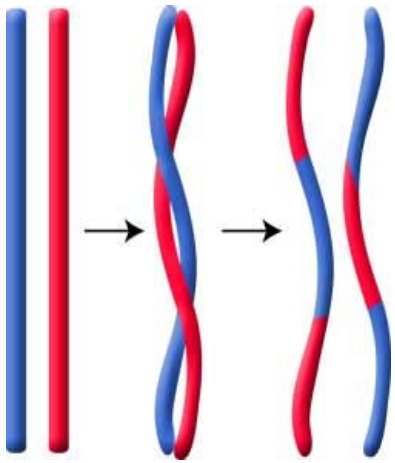
http://pastime.cgu.edu.tw/petang/index.htm



Bioinformatics

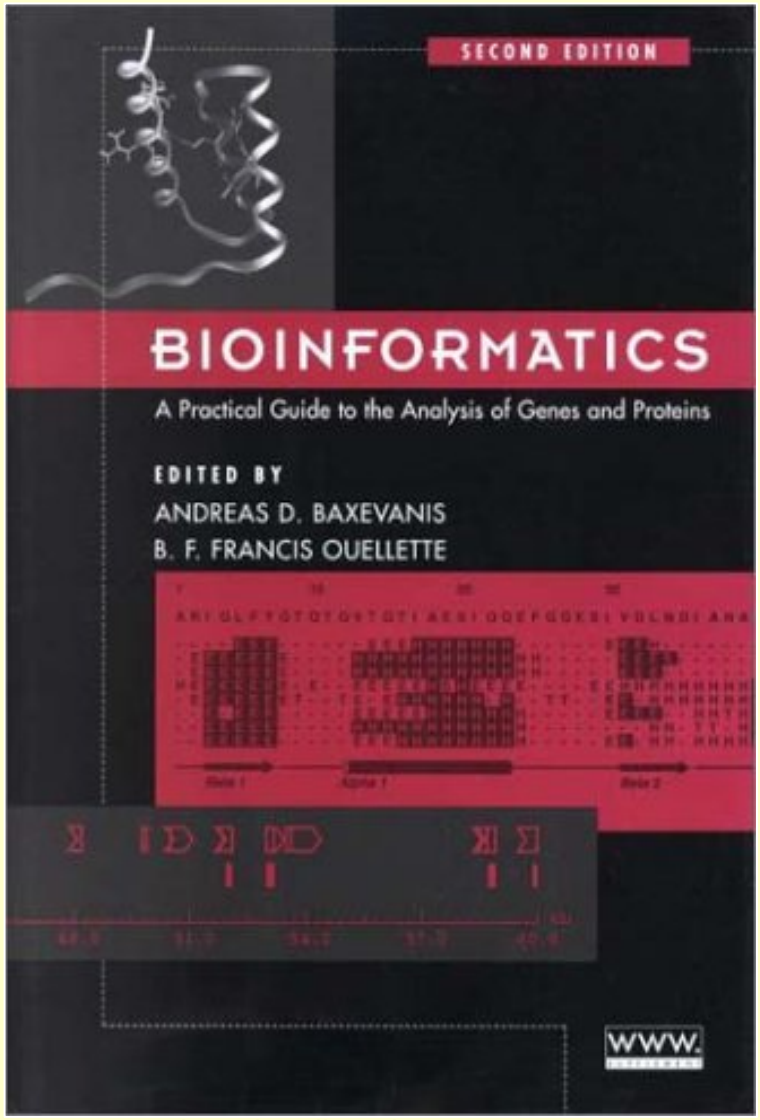
Lecture 1 – Introduction to Bioinformtics

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

助教：
方怡凱（分機5690）
陳玉純（分機5690）

Bioinformatics: A Practical Guide to the Analysis of Genes & Proteins

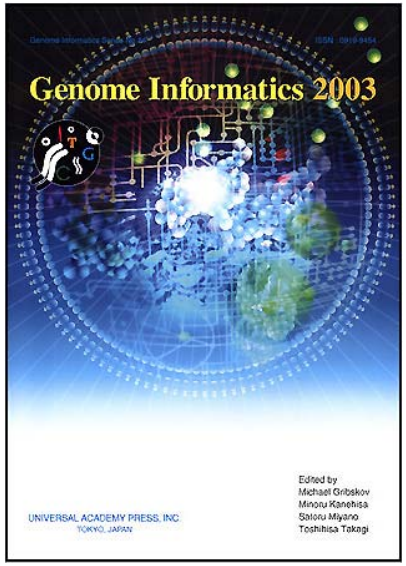


432 pages (2001) Wiley-Liss; ISBN: 0471383910

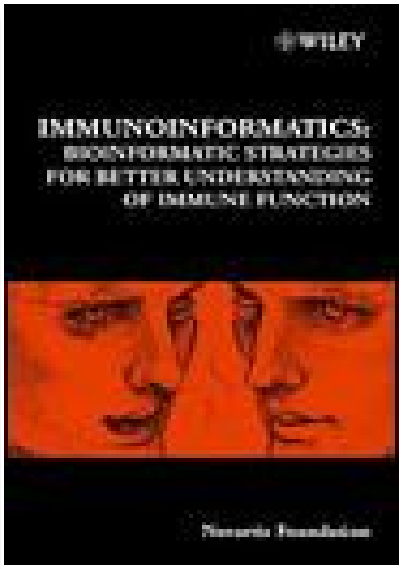
Contents

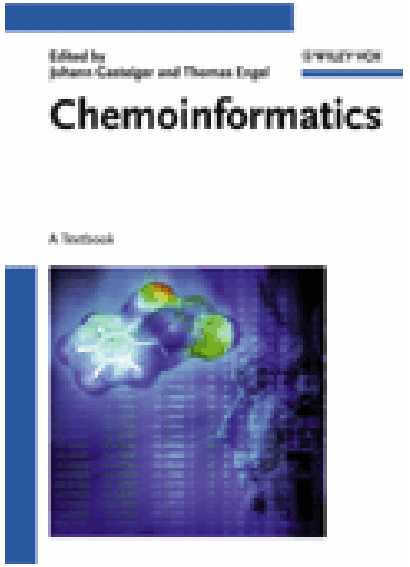
- *Bioinformatics and the Internet*
- *The NCBI Data Model*
- *The GenBank Sequence Database*
- *Structure Databases*
- *Genomic Mapping and Mapping Databases*
- *Information Retrieval from Biological*
- *Databases*
- *Sequence Alignment and Database*
- *Searches*
- *Multiple Sequence Alignment*
- *Predictive Methods using DNA Sequences*
- *Predictive Methods using Protein*
- *Sequences Expressed Sequence Tags*
- *Sequence Assembly and Finishing*
- *Methods Phylogenetic Analysis*
- *Comparative Genome Analysis*
- *Using Perl to Facilitate Biological Analysis*

Bio informatics







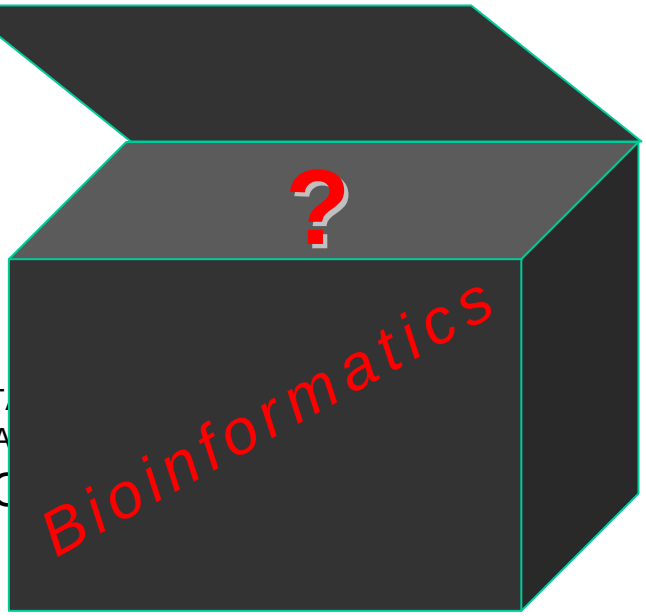


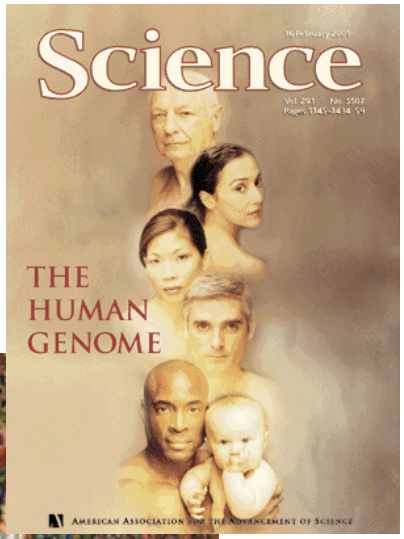
-Omics Mania *biome, cellomics, chronomics, clinomics, complexome, crystallomics, cytomics, degradomics, diagnomics, enzymome, epigenome, expressome, fluxome, foldome, secretome, functome, functomics, genomics, glycomics, immunome, transcriptomics, integromics, interactome, kinome, ligandomics, lipoproteomics, 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...*

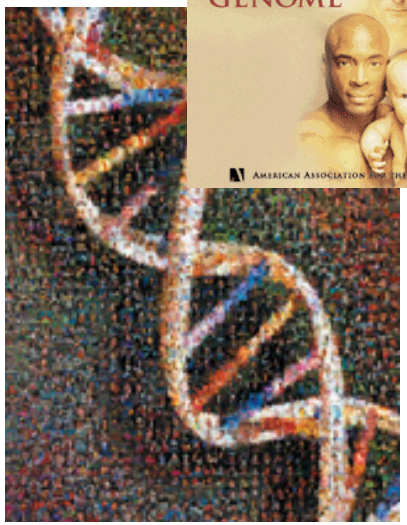
WHAT IS BIOINFORMATICS?

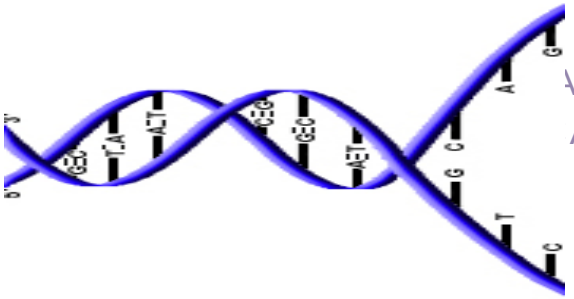


CTAGCTAGCTAGCTA
CTAGCTAGCTAGCTATCGA
CATGCACTAGCTAG





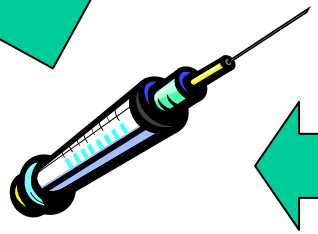


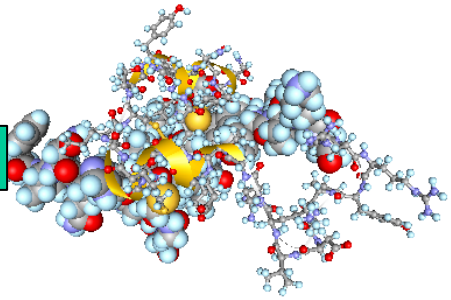


AAATGGCCAATTGATGACCAGAGATTTAGGCCAATTAA
AATGTGAAATGGCCAATTGATGACCAGAGA





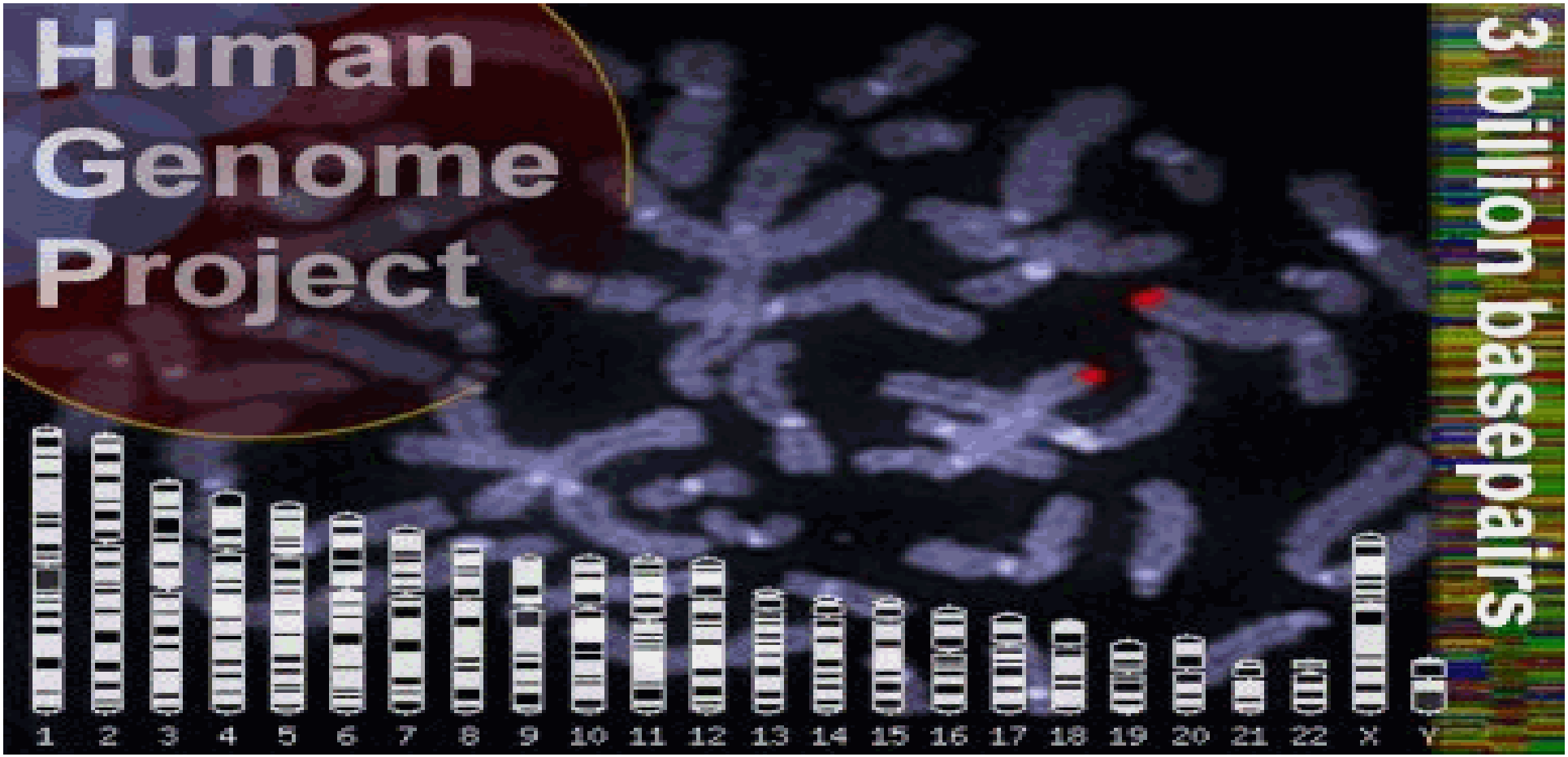


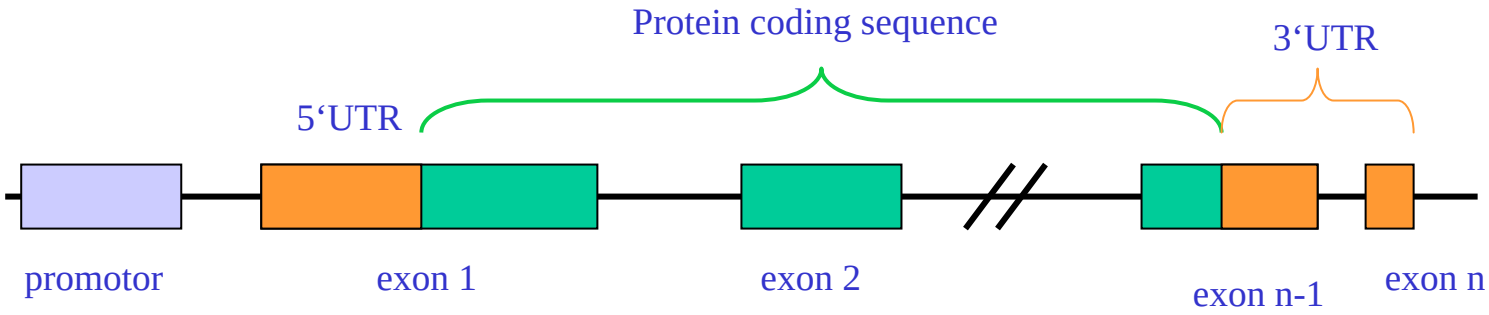


What is Bioinformatics?

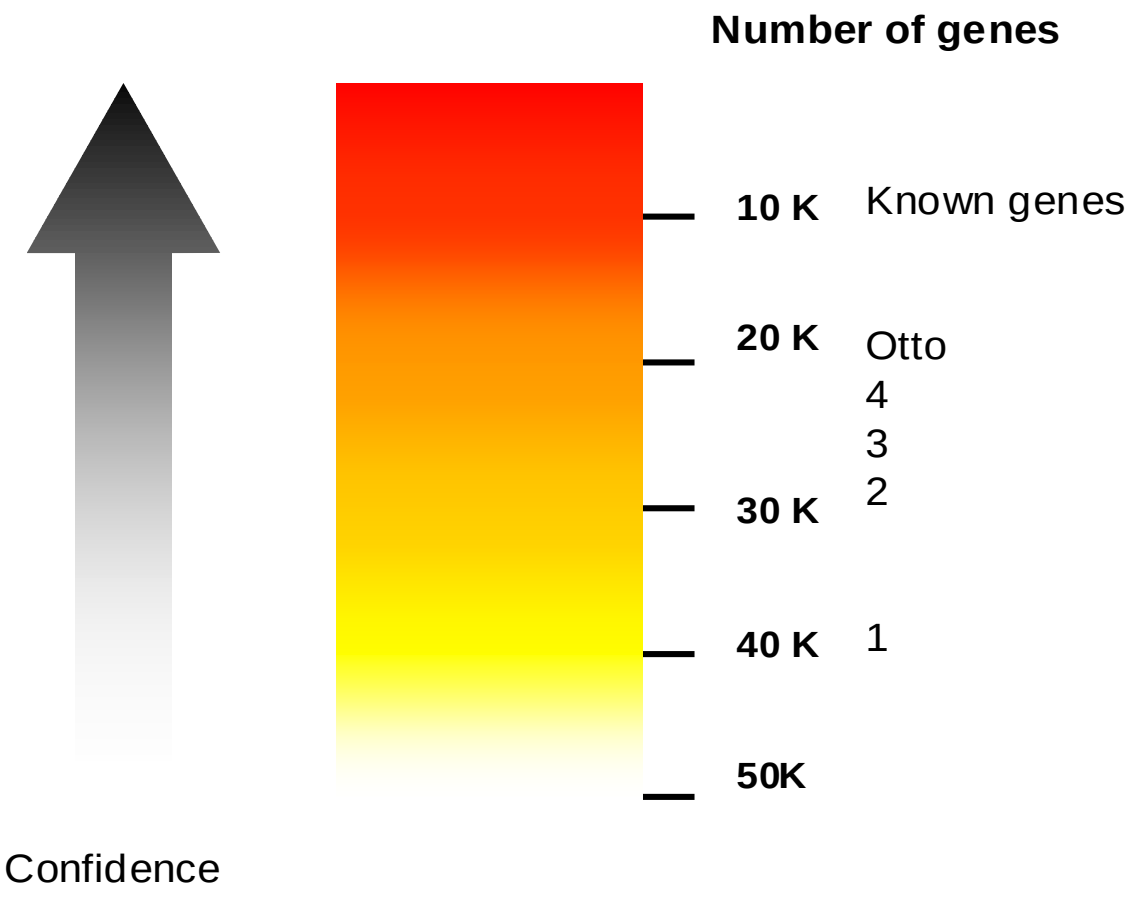
- Development of methods & algorithms to organize, integrate, analyze and interpret biological and biomedical data
- Study of the inherent structure & flow of biological information
- Goals of bioinformatics:
  - Identify patterns
  - Classify
  - Make predictions
  - Create models
  - Better utilize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

結合生物學、計算機科學與資訊學的技術, 應用於生物化學資料的處理，
將繁瑣無意的資料轉化成有意義、有價值的訊息。


Human
Genome
Project
3 billion basepairs
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y

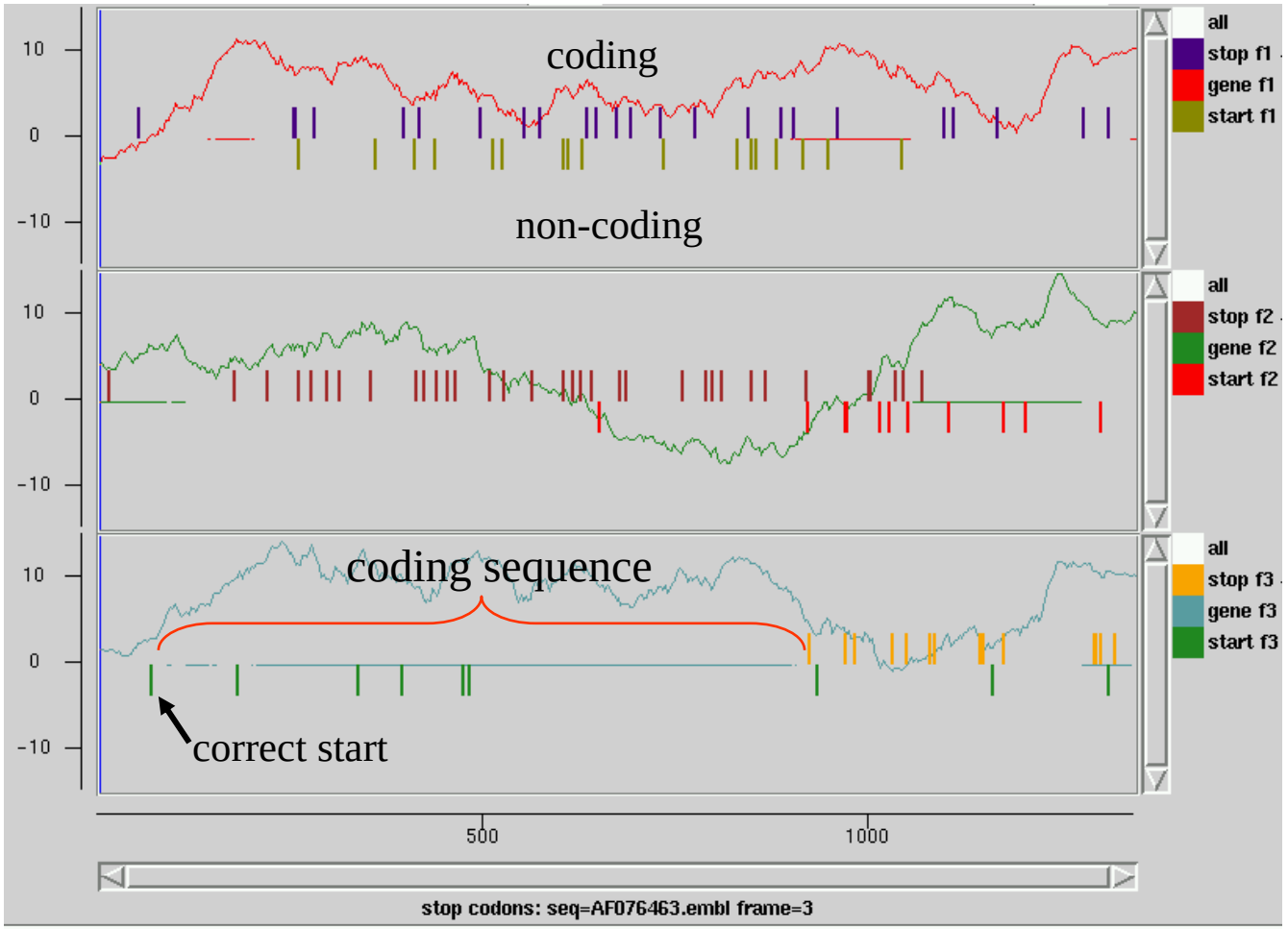

Protein coding sequence
3‘UTR
5‘UTR
promotor
exon 1
exon 2
exon n-1
exon n

Gene Number in the Human Genome



Gene prediction

Codon usage (single exon)



Gene prediction

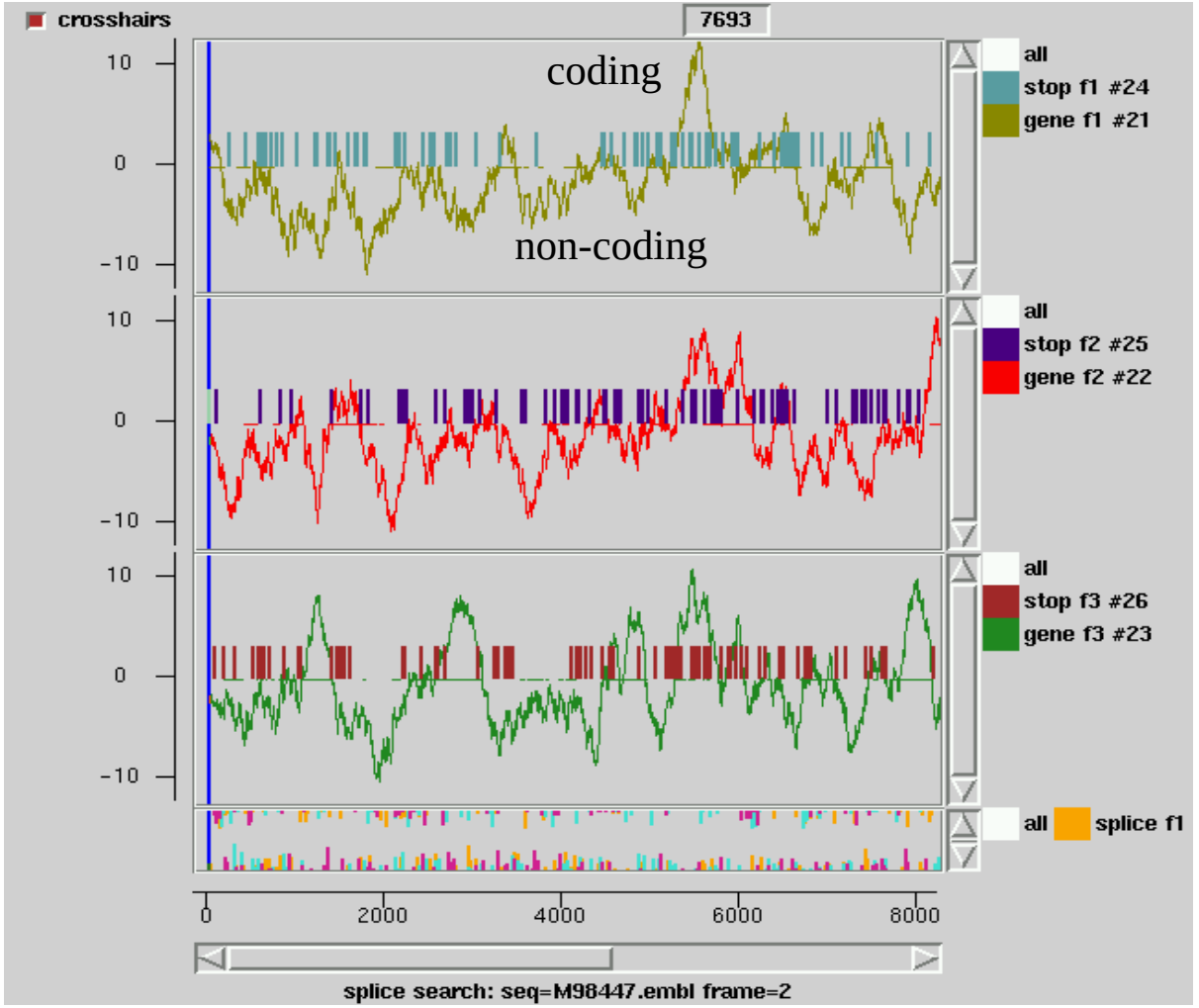
Codon usage (multiple exons)

Frame 1

Frame 2

Frame 3

Splice sites



Exons:

208. .295

1029. .1349

1500. .1688

2686. .2934

3326. .3444

3573. .3680

4135. .4309

4708. .4846

4993. .5096

7301. .7389

7860. .8013

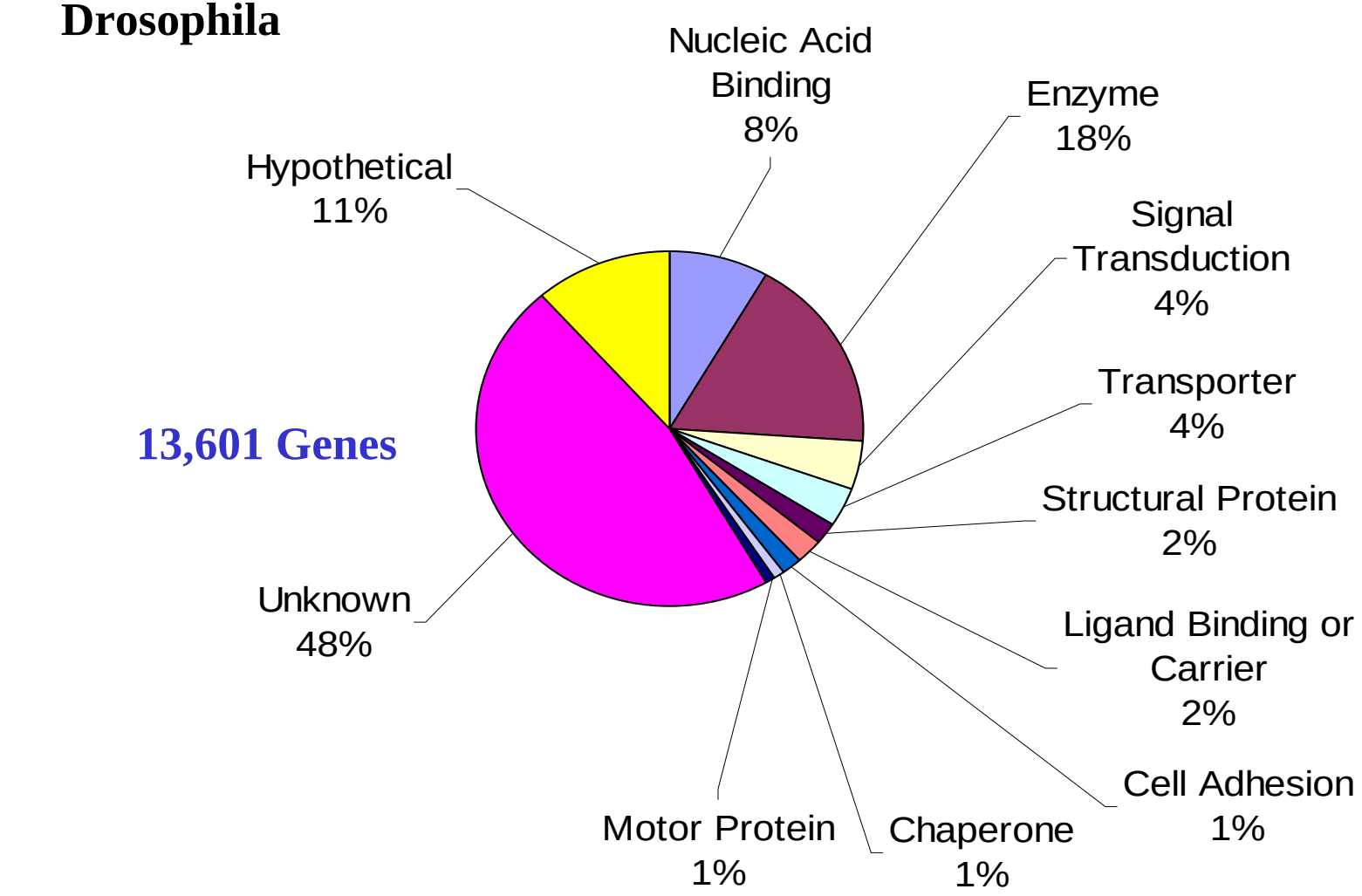
8124. .8405

8553. .8713

9089. .9225

13841. .14244

Drosophila


Nucleic Acid Binding 8%
Enzyme 18%
Signal Transduction 4%
Transporter 4%
Structural Protein 2%
Ligand Binding or Carrier 2%
Cell Adhesion 1%
Chaperone 1%
Motor Protein 1%
Unknown 48%
Hypothetical 11%
13,601 Genes
Nucleic Acid Binding
Enzyme
Signal Transduction
Transporter
Structural Protein
Ligand Binding or Carrier
Cell Adhesion
Chaperone
Motor Protein
Unknown
Hypothetical



Experiment Driven

Hypothesis

↓

Experiments

↓

Results



Information Driven

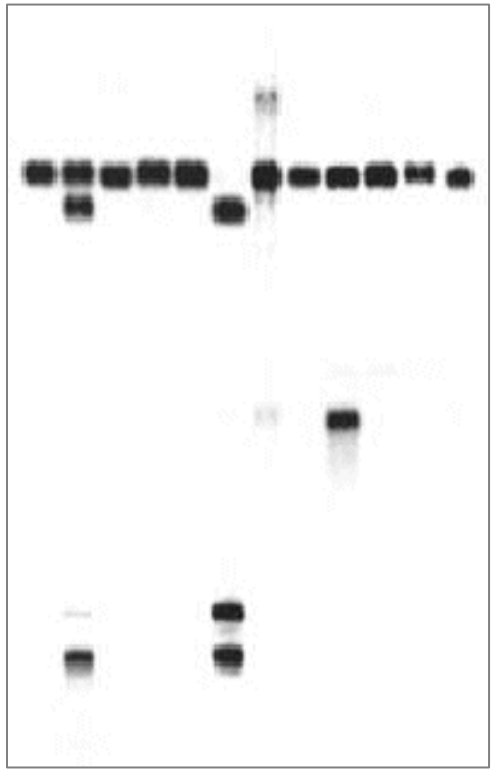
Experiments

↓

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

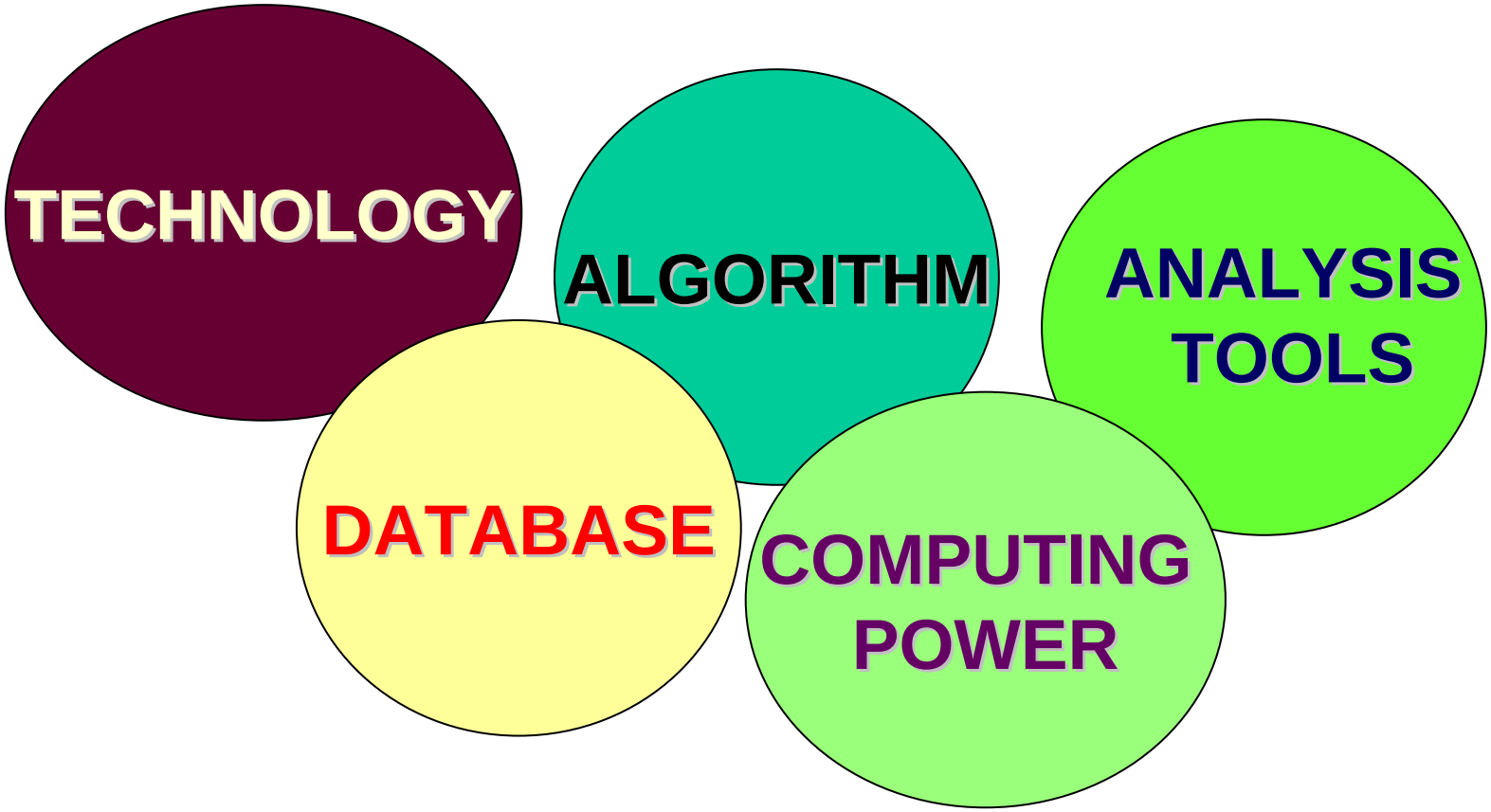
- Hypothesis-driven research to data-driven research
- Expertise-driven approach versus information-driven approach
- Reductionist versus integrationist
- How to answer the question becomes how to question an answer
- Algorithmic approaches for filtering, normalizing, analyzing and interpreting become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

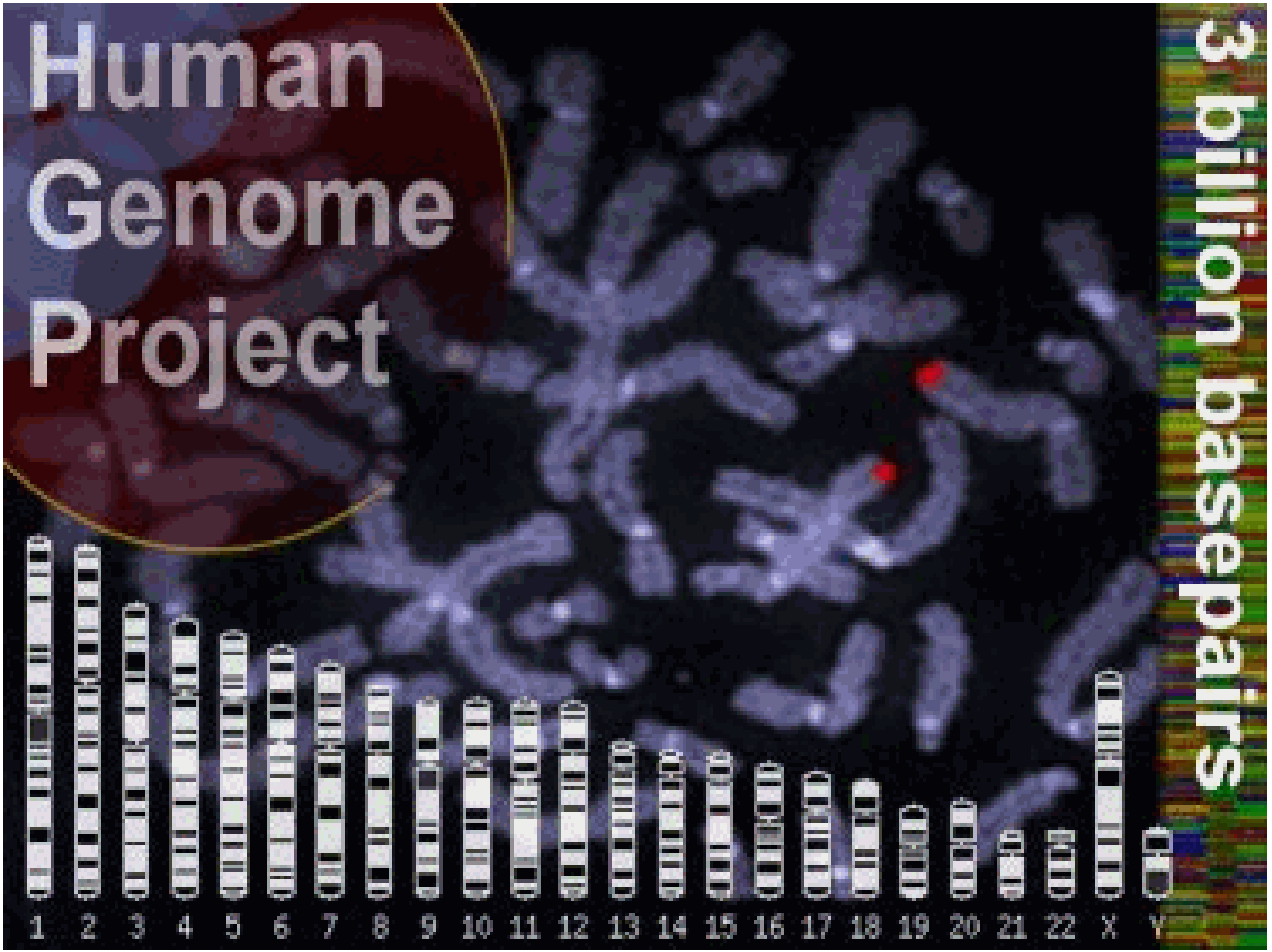
Data-driven Science Done Wrong

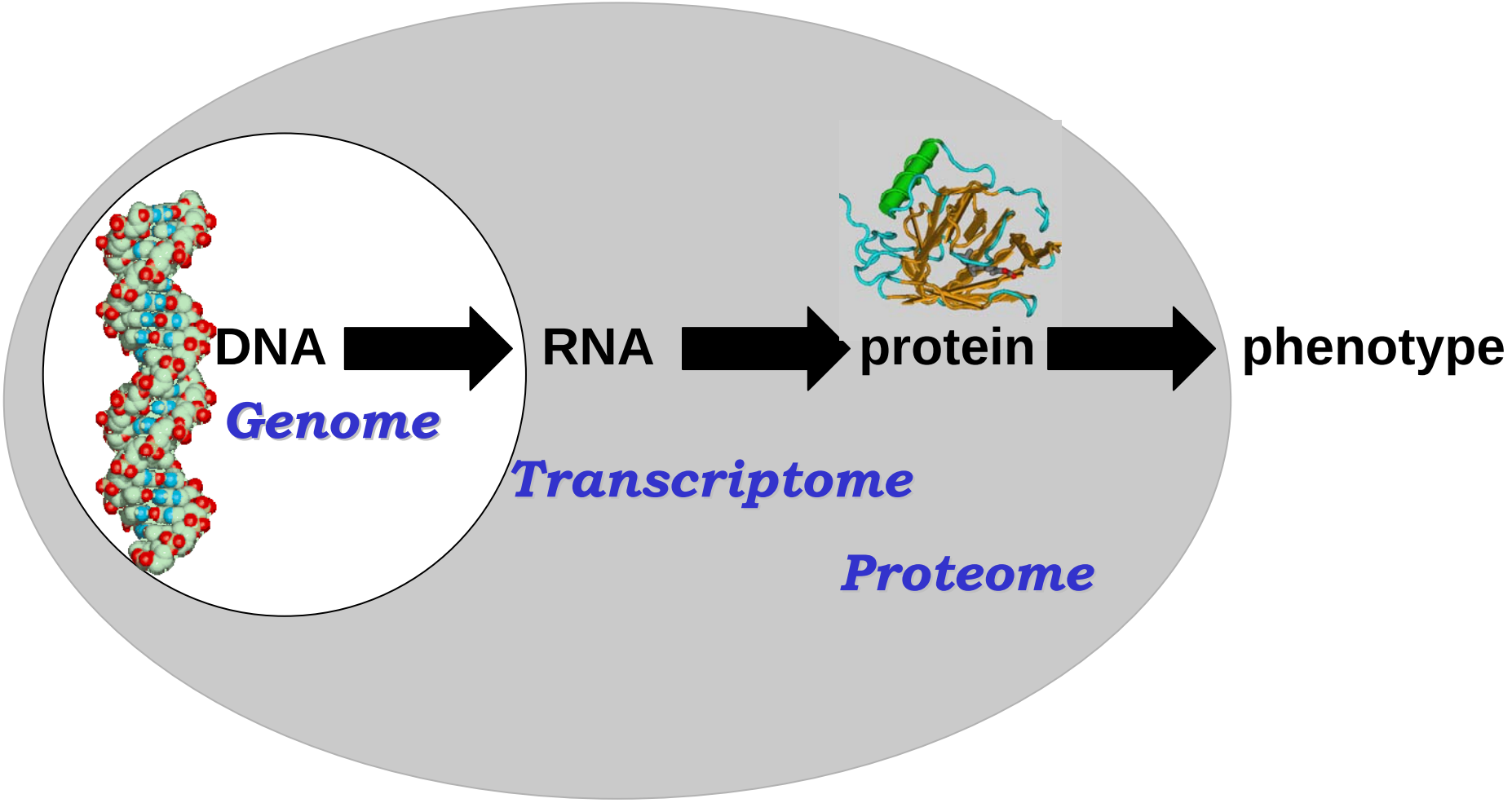
- Must have some hypothesis – data is not the end goal of science
- Finding patterns in the data is where analysis starts, not ends
- Must understand the limits of high-throughput technology (e.g. microarrays measure transcription only, one genome does not tell you about species variation, etc.)
- Must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



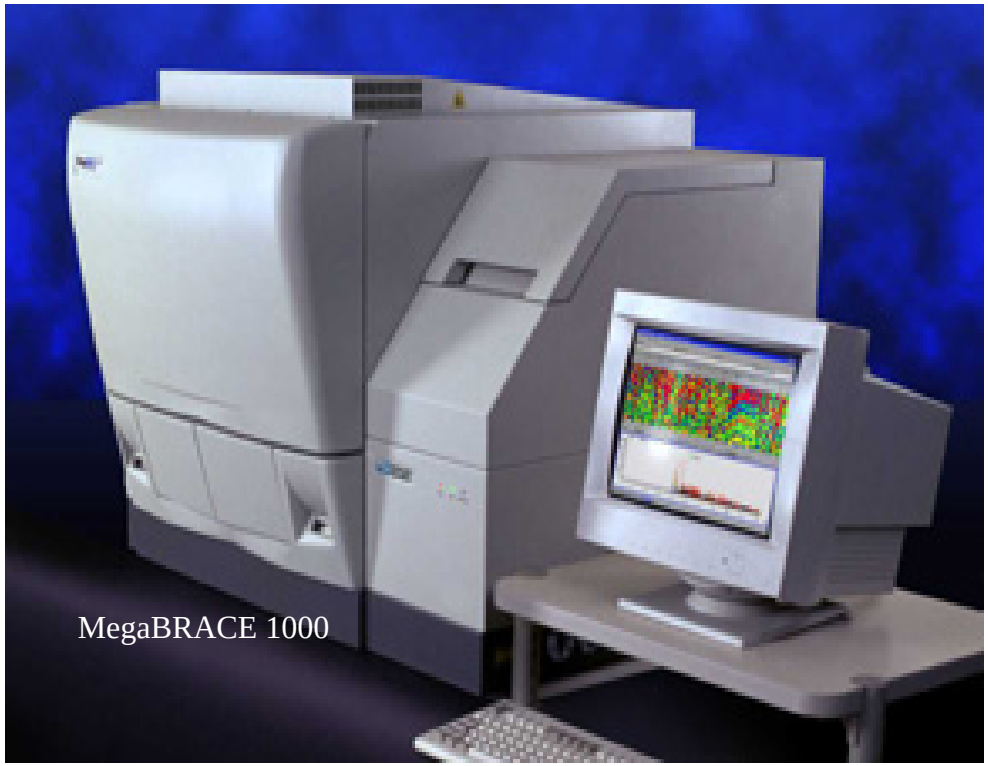
THE COMPONENTS OF BIOINFORMATICS




Human
Genome
Project
3 billion base pairs
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y


DNA
RNA
protein
phenotype
Genome
Transcriptome
Proteome

DNA Sequencing

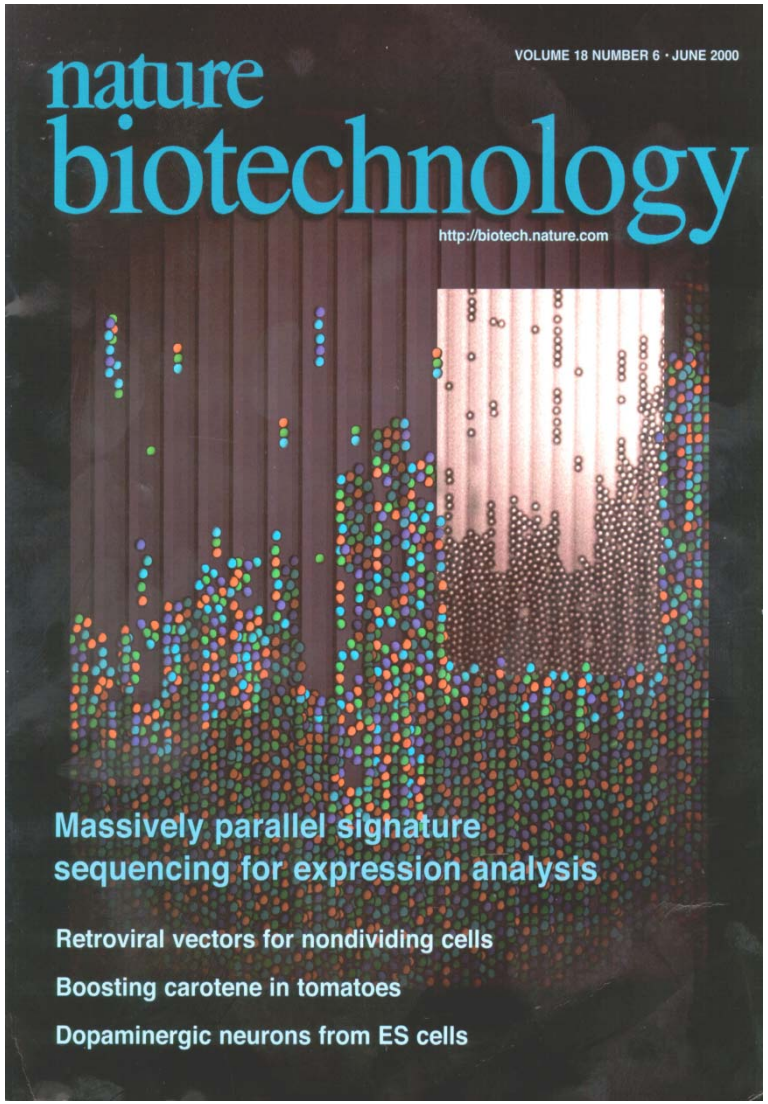




96 DNA sequencing in 2 hrs, approximately 600-800 readable bps per run.

1,000,000 bps in 24 hrs.

Next Generation Sequencing Technology



Massively **P**arallel **S**ignature **S**equencing (MPSS)

Roche 454 GS FLX
http://www.454.com/

Illumina SOLEXA
http://www.illumina.com/pages.ilmn?ID=250

Applied Biosystems SOLID
http://marketing.appliedbiosystems.com/

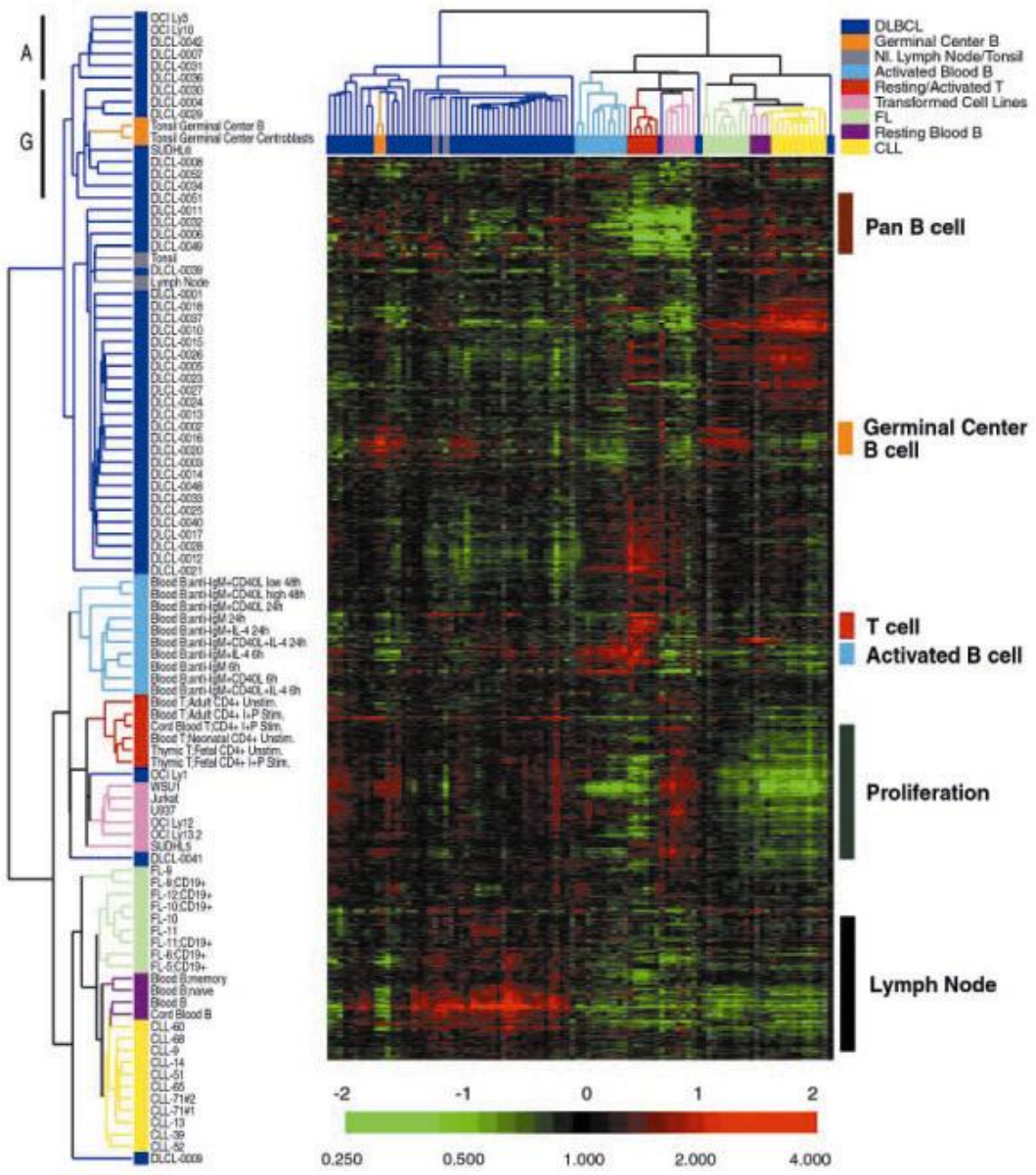
VisiGen Biotechnologies
http://visigenbio.com/

Helicos BioSciences
http://www.helicosbio.com/

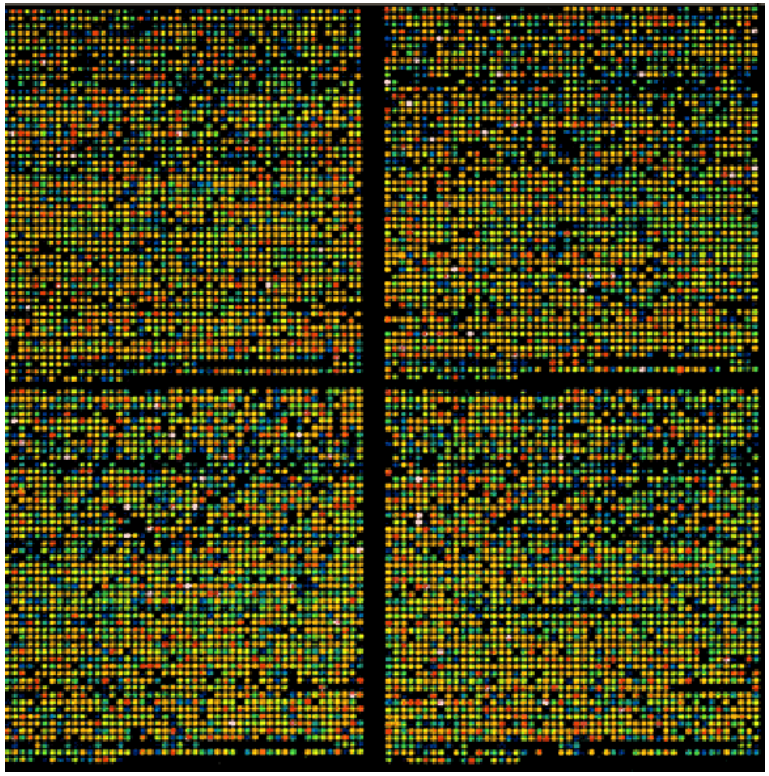
Reveo Inc.
http://www.reveo.com/

1000MB per run, Human genome in 3 months

Microarray

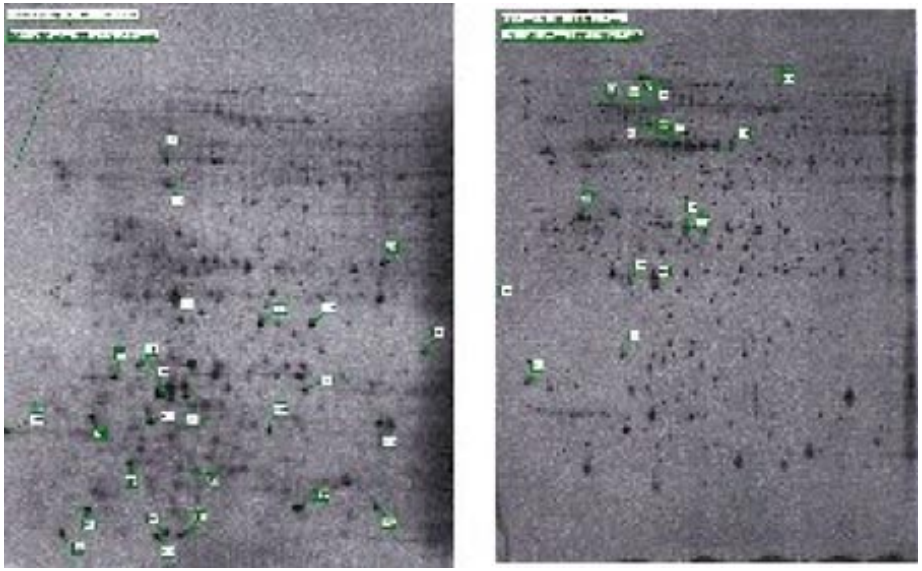


20,000-40,000 Clones per slide



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

Proteomics

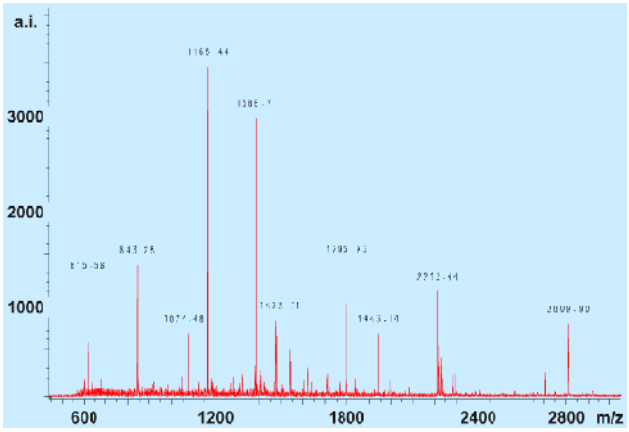


2 Dimensional Electrophoresis gels, differences that are characteristics of the individual starting states recognized by comparison of two protein pattern

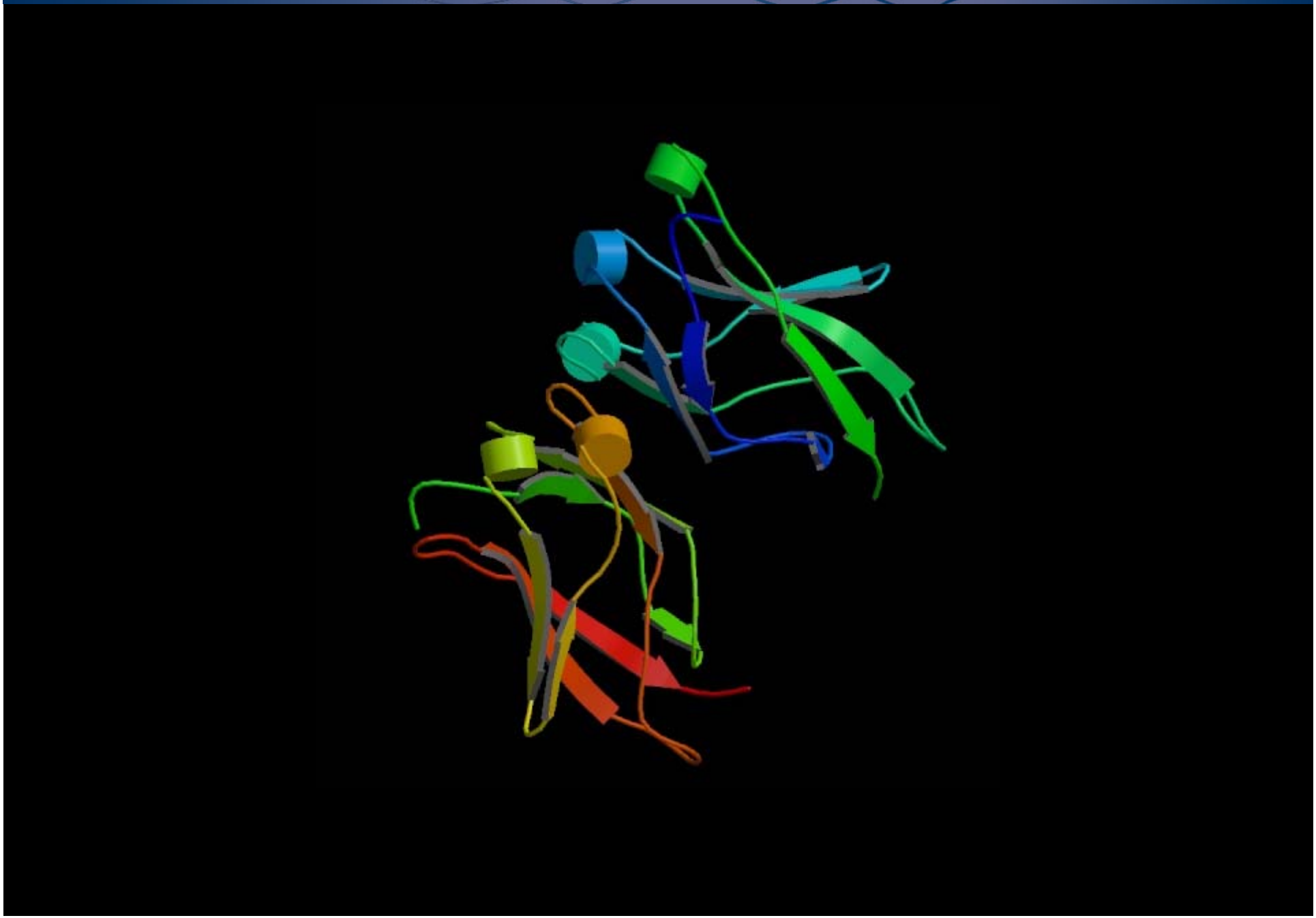
6,000 protein spots per gel

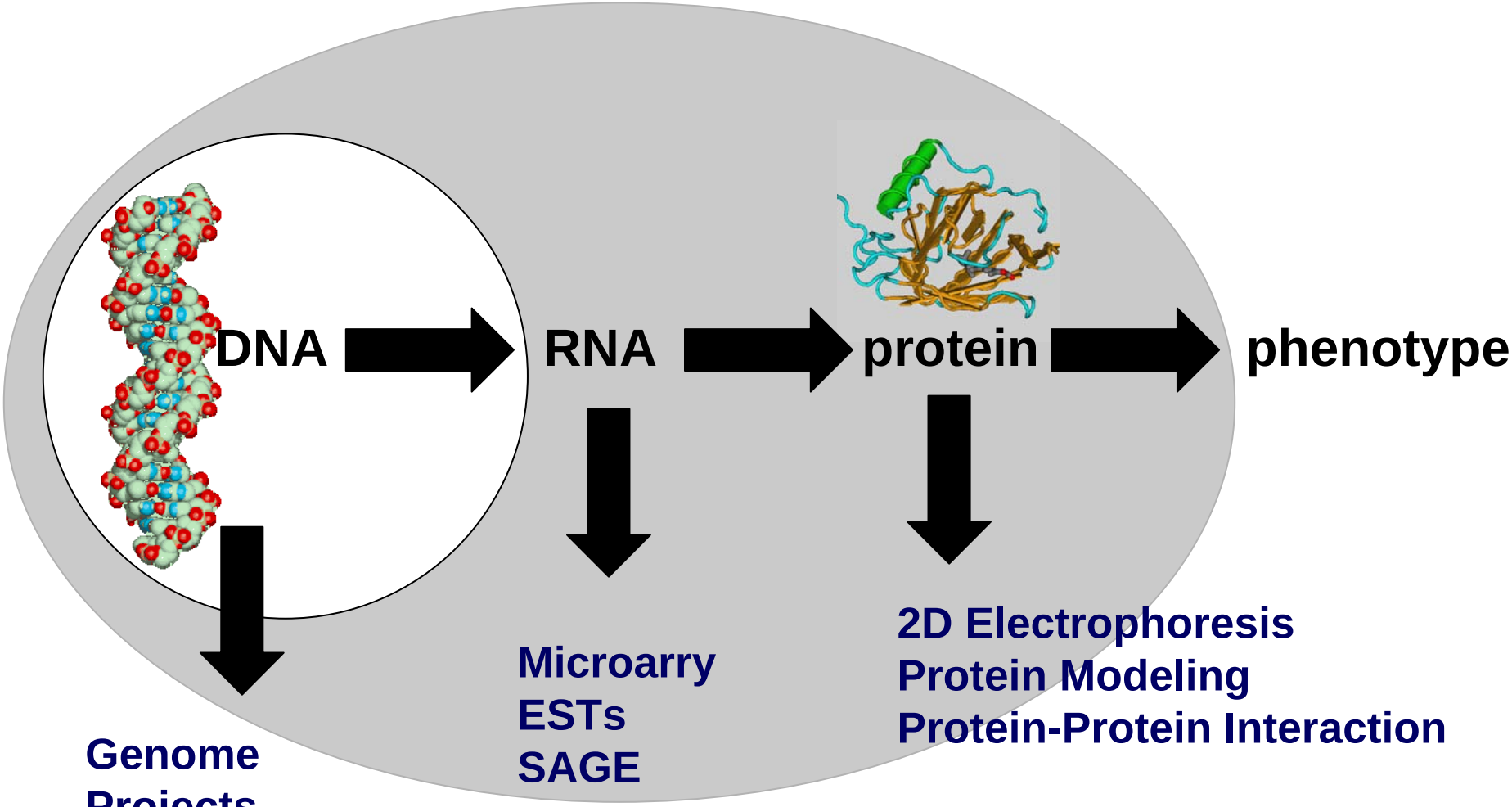


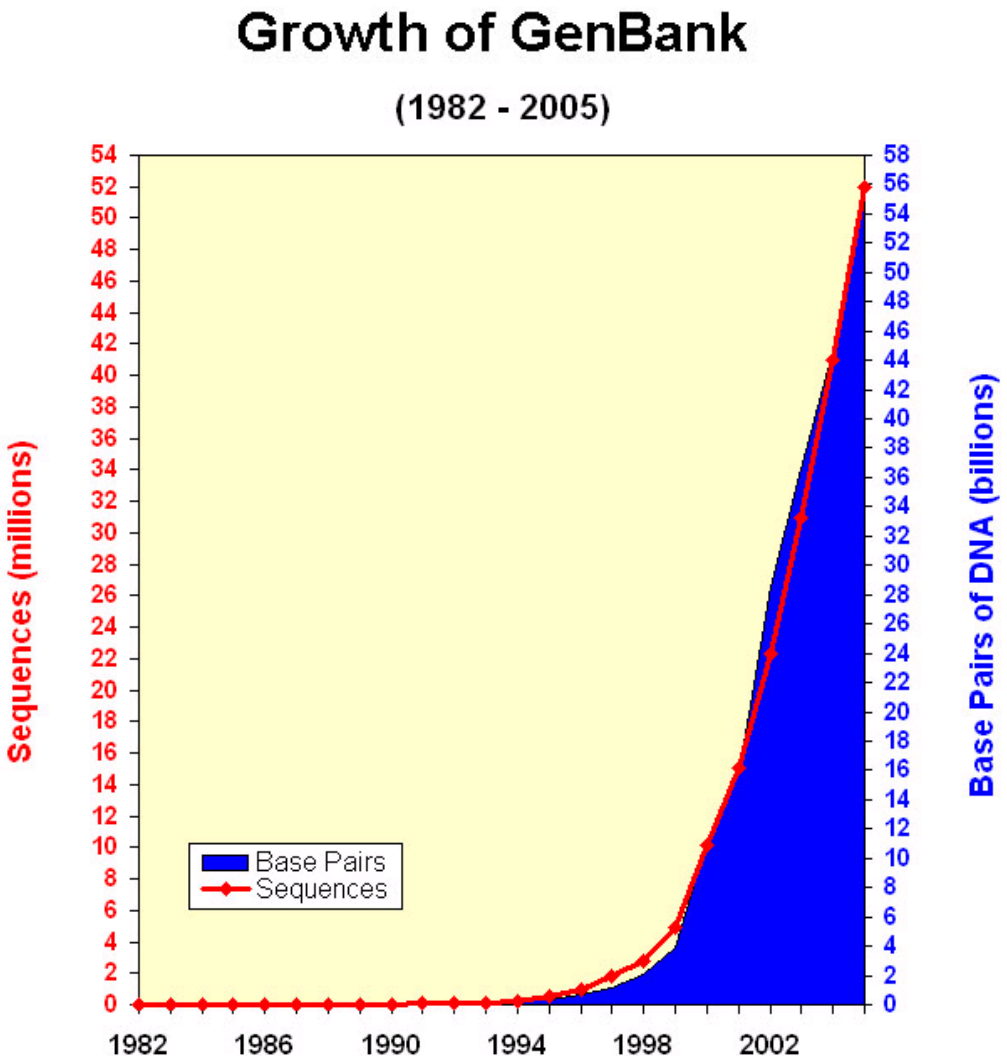
MALDI-MS peptide mass fingerprint, for identification of proteins separated by 2D electrophoresis



3D Modeling




DNA
RNA
protein
phenotype
Genome
Projects
Microarry
ESTs
SAGE
2D Electrophoresis
Protein Modeling
Protein-Protein Interaction



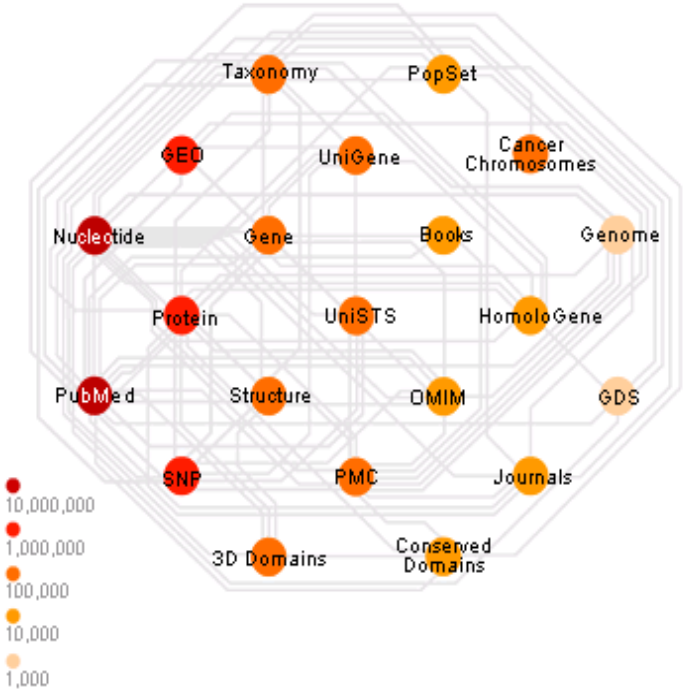
Genetic Sequence Data Bank Aug 15 2008, Release 167.0

95,033,791,652 bases, from
92,748,599 reported sequences

Homo sapiens
13,124,444,947 bases from
11,535,248 sequences

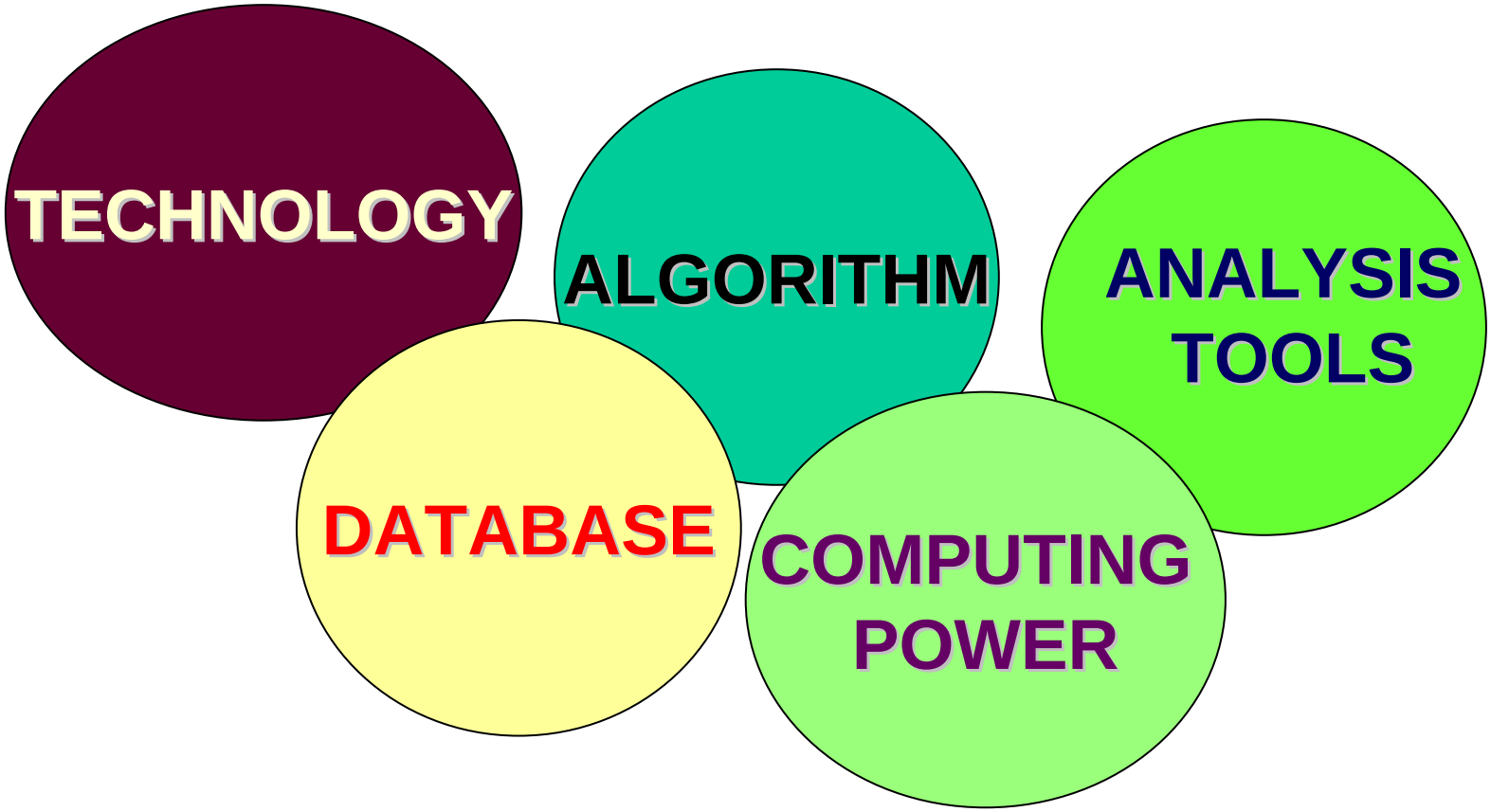
Recent years have seen an explosive growth in biological data. Large sequencing projects are producing increasing quantities of nucleotide sequences. The contents of nucleotide databases are doubling in size approximately every 14 months. The latest release of GenBank (V.139) exceeded two billion base 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
11535248	13124444947	*Homo sapiens*
7252378	8358715455	*Mus musculus*
1642662	5991517925	*Rattus norvegicus*
2086180	5228482576	*Bos taurus*
3188970	4578968522	*Zea mays*
2126845	3141652150	*Sus scrofa*
1588532	2932513510	*Danio rerio*
1205445	1533452587	*Oryza sativa Japonica Group*
227973	1352646211	*Strongylocentrotus purpuratus*
1673014	1142506965	*Nicotiana tabacum*
1410967	1044923875	*Xenopus (Silurana) tropicalis*
212933	996033334	*Pan troglodytes*
779849	911708853	*Drosophila melanogaster*
2210667	911688499	*Arabidopsis thaliana*
650352	905008645	*Vitis vinifera*
803827	869211632	*Gallus gallus*
76854	802815723	*Macaca mulatta*
1215317	748029713	*Ciona intestinalis*
1223247	706524422	*Canis lupus familiaris*
1111132	667180484	*Triticum aestivum*





THE COMPONENTS OF BIOINFORMATICS



The International Nucleotide Sequence Database Collaboration



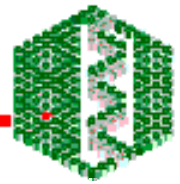
GenBank: *http://www.ncbi.nlm.nih.gov/*
National Center for Biotechnology Information (NCBI)



DDBJ: *http://www.ddbj.nig.ac.jp/*
National Institute of Genetics (NIG)

EMBL: *http://www.ebi.ac.uk*
European Bioinformatics Institute (EBI)

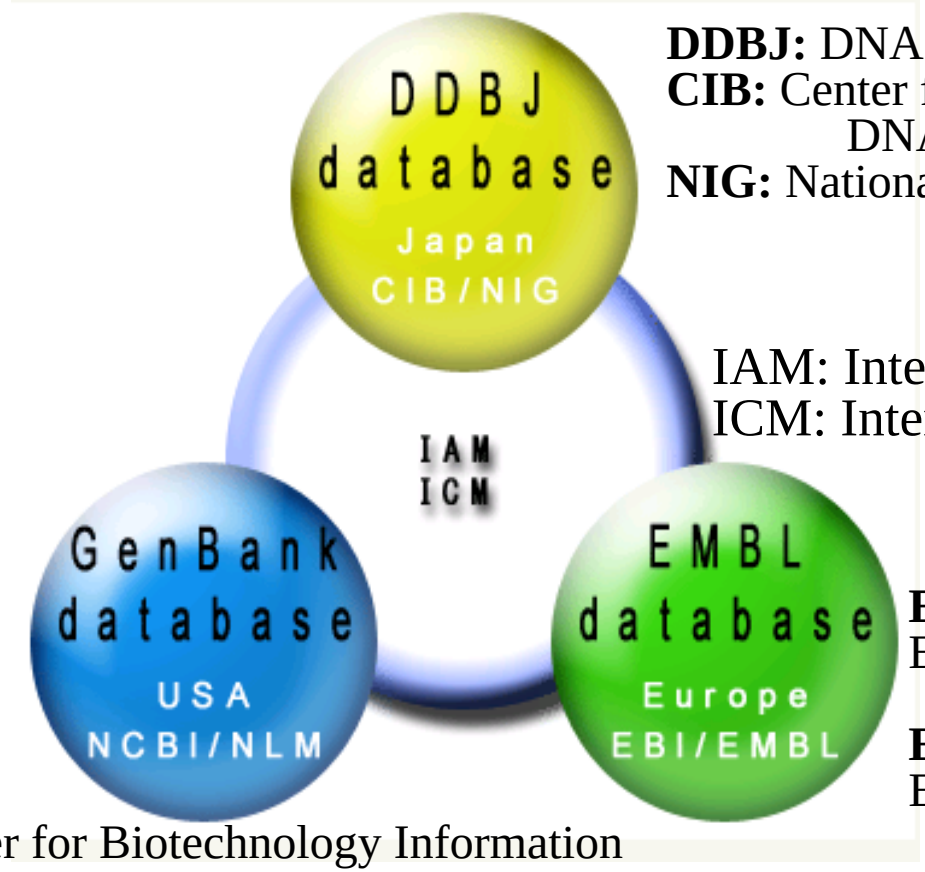






ExPASy: *http://tw.expasy.org*
Expert Protein Analysis System

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

Protein Databases

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



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

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



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

Protein Databases

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



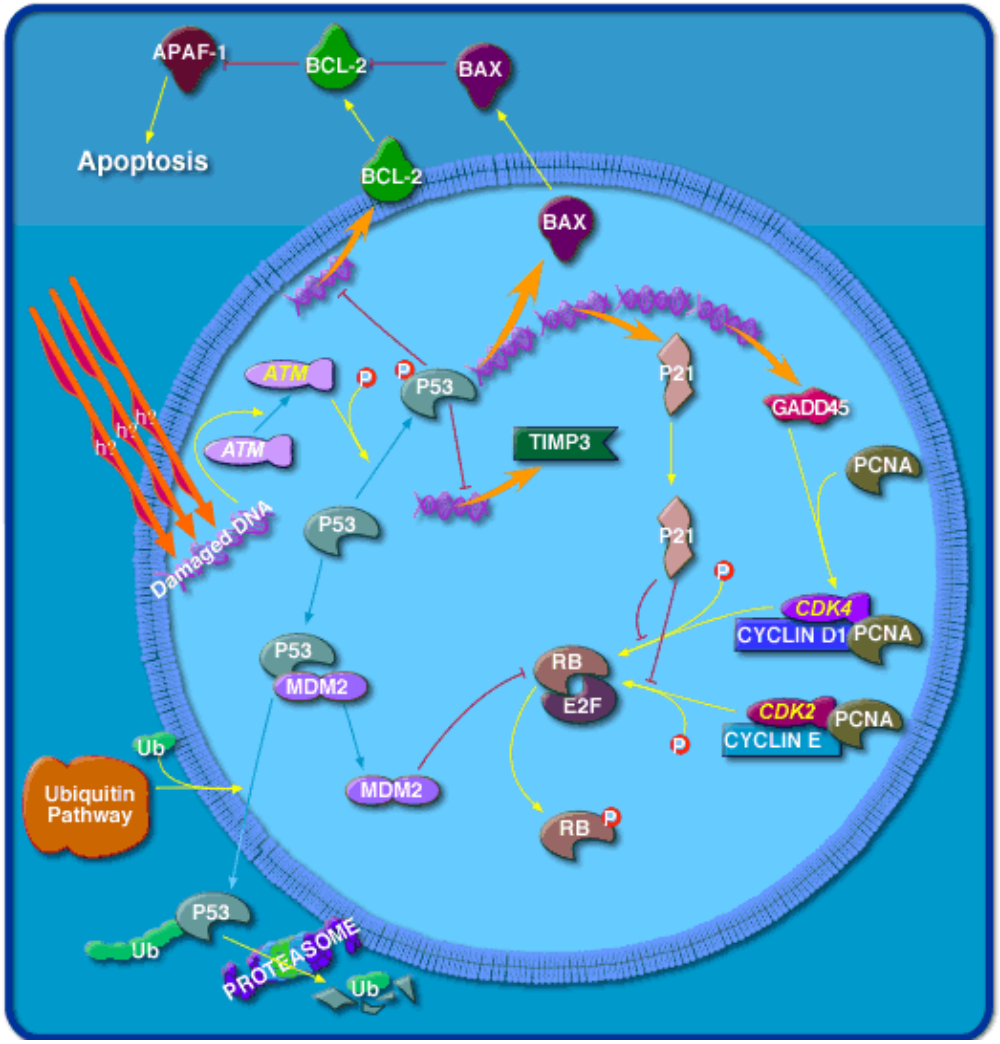
The ExPASy (**Ex**pert **P**rotein **A**nalysis **Sy**stem) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE

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

PROTEIN DATA BANK

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.

Metabolic & Signalling Pathways



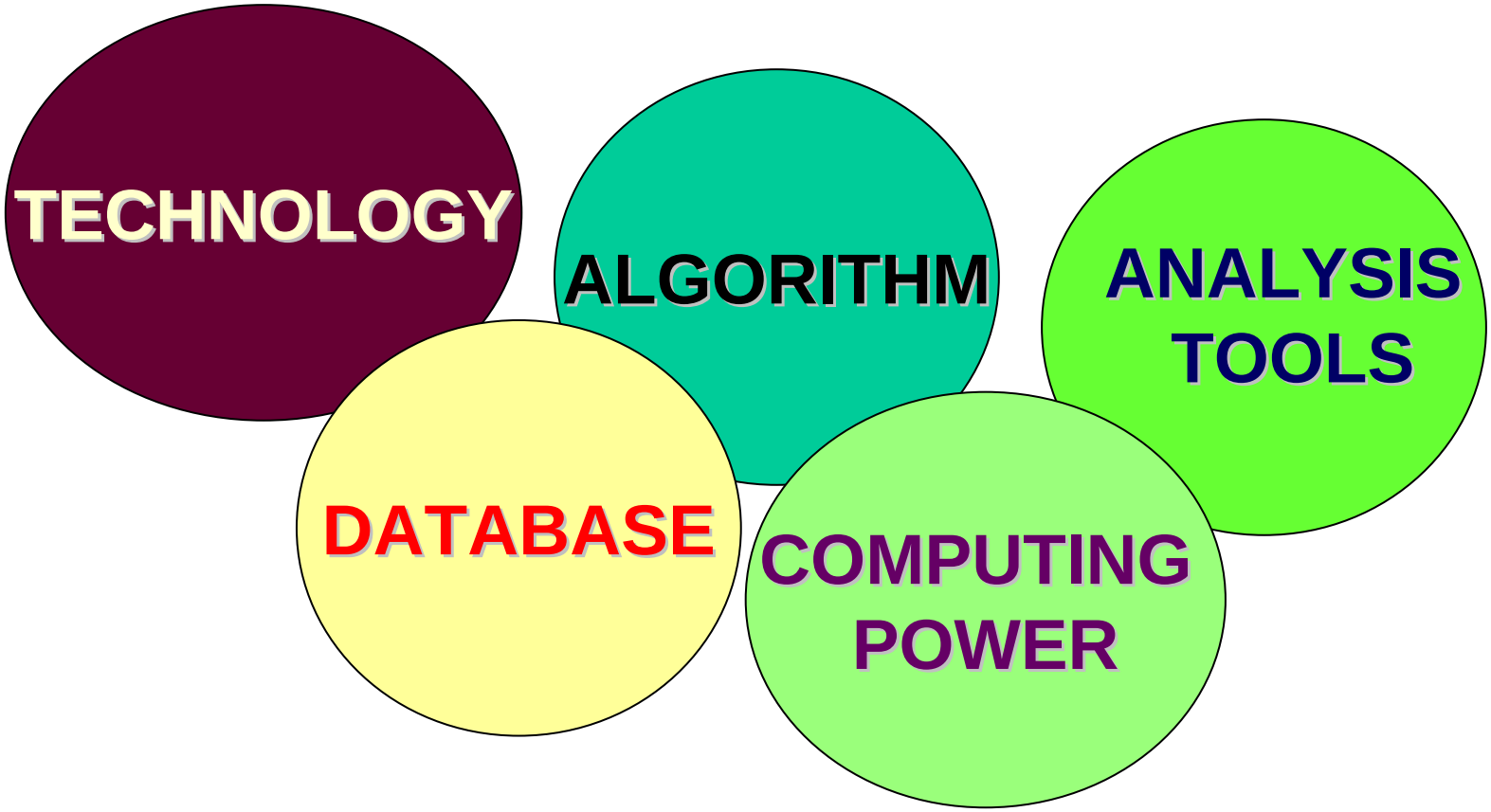
BioCarta
(*http://biocarta.com*)

Kyto Encyclopedia of Genes &Genomes
http://www.genome.ad.jp/kegg/

The Cancer Genome Anatomy Project (CGAP) *http://cgap.nci.nih.gov/*



THE COMPONENTS OF BIOINFORMATICS





BIOINFORMATICS ANALYSIS TOOLS

- $ Vector NTI suite, Omiga, DNAsis
- $ Staden Package, EMBOSIS, BLAST, FASTA

On line analysis tools



NHRI 生物資訊 Bioinformatics

國家衛生研究院巨分子序列分析服務

巨分子序列分析服務 GCG

在 Unix 系統下以 Command Mode 進行核酸或蛋白質的序列分析。
(telnet://bioinfo.nhri.org.tw)

巨分子序列分析服務 SeqWeb

連線至 SeqWEB 以瀏覽器進行核酸或蛋白質的序列分析。
(http://bioinfo.nhri.org.tw/)

EMBOSS

連線至 SeqWEB 以瀏覽器進行核酸或蛋白質的序列分析
(http://srs.nchc.org.tw/EMBOSS/)

Smith-Waterman 快速序列搜尋系統 GenWEB

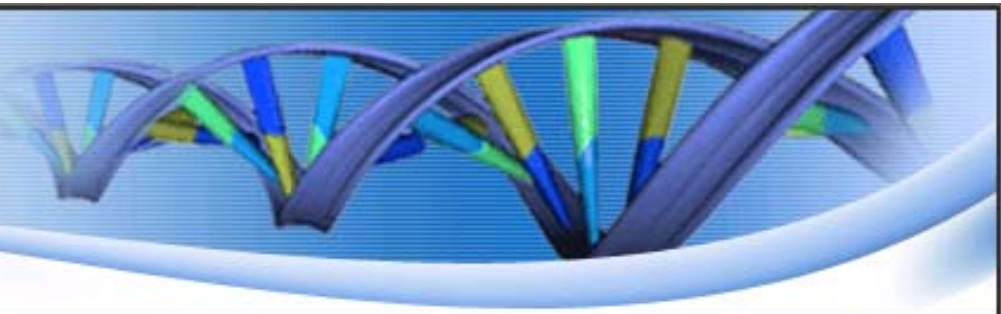
直接連線至 GenWeb 以瀏覽器進行核酸或蛋白質的快速序列搜尋。以特殊設計的硬體加速序列搜尋的速度，可進行 Smith-Waterman 及 FrameSearch 等搜尋功能。
(http://sw.nhri.org.tw/cgi-bin/genweb/bin/login.cgi)

ExPASy (Expert Protein Analysis System)

連線至 ExPASy 以瀏覽器進行蛋白質的序列分析。
(http://tw.expasy.org)



國家衛生研究院
TBI生物資訊核心設施



SEARCH

Go!!

powered by Google 站內搜尋

- GCG
- EMBOSS
- BLAST
- FTP Server
- PDA
- Discussion group
- Site Map
- GMBD Core TBI

HOME

最新消息 | 分析工具 | 資料庫 | 課程訊息 | 常見問題 | 帳號申請 | 全球網站

最新消息

國家衛生研究院生物資訊服務，提供**新版 Wisconsin Package**：(含 **GCG 11.1**與 **SeqWeb 3.1**) 上線使用，詳情請參考 新版公告。(公佈日期 2006/07/03)

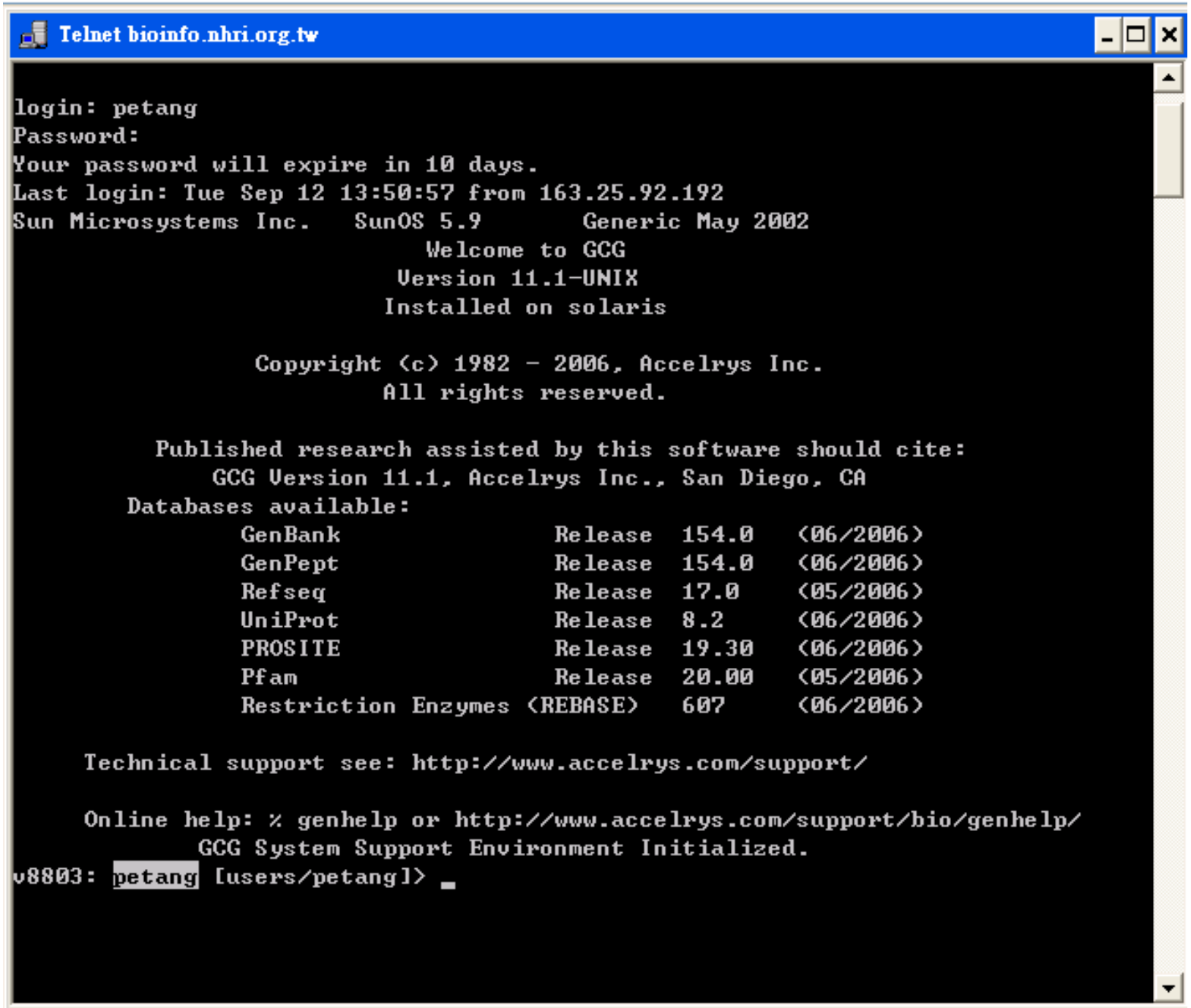
國家衛生研究院「研究資源2006系列活動」，將在 9/20-21 於「中研院生命科學圖書館電腦教室」，舉辦 **JEMBOSS基礎課程**，報名截止，請見學員名單！ (公佈日期 2006/09/18)

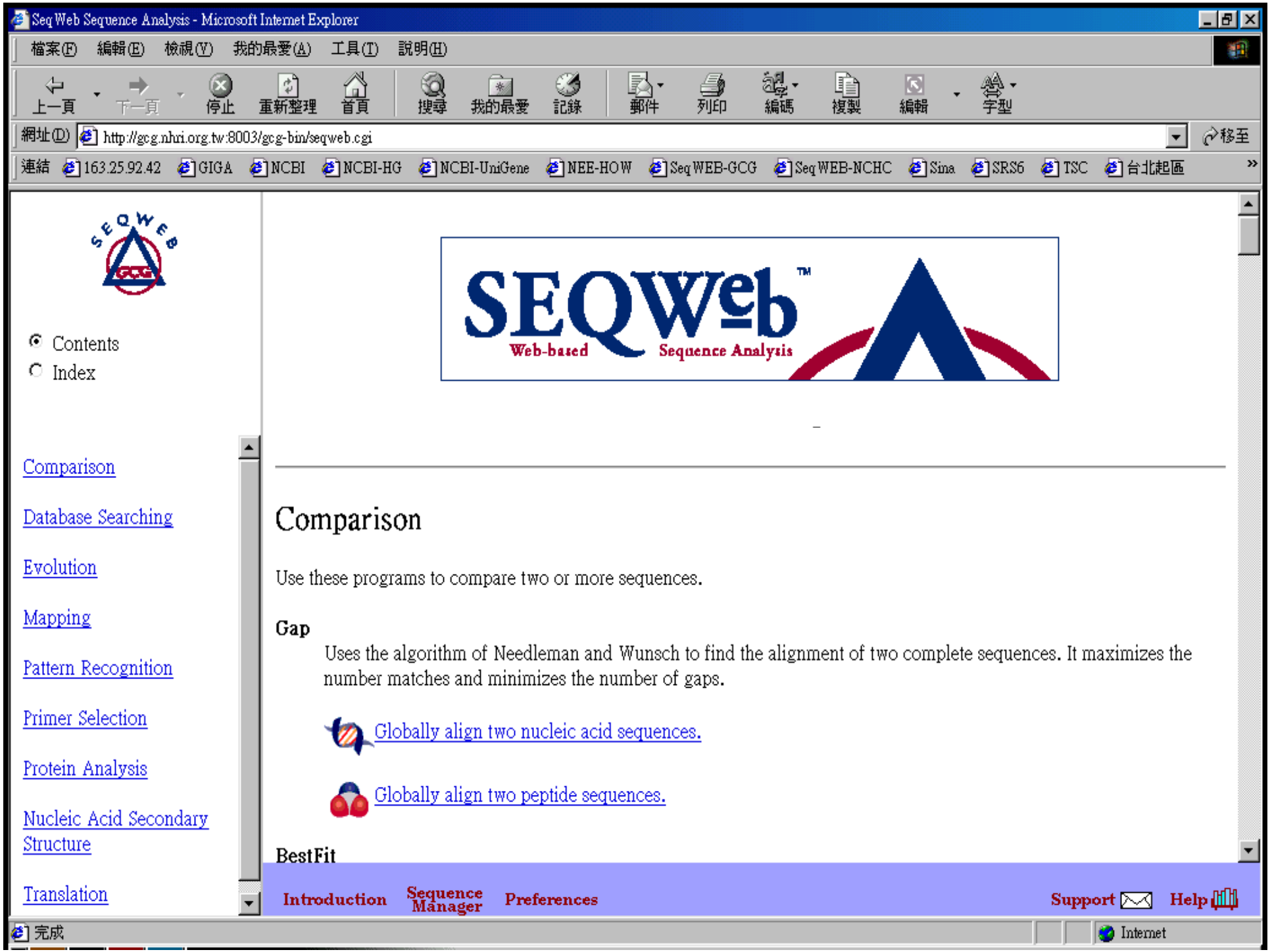
「**基因體研究之分析方法研習會**」，影音內容上線，請進入議程網頁後點選**Video圖示**，即可線上觀看演講內容！ 本項活動由基因體醫學國家型科技計畫支持。 (公佈日期 2006/08/31)

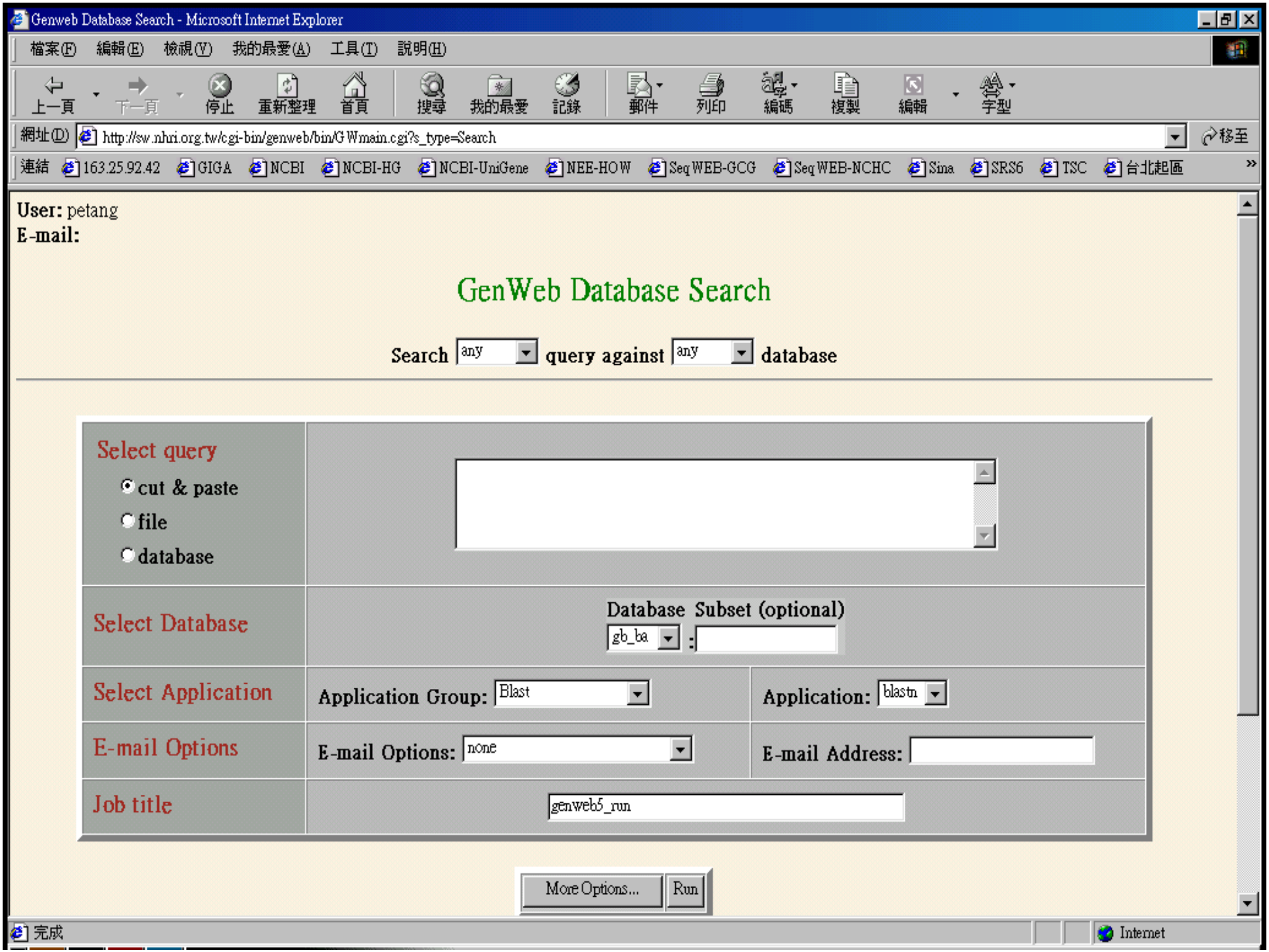
國家衛生研究院 GCG 程式及資料庫更新至 **2006-4** 版！請見詳細內容！(公佈日期 2006/08/30)

國立中興大學、國立陽明大學、教育部顧問室、國科會生命科學推動中心、生物資訊與系統生物學教學資源中心、與台灣生物資訊學會，將於 **2006年9月12日至15日**，假**「中興大學圖書館國際會議廳」**舉辦 2006台灣生物資訊研討會 (Symposium and Workshop of Bioinformatics in Taiwan 2006)， 請見大會網址！(公佈日期 2006/08/16)

GMBD 基因體醫學生技研發生物資訊核心教育訓練課程，將在 8/14-15 於「中研院生命科學圖書館電腦教室」，舉辦 **巨分子序列分析基礎課程**，學員名單公佈，請參考名單

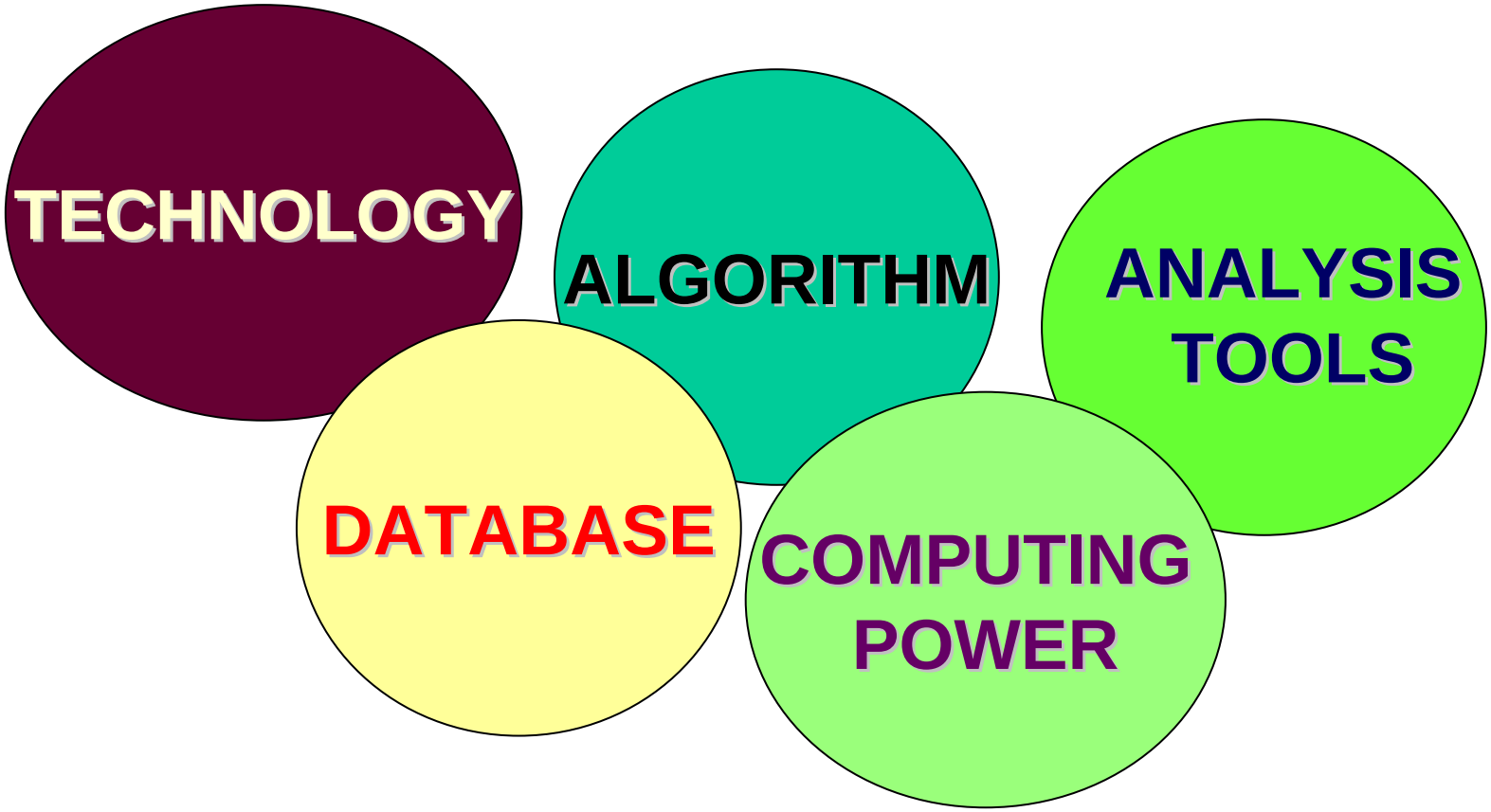

Telnet bioinfo.nhri.org.tw
login: petang
Password:
Your password will expire in 10 days.
Last login: Tue Sep 12 13:50:57 from 163.25.92.192
Sun Microsystems Inc.   SunOS 5.9   Generic May 2002
Welcome to GCG
Version 11.1-UNIX
Installed on solaris
Copyright (c) 1982 - 2006, Accelrys Inc.
All rights reserved.
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]>


Seq Web Sequence Analysis - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 下一頁 停止 重新整理 首頁 搜尋 我的最愛 記錄 郵件 列印 編碼 複製 編輯 字型
網址(D) http://gcg.nhri.org.tw:8003/gcg-bin/seqweb.cgi
移至
連結 163.25.92.42 GIGA NCBI NCBI-HG NCBI-UniGene NEE-HOW Seq WEB-GCG Seq WEB-NCHC Sina SRS6 TSC 台北起區
SEQWeb
GCG
Contents
Index
Comparison
Database Searching
Evolution
Mapping
Pattern Recognition
Primer Selection
Protein Analysis
Nucleic Acid Secondary Structure
Translation
SEQWeb™
Web-based Sequence Analysis
Comparison
Use these programs to compare two or more sequences.
Gap
Uses the algorithm of Needleman and Wunsch to find the alignment of two complete sequences. It maximizes the number matches and minimizes the number of gaps.
Globally align two nucleic acid sequences.
Globally align two peptide sequences.
BestFit
Introduction
Sequence Manager
Preferences
Support
Help
完成
Internet


Genweb Database Search - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 下一頁 停止 重新整理 首頁 搜尋 我的最愛 記錄 郵件 列印 編碼 複製 編輯 字型
網址(D) http://sw.nhri.org.tw/cgi-bin/genweb/bin/GWmain.cgi?s_type=Search
移至
連結 163.25.92.42 GIGA NCBI NCBI-HG NCBI-UniGene NEE-HOW SeqWEB-GCG SeqWEB-NCHC Sina SRS6 TSC 台北起區
User: petang
E-mail:
GenWeb Database Search
Search any query against any database
Select query
cut & paste
file
database
Select Database
Database Subset (optional)
gb_ba :
Select Application
Application Group: Blast
Application: blastn
E-mail Options
E-mail Options: none
E-mail Address:
Job title
genweb5_run
More Options...
Run
完成
Internet



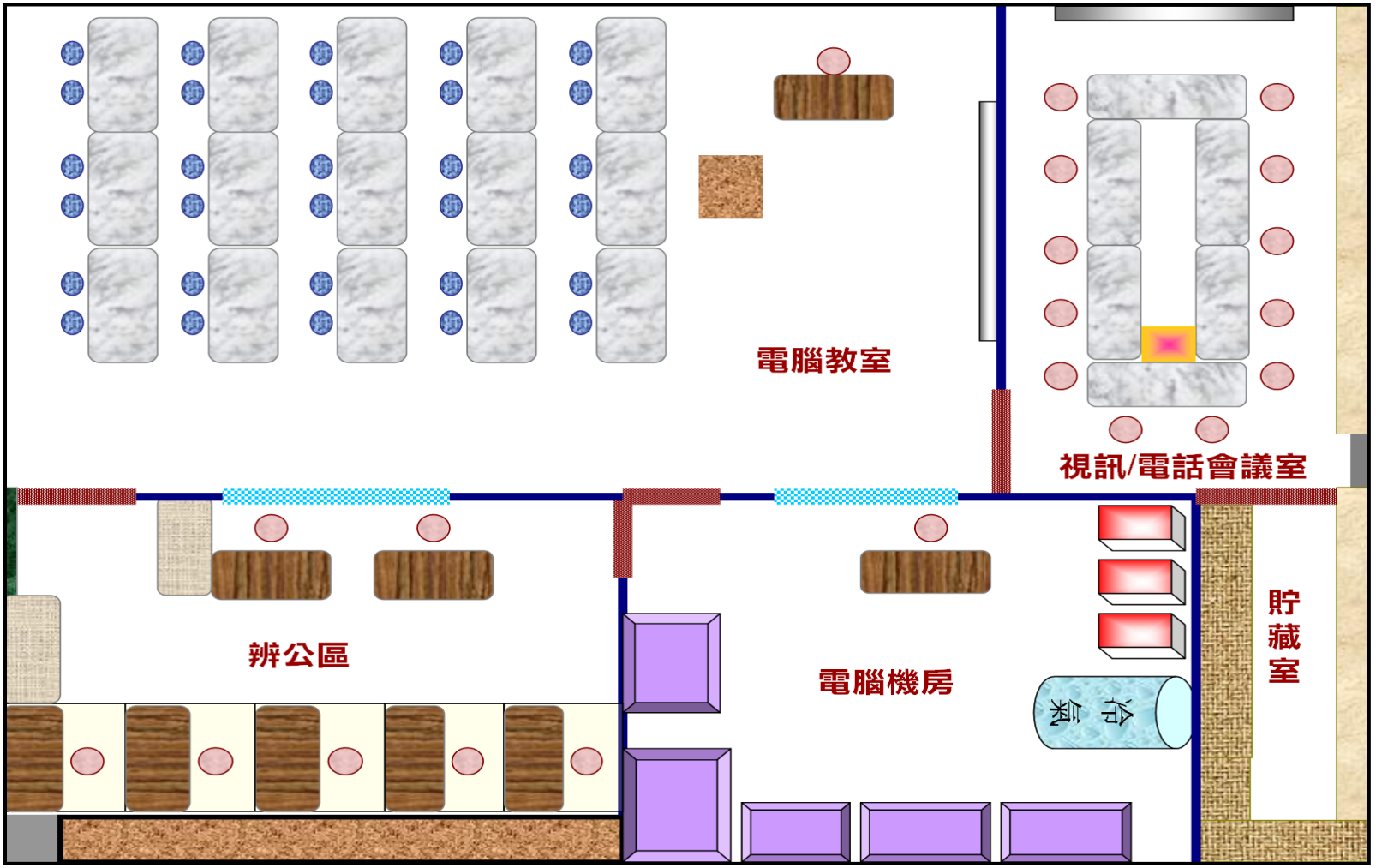
THE COMPONENTS OF BIOINFORMATICS



長庚生物資訊中心
bioinfo.cgu.edu.tw

設備

醫學大樓9樓0917



長庚生物資訊中心

bioinfo.cgu.edu.tw

設備

COMPUTER	CPU	NO.	MEMMORY
SunFire 6800	Sparc 750 MHz	24	48 GB
Sun V60 Cluster	Xeon 2.8 GHz	20	20 GB
IBM X336 Cluster	Xeon 3.2 GHz	14	14 GB
IBM X225 Cluster	Xeon 2.4 GHz	2	1.5GB
HP DL580G3 Cluster	Xeon 3.0 GHz	16	16 GB
LunuxWorX Cluster	Xeon 2.4 GHz	8	8 GB
IBM Z-pro Graphic Station	Xeon 3.2 GB x 2	2	3 GB
教學電腦	P4 2.4 GHz	15	512 MB
教學電腦	P4 3.2 GHz	15	1 GB

ITEMS	SPECIFICATION	NO
Proware RAID System	250 GB x 16 (4 TB)	1
Petastor Fibre RAID System	400 GB x 16 (6.4 TB x 4)	4
Proware NAS System	80 GB x 8 (640 GB)	1
Brocad silkworm 2G Fibre switch	12 ports	1
UPS	10 KVA	1
UPS	30 KVA	2
Video Conference System	Centura	50
Telephone Conference System	Polycom sound station	1

長庚生物資訊中心
bioinfo.cgu.edu.tw

設備

InforMax™ Invitrogen™ life science software [*Vector NTI Advanced Server*]
[*GENOMAX High-Throughout Sequence Analysis System*]

PARACEL [*Paracel BLAST*] [*Paracel TranscriptAssembler*]

[*Bioinformatics Linux Cluster*]
[*Expression Sequence Tag Analysis Pipeline*]

[*Protein Sequence Analysis Pipeline*]

Tripos *[Protein Modeling & Docking System]*
[Lead Compound Database]



emboss [*The European Molecular Biology Open Software Suite*]

LION [*Sequence Retrieval System*]

Gene *[MetaCore: PPI Network]*

Genedata solutions in-silico *[Expressionist]*



Steps to Identify a Gene

- Gene-Search
- Protein-Search
- Annotation





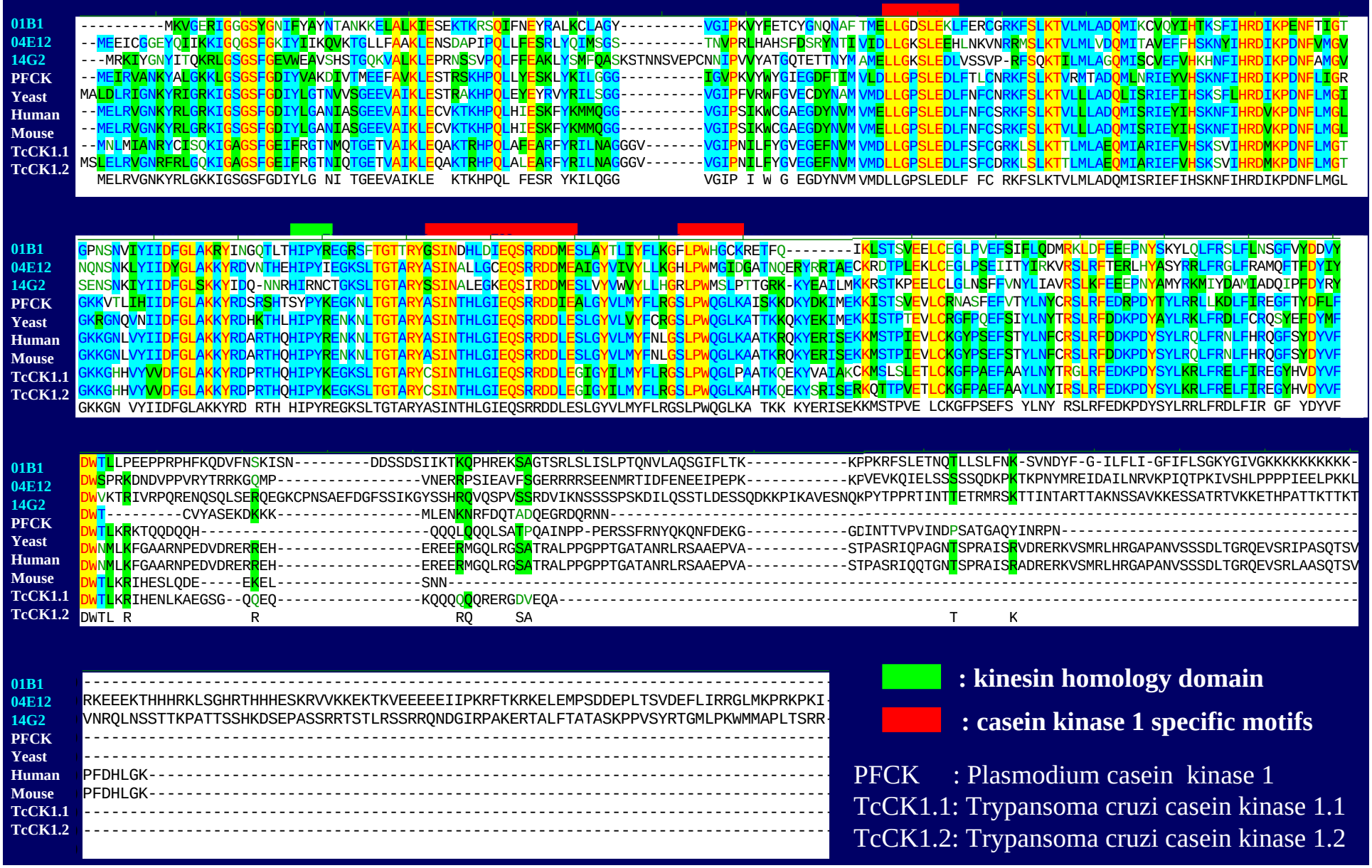




Full length ORF of *TvEST-14G2*

```
  -2          …AGATGCGAAAAA TCTACGGCAA TTACATTACG CAGAAGCGTC TCGGTTCAGG
                 AAGTTTCGGA GAGGTTTGGG AAGCTGTCAG TCATTCGACC GGACAAAAGG
 101             TTGCTCTCAA ATTAGAGCCC CGAAACTCTA GTGTTCCACA ATTATTTTTC
                 GAAGCCAAGC TATACTCAAT GTTTCAGGCT TCAAAATCCA CAAATAATAG
 201             TGTAGAACCA TGCAACAACA TTCCAGTTGT TTATGCGACT GGTCAAACAG
                 AGACAACTAA CTACATGGCC ATGGAATTAC TTGGCAAGTC TCTGGAAGAT
 301             TTAGTTTCAT CGGTCCCTAG ATTTTCCCAA AAGACAATAT TAATGCTTGC
                 CGGACAAATG ATTTCCTGTG TTGAATTCGT TCACAAACAT AATTTTATTC
 401             ACCGCGACAT CAAGCCAGAT AATTTTGCGA TGGGAGTCAG TGAGAACTCA
                 AACAAAATTT ATATTATCGA TTTTGGACTT TCCAAGAAGT ACATTGACCA
 501             AAATAATCGT CATATTAGAA ATTGCACAGG AAAATCACTT ACCGGAACCG
                 CAAGATATTC ATCAATTAAT GCGCTCGAAG GAAAGGAACA GTCTATAAGA
 601             GATGACATGG AATCTTTGGT ATATGTCTGG GTTTATTTAC TTCATGGACG
                 TCTTCCTTGG ATGAGCTTAC CTACAACAGG CCGCAAGAAG TATGAGGCCA
 701             TTTTAATGAA GAAGAGATCA ACGAAACCCG AAGAATTATG TTTAGGACTT
                 AATAGTTTCT TTGTAAACTA CTTAATAGCA GTTCGCTCAT TGAAATTTGA
 801             AGAAGAACCA AATTACGCGA TGTACAGGAA AATGATATAC GACGCAATGA
                 TTGCTGATCA AATTCCTTTT GATTATCGCT ATGATTGGGT CAAAACGAGA
 901             ATTGTTCGCC CACAACGTGA AAACCAATCA CAGTTGTCCG AACGTCAAGA
                 AGGAAAATGT CCAAACTCAG CTGAGTTTGA TGGTTTCTCC TCCATCAAAG
1001             GATATTCTTC GCACAGACAA GTACAAAGCC CCGTTTCATC TAGAGATGTC
                 ATTAAGAACA GTAGTTCAAG TCCATCAAAG GATATTTTGC AATCATCAAC
1101             CCTTGATGAA TCATCTCAAG ATAAAAAGCC AATCAAAGCT GTCGAATCGA
                 ATCAGAAACC ATATACACCG CCACGTACAA TTAATACTAC CGAAACAAGA
1201             ATGAGATCAA AGACTACAAT CAATACTGCA AGAACAACAG CAAAGAACTC
                 TTCGGCAGTT AAGAAAGAAT CGTCAGCAAC AAGGACTGTT AAGAAAGAAA
1301             CACATCCTGC AACTACAAAA ACAACAAAAA CTGTAAATAG ACAATTGAAC
                 TCTTCTACAA CGAAACCGGC AACTACGAGC TCTCACAAAG ACTCAGAACC
1401             GGCTTCATCA AGACGTACAT CAACTCTACG TTCAAGTCGC CGCCAAAATG
                 ACGGAATTCG CCCTGCAAAG GAAAGAACTG CGCTTTTCAC AGCTACAGCC
1501             AGTAAGCCTC CGGTATCTTA CCGTACTGGA ATGCTTCCGA AATGGATGAT
                 GGCTCCTCTC ACATCTCGTC GCTGAAATAT ATTTTTTATA TTATTTATTT
1601             TTTTCTTTTT CTATCTGTAT ATTAAATGTA TTTCTATATT ATTAAAAAAA
```

Amino Acid Sequence Comparison



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



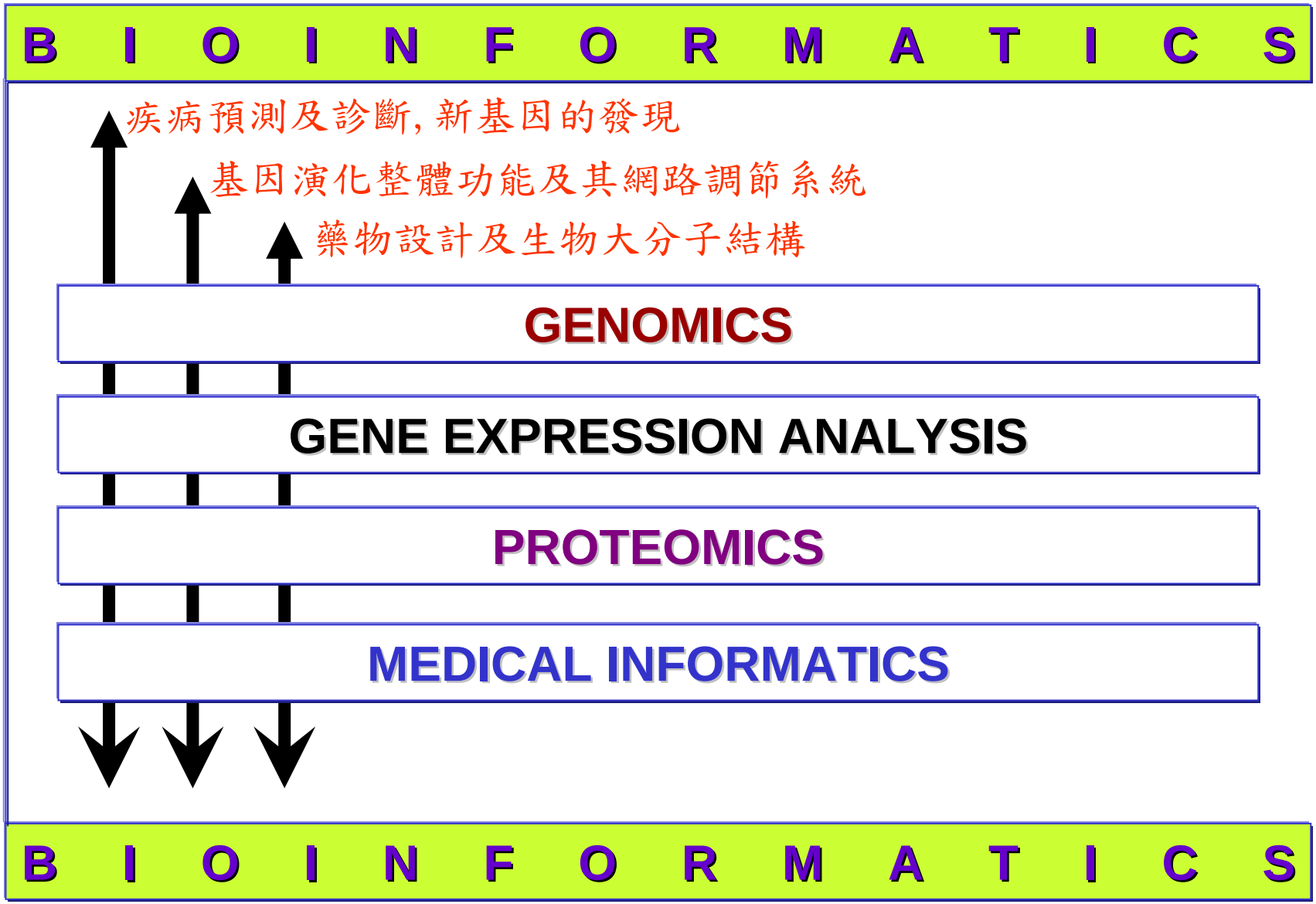
Yeast CK1



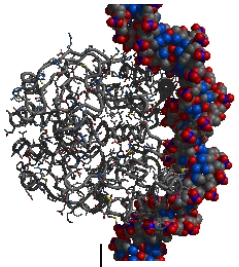
Mouse CK1



Human CK1-δ


B I O I N F O R M A T I C S
疾病預測及診斷, 新基因的發現
基因演化整體功能及其網路調節系統
藥物設計及生物大分子結構
GENOMICS
GENE EXPRESSION ANALYSIS
PROTEOMICS
MEDICAL INFORMATICS
B I O I N F O R M A T I C S

Focuses in Bioinformatics

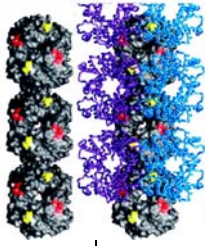


Perturbation
Environment
Medication
Genetic Engineering

Dynamic Response
Gene Expression
Protein Expression

BioChip

Virtual Cell



Analysis

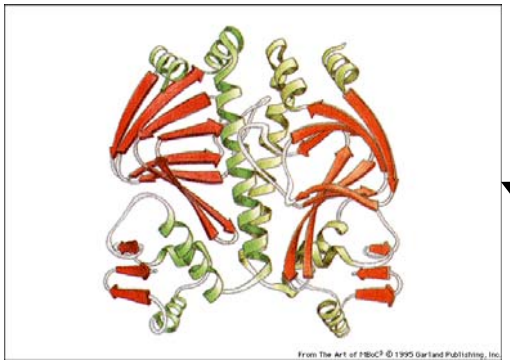
DataBase
Genotype/Phenotype

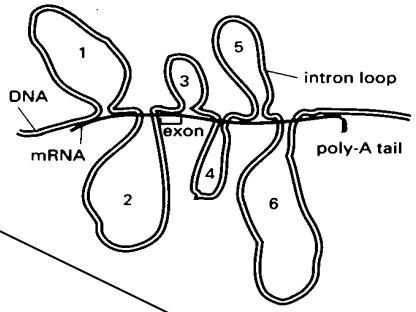
Biology
Molecular Biology
Bio Chemistry
Genetics

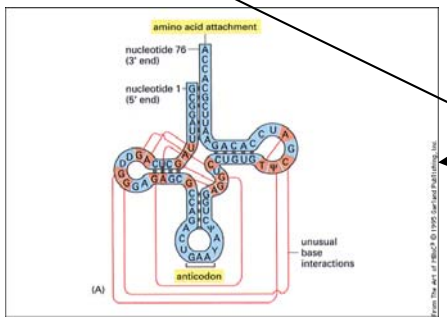
Symbolic
Algorithms/
Computing

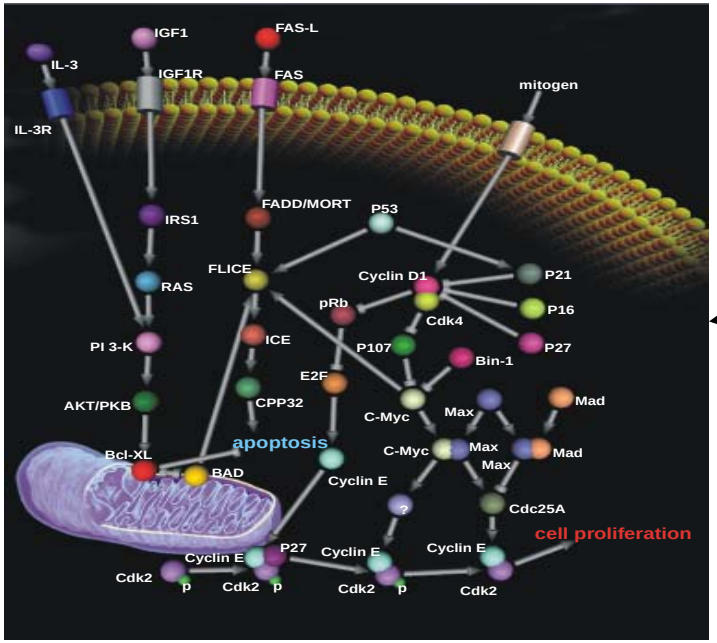
Genome Sequencing

Goals Leading Toward Predictive Biology









Gene Sequence Data

Gene Identification

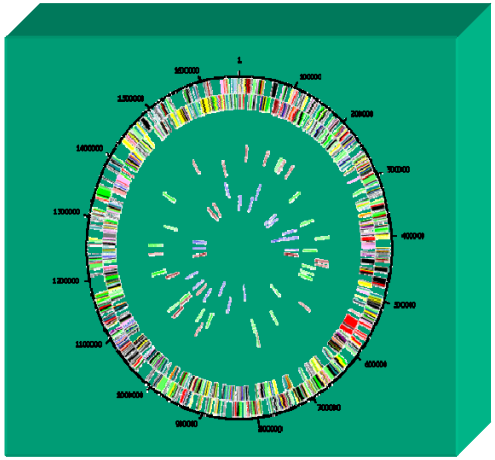
Structure Prediction

Protein Circuit & Regulatory Network Discovery

Biosimulation

Reconstructing Cellular Functions





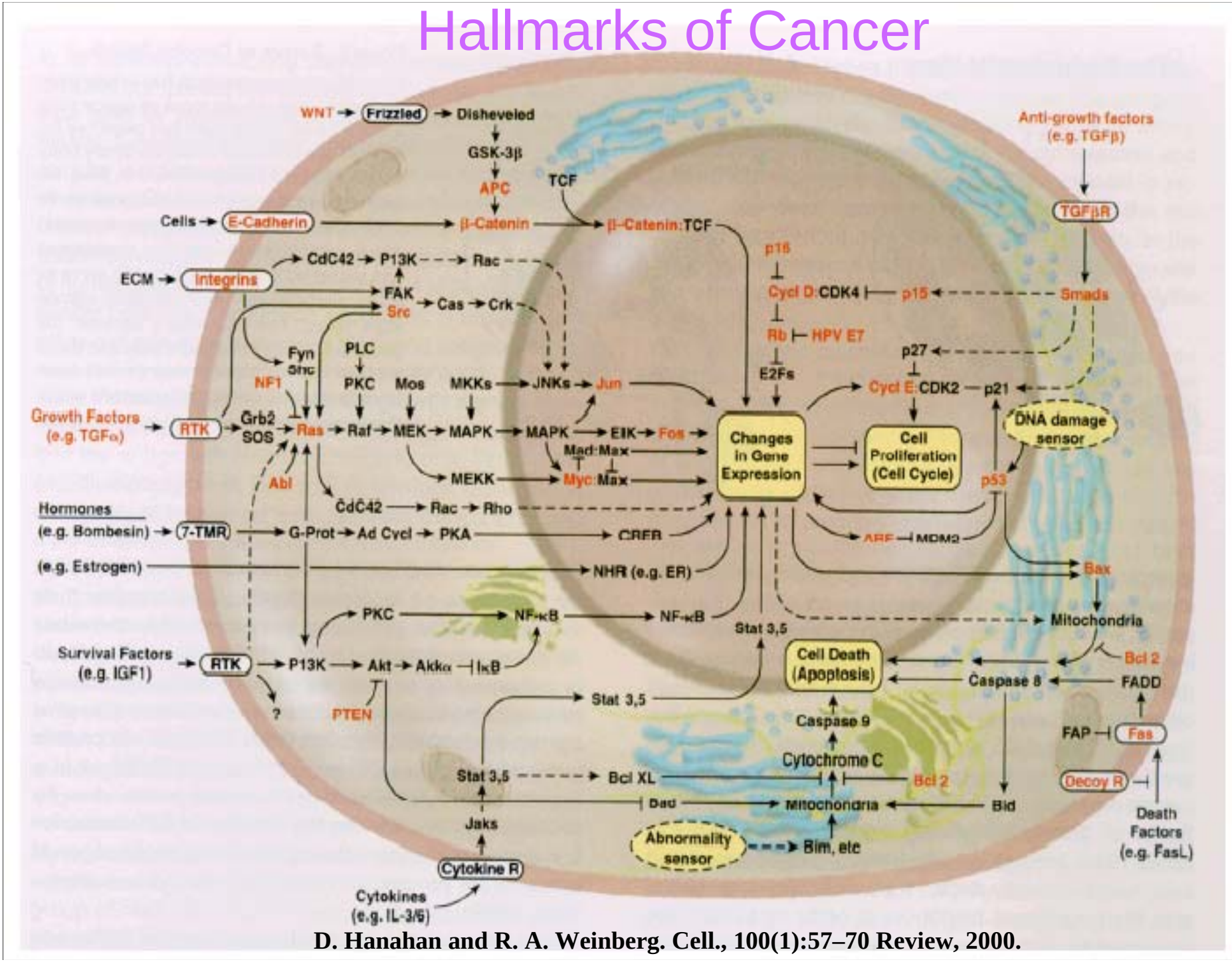
Reductionistic Approach
(Genome Sequencing, DNA arrays, proteomics)

Integrative Approach
(Bioinformatics, Systems Science, modeling & simulation)

20th Century Biology

21th Century Biology

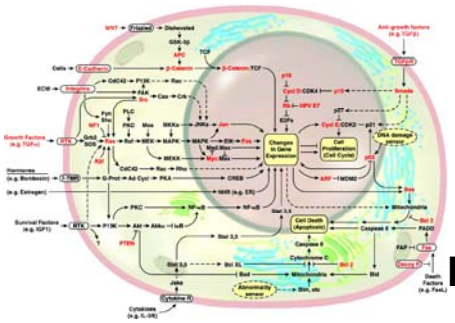
Hallmarks of Cancer



D. Hanahan and R. A. Weinberg. Cell., 100(1):57–70 Review, 2000.

Platform for Systems Biology

- Objective is to link gene response, protein activity, metabolite dynamics to disease and interventions



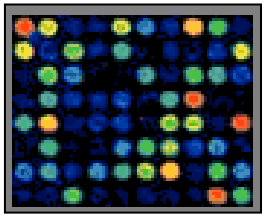
Complex Cellular Samples
bodyfluids, tissue

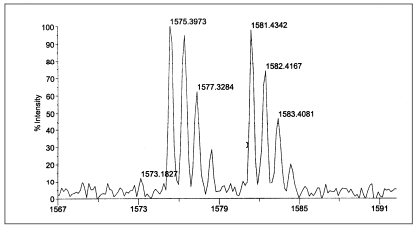
Dynamics
i.e. environmental + time

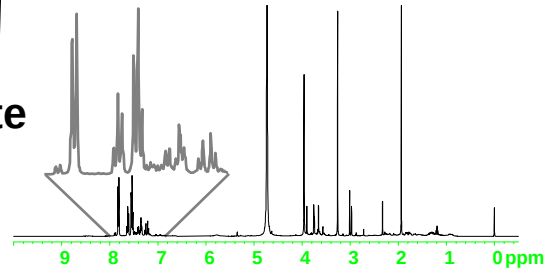
Gene

Protein

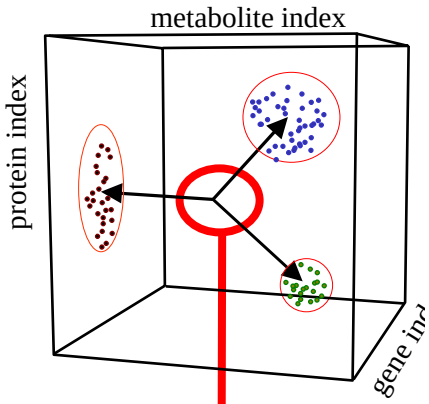
Metabolite







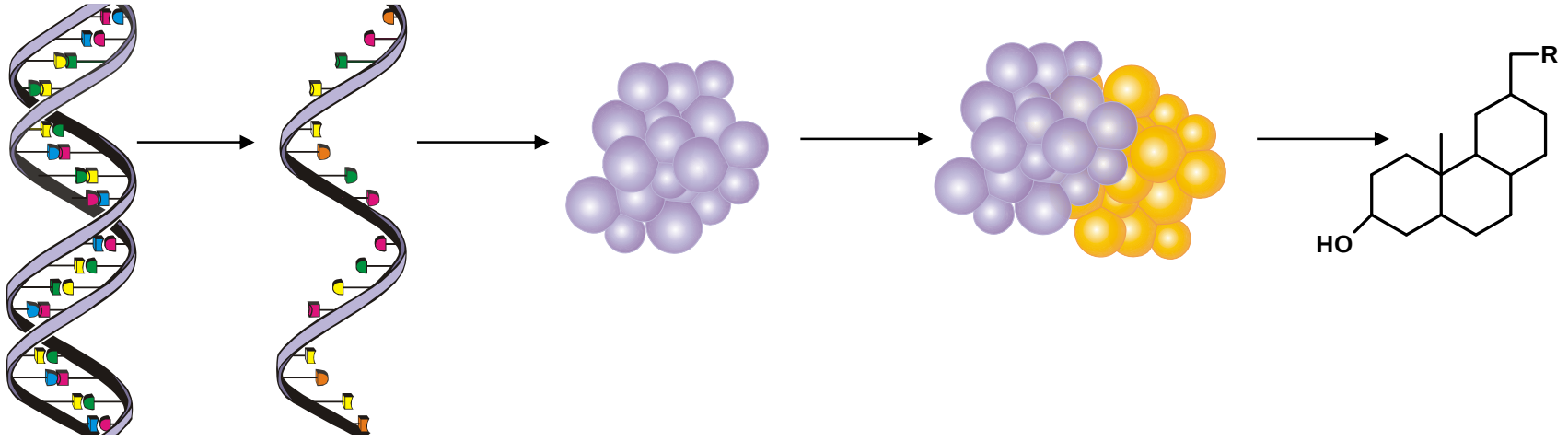
Quantitative Comparisons



BioSystematics™

Targets Biomarkers

SYSTEMS BIOLOGY



Genomics

Proteomics

Metabolomics

Transcriptomics

Functional Proteomics/Genomics

Systems Biology

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.