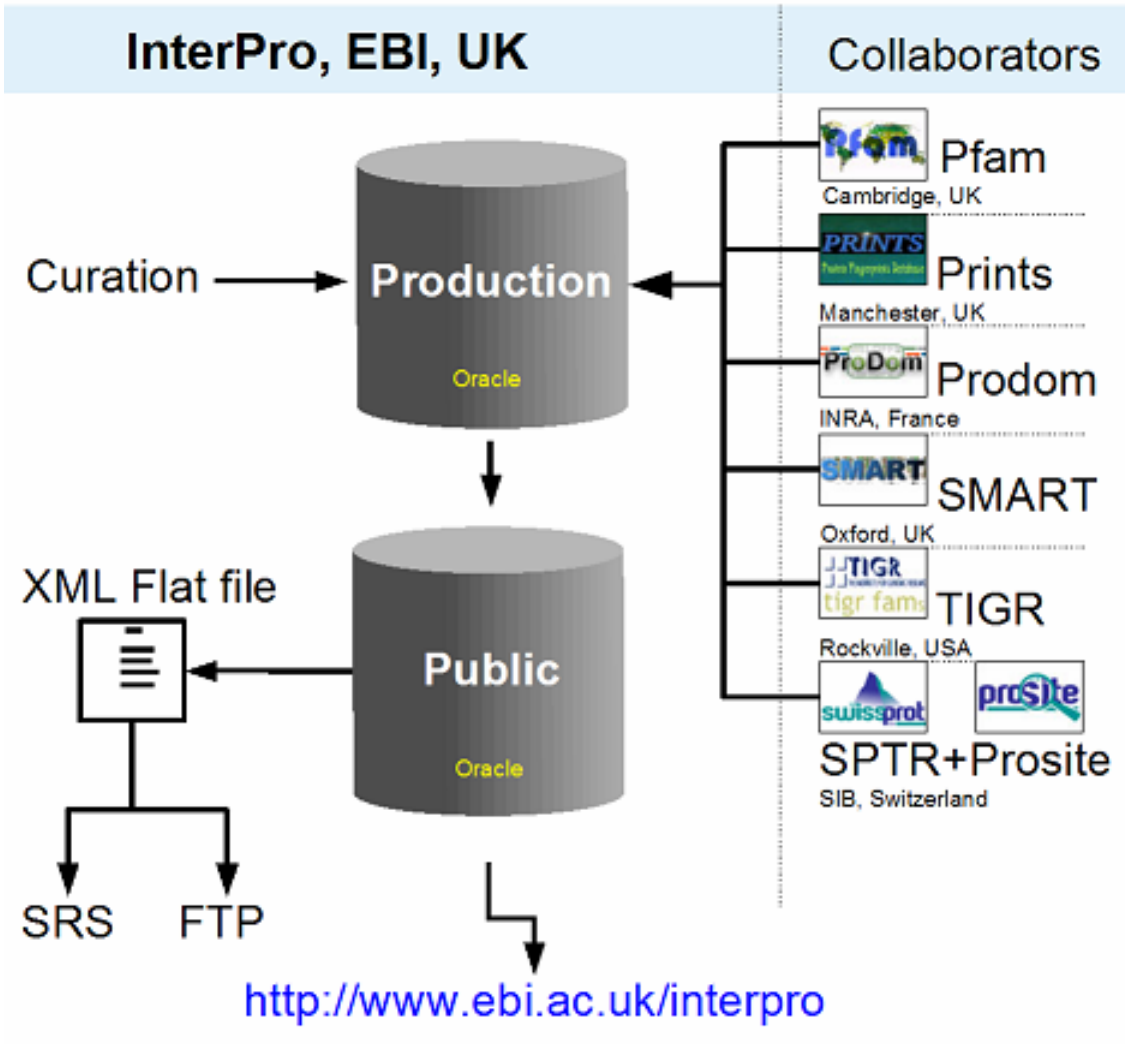


InterPro

- InterPro is a database of **protein families, domains and functional sites.**
- InterPro combines a number of databases that use different methodologies and a varying degree of biological information on well-characterized proteins **to derive protein signatures**.
- member databases
  - Sequence-motif methods, **PROSITE, PRINTS, Pfam, SMART, TIGRFAMs, PIR SuperFamily (PIRSF) and SUPERFAMILY**
    - PROSITE, home of regular expressions and profiles;
    - Pfam, SMART, TIGRFAMs, PIR SuperFamily and SUPERFAMILY keepers of hidden Markov models (HM
    - PRINTS, provider of fingerprints

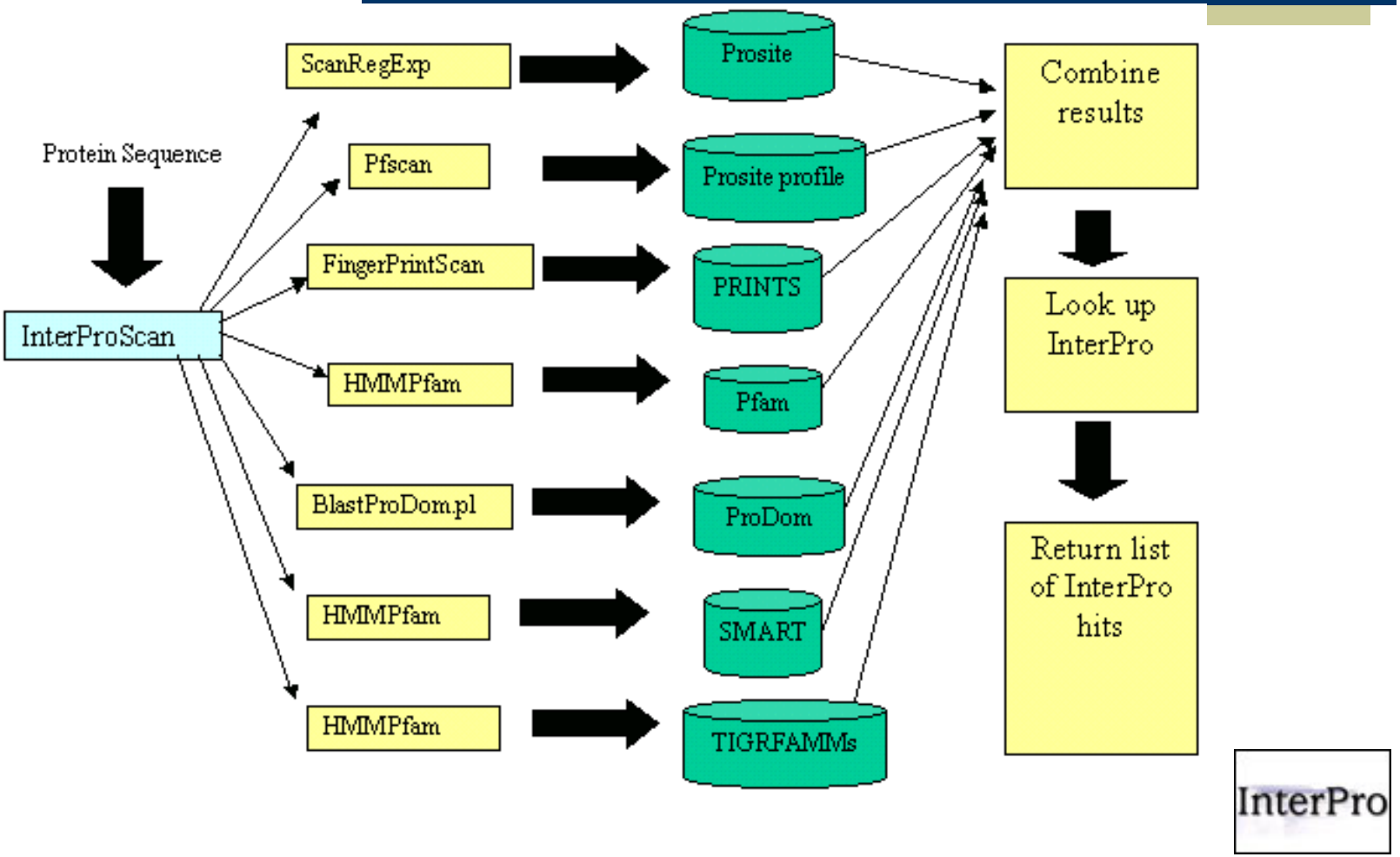
InterPro

InterPro



InterPro

InterPro


Protein Sequence
InterProScan
ScanRegExp
Pfscan
FingerPrintScan
HMMPfam
BlastProDom.pl
HMMPfam
HMMPfam
Prosite
Prosite profile
PRINTS
Pfam
ProDom
SMART
TIGRFAMMs
Combine results
Look up InterPro
Return list of InterPro hits
InterPro

Gene Ontology database

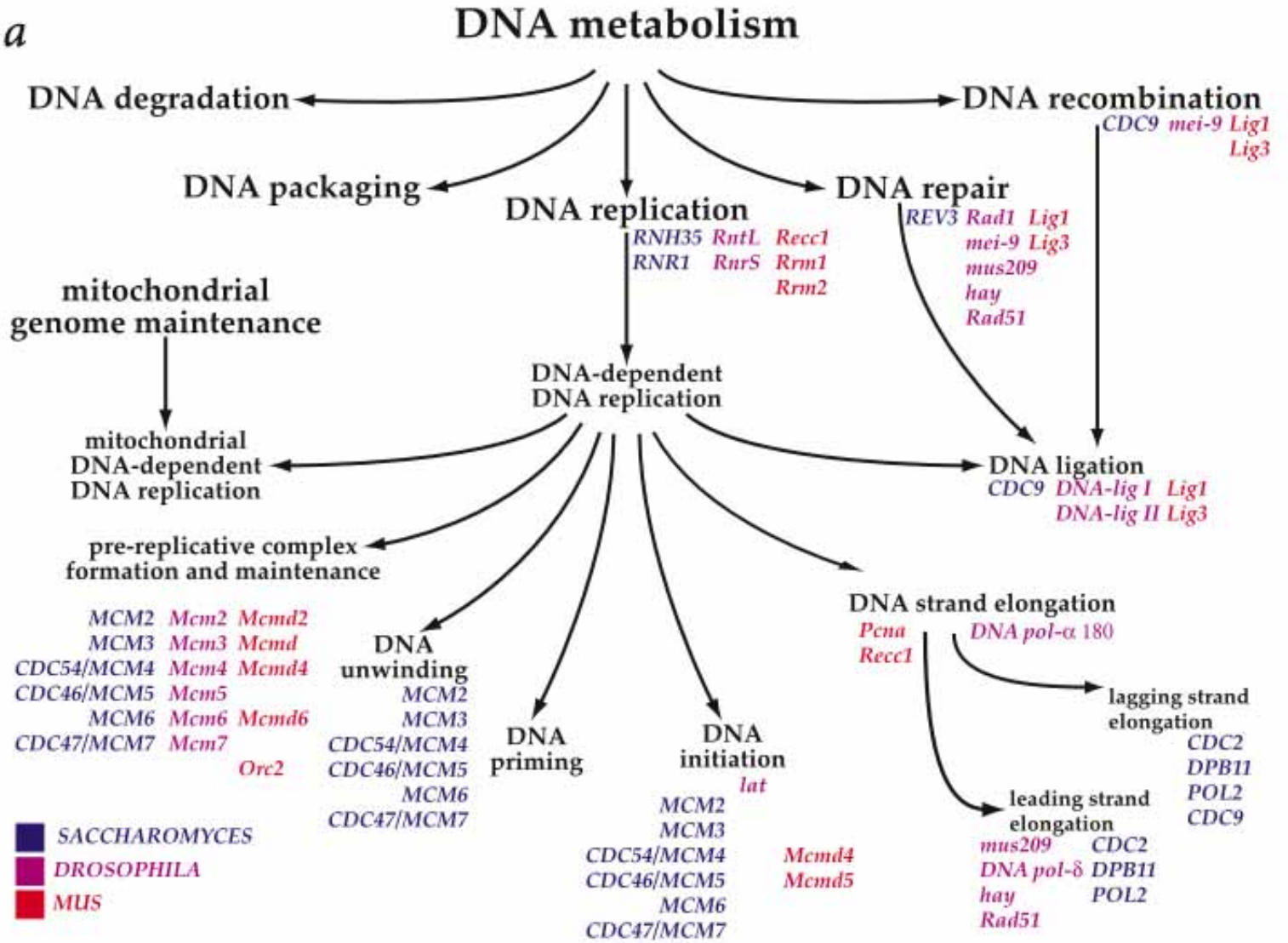
"The Gene Ontology (GO) project seeks to provide a set of structured vocabularies for specific biological domains that can be used to describe gene products in any organism."

A few key points:
GO is a "structured" vocabulary, which is really a specialized type of a "controlled" vocabulary.

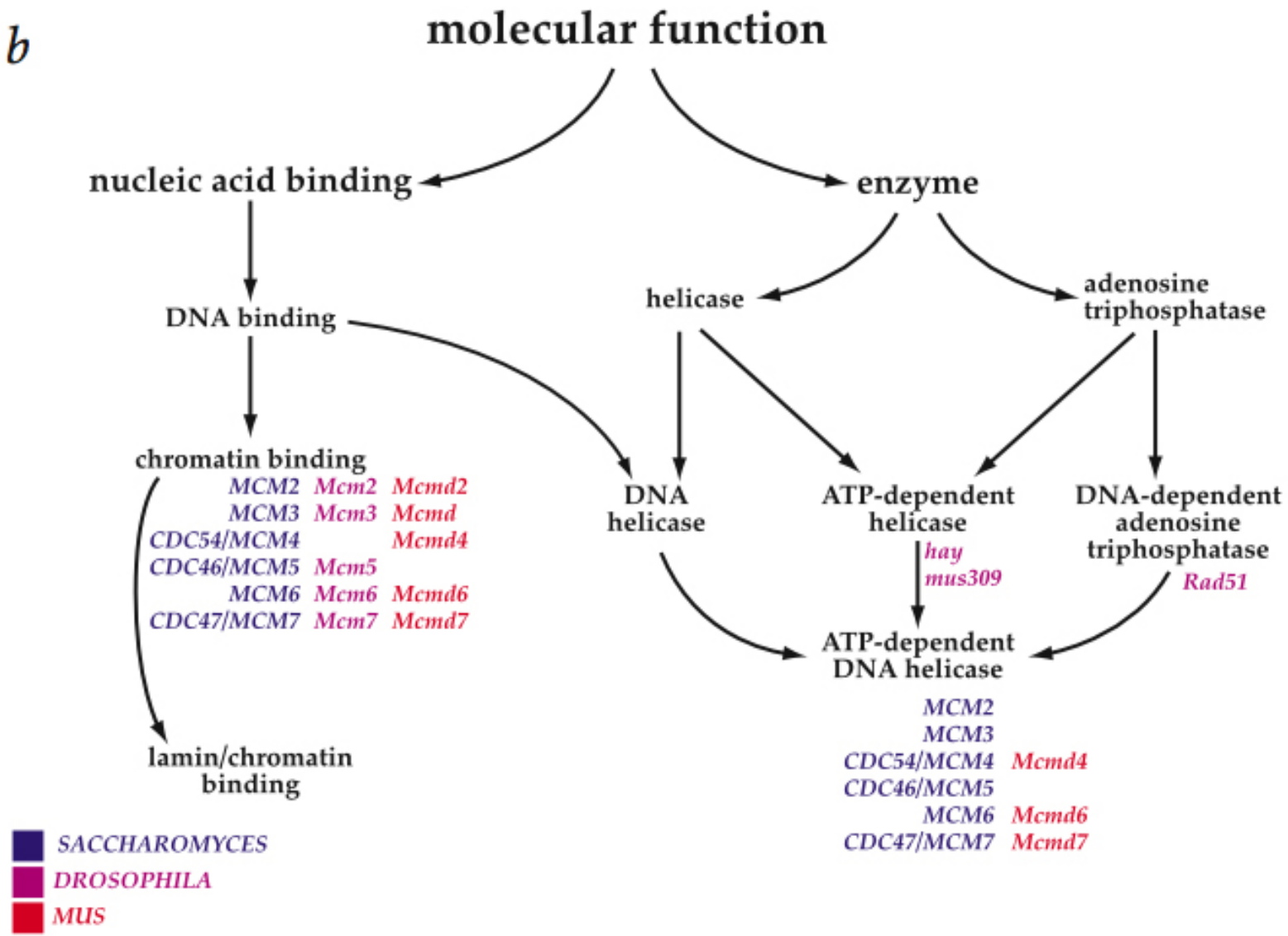
The ontologies in GO are intended to describe three biological areas, "molecular function", "biological processes" and "cellular components".

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.

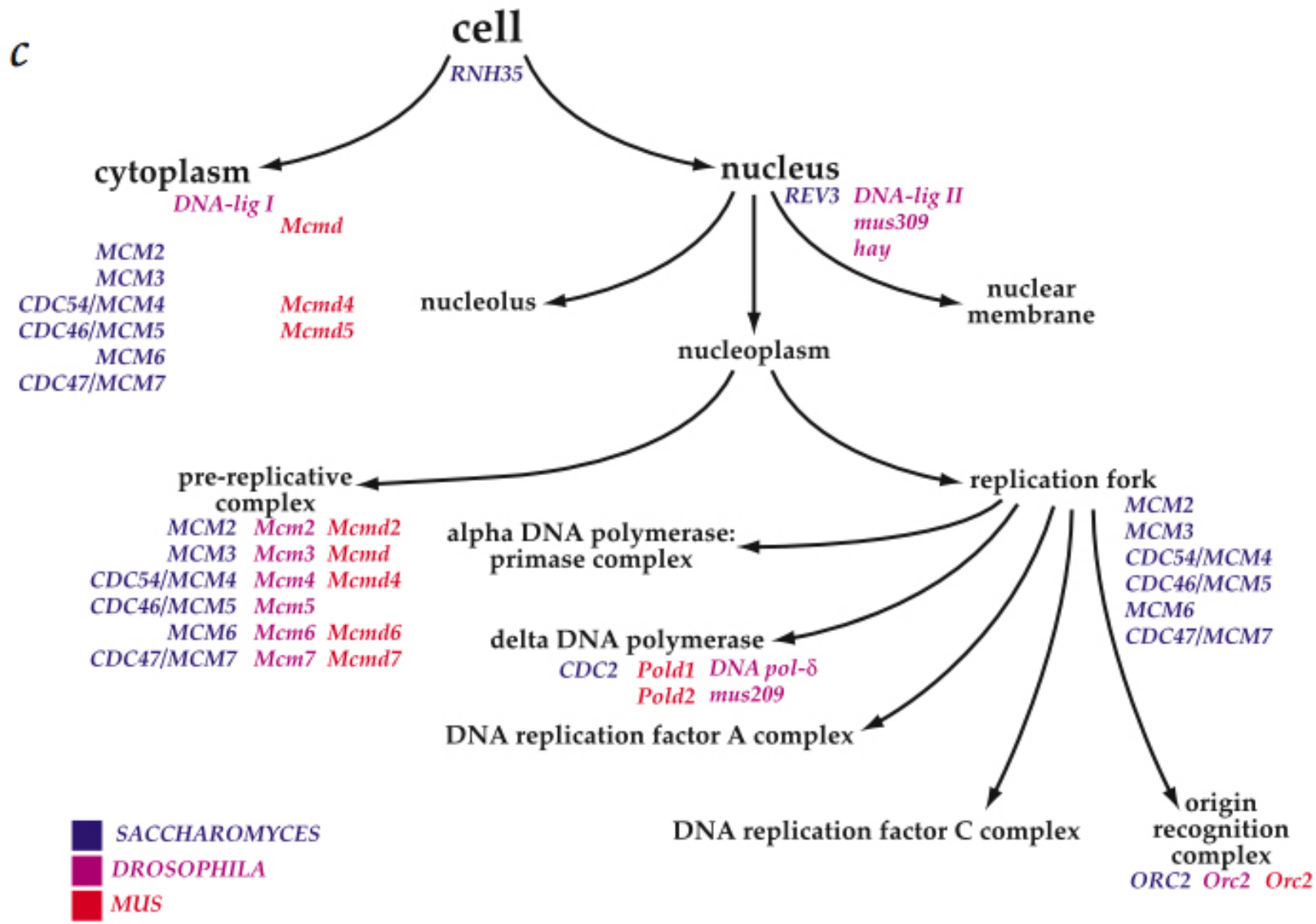
Biological Process Ontology



Molecular Function Ontology



Cellular Component Ontology



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.

The application to biological databases: knowledge base

- Current life science information management systems
  - *information-rich*
  - *knowledge-poor*

Data → Information → Knowledge → Intelligence

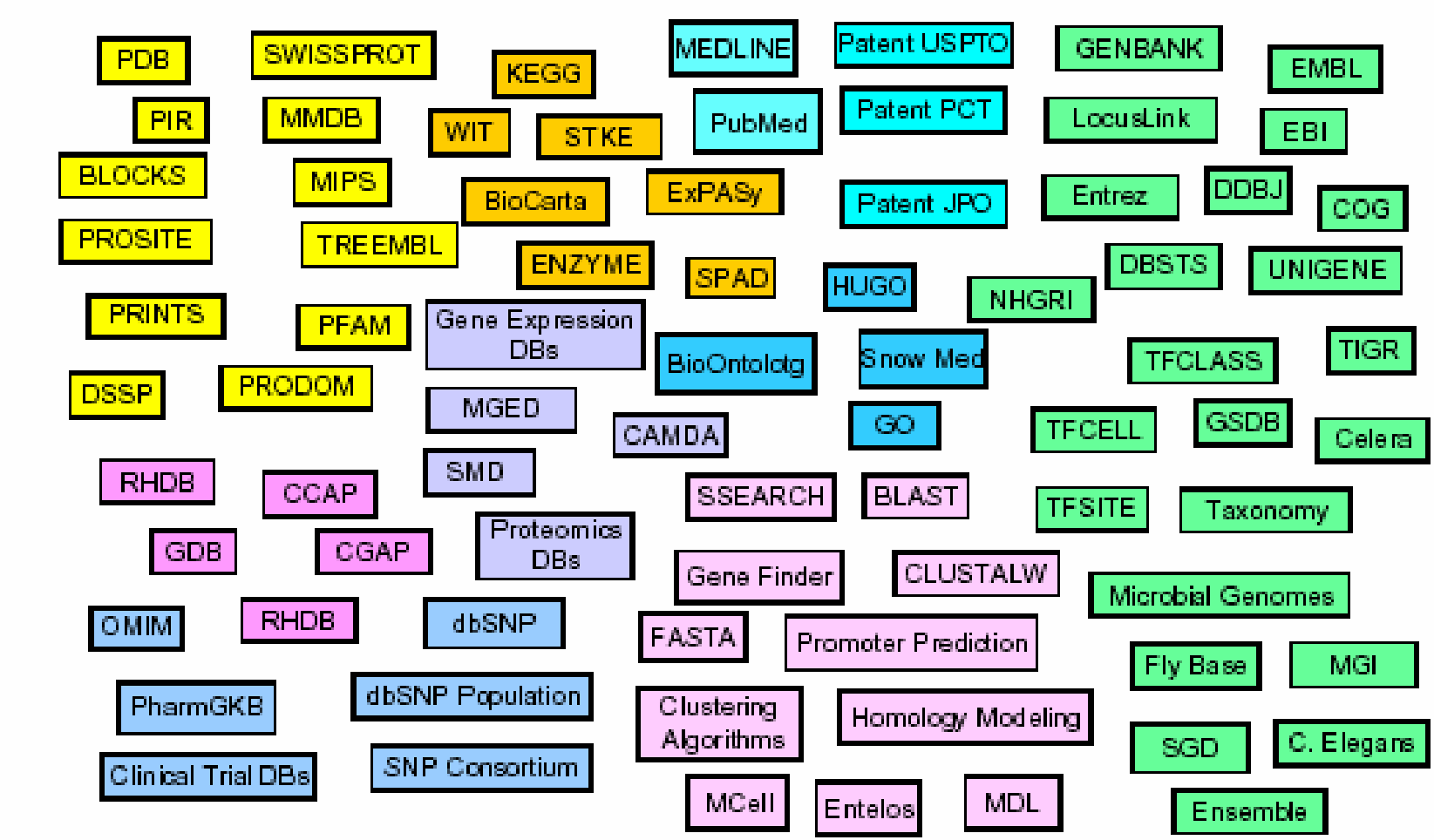
Knowledge discovery in databases

- **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

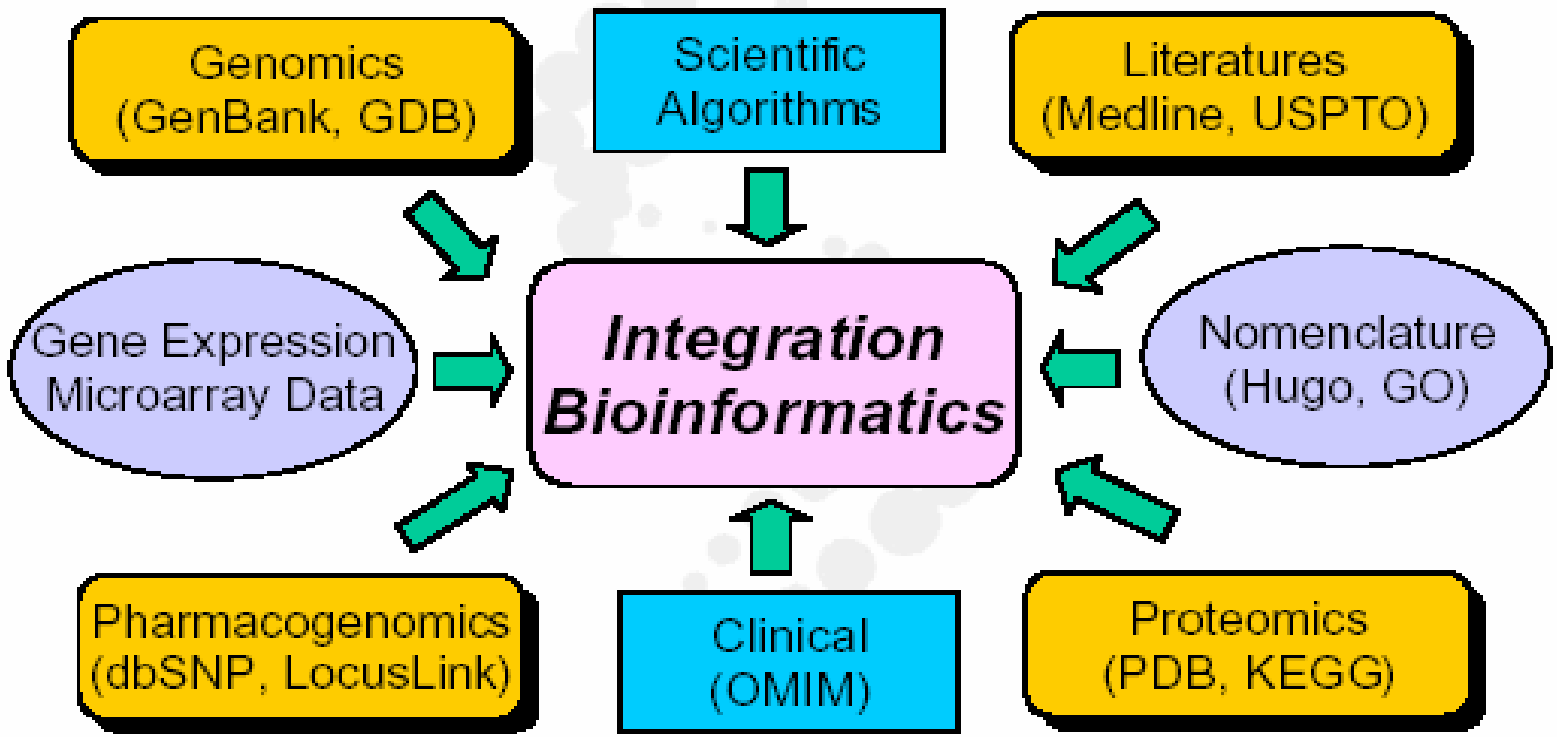
Swimming in Data Sources



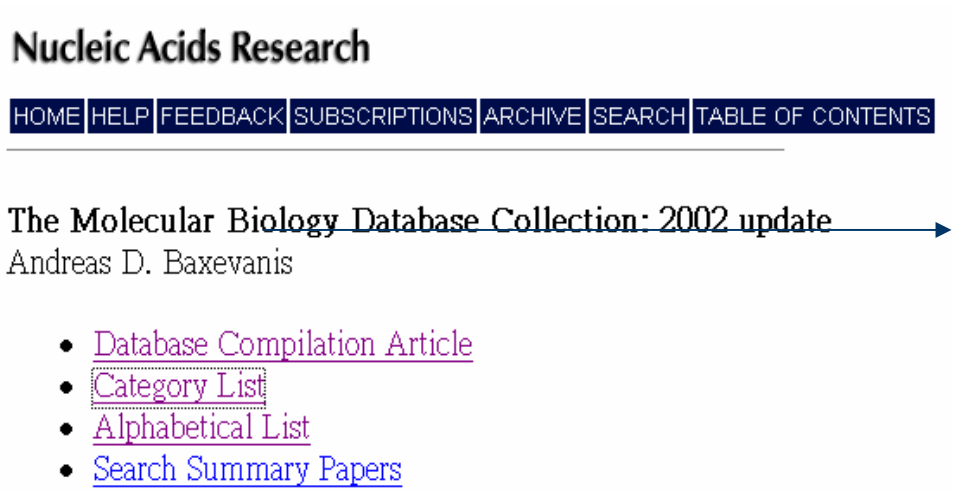
Extract knowledge from various information sources

- A typical example
  - geneA is reported to associated with diseaseD
  - geneB interacts with geneC *in vivo*
  - BacteriaM over expresses substractX in conditionP
  - substractY down regulate geneB
  - 3D chemical prediction sugguests substractX similar to substractY
  - geneC is the alias of geneA
- Simplified Model
  - D – A(C) – B – Y – X

Database Integration



Nucleic Acids Research Database Issue


Nucleic Acids Research

HOME HELP FEEDBACK SUBSCRIPTIONS ARCHIVE SEARCH TABLE OF CONTENTS

The Molecular Biology Database Collection: 2002 update

Andreas D. Baxevanis

- Database Compilation Article
- Category List
- Alphabetical List
- Search Summary Papers

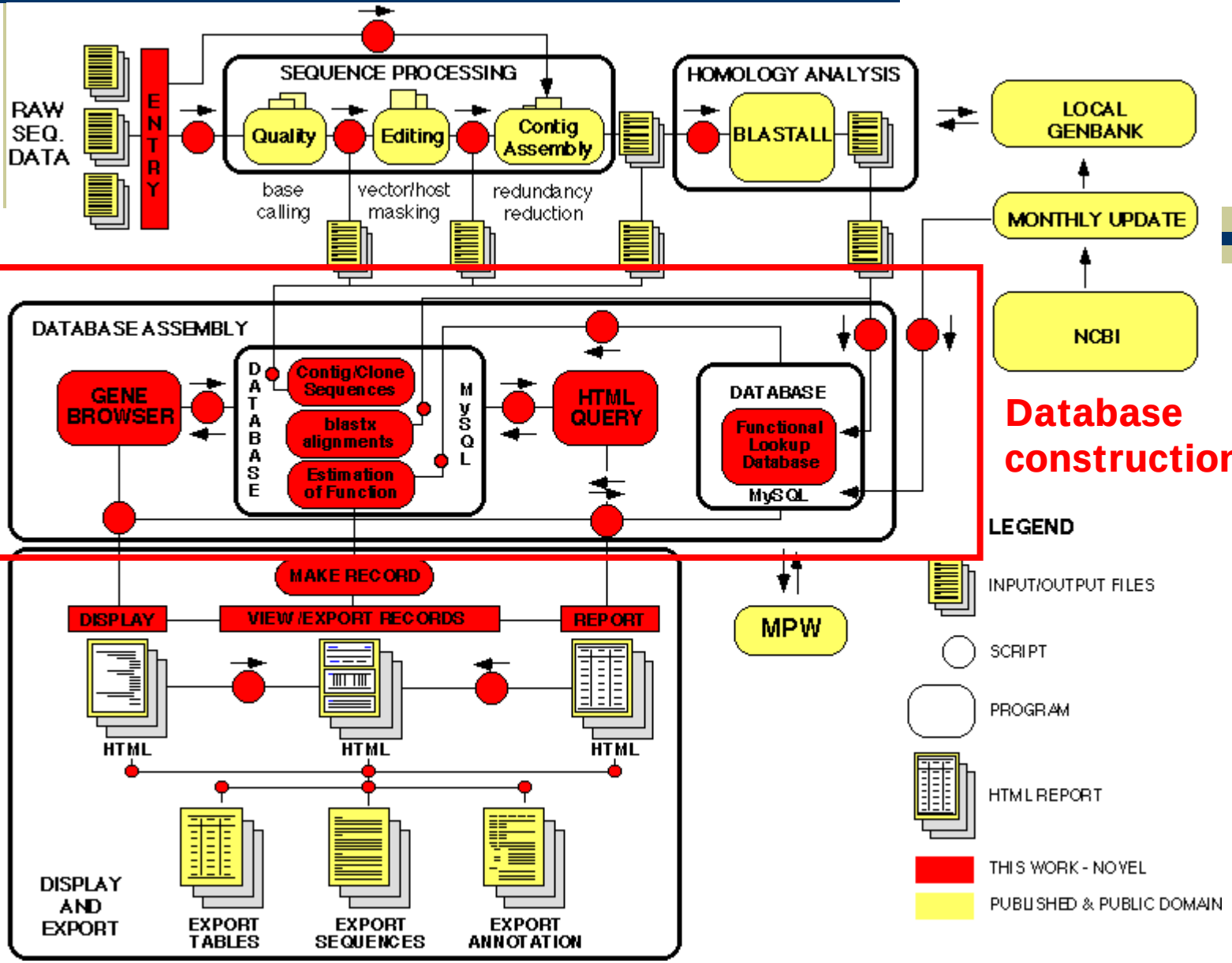
Database Categories List

- Major Sequence Repositories
- Comparative Genomics
- Gene Expression
- Gene Identification and Structure
- Genetic and Physical Maps
- Genomic Databases
- Intermolecular Interactions
- Metabolic Pathways and Cellular Regulation
- Mutation Databases
- Pathology
- Protein Databases
- Protein Sequence Motifs
- Proteome Resources
- RNA Sequences
- Retrieval Systems and Database Structure
- Structure
- Transgenics
- Varied Biomedical Content

Bioinformatics in Monascus Genome

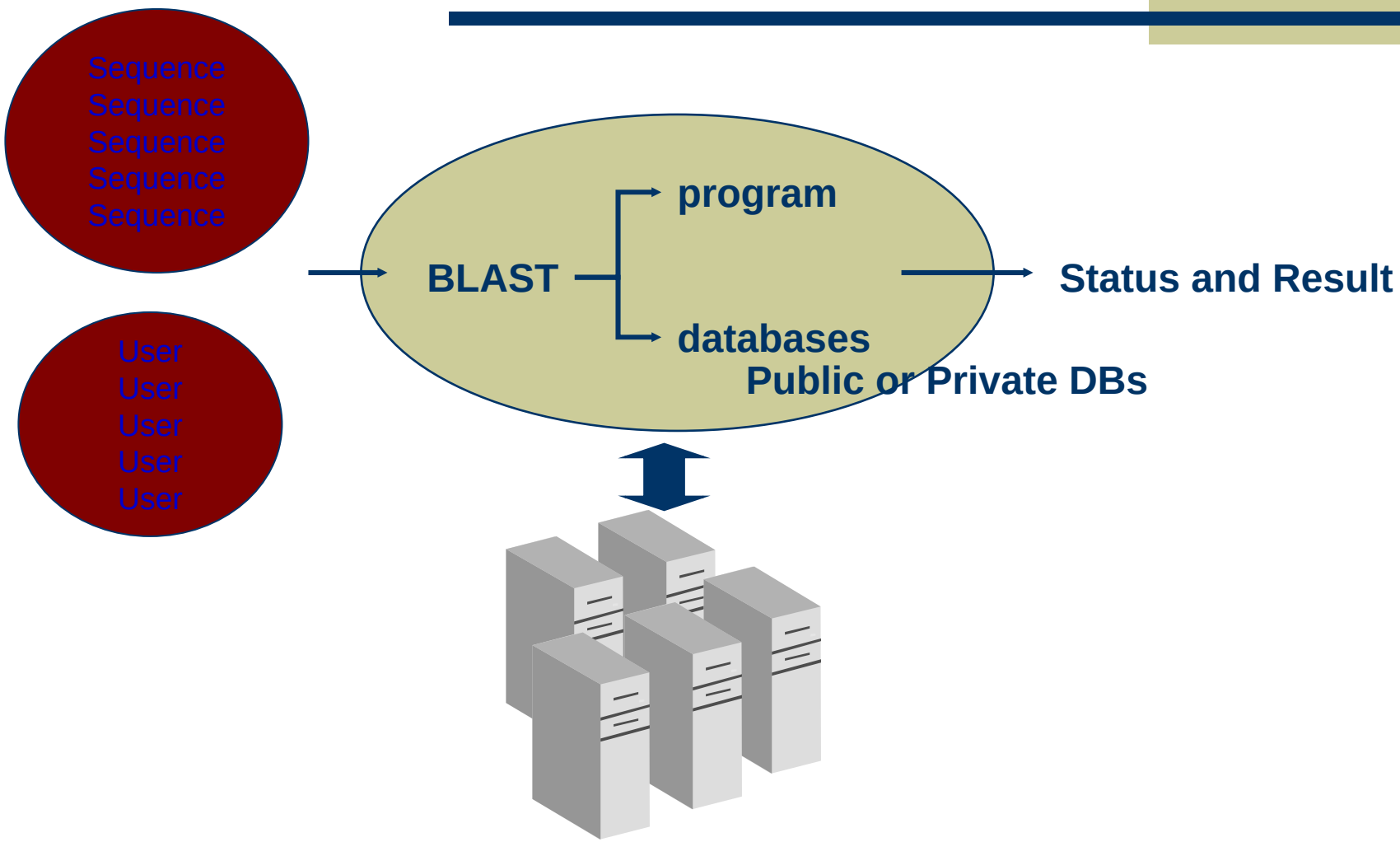
Bioinformation works in the Genome Project

- Infrastructure
  - Computing environment
  - Database System
  - Network Security
- Analysis Platform
  - High throughtput Sequence QC
  - EST clustering
  - Whole Genome Shutgun Assemble
  - Sequences Annotation
    - Similarity Search (Primary uses BLAST)
    - Gene Finding
    - EST to Genome Alignment
  - Web-based GUI Interface



Database construction

Parallel and Partition Blast



Why *Monascus*?

- Monascus sp.
  - 俗稱紅麴 (米麴)
  - 絲狀真菌
  - 東方常用食用真菌
  - 中國傳統醫藥本草綱目中入藥
    - 清血降血壓
    - 降膽固醇
    - 婦女生產後的調理
  - 具產業利用潛力
- Monascus Genome
  - Estimate 30Mb in size (human ~ 3000Mb bacteria ~ 5Mb)
  - Eukaryotic
  - 單細胞多核
  - 豐富的二次代謝產物

Eubacteria vs Fungi Genome

- Typical Eubacteria Genome
  - Less than 10Mb in size
  - Single circular chromatin
  - Single exon gene (no introns)
  - ORF ~ genes (pseudogenes)
  - More compact, few short repeat sequence
- Typical Fungi Genome
  - 20 ~ 100 Mb in size
  - More than one chromosomes
  - Multi exon gene
  - ORF != genes
  - Compact, few repeat sequences than other eukaryotic organisms

Why Genome Assemble?

- Till today, chromosome size is still much larger than single sequencing reaction.

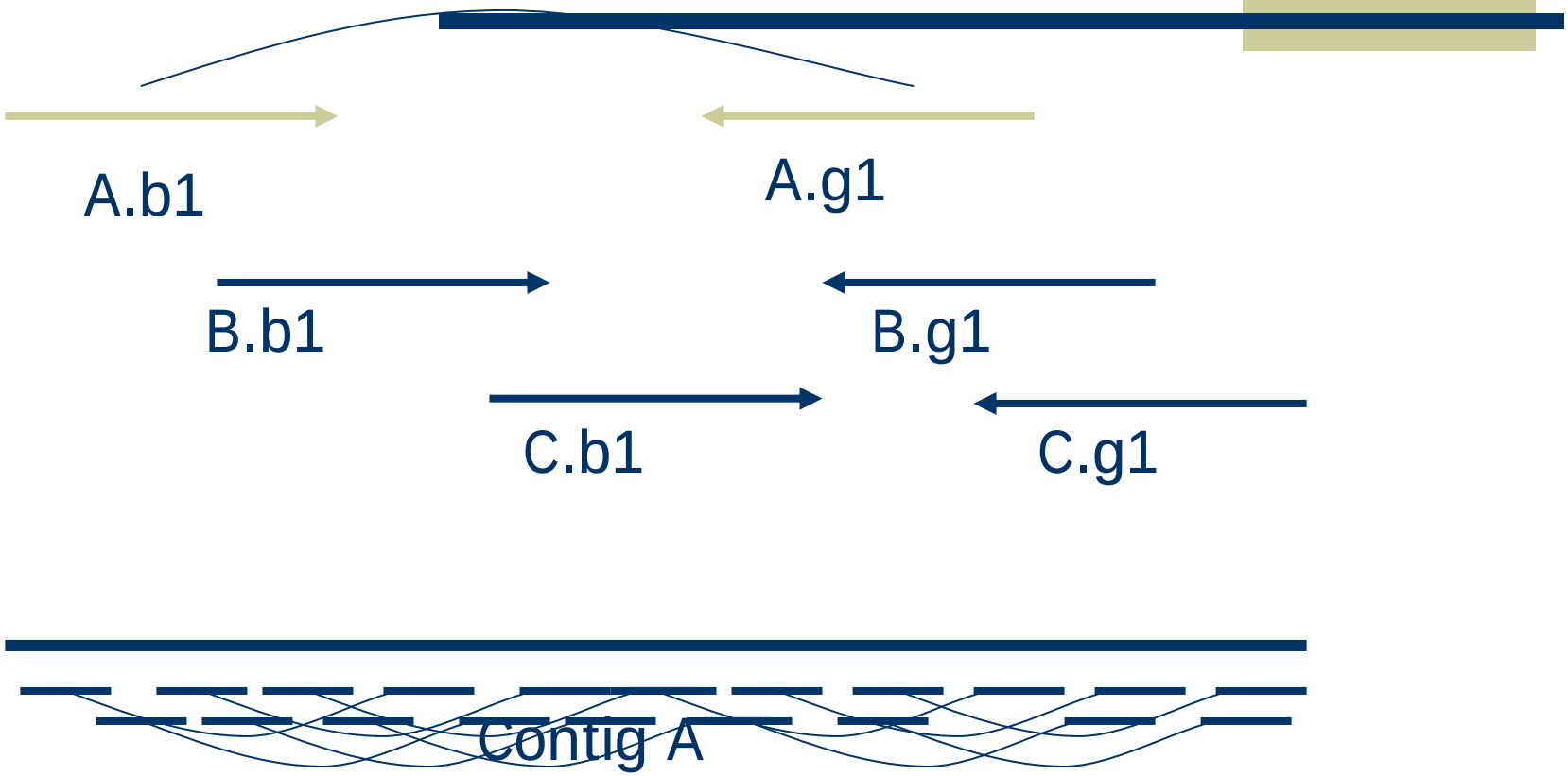
Two Assemble Approach

- BAC approach
  - 發展較早
  - BAC (bacteria artificial chromosome)
  - Large segments of DNA, 100,000 to 200,000 bases, from another species cloned into bacteria
  - Target species genome is mapping to ordered BAC clones
  - Sequencing shotgun library of each BAC clones
  - 優點- 利於不同實驗室間的分工合作
  - 缺點- physical map of the target genome needed

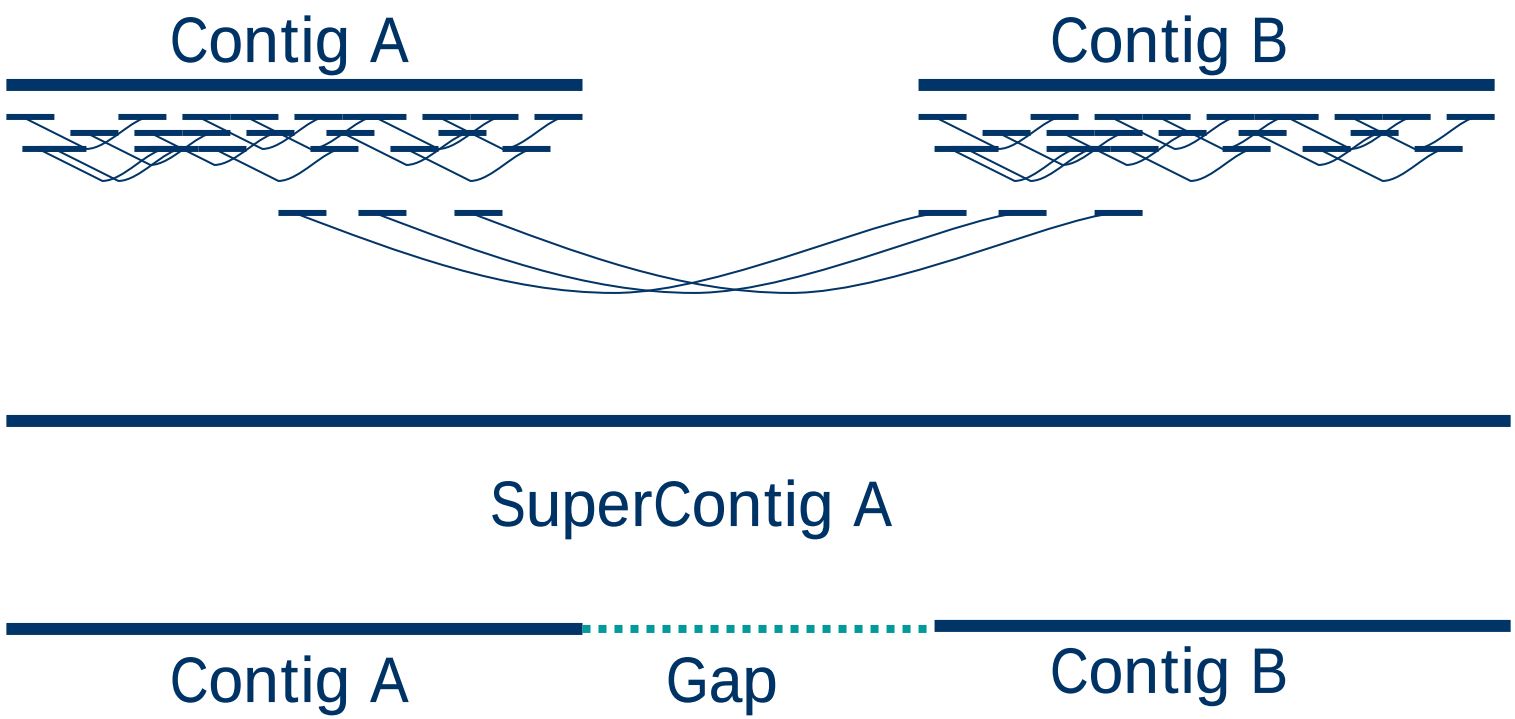
Two Assemble Approach

- Whole genome Shotgun(WGS) approach
  - 因爲需要較大的電腦硬體 後期才開始使用
  - Using shotgun library construction, target genome is cloned in plasmid, fosmid, cosmid of various sizes.
  - Sequencing the shotgun libraries directly.
  - no physical map needed
  - Faster in project processing but more complex

Assemble--contig



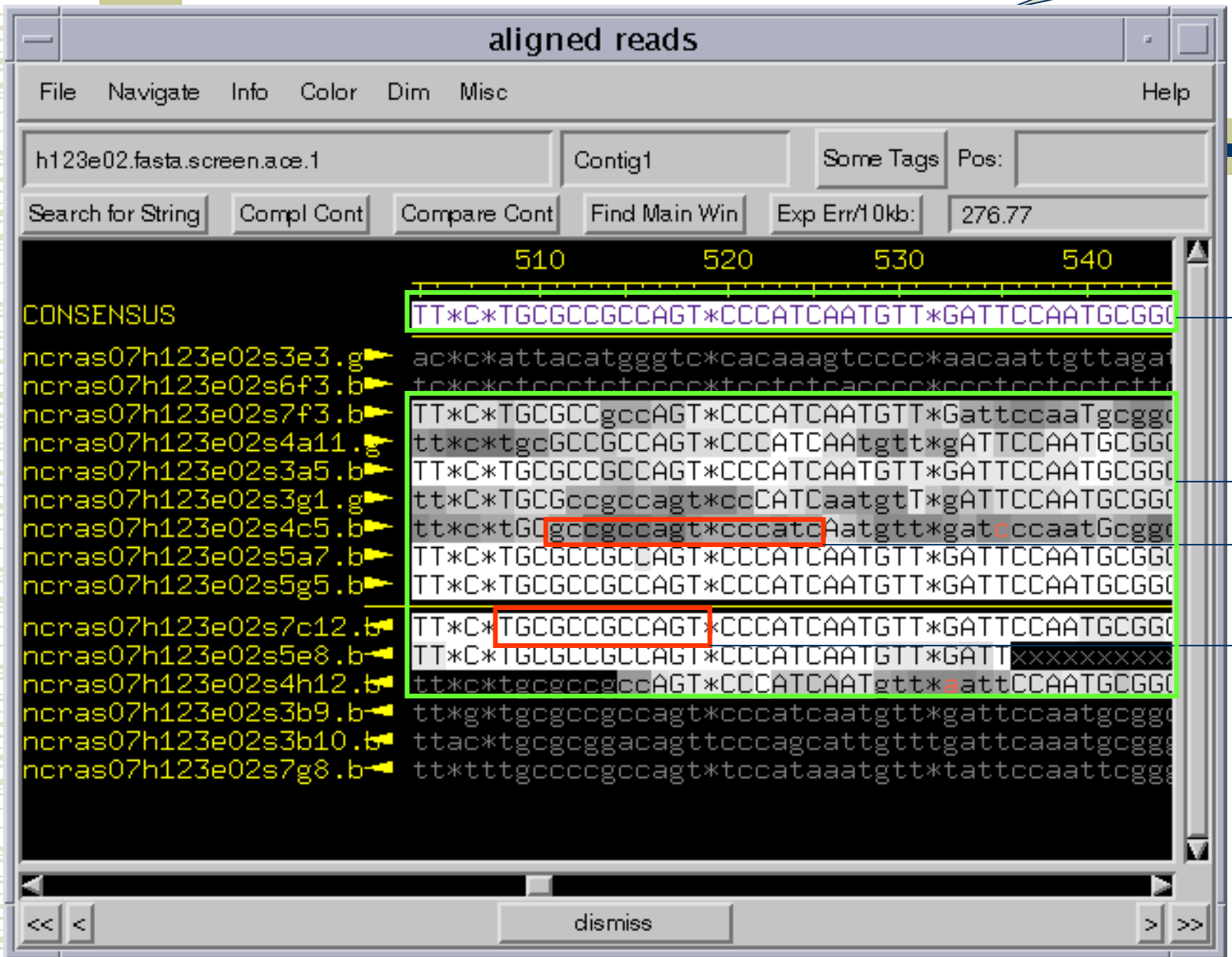
Assemble--supercontig



Draft: gap allowed

Finish: no gap and 0.01 % error rate

Browser/editing

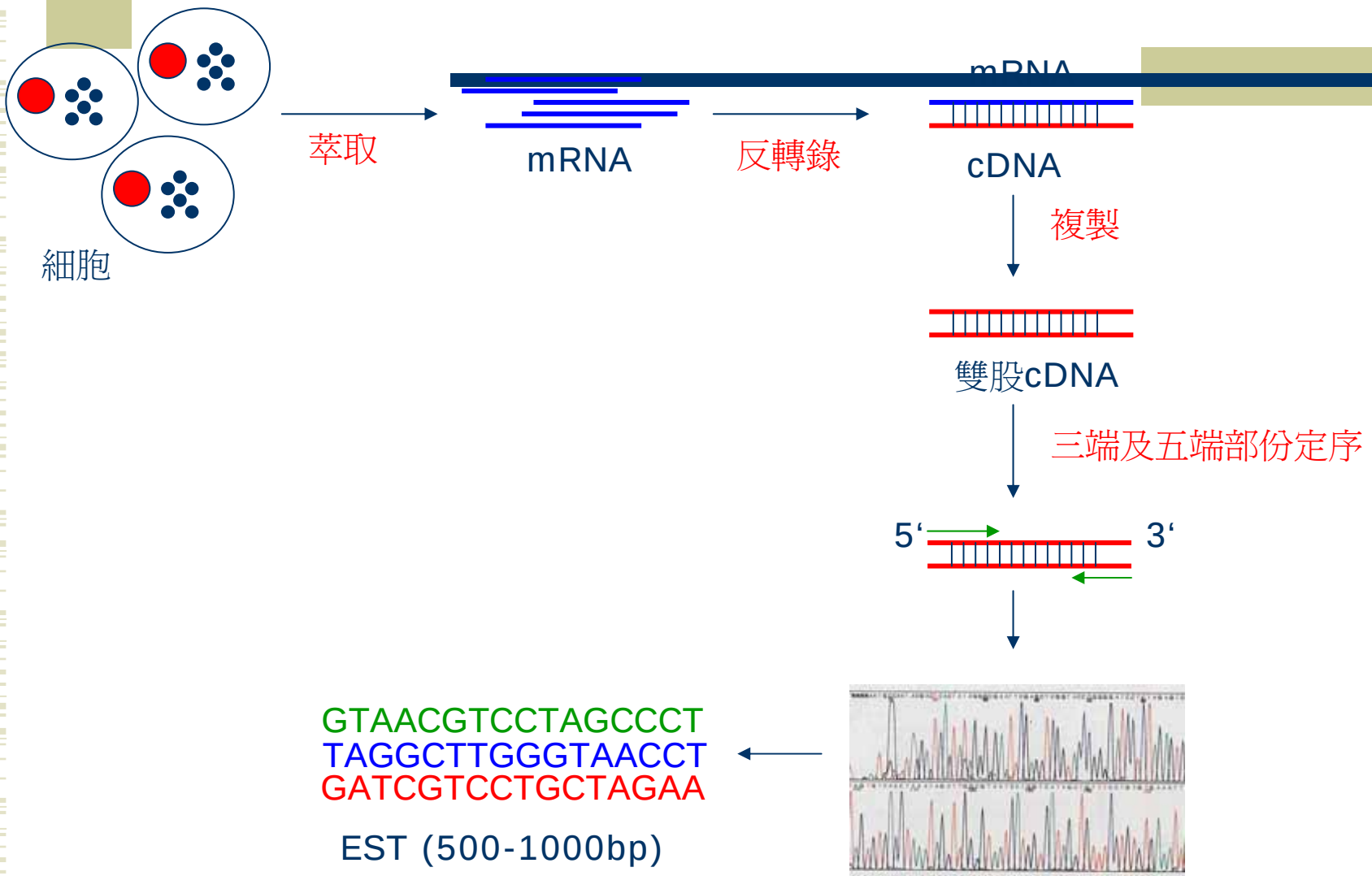


assembly

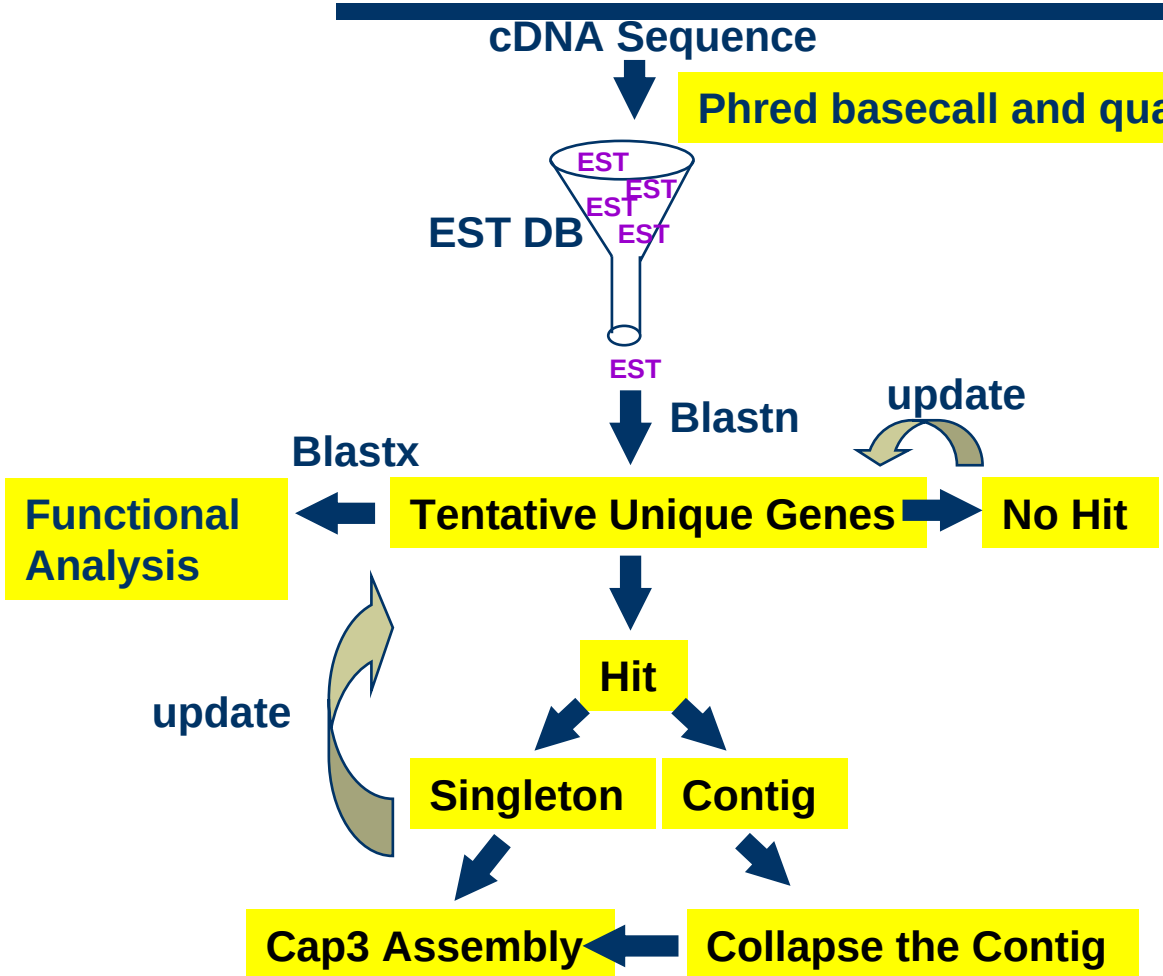
low quality

high quality

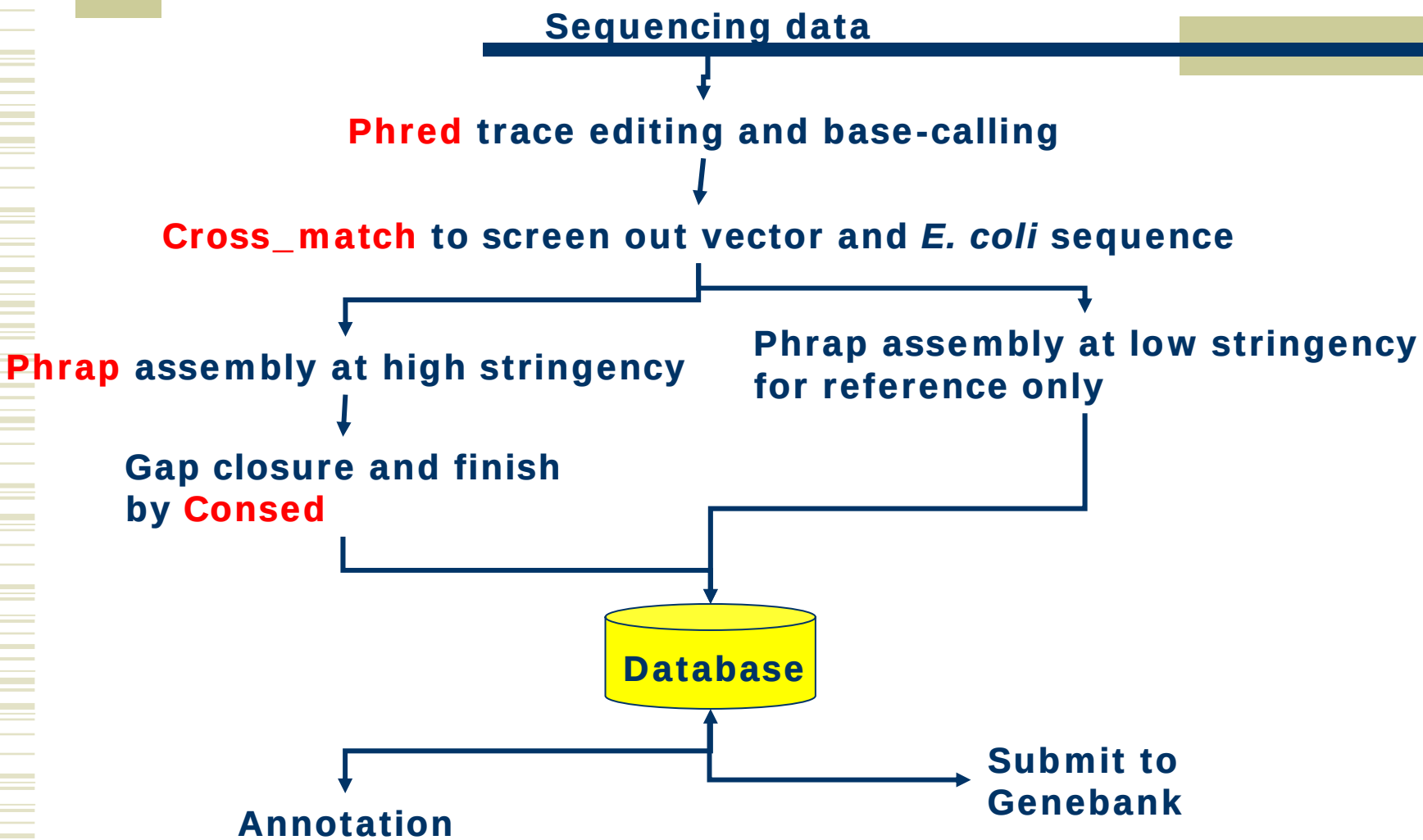
EST (expressed sequence tag)



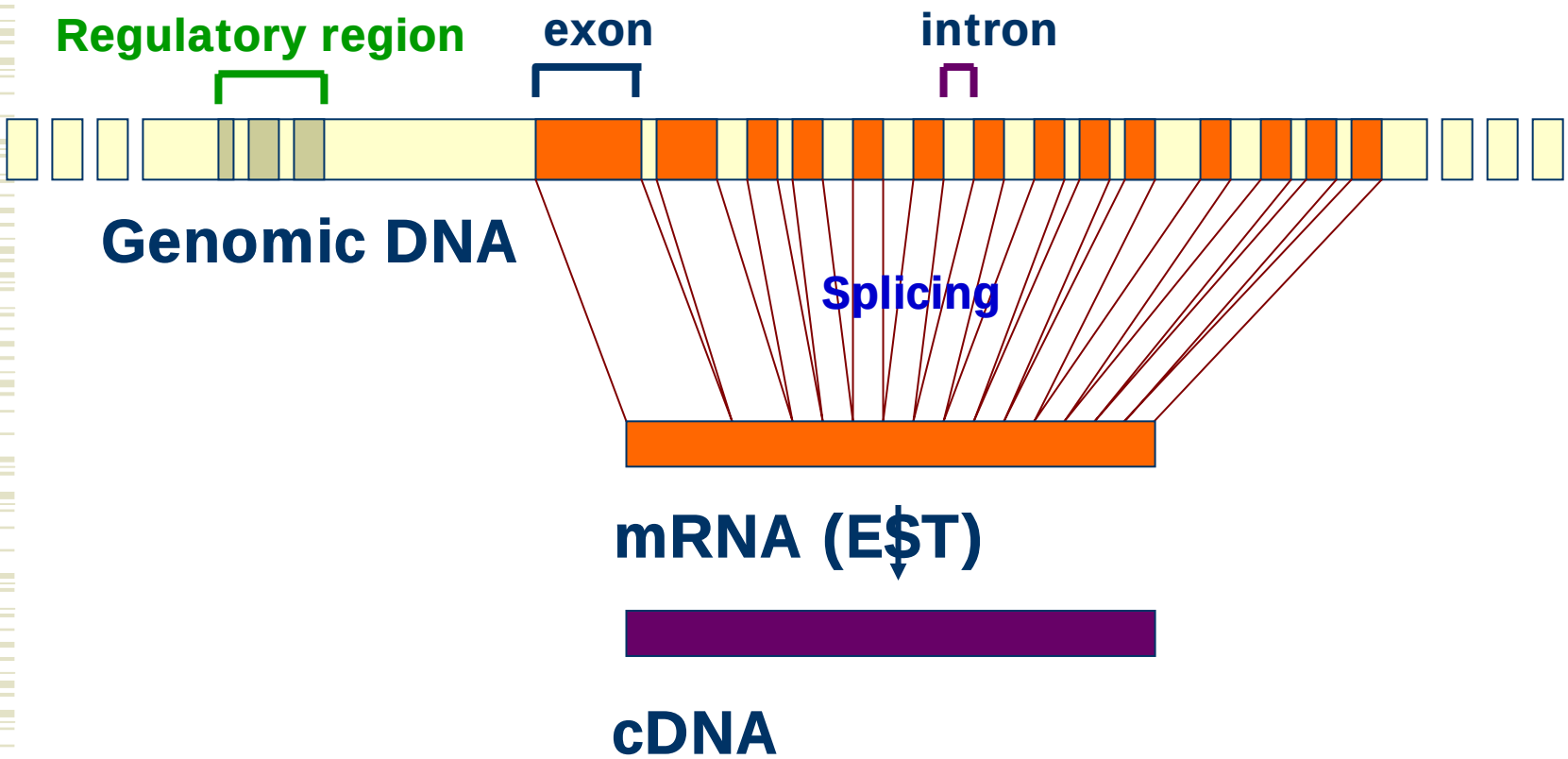
EST clustering and assembly



Automatic Sequencing Process



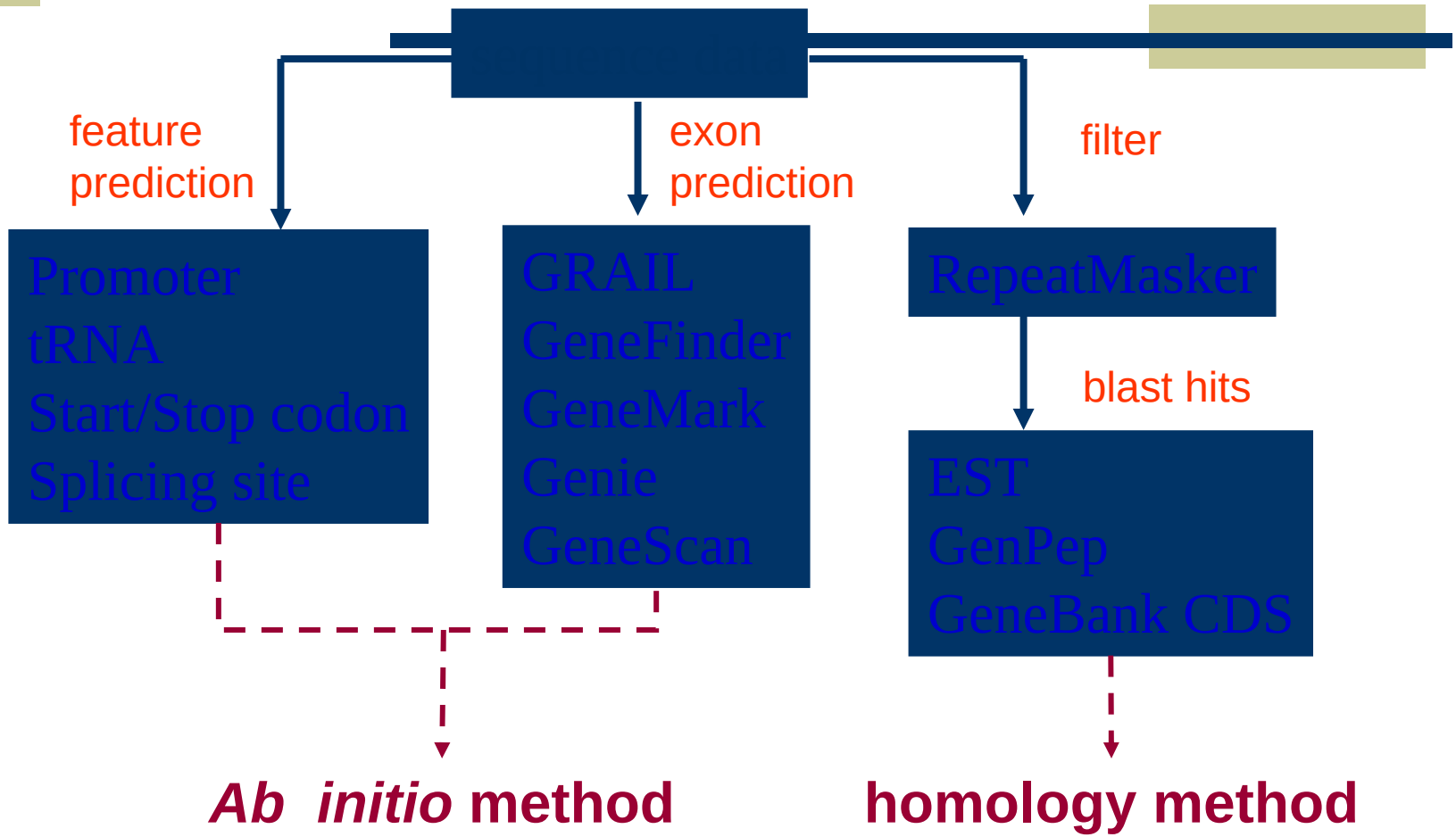
EST to Genome Alignment



Gene Finding (Annotation)

- Typically, Genes are defined as protein coding region
- Typically, Annotation refer to define each genes' function found on genome
- First Step – Locate the gene on genome
- Second Step – Define their function

Locate the Genes



Ab initio method vs homology method

- *Ab initio* method
  - Based on statistics
  - Less accurate than homology method
  - No previous known similar genes needed
- Homolgy method
  - Based on sequence similarity(orthologs)
  - More accurate to mapping ortholog genes between species
  - Need previous known genes with similar sequences

Ab initio method vs homology method

- *Ab initio* programs
  - Genscan
  - GlimmerM
  - GeneID
  - …
- Homology method
  - BLAST
    - Nucleotides comparison
    - Peptides comparison

Annotate Genes Function

- Primary Based on homology search
  - Run BLAST against multi dbs
  - InterProScan
  - GeneOntolgy
  - KEGG
  - Automatic order by some rules to determine their ‘tentative annotation’

Annotation Table Viewer

<table>
<tr><td>Protein Name and ID</td><td>PIR-NREF: NF00626778
GenPept: AAC41675.1 ; AAC41674.1
<table>
<tr><th>Database</th><th>ID</th><th>Protein Name</th></tr>
<tr><td>PIR-PSD</td><td>T17490</td><td>polyketide synthase</td></tr>
<tr><td>SwissProt</td><td>PKS1_ASPPA</td><td>Aflatoxin biosynthesis polyketide synthase (PKS)</td></tr>
</table></td></tr>
<tr><td>Taxonomy</td><td>Source Organism: Aspergillus parasiticus
Taxon Group: Euk/Fungi
NCBI Taxon: 5067
Lineage: cellular organisms; Eukaryota; Fungi/Metazoa group; Fungi; Ascomycota; Pezizomycotina; Eurotiomycetes; Eurotiales; Trichocomaceae; mitosporic Trichocomaceae; Aspergillus</td></tr>
<tr><td>Gene Name</td><td>pksL1</td></tr>
<tr><td>Keywords</td><td>acyltransferase; multifunctional enzyme; phosphopantetheine; transferase</td></tr>
</table>

Genome Annotation

- Gene function and location
- Type
  - Automatic:
    - **IEA** inferred from electronic annotation
    - Based on the sequence similarity
  - Manual
    - **IC** inferred by curator
    - **IDA** inferred from direct assay
    - **IEP** inferred from expression pattern
    - **IGI** inferred from genetic interaction
    - **IMP** inferred from mutant phenotype
    - **IPI** inferred from physical interaction
    - **ISS** inferred from sequence or structural similarity

Web-based GUI interfaces

- MGD
- Genome Browser

How Bioinformatics works

- In a simple words, bioinformatics uses information technology to solve biology questions.
- Bioinformatics research work is the cooperative teamwork on biologists & computer scientists
- As a biologist, you will need,
  - A little skill in computer programming
  - Master your knowledge in the life science fields

Thank You
for Your Attention