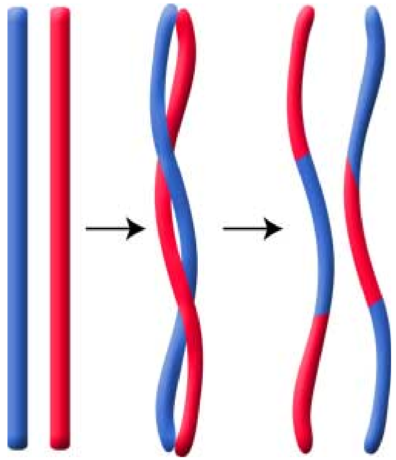
http://petang.cgu.edu.tw/



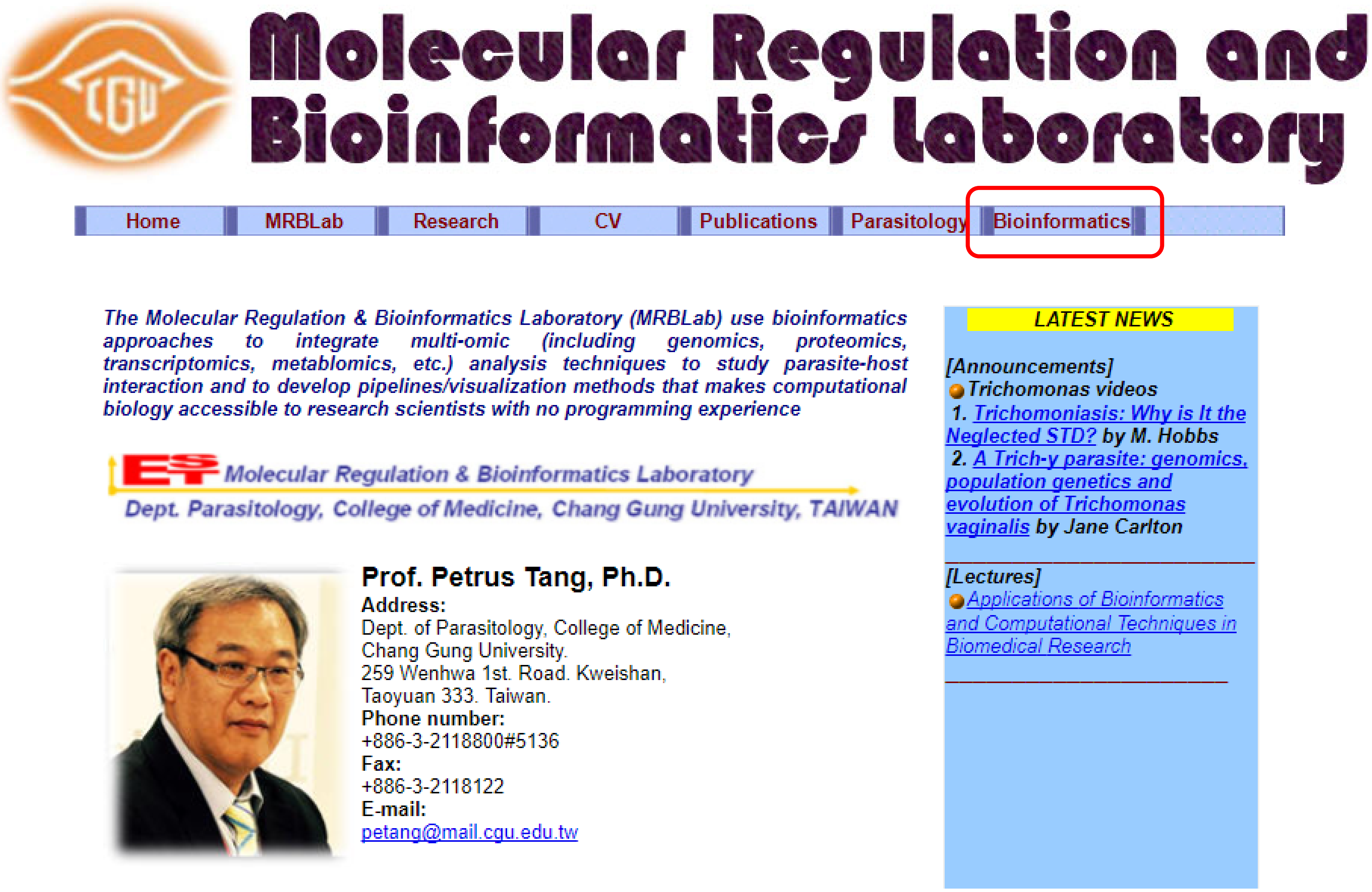
Bioinformatics & Biostatics

Lecture 1 – Introduction to Bioinformatics

AGTCCGCGAATACAGGCTCGGT AGTCCGCGAATACAGGCTCGGT

Petrus Tang, Ph.D. (鄧致剛)
Graduate Institute of Basic Medical Sciences
and
Bioinformatics Center, Chang Gung University.
petang@mail.cgu.edu.tw
EXT: 5136

http://petang.cgu.edu.tw/



Molecular Regulation and Bioinformatics Laboratory

The Molecular Regulation & Bioinformatics Laboratory (MRBLab) use bioinformatics approaches to integrate multi-omic (including genomics, proteomics, transcriptomics, metablomics, etc.) analysis techniques to study parasite-host interaction and to develop pipelines/visualization methods that makes computational biology accessible to research scientists with no programming experience

Molecular Regulation & Bioinformatics Laboratory

Dept. Parasitology, College of Medicine, Chang Gung University, TAIWAN

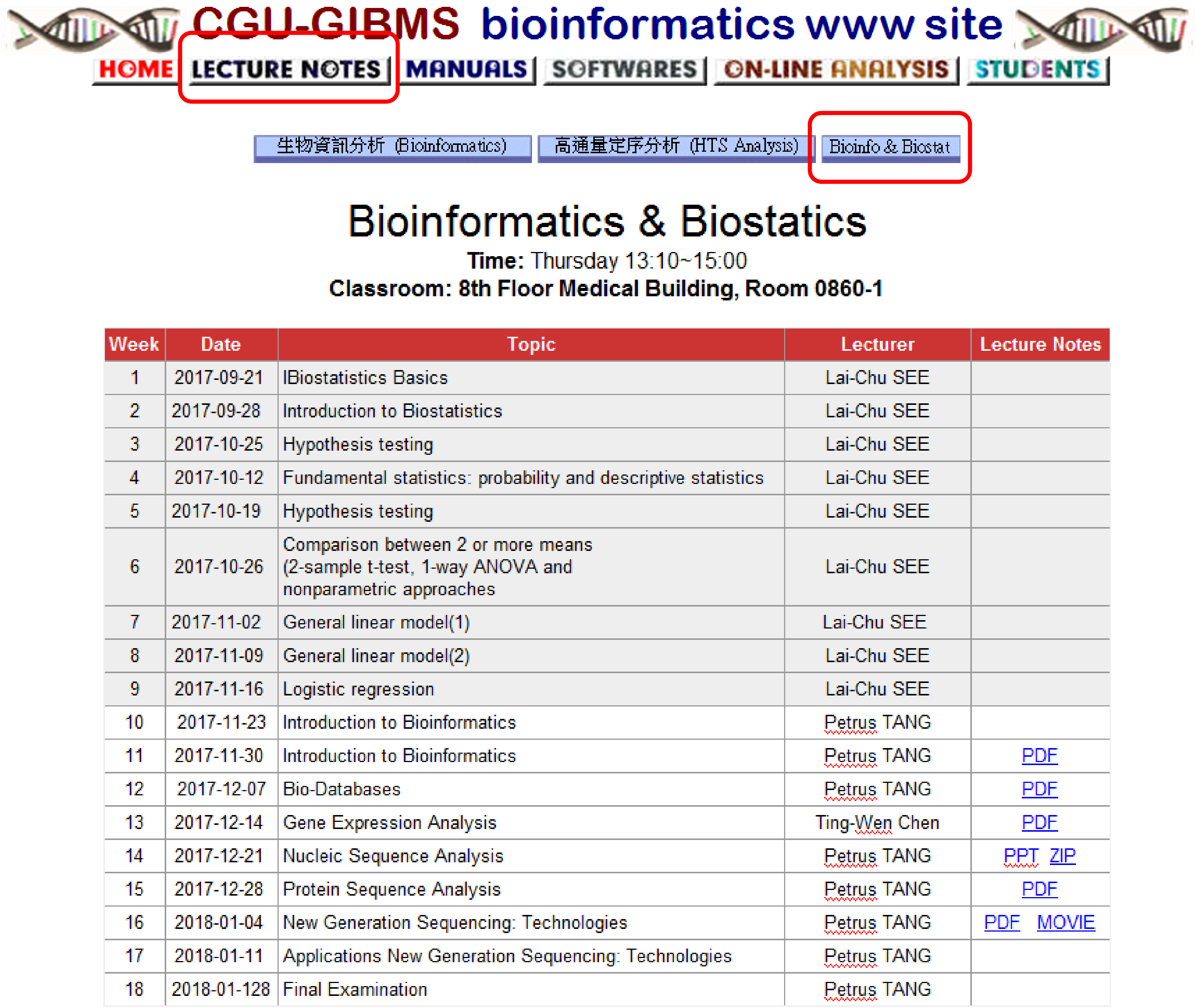
Prof. Petrus Tang, Ph.D.

Address:
Dept. of Parasitology, College of Medicine,
Chang Gung University.
259 Wenhwa 1st. Road. Kweishan,
Taoyuan 333. Taiwan.
Phone number:
+886-3-2118800#5136
Fax:
+886-3-2118122
E-mail:
petang@mail.cgu.edu.tw

LATEST NEWS

[Announcements]
● Trichomonas videos
1. *Trichomoniasis: Why is It the Neglected STD?* by M. Hobbs
2. *A Trich-y parasite: genomics, population genetics and evolution of Trichomonas vaginalis* by Jane Carlton

[Lectures]
● *Applications of Bioinformatics and Computational Techniques in Biomedical Research*



CGU-GIBMS bioinformatics www site

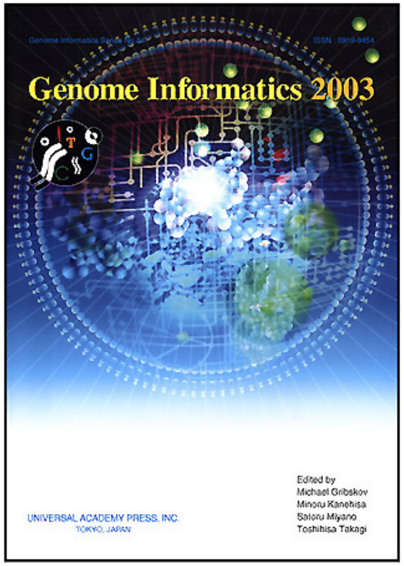
Bioinformatics & Biostatics

Time: Thursday 13:10~15:00

Classroom: 8th Floor Medical Building, Room 0860-1

Week	Date	Topic	Lecturer	Lecture Notes
1	2017-09-21	IBiostatistics Basics	Lai-Chu SEE	
2	2017-09-28	Introduction to Biostatistics	Lai-Chu SEE	
3	2017-10-25	Hypothesis testing	Lai-Chu SEE	
4	2017-10-12	Fundamental statistics: probability and descriptive statistics	Lai-Chu SEE	
5	2017-10-19	Hypothesis testing	Lai-Chu SEE	
6	2017-10-26	Comparison between 2 or more means (2-sample t-test, 1-way ANOVA and nonparametric approaches	Lai-Chu SEE	
7	2017-11-02	General linear model(1)	Lai-Chu SEE	
8	2017-11-09	General linear model(2)	Lai-Chu SEE	
9	2017-11-16	Logistic regression	Lai-Chu SEE	
10	2017-11-23	Introduction to Bioinformatics	Petrus TANG	
11	2017-11-30	Introduction to Bioinformatics	Petrus TANG	PDF
12	2017-12-07	Bio-Databases	Petrus TANG	PDF
13	2017-12-14	Gene Expression Analysis	Ting-Wen Chen	PDF
14	2017-12-21	Nucleic Sequence Analysis	Petrus TANG	PPT ZIP
15	2017-12-28	Protein Sequence Analysis	Petrus TANG	PDF
16	2018-01-04	New Generation Sequencing: Technologies	Petrus TANG	PDF MOVIE
17	2018-01-11	Applications New Generation Sequencing: Technologies	Petrus TANG	
18	2018-01-128	Final Examination	Petrus TANG	

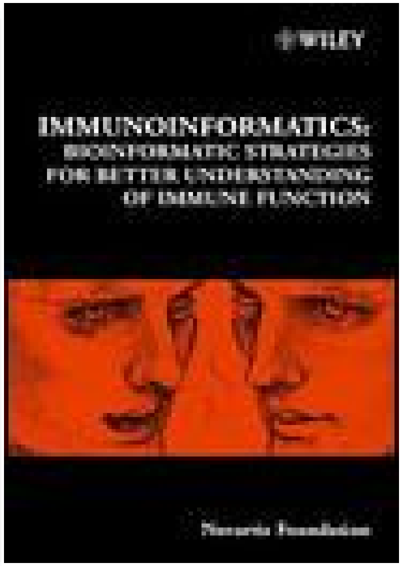
Bio informatics


Genome Informatics 2003

Edited by
Michael Gribskov
Minoru Kanehisa
Satoru Miyano
Toshihisa Takagi

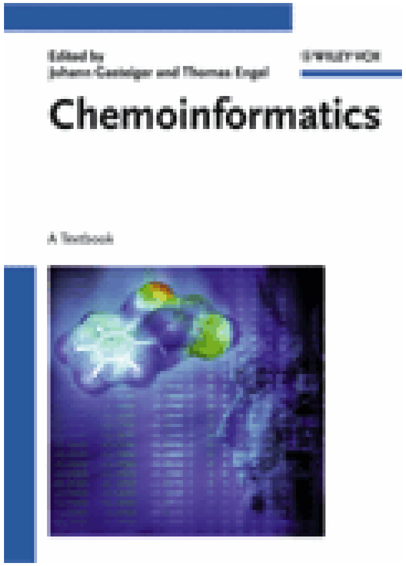
UNIVERSAL ACADEMY PRESS, INC.
TOKYO, JAPAN


Neuroinformatics


WILEY

IMMUNOINFORMATICS:
BIOINFORMATIC STRATEGIES
FOR BETTER UNDERSTANDING
OF IMMUNE FUNCTION

Novartis Foundation


Edited by
Johann Gasteiger and Thomas Engel

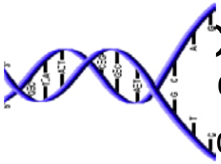
WILEY-VCH

Chemoinformatics

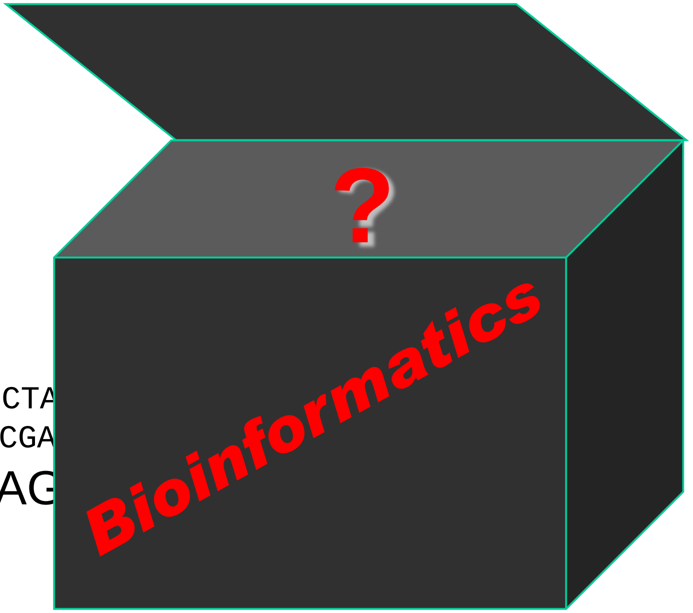
A Textbook

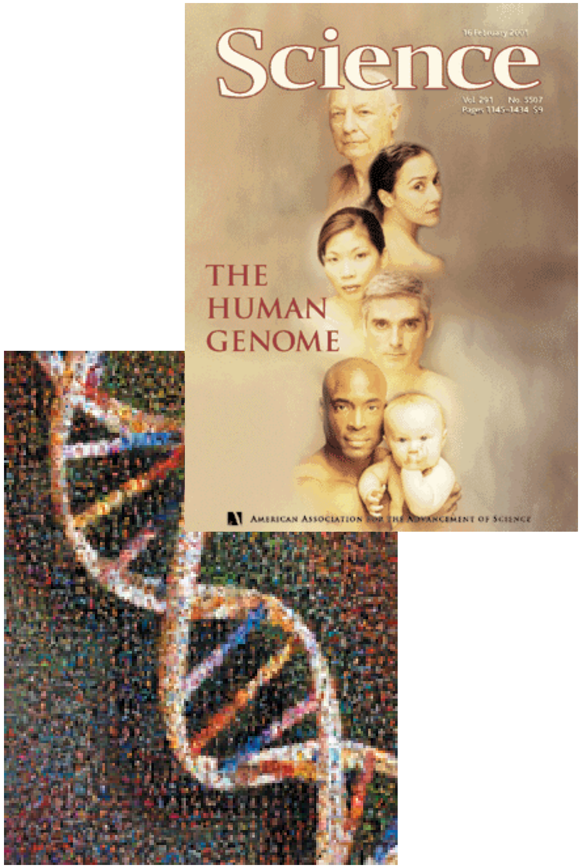
-Omics Mania *biome, cellomics, chronomics, clinomics, complexome, crystallomics, cytomics, degradomics, diagnomics, enzymome, epigenome, expressome, fluxome, foldome, secretome, functome, functomics, genomics, glycomics, immunome, transcriptomics, integromics, interactome, kinome, ligandomics, lipoproteomics, localizome, phenomics, metabolome, pharmacometabonomics, methylome, microbiome, morphome, neurogenomics, nucleome, secretome, oncogenomics, operome, transcriptomics, ORFeome, parasitome, pathome, peptidome, pharmacogenome, pharmacomethylomics, phenomics, phylome, physiogenomics, postgenomics, predictome, promoterome, proteomics, pseudogenome, secretome, regulome, resistome, ribonome, ribonomics, riboproteomics, saccharomics, secretome, somatonome, systeome, toxicomics, transcriptome, translatome, secretome, unknome, vaccinome, variomics...*

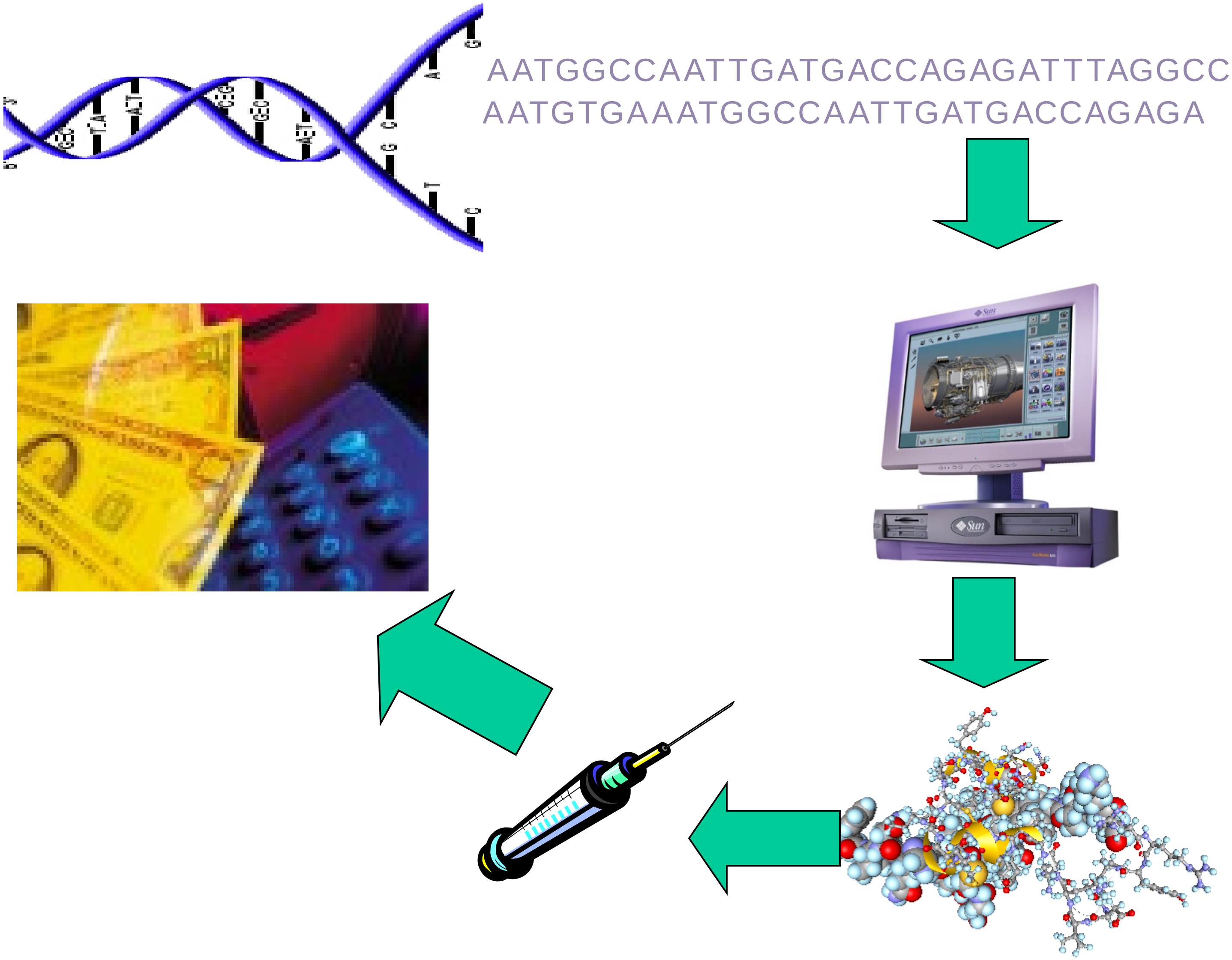
WHAT IS BIOINFORMATICS?



TAGCTAGCTAGCTA
CTAGCTAGCTAGCTATCGA
CATGCACTAGCTAG







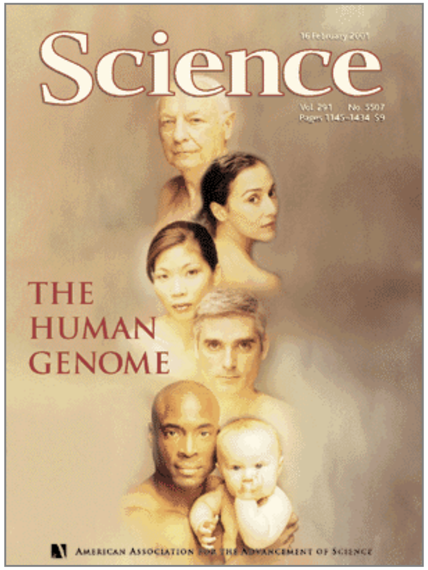
AATGGCCAATTGATGACCAGAGATTTAGGCCAATTAA
AATGTGAAATGGCCAATTGATGACCAGAGA

What is Bioinformatics?

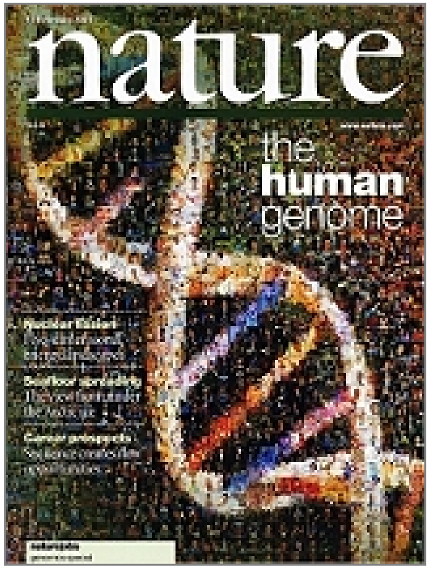
- Development of methods & algorithms to organize, integrate, analyze and interpret biological and biomedical data
- Study of the inherent structure & flow of biological information
- Goals of bioinformatics:
    - Identify patterns
    - Classify
    - Make predictions
    - Create models
    - Better utilize existing knowledge

February 2001: Completion of the Draft Human Genome

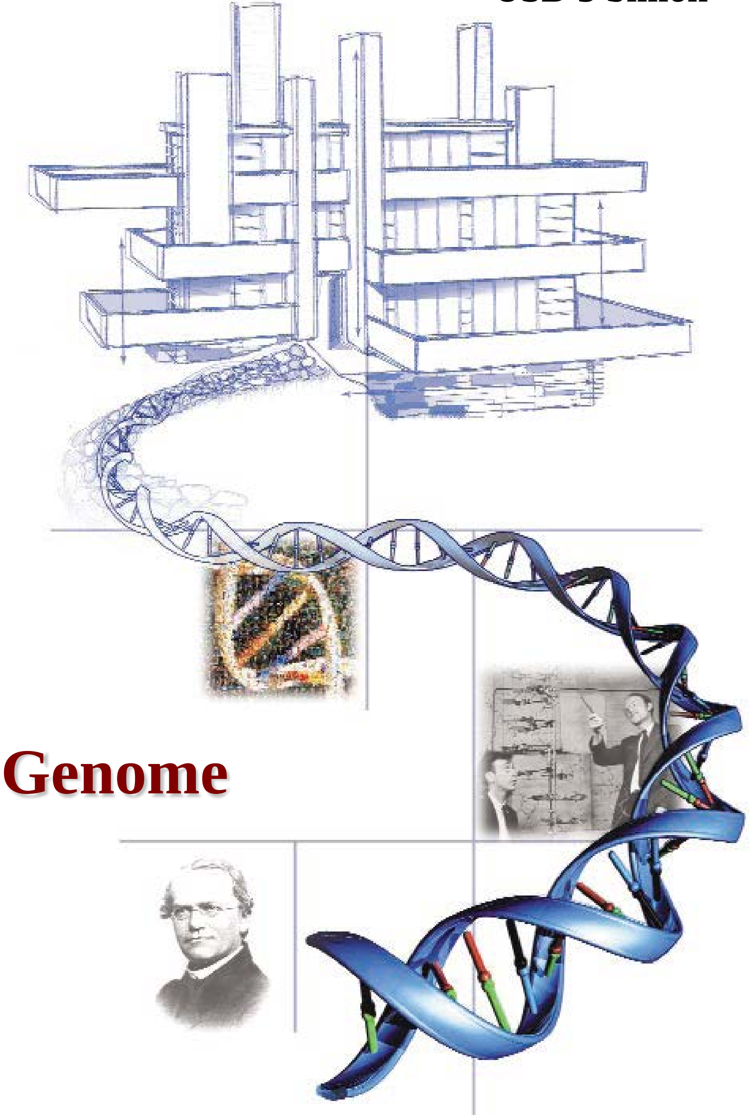
>10 years to finish
USD 3 billion



Science, 16 February 2001
Vol. 291, Pages 1145-1434



Nature, 15 February 2001
Vol. 409, Pages 813-960



50 Years of DNA: *From Double Helix to Health*
A Celebration of the Genome

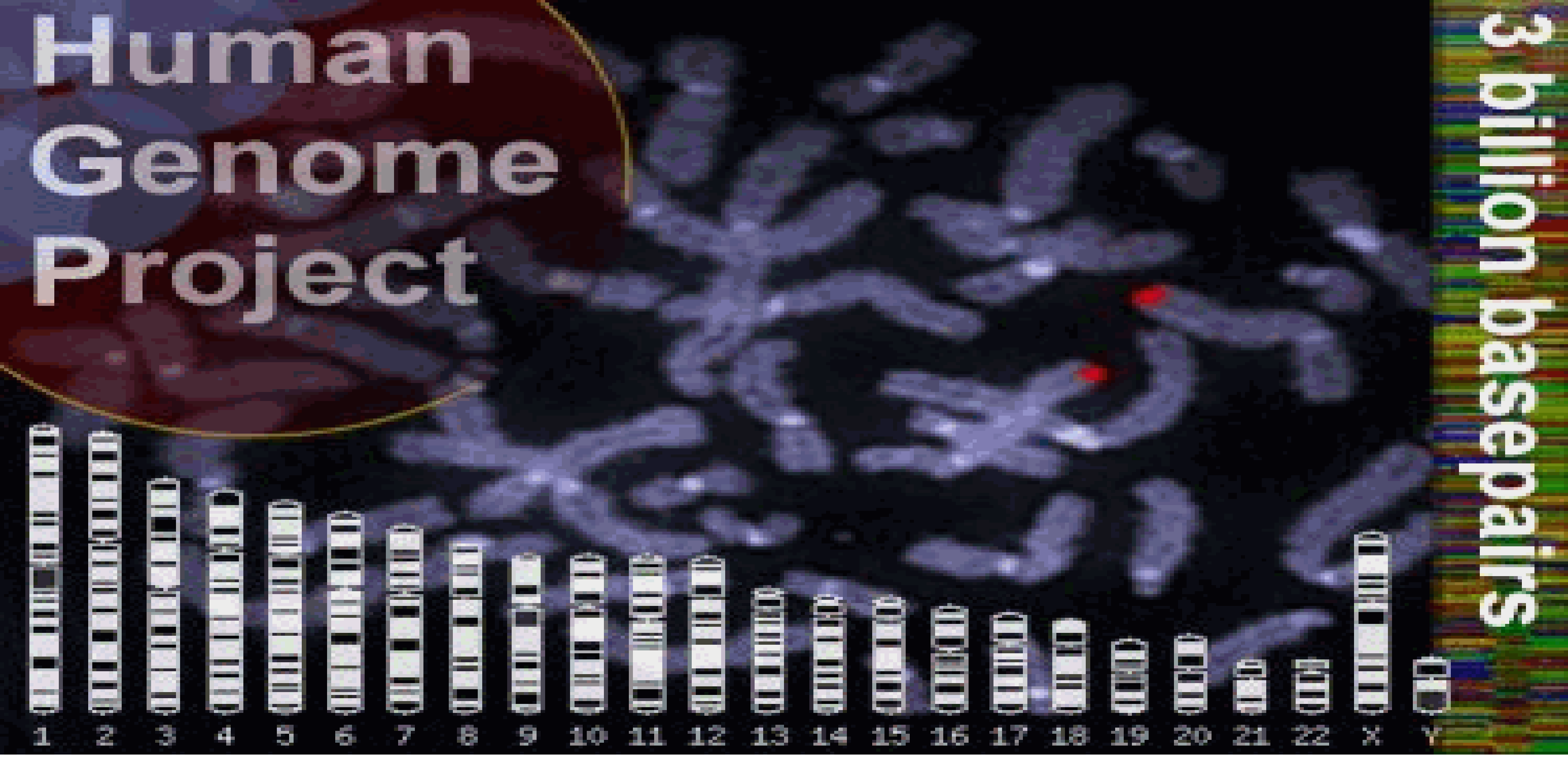
Nature, 23 April 2003
Vol. 422, Pages 1-13

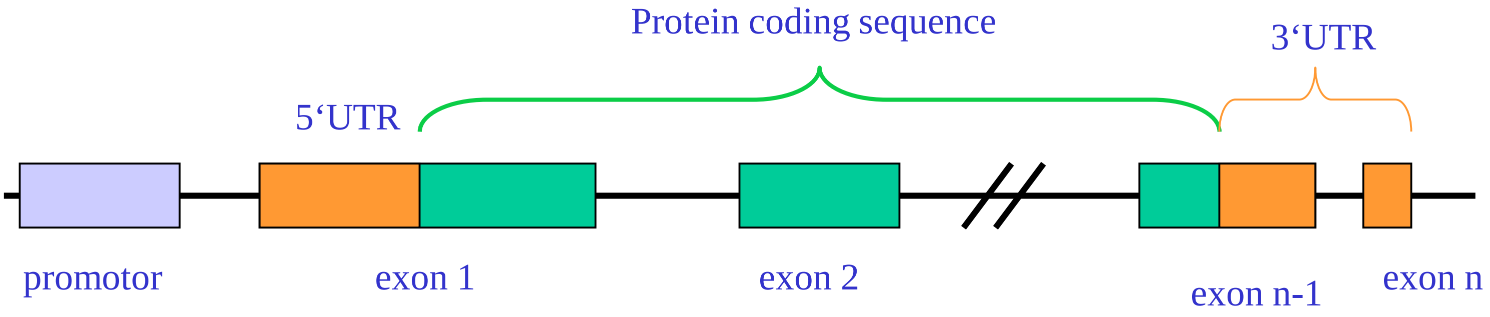
April 2003: High-Resolution Human Genome

A vision for the future of genomics research

A blueprint for the genomic era.

Francis S. Collins, Eric D. Green, Alan E. Guttmacher and Mark S. Guyer on behalf of the US National Human Genome Research Institute*





Protein coding sequence

3‘UTR

5‘UTR

promotor

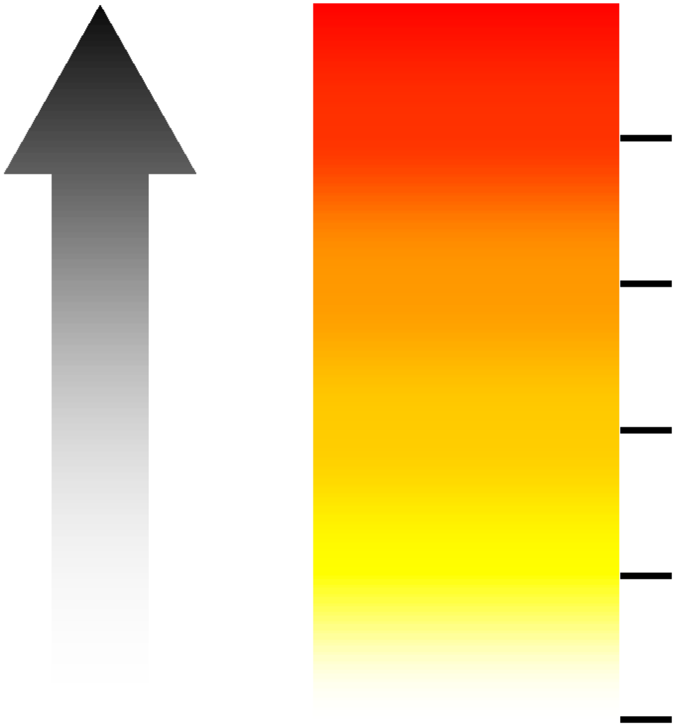
exon 1

exon 2

exon n-1

exon n

Gene Number in the Human Genome



Number of genes

10 K Known genes

20 K Otto
4
3
2

30 K

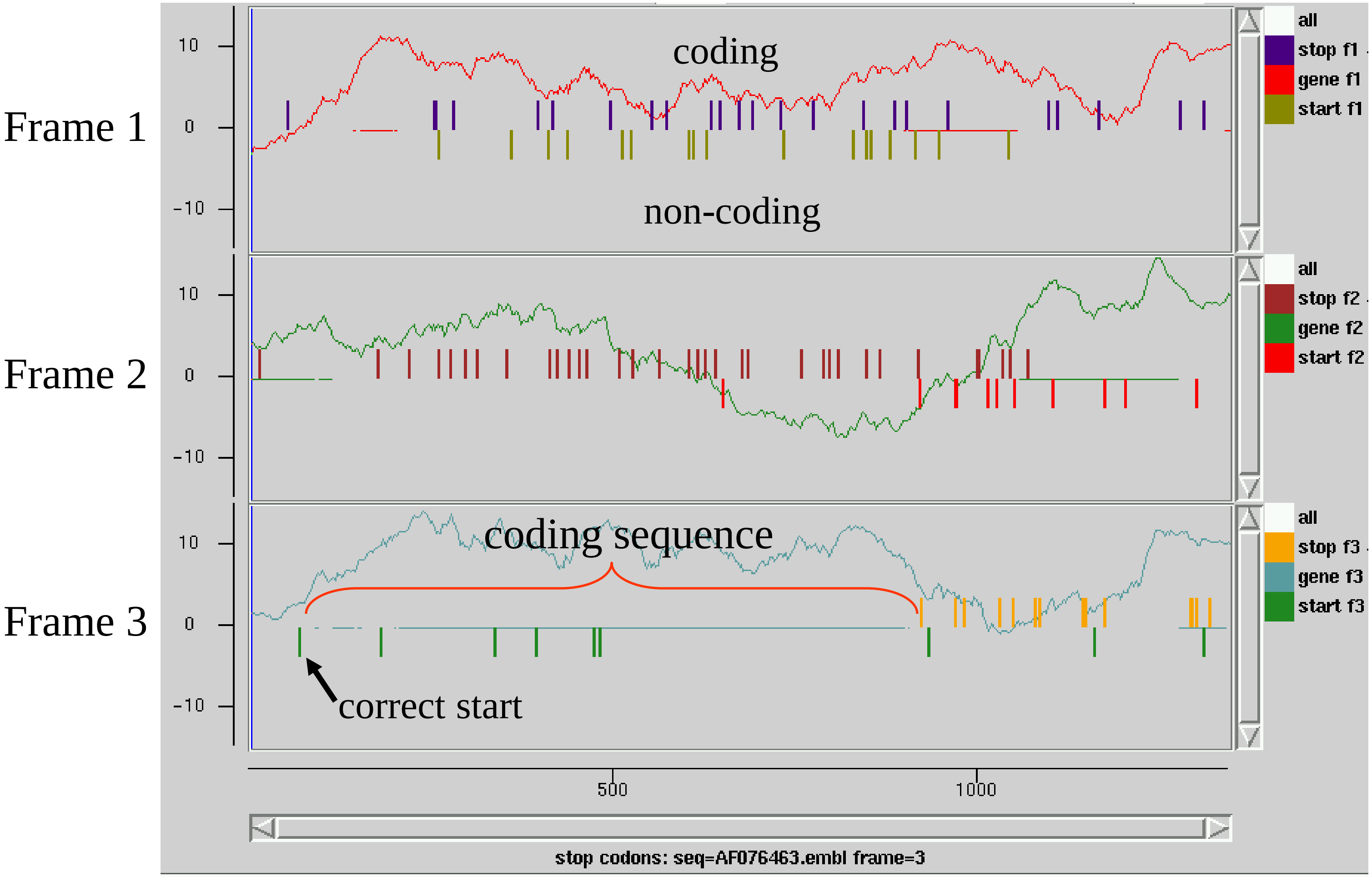
40 K 1

50K

Confidence

Gene prediction

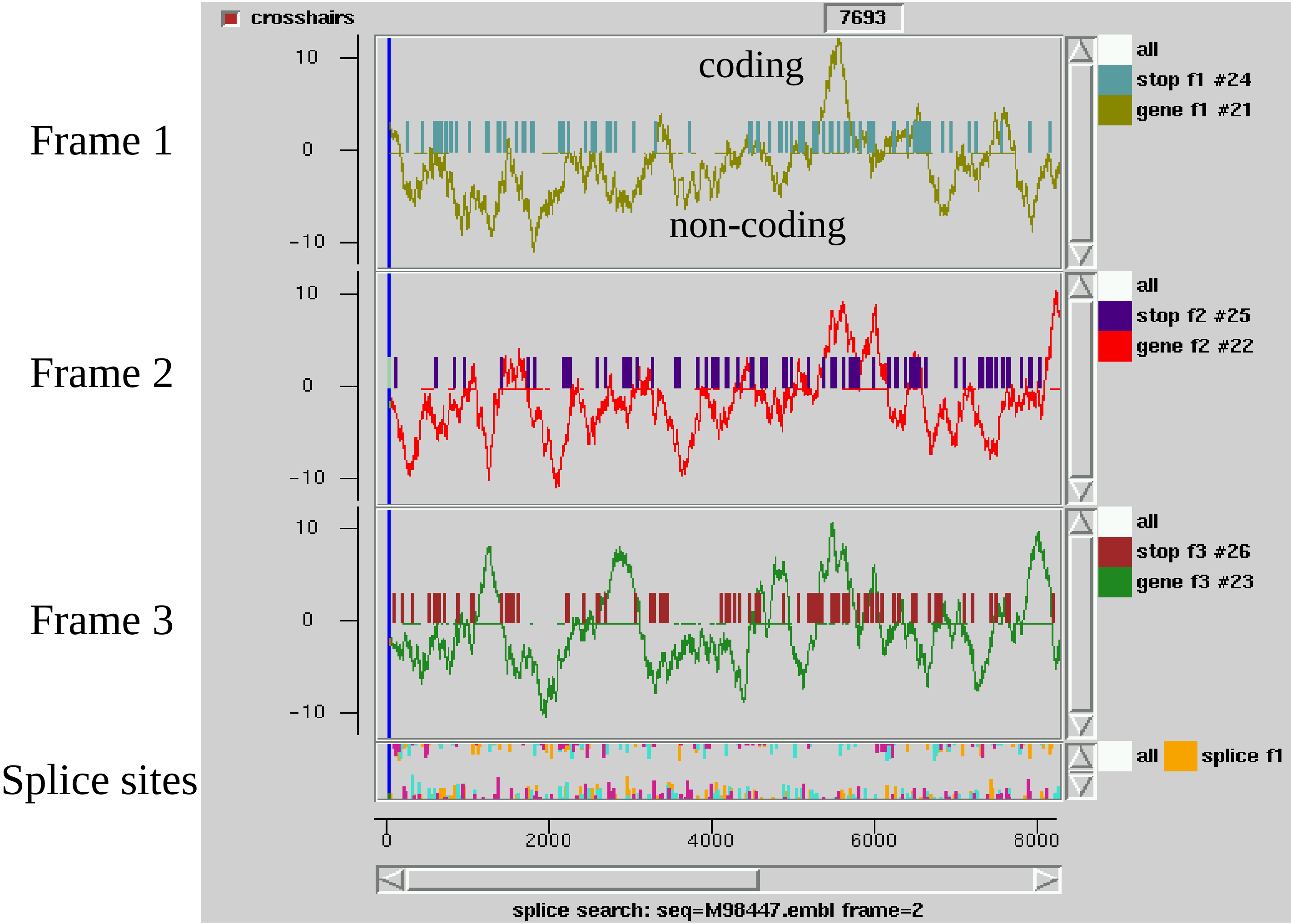
Codon usage (single exon)



stop codons: seq=AF076463.embl frame=3

Gene prediction

Codon usage (multiple exons)

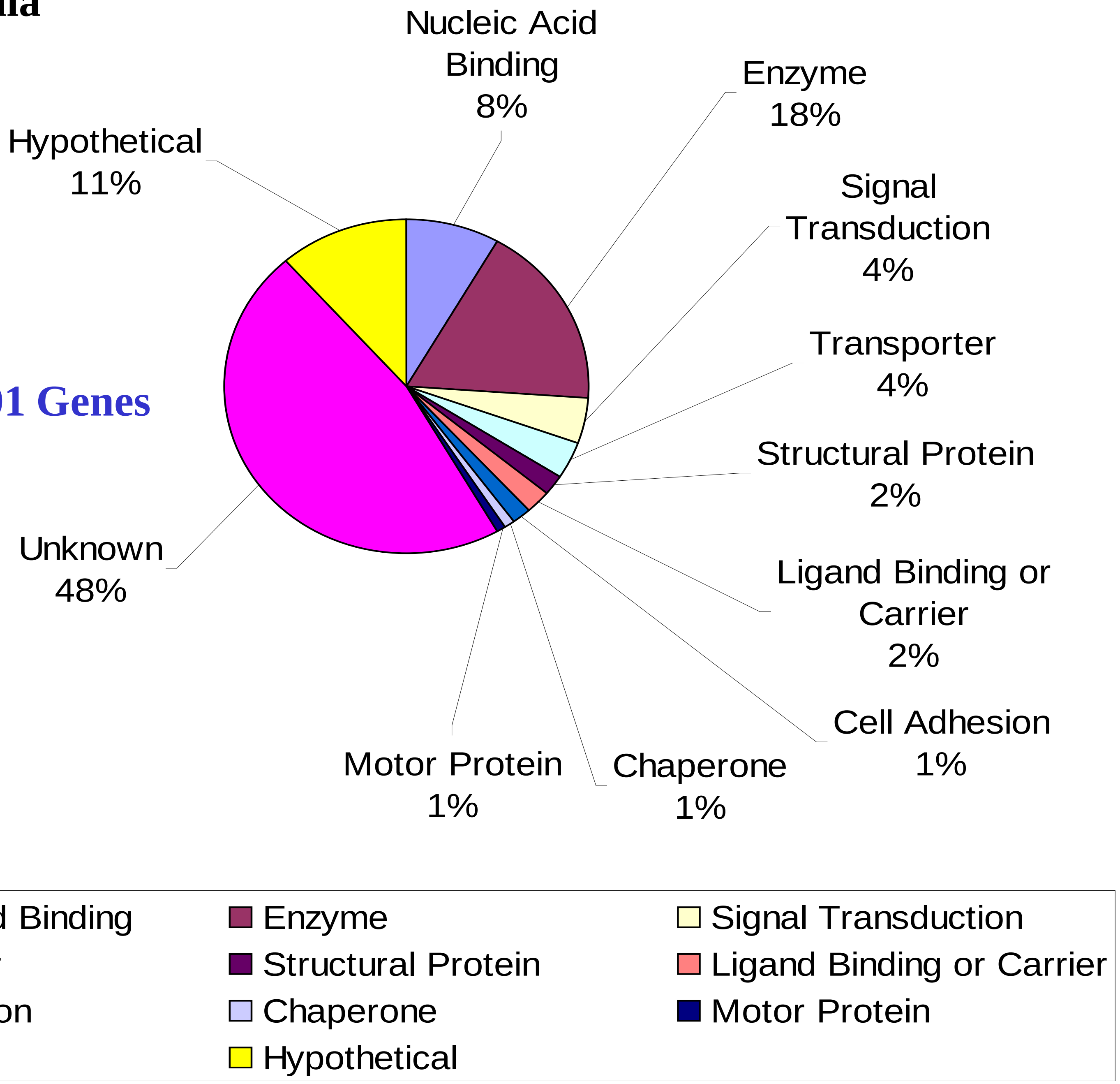


splice search: seq=M98447.embl frame=2

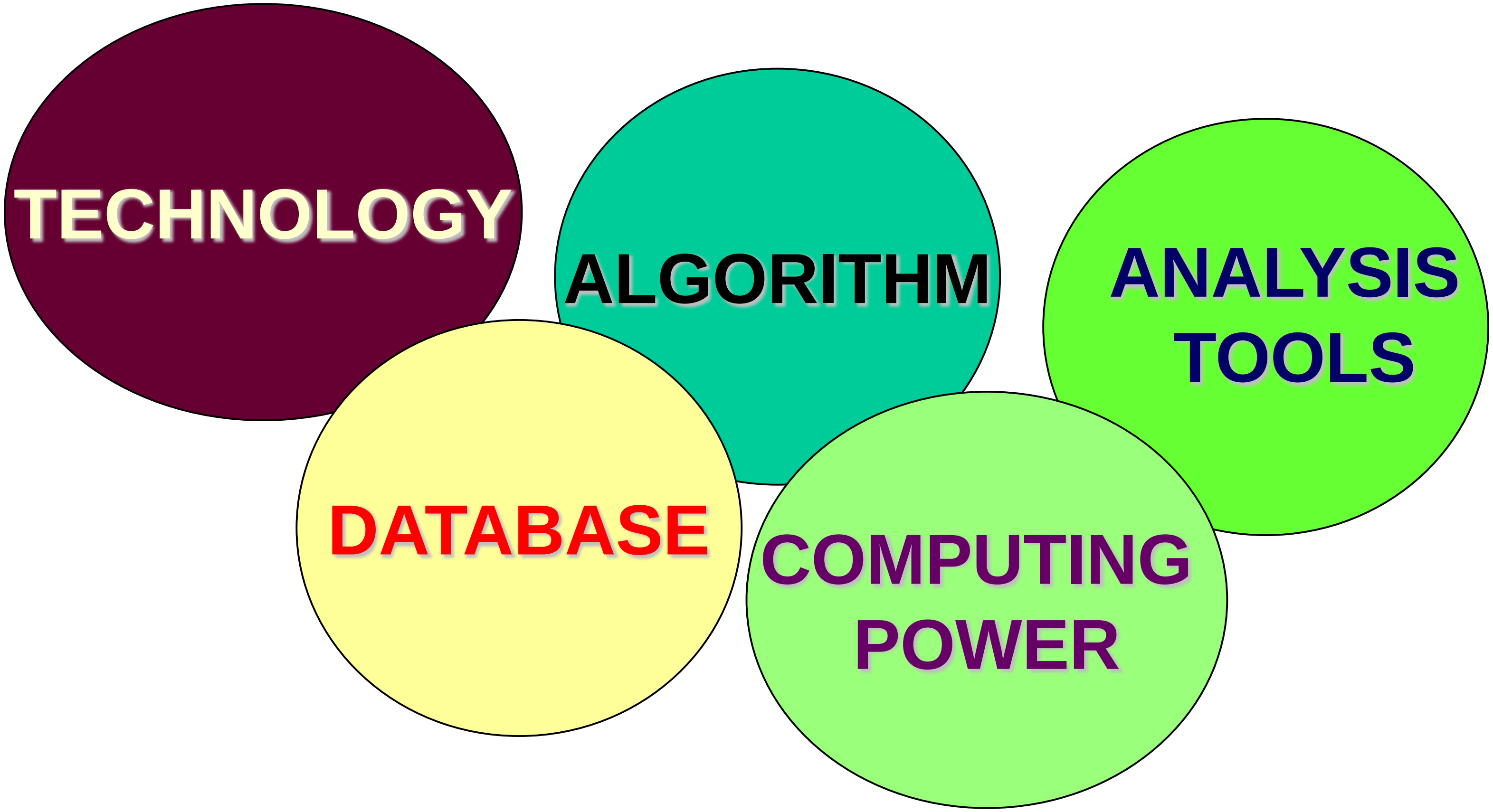
Exons:
208. .295
1029. .1349
1500. .1688
2686. .2934
3326. .3444
3573. .3680
4135. .4309
4708. .4846
4993. .5096
7301. .7389
7860. .8013
8124. .8405
8553. .8713
9089. .9225
13841. .14244

Drosophila

13,601 Genes



THE COMPONENTS OF BIOINFORMATICS



TECHNOLOGY

ALGORITHM

ANALYSIS TOOLS

DATABASE

COMPUTING POWER

THE COMPONENTS OF BIOINFORMATICS

TECHNOLOGY

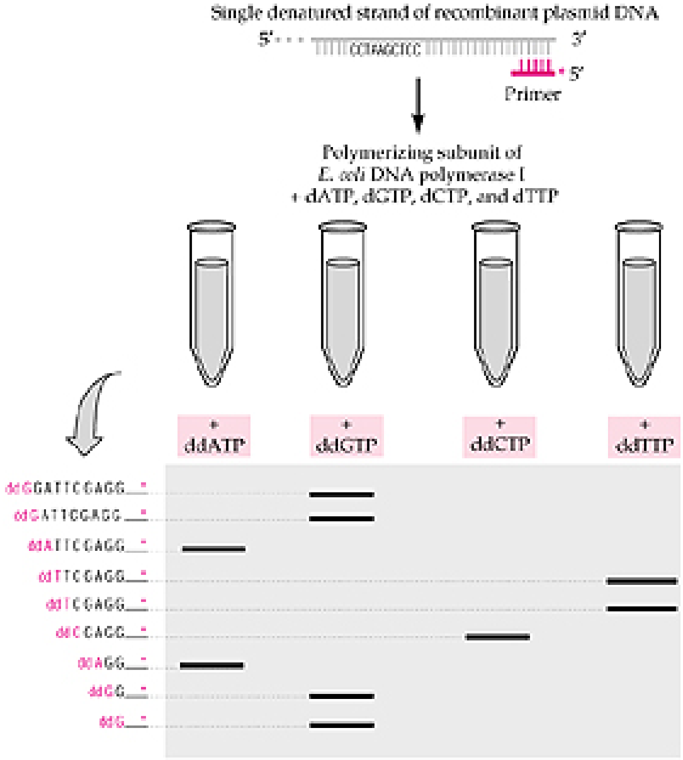
ALGORITHM

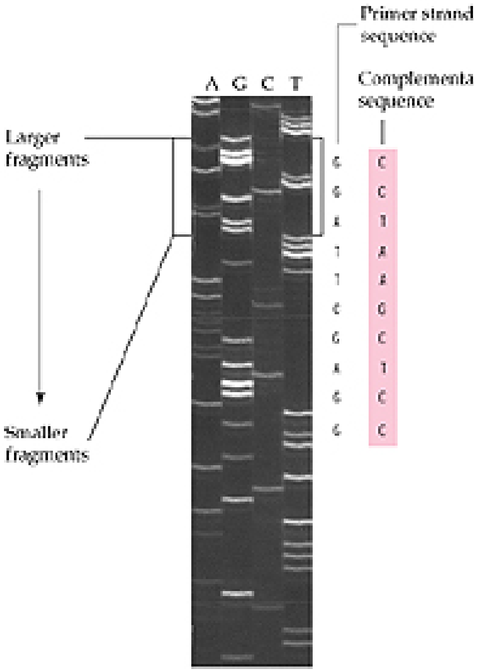
ANALYSIS TOOLS

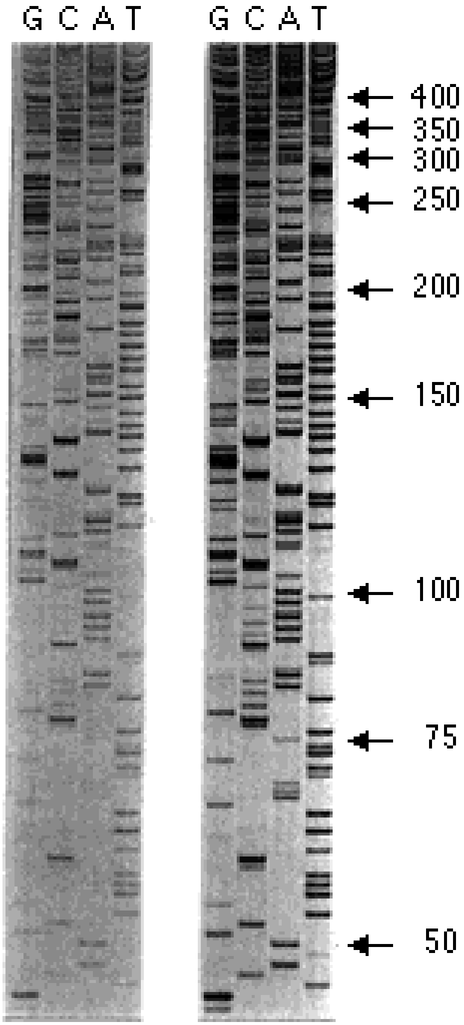
DATABASE

COMPUTING POWER

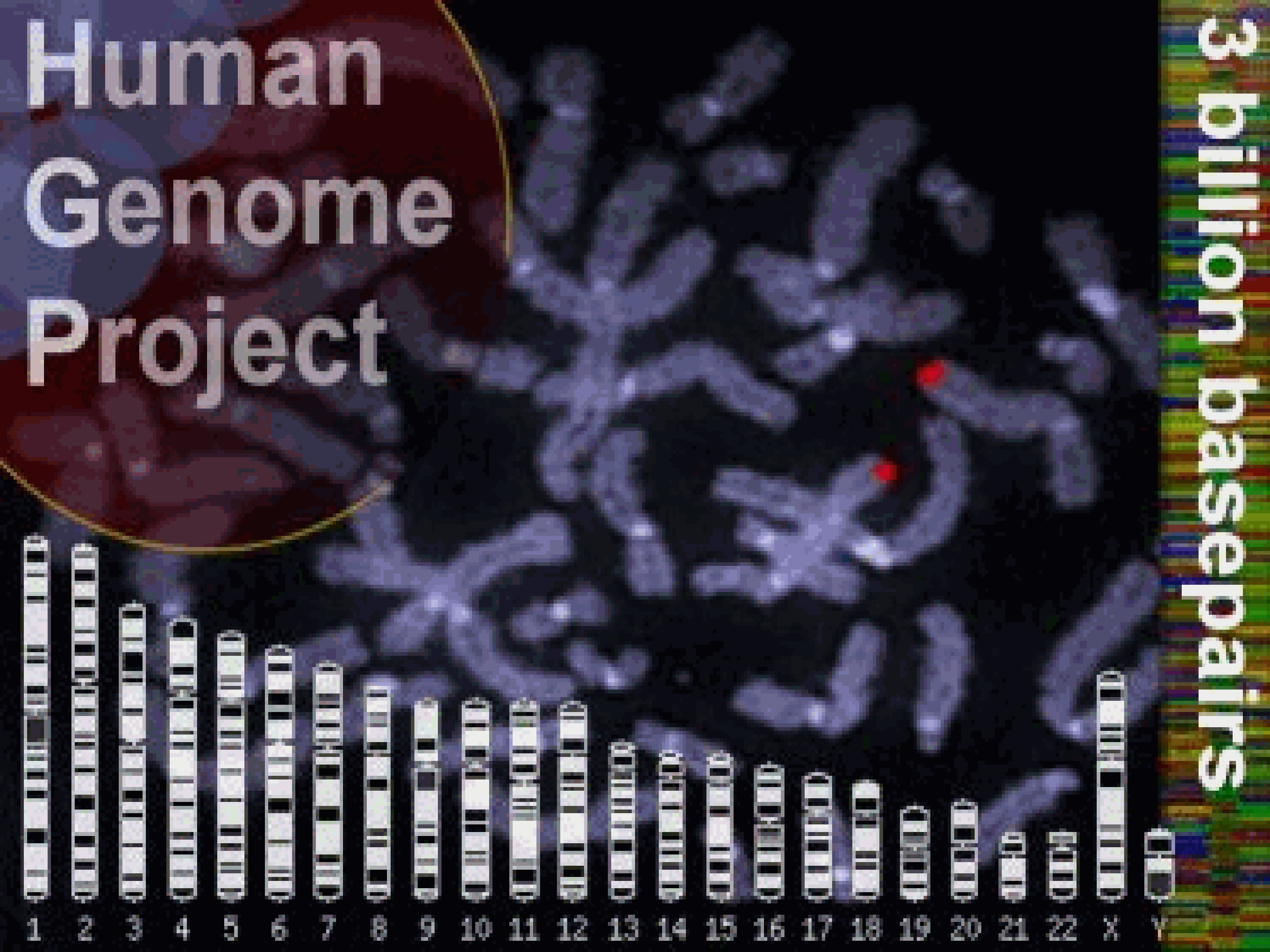
Sanger Dideoxy Sequencing





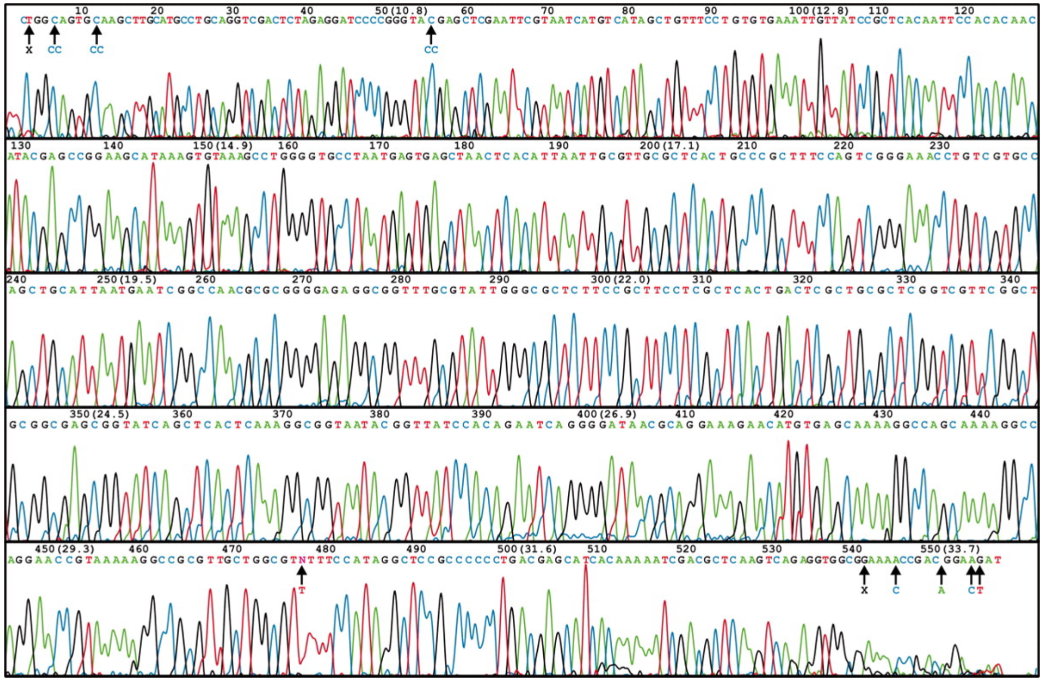


Human Genome Project



3 billion basepairs

ABI 3730 XL DNA Sequencer

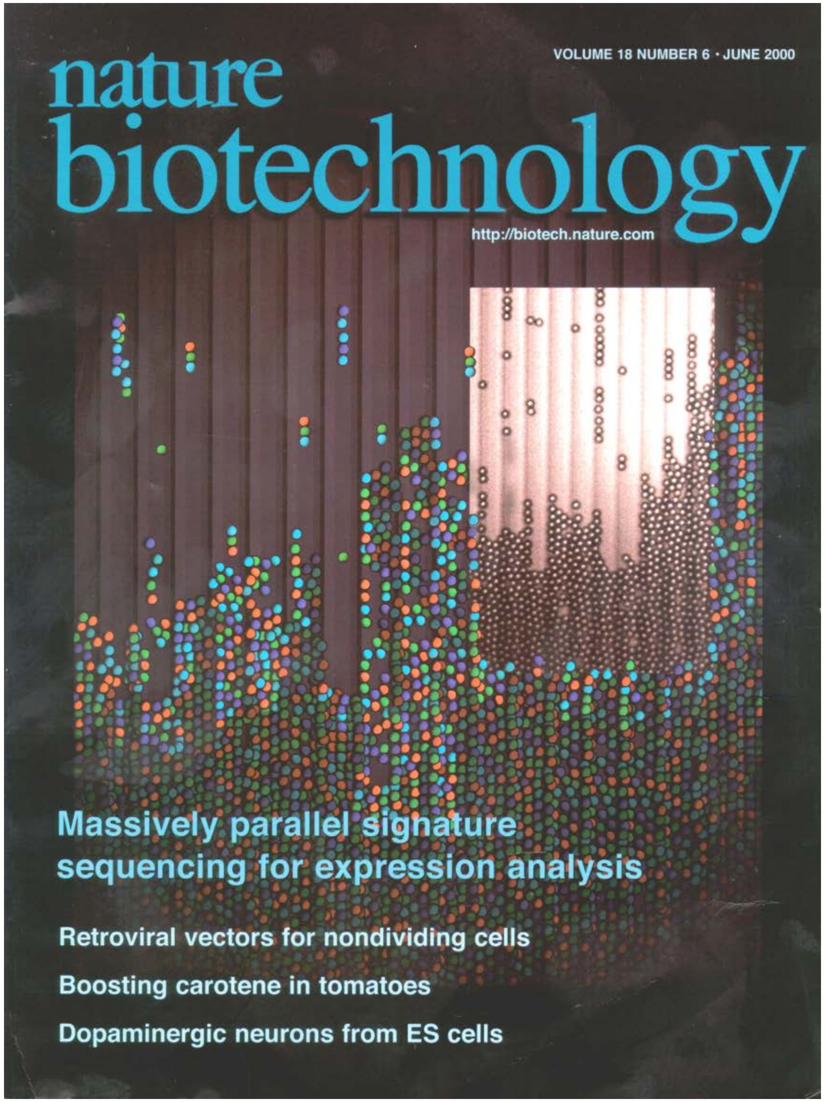


96/384 DNA sequencing in 2 hrs, approximately 600-1000 readable bps per run.

1-4 MB bps/day

A human genome of 3GB need 750 days to finish

Next Generation Sequencing (NGS) Technology



VOLUME 18 NUMBER 6 · JUNE 2000

nature biotechnology

http://biotech.nature.com

Massively parallel signature sequencing for expression analysis

Retroviral vectors for nondividing cells

Boosting carotene in tomatoes

Dopaminergic neurons from ES cells

ISSN 1088-9051

September 2008

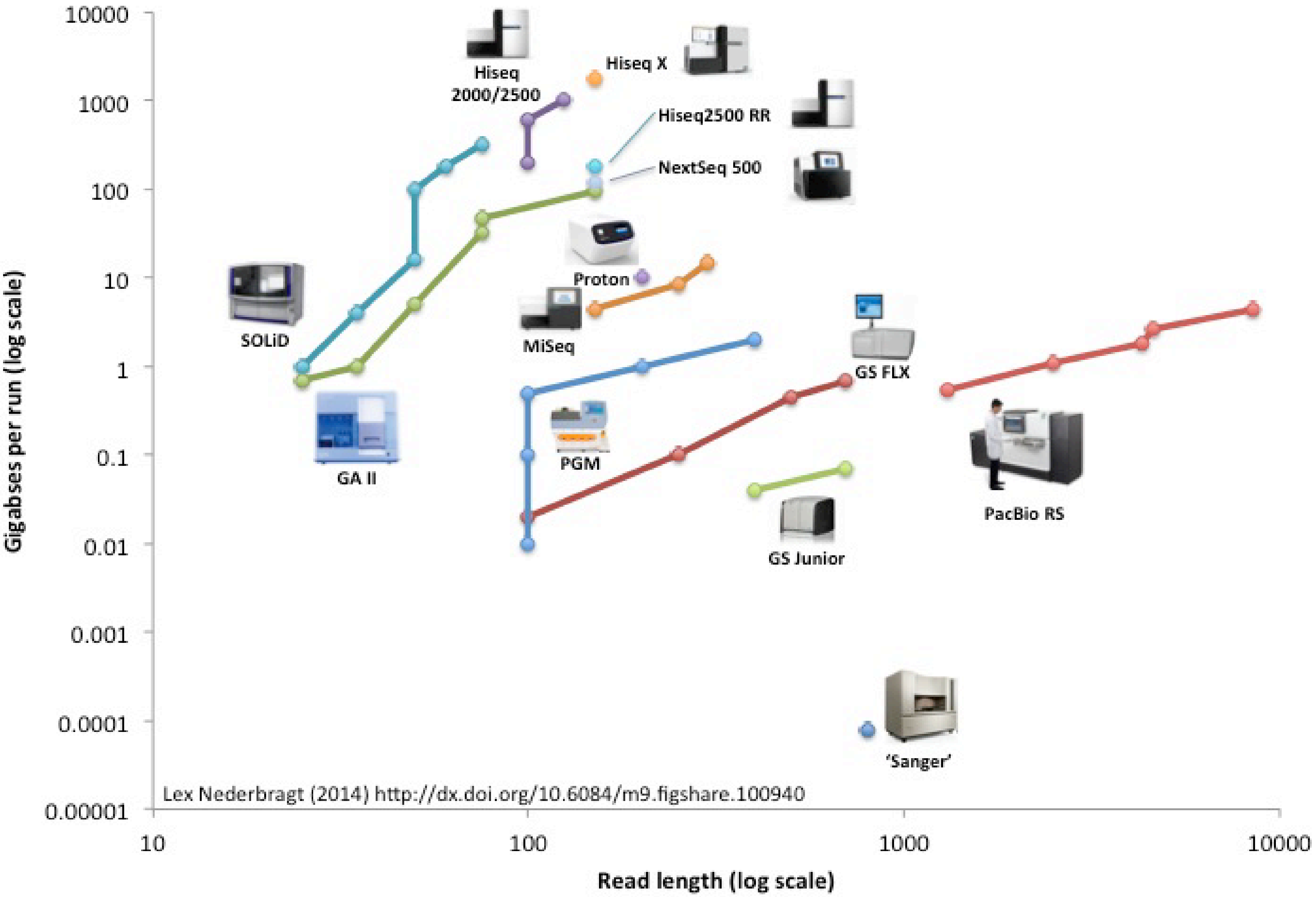
GENOME RESEARCH

Volume 18 Number 9

Non-coding RNA in Stem Cell Differentiation • Human Tissue-Specific Differentially Methylated Regions • Very Small Mobile Elements • Dachshund Cone-Rod Dystrophy • EagleView Genome Assembly Viewer

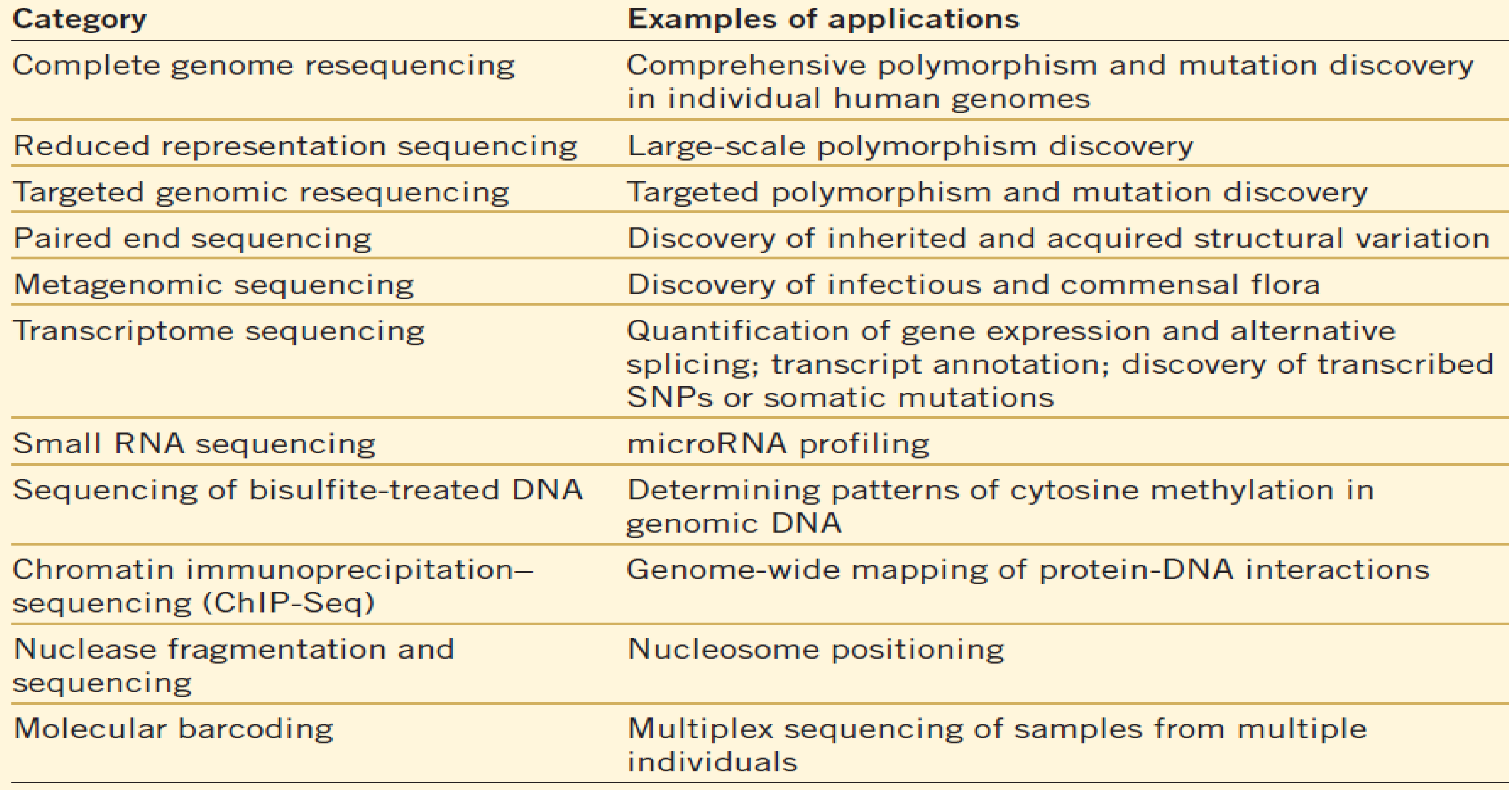
Cold Spring Harbor Laboratory Press

Throughput of NGS machines (2014)

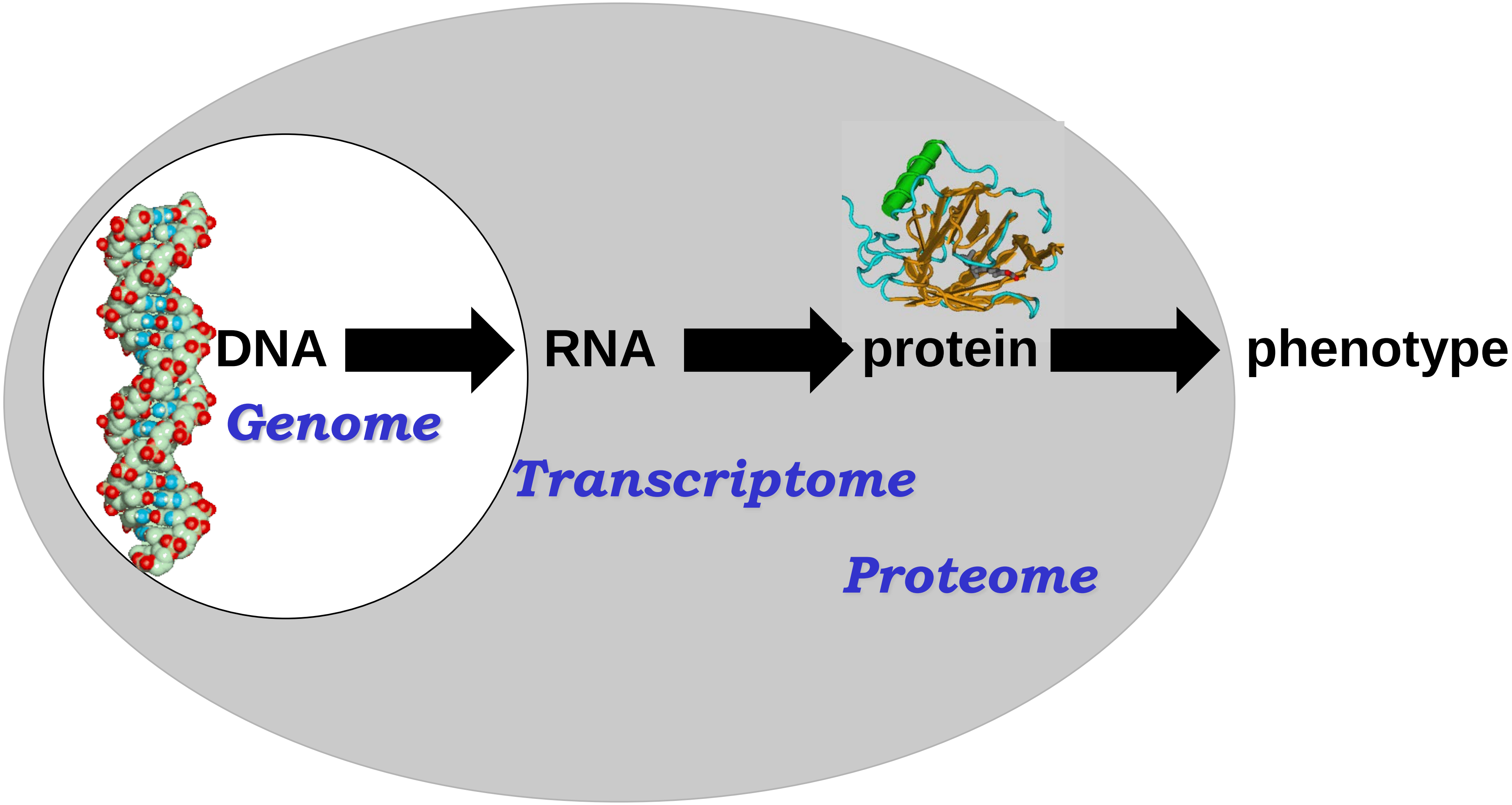


Lex Nederbragt (2014) http://dx.doi.org/10.6084/m9.figshare.100940

Applications on Biomedical Sciences



Category	Examples of applications
Complete genome resequencing	Comprehensive polymorphism and mutation discovery in individual human genomes
Reduced representation sequencing	Large-scale polymorphism discovery
Targeted genomic resequencing	Targeted polymorphism and mutation discovery
Paired end sequencing	Discovery of inherited and acquired structural variation
Metagenomic sequencing	Discovery of infectious and commensal flora
Transcriptome sequencing	Quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations
Small RNA sequencing	microRNA profiling
Sequencing of bisulfite-treated DNA	Determining patterns of cytosine methylation in genomic DNA
Chromatin immunoprecipitation–sequencing (ChIP-Seq)	Genome-wide mapping of protein-DNA interactions
Nuclease fragmentation and sequencing	Nucleosome positioning
Molecular barcoding	Multiplex sequencing of samples from multiple individuals



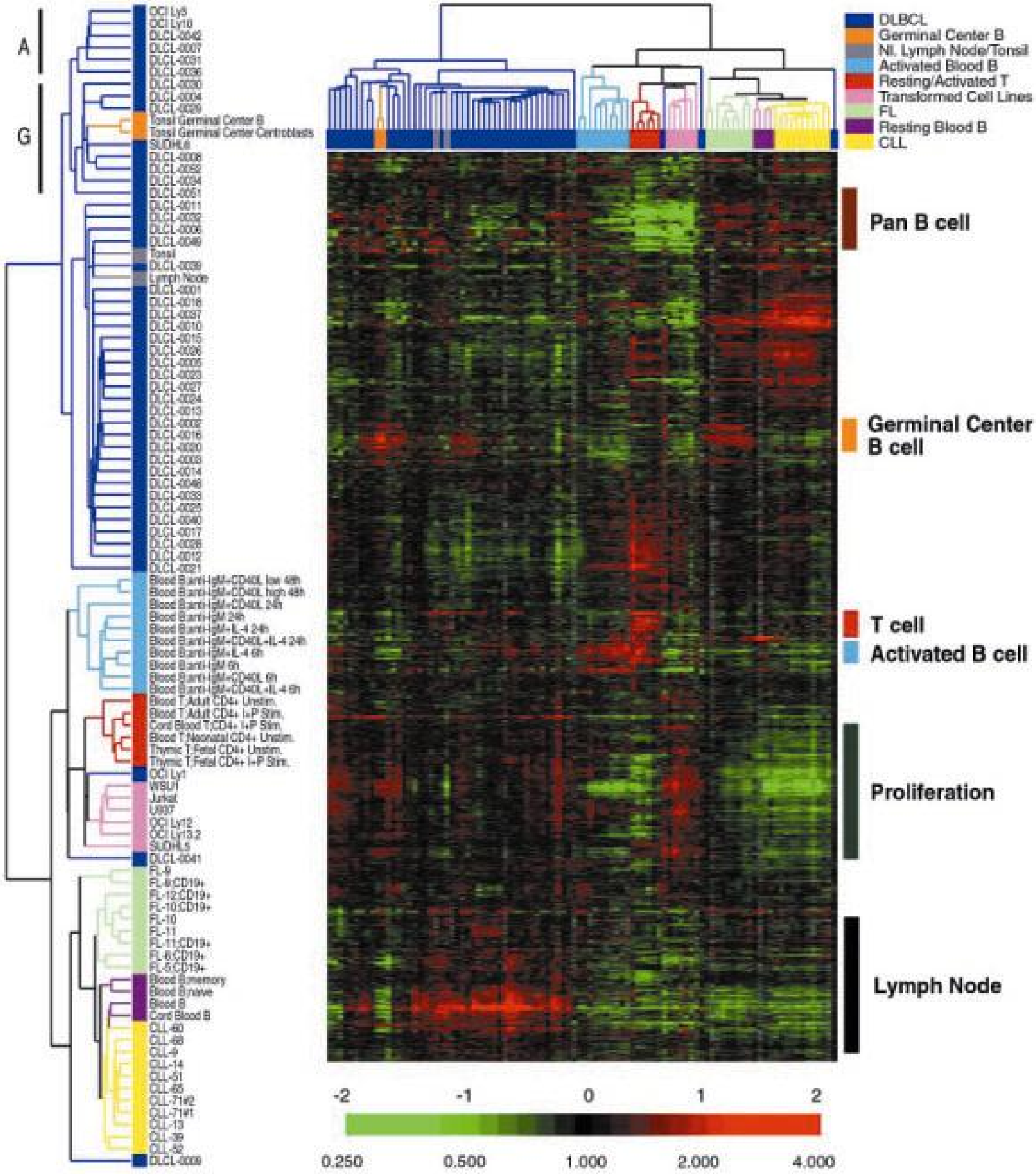
DNA → RNA → protein → phenotype

Genome

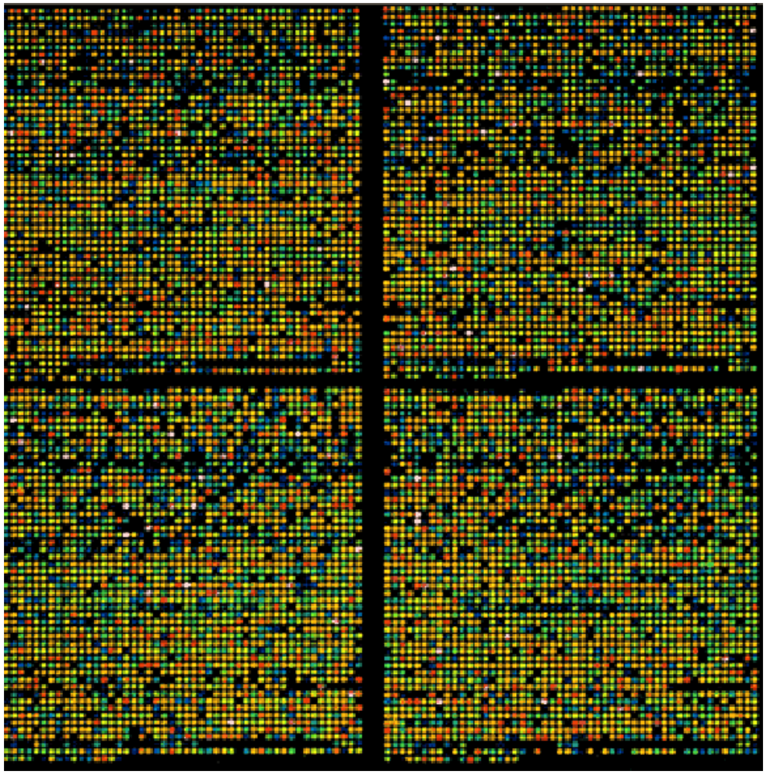
Transcriptome

Proteome

Microarray

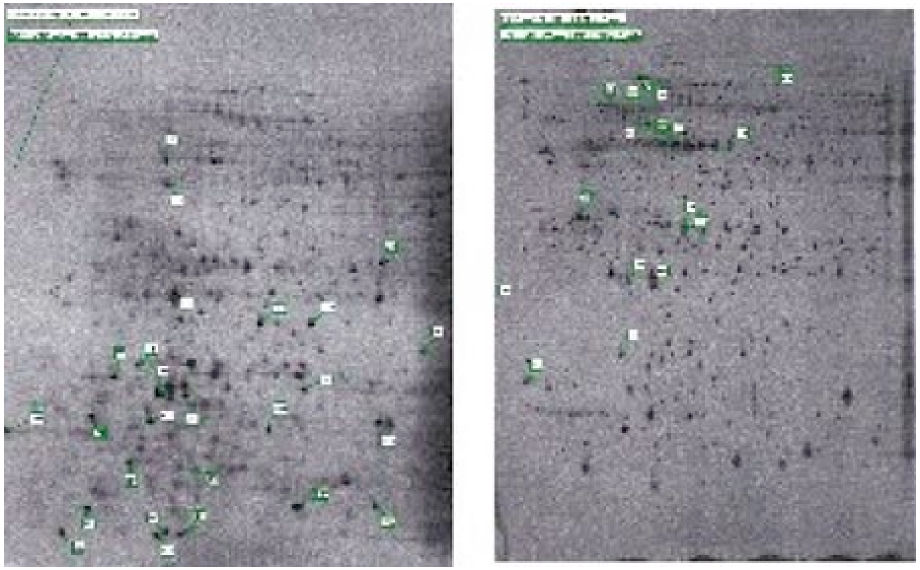


20,000-40,000
Clones
per
slide



http://llmpp.nih.gov/lymphoma/images/figure1.html (1 of 2) [2000/11/14 AM 08:54:21]

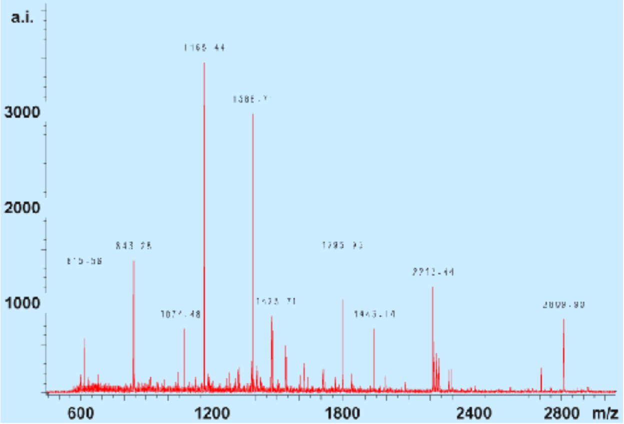
Proteomics



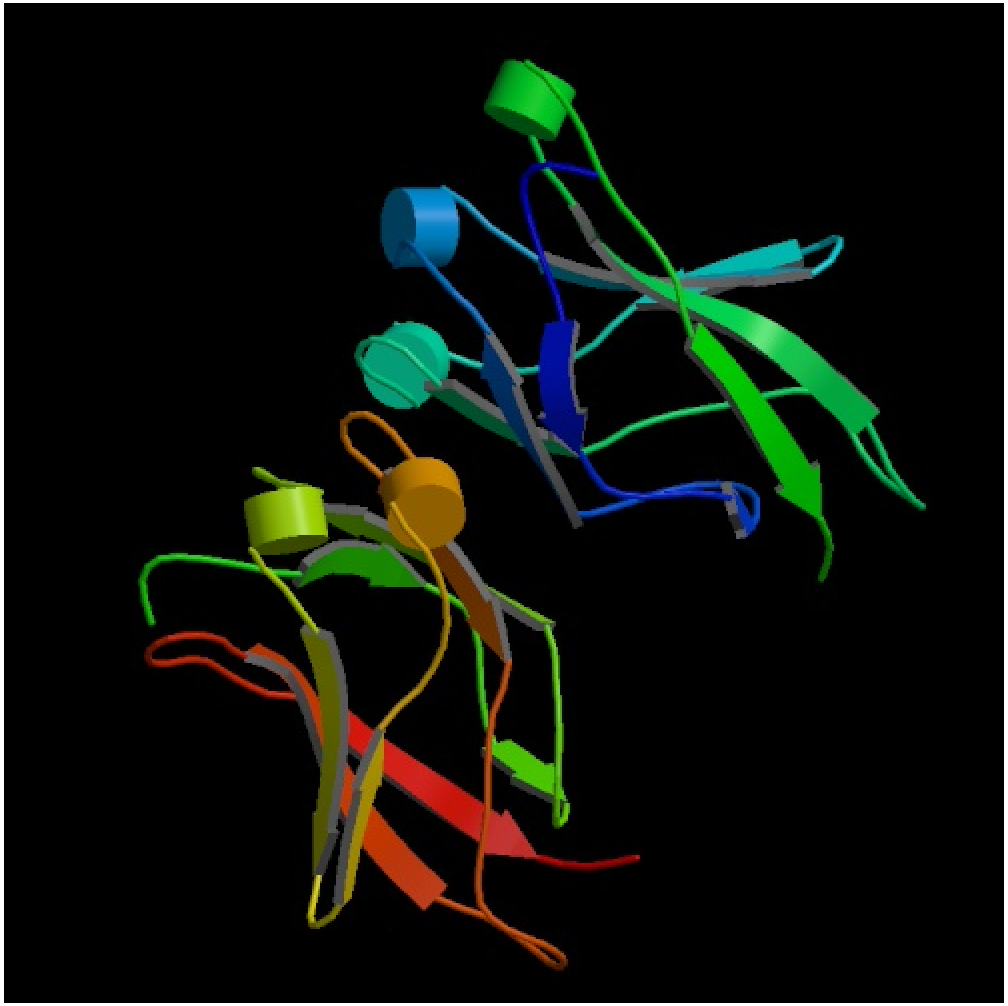
2 Dimensional Electrophoresis gels, differences that are characteristics of the individual starting states recognized by comparison of two protein pattern

6,000
protein spots
per gel

MALDI-MS peptide mass fingerprint, for identification of proteins separated by 2D electrophoresis



3D Modeling



THE COMPONENTS OF BIOINFORMATICS

TECHNOLOGY

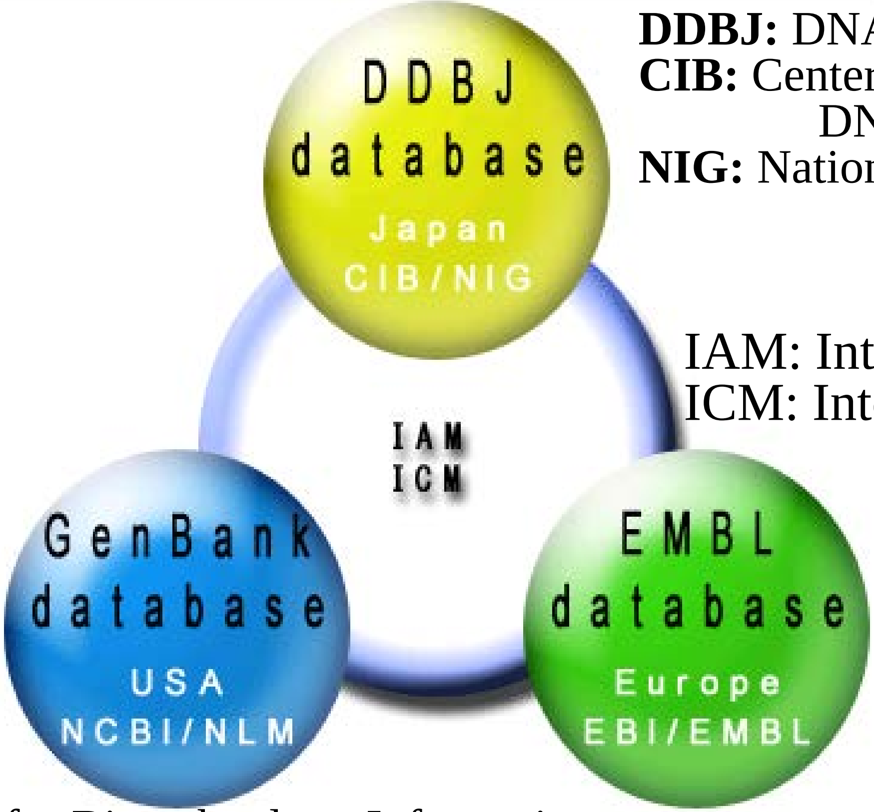
ALGORITHM

ANALYSIS TOOLS

DATABASE

COMPUTING POWER

GenBank/EMBL/DDBJ International Nucleotide Sequence Database



DDBJ: DNA Data Bank of Japan
CIB: Center for Information Biology and DNA Data Bank of Japan
NIG: National Institute of Genetics

IAM: International Advisory Meeting
ICM: International Collaborative Meeting

EMBL:
European Molecular Biology Laboratory
EBI:
European Bioinformatics Institute

NCBI:
National Center for Biotechnology Information
NLM:
National Library of Medicine

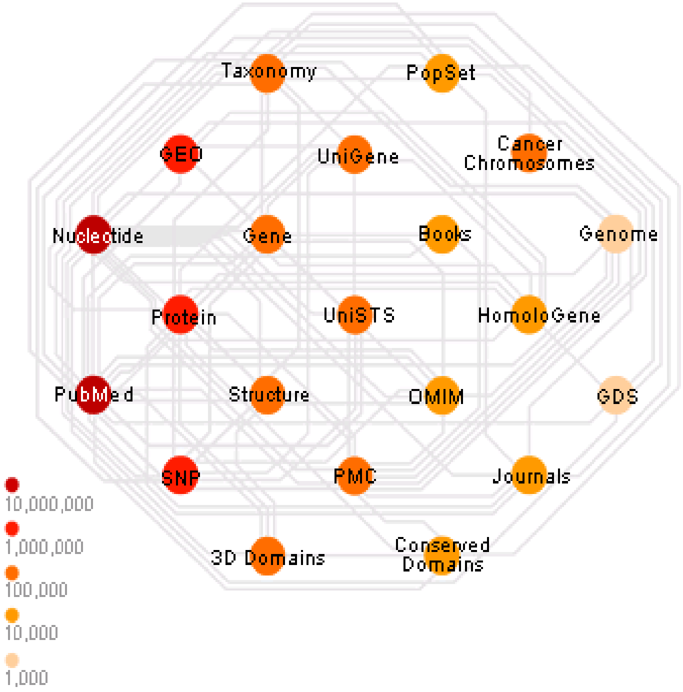
GenBank

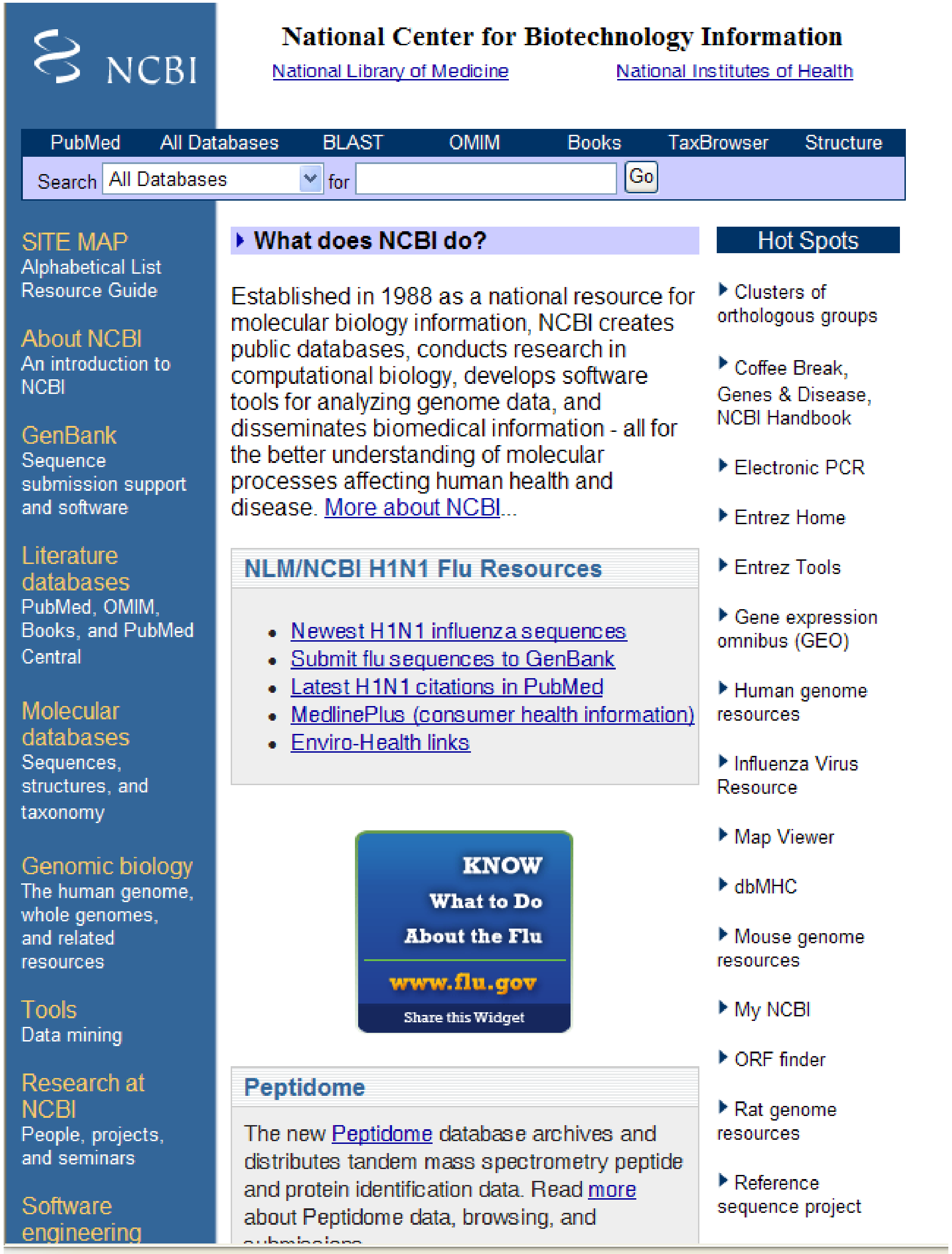
Recent years have seen an explosive growth in biological data. Large sequencing projects are producing increasing quantities of nucleotide sequences. The contents of nucleotide databases are doubling in size approximately every 14 months. The latest release of GenBank exceeded 165 billion base pairs. Not only the size of sequence data is rapidly increasing, but also the number of characterized genes from many organisms and protein structures doubles about every two years. To cope with this great quantity of data, a new scientific discipline has emerged: ***bioinformatics, biocomputing or computational biology***

ENTRIES	**BASES**	**SPECIES**
20614460	17575474103	*Homo sapiens*
9724856	9993232725	*Mus musculus*
2193460	6525559108	*Rattus norvegicus*
2203159	5391699711	*Bos taurus*
3967977	5079812801	*Zea mays*
3296476	4894315374	*Sus scrofa*
1727319	3128000237	*Danio rerio*
1796154	1925428081	*Triticum aestivum*
744380	1764995265	*Solanum lycopersicum*
1332169	1617554059	*Hordeum vulgare subsp. vulgare*
257614	1435261003	*Strongylocentrotus purpuratus*
456726	1297237624	*Macaca mulatta*
1376132	1265215013	*Oryza sativa Japonica Group*
1588338	1249788384	*Xenopus (Silurana) tropicalis*
1778369	1200025462	*Nicotiana tabacum*
2398266	1165816533	*Arabidopsis thaliana*
1267298	1155228906	*Drosophila melanogaster*
809463	1071458039	*Vitis vinifera*
2104483	1020646789	*Glycine max*
217105	1010316029	*Pan troglodytes*

Genetic Sequence Data Bank
Aug 15 2014,
Release 203.0

165,722,980,375 bases, from
174,108,750 reported sequence





NCBI

National Center for Biotechnology Information

National Library of Medicine   National Institutes of Health

PubMed | All Databases | BLAST | OMIM | Books | TaxBrowser | Structure

Search [All Databases] for [ ] Go

SITE MAP
Alphabetical List
Resource Guide

About NCBI
An introduction to NCBI

GenBank
Sequence submission support and software

Literature databases
PubMed, OMIM, Books, and PubMed Central

Molecular databases
Sequences, structures, and taxonomy

Genomic biology
The human genome, whole genomes, and related resources

Tools
Data mining

Research at NCBI
People, projects, and seminars

Software engineering

What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More about NCBI...

NLM/NCBI H1N1 Flu Resources

- Newest H1N1 influenza sequences
- Submit flu sequences to GenBank
- Latest H1N1 citations in PubMed
- MedlinePlus (consumer health information)
- Enviro-Health links

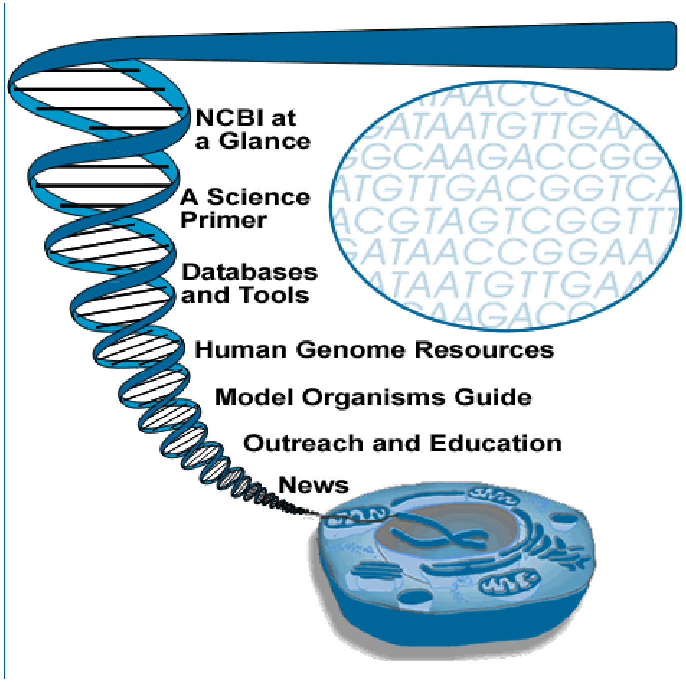
KNOW
What to Do
About the Flu
www.flu.gov
Share this Widget

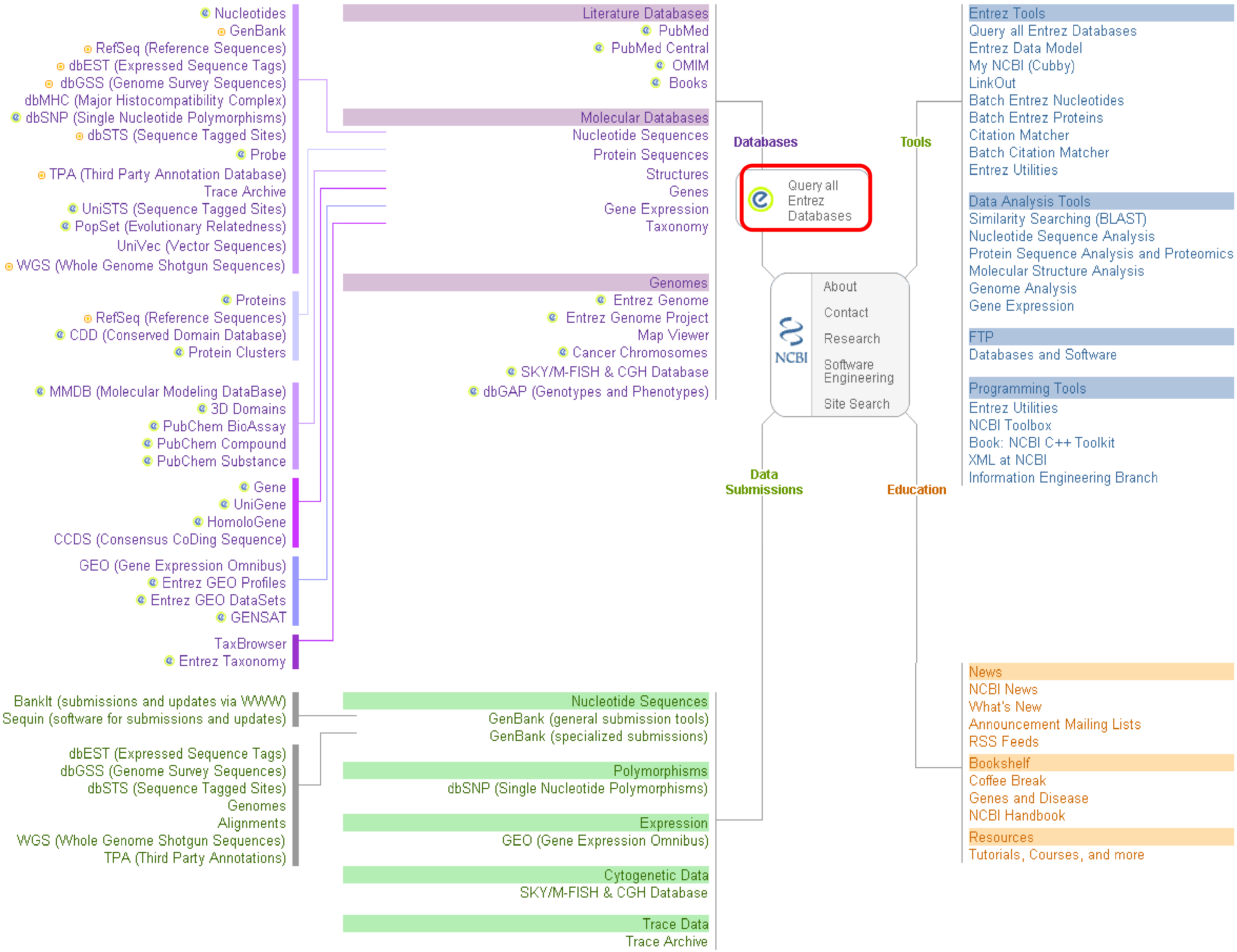
Peptidome

The new Peptidome database archives and distributes tandem mass spectrometry peptide and protein identification data. Read more about Peptidome data, browsing, and

Hot Spots

- Clusters of orthologous groups
- Coffee Break, Genes & Disease, NCBI Handbook
- Electronic PCR
- Entrez Home
- Entrez Tools
- Gene expression omnibus (GEO)
- Human genome resources
- Influenza Virus Resource
- Map Viewer
- dbMHC
- Mouse genome resources
- My NCBI
- ORF finder
- Rat genome resources
- Reference sequence project





Query all Entrez Databases

NCBI
About
Contact
Research
Software Engineering
Site Search

Databases

Literature Databases
PubMed
PubMed Central
OMIM
Books

Molecular Databases
Nucleotide Sequences
Protein Sequences
Structures
Genes
Gene Expression
Taxonomy

Nucleotides
GenBank
RefSeq (Reference Sequences)
dbEST (Expressed Sequence Tags)
dbGSS (Genome Survey Sequences)
dbMHC (Major Histocompatibility Complex)
dbSNP (Single Nucleotide Polymorphisms)
dbSTS (Sequence Tagged Sites)
Probe
TPA (Third Party Annotation Database)
Trace Archive
UniSTS (Sequence Tagged Sites)
PopSet (Evolutionary Relatedness)
UniVec (Vector Sequences)
WGS (Whole Genome Shotgun Sequences)

Proteins
RefSeq (Reference Sequences)
CDD (Conserved Domain Database)
Protein Clusters

MMDB (Molecular Modeling DataBase)
3D Domains
PubChem BioAssay
PubChem Compound
PubChem Substance

Gene
UniGene
HomoloGene
CCDS (Consensus CoDing Sequence)

GEO (Gene Expression Omnibus)
Entrez GEO Profiles
Entrez GEO DataSets
GENSAT

TaxBrowser
Entrez Taxonomy

Genomes
Entrez Genome
Entrez Genome Project
Map Viewer
Cancer Chromosomes
SKY/M-FISH & CGH Database
dbGAP (Genotypes and Phenotypes)

Data Submissions

Nucleotide Sequences
GenBank (general submission tools)
GenBank (specialized submissions)

BankIt (submissions and updates via WWW)
Sequin (software for submissions and updates)

dbEST (Expressed Sequence Tags)
dbGSS (Genome Survey Sequences)
dbSTS (Sequence Tagged Sites)
Genomes
Alignments
WGS (Whole Genome Shotgun Sequences)
TPA (Third Party Annotations)

Polymorphisms
dbSNP (Single Nucleotide Polymorphisms)

Expression
GEO (Gene Expression Omnibus)

Cytogenetic Data
SKY/M-FISH & CGH Database

Trace Data
Trace Archive

Tools

Entrez Tools
Query all Entrez Databases
Entrez Data Model
My NCBI (Cubby)
LinkOut
Batch Entrez Nucleotides
Batch Entrez Proteins
Citation Matcher
Batch Citation Matcher
Entrez Utilities

Data Analysis Tools
Similarity Searching (BLAST)
Nucleotide Sequence Analysis
Protein Sequence Analysis and Proteomics
Molecular Structure Analysis
Genome Analysis
Gene Expression

FTP
Databases and Software

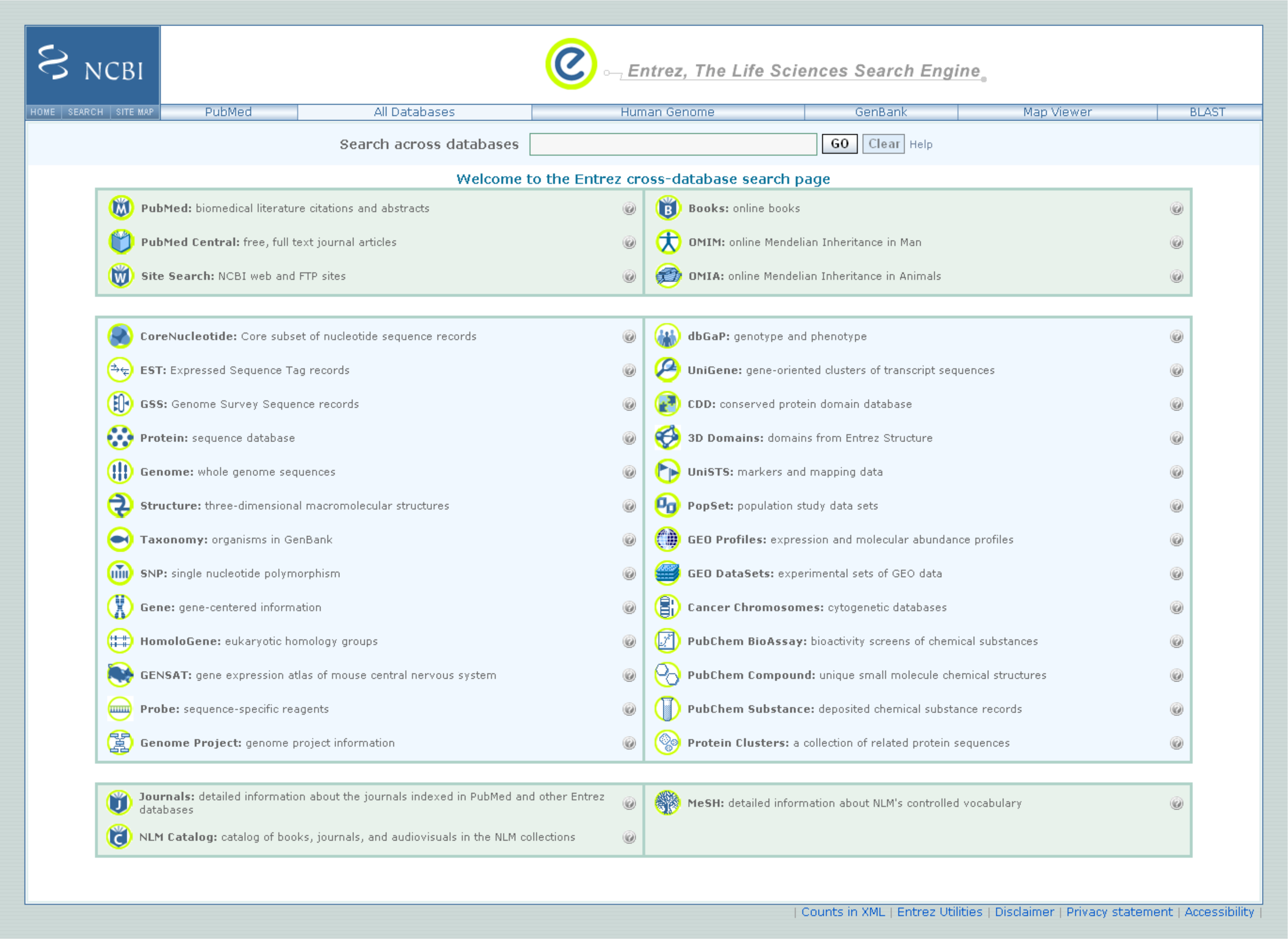
Programming Tools
Entrez Utilities
NCBI Toolbox
Book: NCBI C++ Toolkit
XML at NCBI
Information Engineering Branch

Education

News
NCBI News
What's New
Announcement Mailing Lists
RSS Feeds

Bookshelf
Coffee Break
Genes and Disease
NCBI Handbook

Resources
Tutorials, Courses, and more

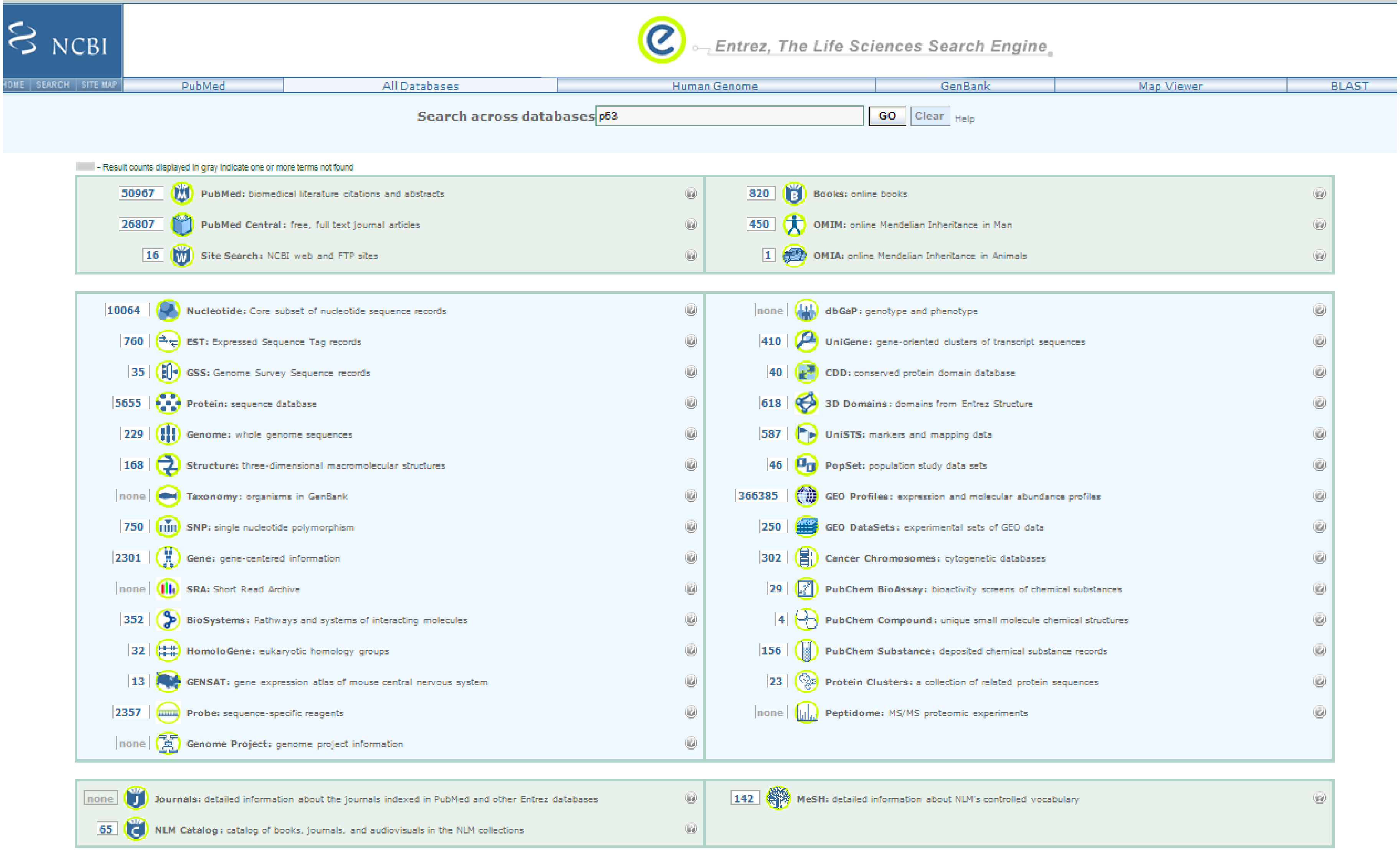


Welcome to the Entrez cross-database search page

- **PubMed:** biomedical literature citations and abstracts
- **PubMed Central:** free, full text journal articles
- **Site Search:** NCBI web and FTP sites
- **Books:** online books
- **OMIM:** online Mendelian Inheritance in Man
- **OMIA:** online Mendelian Inheritance in Animals

- **CoreNucleotide:** Core subset of nucleotide sequence records
- **EST:** Expressed Sequence Tag records
- **GSS:** Genome Survey Sequence records
- **Protein:** sequence database
- **Genome:** whole genome sequences
- **Structure:** three-dimensional macromolecular structures
- **Taxonomy:** organisms in GenBank
- **SNP:** single nucleotide polymorphism
- **Gene:** gene-centered information
- **HomoloGene:** eukaryotic homology groups
- **GENSAT:** gene expression atlas of mouse central nervous system
- **Probe:** sequence-specific reagents
- **Genome Project:** genome project information
- **dbGaP:** genotype and phenotype
- **UniGene:** gene-oriented clusters of transcript sequences
- **CDD:** conserved protein domain database
- **3D Domains:** domains from Entrez Structure
- **UniSTS:** markers and mapping data
- **PopSet:** population study data sets
- **GEO Profiles:** expression and molecular abundance profiles
- **GEO DataSets:** experimental sets of GEO data
- **Cancer Chromosomes:** cytogenetic databases
- **PubChem BioAssay:** bioactivity screens of chemical substances
- **PubChem Compound:** unique small molecule chemical structures
- **PubChem Substance:** deposited chemical substance records
- **Protein Clusters:** a collection of related protein sequences

- **Journals:** detailed information about the journals indexed in PubMed and other Entrez databases
- **NLM Catalog:** catalog of books, journals, and audiovisuals in the NLM collections
- **MeSH:** detailed information about NLM's controlled vocabulary



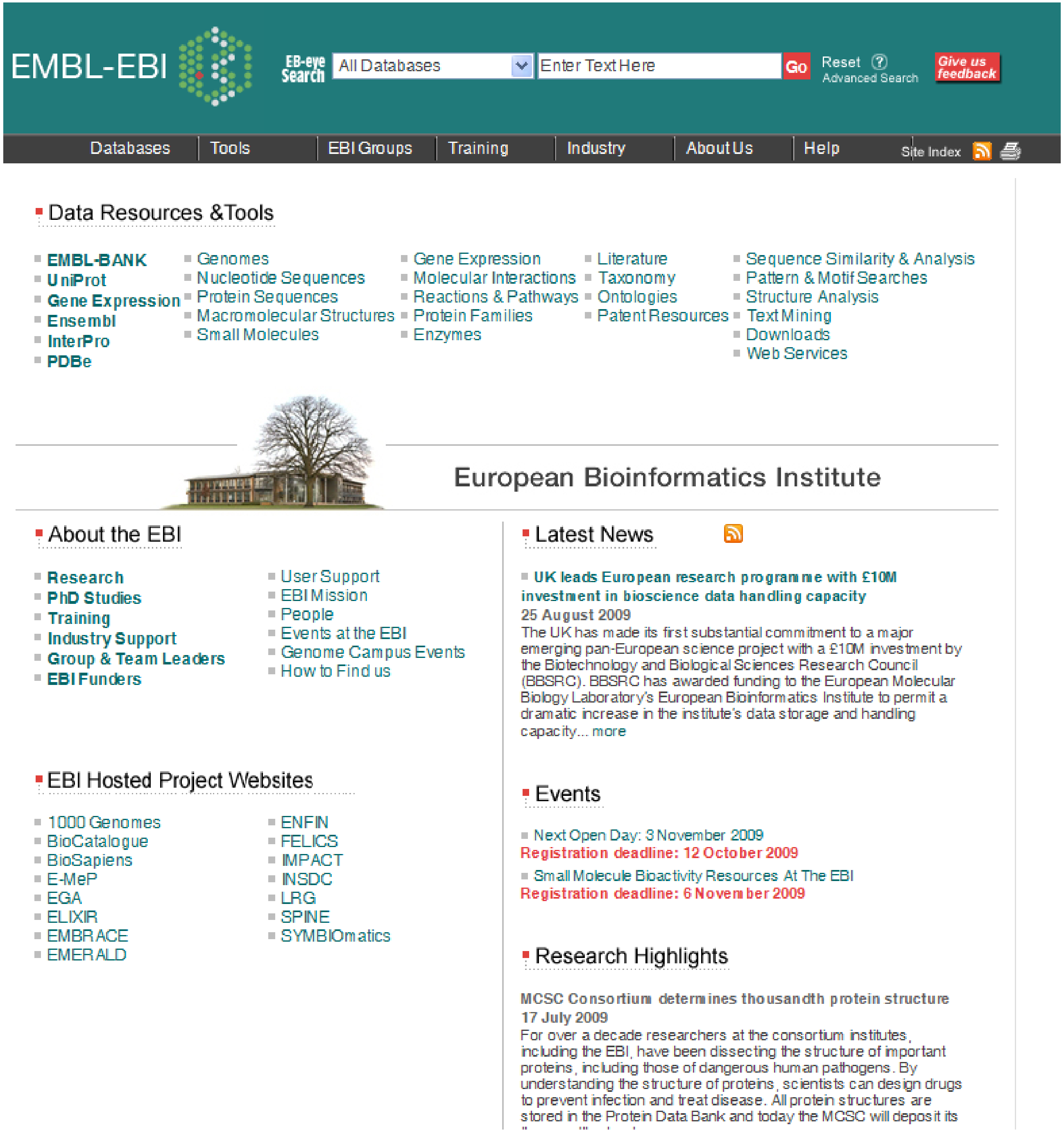
Search across databases p53 GO Clear Help

- Result counts displayed in gray indicate one or more terms not found

50967	**PubMed:** biomedical literature citations and abstracts	820	**Books:** online books
26807	**PubMed Central:** free, full text journal articles	450	**OMIM:** online Mendelian Inheritance in Man
16	**Site Search:** NCBI web and FTP sites	1	**OMIA:** online Mendelian Inheritance in Animals

10064	**Nucleotide:** Core subset of nucleotide sequence records	none	**dbGaP:** genotype and phenotype
760	**EST:** Expressed Sequence Tag records	410	**UniGene:** gene-oriented clusters of transcript sequences
35	**GSS:** Genome Survey Sequence records	40	**CDD:** conserved protein domain database
5655	**Protein:** sequence database	618	**3D Domains:** domains from Entrez Structure
229	**Genome:** whole genome sequences	587	**UniSTS:** markers and mapping data
168	**Structure:** three-dimensional macromolecular structures	46	**PopSet:** population study data sets
none	**Taxonomy:** organisms in GenBank	366385	**GEO Profiles:** expression and molecular abundance profiles
750	**SNP:** single nucleotide polymorphism	250	**GEO DataSets:** experimental sets of GEO data
2301	**Gene:** gene-centered information	302	**Cancer Chromosomes:** cytogenetic databases
none	**SRA:** Short Read Archive	29	**PubChem BioAssay:** bioactivity screens of chemical substances
352	**BioSystems:** Pathways and systems of interacting molecules	4	**PubChem Compound:** unique small molecule chemical structures
32	**HomoloGene:** eukaryotic homology groups	156	**PubChem Substance:** deposited chemical substance records
13	**GENSAT:** gene expression atlas of mouse central nervous system	23	**Protein Clusters:** a collection of related protein sequences
2357	**Probe:** sequence-specific reagents	none	**Peptidome:** MS/MS proteomic experiments
none	**Genome Project:** genome project information		

none	**Journals:** detailed information about the journals indexed in PubMed and other Entrez databases	142	**MeSH:** detailed information about NLM's controlled vocabulary
65	**NLM Catalog:** catalog of books, journals, and audiovisuals in the NLM collections		



Data Resources & Tools

- **EMBL-BANK**
- **UniProt**
- **Gene Expression**
- **Ensembl**
- **InterPro**
- **PDBe**

- Genomes
- Nucleotide Sequences
- Protein Sequences
- Macromolecular Structures
- Small Molecules

- Gene Expression
- Molecular Interactions
- Reactions & Pathways
- Protein Families
- Enzymes

- Literature
- Taxonomy
- Ontologies
- Patent Resources

- Sequence Similarity & Analysis
- Pattern & Motif Searches
- Structure Analysis
- Text Mining
- Downloads
- Web Services

European Bioinformatics Institute

About the EBI

- **Research**
- **PhD Studies**
- **Training**
- **Industry Support**
- **Group & Team Leaders**
- **EBI Funders**

- User Support
- EBI Mission
- People
- Events at the EBI
- Genome Campus Events
- How to Find us

EBI Hosted Project Websites

- 1000 Genomes
- BioCatalogue
- BioSapiens
- E-MeP
- EGA
- ELIXIR
- EMBRACE
- EMERALD

- ENFIN
- FELICS
- IMPACT
- INSDC
- LRG
- SPINE
- SYMBIOmatics

Latest News

- **UK leads European research programme with £10M investment in bioscience data handling capacity**

25 August 2009

The UK has made its first substantial commitment to a major emerging pan-European science project with a £10M investment by the Biotechnology and Biological Sciences Research Council (BBSRC). BBSRC has awarded funding to the European Molecular Biology Laboratory's European Bioinformatics Institute to permit a dramatic increase in the institute's data storage and handling capacity... more

Events

- Next Open Day: 3 November 2009

Registration deadline: 12 October 2009

- Small Molecule Bioactivity Resources At The EBI

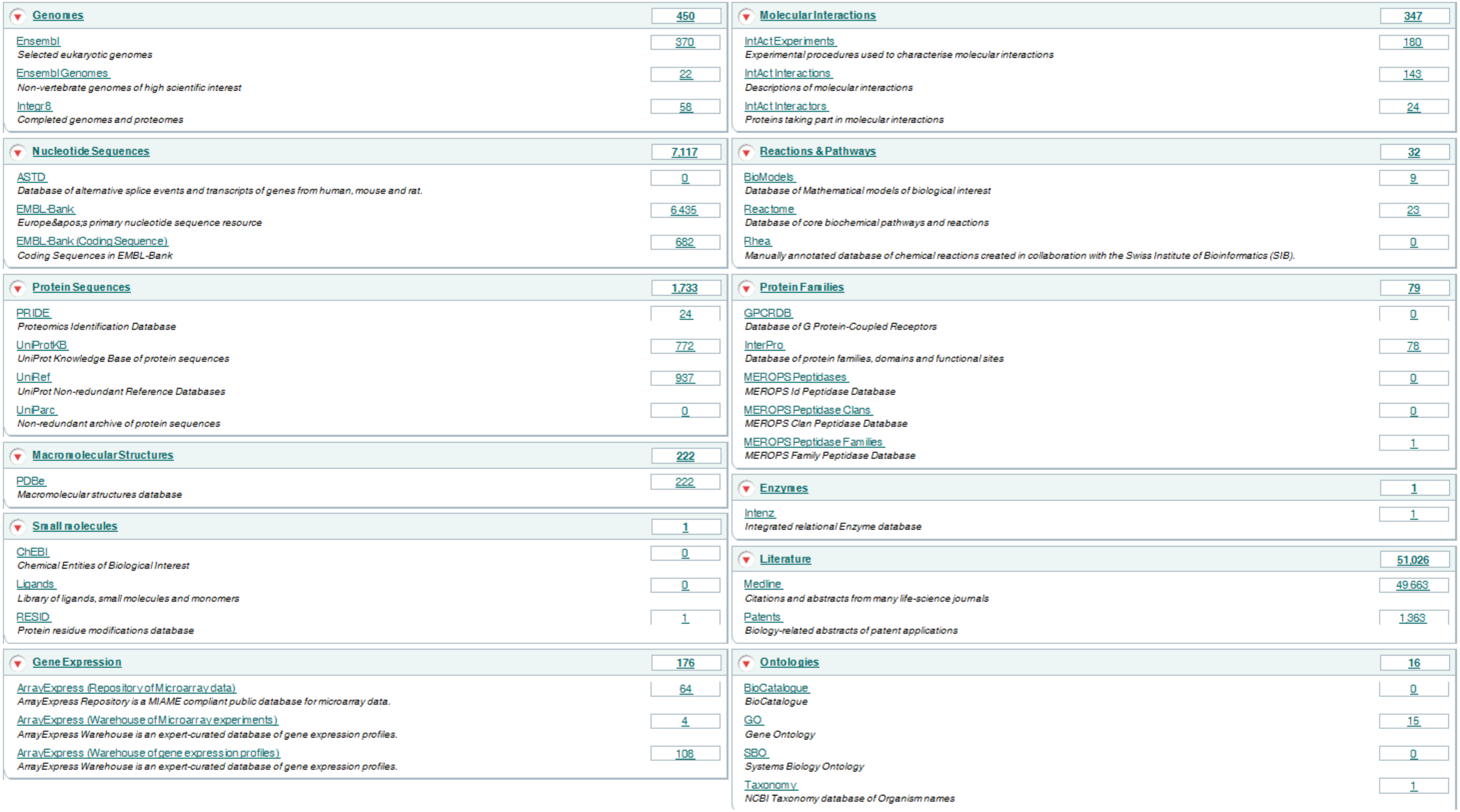
Registration deadline: 6 November 2009

Research Highlights

MCSC Consortium determines thousandth protein structure

17 July 2009

For over a decade researchers at the consortium institutes, including the EBI, have been dissecting the structure of important proteins, including those of dangerous human pathogens. By understanding the structure of proteins, scientists can design drugs to prevent infection and treat disease. All protein structures are stored in the Protein Data Bank and today the MCSC will deposit its



Genomes		450
Ensembl	Selected eukaryotic genomes	370
Ensembl Genomes	Non-vertebrate genomes of high scientific interest	22
Integr8	Completed genomes and proteomes	58

Nucleotide Sequences		7,117
ASTD	Database of alternative splice events and transcripts of genes from human, mouse and rat.	0
EMBL-Bank	Europe's primary nucleotide sequence resource	6,435
EMBL-Bank (Coding Sequence)	Coding Sequences in EMBL-Bank	682

Protein Sequences		1,733
PRIDE	Proteomics Identification Database	24
UniProtKB	UniProt Knowledge Base of protein sequences	772
UniRef	UniProt Non-redundant Reference Databases	937
UniParc	Non-redundant archive of protein sequences	0

Macromolecular Structures		222
PDBe	Macromolecular structures database	222

Small molecules		1
ChEBI	Chemical Entities of Biological Interest	0
Ligands	Library of ligands, small molecules and monomers	0
RESID	Protein residue modifications database	1

Gene Expression		176
ArrayExpress (Repository of Microarray data)	ArrayExpress Repository is a MIAME compliant public database for microarray data.	64
ArrayExpress (Warehouse of Microarray experiments)	ArrayExpress Warehouse is an expert-curated database of gene expression profiles.	4
ArrayExpress (Warehouse of gene expression profiles)	ArrayExpress Warehouse is an expert-curated database of gene expression profiles.	108

Molecular Interactions		347
IntAct Experiments	Experimental procedures used to characterise molecular interactions	180
IntAct Interactions	Descriptions of molecular interactions	143
IntAct Interactors	Proteins taking part in molecular interactions	24

Reactions & Pathways		32
BioModels	Database of Mathematical models of biological interest	9
Reactome	Database of core biochemical pathways and reactions	23
Rhea	Manually annotated database of chemical reactions created in collaboration with the Swiss Institute of Bioinformatics (SIB).	0

Protein Families		79
GPCRDB	Database of G Protein-Coupled Receptors	0
InterPro	Database of protein families, domains and functional sites	78
MEROPS Peptidases	MEROPS Id Peptidase Database	0
MEROPS Peptidase Clans	MEROPS Clan Peptidase Database	0
MEROPS Peptidase Families	MEROPS Family Peptidase Database	1

Enzymes		1
Intenz	Integrated relational Enzyme database	1

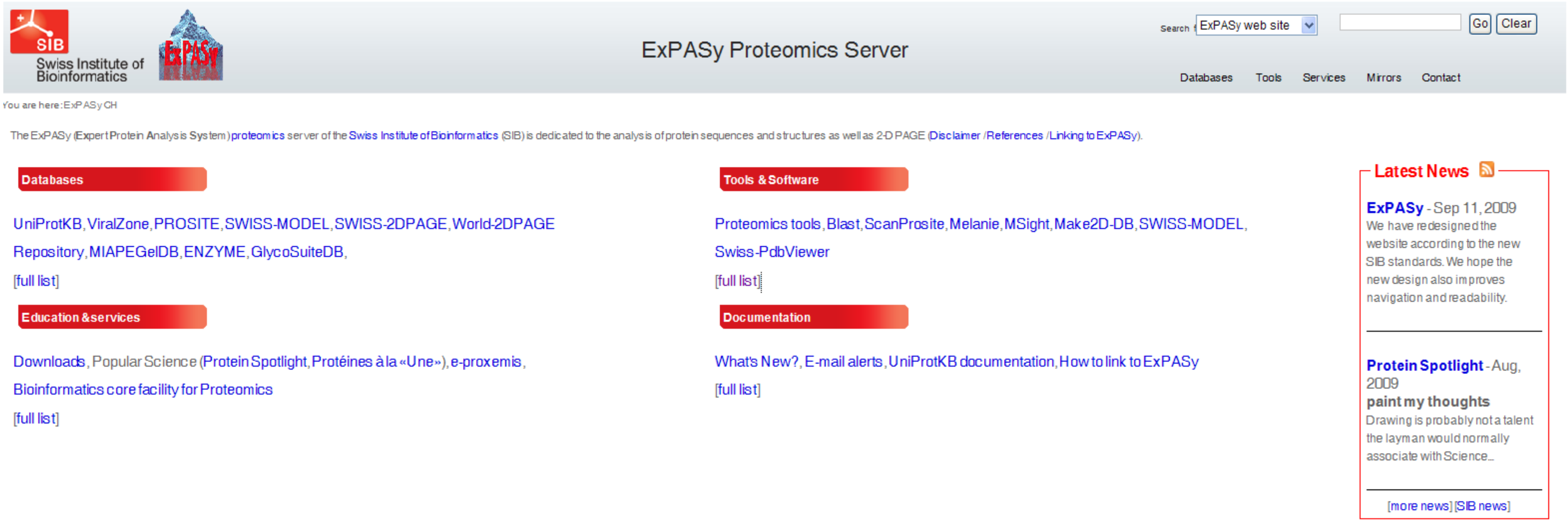
Literature		51,026
Medline	Citations and abstracts from many life-science journals	49,663
Patents	Biology-related abstracts of patent applications	1,363

Ontologies		16
BioCatalogue	BioCatalogue	0
GO	Gene Ontology	15
SBO	Systems Biology Ontology	0
Taxonomy	NCBI Taxonomy database of Organism names	1

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



ExPASy Proteomics Server

You are here: ExPASy CH

The ExPASy (Expert Protein Analysis System) 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 (Disclaimer / References / Linking to ExPASy).

Databases

UniProtKB, ViralZone, PROSITE, SWISS-MODEL, SWISS-2DPAGE, World-2DPAGE Repository, MIAPEGelDB, ENZYME, GlycoSuiteDB,

[full list]

Education & services

Downloads, Popular Science (Protein Spotlight, Protéines à la «Une»), e-proxemis, Bioinformatics core facility for Proteomics

[full list]

Tools & Software

Proteomics tools, Blast, ScanProsite, Melanie, MSight, Make2D-DB, SWISS-MODEL, Swiss-PdbViewer

[full list]

Documentation

What's New?, E-mail alerts, UniProtKB documentation, How to link to ExPASy

[full list]

Latest News

ExPASy - Sep 11, 2009
We have redesigned the website according to the new SIB standards. We hope the new design also improves navigation and readability.

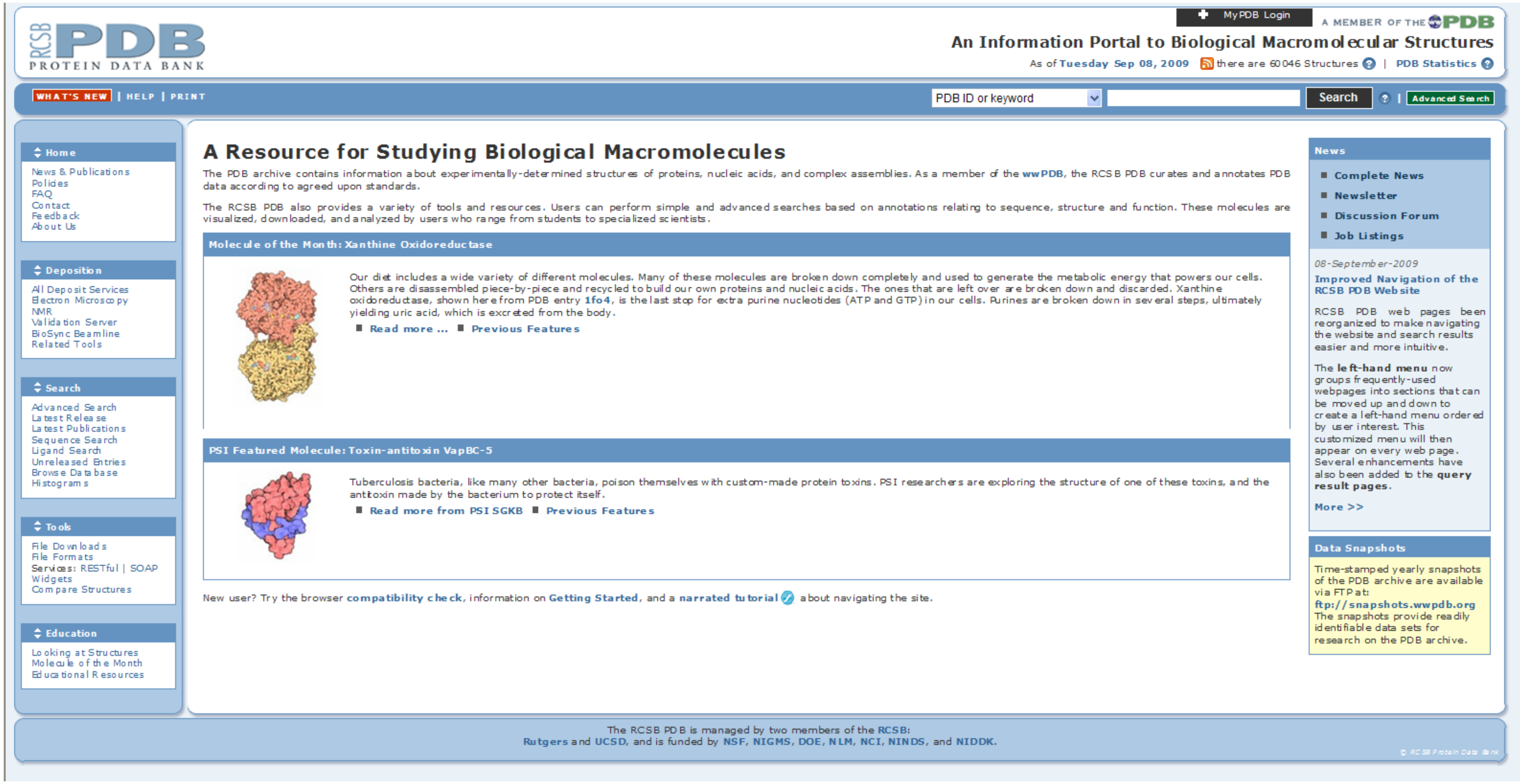
Protein Spotlight - Aug, 2009
paint my thoughts
Drawing is probably not a talent the layman would normally associate with Science...

[more news] [SIB news]

Protein Databases

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.



An Information Portal to Biological Macromolecular Structures

As of **Tuesday Sep 08, 2009** there are 60046 Structures | **PDB Statistics**

A Resource for Studying Biological Macromolecules

The PDB archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the **wwPDB**, the RCSB PDB curates and annotates PDB data according to agreed upon standards.

The RCSB PDB also provides a variety of tools and resources. Users can perform simple and advanced searches based on annotations relating to sequence, structure and function. These molecules are visualized, downloaded, and analyzed by users who range from students to specialized scientists.

Molecule of the Month: Xanthine Oxidoreductase

Our diet includes a wide variety of different molecules. Many of these molecules are broken down completely and used to generate the metabolic energy that powers our cells. Others are disassembled piece-by-piece and recycled to build our own proteins and nucleic acids. The ones that are left over are broken down and discarded. Xanthine oxidoreductase, shown here from PDB entry **1fo4**, is the last stop for extra purine nucleotides (ATP and GTP) in our cells. Purines are broken down in several steps, ultimately yielding uric acid, which is excreted from the body.

■ **Read more ...** ■ **Previous Features**

PSI Featured Molecule: Toxin-antitoxin VapBC-5

Tuberculosis bacteria, like many other bacteria, poison themselves with custom-made protein toxins. PSI researchers are exploring the structure of one of these toxins, and the antitoxin made by the bacterium to protect itself.

■ **Read more from PSI SGKB** ■ **Previous Features**

New user? Try the browser **compatibility check**, information on **Getting Started**, and a **narrated tutorial** about navigating the site.

08-September-2009

Improved Navigation of the RCSB PDB Website

RCSB PDB web pages been reorganized to make navigating the website and search results easier and more intuitive.

The **left-hand menu** now groups frequently-used webpages into sections that can be moved up and down to create a left-hand menu ordered by user interest. This customized menu will then appear on every web page. Several enhancements have also been added to the **query result pages**.

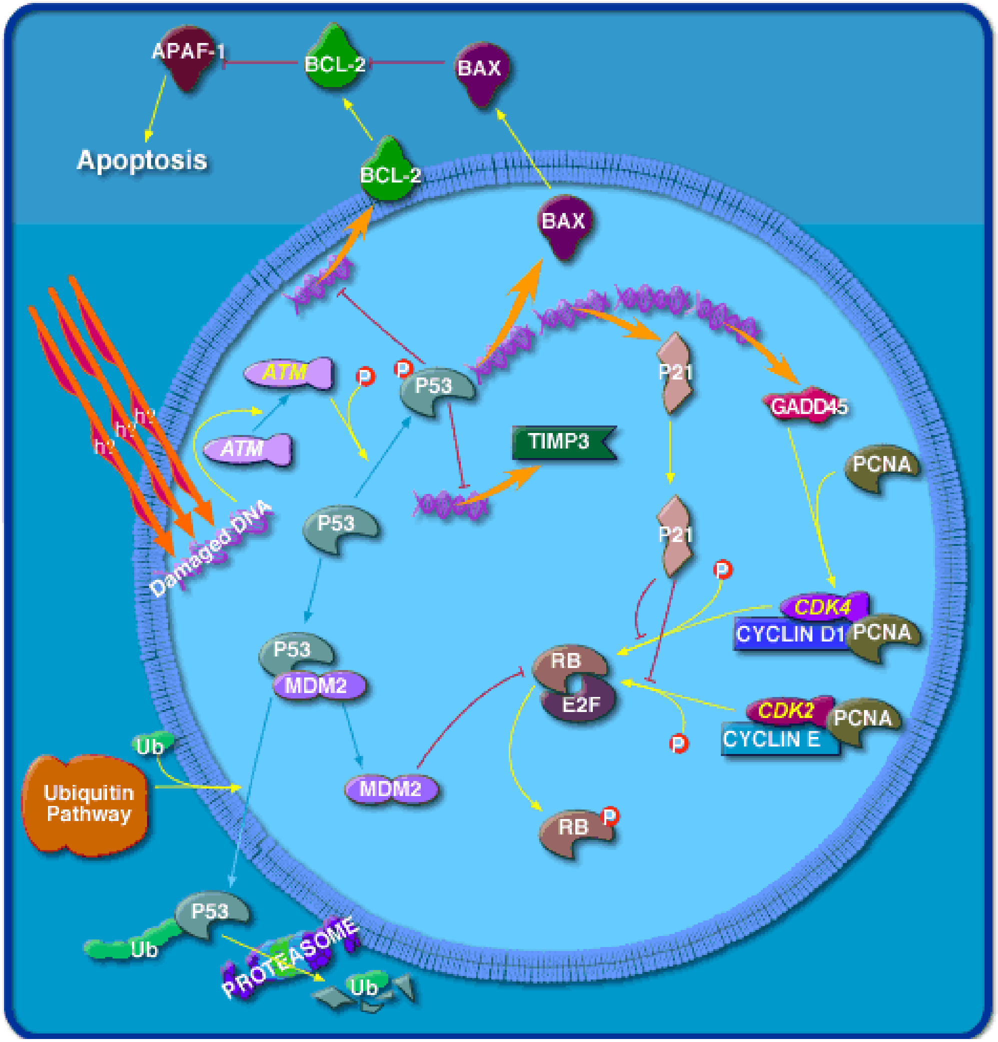
More >>

Data Snapshots

Time-stamped yearly snapshots of the PDB archive are available via FTP at: **ftp://snapshots.wwpdb.org** The snapshots provide readily identifiable data sets for research on the PDB archive.

The RCSB PDB is managed by two members of the **RCSB**:
Rutgers and **UCSD**, and is funded by **NSF**, **NIGMS**, **DOE**, **NLM**, **NCI**, **NINDS**, and **NIDDK**.

Metabolic & Signalling Pathways

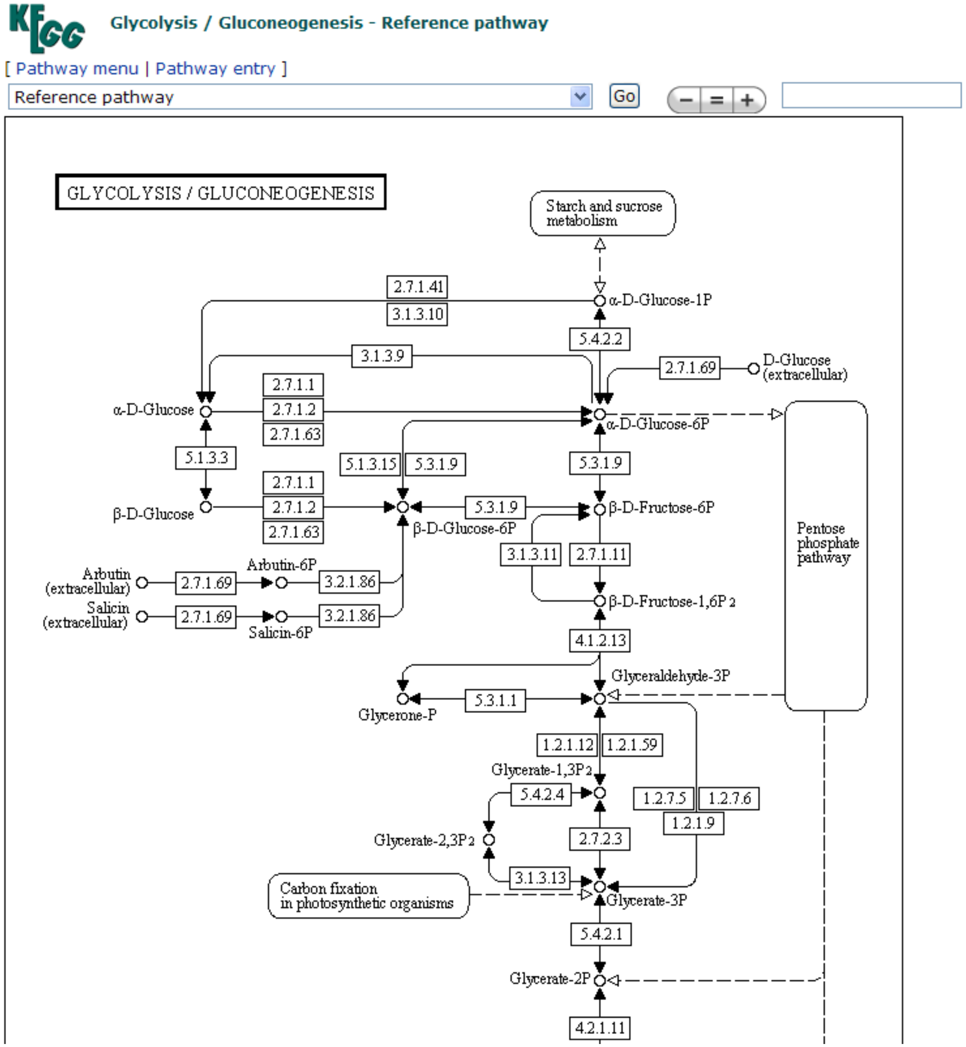


Biocarta
(*http://biocarta.com*)

Metabolic & Signalling Pathways

Kyoto Encyclopedia of Genes &Genomes

http://www.genome.ad.jp/kegg/



THE COMPONENTS OF BIOINFORMATICS

TECHNOLOGY

ALGORITHM

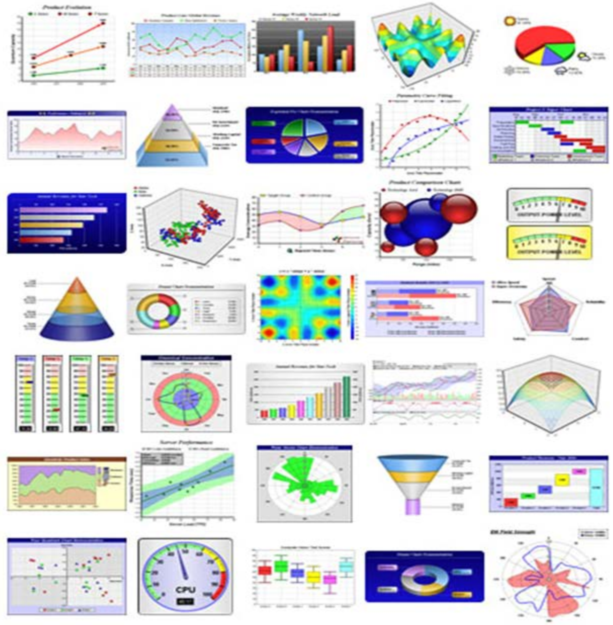
ANALYSIS TOOLS

DATABASE

COMPUTING POWER

BIOINFORMATICS ANALYSIS TOOLS

CentOS
SOAP
gatk
BowTIE
Galaxy
BEDTOOLS
circos
R
SAMtools
VarScan
Bioconductor
Integrative Genomics Viewer
COSMIC Catalogue of somatic mutations in cancer
swissprot
Pfam
KEGG
miRBase
WormBase
MGI
ZFIN
UniProt
Cytoscape
FlyBase
EuPathDB
WebLogo
PDB PROTEIN DATA BANK
Mascot Insight
1000 Genomes A Deep Catalog of Human Genetic Variation
x!



CLC Genomics
Vector NTI suite
DNA Star

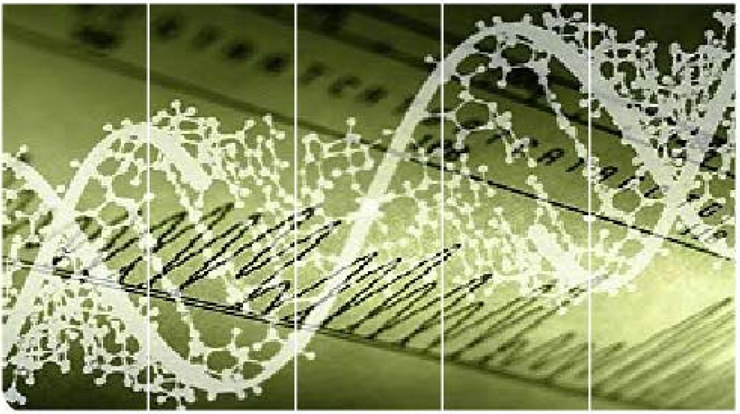
Staden Package
EMBOSS
BLAST
SMS….

On line analysis tools

GMBD Bioinformatics Core

Bioinformatics Core for Genomic Medicine and Biotechnology Development (GMBD Bioinformatics Core) is a consortium formed by joint efforts of bioinformatics research groups in Taiwan, aiming at becoming bioinformatics infrastructure for our research community.

Welcome and enjoy your visit!



Unit 1 Unit 2 Unit 3 Unit 4 Unit 5

Coordination Unit

Since 1997, Department of Research Resources at NHRI has devoted considerable amount of resources in providing bioinformatics services in Taiwan. Coordination Unit will serve as the administration, education and service center of GMBD Bioinformatics Core.

基因體醫學生技研發生物資訊核心

Bioinformatics Core for Genomic Medicine and Biotechnology Development

❖ Databases

CPDB	Circular Permutation Database [NTHU]
dbPTM	Protein Post-Translational Modification Database [NCTU]
FlyDPI	Drosophila Database of Protein Interactomes [NHRI]
GPDB	Genome Profile Database [NTHU]
hp-DPI	H. pylori Database of protein interactomes [NHRI]
HPSG	Structural Genomics Databases of Helicobacter pylori [NTHU]
iGenome	A database index for the genome summarized information [NCTU]
KPSG	Structural Genomics Databases of Klebsiella pneumoniae [NTHU]
miRNAMap	Genomic Maps for microRNA [NCTU]
RegRNA	Regulatory RNA Motifs and Elements Database [NCTU]
RepeatMasker	Screening for low complexity DNA sequences and interspersed repeats [mirror at NCKU]
SM	Structural Genomics Databases of Stenotrophomonas maltophili [NTHU]
SNP	SNP value-added database [NCKU]
SSDB	Disulfide Proteins Database [NCTU]
TAG	Tumor Associated Gene Database [NCKU]
The BEST	The Binding Element Searching Tool [NCKU]
TPMD	Taiwan Polymorphism Marker Database [NHRI]
XCC	Structural Genomics Databases of Xanthomonas campestris [NTHU]

❖ AnalysisTools Online

- Homology & Similarity

BLAST	BLAST Search at NCKU
PDB-BLAST	BLAST against PDB protein databank at NTHU
RPS-BLAST	BLAST against Motif/Domain database at NTHU
MyBLAST	BLAST against customized database (the databases built by yourself) [NHRI, Sinica]

- Sequence Analysis

Chameleon Sequence Finder	Retrive a small peptide fragment sequence from a collection of PDB sequences [NCTU]
ClustalW	Web-based Multiple Sequence Alignment [NTHU]
ENACE System	Identification and evolutionary analysis of novel exons and alternative splicing events [Sinica, NHRI]
ESTviewer	A Web interface for visualizing ESTs [Sinica, NHRI]
GenePredict	Web-based gene prediction service [NTHU]
GPRM	Genetic Programming for RNA Motifs [NCTU]
KinasePhos	Predict phosphorylation sites within given protein sequences [NCTU]
MuSiC	Multiple Sequence Alignment with Constraints [NCTU]
PDA	Primer Design Assistant [NHRI]
PSEP System	A Comparative Method for Identification of Gene Structures and Alternatively Spliced Variants [Sinica, NHRI]
RNAMST	RNA Motif Search Tool [NCTU]
SpliceInfo	An information repository for mRNA alternative splicing in human genome [NCTU]
SpliceMotif	Identify conserved sequence elements associated to mRNA splicing [NCTU]
UPS	Unique Probe Selector [Sinica, NHRI]

- Phylogenetic Analysis

Palm	Phylogenetic reconstruction by Automatic Likelihood Model selector [Sinica, NHRI]
POWER	Phylogenetic Web Repeater [NHRI]

- Topology Prediction

CELLO	Prediction of protein subcellular localization [NCTU]
pKNOT	pKNOT: the protein KNOT web server [NCTU]

- Structural Analysis

(PS)2	Protein Structure Prediction Server [NCTU]
3D-BLAST	Protein structure search [NCTU]
CPSARST	Circular Permutation Search Aided by Ramachandran Sequential Transformation [NTHU]
GEMDOCK	Generic Evolutionary Method for molecular DOCKing [NCTU]
PredictProtein	Sequence analysis and structure prediction [mirror at NTHU]
SARST	Structure Alignment by Ramachandran Search Tool [NTHU]
SDSE	Sequence Derived Structure Entropy [NCTU]
StEQ	Structural Entropy Query [NCTU]
Tm Predictor	Melting Temperature Prediction : Predict thermal stability of proteins [NTHU]

- Miscellaneous Tools

EMBOSS	European Molecular Biology Open Software Suite [NHRI]
EMBOSS GUI	EMBOSS Graphical User Interface: EMBOSS explorer [NHRI]
GCG	Wisconsin Sequence Analysis Package [NHRI]
Hubba-Hubba	Hub OBjects Analyser, a web-based service designed to explore the essential nodes in a network [Sinica, NHRI]
JEMBOSS	Java user interface of EMBOSS [NHRI]
KPST	KEGG Pathway Search Tool [NTHU]
RMA	Reinforced Merging Algorithms [NTHU]
SeqWeb	Web-based GCG [NHRI]

http://www.tbi.org.tw/tools/application01.htm

Home/ Tools / GCG Application

GCG, EMBOSS 序列分析工具帳號申請

- **帳號申請**

The Wisconsin Package (GCG/SeqWeb) 爲商業軟體，經 Accelrys公司授權予國家衛生研究院，提供給非營利之醫療衛生與研究單位之實驗室主持人 (principal investigator, PI) 和博士後研究員 (Postdoc)申請使用。

非實驗室主持人申請時，需請實驗室主持人簽名 (或蓋職章) 作爲確認。

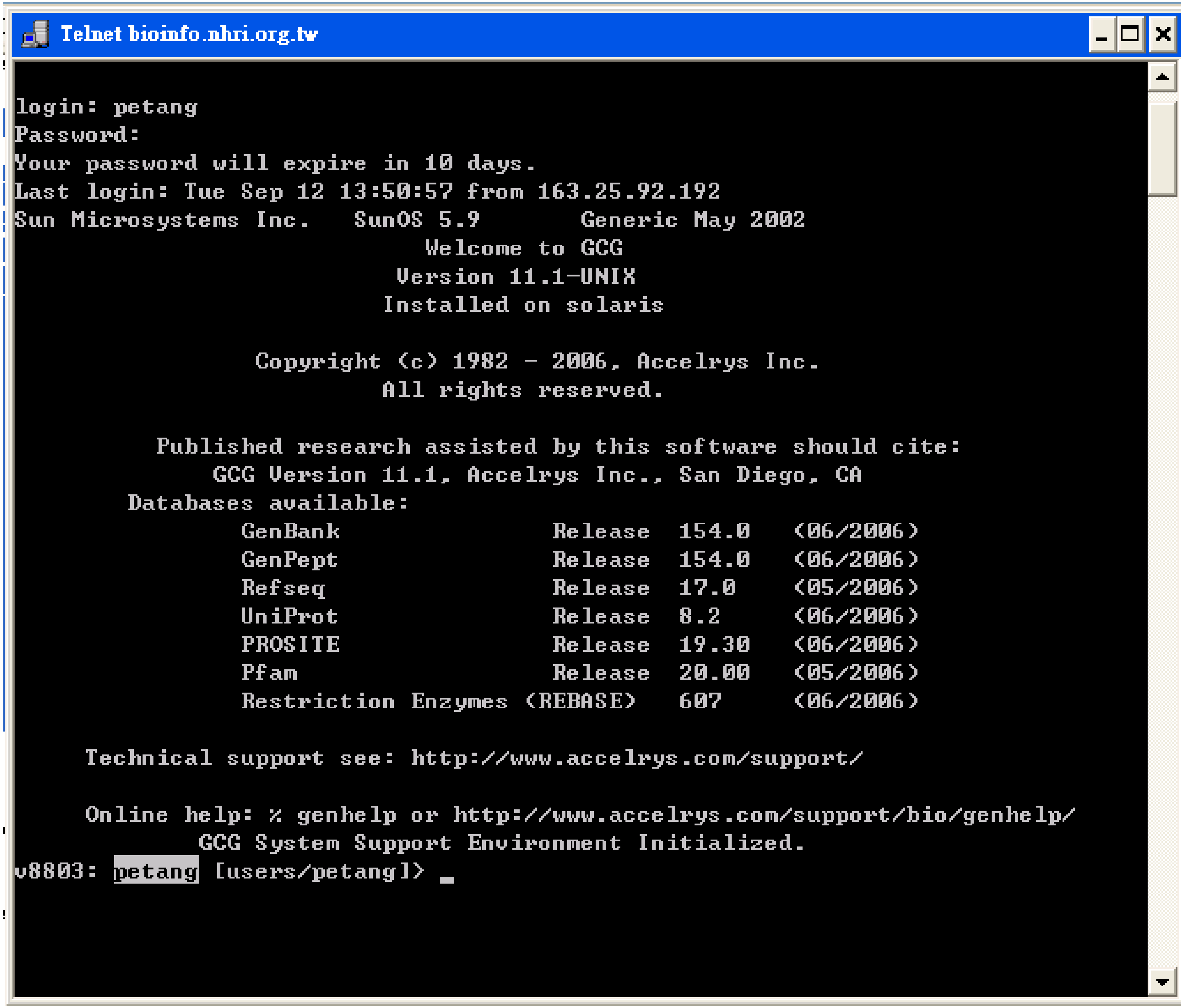
※ 本院恕不提供帳號申請給營利性質之單位與公司。

本院提供申請表格式檔案，請自行下載使用。申請表內容請務必填寫正確，Email 欄位僅接受所屬機構之信箱。

填妥後，請以傳真方式：(037) 586410 或郵寄方式至：**350 苗栗縣竹南鎮頂埔里科研路 35 號 生醫醫資料庫辦公室** 收，我們將儘速爲您處理！

下載申請表檔案

bioinfo-form.doc
(size: 49KB) **Word 檔案格式**



```
login: petang
Password:
Your password will expire in 10 days.
Last login: Tue Sep 12 13:50:57 from 163.25.92.192
Sun Microsystems Inc.   SunOS 5.9       Generic May 2002
                                     Welcome to GCG
                                   Version 11.1-UNIX
                                  Installed on solaris

                         Copyright (c) 1982 - 2006, Accelrys Inc.
                                  All rights reserved.

          Published research assisted by this software should cite:
              GCG Version 11.1, Accelrys Inc., San Diego, CA
        Databases available:
                GenBank                 Release  154.0   (06/2006)
                GenPept                 Release  154.0   (06/2006)
                Refseq                  Release  17.0    (05/2006)
                UniProt                 Release  8.2     (06/2006)
                PROSITE                 Release  19.30   (06/2006)
                Pfam                    Release  20.00   (05/2006)
                Restriction Enzymes (REBASE)     607     (06/2006)

     Technical support see: http://www.accelrys.com/support/

     Online help: % genhelp or http://www.accelrys.com/support/bio/genhelp/
             GCG System Support Environment Initialized.
v8803: petang [users/petang]> _
```

About • Applications • GUIs • Servers • Downloads • Licence • User docs • Developer docs • Administrator docs • Get involved • Support • Meetings • News • Credits

EMBOSS is funded from May 2009 by BBSRC grant BBR/G02264X/1

Funded from May 2006 to April 2009 by BBSRC grant BB/D018358/1

About EMBOSS Overview • Uses • FAQ Citing EMBOSS

A high-quality package of free, Open Source software for molecular biology ... *more >*

Applications EMBOSS • EMBASSY • Groups Proposed

Hundreds of useful, well documented applications for molecular sequence and other analyses ... *more >*

GUIs Jemboss • GUIs • Web • Others

We support the Jemboss GUI but many others are available... *more >*

Servers Portals • Servers • Mirrors • Misc

Many EMBOSS portals, servers and mirrors are available ... *more >*

Downloads Stable release • Developers (CVS) version • Getting started

EMBOSS is open source software and is freely available to all ... *more >*

Licence Licensing terms

EMBOSS uses the General Public Licence (GPL) and Library GPL ... *more >*

User documents FAQ • Tutorial • Running applications • Themes • Citing

Help for biologists, EMBOSS tutorial, guides to major bioinformatics themes ... *more >*

生物資訊分析

上課時間：每週五上午 08:10~12:00 （08:10~09:00—授課，09:10~12:00—實習）
上課地點：醫學院 9F 生物資訊中心 電腦教室

助教：方怡凱(分機5690) ，鄭尉弘(分機5690)
生物資訊中心：黃栢榕 老師 (分機3166)

進度	日期	課程進度	授課老師	講義下載
1	0926	Introduction to Bioinformatics	鄧致剛	PDF RAR Ex
2	1003	Bio-computing and Linux Basics	黃柏榕	PDF RAR Ex Unix cmd
3	1010	HOLIDAY		
4	1017	Bioinformatics Databases: searching & Management	鄧致剛	PDF RAR Ex
5	1024	Comparing Sequences and Multiple Sequence Alignment	鄧致剛	PDF RAR Ex
6	1031	DNA Sequence Analysis & Assembling	鄧致剛	PDF RAR Ex1 Ex2
7	1107	Phylogenetic Analysis	辛致煒	Phylip CLC Gen
8	1114	DNA Sequence Analysis (II)	林文昌	PPT RAR Ex
9	1121	Special Topic-Bioinformatics on the Internet	黃柏榕	PDF RAR Ex
10	1128	Protein Sequence Analysis	鄧致剛	PDF Ex ASS03
11	1205	Protein 3D Structure Modeling	呂平江	Rasmol2721 教學網
12	1212	Special Topic-Gene Chip Technology & Microarray Analysis	辛致煒	PDF MeV4 Dataset
13	1219	Special Topic-Integration of Omics (1) Metacore	鄧致剛	PDF RAR Ex
14	1226	Special Topic-Integration of Omics (2) David	鄧致剛	PDF RAR Ex
15	0102	Special Topic-Next Generation Sequencing Technology	劉軒	PDF RAR Ex
16	0109	Special Topic-microRNA analysis and databases	黃柏榕	PDF
17	0116	Student Project Presentation	鄧致剛	PDF RAR Ex
18	0123	Final Examination	鄧致剛	

THE COMPONENTS OF BIOINFORMATICS

TECHNOLOGY

ALGORITHM

ANALYSIS TOOLS

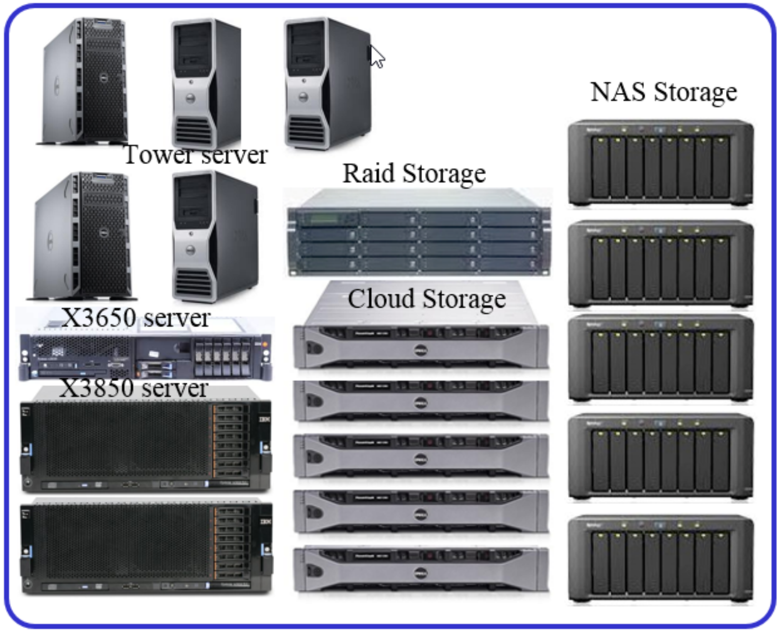
DATABASE

COMPUTING POWER

Chang Gung Bioinformatics Core

Server CPU/MEM: 436 Cores/3.04 TB
Workstation GPU/MEM: 12 Cores/192 GB
Workstation CPU/MEM: 66 Cores/ 512 GB
Storage: 736 TB

NGS Core

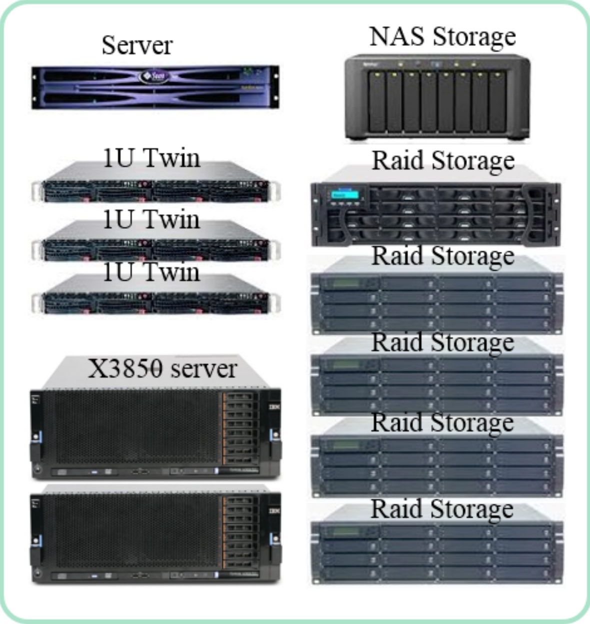


Tower Server	**5**
Server	**3**
Raid Storage	**1**
Cloud Storage	**5**
NAS Storage	**5**
Total	**19**

Software

Life Scope, Torrent Suite, Partek and Galaxy

Bioinformatics Core



Server	**13**
Raid Storage	**5**
NAS Storage	**1**
Total	**19**

Software

VectorNTI
GeneGo (MetaCore)
CLC Genomics Workbench
CPAP, DSAP, FASTAnnotator & VIPDB
VANNO, HMPD, CRSAP

Proteomics Core

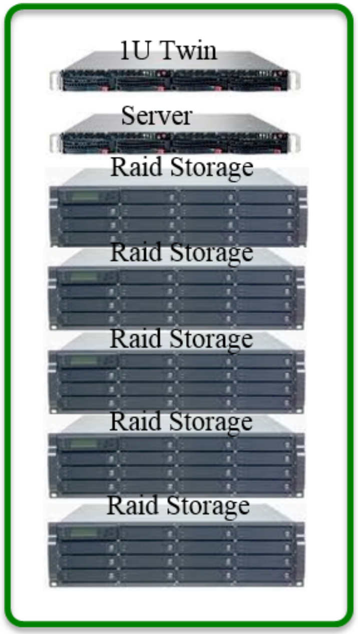


Image Center

Server	**3**
Raid	**5**
Server	**2**
Total	**10**

Small Genomes & Transcriptomes

送LED燈+1TB雲端硬碟

華碩 4代i7四核Win8.1電玩旗艦機

最新第4代Core i處理器

- 處理器：Intel Core i7-4790(3.4GHz)
- 晶片：Intel H97
- 記憶體：16GB DDR3 (8G*2) → **64GB**
- 硬碟：SATAIII 2TB+8G SSD(Hybrid) → **4TB**
- 顯示介面：NVIDIA GTX750 1G GDDR5 獨立顯卡
- 光碟機：超薄Super Multi DVD燒錄機
- 作業系統：Windows 8.1 (64bit)
- 其他：USB3.0、HDMI
- 保固：三年保固(含零件、人工/三年到府收送)

An Example

Steps to Identify a Gene

- Gene-Search
- Protein-Search
- Annotation

Full length ORF of *TvEST-14G2*

…AGATG

TGAAATAT ATTTTTTATA TTATTTATTT
TTTTCTTTTT CTATCTGTAT ATTAAATGTA TTTCTATATT ATTAAAAAAA

Amino Acid Sequence Comparison

01B1
04E12
14G2
PFCK
Yeast
Human
Mouse
TcCK1.1
TcCK1.2

01B1
04E12
14G2
PFCK
Yeast
Human
Mouse
TcCK1.1
TcCK1.2

01B1
04E12
14G2
PFCK
Yeast
Human
Mouse
TcCK1.1
TcCK1.2

01B1
04E12
14G2
PFCK
Yeast
Human
Mouse
TcCK1.1
TcCK1.2

🟩 **: kinesin homology domain**

🟥 **: casein kinase 1 specific motifs**

PFCK : Plasmodium casein kinase 1
TcCK1.1: Trypansoma cruzi casein kinase 1.1
TcCK1.2: Trypansoma cruzi casein kinase 1.2

Similarity of Various CK1s from Different Species

	TvEST-04E12	TvEST-14G2	TvEST-01B1	*T. cruzi* CK1.1	*T. cruzi* CK1.2	PFCK	Yeast CK1	Mouse CK1	Human CK1
TvEST-04E12	100	32	32	34	34	34	37	37	37
TvEST-14G2		100	24	24	23	24	24	26	25
TvEST-01B1			100	47	47	48	48	38	38
T. cruzi CK1.1				100	23	73	24	61	61
T. cruzi CK1.2					100	74	70	63	63
PFCK						100	69	62	62
Yeast CK1							100	69	67
Mouse CK1								100	99
Human CK1									100

3-D Structure of TvEST-14G2 and other CK1s

```
  1   MRKIYGNYIT QKRLGSGSFG EVWEAVSHST GQKVALKLEP RNSSVPQLFF
 51   EAKLYSMFQA SKSTNNSVEP CNNIPVVYAT GQTETTNYMA MELLGKSLED
101   LVSSVPRFSQ KTILMLAGQM ISCVEFVHKH NFIHRDIKPD NFAMGVSENS
151   NKIYIIDFGL SKKYIDQNNR HIRNCTGKSL TGTARYSSIN ALEGKEQSIR
201   DDMESLVYVW VYLLHGRLPW MSLPTTGRKK YEAILMKKRS TKPEELCLGL
251   NSFFVNYLIA VRSLKFEEEP NYAMYRKMIY DAMIADQIPF DYRYDWVKTR
301   IVRPQRENQS QLSERQEGKC PNSAEFDGFS SIKGYSSHRQ VQSPVSSRDV
351   IKNSSSSPSK DILQSSTLDE SSQDKKPIKA VESNQKPYTP PRTINTTETR
401   MRSKTTINTA RTTAKNSSAV KKESSATRTV KKETHPATTK TTKTVNRQLN
451   SSTTKPATTS SHKDSEPASS RRTSTLRSSR RQNDGIRPAK ERTALFTATA
501   SKPPVSYRTG MLPKWMMAPL TSRR
```

TVEST-14G2

TcCK1.1

TcCK1.2

PfCK1

Yeast CK1

Mouse CK1

Human CK1-δ

Experiment Driven

Hypothesis

↓

Experiments

↓

Results

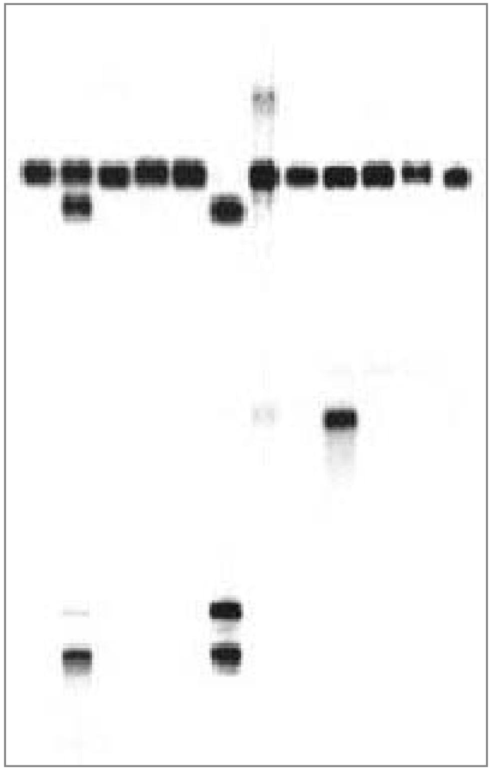
Information Driven

Experiments

↓

Hypothesis

The “old” biology



The most challenging task for a scientist is to get good data

The “new” biology

The most challenging task for a scientist is to make sense of lots of data

Old vs New – What's the difference? (1) Economics

- Miniaturize – less cost
- Multiplex – more data
- Parallelize – save time
- Automate – minimize human intervention
- Thus, you must be able to deal with large amounts of data and trust the process that generated it

What’s the difference?
(2) Scale

- From gene sequencing (~ 1 KB) to genome sequencing (many MB, even GB)
- From picking several genes for expression studies to analyzing the expression patterns of all genes
- From a catalog of key genes in a few key species to a catalog of all genes in many species
- Analyzing your data in isolation makes less sense when you can make much more powerful statements by including data from others

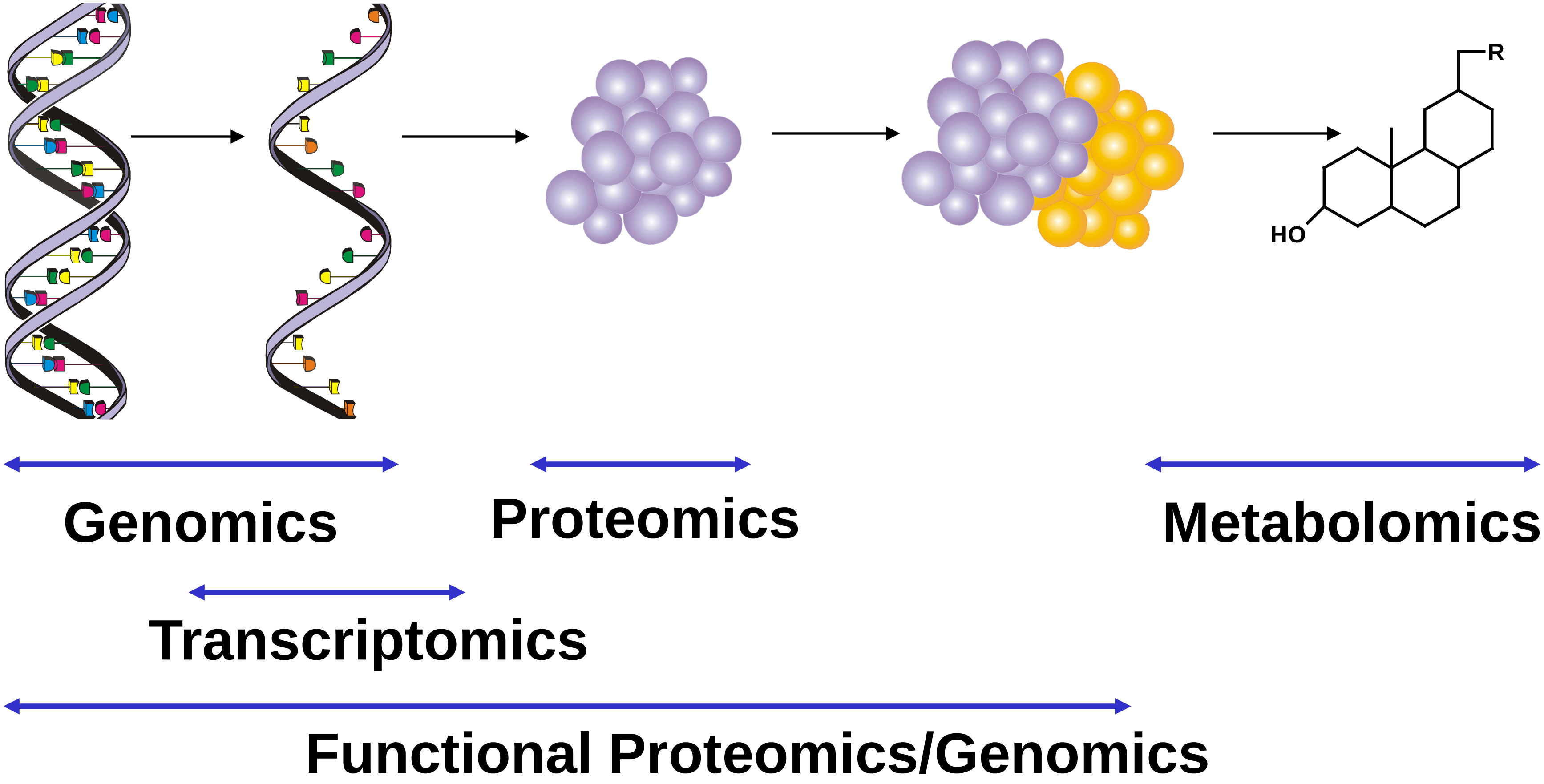
What’s the difference?
(3) Logic

- Hypothesis-driven research to data-driven research
- Expertise-driven approach versus information-driven approach
- Reductionist versus integrationist
- How to answer the question becomes how to question an answer
- Algorithmic approaches for filtering, normalizing, analyzing and interpreting become increasingly important

Data-driven Science Done Wrong

- Must have some hypothesis – data is not the end goal of science
- Finding patterns in the data is where analysis starts, not ends
- Must understand the limits of high-throughput technology (e.g. microarrays measure transcription only, one genome does not tell you about species variation, etc.)
- Must understand or explore the limits of your algorithm

Omics



Genomics

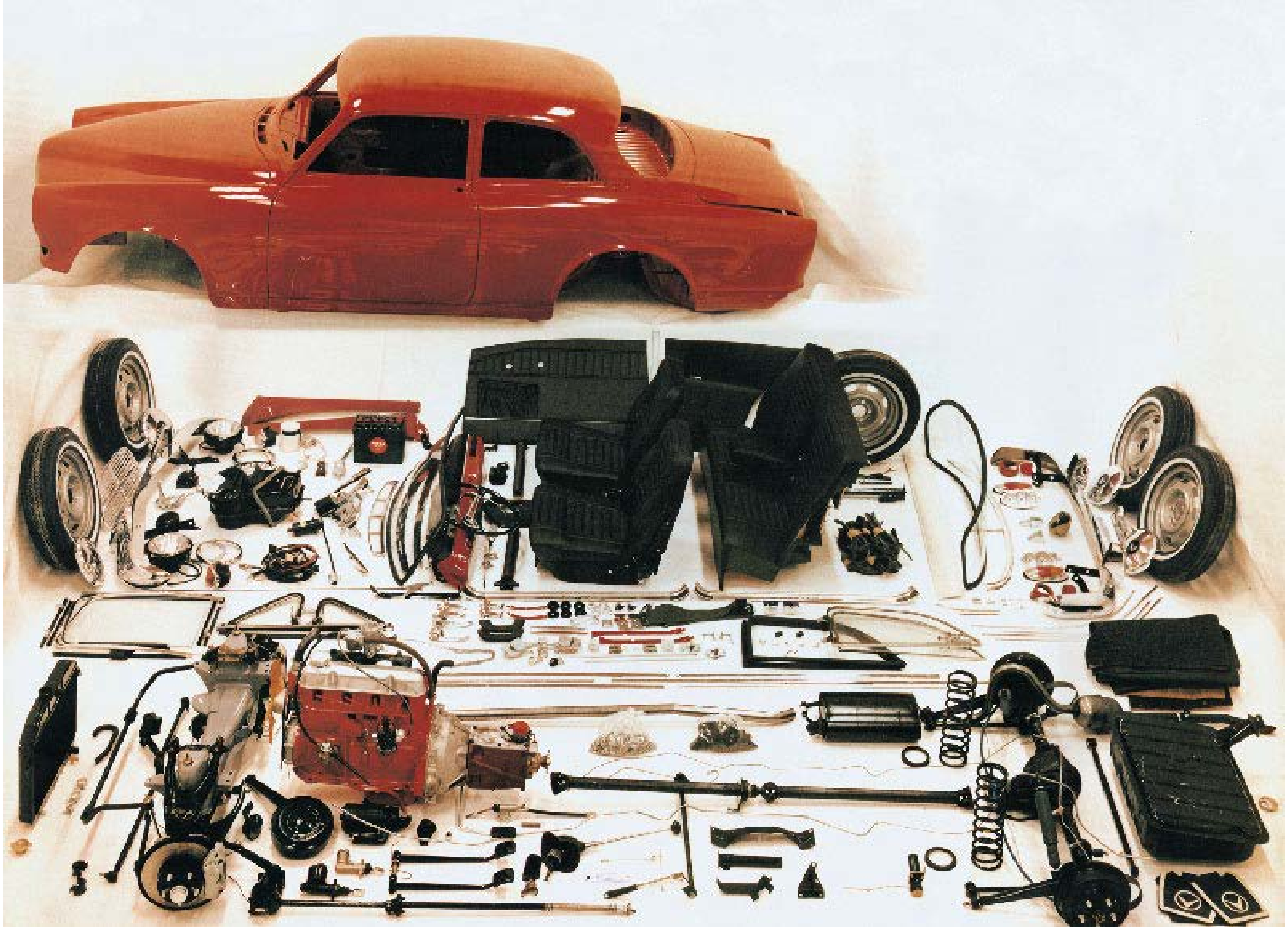
Proteomics

Metabolomics

Transcriptomics

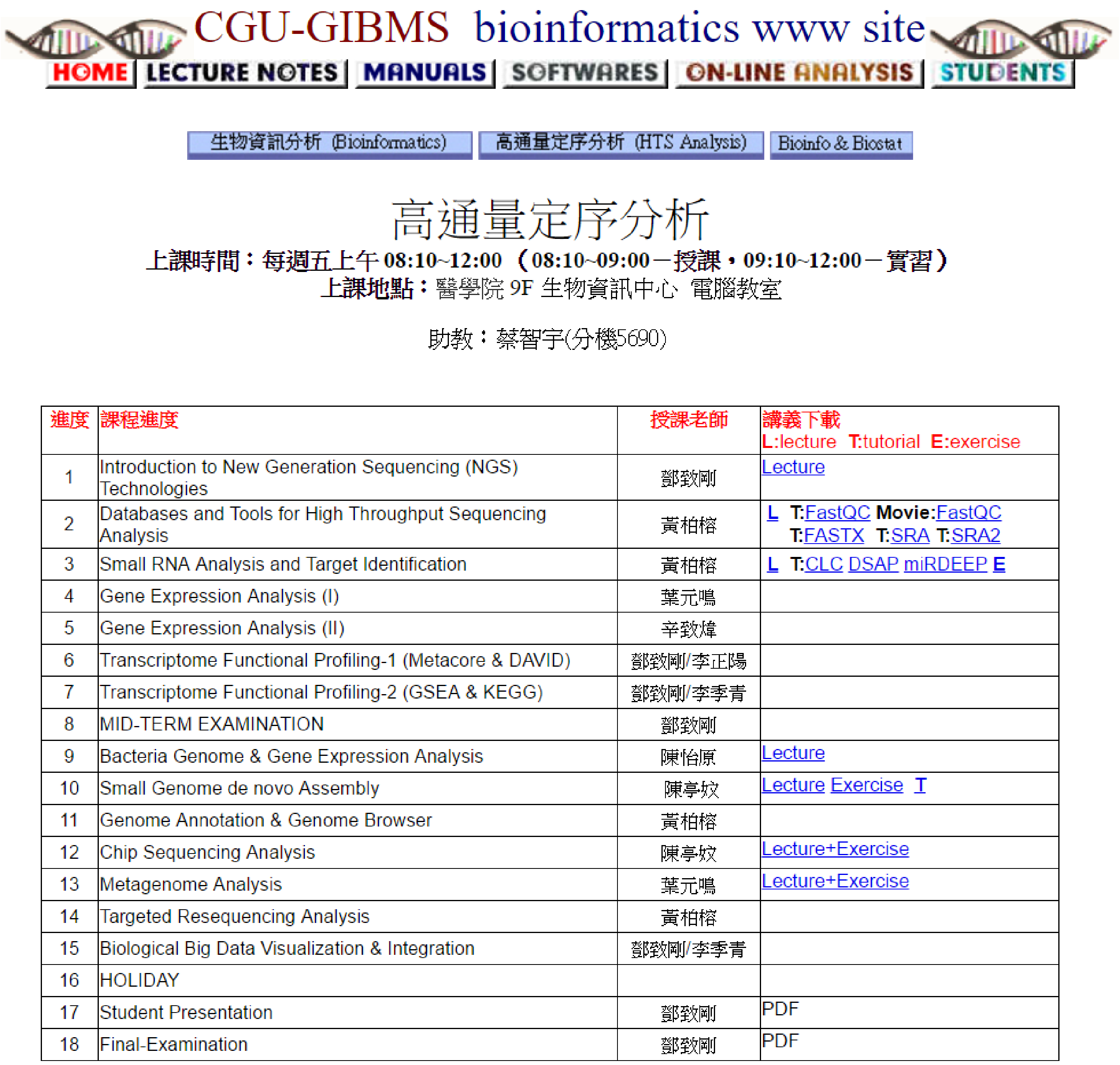
Functional Proteomics/Genomics

SYSTEMS BIOLOGY



In 20 Jan 2015, President Obama called for a new initiative to fund precision medicine

I want the country that eliminated polio and mapped the human genome to lead a new era of medicine—one that delivers the right treatment at the right time. In some patients with cystic fibrosis, this approach has reversed a disease once thought unstoppable. Tonight, I'm launching a new Precision Medicine Initiative to bring us closer to curing diseases like cancer and diabetes—and to give all of us access to the personalized information we need to keep ourselves and our families healthier.



高通量定序分析

上課時間：每週五上午 08:10~12:00（08:10~09:00－授課，09:10~12:00－實習）

上課地點：醫學院 9F 生物資訊中心 電腦教室

助教：蔡智宇(分機5690)

進度	課程進度	授課老師	講義下載 L:lecture T:tutorial E:exercise
1	Introduction to New Generation Sequencing (NGS) Technologies	鄧致剛	Lecture
2	Databases and Tools for High Throughput Sequencing Analysis	黃柏榕	L T:FastQC **Movie:**FastQC T:FASTX T:SRA T:SRA2
3	Small RNA Analysis and Target Identification	黃柏榕	L T:CLC DSAP miRDEEP E
4	Gene Expression Analysis (I)	葉元鳴	
5	Gene Expression Analysis (II)	辛致煒	
6	Transcriptome Functional Profiling-1 (Metacore & DAVID)	鄧致剛/李正陽	
7	Transcriptome Functional Profiling-2 (GSEA & KEGG)	鄧致剛/李季青	
8	MID-TERM EXAMINATION	鄧致剛	
9	Bacteria Genome & Gene Expression Analysis	陳怡原	Lecture
10	Small Genome de novo Assembly	陳亭妏	Lecture Exercise T
11	Genome Annotation & Genome Browser	黃柏榕	
12	Chip Sequencing Analysis	陳亭妏	Lecture+Exercise
13	Metagenome Analysis	葉元鳴	Lecture+Exercise
14	Targeted Resequencing Analysis	黃柏榕	
15	Biological Big Data Visualization & Integration	鄧致剛/李季青	
16	HOLIDAY		
17	Student Presentation	鄧致剛	PDF
18	Final-Examination	鄧致剛	PDF

Q. As a biologist, what skills do I need to make the transition to bioinformatics?

The fact is that many of the jobs available CURRENTLY involve the design and implementation of programs and systems for the storage, management and analysis of vast amounts of DNA sequence data. Such positions require in-depth programming and relational database skills which very few biologists possess, and so it is largely the computational specialists who are filling these roles. This is not to say the computer-savvy biologist doesn't play an important role. As the bioinformatics field matures there will be a huge demand for outreach to the biological community, as well as the need for individuals with the in-depth biological background necessary to sift through gigabases of genomic sequence in search of novel targets. It will be in these areas that biologists with the necessary computational skills will find their niche.

A. Molecular biology packages (GCG, BLAST etc),
Web and programming skills including HTML, Perl, JAVA and C++,
Familiar with a variety of operating systems (especially UNIX),
Relational database skills such as SQL, Sybase or Oracle,
Statistics,
Structural biology and modeling,
Mathematical optimization,
Computer graphics theory and linear algebra.

You will need to be able to readily pick up, use and understand the tools and databases designed by computer programmers, and
To communicate biological science requirements to core computer scientists.