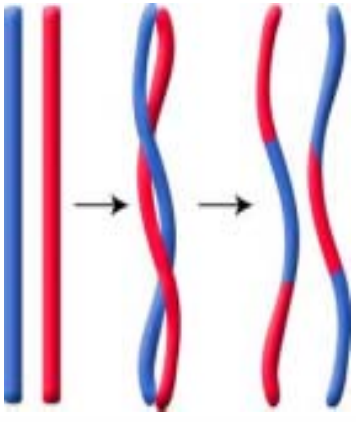
http://pastime.cgu.edu.tw/petang/index.htm

ftp://163.25.92.192 DOWNLOAD Bioinfo92-1.exe to your desktop
Open bioinfo92-1-01.htm



Bioinformatics 92-1

Lecture 1 – Introduction to Bioinformtics

AGTCCGCGAATACAGGCTCGGTAGTCCGCGAATACAGGCTCGGT

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

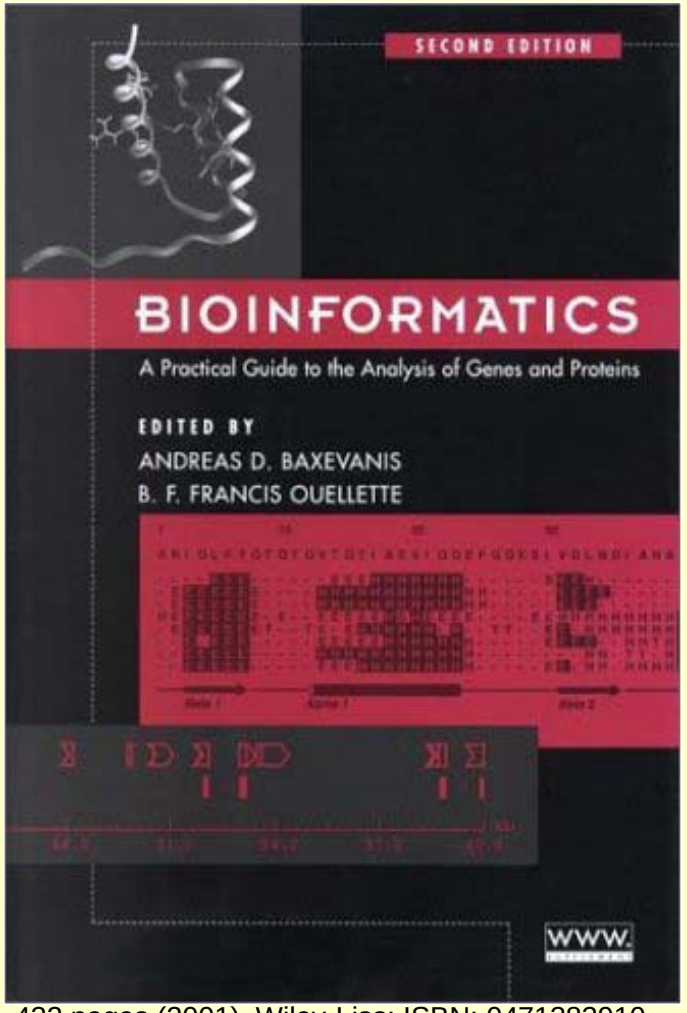
助教：
王琇儀 (分機5115)
鄭羽昕 (分機5690)

19th September 2003

選修本課程必需先申請
「國家衛生研究院巨分子序列分析服務帳號」
（http://bioinfo.nhri.org.tw/gcg/gcgform.htm）

週	日　期	課程進度 - 授課老師
1	2003-09-19	Introduction to Bioinformatics - 鄧致剛
2	2003-09-26	Unix, Html and WWW Basics - 鄧致剛
3	2003-10-03	Bioinformatics Databases: Database searching and Sequence Management - 鄧致剛
4	2003-10-10	HOLIDAY
5	2003-10-17	Comparing Sequences and Multiple Sequence Alignment - 鄧致剛
6	2003-10-24	RNA Secondarily Structure & Primer Design - 鄧致剛
7	2003-10-31	DNA Sequence Analysis - 林文昌
8	2003-11-07	DNA Fragment Assembly & Clustering - 鄧致剛
9	2003-11-14	Protein Sequence Analysis - 鄧致剛
10	2003-11-21	Protein 3D Structure & Modelling - 呂平江
11	2003-11-28	Phylogenetic Analysis - 劉信孚
12	2003-12-05	Special Topic-Bioinformatics on the Internet -鄧致剛
13	2003-12-12	Special Topic-Genome Analysis - 林文昌
14	2003-12-19	Special Topic-Gene Expression Analysis - 鄧致剛
15	2003-12-26	Special Topic-Gene Chip Technology & Microarray Analysis - 李御賢
16	2004-01-02	Special Topic-Proteome Analysis - 潘台龍
17	2004-01-09	Student Presentation - 鄧致剛
18	2004-01-16	Final Examination - 鄧致剛

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*

- 鄧致剛老師教學網
  (http://pastime.cgu.edu.tw/petang/index.htm)
- 國立陽明大學楊永正教授POST 系統
  (http://binfo.ym.edu.tw/post/)
- 國立清華大學呂平江教授Bioinformatics教學網
  (http://www.life.nthu.edu.tw/~lslpc/bioinfo.html)
- 國家衛生研究院巨分子序列分析服務使用說明
  (http://bioinfo.nhri.org.tw/)

BioInfoFab

http://blast.ym.edu.tw/indexEasy.php

生物引擎 www.bio-engine.com

華文生技網 bioweb.com.tw

BioSino *http://www.biosino.org/*

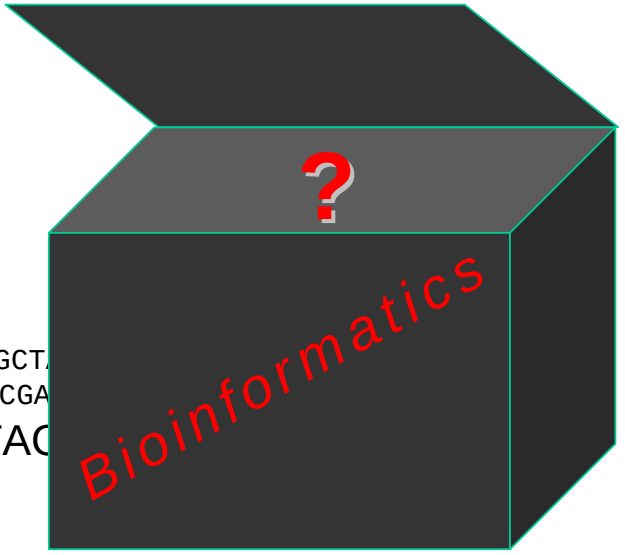
台灣生物資訊學會 Bioinformatics Society Taiwan

http://dblab8.csie.ncu.edu.tw/home.htm

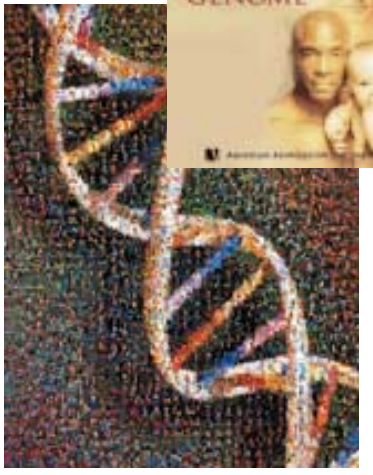
WHAT IS BIOINFORMATICS?

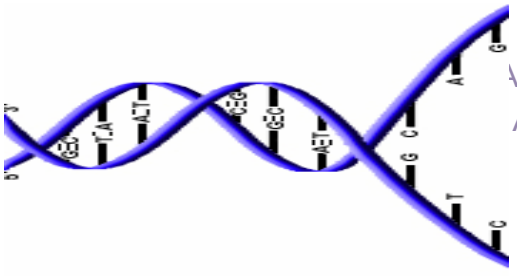


CTAGCTAGCTAGCT
CTAGCTAGCTAGCTATCGA
CATGCACTAGCTAC





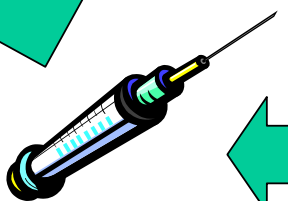


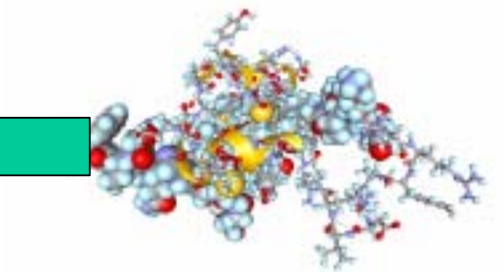


AAATGGCCAATTGATGACCAGAGATTTAGGCCAATTAA
AATGTGAAATGGCCAATTGATGACCAGAGA









What is Bioinformatics?

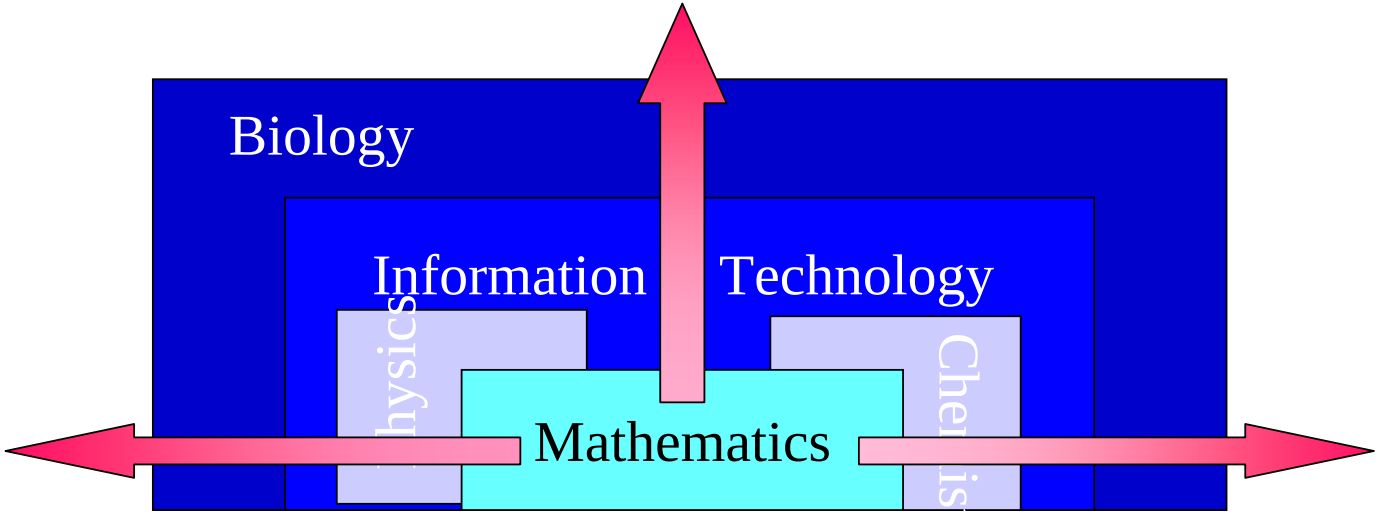
The answer to this question depends on whether you are talking to

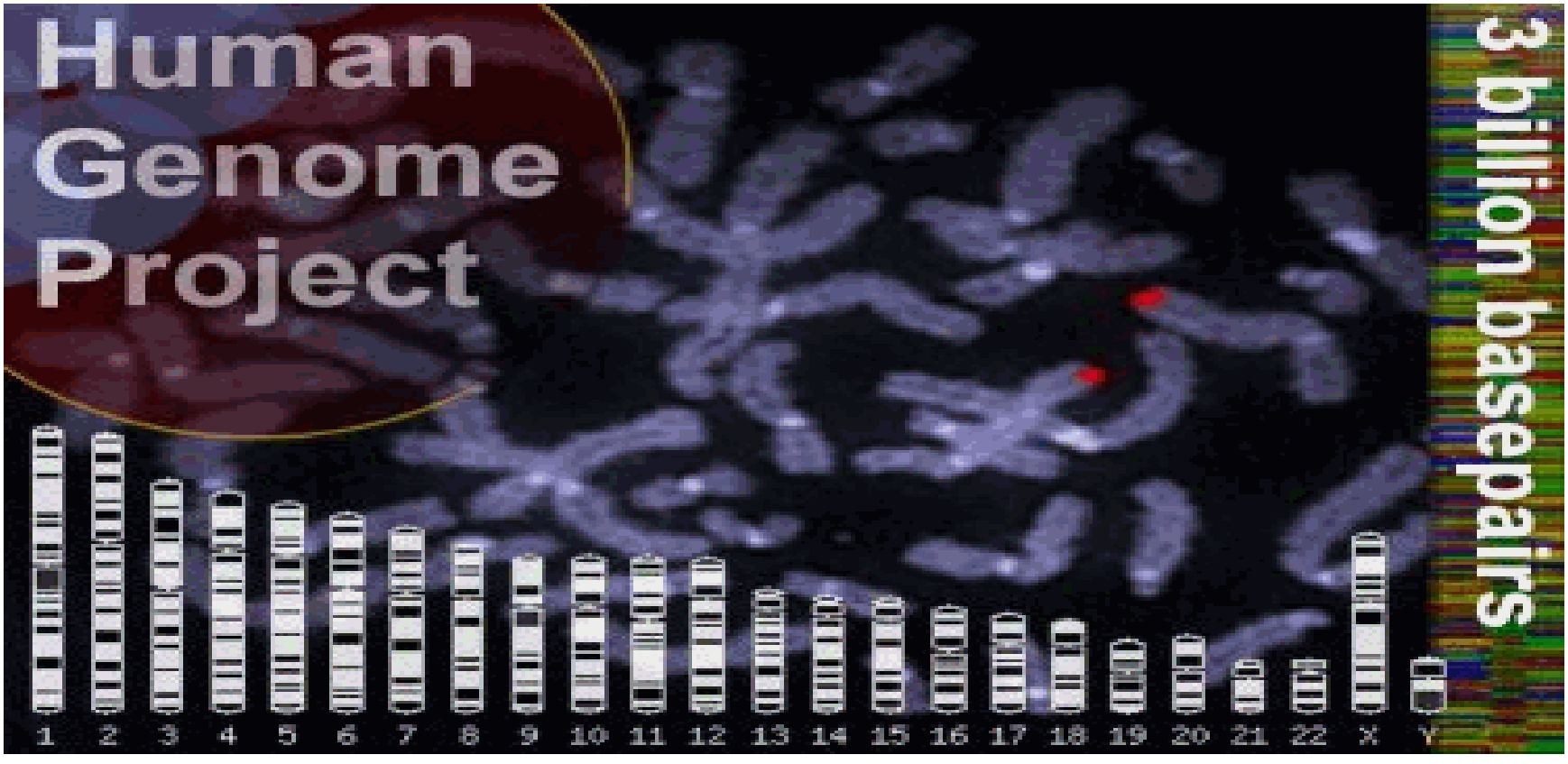
A computer scientist who 'does' biology, or

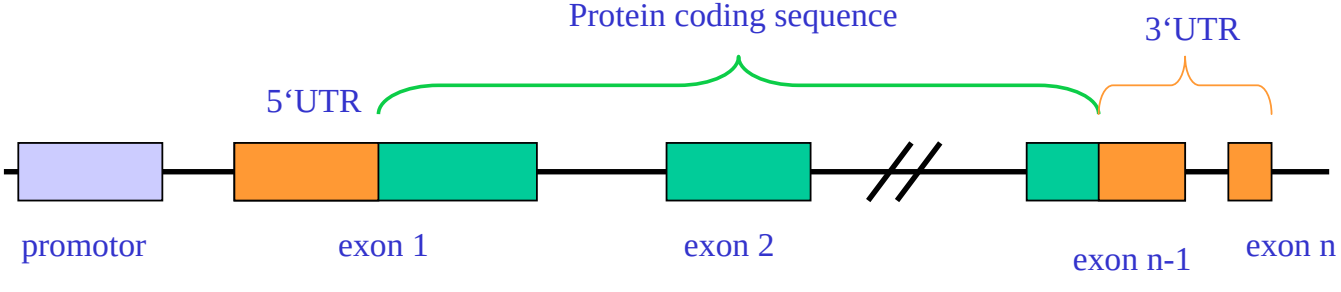
A molecular biologist who 'does' computing.

Bioinformatics is the application of computer technology to the management of biological information. Computers are used to gather, store, analyze and integrate biological and genetic information which can then be applied to gene-based drug discovery and development.

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。

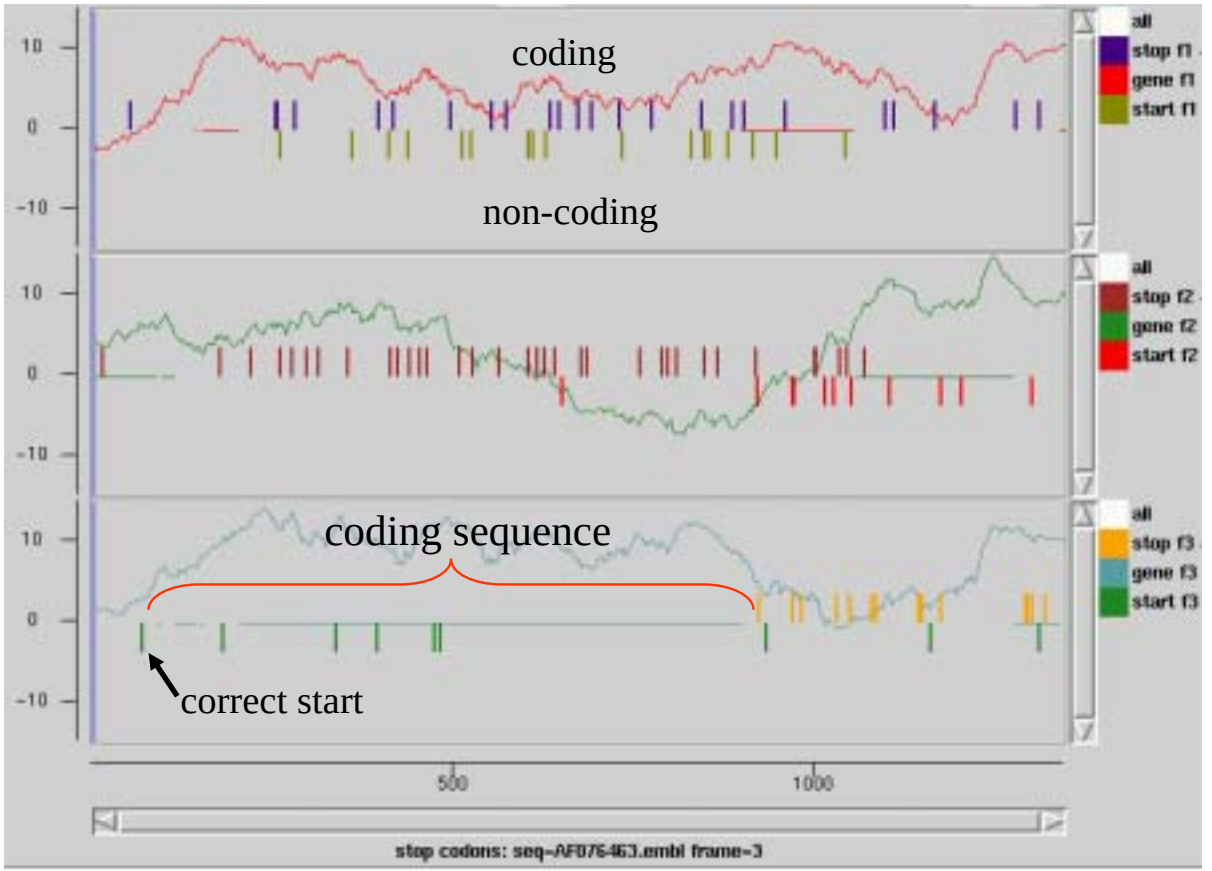



Human
Genome
Project
3 billion basepairs
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y


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

Gene prediction

Codon usage (single exon)



Gene prediction

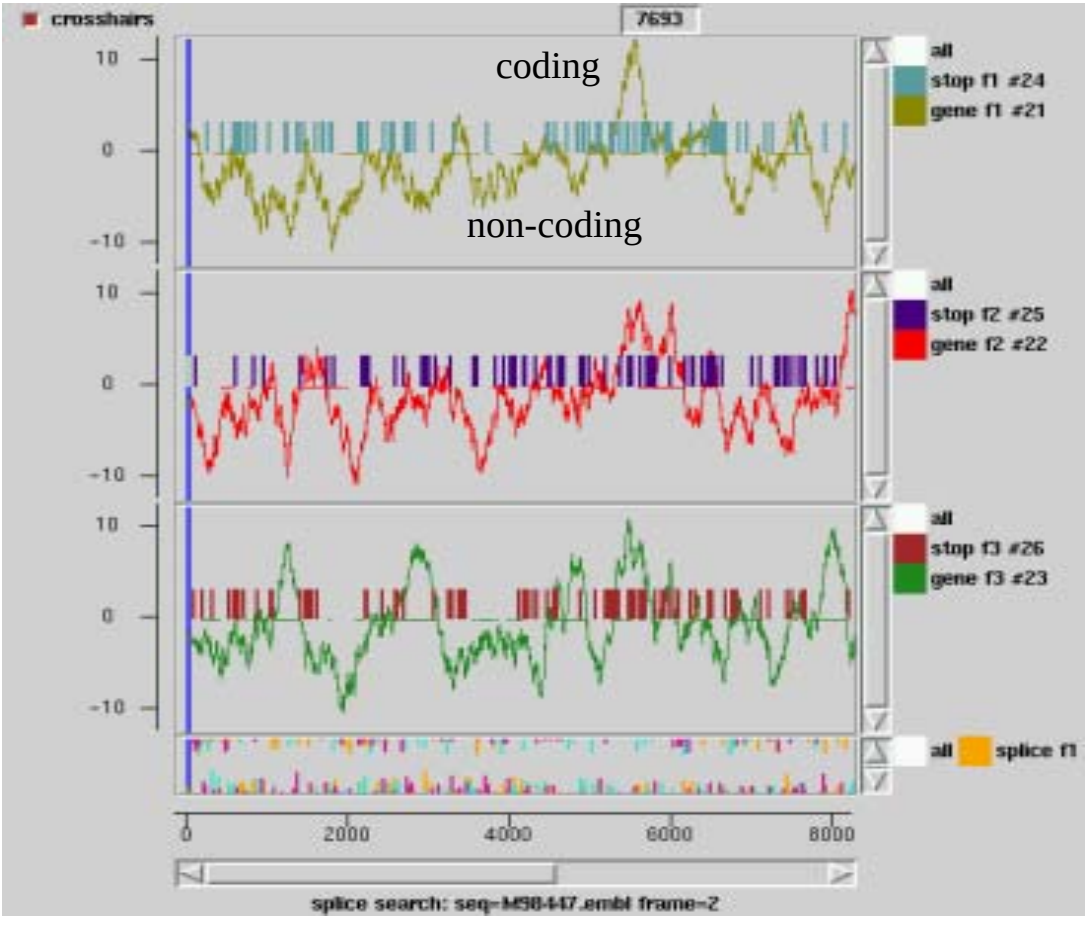
Codon usage (multiple exons)

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

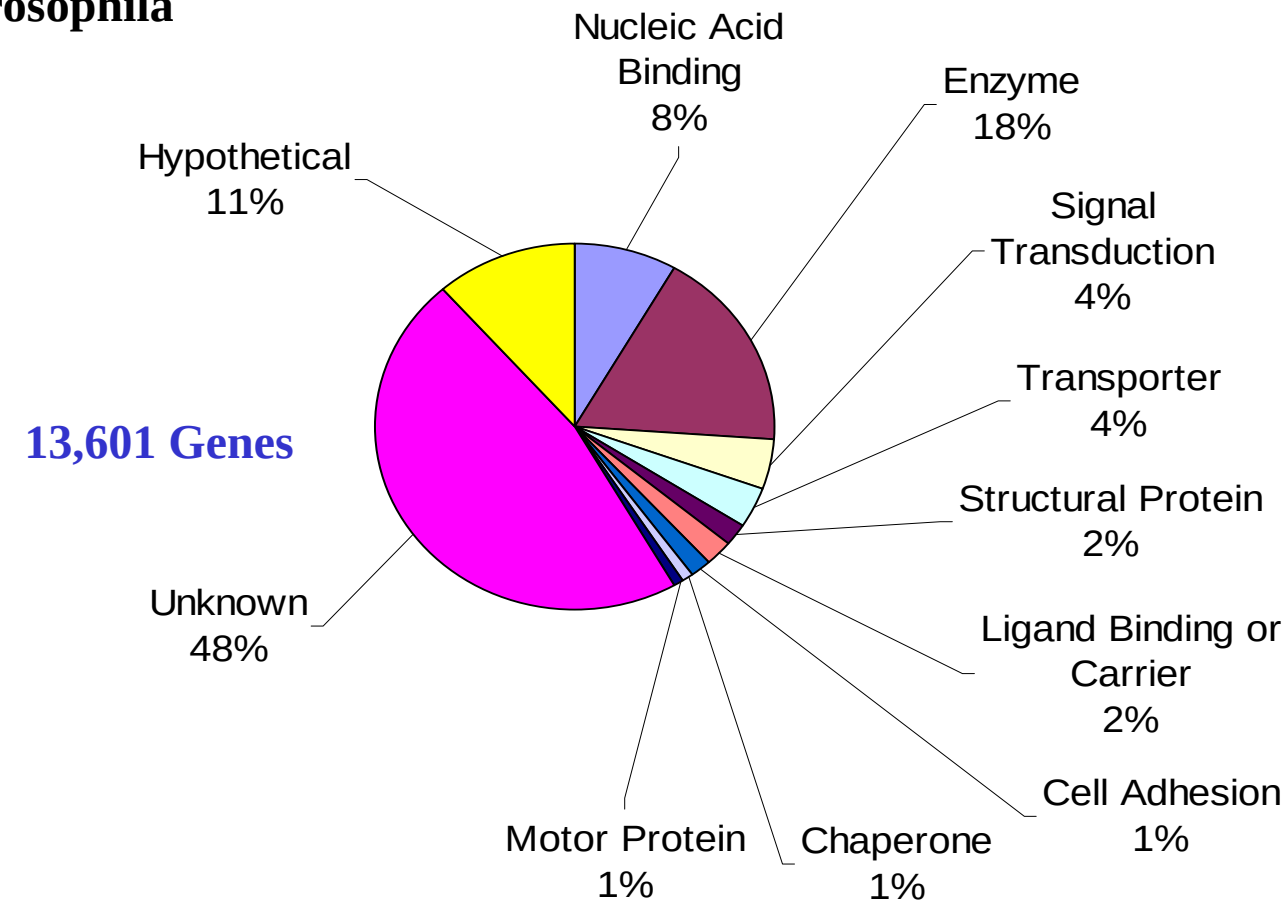
8124. .8405

8553. .8713

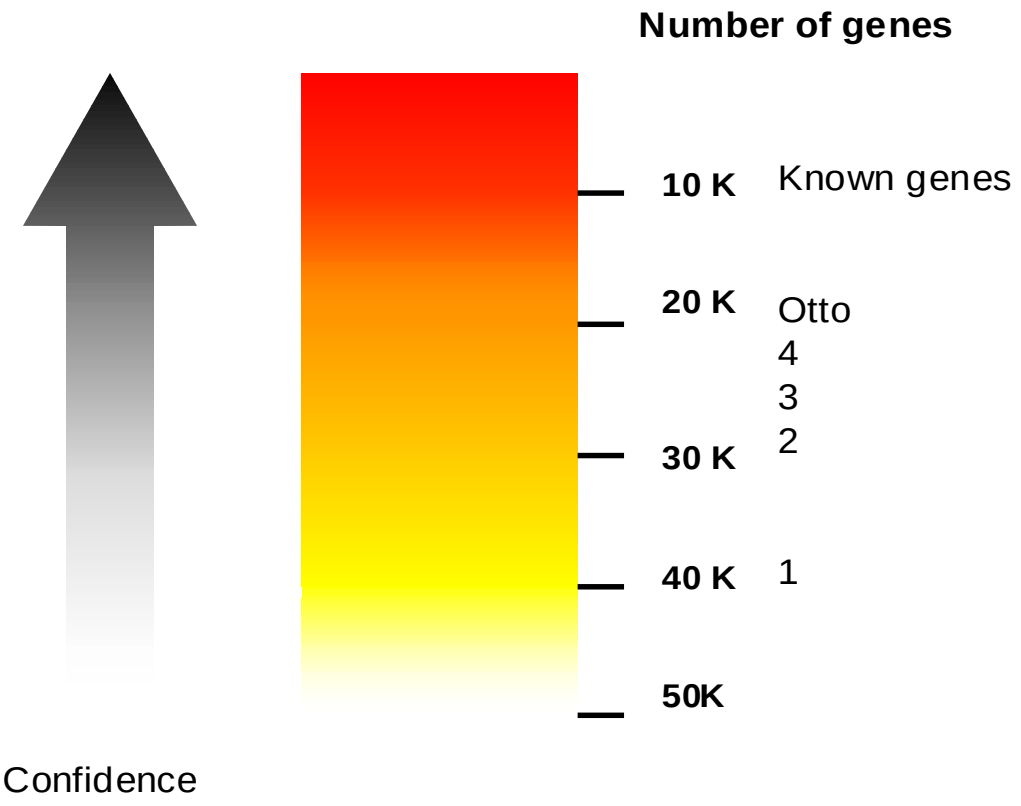
9089. .9225

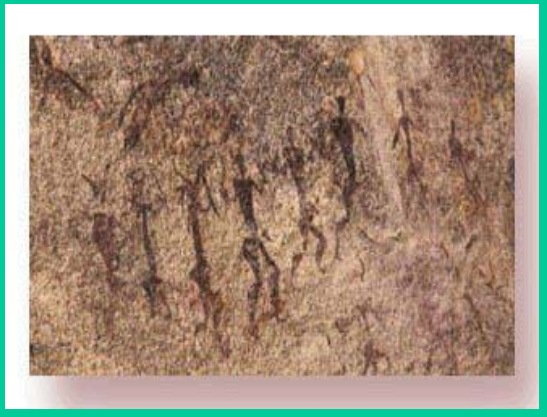
13841. .14244

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

Gene Number in the Human Genome





Experiment Driven

Hypothesis

↓

Experiments

↓

Results



Information Driven

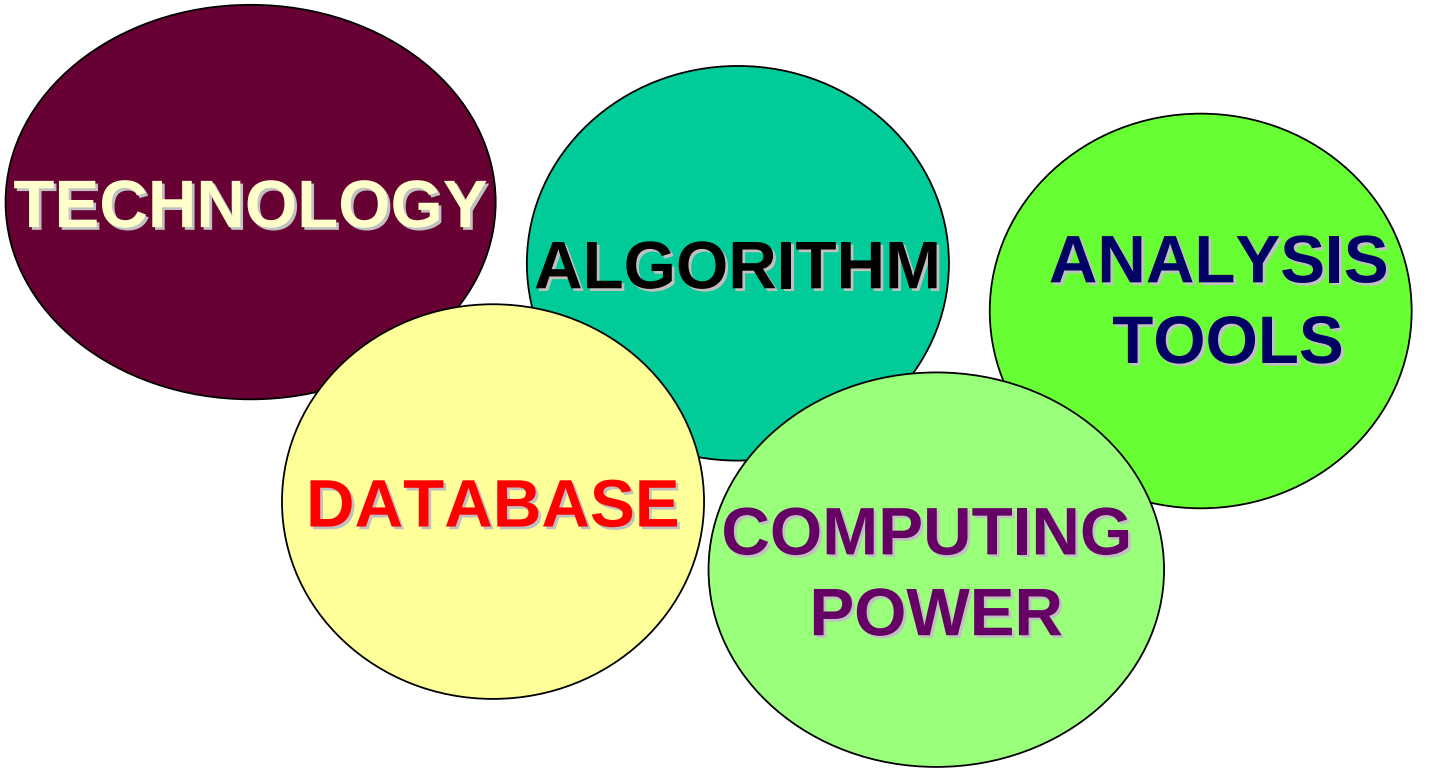
Experiments

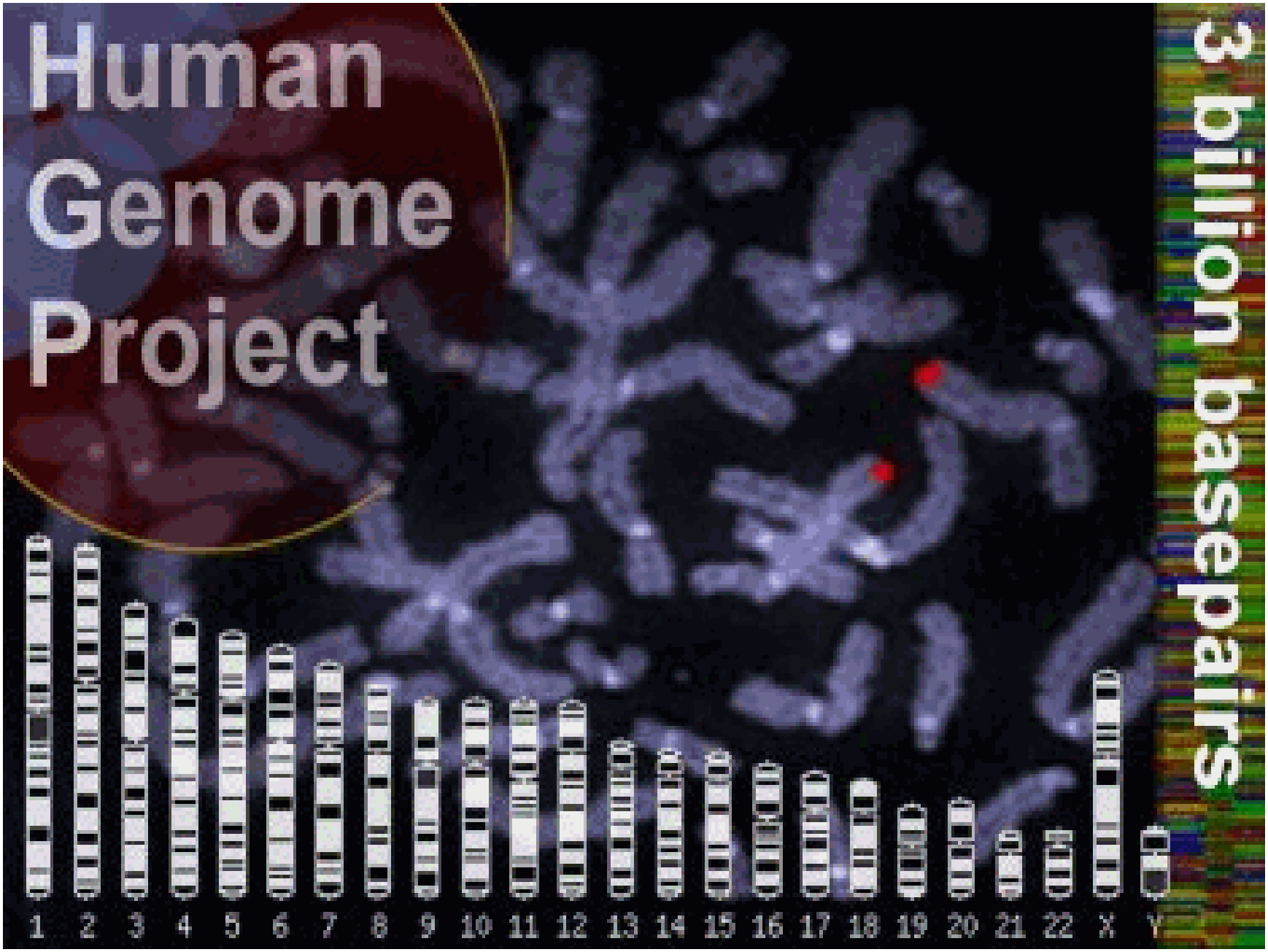
↓

Hypothesis



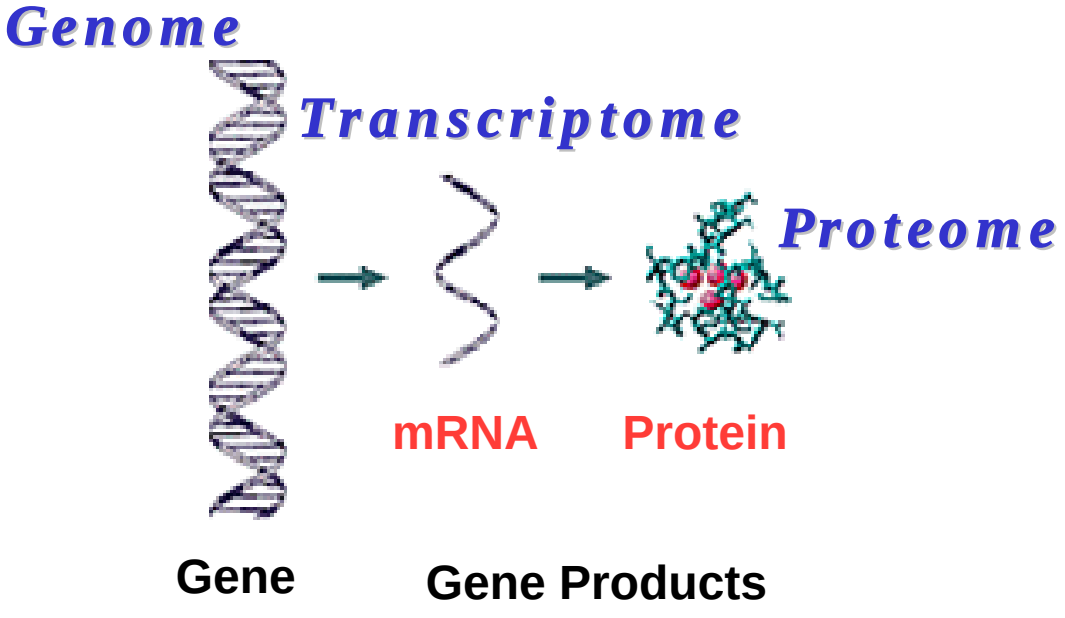
THE COMPONENTS OF BIOINFORMATICS



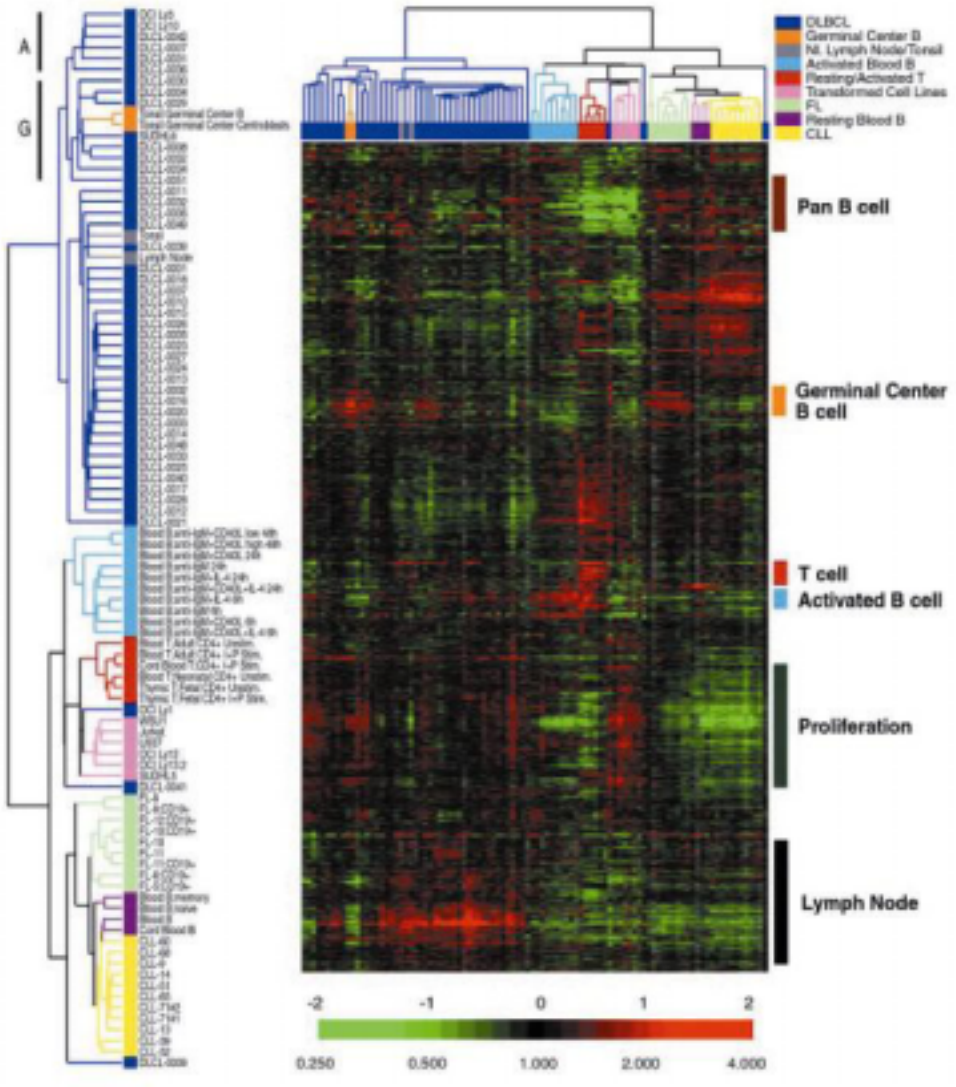

Human
Genome
Project
3 billion base pairs
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y

High Throughput Technologies: *The future of Molecular Medicine*

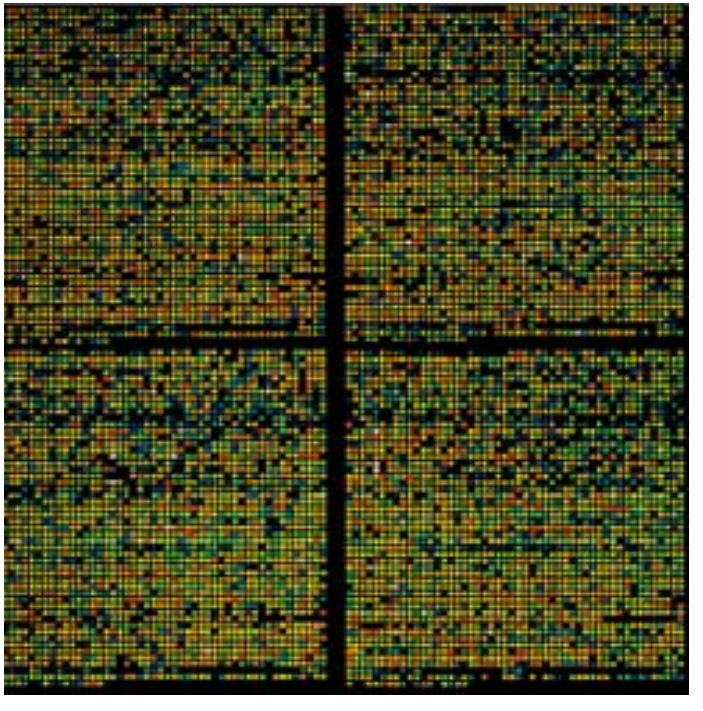
High Throughput Technologies (HTTs) are developed to produce huge amount of information from genome projects, but they have clear potential in mass screening and diagnostics of Infectious Diseases. The application of HTTs may revolutionize diagnostic techniques and replacing multiple individual assays.



Microarray



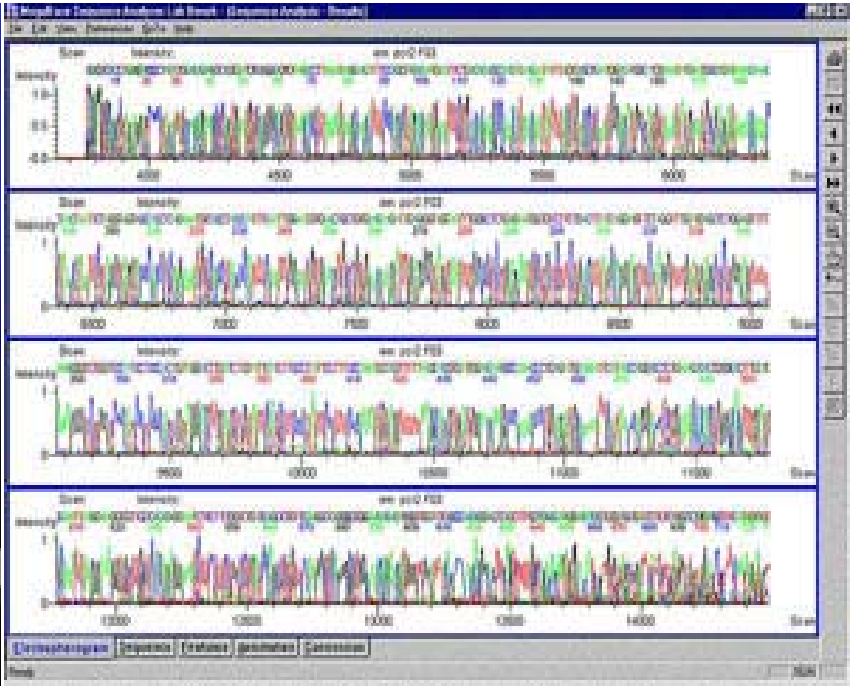
10,000 Clones per slide



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

DNA Sequencing

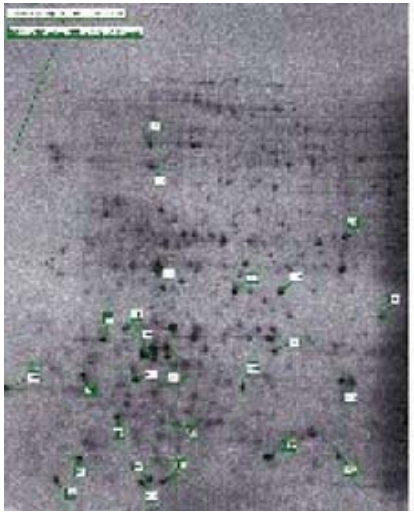


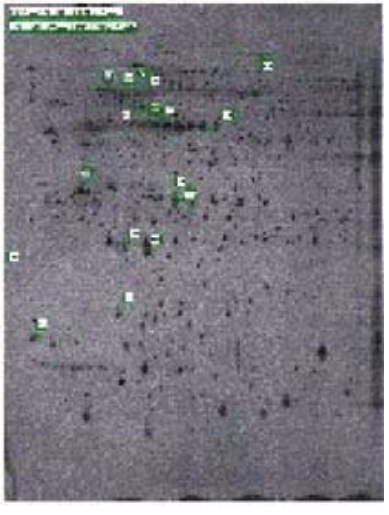


96 DNA sequencing in 2 hrs, approximately 600-800 readable bps per run.

1,000,000 bps in 24 hrs.

Proteomics



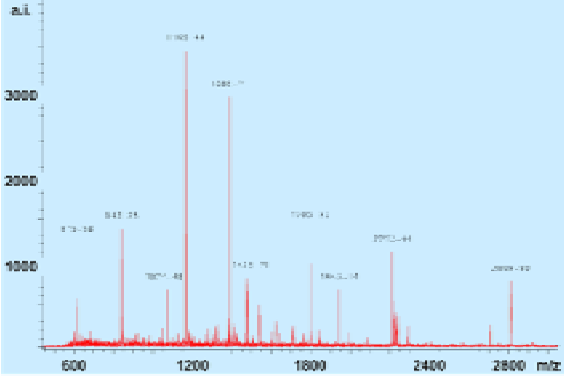


2 Dimensional Electrophoresis gels, differences that are characteristics of the individual starting states recognized by comparison of two protein pattern

6,000 protein spots per gel



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

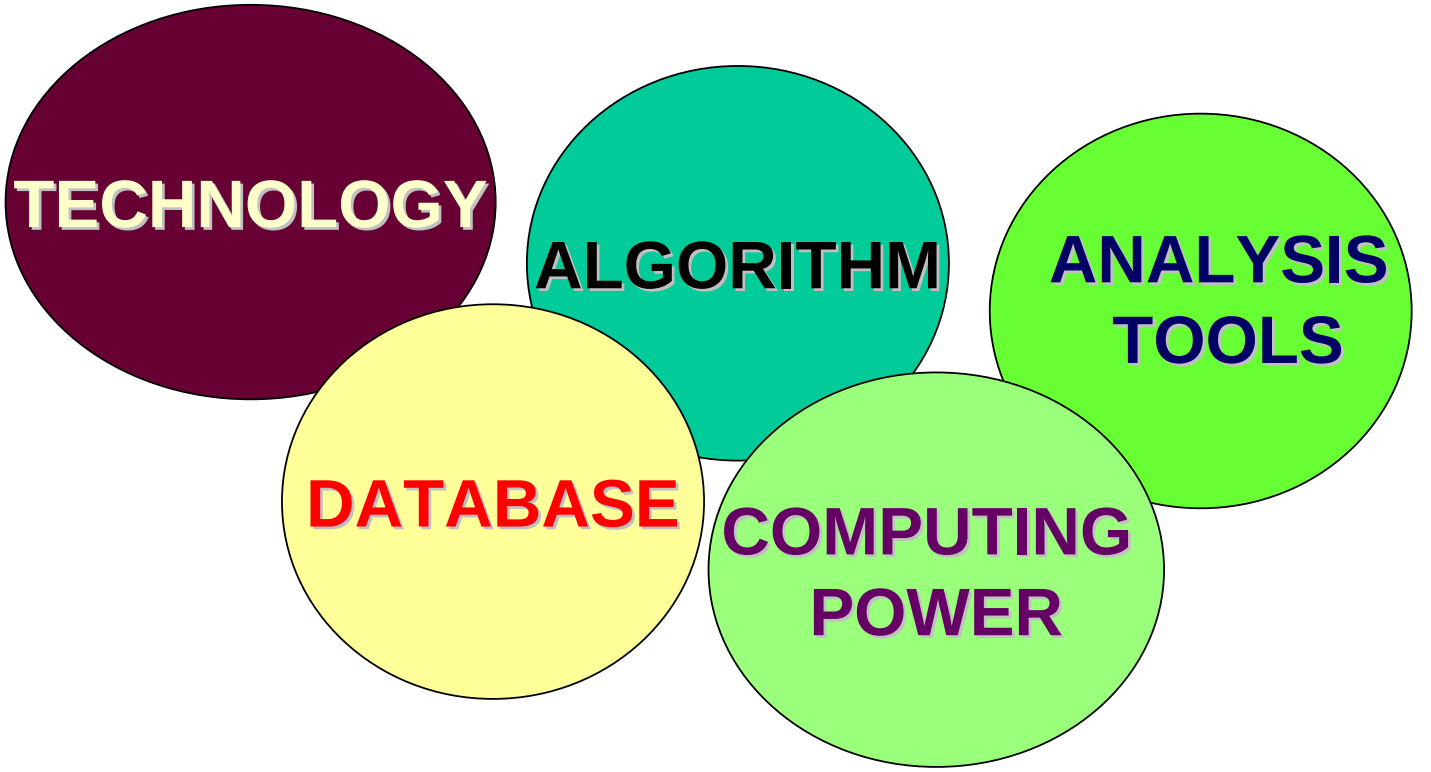


3D Modeling





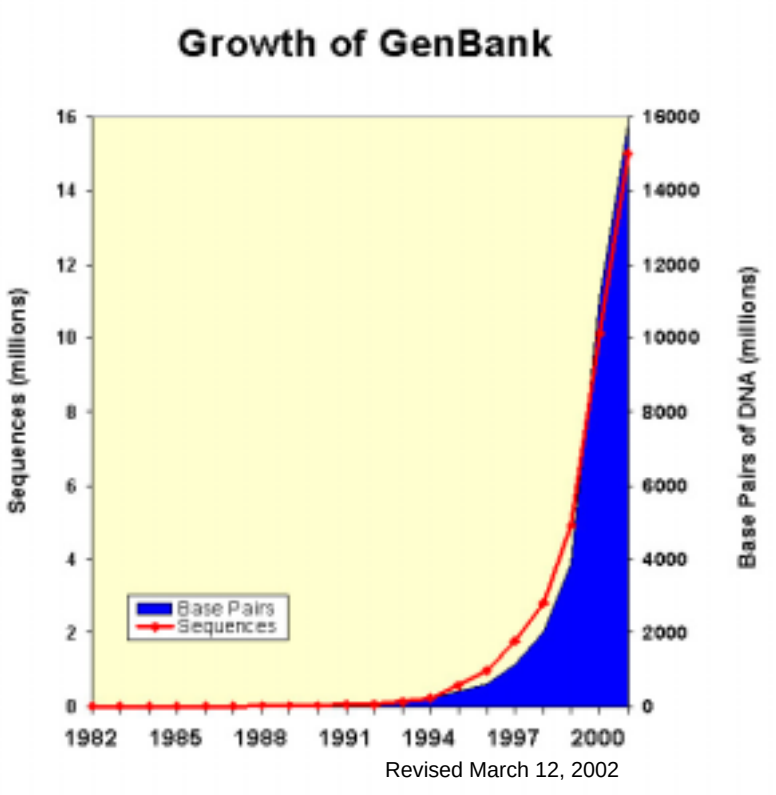
THE COMPONENTS OF BIOINFORMATICS



NCBI-GenBank Flat File Release 131.0

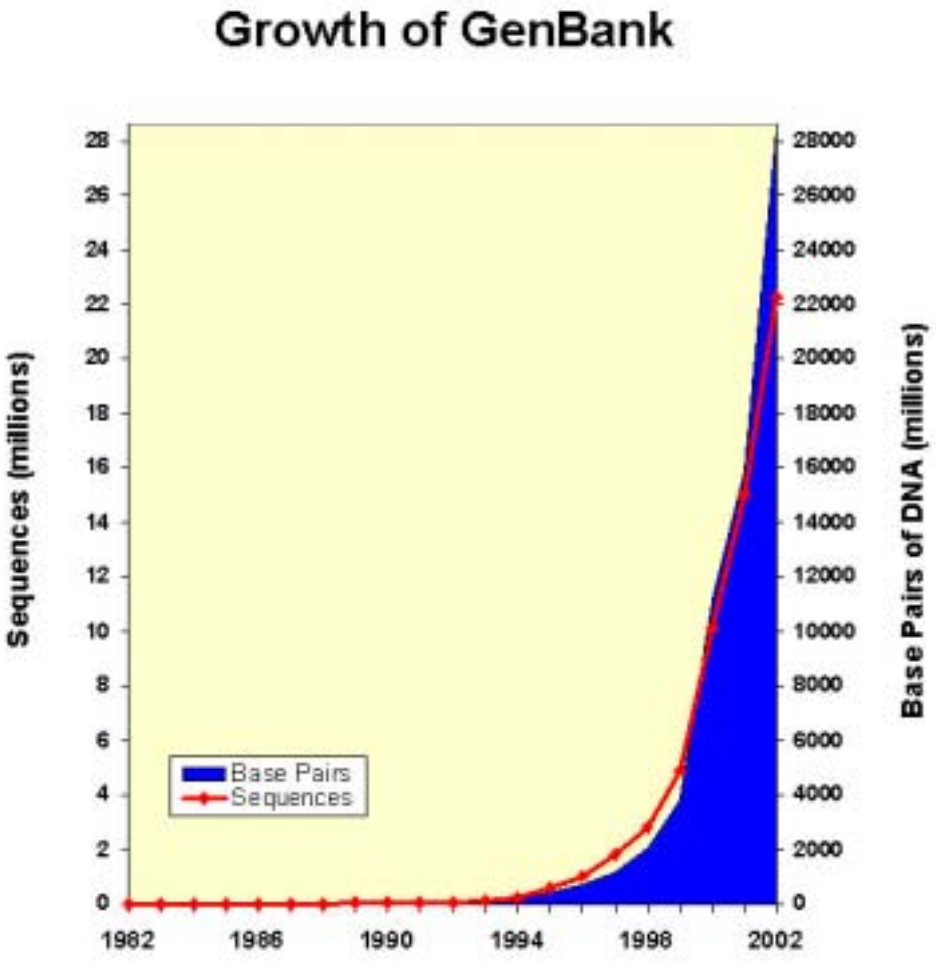
(August 15 2002)

[18,197,119 Genes] [22,616,937,182 Bases]



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.131) 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***

GenBank Data		
Year	**Base Pairs**	**Sequences**
1982	680338	606
1983	2274029	2427
1984	3368765	4175
1985	5204420	5700
1986	9615371	9978
1987	15514776	14584
1988	23800000	20579
1989	34762585	28791
1990	49179285	39533
1991	71947426	55627
1992	101008486	78608
1993	157152442	143492
1994	217102462	215273
1995	384939485	555694
1996	651972984	1021211
1997	1160300687	1765847
1998	2008761784	2837897
1999	3841163011	4864570
2000	11101066288	10106023
2001	15849921438	14976310



Genetic Sequence Data Bank June 15 2003, Release 136.0

25,592,865 loci,
32,528,249,295 bases, from
25,592,865 reported sequences

Homo sapiens: 6,859,992 sequences
EST: 5,399,835 sequences

The International Nucleotide Sequence Database Collaboration



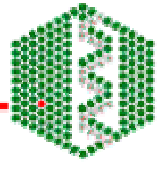
GenBank: *http://www.ncbi.nlm.nih.gov/*
National Center for Biotechnology Information (NCBI)

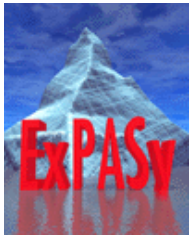


DDBJ: *http://www.ddbj.nig.ac.jp/*
National Institute of Genetics (NIG)

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

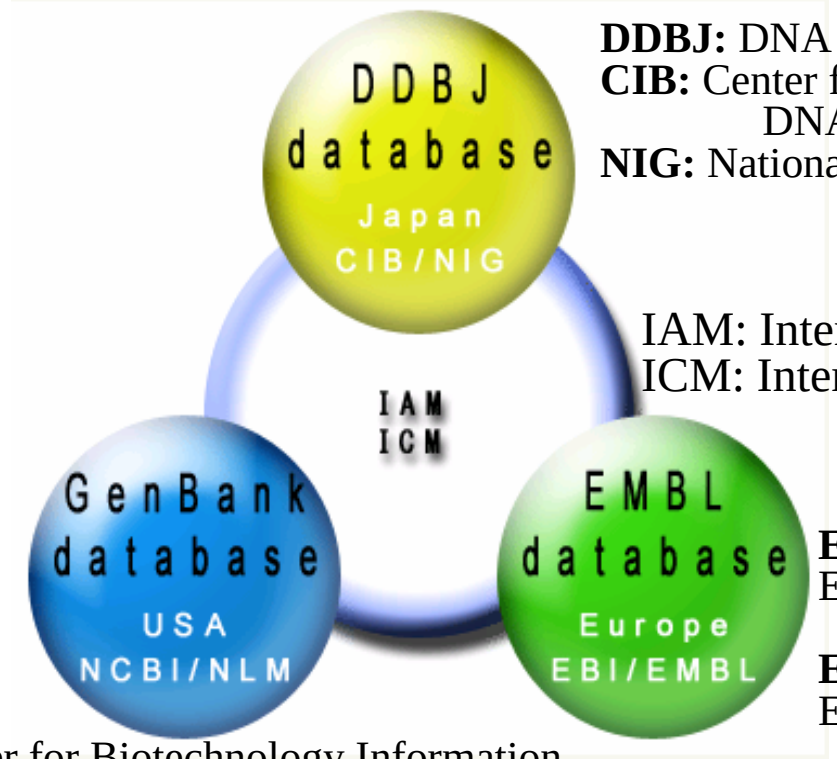
EMBL
European Bioinformatics Institute





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

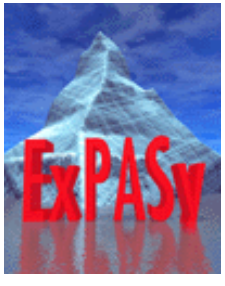
SWISSPROT *http://www.ebi.ac.uk/swissprot/*



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*



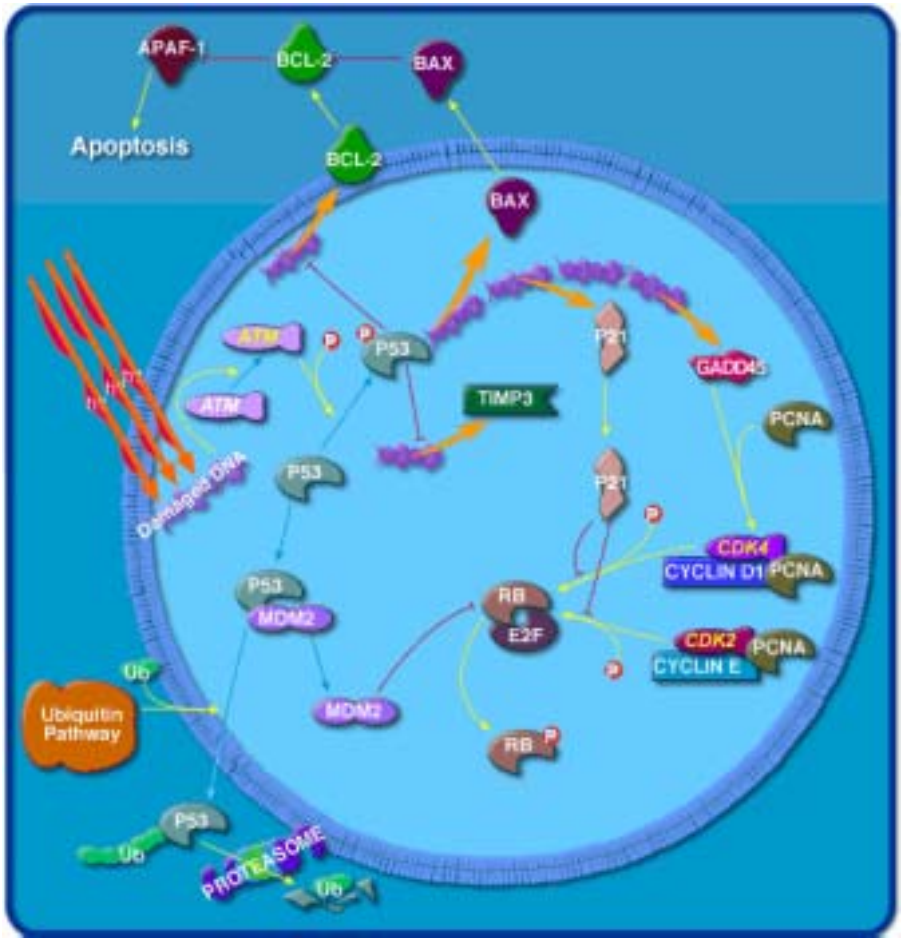
The ExPASy (**Ex**pert **P**rotein **A**nalysis **Sy**stem) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE

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

PROTEIN DATA BANK PDB

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



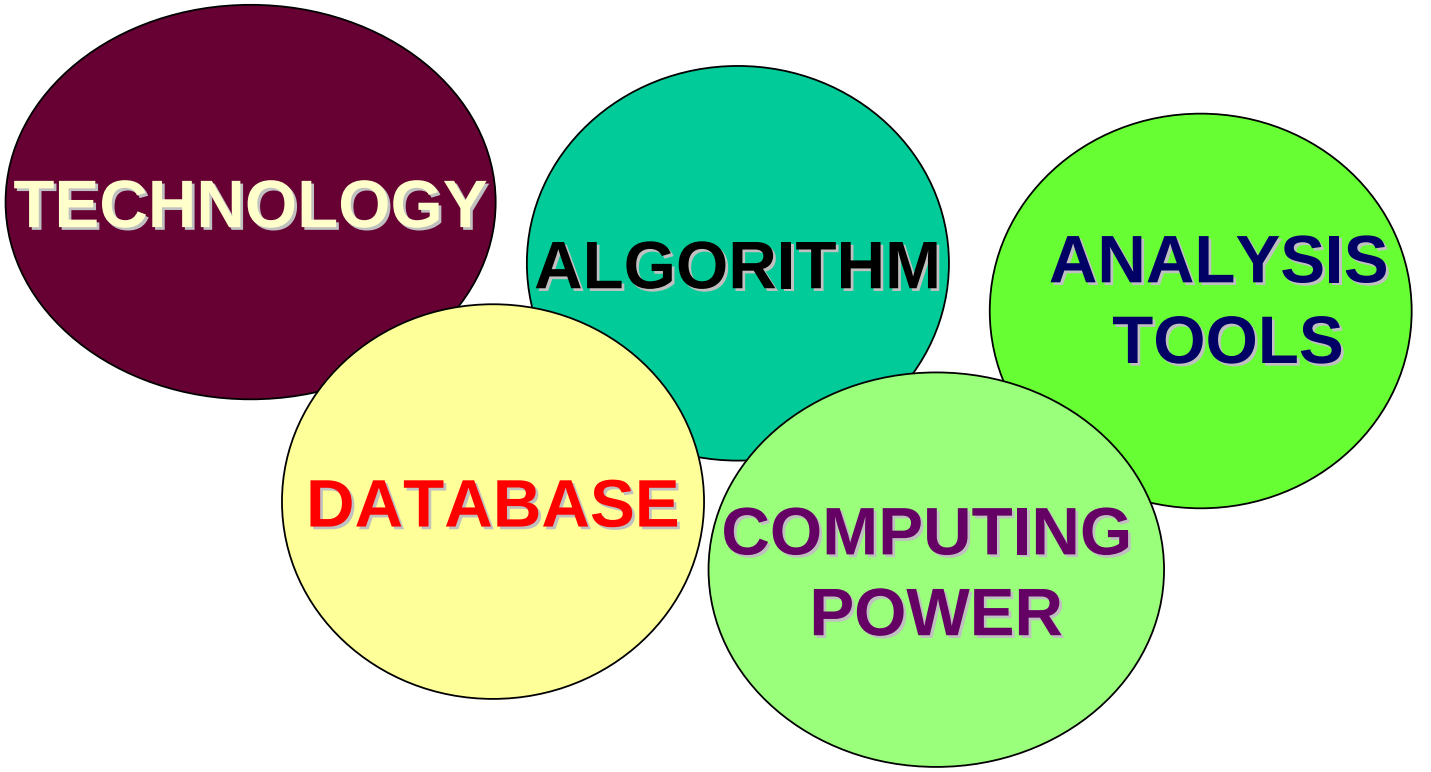
Biocarta
(*http://biocarta.com*)

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://apdb1.nhri.org.tw:8003/)

EMBOSS

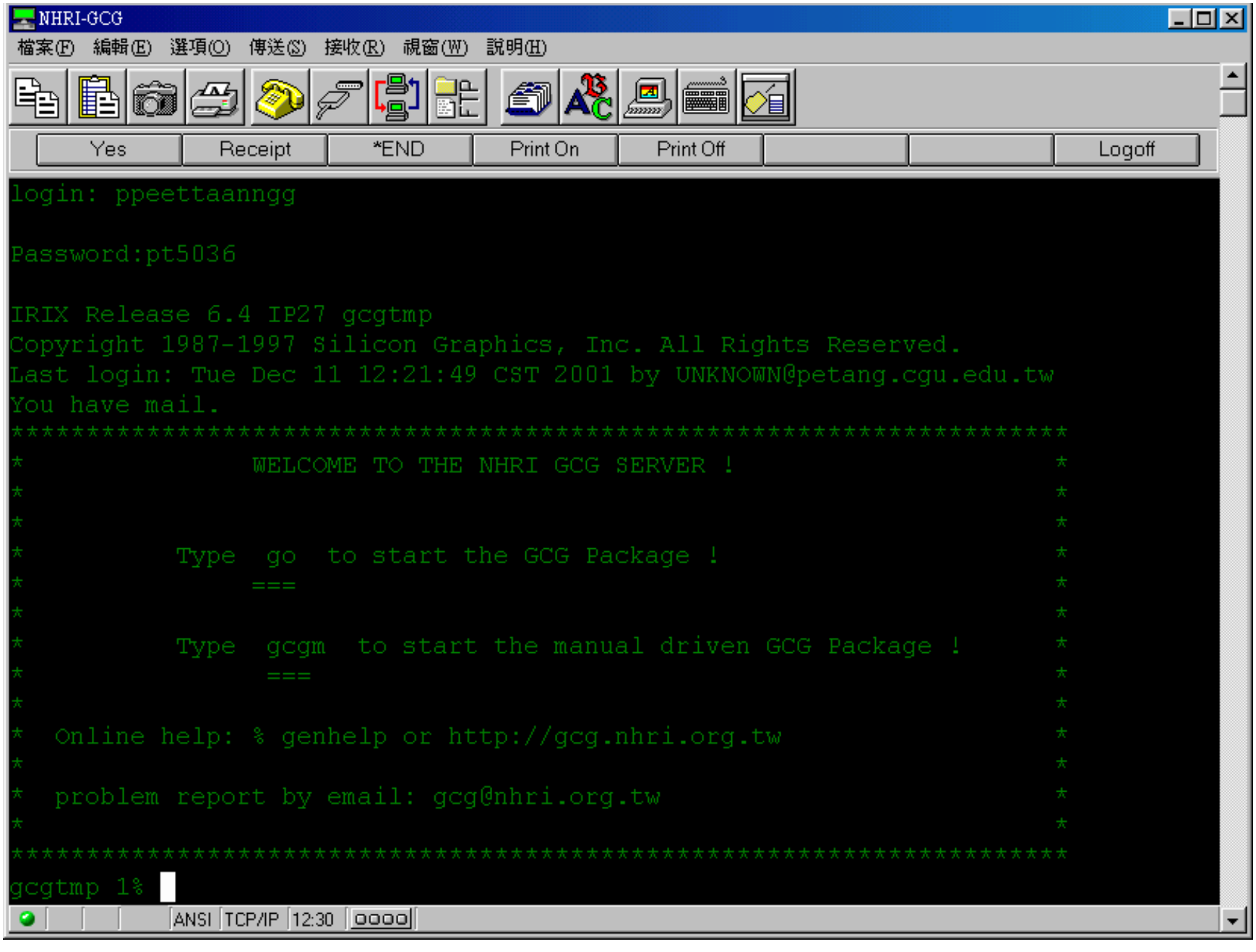
連線至 SeqWEB 以瀏覽器進行核酸或蛋白質的序列分析
(http://emboss.nhri.org.tw/)

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 (**Ex**pert **P**rotein **A**nalysis **Sy**stem)

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


NHRI-GCG
檔案(F) 編輯(E) 選項(O) 傳送(S) 接收(R) 視窗(W) 說明(H)
Yes
Receipt
*END
Print On
Print Off
Logoff
login: ppeettaanngg
Password:pt5036
IRIX Release 6.4 IP27 gcgtmp
Copyright 1987-1997 Silicon Graphics, Inc. All Rights Reserved.
Last login: Tue Dec 11 12:21:49 CST 2001 by UNKNOWN@petang.cgu.edu.tw
You have mail.
WELCOME TO THE NHRI GCG SERVER !
Type go to start the GCG Package !
Type gcgm to start the manual driven GCG Package !
Online help: % genhelp or http://gcg.nhri.org.tw
problem report by email: gcg@nhri.org.tw
gcgtmp 1%
ANSI TCP/IP 12:30

SeqWeb 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 SeqWEB-GCG SeqWEB-NCHC Sina SRS6 TSC 台北起區



- Contents
- Index

Comparison

Database Searching

Evolution

Mapping

Pattern Recognition

Primer Selection

Protein Analysis

Nucleic Acid Secondary Structure

Translation



SeqWeb Version 1.2
For use with the Wisconsin Package Version 10.1

Comparison

Use these programs to compare two or more sequences.

Gap

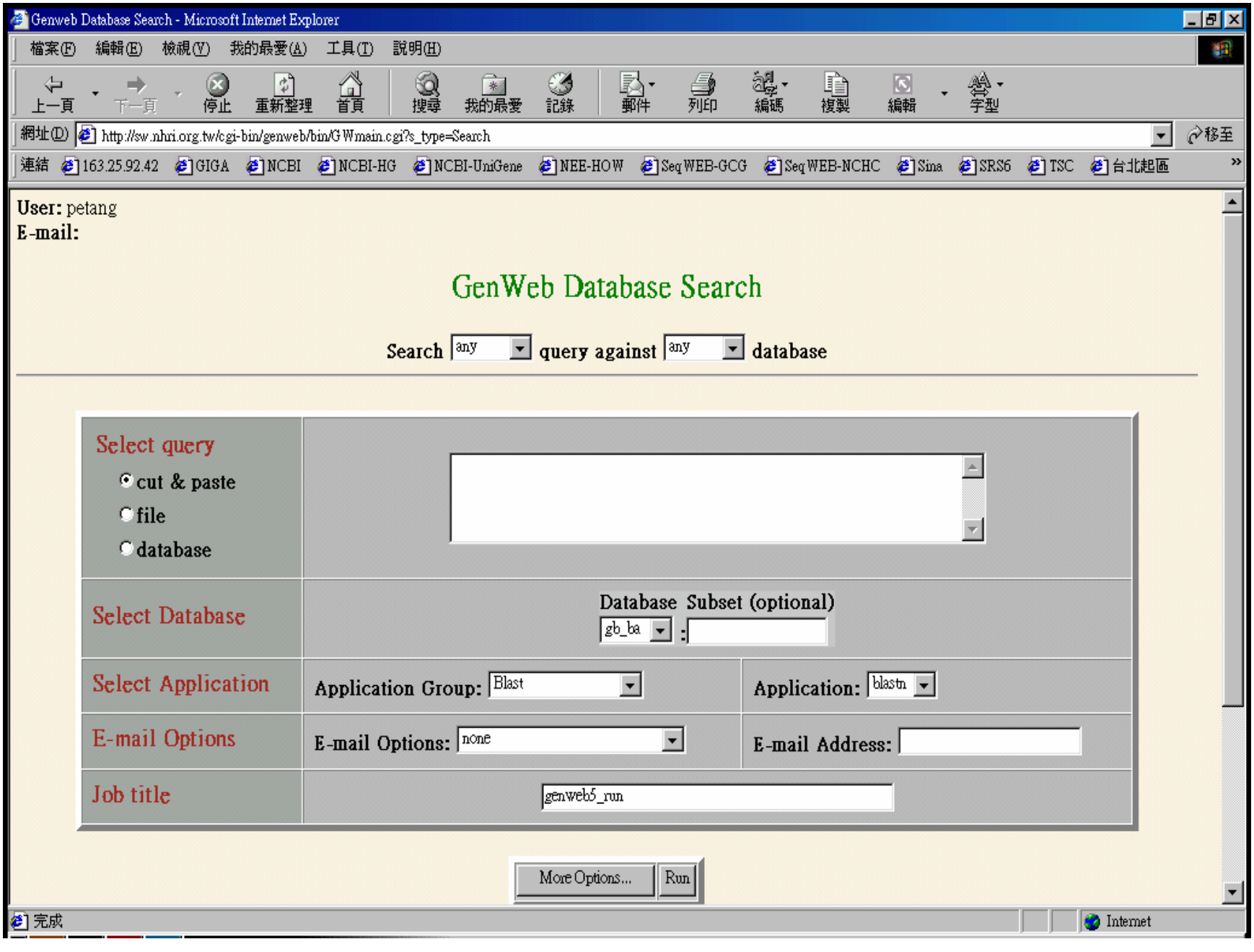
Uses the algorithm of Needleman and Wunsch to find the alignment of two complete sequences. It maximizes the number matches and minimizes the number of gaps.


Globally align two nucleic acid sequences.


Globally align two peptide sequences.

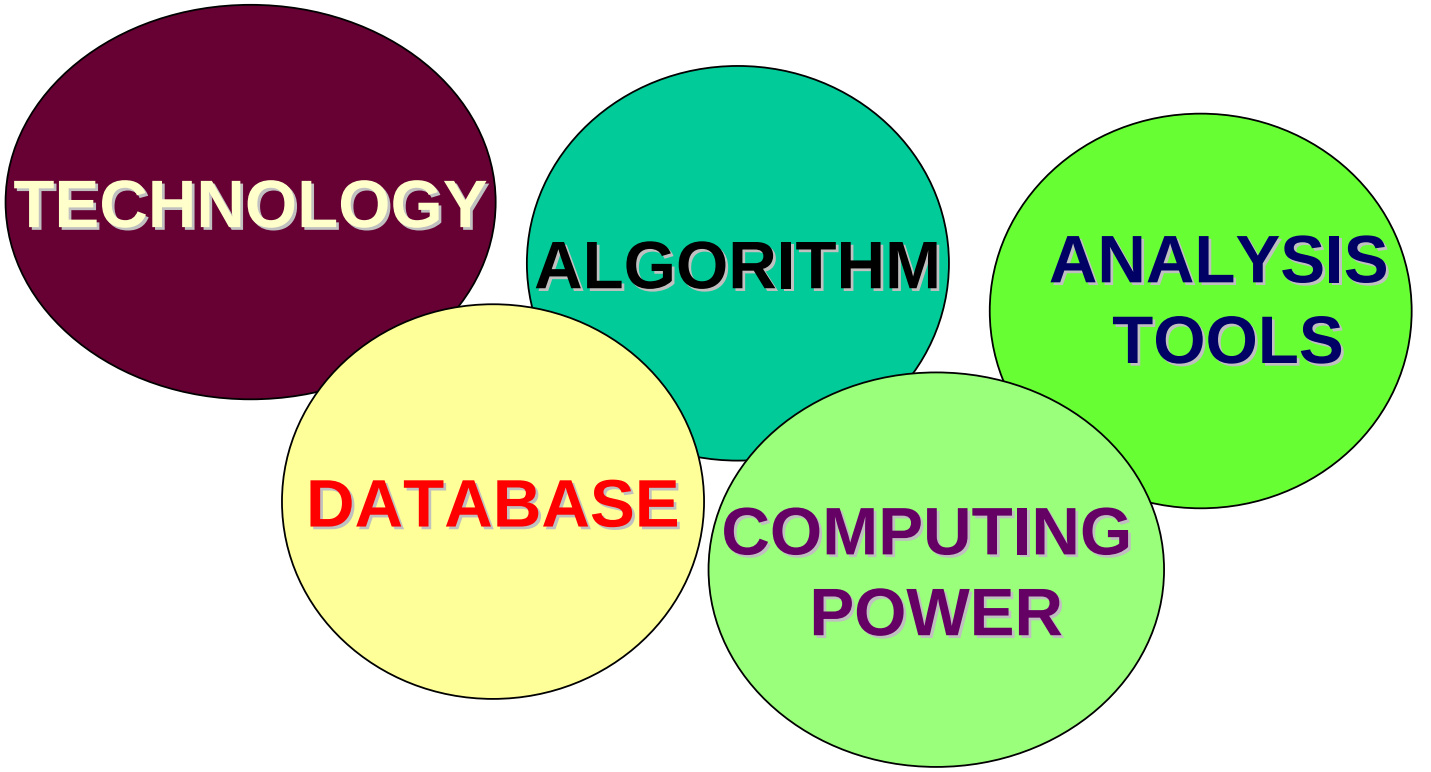
BestFit

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 Seq WEB-GCG Seq WEB-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

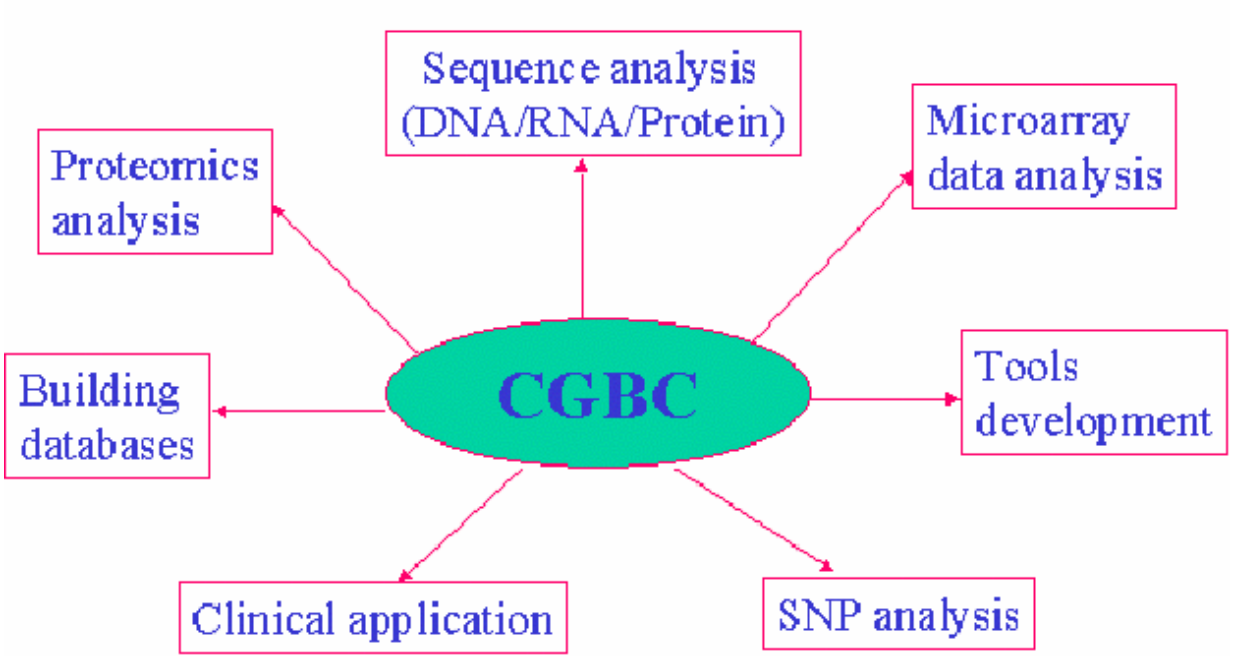


THE COMPONENTS OF BIOINFORMATICS



Chang Gung Bioinformatics Center (CGBC)

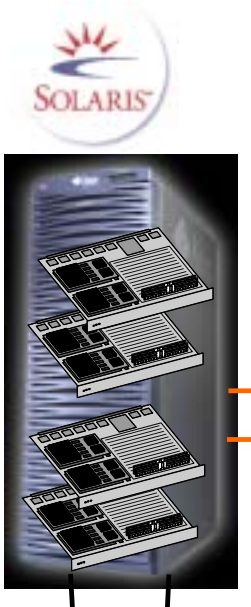


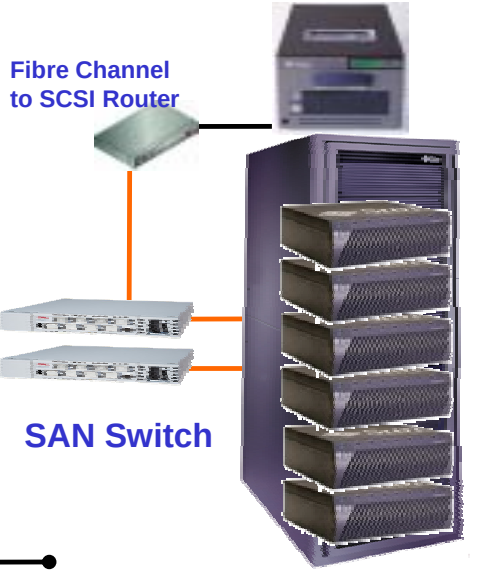


長庚生物資訊中心核心硬體架構

Sun L20 Tape Library
- 1 x DLT8000, 20 slots
- Backup Capacity : 800/1600GB
- Solstice Backup software

Sun Fire 6800 Server
(Append 8 CPUs 8 GB Memory)
- Four System Domain
- 4 x system board
- 16 x 750MHz Ultra SPARC III CPU
- 32GB Memory
- 6 x Power Supply, 4 x Fan tray
- 2 x System Controller
- 4 x PCI Assembly with 8 slots
- 4 x 10/100 FastEthernet Interface
- 4 Media tray (DVD, DDS-4 tape driver)
- 4 x2x18GB HDD (OS Mirror protect)
- 4 x PCI-Bus FC-AL Network Adapter





Local Area Network (Fast Ethernet Switch)



x6

Sun Blade 1000
- 750MHz Ultra SPARC III CPU
- 512MB RAM, 18GB SCSI HDD



x 20

PC P4
- 1.5GHz P4 CPU
- 2GB RAM, 80GB IDE HDD

Sun StorEdge T3 FC-AL Disk Array
(Append 1152GB Capacity)
- 2 x 9 x 18GB FC-AL HDD
- 2 x 9 x 36GB FC-AL HDD
- 2 x 9 x 72GB FC-AL HDD
- RAID 5 formatted Capacity : 2016GB
- 6 x Fibre Channel RAID Controller
- 6 x 258MB Cache Memory



Steps to Identify a Gene

- Gene-Search
- Protein-Search
- Annotation









Translation of a 5'-EST nucleotide sequence

```
      taaccatcagatcgatcgacttcatgcatcgaacaacaaaatgttcatgtcaatttctga
    1 ---------+---------+---------+---------+---------+---------+ 60
      attggtagtctagctagctgaagtacgtagcttgttgttttacaagtacagttaaagact

a       * P S D R S T S C I E Q Q N V H V N F *  - a
b        N H Q I D R L H A S N N K M F M S I S D - b
c         T I R S I D F M H R T T K C S C Q F L M - c

      tgagcacggtgatccaactggcgaattgtcagtgtccgacaatggattgatgagagttct
   61 ---------+---------+---------+---------+---------+---------+ 120
      actcgtgccactaggttgaccgcttaacagtcacaggctgttacctaactactctcaaga

a       * A R * S N W R I V S V R Q W I D E S S  - a
b        E H G D P T G E L S V S D N G L M R V L - b
c         S T V I Q L A N C Q C P T M D * * E F F - c

      tcgagaaaatgtcattggagtgagcaacgtggcagtcgactggattggtggaaacgtttt
  121 ---------+---------+---------+---------+---------+---------+ 180
      agctcttttacagtaacctcactcgttgcaccgtcagctgacctaaccacctttgcaaaa

a       S R K C H W S E Q R G S R L D W W K R F  - a
b        R E N V I G V S N V A V D W I G G N V F - b
c         E K M S L E * A T W Q S T G L V E T F S - c

      cttcacacaaaaatctccatctccaagcgctgggatttccatctgcacaatgagcggaat
  181 ---------+---------+---------+---------+---------+---------+ 240
      gaagtgtgtttttagaggtagaggttcgcgaccctaaaggtagacgtgttactcgcctta

a       L H T K I S I S K R W D F H L H N E R N  - a
b        F T Q K S P S P S A G I S I C T M S G M - b
c         S H K N L H L Q A L G F P S A Q * A E C - c

      gttctgtcgccgagttatcgaaggcaaagaacaaggacaatcctatcgtggtcttgttgt
  241 ---------+---------+---------+---------+---------+---------+ 300
      caagacagcggctcaatagcttccgtttcttgttcctgttaggatagcaccagaacaaca

a       V L S P S Y R R Q R T R T I L S W S C C  -
b        F C R R V I E G K E Q G Q S Y R G L V V -
c         S V A E L S K A K N K D N P I V V L L F -
```

```
      tcacccgatgcgcggtctcatcatctggatcgattcttatcagaaatatcatcgcatcat
  301 ---------+---------+---------+---------+---------+---------+ 360
      agtgggctacgcgccagagtagtagacctagctaagaatagtctttatagtagcgtagta

a       S P D A R S H H L D R F L S E I S S H H  -
b        H P M R G L I I W I D S Y Q K Y H R I M -
c         T R C A V S S S G S I L I R N I I A S * -

      gatggctaatatggatgggtctcaggtcagaatccttctcgacaacaagttngaagttcc
  361 ---------+---------+---------+---------+---------+---------+ 420
      ctaccgattatacctacccagagtccagtcttaggaagagctgttgttcaancttcaagg

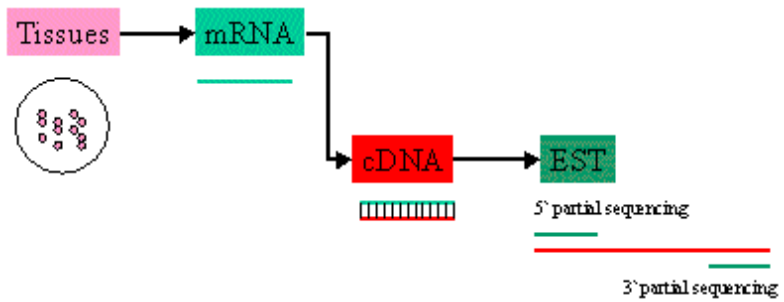
a       D G * Y G W V S G Q N P S R Q Q V ? S S  -
b        M A N M D G S Q V R I L L D N K ? E V P -
c         W L I W M G L R S E S F S T T S ? K F H -

      atcagcttttgccatcgnctacatccgccacgatgtctatttttggggntgttgancgca
  421 ---------+---------+---------+---------+---------+---------+ 480
      tagtcgaaaacggtagcngatgtaggcggtgctacagataaaaaccccnacaactngcgt

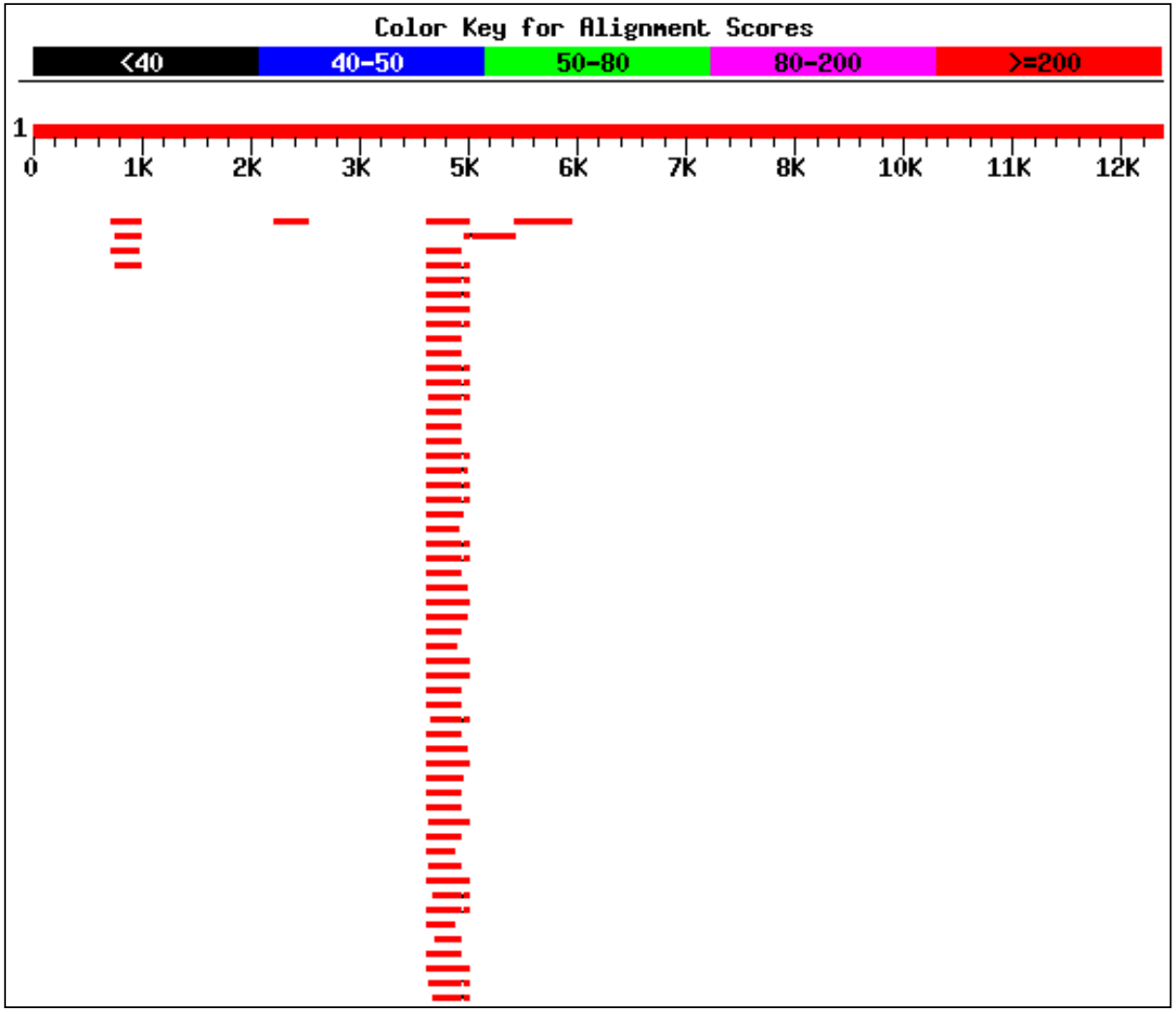
a       I S F C H R L H P P R C L F L G ? L ? A  -
b        S A F A I ? Y I R H D V Y F W G C * ? H -
c         Q L L P S ? T S A T M S I F G ? V ? R I -

      ttgatcgaaagngtcatttcgncacggaagngtg
  481 ---------+---------+---------+---- 514
      aactagctttcncagtaaagcngtgccttcncac

a       L I E ? V I S ? R K ?   -
b        * S K ? S F R H G ? V  -
c         D R K ? H F ? T E ?   -
```



5’-EST Genomic Blast



```
  1  M  R  T  M  R  L  A  W  L  L  P  L  F  I  H  I  L  I  K  N
 21  T  A  Q  A  P  A  V  N  N  S  T  C  D  Q  A  K  E  F  D  C
 41  G  N  G  R  L  R  C  I  P  A  E  W  Q  C  D  N  V  A  D  C
 61  D  K  G  R  D  E  S  G  C  S  Y  A  H  H  C  S  T  S  F  M
 81  L  C  K  N  G  L  C  V  A  N  E  F  K  C  D  G  E  D  D  C
101  R  D  G  S  D  E  Q  H  C  E  Y  N  I  L  K  S  R  F  D  G
121  S  N  P  S  A  P  T  T  F  L  G  H  N  G  P  E  C  H  P  P
141  R  L  R  C  R  S  G  Q  C  I  Q  P  D  L  V  C  D  G  H  Q
161  D  C  S  G  G  D  D  E  V  N  C  T  R  R  G  H  E  N  M  Q
181  S  S  T  D  F  H  D  D  V  H  L  V  D  P  T  F  F  A  N  E
201  D  N  K  C  R  S  G  Y  T  M  C  H  S  G  D  V  C  I  P  D
221  S  F  L  C  D  G  D  L  D  C  D  D  A  S  D  E  K  N  C  Q
241  T  N  A  P  S  E  E  E  Y  L  S  G  Q  A  D  H  M  H  S  C
261  S  A  A  G  M  Y  S  C  G  R  K  G  S  E  I  G  V  C  I  P
281  M  N  A  T  C  N  G  I  K  E  C  P  L  G  D  D  E  S  K  H
301  C  S  E  C  A  R  K  R  C  D  H  T  C  M  N  T  P  H  G  A
321  R  C  I  C  Q  E  G  Y  K  L  A  D  D  G  L  T  C  E  D  E
341  D  E  C  A  T  H  G  H  L  C  Q  H  F  C  E  D  R  L  G  S
361  F  A  C  K  C  A  N  G  Y  E  L  E  T  D  G  H  S  C  K  Y
381  E  A  T  T  T  P  E  G  Y  L  F  I  S  L  G  G  E  V  R  Q
401  M  P  L  A  D  F  T  D  G  S  N  Y  S  A  I  Q  K  F  A  G
421  H  G  T  I  R  S  I  D  F  M  H  R  N  N  K  M  F  M  S  I
441  S  D  E  H  G  D  P  T  G  E  L  S  V  S  D  N  G  L  M  R
461  V  L  R  E  N  V  I  G  V  S  N  V  A  V  D  W  I  G  G  N
481  V  F  F  T  Q  K  S  P  S  P  S  A  G  I  S  I  C  T  M  S
501  G  M  F  C  R  R  V  I  E  G  K  E  Q  G  Q  S  Y  R  G  L
521  V  V  H  P  M  R  G  L  I  I  W  I  D  S  Y  Q  K  Y  H  R
541  I  M  M  A  N  M  D  G  S  Q  V  R  I  L  L  D  N  K  L  E
561  V  P  S  A  L  A  I  D  Y  I  R  H  D  V  Y  F  G  D  V  E
581  R  Q  L  I  E  R  V  N  I  D  T  K  E  R  R  V  V  I  S  N
601  G  V  H  H  P  Y  D  M  A  Y  F  N  G  F  L  Y  W  A  D  W
621  G  S  E  S  L  K  V  Q  E  M  T  H  H  H  S  S  P  Q  V  I
641  H  T  F  N  R  Y  P  Y  G  I  A  V  N  H  S  L  Y  Q  T  G
661  P  P  S  N  P  C  L  E  L  E  C  P  W  L  C  V  I  V  P  K
681  S  D  F  I  M  T  A  K  C  V  C  P  D  G  Y  T  H  S  V  T
701  E  N  S  C  I  P  P  V  T  I  E  D  E  E  N  L  E  K  L  S
721  H  I  G  S  A  L  M  A  E  Y  C  E  A  G  V  A  C  M  N  G
741  G  A  C  R  E  L  Q  N  E  H  G  R  A  H  R  I  V  C  D  C
761  E  G  P  Y  D  G  Q  Y  C  E  R  L  N  P  E  K  F  S  A  M
781  E  E  E  D  S  S  L  W  L  I  V  L  L  L  I  F  L  I  I  V
801  A  V  V  G  I  I  A  F  L  W  F  S  Q  Q  E  H  M  K  D  L
821  I  S  T  A  R  V  R  V  D  N  M  A  R  K  A  E  D  A  A  A
841  P  I  V  E  K  F  R  K  V  T  D  K  Q  R  S  T  P  P  R  E
861  G  C  Q  T  A  T  N  V  D  F  V  S  Y  E  T  N  A  E  K  R
881  I  R  M  D  S  S  P  T  S  Y  G  N  P  M  Y  D  E  V  P  E
901  S  S  T  G  F  V  R  S  A  S  A  P  F  A  G  V  I  R  F  E
921  N  D  S  L  L
```

A Full-length cDNA clone

Amino Acid Composition

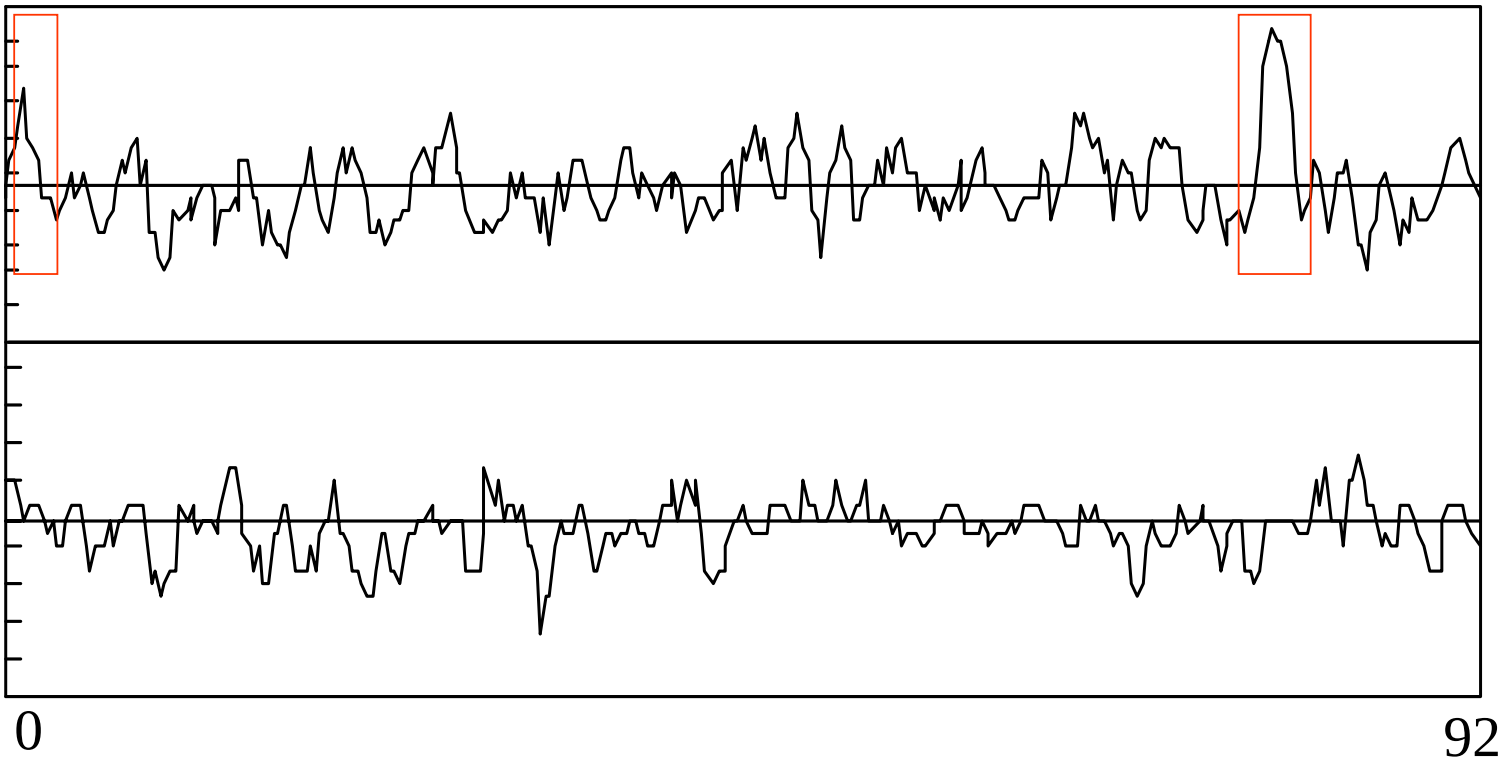
Amino acid	Number	%	Amino acid	Number	%
Alanine	56	6.0	Leucine	56	6.0
Arginine	49	5.3	Lysine	33	3.6
Asparagine	45	4.9	Methionine	30	3.2
Aspartic acid	70	7.6	Phenylalanine	34	3.7
Cysteine	57	6.2	Proline	42	4.5
Glutamic acid	66	7.1	Serine	72	7.8
Glutamine	29	3.1	Threonine	41	4.4
Glycine	70	7.6	Tryptophan	9	1.0
Histidine	38	4.1	Tyrosine	28	3.0
Isoleucine	49	5.3	Valine	51	5.5

Total molecular weight of premature protein = 103 KDa
Total molecular weight of mature protein = 100 KDa

PROTEIN DATABASE SEARCH

Prorein	Similarity	Identity
LDL receptors		
Human	52.30%	27.59%
Mouse	50.06%	25.90%
Rat	50.44%	26.54%
Hamster	52.30%	27.59%
Rabbit	50.70%	26.19%
Xenopus1	48.63%	26.28%
Xenopus2	48.37%	26.28%
VLDL receptors		
Rabbit	49.06%	27.28%

Hydrophobicity (upper) and charge (lower) plots



Signal Sequence

possible clevage sites

(1)M R T M R L A W L L P L F I H I L I K N T A Q A P A(26)

-3 -1

-3 -1

Transmembrane segment

NH_2-

(673)WLCVIVPKSDFIMT

(687)AKCVCPDGYTHSVTENSCIPPVTIEDEENLEKLSHIGSALMAEYCEAGVA

(727)CMNGGACRELQNEHGRAHRIVCDCEGPYDGQYCERLNPEKFSAMEEEDSS

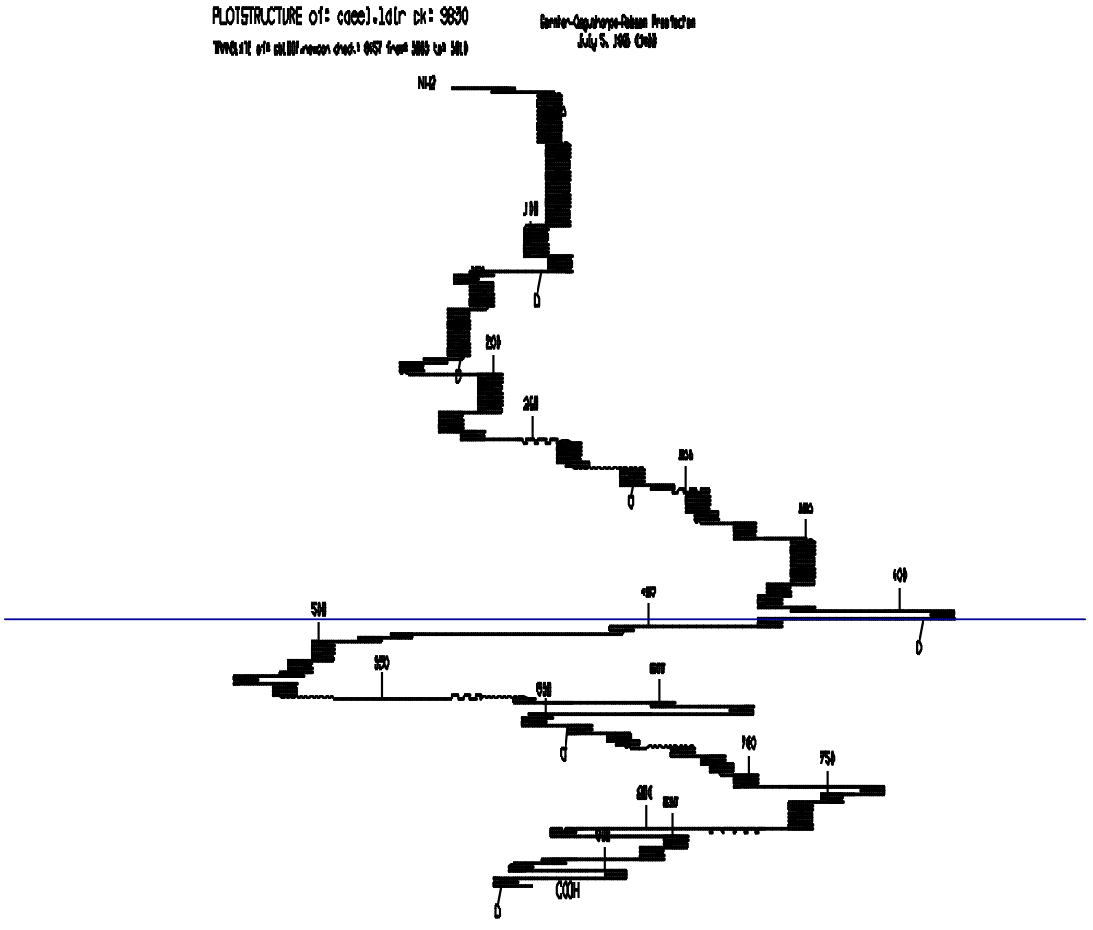
(787)LWLIVLLLIFLIIVAVVGIIAFLWF

(812)SQQEHMKDVISTARVRVDNMARKAEDAAAPIVEKFRKVTDKQRSTPPREG

(862)CQTATNVDFVSYETNAEKRIRMDSSPTSYGNPMYDEVPESSTGFVRSASA

(912)PFAGVIRFENDSLL(925)-COOH

PEPTIDE PLOT



Press <Return>

MOTIFS SEARCH

(A)Class A repeats

```
I     (32)CDQAKEFDCGNGRLR---CIPAEWQCDNVADC-DKRDE-SGC(69)
          -  +- -    + +       -   -   -  -++--
II    (75)CSTSFML-CKNGL-----CVANEFKCDGEDDCRDGSDE-QHC(109)
                  +           - + - --- +-  --
III  (137)CHPPRLR-CRSGQ-----CIQPDLVCDGHQDCSGGDDE-VNC(171)
             + +  +           -   -   -    ---
IV   (204)CRSGYTM-CHSGDV----CIPDSFLCDGDLDCDDASDE-KNC(239)
          +          -       +    - - - --  -- +
V    (260)CSAAGYS-CGTKGSEIGVCIPMNATCNGIKECPLGDDESKHC(301)
                   +  -              +-    --- +
          C       C  G      CIP    CDG  DC   G DE  C  C. elegans consensus
          C    EF C  G      CI   W CD   DC  DGSDE  C  Human LDL receptor
          C     F C  G     RCIP  W CD   DC  D SDE  C  Human LRP
          C     F C N       CI     CDG  DC  DGSDE  C  C. elegans LRP
```

(B)Class B.2 repeats

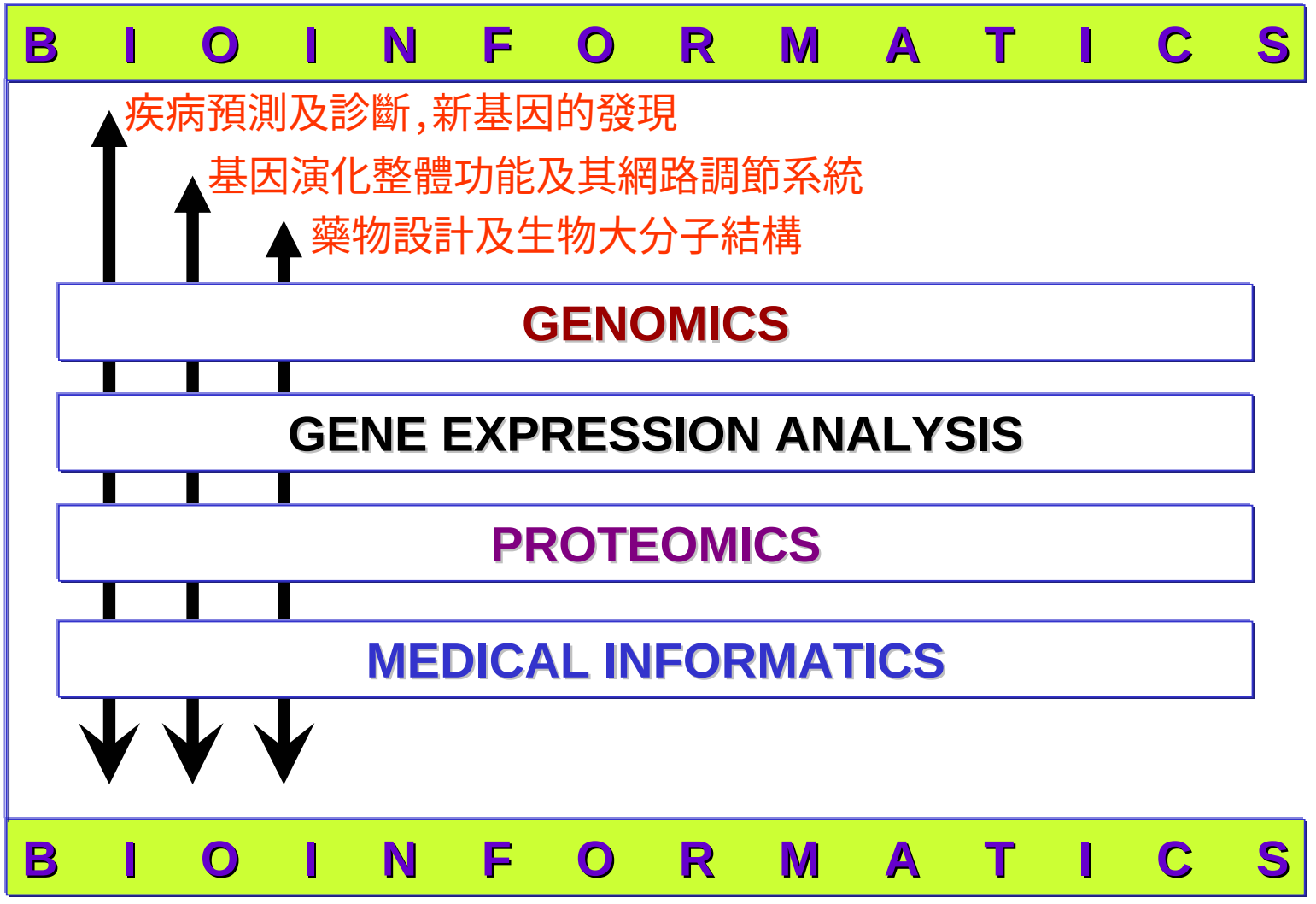
```
A    (304)CARKR--CDHTCMNTPHGAR-----CICQEGYKLADDGLTC(337)
B    (343)CATHGHLCQHFCEDRLGSFA-----CKCANGYELETDGHSC(378)
C    (666)CLELE--CPWLCVIVPKSDFIMTAKCVCPDGYTHSVTENSC(704)
          C      C   C            C C  GY        C   C. elegans consensus
```

(C)Class B.1 repeat

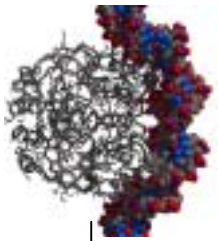
```
D    (731)CEAGVACNGGACRELQNEHGRAHRIVCDCEGPYDGQYC(769)
```

Conclusions

1. This a Type-I transmembrane protein
2. It is similar to the human low-density lipoprotein receptor
3. Contains unique cysteine-rich repeats
4. A new class of the LDL-receptor family


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

Our Focus in Bioinformatics

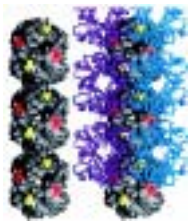


Perturbation
Environment
Medication
Genetic Engineering

Dynamic Response
Gene Expression
Protein Expression

BioChip

Virtual Cell



Analysis

DataBase
Genotype/Phenotype

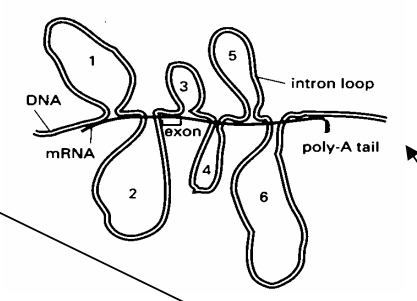
Biology
Molecular Biology
Bio Chemistry
Genetics

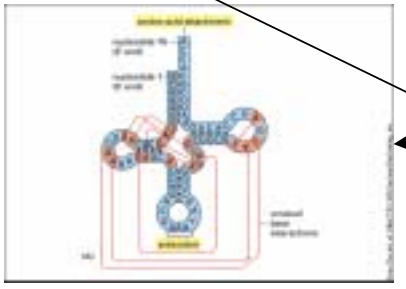
Symbolic
Algorithms/
Computing

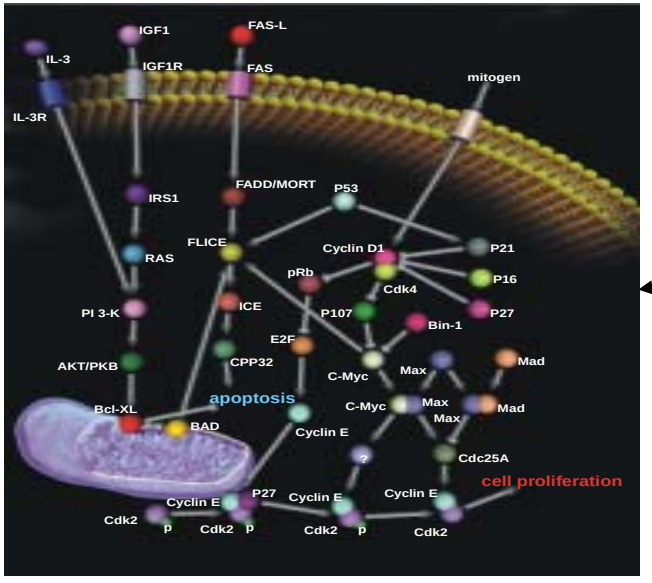
Genome Sequencing

Goals Leading Toward Predictive Biology









Gene Sequence Data

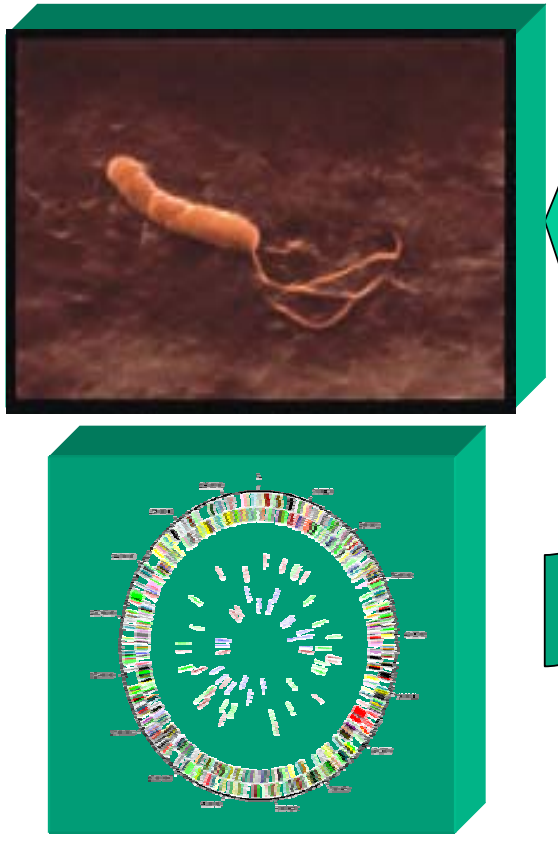
Gene Identification

Structure Prediction

Protein Circuit & Regulatory Network Discovery

Biosimulation

Reconstructing Cellular Functions



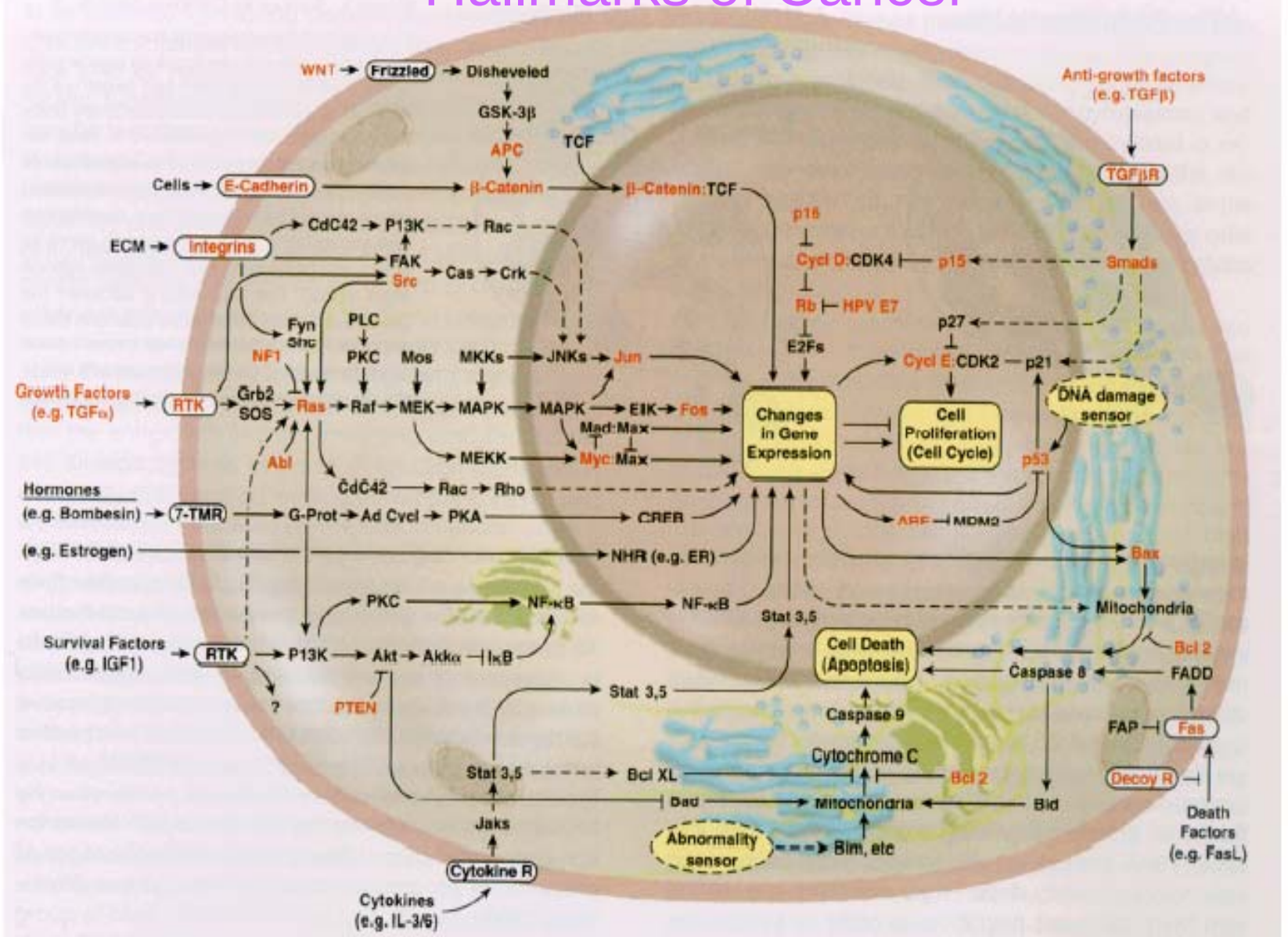
Reductionistic Approach
(Genome Sequencing, DNA arrays, proteomics)

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



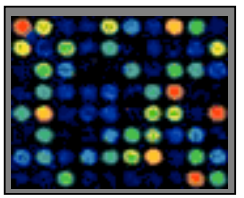
Complex Cellular Samples
bodyfluids, tissue

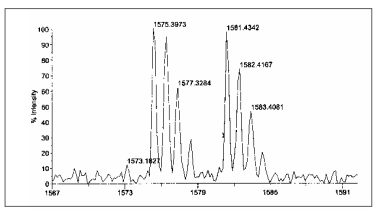
Dynamics
i.e. environmental + time

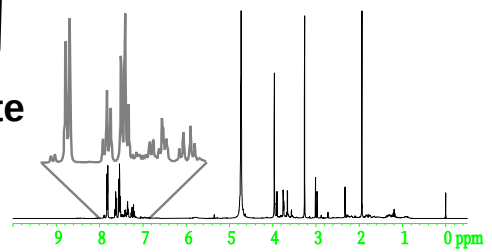
Gene

Protein

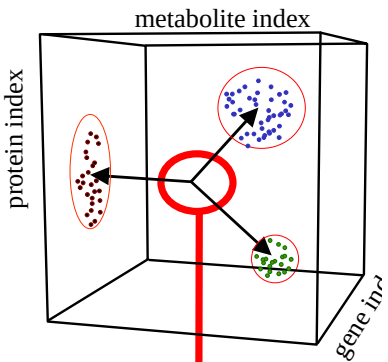
Metabolite







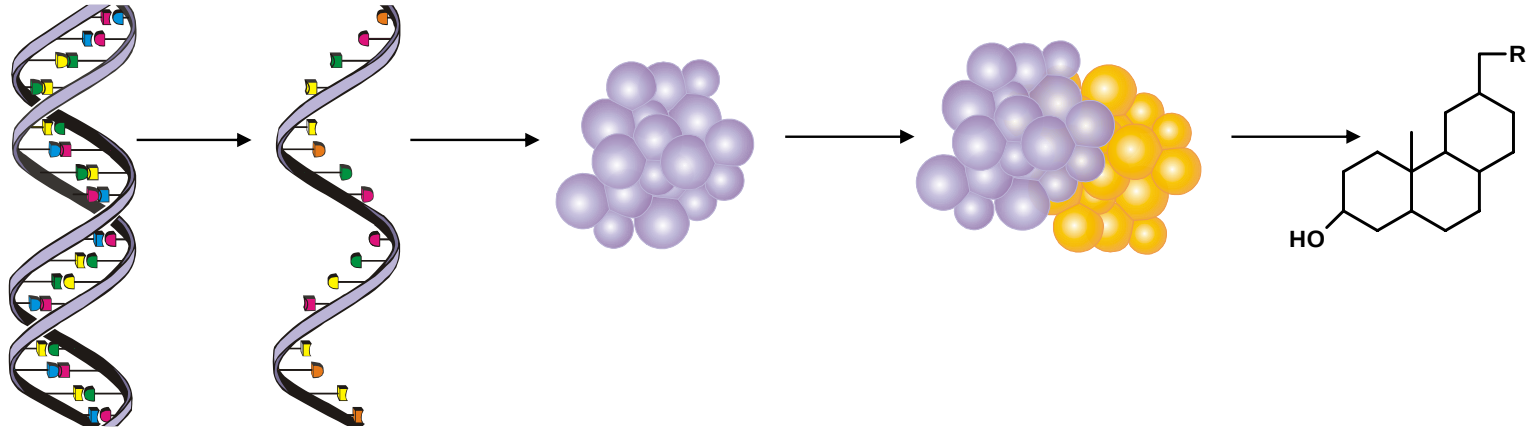
Quantitative Comparisons



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.