Web Resources for Bioinformatics, Biotechnology and Life Sciences Research

Presented at the 4th International Conference and Exhibition on Biometrics & Biostatistics , San Antonio, USA

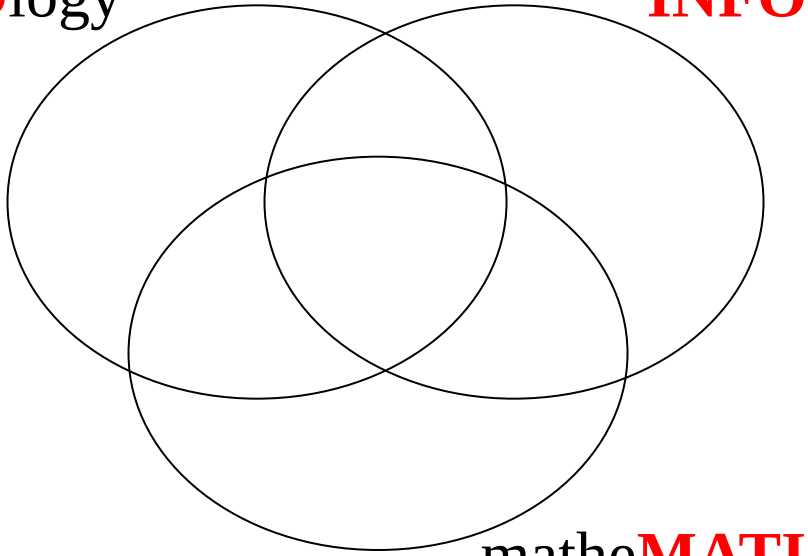
November 16, 2015 :16:55-17:15

Ajit Kumar Roy
Central Agricultural University,
Tripura, India

Outline/Contents

- **Introduction to bioinformatics**
- **Bioinformatics in Post genomic Era**
- **High-throughput technologies and experimental options**
- **Useful bioinformatics issues and databases**
- **Use of algorithm/ utilities, other useful tools**
- **Statistical methods**
- **Bioinformatics Links for Tutorials**
- **Bioinformatics Tutorial videos**
- **Bioinformatics Blogs**
- **Bioinformatics Tutorials**
- **Bioinformatics in Social Media**

Bioinformatics



BIOlogy

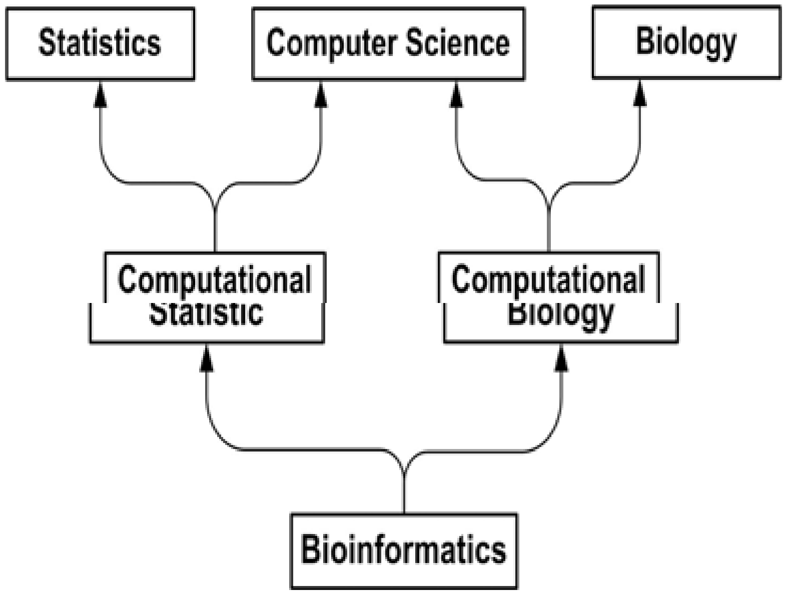
INFORmatics

mathe**MATICS**

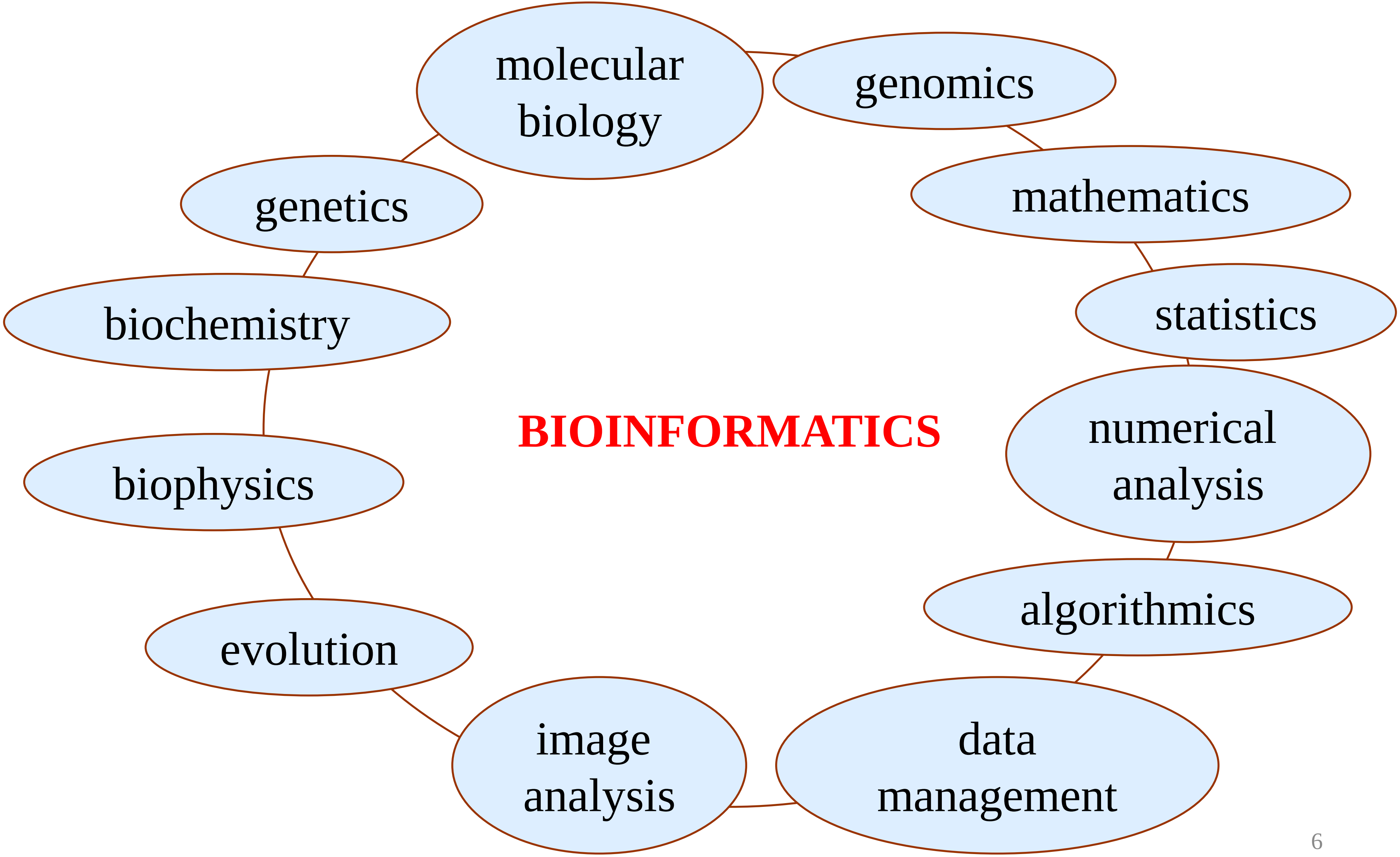
Simple definition for bioinformatics

- *“The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information“.*
- *Extraction of biological knowledge from complex data*

Interconnection Between Computer Science, Computational Statistics and Bioinformatics



Bioinformatics- A Multidisciplinarity-----



molecular biology

genomics

genetics

mathematics

biochemistry

statistics

BIOINFORMATICS

biophysics

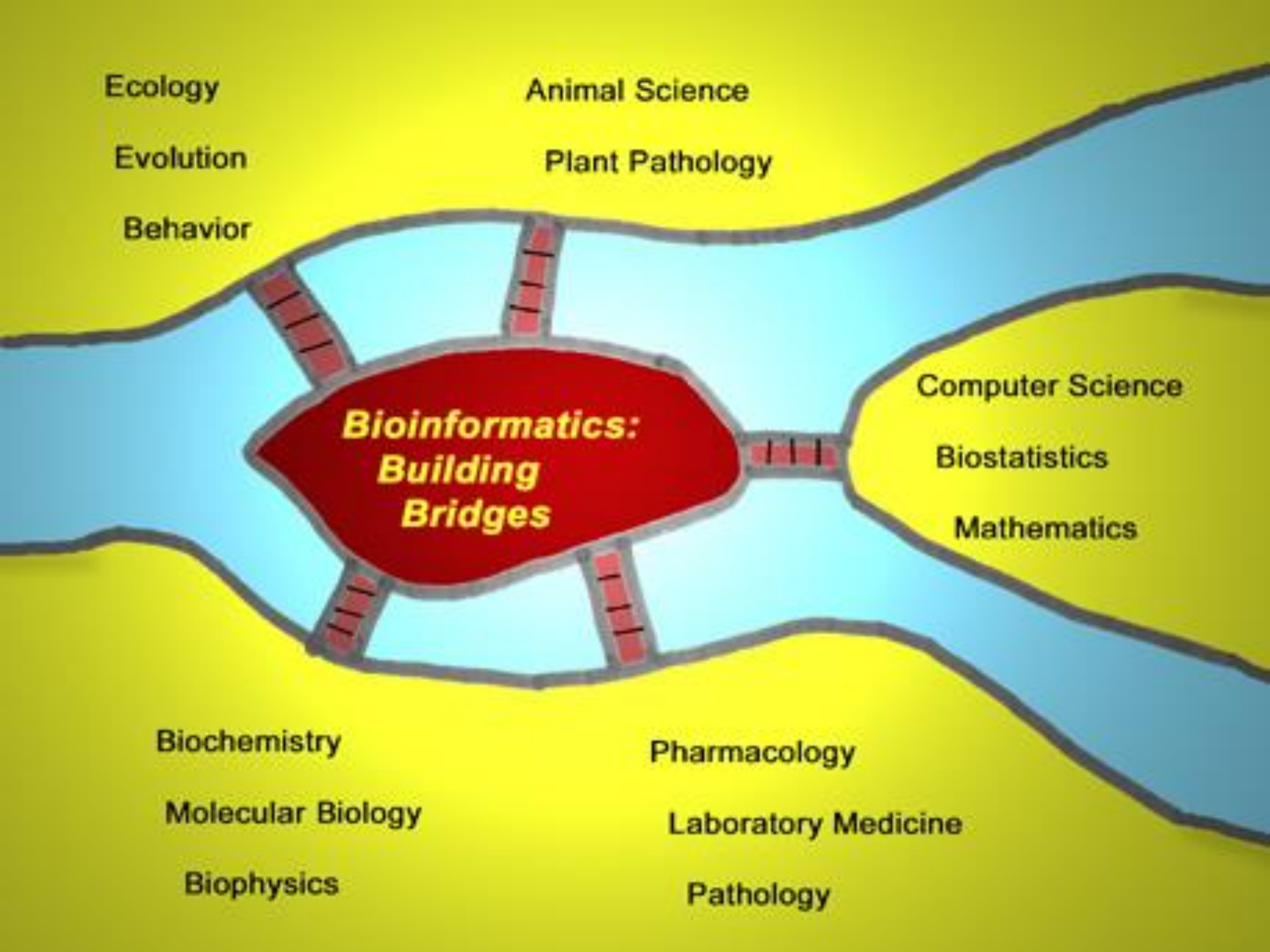
numerical analysis

evolution

algorithmics

image analysis

data management



Bioinformatics: Building Bridges

Ecology

Evolution

Behavior

Animal Science

Plant Pathology

Computer Science

Biostatistics

Mathematics

Biochemistry

Molecular Biology

Biophysics

Pharmacology

Laboratory Medicine

Pathology

Bioinformatics-*a merger of computer science and biology.*

- *Revolutionized our ability to do biology in much the same way as PCR and robotics changed the bench science.*
- *“the computational branch of molecular biology”*
- *a merger of computer science and biology.*

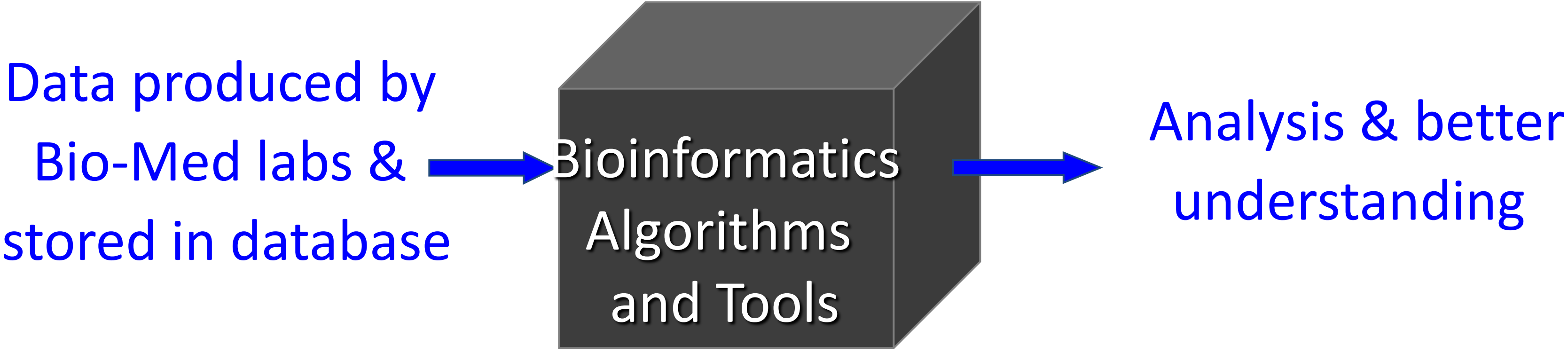
Goals of bioinformatics

- *Classify*
- *Identify patterns/ Pattern Recognition*
- *Make predictions*
- *Data Modelling, Creation of models & Prediction*
- *Assessment and Comparison*
- *Optimization*
- *Better utilize existing knowledge*

Why Bioinformatics ?

Enables large scale analysis and interpretation of data.

Provides computational methods for global understanding of biological data.



Data produced by Bio-Med labs & stored in database

Bioinformatics Algorithms and Tools

Analysis & better understanding

Goal: Make sense of bio-medical data using computer tools, and thereby bridge the gap between molecular biology and computer science.

In post genomic era a new language has been created for new biology

- *Genomics*
- *Functional Genomics*
- *Proteomics*
- *cDNA microarrays*
- *Global Gene Expression Patterns*

The human genome project (HGP)

1990-2003:
Aim: to reveal the blueprint of human biology (3,000,000,000 letters).

Rex Babin, California – The award winning cartoonist for the **Sacramento Bee**. Visit Rex's cartoon archives – E-Mail Rex

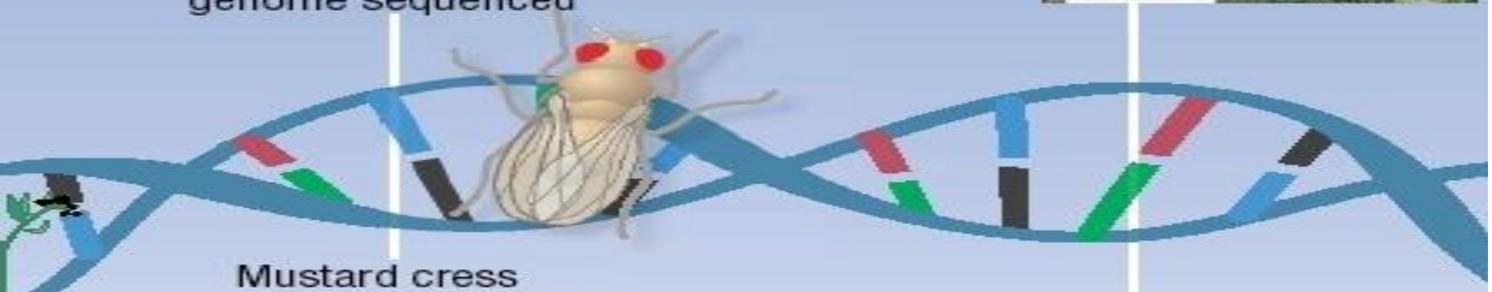
Francis Collins (chairman of the international project from the NIH):
"I think this is probably the most important scientific effort that mankind has ever mounted. That includes splitting the atom and going to the moon".

2000

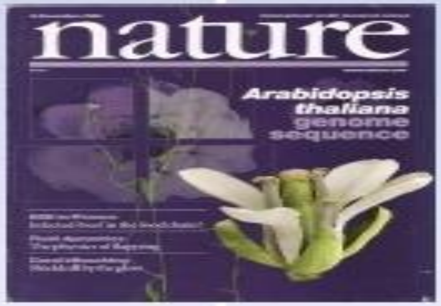
Draft version of human genome sequence completed

US President Bill Clinton and UK Prime Minister Tony Blair support free access to genome information

Fruitfly (*Drosophila melanogaster*) genome sequenced



Mustard cress (*Arabidopsis thaliana*) genome sequenced



2001

Draft version of human genome sequence published

10,000 full-length human cDNAs sequenced

Mammalian
Gene
Collection

June 26, 2000 at the Whitehouse

2001

The HGP consortium publishes its working draft in *Nature* (15 February), and Celera publishes its draft in *Science* (16 February).

Human genome: 3 billion bases

In total 23 billion bases were sequenced (7.5-fold coverage)

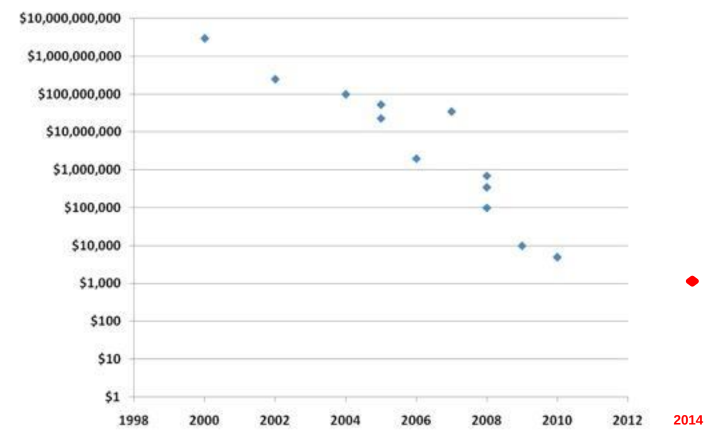
This comprises 23 Gbyte of data

Public Human Genome project
- -1990-2003
- - 3 billion US dollar

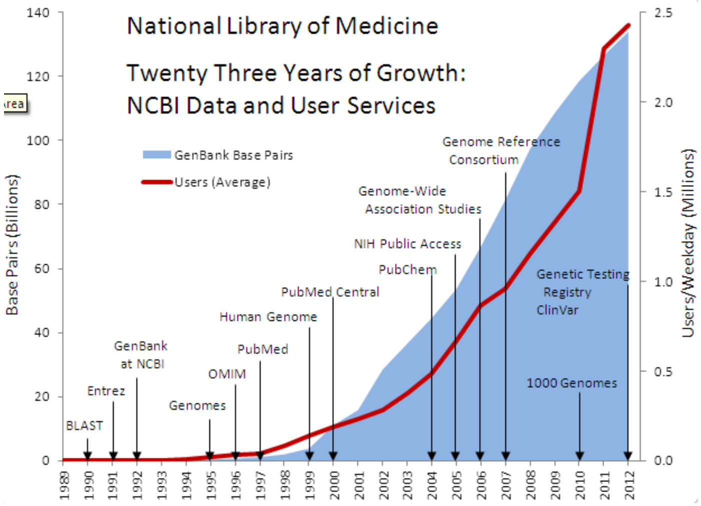
Private Celera genome project
- -1998 - 2001
- - Craig Venter
- - 300 million US dollar 2001

Genomes for everyone

The Cost of Sequencing a Human

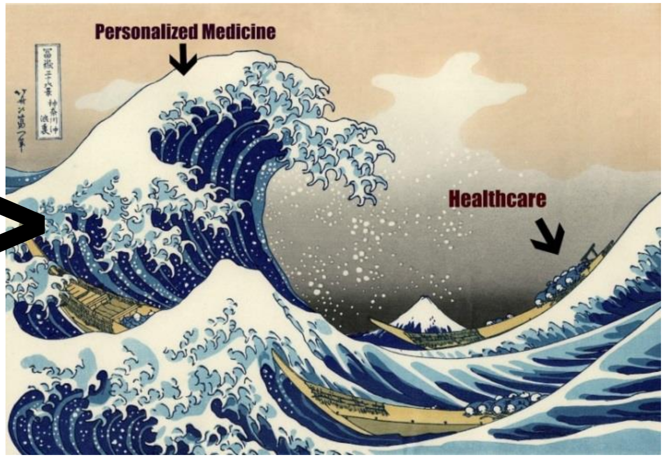


+



=>

=>



Five domains of Bioinformatics:

Large-scale biological data can be represented in different forms for different computation and analysis.

The common ones are:

- Sequence
- Structure
- Interaction
- Expression
- Function

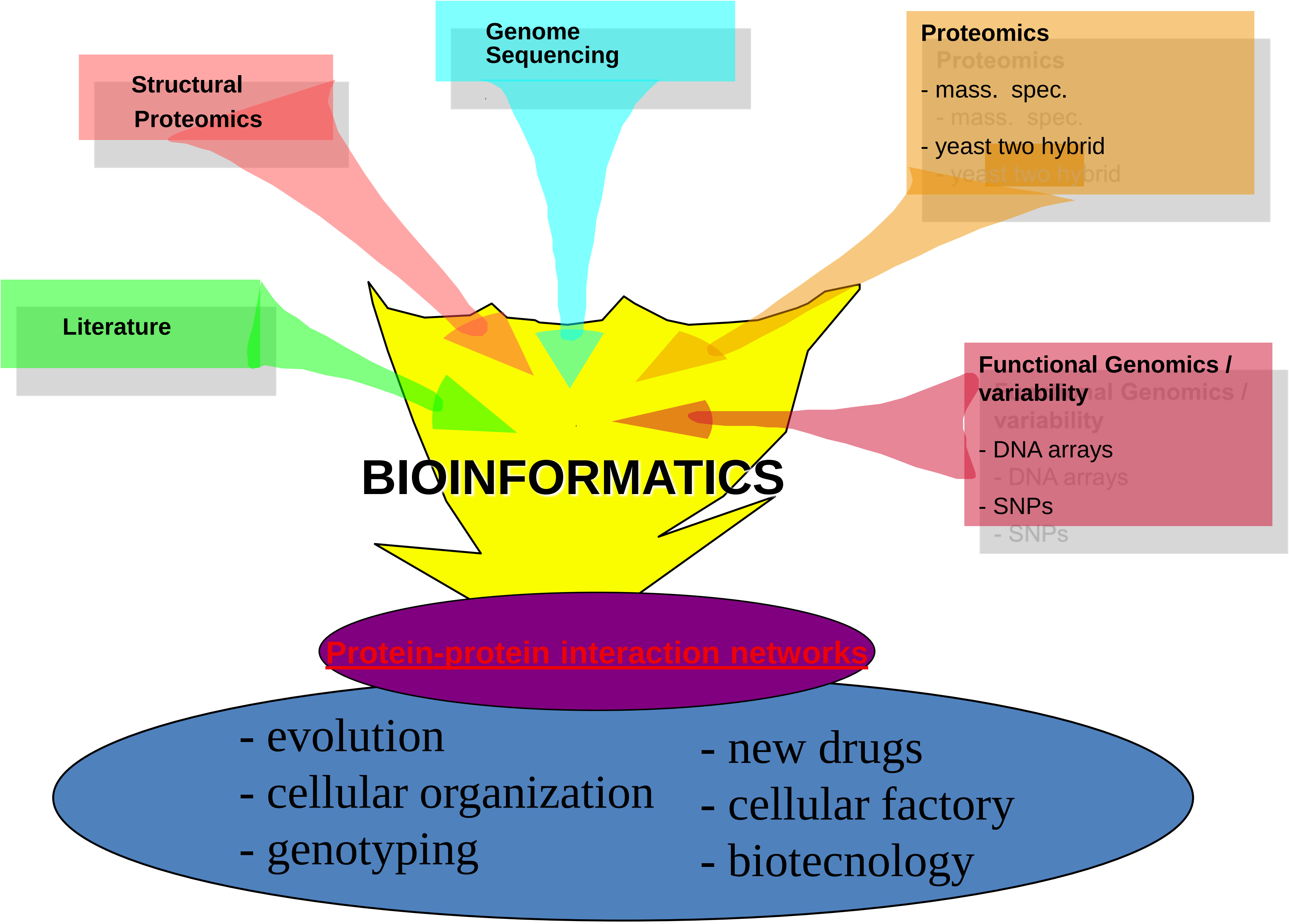
New Computational Tools are Needed for---

- *Sequencing*
- *Analyzing experimental data*
- *Representing vast quantities of information*
- *Searching*
- *Pattern matching*
- *Data mining*
- *Gene discovery*
- *Function discovery*

Before bioinformatics

- *In vivo* experiments
  - In the living organism
- *In vitro* experiments
  - In a test tube

Post genomic technologies



Structural Proteomics

Genome Sequencing

Proteomics

Proteomics

- mass. spec.

- mass. spec.

- yeast two hybrid

- yeast two hybrid

Literature

Functional Genomics / variability

Functional Genomics / variability

- DNA arrays

- DNA arrays

- SNPs

- SNPs

BIOINFORMATICS

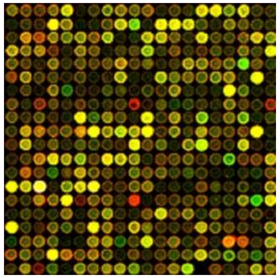
Protein-protein interaction networks

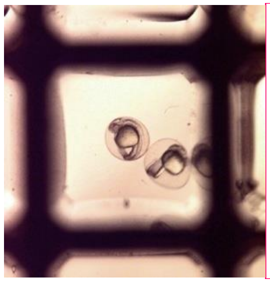
- evolution
- cellular organization
- genotyping

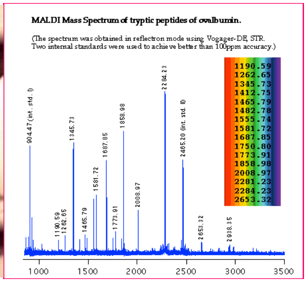
- new drugs
- cellular factory
- biotecnology

Phenotype

- *mRNA level*
- *Protein expression*
- *Protein structure*
- *Cell morphology*
- *Tissue morphology*
- *System physiological functions*
- *Behavior*
- …







MALDI Mass Spectrum of tryptic peptides of ovalbumin.

(The spectrum was obtained in reflectron mode using Vogager-DE, STR. Two internal standards were used to achieve better than 100ppm accuracy.)

High-throughput techniques

High throughput protein crystalization

Massive parallel sequencing

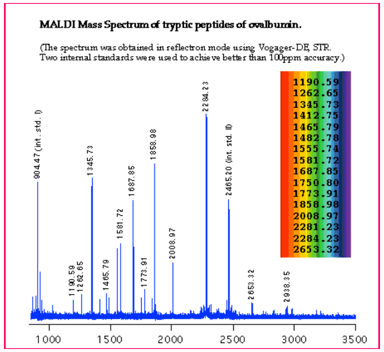
Mass spectrometry

Microarray

High throughput cell imaging

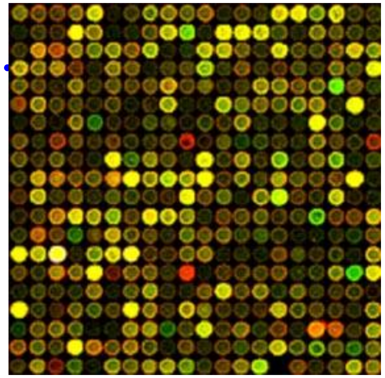
High throughput in vivo screening

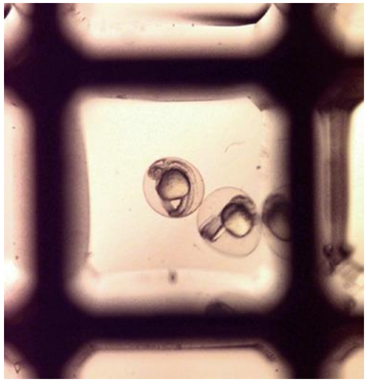
...



MALDI Mass Spectrum of tryptic peptides of ovalbumin.

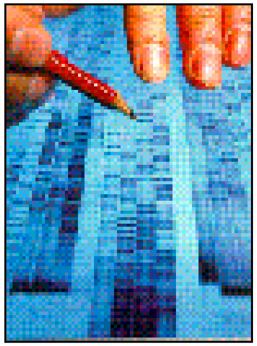
(The spectrum was obtained in reflectron mode using Vogager-DE, STR. Two internal standards were used to achieve better than 100ppm accuracy.)



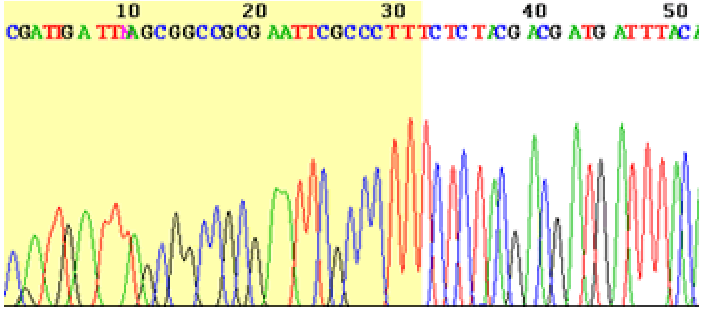


The Biotechnology Revolution Creates High-throughput Data

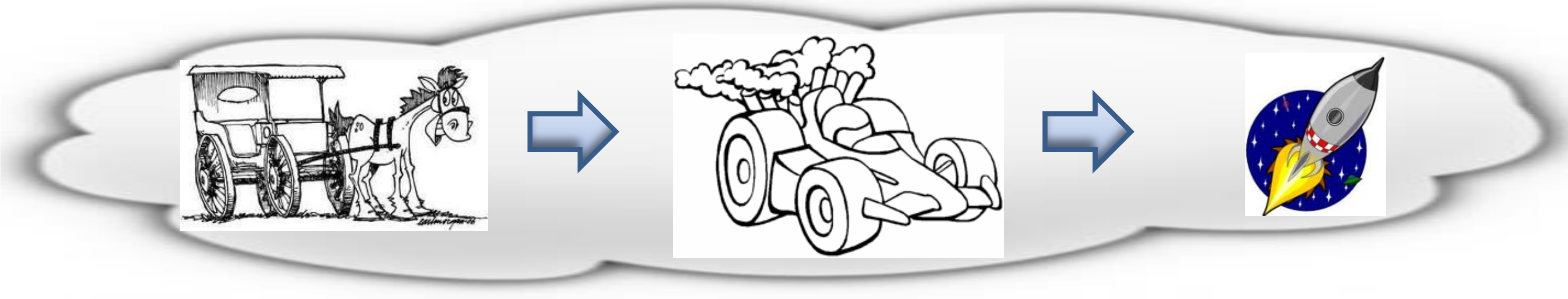
Late 60's, early 70's



1980s



21st century



http://chagall.med.cornell.edu/BioinfoCourse/presentations2010/Lecture1_2010.pdf

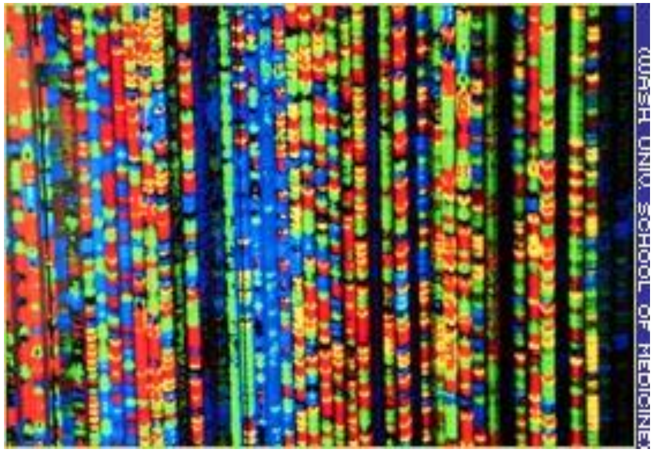
How to extract the information?

Computational tools

- *Building the databases*
- *Perform analysis/extract features*
- *Data mining*
- *Classification/statistical learning*
- *Visualization/representation*

Biological information!!

PG Era-High throughput DNA sequencing Centre



Custom-designed factory-style conveyor belt robots:

perform all functions from purifying DNA from bacterial cultures through setting up and purifying sequencing reactions.

Whitehead institute

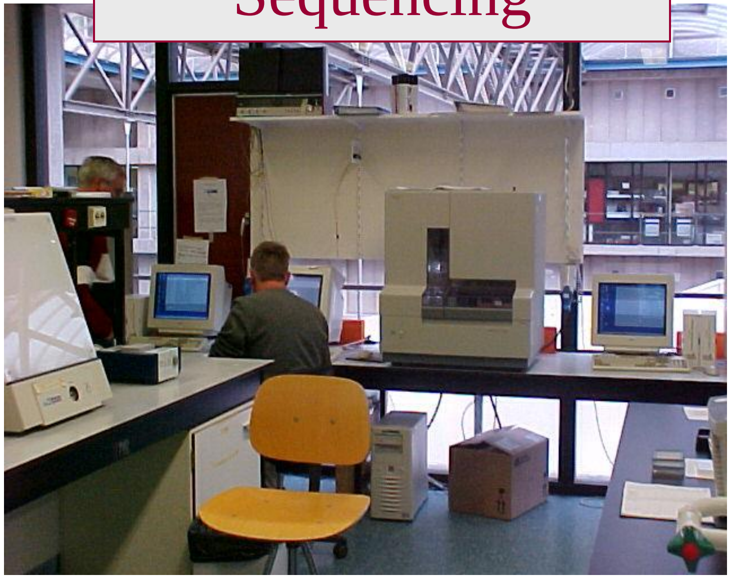
Automated sequencing

DNA isolation

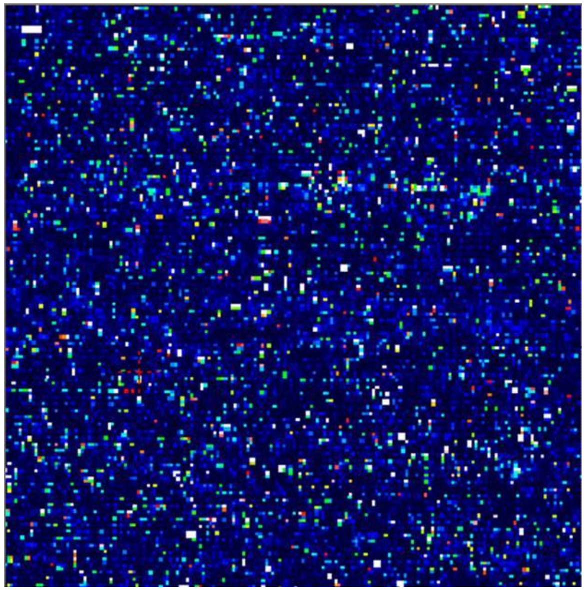
Beckman Biomek FX

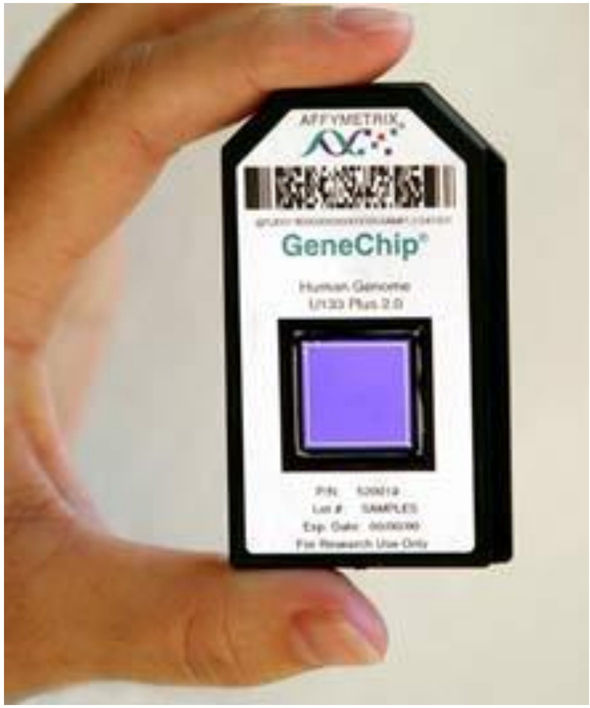
Mass spectrometer

Sequencing

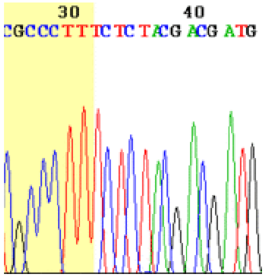


Affymetrix gene expression





Traditional versus high throughput DNA sequencing



http://en.wikipedia.org/

Sanger, one run:

? hours (human genome took 15 years)

1-384 sequences

300-1000 nt per sequence

1 KB - 384 KB data

= 1,000-384,000 bases

Year: 1963 – now

http://www.454.com/

Roche 454, one run:

7.5 hours

1,000,000 sequences

500 nt per sequence

35 GB data (including images)

500 MB data (excluding images)

= 500,000,000 bases

Year: 2005 – now

http://solid.appliedbiosystems.com/

ABI SoLiD, one run:

3-5 days

150,000,000 sequences

50-100 nt per sequence

2-4 TB data (raw data)

= 15,000,000,000 bases

Year: 2005 – end 2011

Three main aspects:

- There are three main aspects need to be addressed to handle data in PGEra.

I. *The first is that it is a very systematic way of dealing with biological data. Therefore, constructing an infrastructure such as large scale database and server systems for genomes and proteomes is an important part of it.*

II. *The second is that it views the processes and mechanisms of life as information processing.*

III. *The third aspect is that it is multi-disciplinary employing experimental biology, theoretical science and computers*

A bioinformatics laboratory

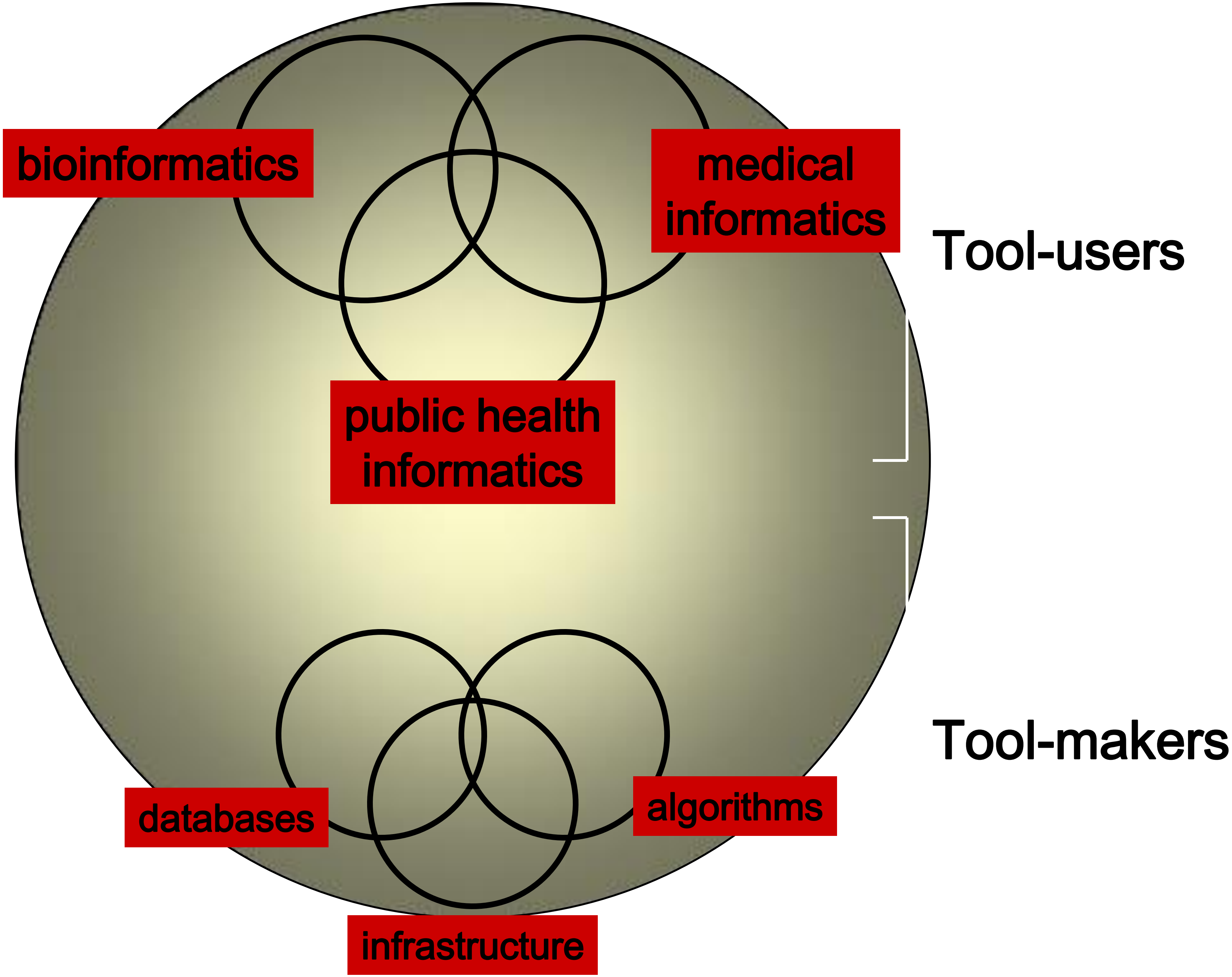
The Cell

Statistics

Algorithms

JAVA, PHP

Biomedical researchers are analyzing and interpreting data



bioinformatics

medical informatics

public health informatics

Tool-users

Tool-makers

databases

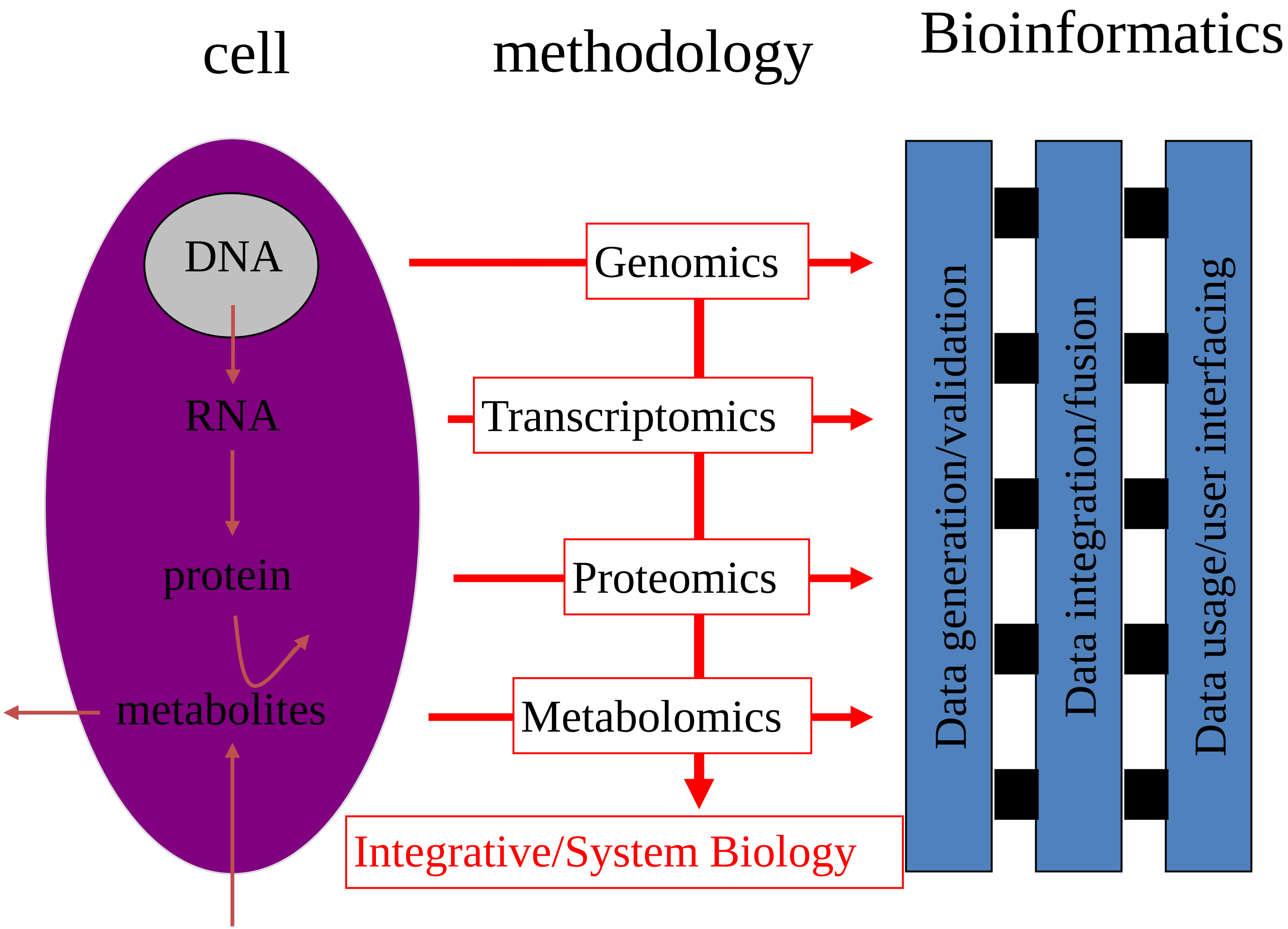
algorithms

infrastructure

The “*omics*” Series

- **Genomics**
  - **Gene identification & charaterisation**
- **Transcriptomics**
  - **Expression profiles of mRNA**
- **Proteomics**
  - **functions & interactions of proteins**
- **Structural Genomics**
  - **Large scale structure determination**
- **Cellinomics**
  - **Metabolic Pathways**
  - **Cell-cell interactions**
- **Pharmacogenomics**
  - **Genome-based drug design**

Role of bioinformatics



cell

methodology

Bioinformatics

Use of Bioinformatics

- DNA analysis
  - Genome sequencing
    - *Sequence assembly*
    - *Sequence/gene annotations*
    - *Genefinding/Sequence translation tools*
    - *Sequence Similarity searching (eg. BLAST, ClustalW)*
    - *Comparison between genomes*
    - *Evolution of sequences (Phylogenetic analysis)*
    - *Gene expression*

Use of Bioinformatics (..contd.)

- Protein analysis
  - Structure
    - *X-ray crystallography*
    - *Homology based models*
    - *Drug designing*
  - Sequence
    - *Sequence similarity*
    - *Protein family assignments*
    - *Conserved motifs*
    - *Proteomics data analysis*
    - *Protein Evolution*

Uses of Bioinformatics (..contd.)

- Other uses:
  - *Drug designing*
  - *Vaccine development*
  - *Dairy technology*
  - *Forensics*
  - *Crop improvement*
  - *Designing enzymes for detergents*
  - *Genetic counseling*

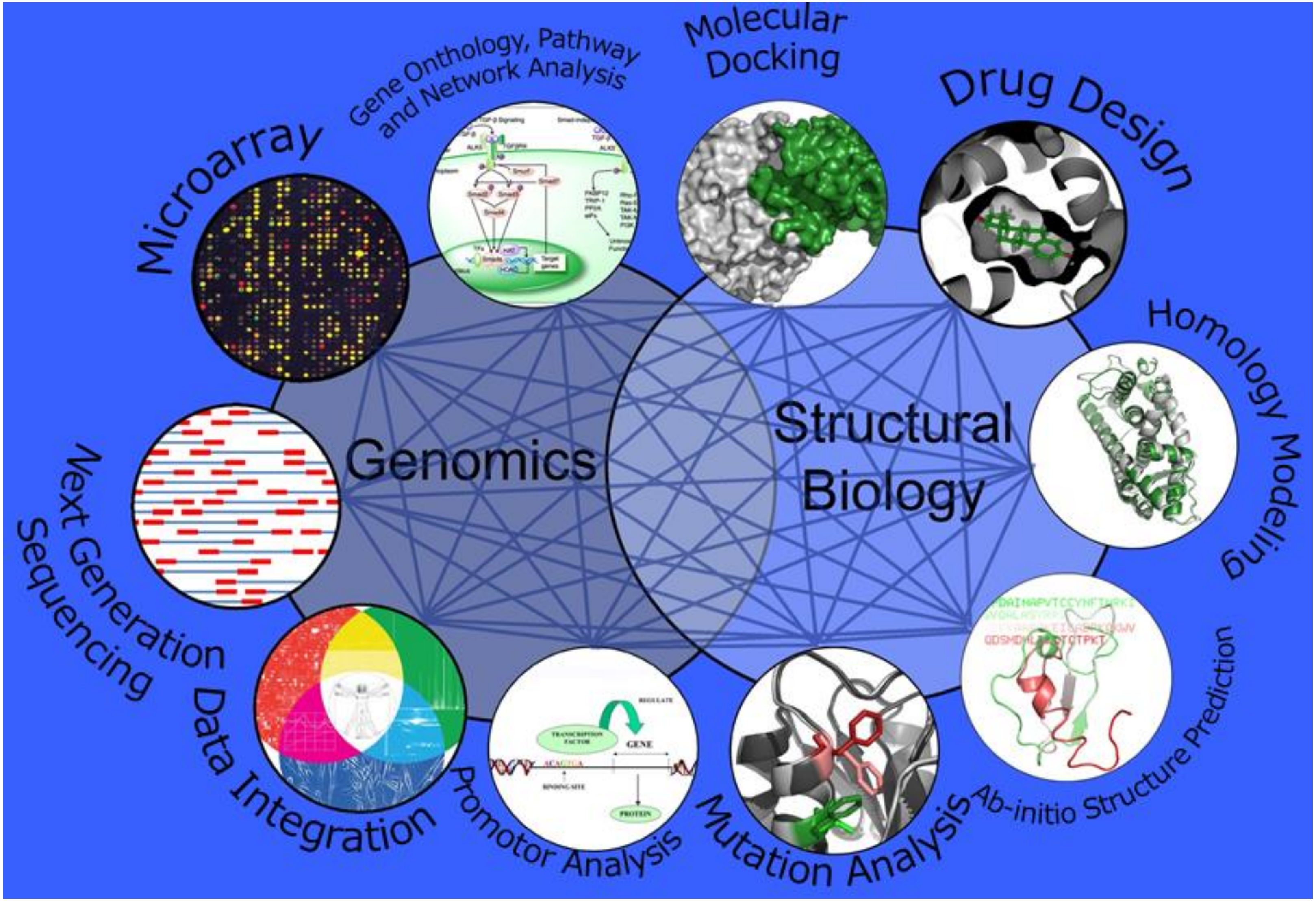
Recent events making bioinformatics more important

- *Exponential expansion of biological information*
- *Expansion of multiple types of information*
- *Cheaper high throughput technologies*
- *Improvement in computation power*
- *Lack of standards/quality*
- *Need for micro and macro analysis*
- *Need for better algorithms*

Related Fields

Computational Biology
Cheminformatics
Medical Informatics
Genomics
Proteomics
Pharmacogenomics
Pharmacogenetics

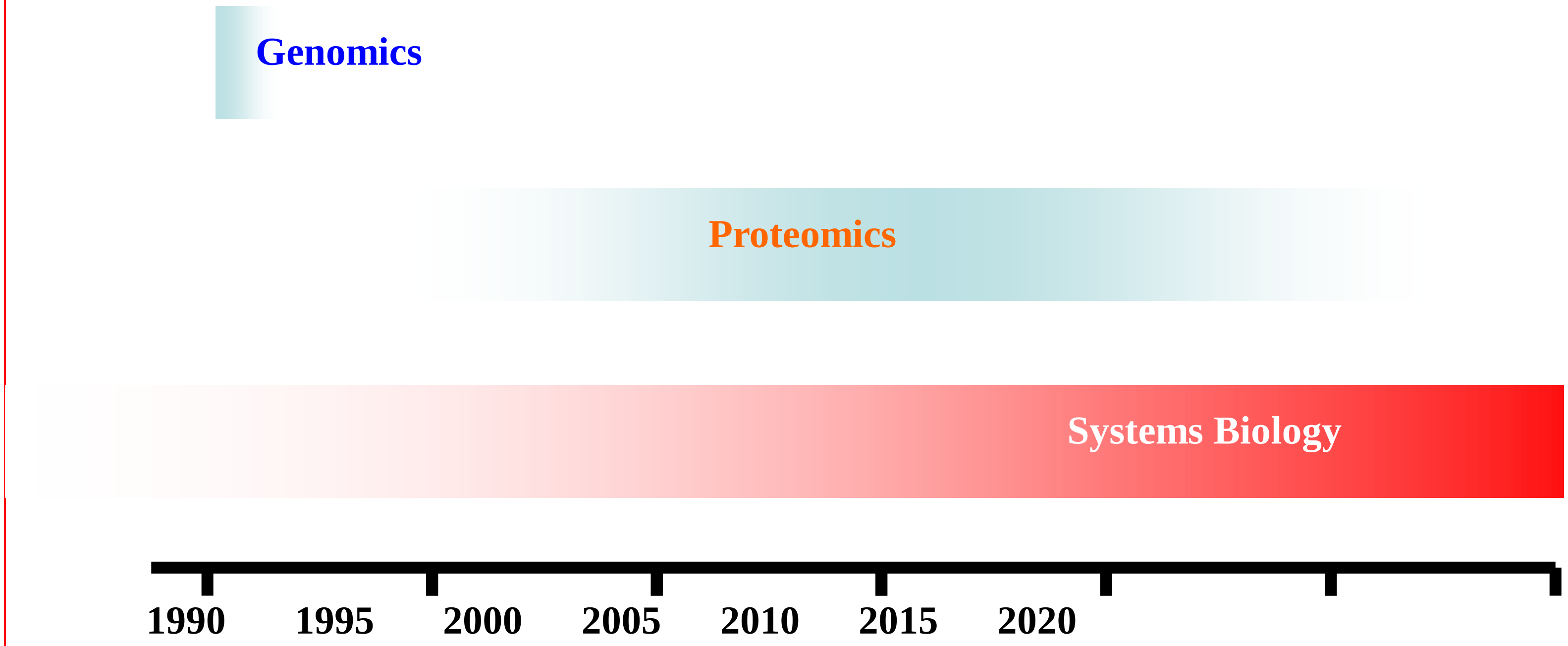
Bioinformatics topics are all linked together



- ***We are in the midst of a "Golden Era" of biology***

- ***The revolution is mostly about treating biology as an information science, not only specific biochemical technologies.***

Genomics, Proteomics & Systems Biology



Genomics

Proteomics

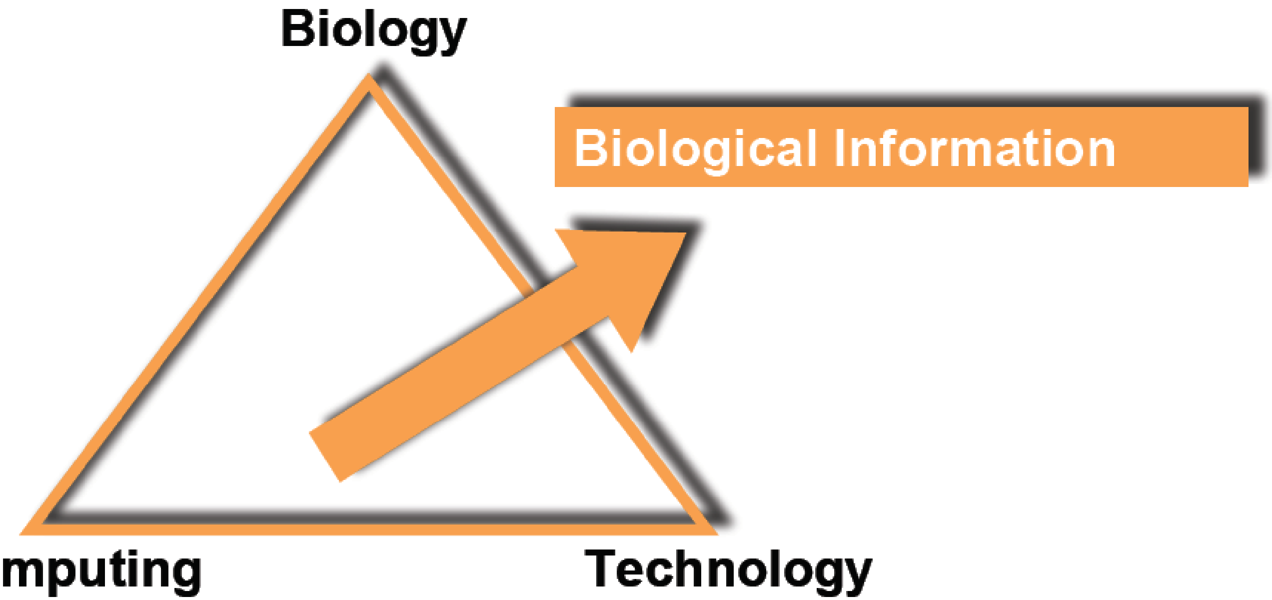
Systems Biology

1990 1995 2000 2005 2010 2015 2020

What is Systems Biology?

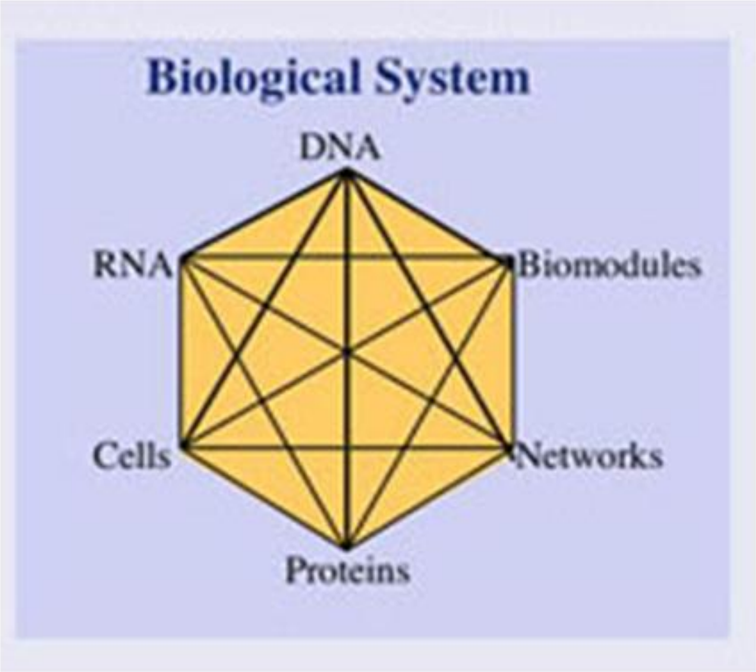
The
Systems Biology
Triangle

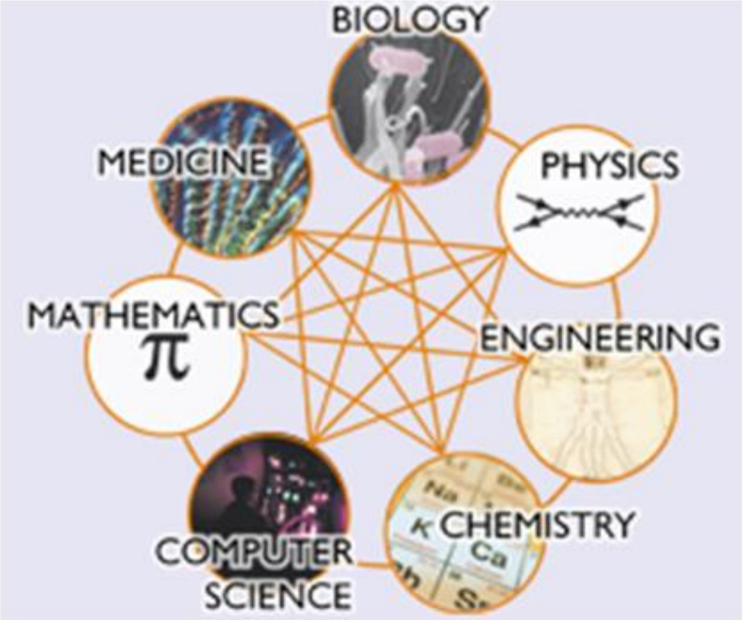
CC-SPMD
Competence Center for
Systems Physiology
and Metabolic Diseases

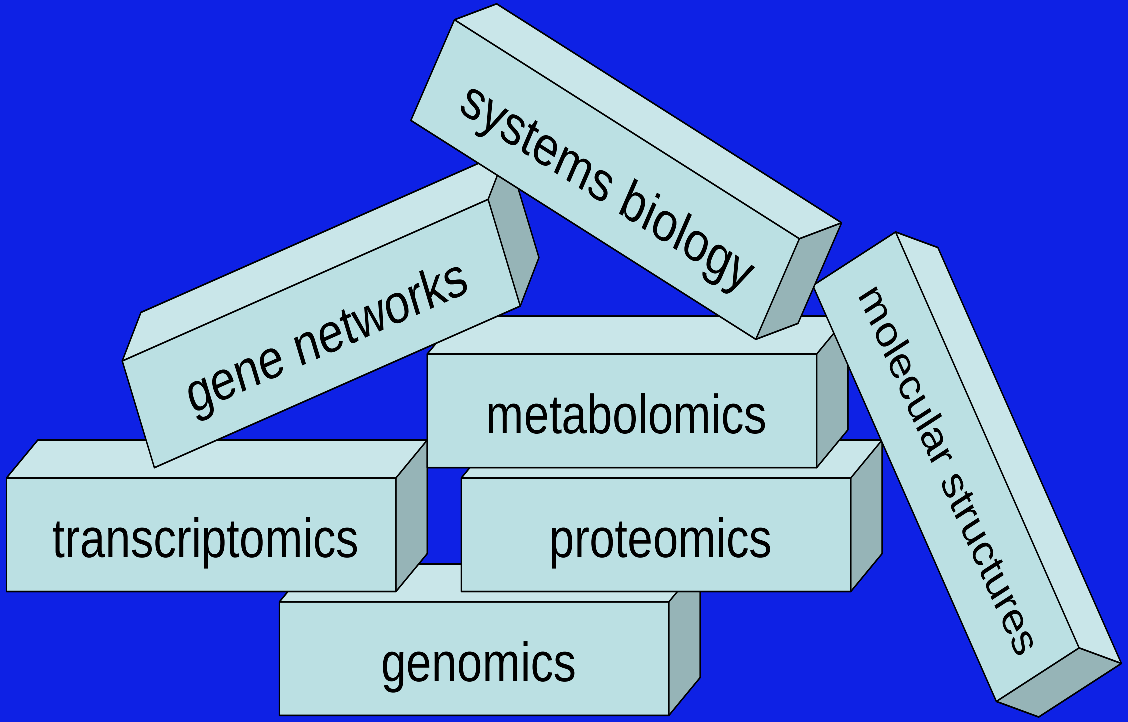


© Ruedi Aebersold

Systems Biology*







systems biology

gene networks

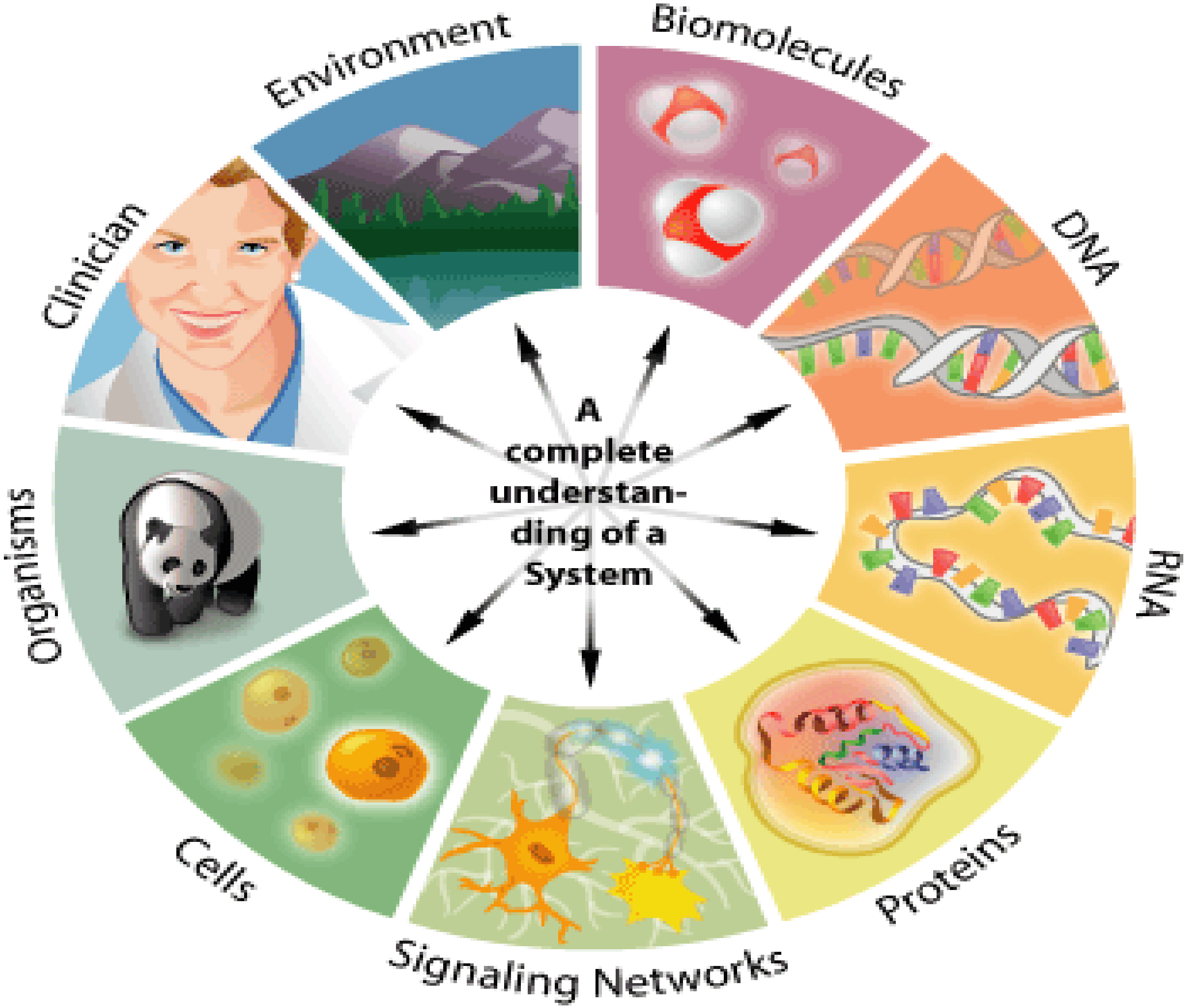
molecular structures

metabolomics

transcriptomics

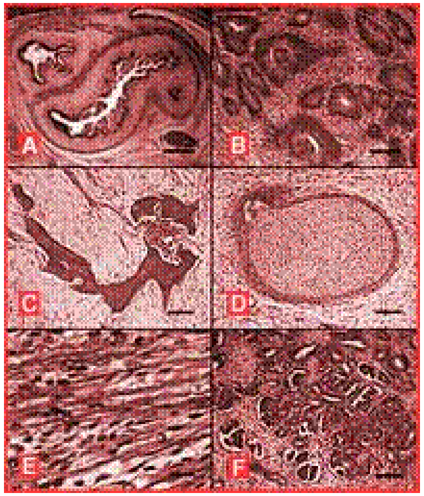
proteomics

genomics

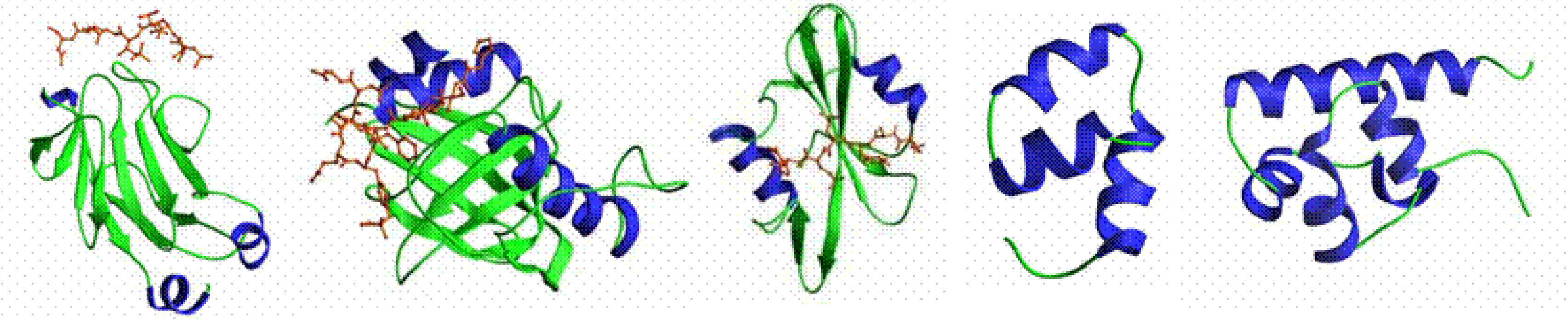

Environment
Biomolecules
DNA
RNA
Proteins
Signaling Networks
Cells
Organisms
Clinician
A complete understanding of a System

Building Blocks in Biology*

10^7 species

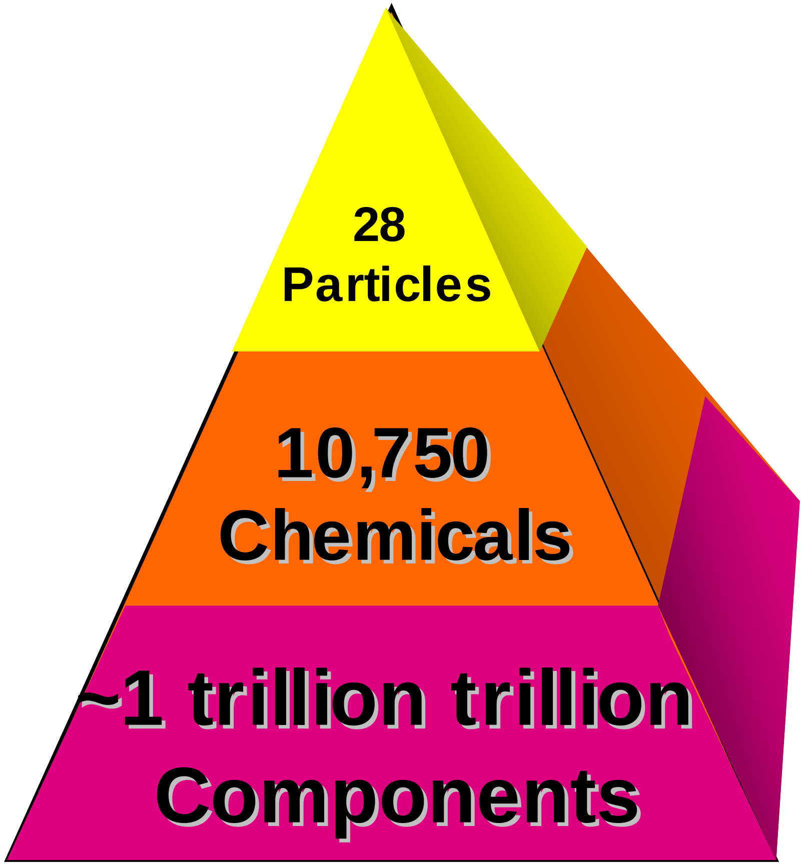


10^{12} cells/ organism



10^5 proteins/species

A Matter of Scale*



28 Particles

10,750 Chemicals

~1 trillion trillion Components

Physics

Chemistry

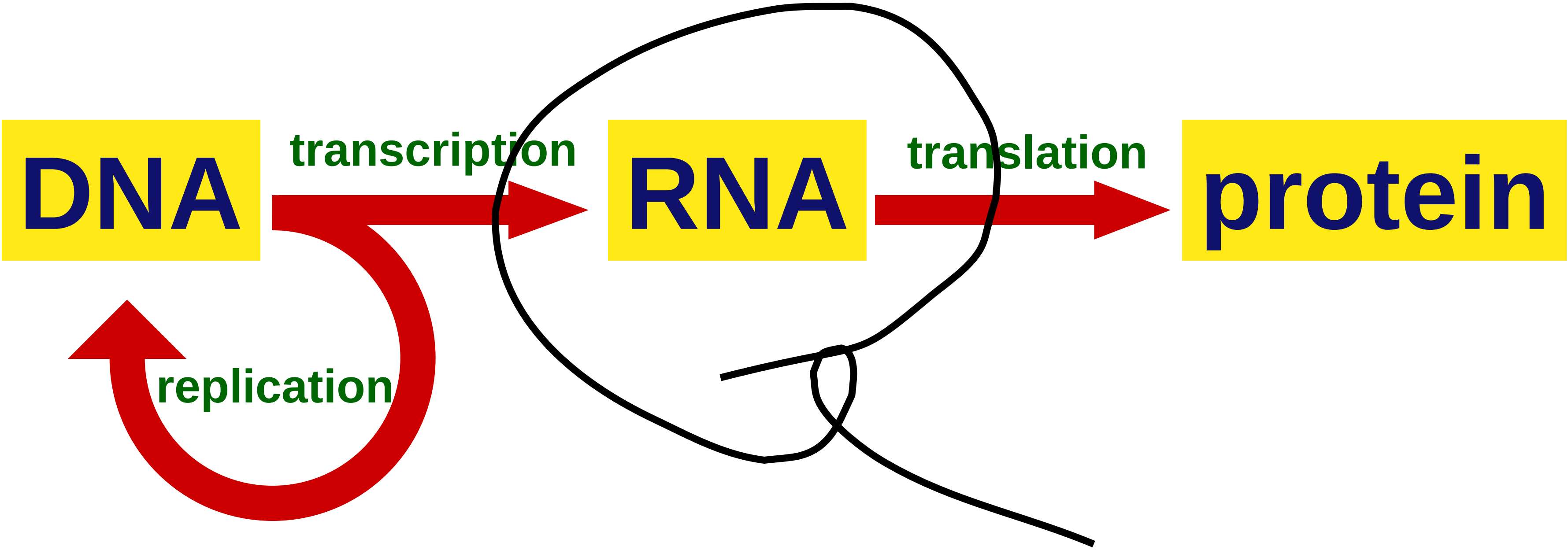
Biology

Some Fact about Molecular Biology

- ***The instruction set for a cell is contained in its chromosomes.***
- ***Each chromosome is a long molecule called DNA.***
- ***Each DNA molecule contains 100s or 1000s of genes.***
- ***Each gene encodes a protein.***
- ***A gene is transcribed to mRNA in the nucleus.***
- ***An mRNA is translated to a protein in a ribosome.***

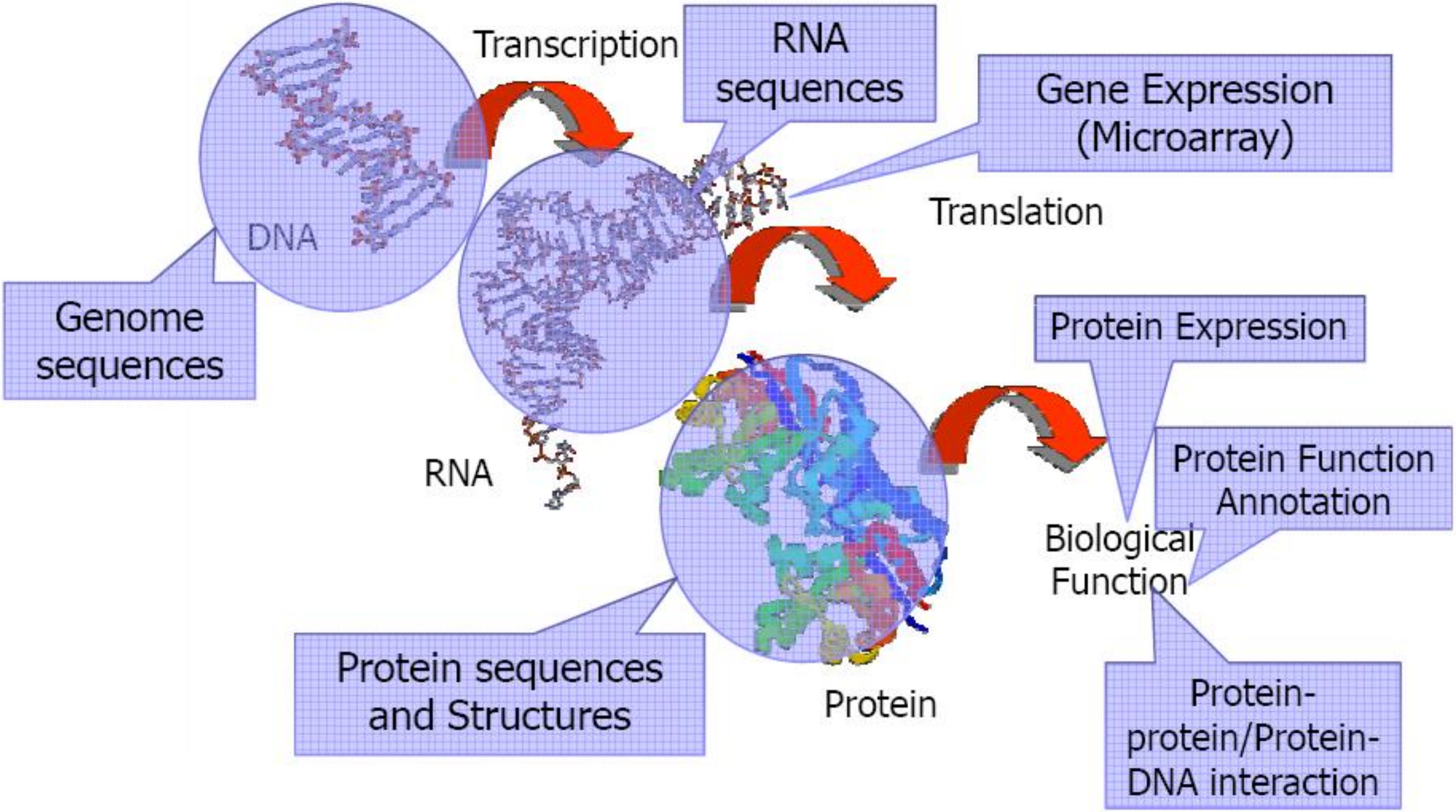
The Central Dogma of Molecular Biology

DNA is ***transcribed*** into RNA which is then ***translated*** into protein



Measured by Microarray

Biological Data



DNA

Transcription

RNA

Translation

Protein

Biological Function

Genome sequences

RNA sequences

Gene Expression (Microarray)

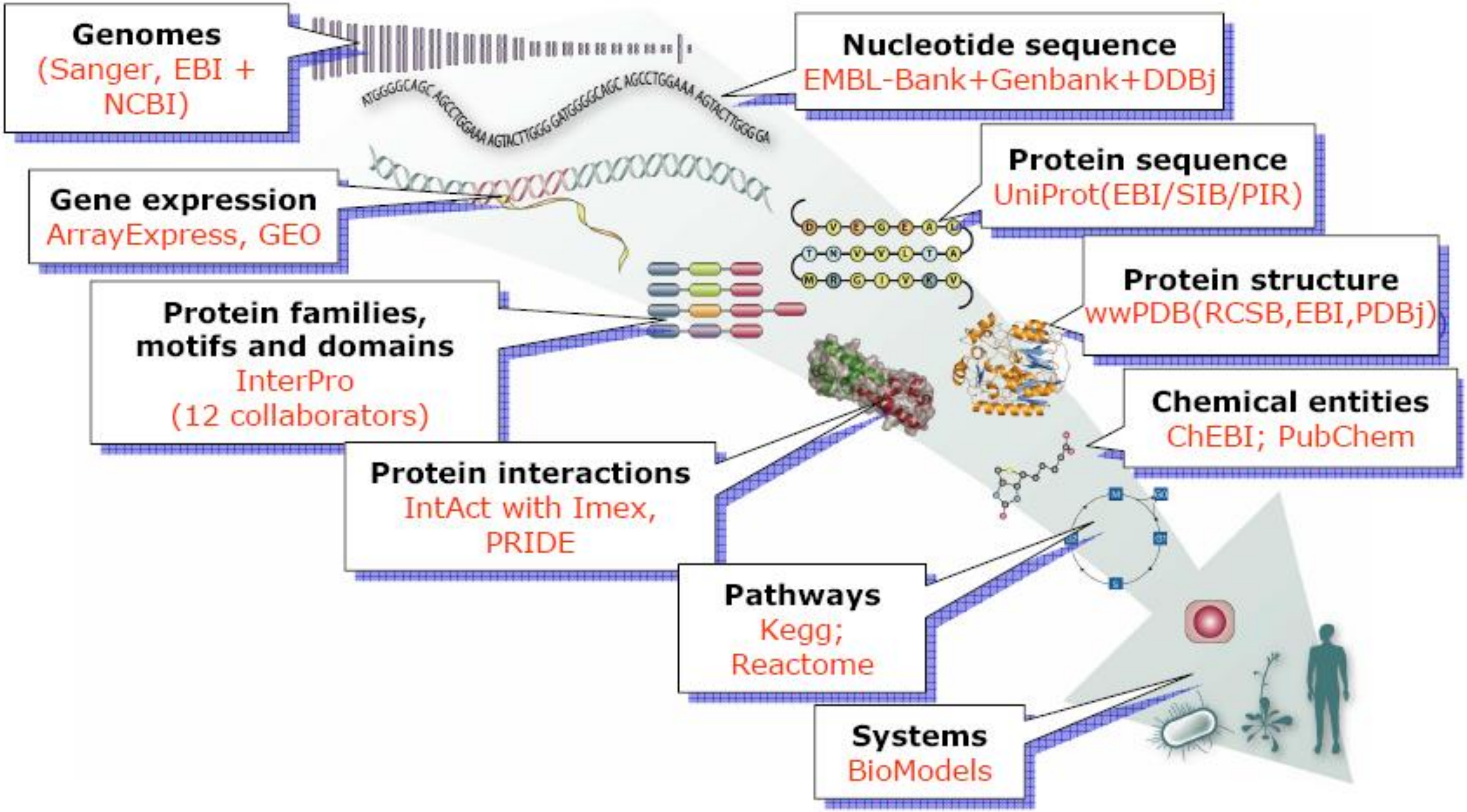
Protein sequences and Structures

Protein Expression

Protein Function Annotation

Protein-protein/Protein-DNA interaction

Core Molecular Data



Genomes
(Sanger, EBI + NCBI)

Nucleotide sequence
EMBL-Bank+Genbank+DDBj

Gene expression
ArrayExpress, GEO

Protein sequence
UniProt(EBI/SIB/PIR)

Protein families, motifs and domains
InterPro
(12 collaborators)

Protein structure
wwPDB(RCSB,EBI,PDBj)

Protein interactions
IntAct with Imex, PRIDE

Chemical entities
ChEBI; PubChem

Pathways
Kegg;
Reactome

Systems
BioModels

Need for databases in Biology?

- ***Need for storing and communicating large datasets has grown.***
- ***Need to disseminate biological information.***
- ***Provide Organized data for analysis friendly retrieval.***
- ***Need to make biological data available in computer-readable form.***

Different classifications of databases

- Type of data
  - *nucleotide sequences*
  - *protein sequences*
  - *proteins sequence patterns or motifs*
  - *macromolecular 3D structure*
  - *gene expression data*
  - *metabolic pathways*
  - *proteomics data*

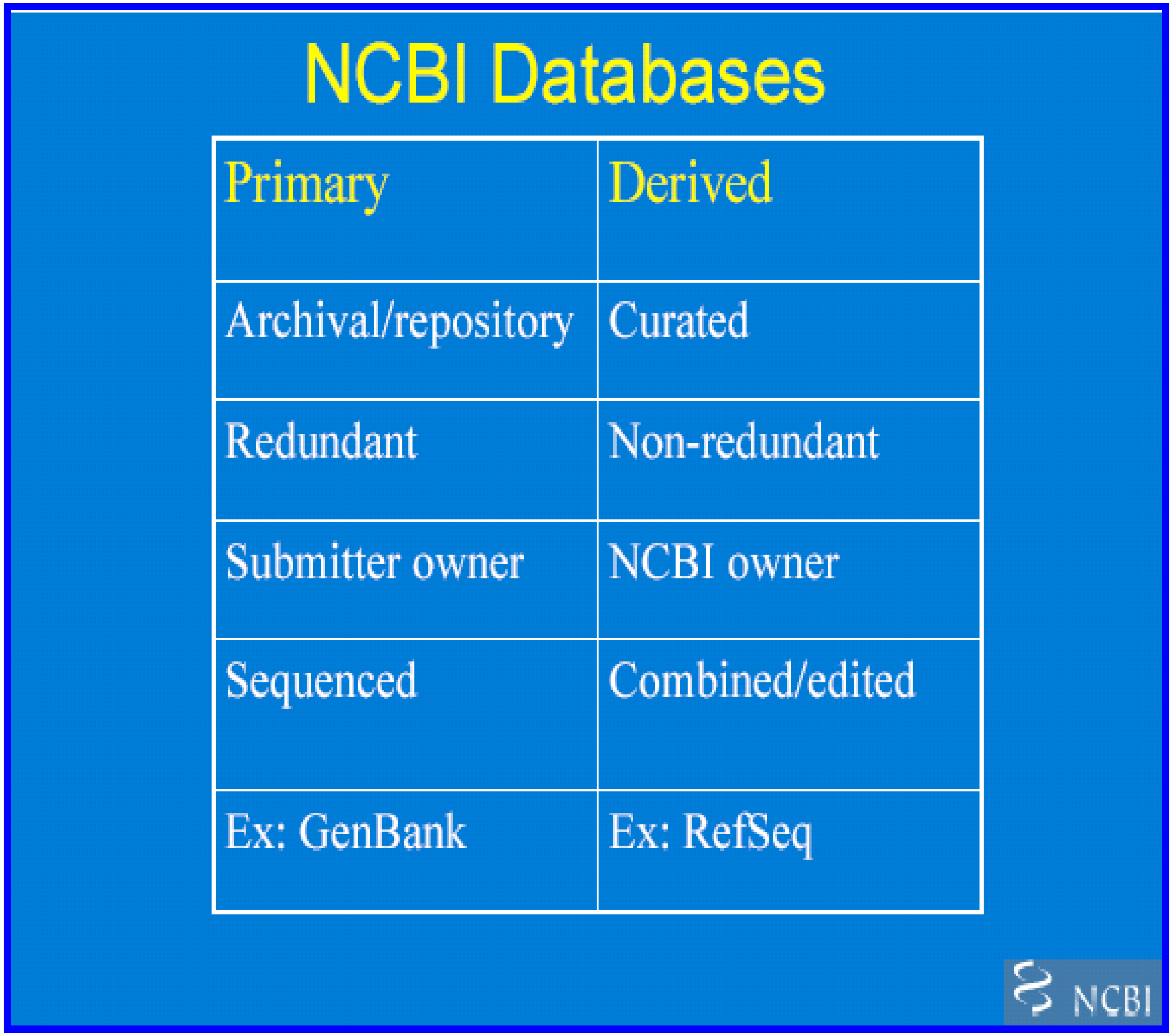
Different classifications of databases….

- Primary or derived databases
  - **Primary databases**: experimental results directly into database
  - **Secondary databases**: results of analysis of primary databases
  - **Aggregate of many databases**
    - Links to other data items
    - Combination of data
    - Consolidation of data

Different classifications of databases….

- Availability
  - Publicly available, no restrictions
  - Available, but with copyright
  - Accessible, but not downloadable
  - Academic, but not freely available
  - Proprietary, commercial; possibly free for academics

Molecular databases



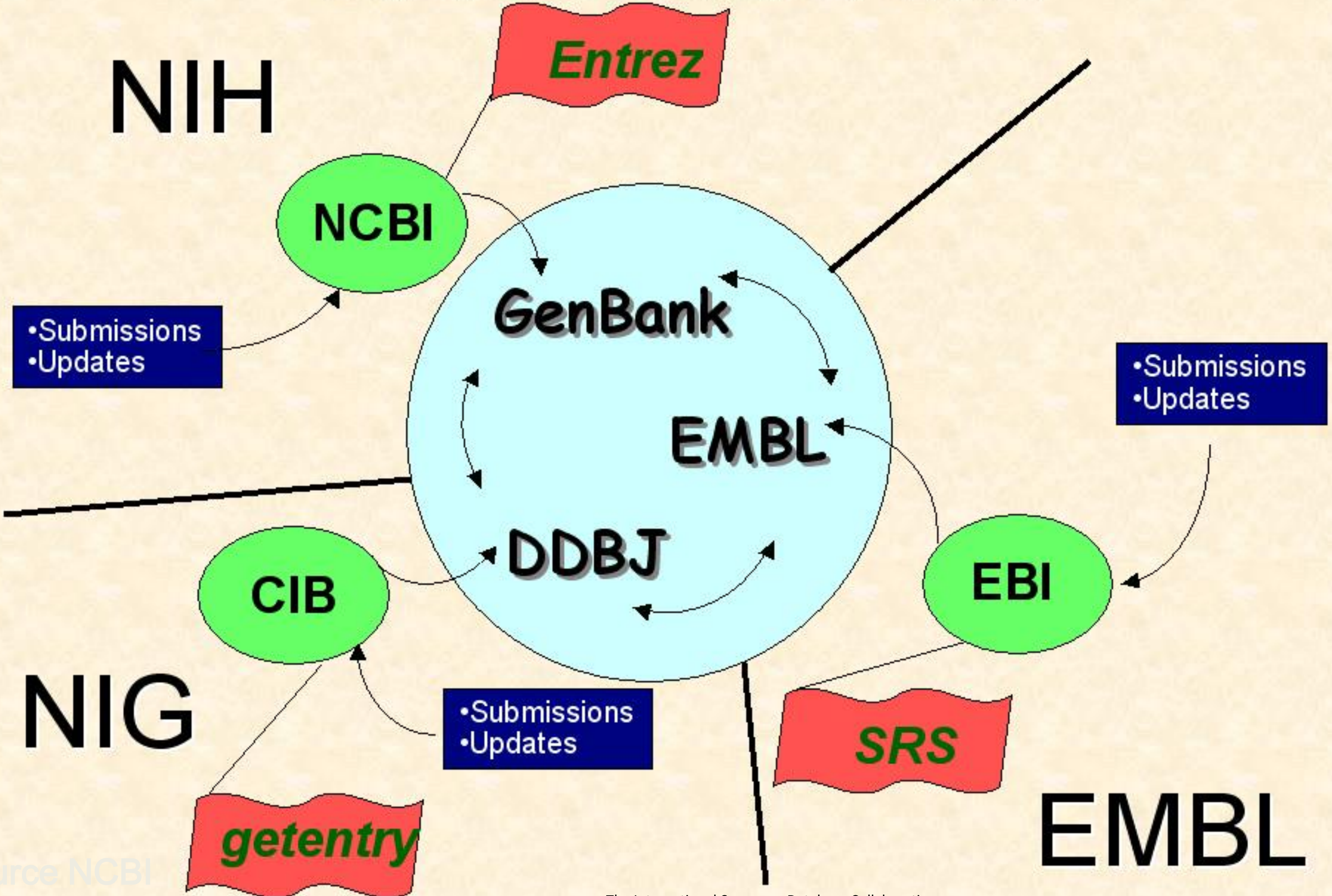
NCBI Databases

Primary	Derived
Archival/repository	Curated
Redundant	Non-redundant
Submitter owner	NCBI owner
Sequenced	Combined/edited
Ex: GenBank	Ex: RefSeq

Where to look for databases?

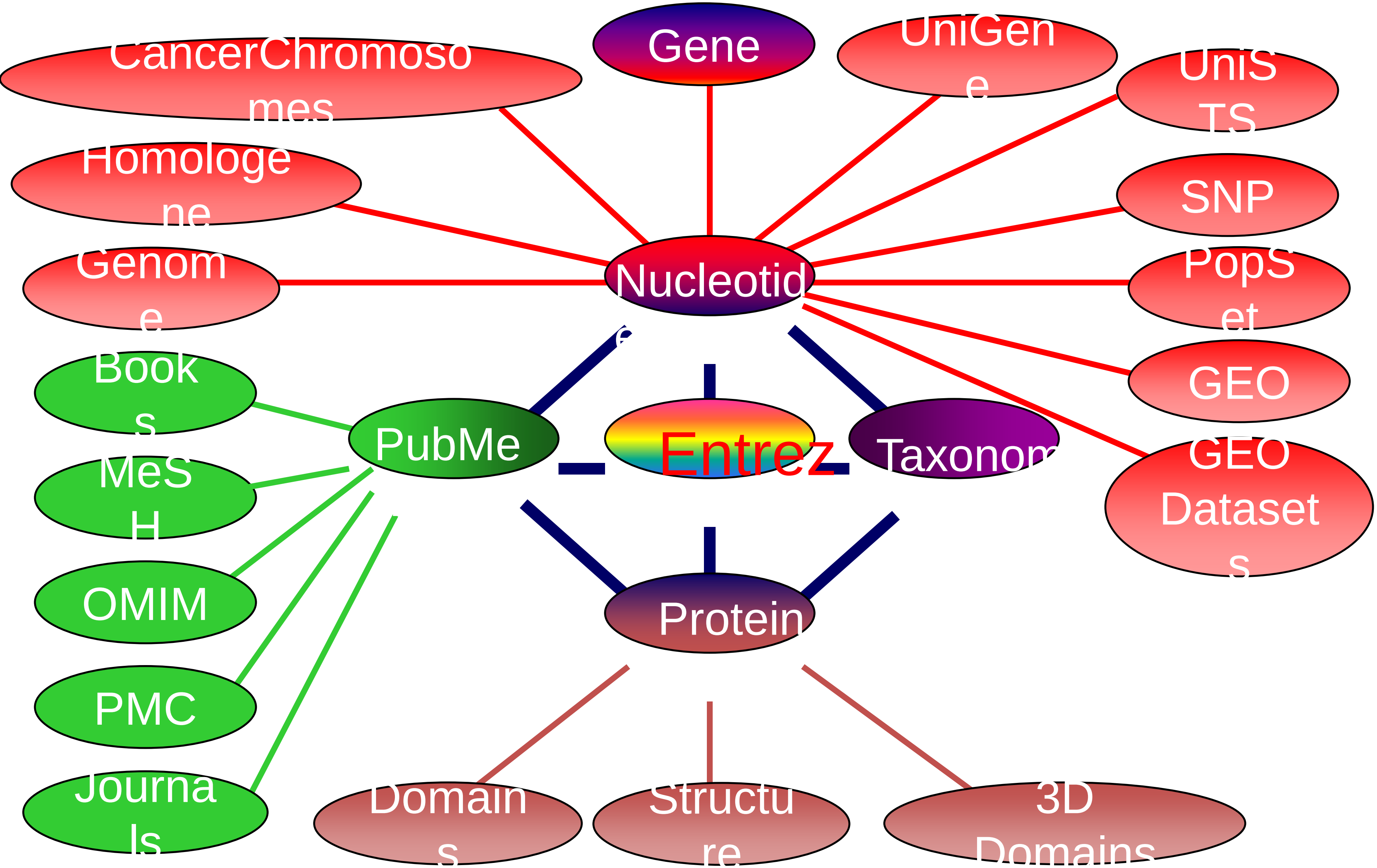
- Search Engines
- Websites like:
  - http://www.biophys.uni-duesseldorf.de/BioNet/Pedro/rt_all.html
  - www.expasy.ch

The International Sequence Database Collaboration



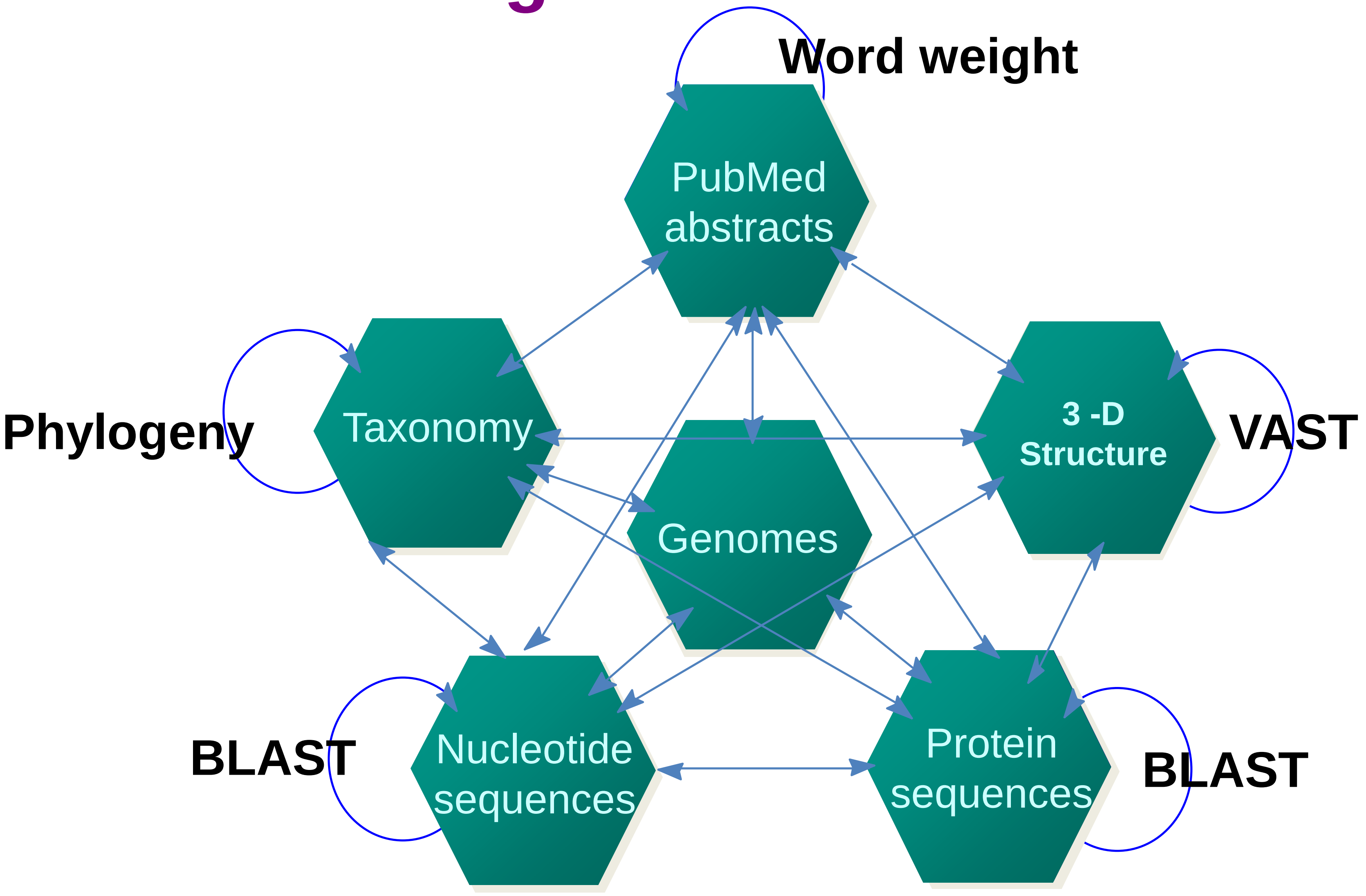
Source NCBI

Entrez Integrates Most of Them!



Gene
CancerChromosomes
UniGene
UniSTS
Homologene
SNP
Genome
Nucleotide
PopSet
Books
GEO
PubMe
Entrez
Taxonom
GEO Datasets
MeSH
OMIM
Protein
PMC
Journals
Domains
Structure
3D Domains

Entrez: Neighbors and Hard Links



Word weight

PubMed abstracts

Phylogeny

Taxonomy

Genomes

3 -D Structure

VAST

BLAST

Nucleotide sequences

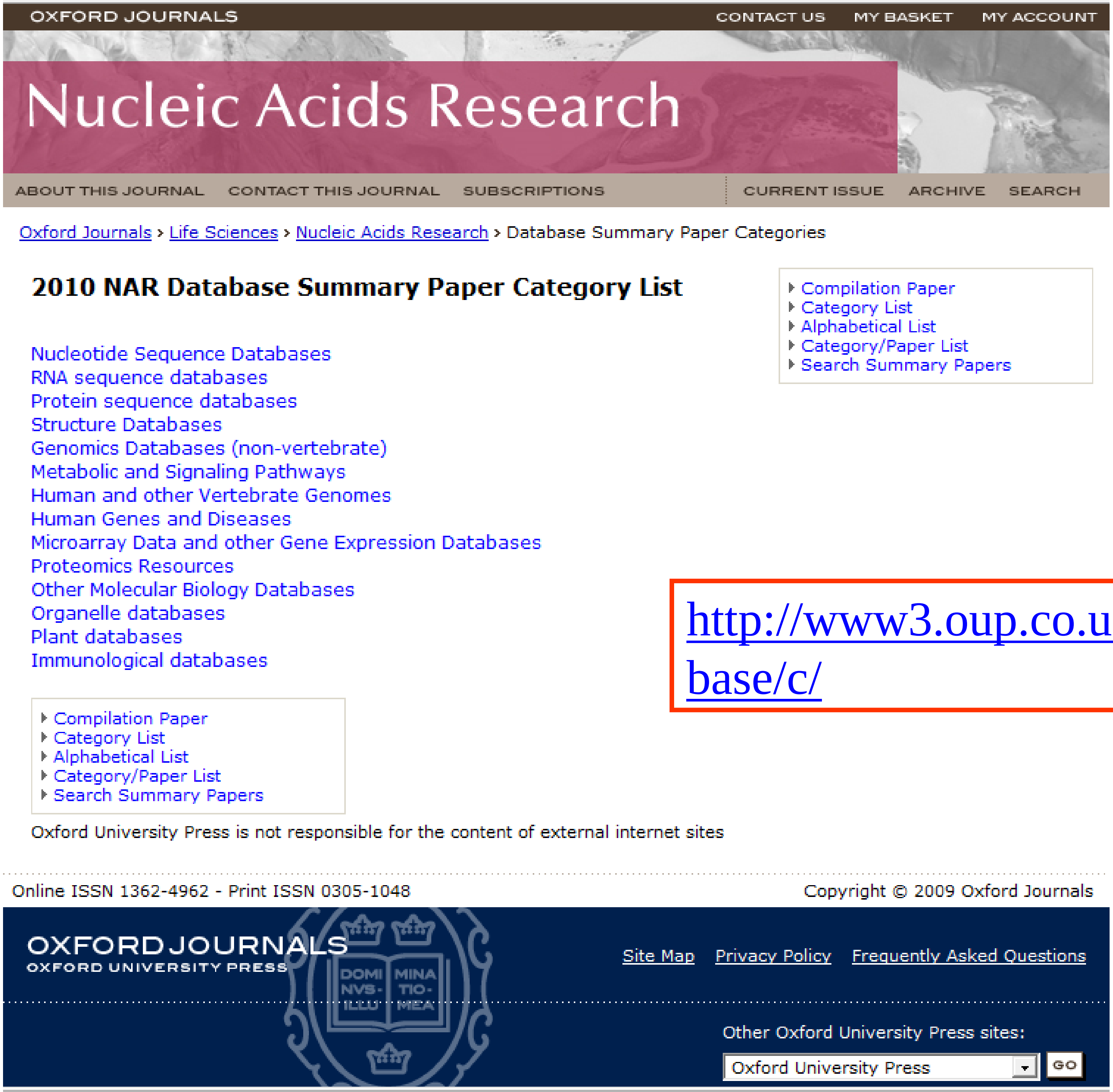
Protein sequences

BLAST

Source NCBI

NAR DB issue 2010

- 58 new dbs since last year!
- Total >1230!
- (http://www.oxfordjournals.org/nar/database/a/
- Complete list
  - Searchable
  - http://nar.oxfordjournals.org/cgi/content/full/gkm1037/DC1/1 (html format), also as downloadable word file)



Oxford Journals > Life Sciences > Nucleic Acids Research > Database Summary Paper Categories

2010 NAR Database Summary Paper Category List

- Compilation Paper
- Category List
- Alphabetical List
- Category/Paper List
- Search Summary Papers

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

- Compilation Paper
- Category List
- Alphabetical List
- Category/Paper List
- Search Summary Papers

Online ISSN 1362-4962 - Print ISSN 0305-1048

http://www3.oup.co.uk/nar/database/c/

Clinical Bioinformatics

Wang and Liotta *Journal of Clinical Bioinformatics* 2011, **1**:1
http://www.jclinbioinformatics.com/content/1/1/1

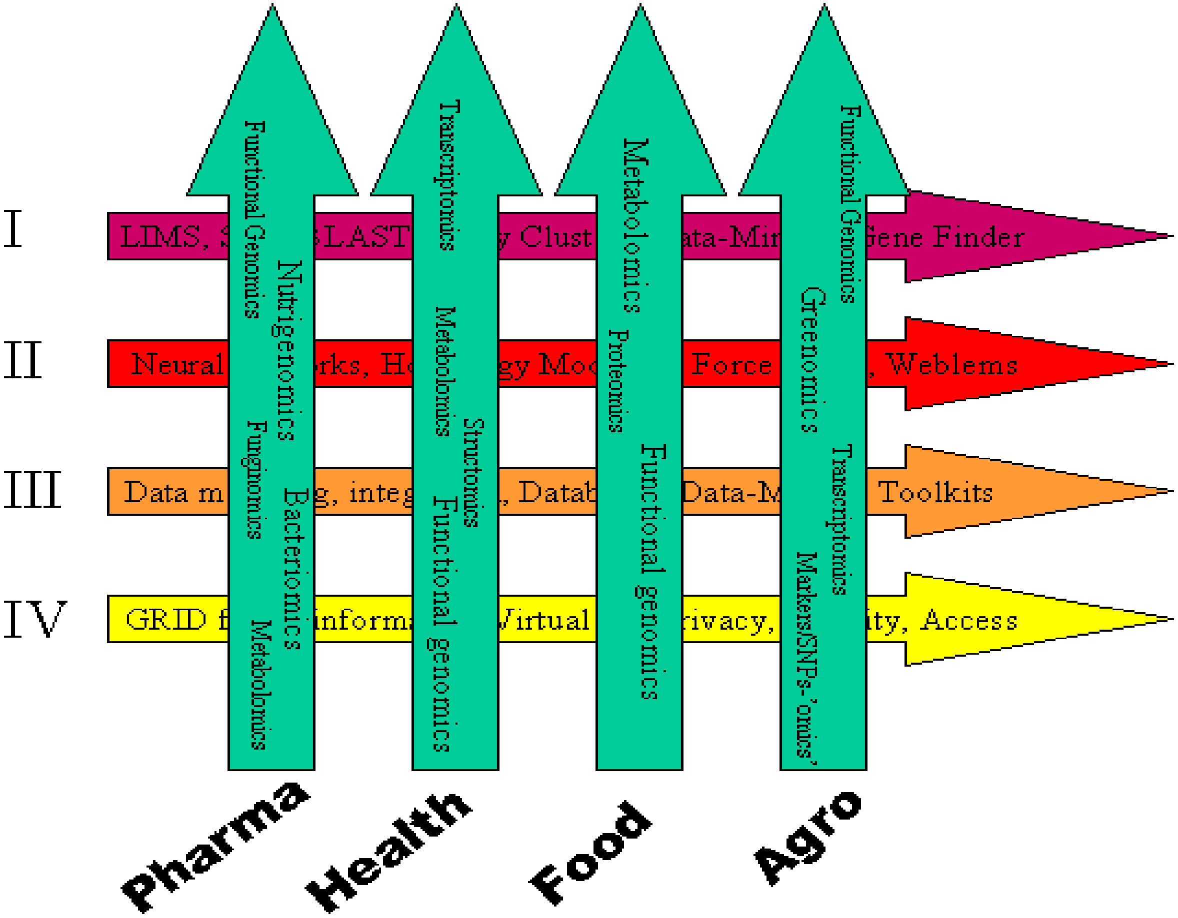
JCB*i* JOURNAL OF
CLINICAL BIOINFORMATICS

EDITORIAL **Open Access**

Clinical bioinformatics: a new emerging science

- Microarrays, omics data (genomics, proteomics, interactomics, metabolomics, …)
- Combination of bioinformatics and medical informatics

Omics impact on the life sciences



Instrumentation Being Used In Omics Experimentation

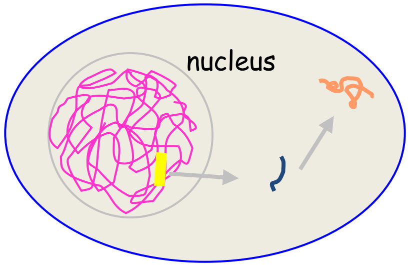
- Transcriptomics via among others; micro-arrays
- Proteomics via among others; Mass Spectroscopy (MS)
- Metabolomics via among others; MS & Nuclear Magnetic Resonance (NMR)

Paradigm shift in Life sciences

- *Past experiments where hypothesis driven*
  - Evaluate hypothesis
  - Complement existing knowledge

- *Present experiments are data driven*
  - Discover knowledge from large amounts of data

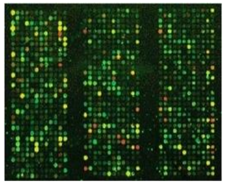
Life sciences research: from gene to function



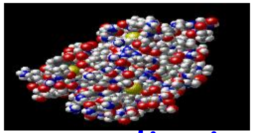
cell

Whole-genome sequence projects

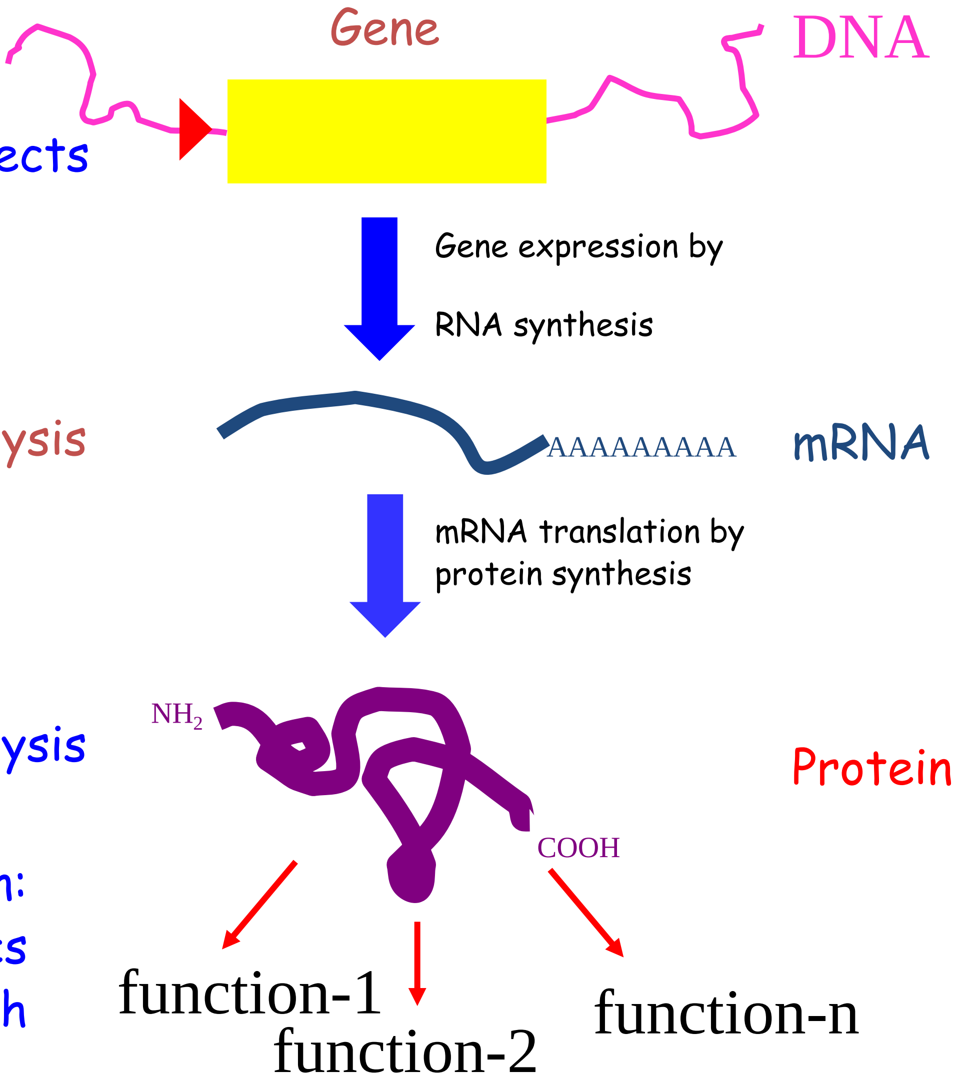
Genome-wide micro-array analysis



“High-throughput” protein-analysis



Protein function:
-prediction by bioinformatics
-proof by laboratory research



Gene

DNA

Gene expression by

RNA synthesis

AAAAAAAAA

mRNA

mRNA translation by
protein synthesis

NH_2

COOH

Protein

function-1
function-2
function-n

GenBank

Public Collections of DNA and RNA Sequence Reach over 145 Gigabases

These **145,000,000,000 bases** of the genetic code, represent both individual genes and partial and complete genomes **(>61,000,000 sequences)** of over **240,000 organisms**, e.g. humans, elephants, earthworms, fruitflies, apple trees, and bacteria. **120 of more than 370** microbial genomes submitted past year. About **3000 new organisms** are added each month. 16% of the sequences is from human origin.

GenBank is maintained by the National Center for Biotechnology Information (NCBI). **Submitters to GenBank currently contribute over 15 million new DNA sequences last year to the database**.

Size of GenBank doubles about every **18 months**

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

Top organisms in GenBank (Release 191)

Organism	base pairs
Homo sapiens	16,310,774,187
Mus musculus	9,974,977,889
Rattus norvegicus	6,521,253,272
Bos taurus	5,386,258,455
Zea mays	5,062,731,057
Sus scrofa	4,887,861,860
Danio rerio	3,120,857,462
Strongylocentrotus purpuratus	1,435,236,534
Macaca mulatta	1,256,203,101
Oryza sativa Japonica Group	1,255,686,573
Nicotiana tabacum	1,197,357,811
Xenopus (Silurana) tropicalis	1,249,938,611
Drosophila melanogaster	1,119,965,220

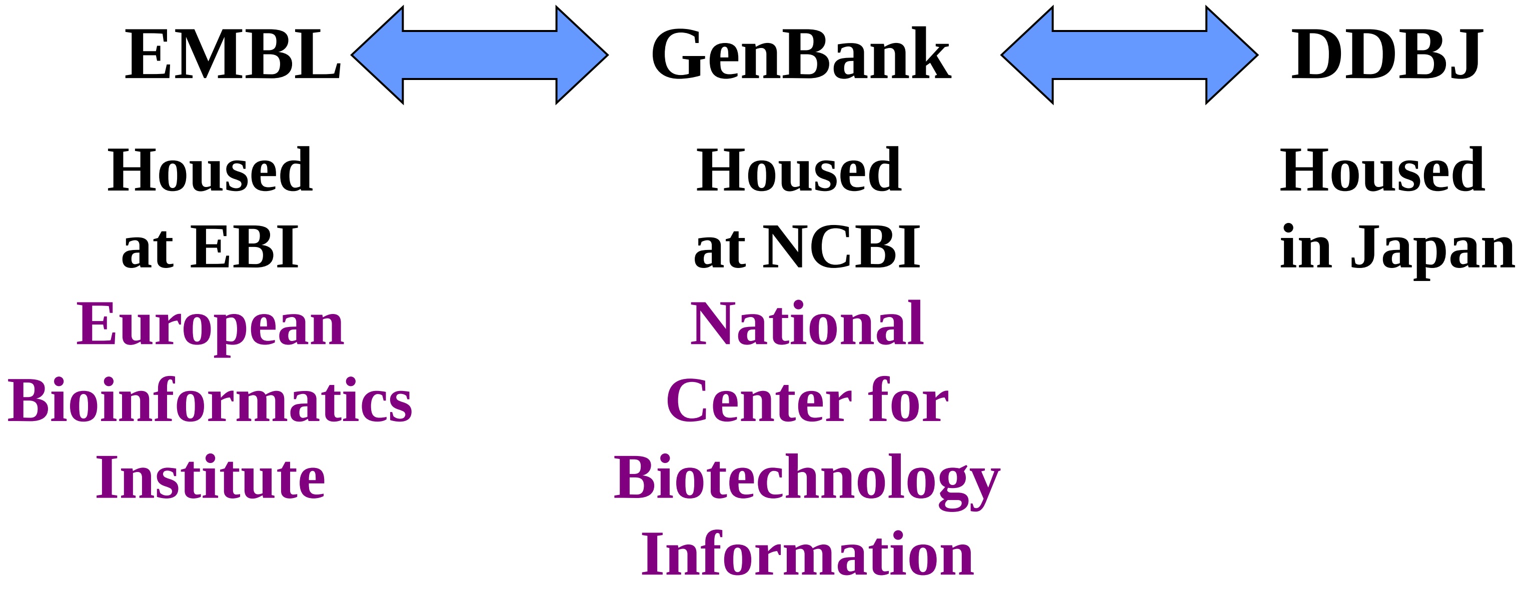
Genome sizes

- Amoeba – 670 Gb (billion base pairs)
- Humans – 3 Gb
- Corn – 2.5 Gb
- Honey Bee – 1.8 Gb
- Soybean – 1.4 Mb (million base pairs)
- Arabidopsis – 135 Mb
- Fruit fly – 130 MB (60% of fruit fly genes conserved with human)
- Yeast – 20 Mb
- E. coli bacterium – 4 Mb
- Viruses – 10^3 – 10^5 bps

Gene bank

- GenBank and its collaborators receive sequences produced in laboratories throughout the world from more than 100,000 distinct organisms.
- GenBank continues to grow at an exponential rate, doubling every 10-15 months.
- Release 134, produced in February 2003, contained over 29.3 billion nucleotide bases in more than 23.0 million sequences.
- GenBank is built by direct submissions from individual laboratories, as well as from bulk submissions from large-scale sequencing centers.

Three major public DNA databases



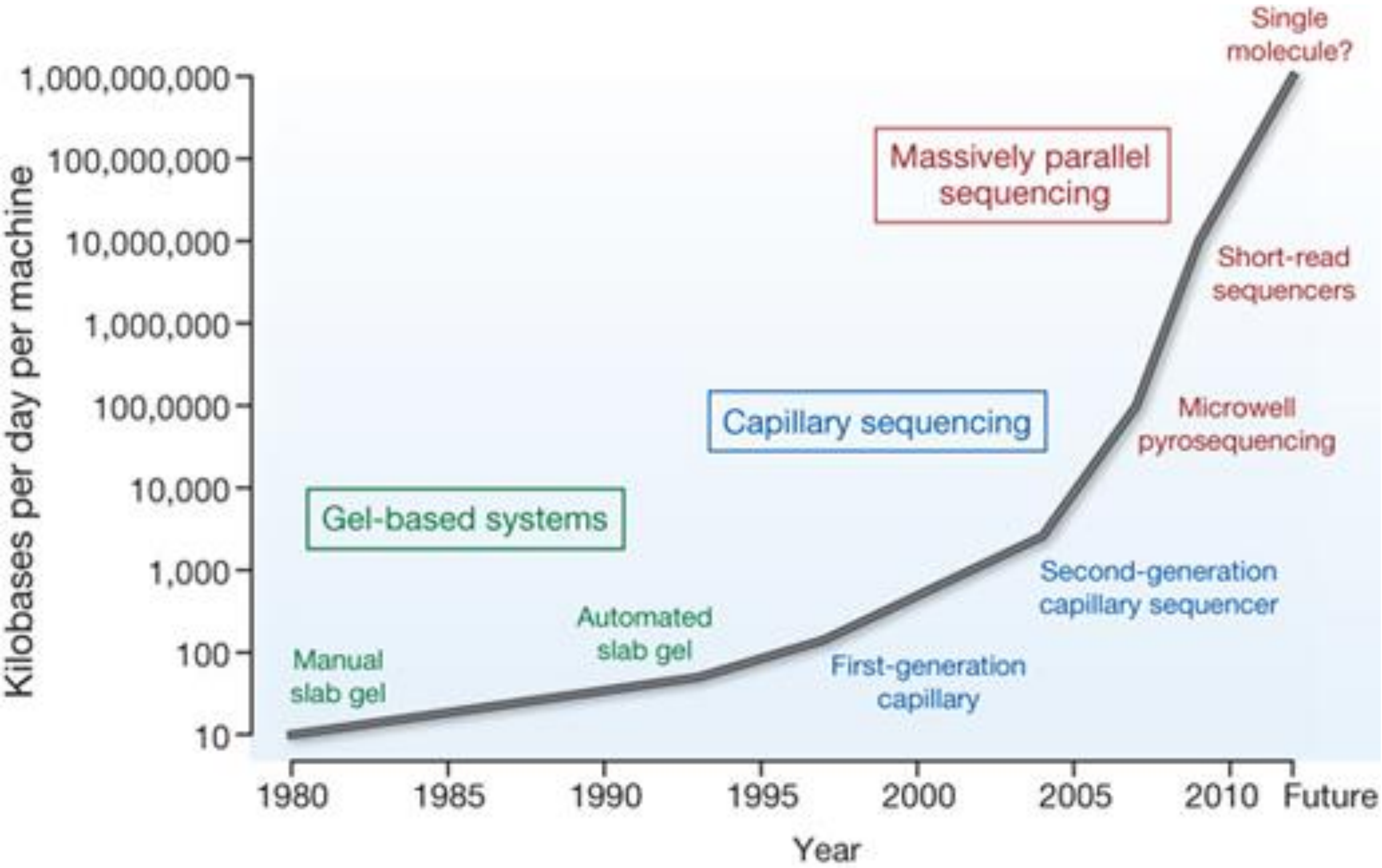
EMBL ⟷ **GenBank** ⟷ **DDBJ**

Housed
at EBI
European
Bioinformatics
Institute

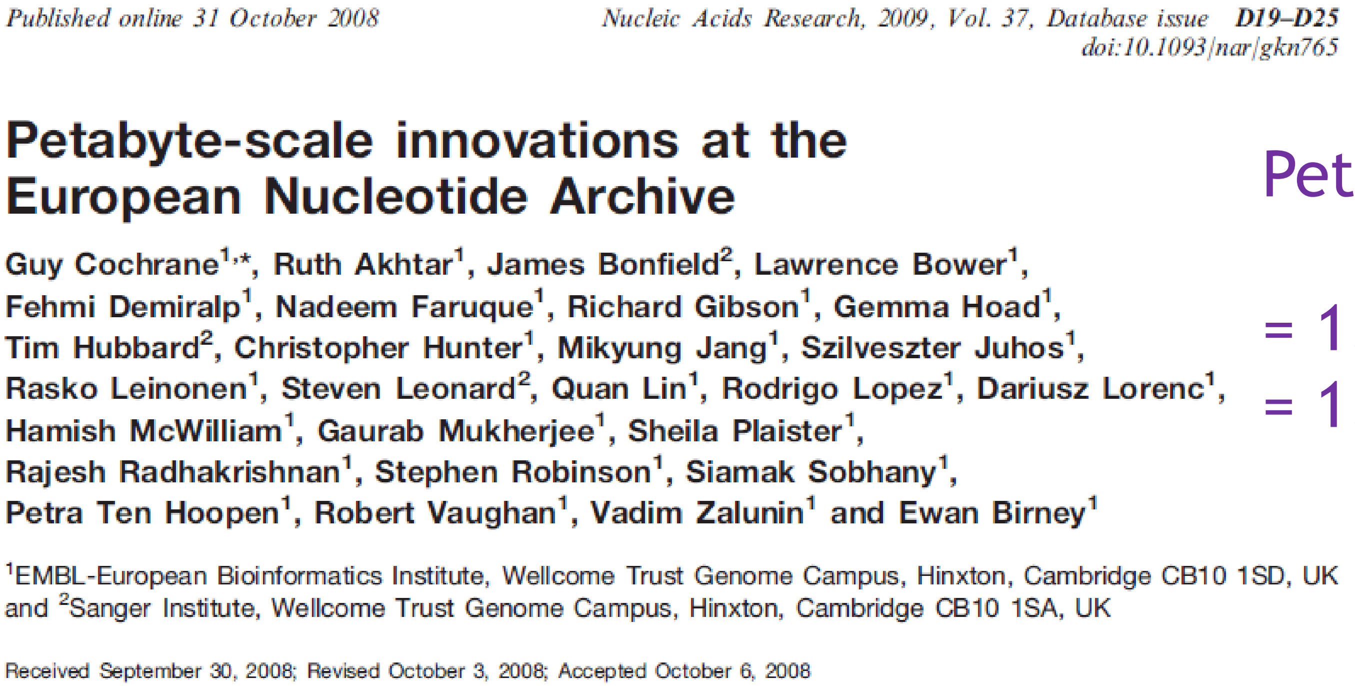
Housed
at NCBI
National
Center for
Biotechnology
Information

Housed
in Japan

DNA sequencing rate



http://www.wellcome.ac.uk/Education-resources/Teaching-and-education/Big-Picture/All-issues/Genes-Genomes-and-Health/WTDV027167.htm



Nucleic Acids Research, 2009, Vol. 37, Database issue **D19–D25**
doi:10.1093/nar/gkn765

Petabyte-scale innovations at the European Nucleotide Archive

Guy Cochrane[1,*], Ruth Akhtar[1], James Bonfield[2], Lawrence Bower[1], Fehmi Demiralp[1], Nadeem Faruque[1], Richard Gibson[1], Gemma Hoad[1], Tim Hubbard[2], Christopher Hunter[1], Mikyung Jang[1], Szilveszter Juhos[1], Rasko Leinonen[1], Steven Leonard[2], Quan Lin[1], Rodrigo Lopez[1], Dariusz Lorenc[1], Hamish McWilliam[1], Gaurab Mukherjee[1], Sheila Plaister[1], Rajesh Radhakrishnan[1], Stephen Robinson[1], Siamak Sobhany[1], Petra Ten Hoopen[1], Robert Vaughan[1], Vadim Zalunin[1] and Ewan Birney[1]

[1]EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [2]Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

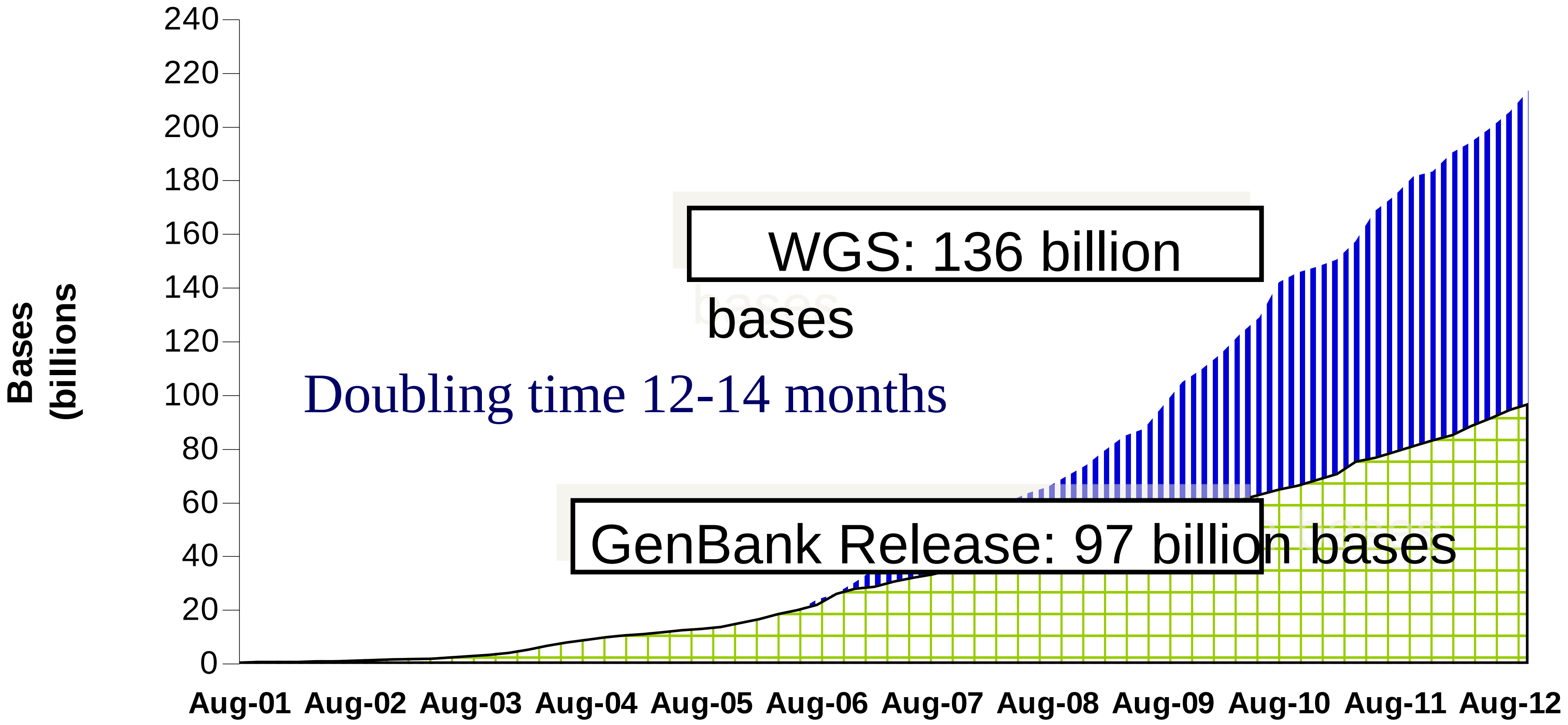
Received September 30, 2008; Revised October 3, 2008; Accepted October 6, 2008

Petabyte = 10^{15} byte

= 1.000 Terra byte
= 1.000.000 Giga byte

- European Nucleotide Archive (ENA)
  - Established for next generation sequencing data
  - Last 3 months the EBI received 10TB of data representing 1/8 the total volume of data accumulated in 28 years
- Challenges
  - Bandwidth to overcome network constraints
  - User tools to submit the data
  - Validation systems to overcome unmanageable demands on manual curation efforts
  - Tools to maximize data utility to users

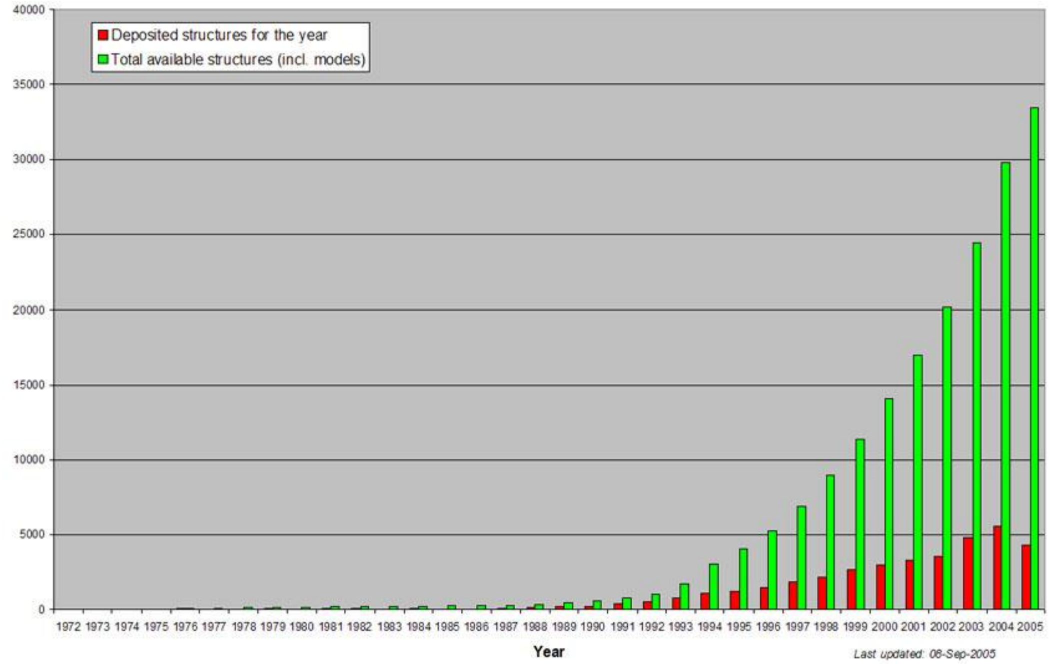
The Growth of GenBank



WGS: 136 billion bases

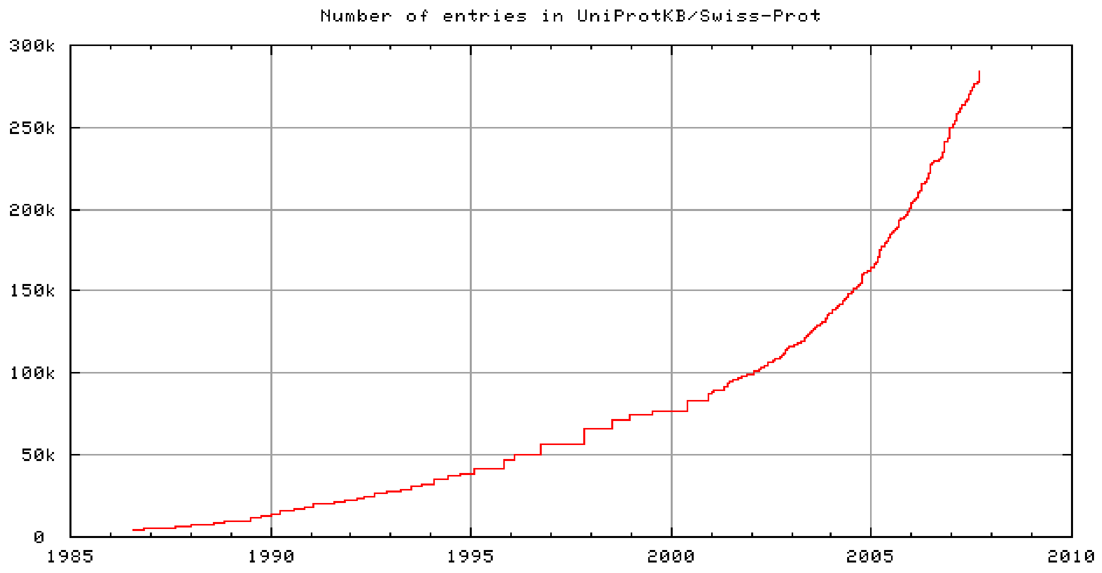
Doubling time 12-14 months

GenBank Release: 97 billion bases



3D Structures Growth

Source
(http://www.rcsb.org/pdb/holdings.html)



Number of Entries in UniProtKB/TrEMBL Protein Database

Source
http://www.ebi.ac.uk/swissprot/sptr_stats/index.html

Informatics Based Infrastructue

(1) Installation of informatics based infrastructure for facilitating research work.

(2) Procurement and development of Hardware, Software and databases, computational tools.

(3) Telecommunications facilities .

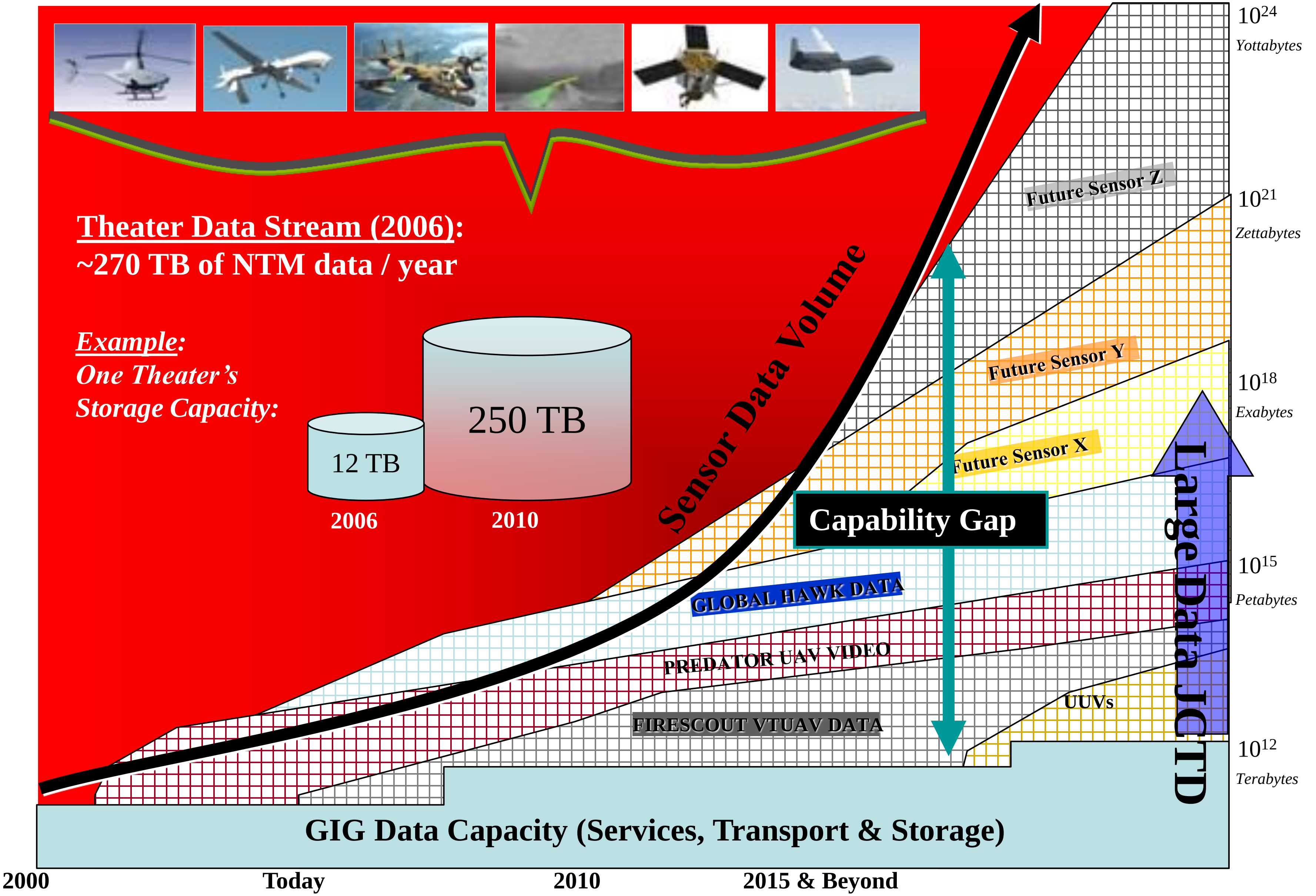
(4) databases

Growth of Data

2008: **2 Tb** 2009: **32 Tb** 2010: **150 Tb** 2011: **165 Tb**

How will we keep up with this?

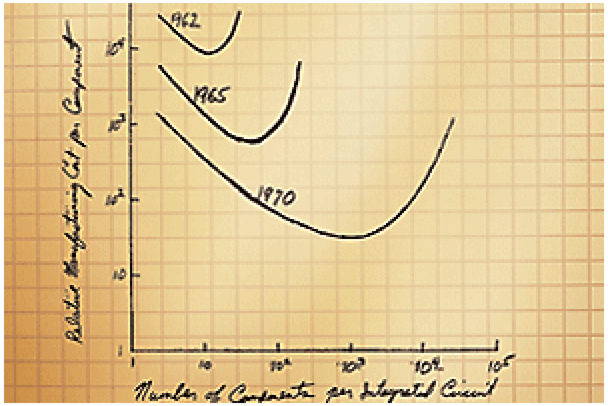
- maintaining/annotating quality
- storage
- communication (network bandwidth)
- analysis (including software and databases)
- integration



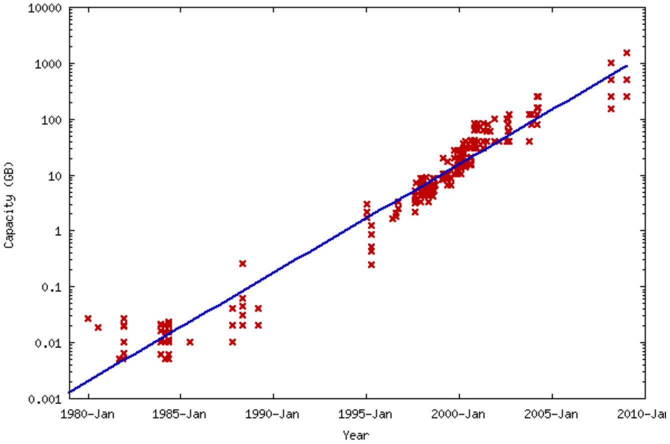
<u>**Theater Data Stream (2006)**</u>**:**
~270 TB of NTM data / year

<u>Example</u>**:**
One Theater's
Storage Capacity:

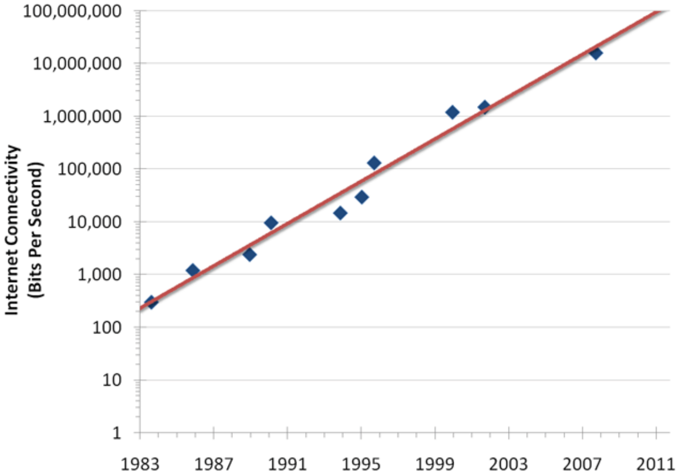
Bob Gourley
http://ctovision.typepad.com/InfoSharingTechnologyFutures.ppt



Moore's law: CPU power doubles in ~18-24 mo.



Hard drive capacity doubles in ~12 mo.

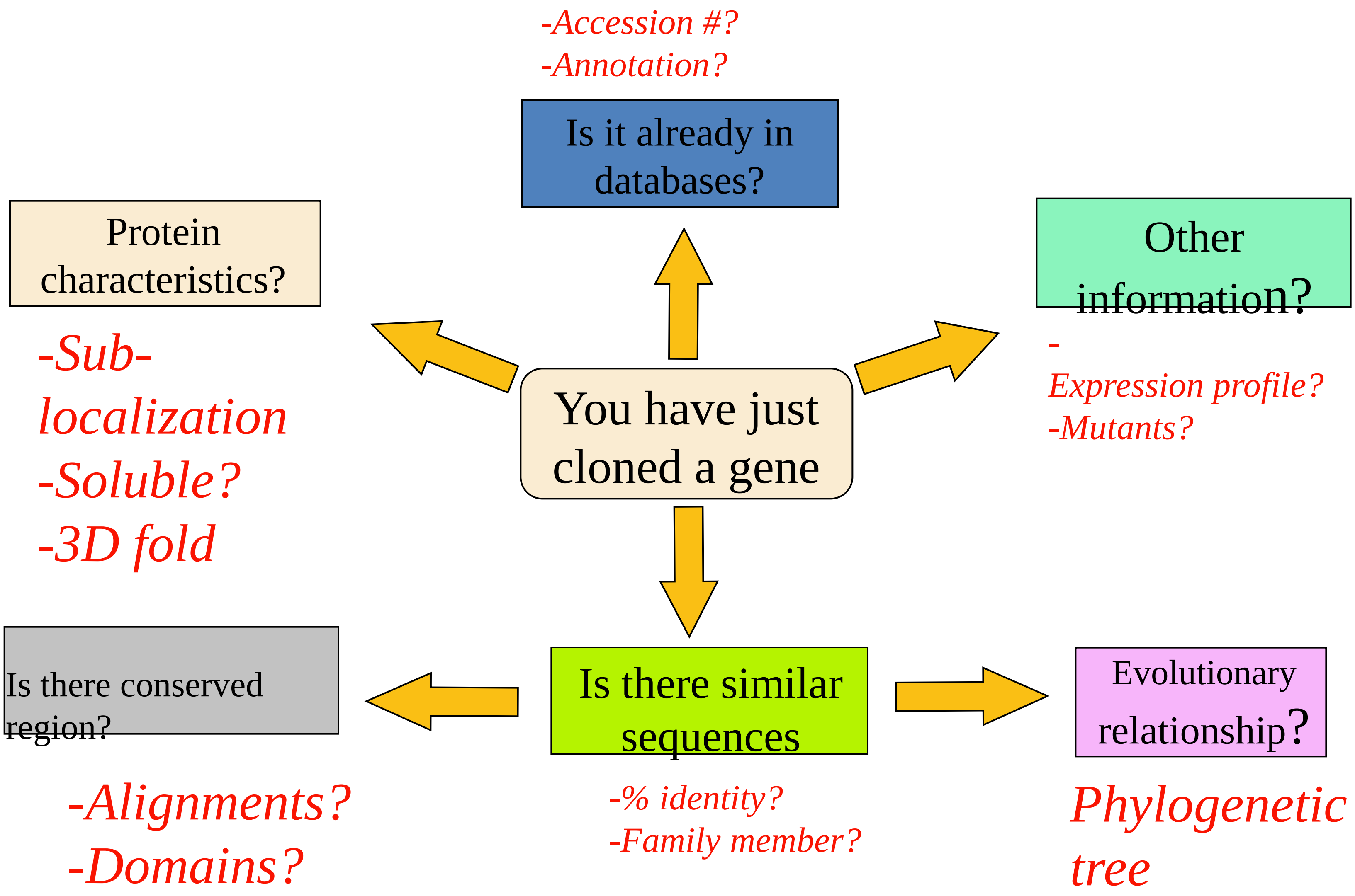


Network bandwidth doubles in ~20 mo.

Challenges for Bioinformatics

- *Analyzing and synthesizing complex experimental data*
- *Representing and accessing vast quantities of information*
- *Pattern matching*
- *Data mining*
- *Gene discovery*
- *Function discovery*
- *Modeling the dynamics of cell function*

Applying algorithms to analyze genomics data



-Accession #?
-Annotation?

Is it already in databases?

Protein characteristics?

-Sub-localization
-Soluble?
-3D fold

You have just cloned a gene

Other information?

-Expression profile?
-Mutants?

Is there conserved region?

-Alignments?
-Domains?

Is there similar sequences

-% identity?
-Family member?

Evolutionary relationship?

Phylogenetic tree

- A critical failure of current bioinformatics is the lack of a <u>single software</u> package that can perform all of these functions.

Development of Algorithms for Sequence Comparison:

- *Phylogenetic Algorithm*
  - *Complex mathematical formula used to determine sequence homology*
  - *All possible ways a large number of sequences can be compared to one another*
- *Fitch and Margoliash*
  - *Sequence comparisons to determine evolutionary trees*
  - *Computer calculates the minimum number of steps to convert one sequence to another and builds possible trees*
- *Needleman and Wunsch*
  - *Similarities in protein sequences*

Major breakthroughs in Bioinformatics

Through innovations in Mathematics or Statistics

- ***FASTA,***
- ***BLAST,***
- ***Phred/Phrap,***
- ***BLOSUM,***
- ***GenScan,***
- ***PSI-BLAST,***
- ***Threading,***
- ***GRAIL etc.***

Statistical, Computational and Data Mining Methods/Techniques with Computational Intelligence

- *Decision Trees*
- *Probability for Data Miners*
- *Gaussians Theory*
- *Probability Density Functions*
- *Maximum Likelihood Estimation*
- *Gaussian Bayes Classifiers*
- *Cross-Validation*
- *Neural Networks*

Statistical, Computational and Data Mining----

- *Regression for Predicting Real-valued Outputs*
- *Game Tree Search Algorithms, including Alpha-Beta Search*
- *Hierarchical Clustering*
- *K-means Clustering*
- *Gaussian Mixture Models*
- *Hidden Markov Models*

Statistical, Computational and Data Mining----

- *Artificial Intelligence*
- *Correspondence Analysis*
- *Principal Component Analysis*
- *Search Algorithms*
- *Star Heuristic Search*
- *Genetic Algorithms*

Statistical software for analysing molecular biology data

- *R (downloadable free from net)*
- *Bio-conductor*
- *SAS*
- *SPSS*
- *Mebiogeme*
- *Dataplot*
- *JMP genomics*
- *Splus and*
- *MATLAB Bioinformatics Toolbox*

Programming Tools in Bioinformatics

(just a few!)

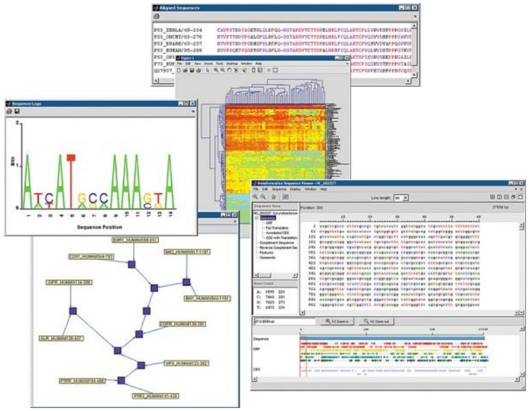
- **JAVA in Bioinformatics:**
  - **BioJava**: (www.biojava.org) Java tools for processing biological data which includes objects for manipulating sequences, dynamic programming, file parsers, simple statistical routines, etc.
- **Perl in Bioinformatics:**
  - **BioPerl**: (www.bioperl.org) The BioPerl project is an international association of developers of Perl tools for bioinformatics and provides an online resource for modules, scripts and web links for developers of Perl-based software.

- **Python** is an interpreted, interactive programming language created by Rossum in 1990..
  - **Biopython (http://biopython.org/) and biojava (www.biojava.org):** Biopython and biojava are open source projects with very similar goals to bioperl

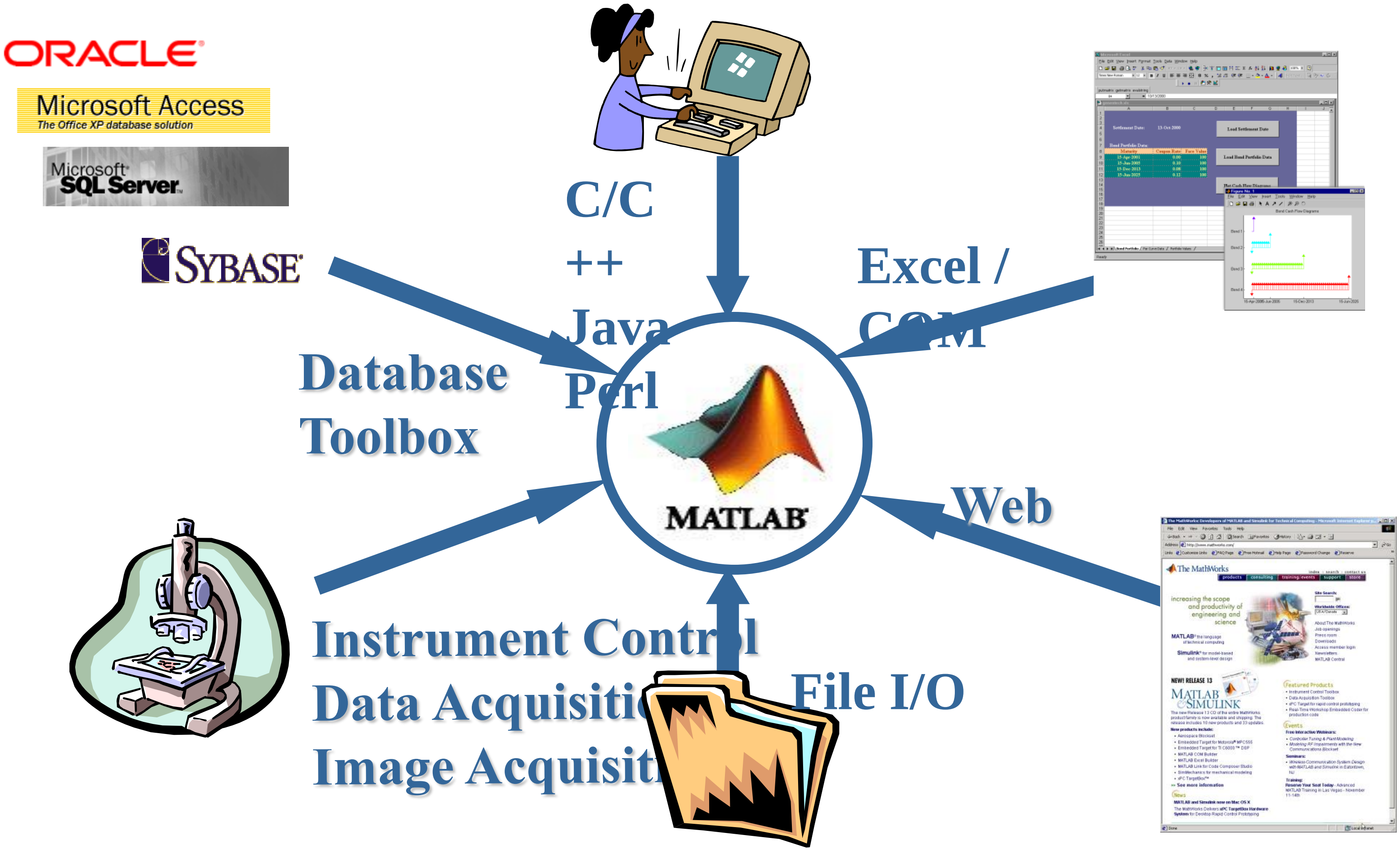
- **MATLAB Bioinformatics Toolbox:**

MATLAB Bioinformatics Toolbox
http://www.mathworks.com/products/bioinfo/

MATLAB



Connecting to MATLAB



C/C
++
Java
Perl

Excel /
COM

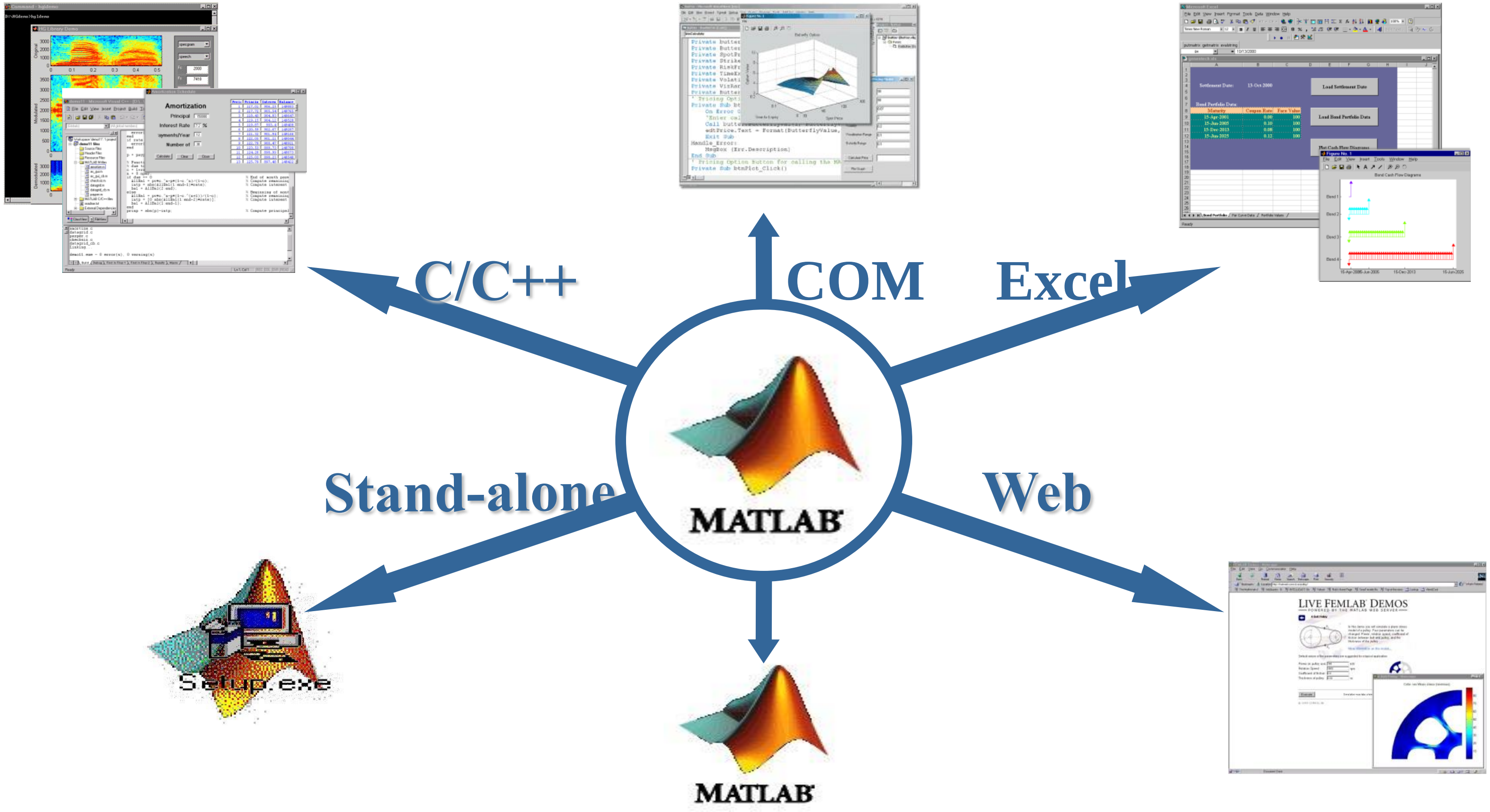
Database
Toolbox

Web

Instrument Contr l
Data Acquisiti
Image Acquisit

File I/O

Deploying with MATLAB



C/C++

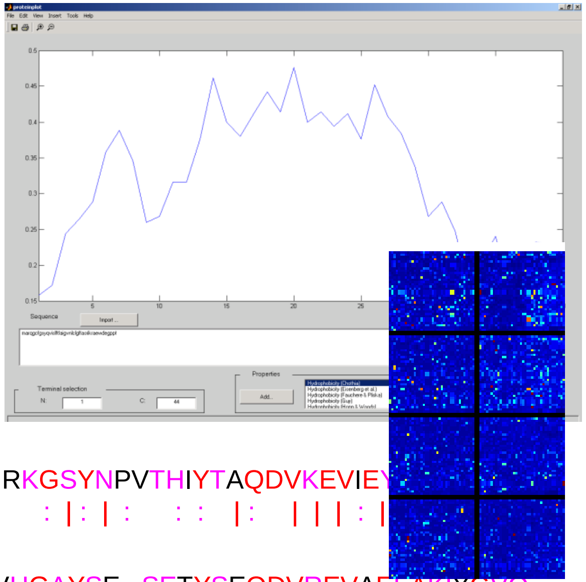
COM

Excel

Stand-alone

Web

Bioinformatics Toolbox



```
212  PYESFTFPELMRKGSYNPVTHIYTAQDVKEVIEY
     | |  |    : | | |    : | : | :   : :  | :   | | | : |
:   | : :
321  PYISRYYPELAVHGAYSE -SETYSEQDVREVAEFAKIYGVQ
```

- File I/O
  - FASTA, PDB, SCF, GPR, GAL
- Web Connectivity
  - GenBank, EMBL, PIR, PDB
- Sequence Analysis & Alignment
  - Needleman-Wunsch, Smith-Waterman
  - DNA/RNA/AA conversions, pattern searching
- Microarray Normalization & Visualization
  - Lowess, global mean, MAD (median absolute deviation)
- Protein Visualization
  - Atomic composition, molecular weight, hydrophobicity profile

- ***R-language*** *for Statistical Computing : R is a free software environment for statistical computing and graphics.*
- ***Bioconductor*: Bioinformatics with R**
  - *The broad goals of Bioconductor are to*
    - *provide access to a wide range of powerful statistical and graphical methods for the analysis of genomic data;*
    - *provide training in computational and statistical methods for the analysis of genomic data.*
- ***Microarray Software Comparison*** *(http://ihome.cuhk.edu.hk/~b400559/arraysoft_rpackages.html)*
  - *A website organizing and commenting on links to R software for gene expression data analysis.*

DNA (nucleotide sequences) databases

- They are big databases and searching either one should produce similar results because they exchange information routinely.

-GenBank (NCBI): *http://www.ncbi.nlm.nih.gov*

-DDBJ (DNA DataBase of Japan): *http://www.ddbj.nig.ac.jp*

-TIGR: *http://tigr.org/tdb/tgi*

-Yeast: *http://yeastgenome.org*

-*E. coli*: *http://colibase.bham.ac.uk/blast/*

- Specialized databases:Tissues, species…

-ESTs (Expressed Sequence Tags)

~at NCBI *http://www.ncbi.nlm.nih.gov/dbEST*

~at TIGR *http://tigr.org/tdb/tgi*

- many more!

Protein Data Bank

DEPOSIT data
DOWNLOAD files
browse LINKS
BETA TEST new features
PDBML/XML files

Current Holdings

32727 Structures
Last Update: 20-Sep-2005
PDB Statistics

Molecule of the Month:
Cholera Toxin

The Protein Data Bank (PDB) is operated by Rutgers, The State University of New Jersey and the San Diego Supercomputer Center at the University of California, San Diego -- two members of the Research Collaboratory for Structural Bioinformatics (RCSB).

Did you find what you wanted?

Welcome to the PDB, the single worldwide repository for the processing and distribution of 3-D biological macromolecular structure data.

ABOUT PDB | NEW FEATURES | USER GUIDES | FILE FORMATS | DATA UNIFORMITY | STRUCTURAL GENOMICS | SOFTWARE | PUBLICATIONS | EDUCATION

Search the Archive

Enter a PDB ID or keyword

● PDB ID ○ Authors ○ Full Text Search
☑ match exact word ☐ remove similar sequences

QuickSearch! search Web pages and structures
SearchLite keyword search form with examples
SearchFields customizable search form
Status Search find entries awaiting release

News

Complete News
Newsletter
pdb-l Archive
Subscribe

20-Sep-2005
New Structural Genomics Portal Available

The RCSB PDB offers online tools, summary reports, and target information related to structural genomics from a new information portal at sg.pdb.org . [MORE...]

PDB Mirrors

Please bookmark a mirror site

San Diego Supercomputer Center, UCSD*
Rutgers University*
Cambridge Crystallographic Data Centre, UK
National University of Singapore
Osaka University, Japan
Max Delbrück Center for Molecular Medicine, Germany

OCA / PDB Lite MORE...

**RCSB partner*

In citing the PDB please refer to:

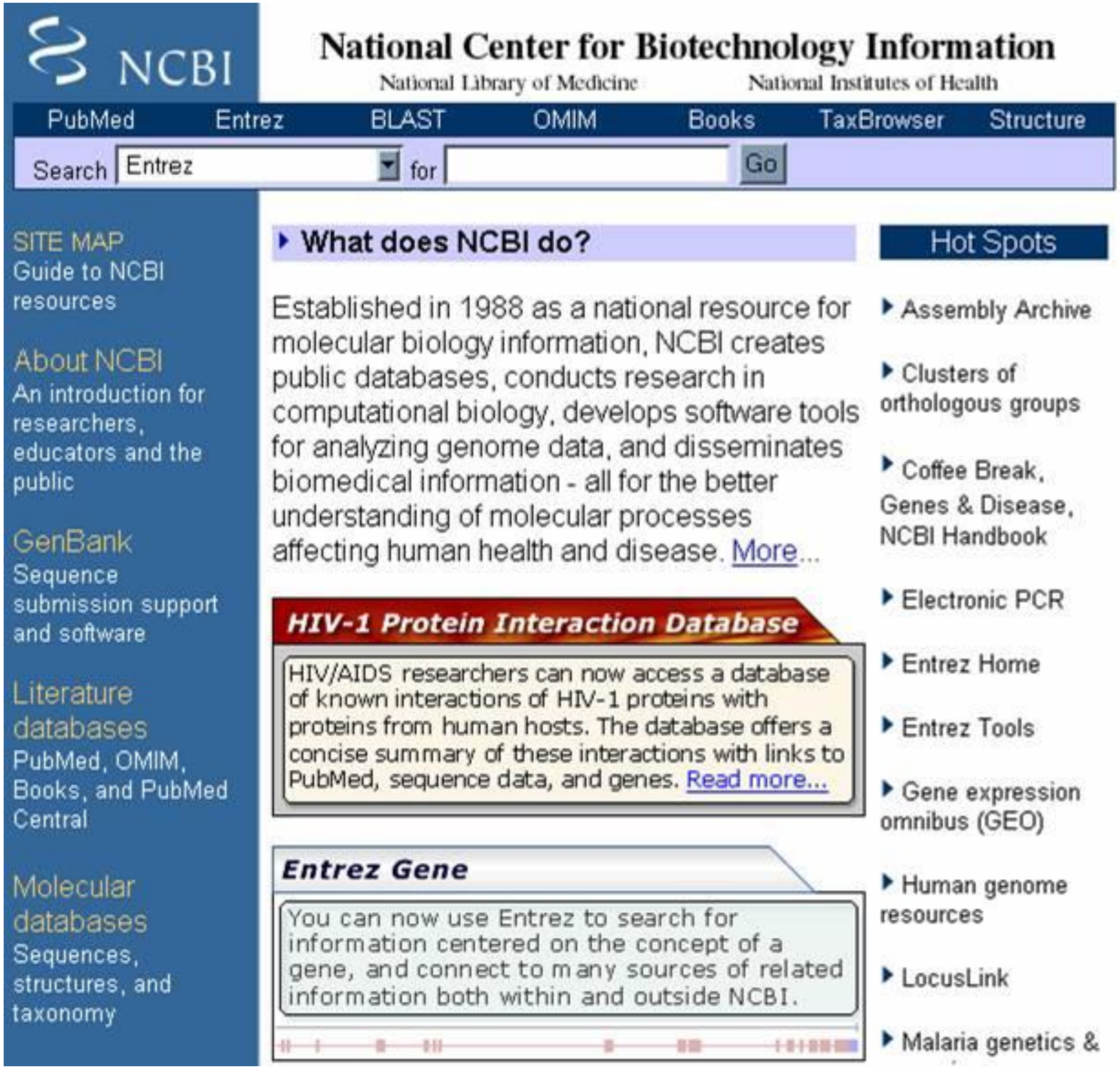
H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne: The Protein Data Bank. *Nucleic Acids Research*, **28** pp. 235-242 (2000)

ABOUT PDB | NEW FEATURES | USER GUIDES | FILE FORMATS | DATA UNIFORMITY | STRUCTURAL GENOMICS | SOFTWARE | PUBLICATIONS | EDUCATION

The RCSB PDB is supported by funds from the National Science Foundation (NSF), the National Institute of General Medical Sciences (NIGMS), the Office of Science, Department of Energy (DOE), the National Library of Medicine (NLM), the National Cancer Institute (NCI), the National Center for Research Resources (NCRR), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), and the National Institute of Neurological Disorders and Stroke (NINDS).

© RCSB

Web Access: http://www.ncbi.nlm.nih.gov



SITE MAP
Guide to NCBI resources

About NCBI
An introduction for researchers, educators and the public

GenBank
Sequence submission support and software

Literature databases
PubMed, OMIM, Books, and PubMed Central

Molecular databases
Sequences, structures, and taxonomy

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...

HIV-1 Protein Interaction Database

HIV/AIDS researchers can now access a database of known interactions of HIV-1 proteins with proteins from human hosts. The database offers a concise summary of these interactions with links to PubMed, sequence data, and genes. Read more...

Entrez Gene

You can now use Entrez to search for information centered on the concept of a gene, and connect to many sources of related information both within and outside NCBI.

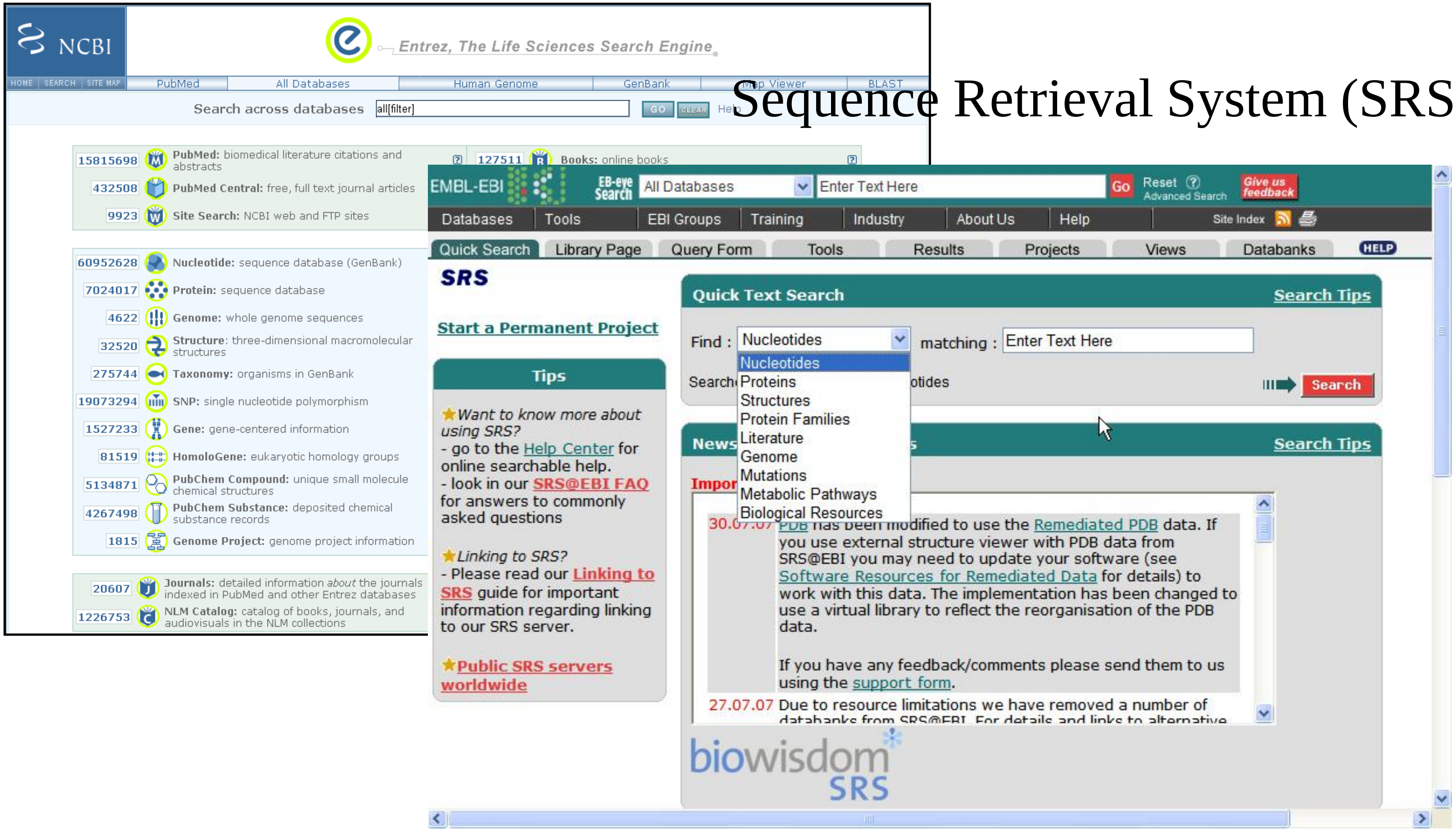
Hot Spots

- Assembly Archive
- Clusters of orthologous groups
- Coffee Break, Genes & Disease, NCBI Handbook
- Electronic PCR
- Entrez Home
- Entrez Tools
- Gene expression omnibus (GEO)
- Human genome resources
- LocusLink
- Malaria genetics &

Text Searches

Entrez System

Sequence Retrieval System (SRS)



all[filter]

15815698	**PubMed:** biomedical literature citations and abstracts
432508	**PubMed Central:** free, full text journal articles
9923	**Site Search:** NCBI web and FTP sites
127511	**Books:** online books
60952628	**Nucleotide:** sequence database (GenBank)
7024017	**Protein:** sequence database
4622	**Genome:** whole genome sequences
32520	**Structure:** three-dimensional macromolecular structures
275744	**Taxonomy:** organisms in GenBank
19073294	**SNP:** single nucleotide polymorphism
1527233	**Gene:** gene-centered information
81519	**HomoloGene:** eukaryotic homology groups
5134871	**PubChem Compound:** unique small molecule chemical structures
4267498	**PubChem Substance:** deposited chemical substance records
1815	**Genome Project:** genome project information
20607	**Journals:** detailed information *about* the journals indexed in PubMed and other Entrez databases
1226753	**NLM Catalog:** catalog of books, journals, and audiovisuals in the NLM collections

Find : Nucleotides matching : Enter Text Here

Nucleotides
Proteins
Structures
Protein Families
Literature
Genome
Mutations
Metabolic Pathways
Biological Resources

Search[...]otides

★*Want to know more about using SRS?*
- go to the Help Center for online searchable help.
- look in our **SRS@EBI FAQ** for answers to commonly asked questions

★*Linking to SRS?*
- Please read our **Linking to SRS** guide for important information regarding linking to our SRS server.

★**Public SRS servers worldwide**

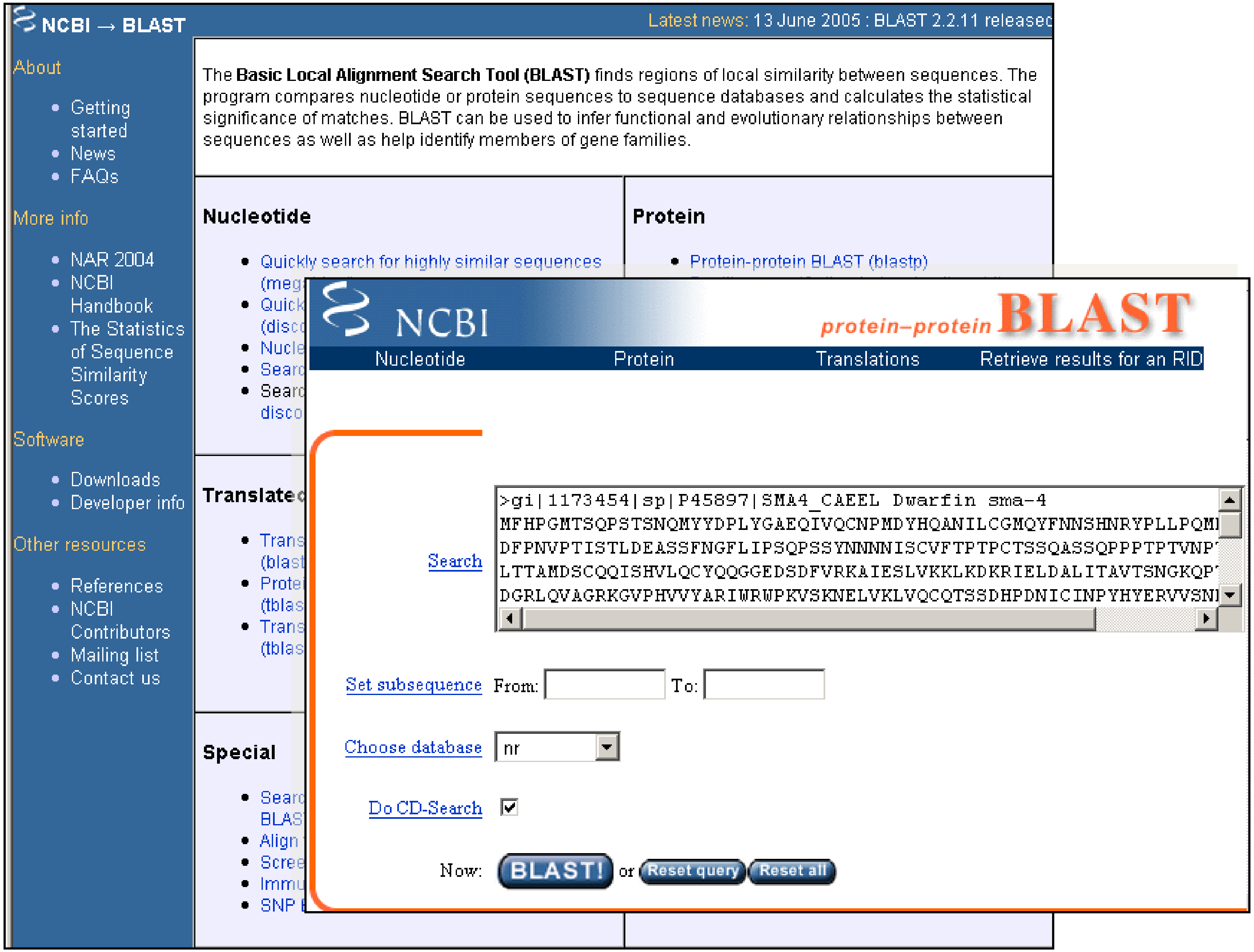
Impor

30.07.07 PDB has been modified to use the Remediated PDB data. If you use external structure viewer with PDB data from SRS@EBI you may need to update your software (see Software Resources for Remediated Data for details) to work with this data. The implementation has been changed to use a virtual library to reflect the reorganisation of the PDB data.

If you have any feedback/comments please send them to us using the support form.

27.07.07 Due to resource limitations we have removed a number of databanks from SRS@EBI. For details and links to alternative

BLAST: Sequence Similarity Searches



The **Basic Local Alignment Search Tool (BLAST)** finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Search

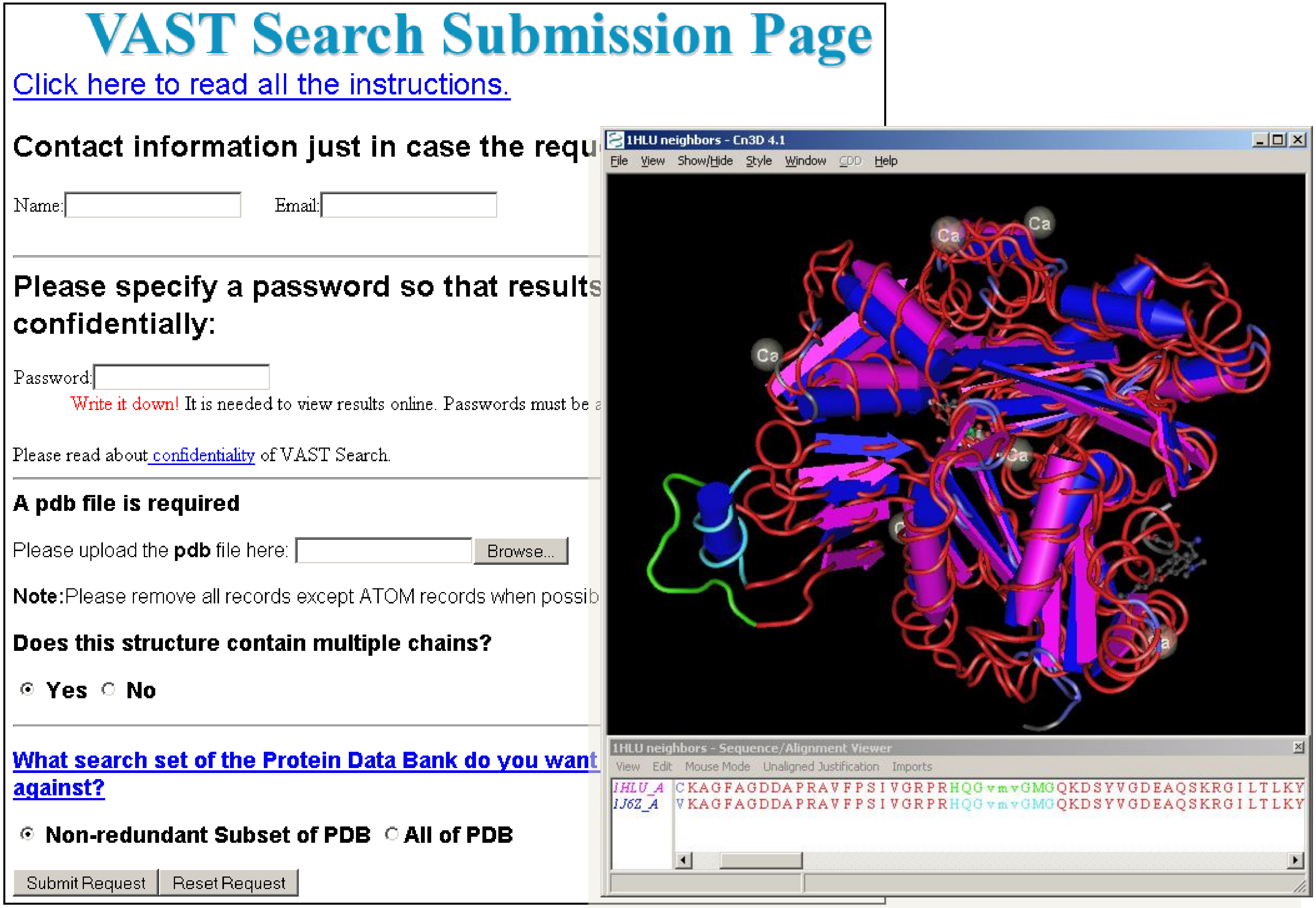
```
>gi|1173454|sp|P45897|SMA4_CAEEL Dwarfin sma-4
MFHPGMTSQPSTSNQMYYDPLYGAEQIVQCNPMDYHQANILCGMQYFNNSHNRYPLLPQM
DFPNVPTISTLDEASSFNGFLIPSQPSSYNNNNISCVFTPTPCTSSQASSQPPPTPTVNP
LTTAMDSCQQISHVLQCYQQGGEDSDFVRKAIESLVKKLKDKRIELDALITAVTSNGKQP
DGRLQVAGRKGVPHVVYARIWRWPKVSKNELVKLVQCQTSSDHPDNICINPYHYERVVSN
```

Set subsequence From: To:

Choose database: nr

Do CD-Search ☑

VAST: Structure Similarity Searches



VAST Search Submission Page

Click here to read all the instructions.

Contact information just in case the requ

Name: 	Email:

Please specify a password so that results confidentially:

Password:

Write it down! It is needed to view results online. Passwords must be a

Please read about confidentiality of VAST Search.

A pdb file is required

Please upload the **pdb** file here: 	Browse...

Note: Please remove all records except ATOM records when possib

Does this structure contain multiple chains?

● Yes ○ No

What search set of the Protein Data Bank do you want against?

● Non-redundant Subset of PDB ○ All of PDB

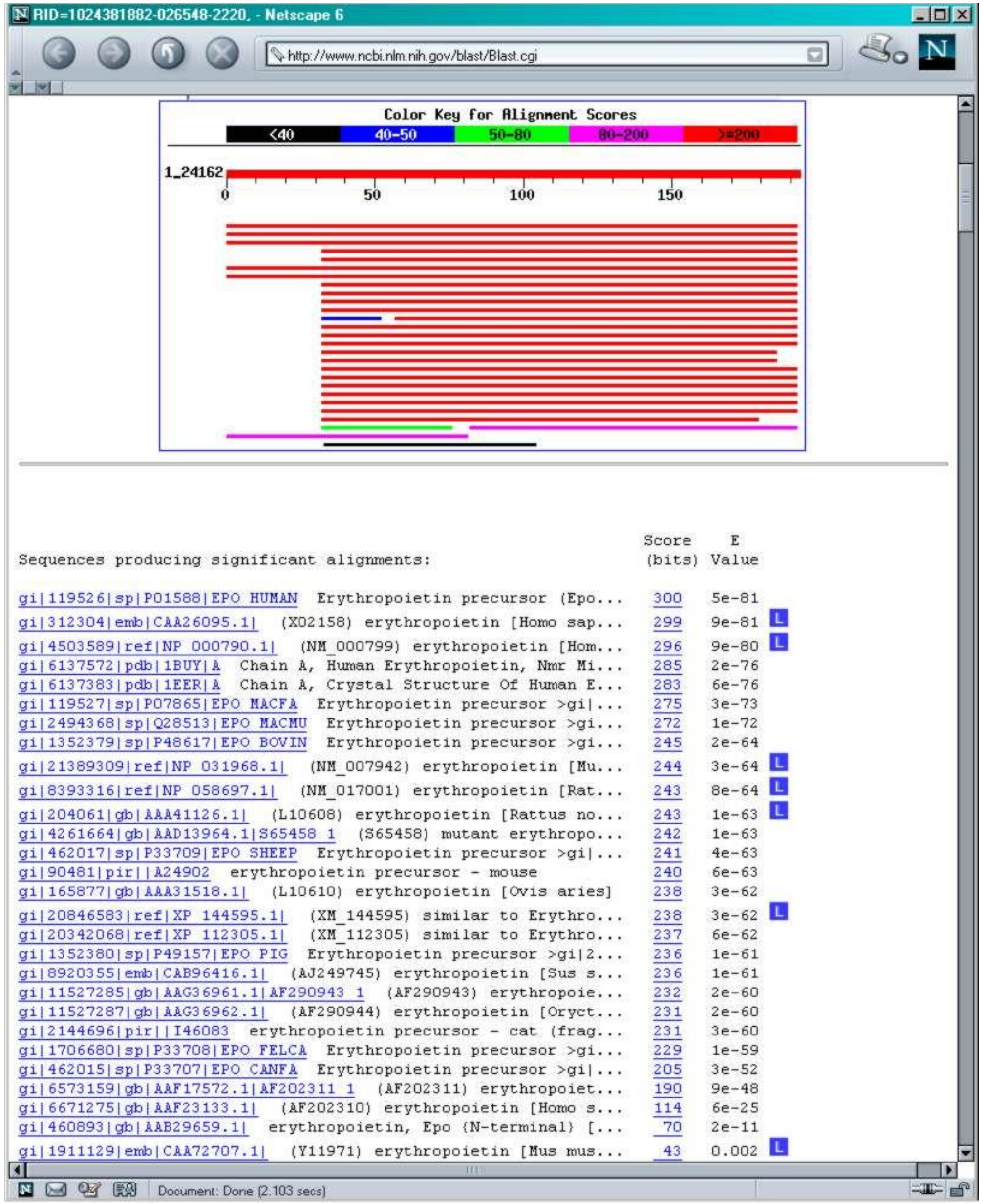
Submit Request 	Reset Request

```
1HLU_A  CKAGFAGDDAPRAVFPSIVGRPRHQGvmvGMGQKDSYVGDEAQSKRGILTLKY
1J6Z_A  VKAGFAGDDAPRAVFPSIVGRPRHQGvmvGMGQKDSYVGDEAQSKRGILTLKY
```

BLAST:

Basic
Local
Alignment
Search
Tool

NCBI BLAST
http://www.ncbi.nlm.nih.gov/blast/

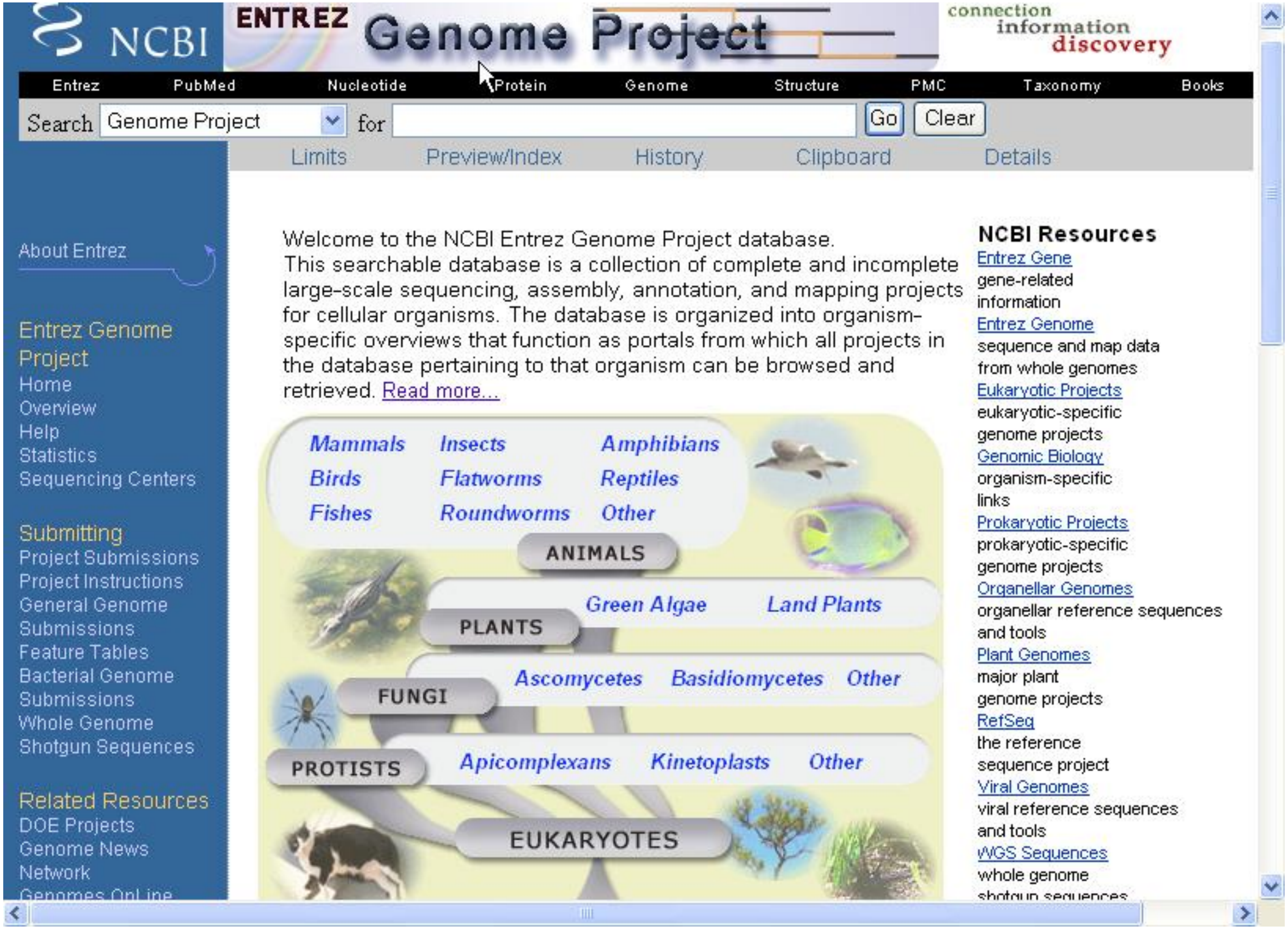


```
                                                               Score    E
Sequences producing significant alignments:                   (bits) Value

gi|119526|sp|P01588|EPO_HUMAN  Erythropoietin precursor (Epo...   300   5e-81
gi|312304|emb|CAA26095.1|  (X02158) erythropoietin [Homo sap...   299   9e-81  L
gi|4503589|ref|NP_000790.1|  (NM_000799) erythropoietin [Hom...   296   9e-80  L
gi|6137572|pdb|1BUY|A  Chain A, Human Erythropoietin, Nmr Mi...   285   2e-76
gi|6137383|pdb|1EER|A  Chain A, Crystal Structure Of Human E...   283   6e-76
gi|119527|sp|P07865|EPO_MACFA  Erythropoietin precursor >gi|...   275   3e-73
gi|2494368|sp|Q28513|EPO_MACMU  Erythropoietin precursor >gi...   272   1e-72
gi|1352379|sp|P48617|EPO_BOVIN  Erythropoietin precursor >gi...   245   2e-64
gi|21389309|ref|NP_031968.1|  (NM_007942) erythropoietin [Mu...   244   3e-64  L
gi|8393316|ref|NP_058697.1|  (NM_017001) erythropoietin [Rat...   243   8e-64  L
gi|204061|gb|AAA41126.1|  (L10608) erythropoietin [Rattus no...   243   1e-63  L
gi|4261664|gb|AAD13964.1|S65458_1  (S65458) mutant erythropo...   242   1e-63
gi|462017|sp|P33709|EPO_SHEEP  Erythropoietin precursor >gi|...   241   4e-63
gi|90481|pir||A24902  erythropoietin precursor - mouse            240   6e-63
gi|165877|gb|AAA31518.1|  (L10610) erythropoietin [Ovis aries]    238   3e-62
gi|20846583|ref|XP_144595.1|  (XM_144595) similar to Erythro...   238   3e-62  L
gi|20342068|ref|XP_112305.1|  (XM_112305) similar to Erythro...   237   6e-62
gi|1352380|sp|P49157|EPO_PIG  Erythropoietin precursor >gi|2...   236   1e-61
gi|8920355|emb|CAB96416.1|  (AJ249745) erythropoietin [Sus s...   236   1e-61
gi|11527285|gb|AAG36961.1|AF290943_1  (AF290943) erythropoie...   232   2e-60
gi|11527287|gb|AAG36962.1|  (AF290944) erythropoietin [Oryct...   231   2e-60
gi|2144696|pir||I46083  erythropoietin precursor - cat (frag...   231   3e-60
gi|1706680|sp|P33708|EPO_FELCA  Erythropoietin precursor >gi...   229   1e-59
gi|462015|sp|P33707|EPO_CANFA  Erythropoietin precursor >gi|...   205   3e-52
gi|6573159|gb|AAF17572.1|AF202311_1  (AF202311) erythropoiet...   190   9e-48
gi|6671275|gb|AAF23133.1|  (AF202310) erythropoietin [Homo s...   114   6e-25
gi|460893|gb|AAB29659.1|  erythropoietin, Epo (N-terminal) [...    70   2e-11
gi|1911129|emb|CAA72707.1|  (Y11971) erythropoietin [Mus mus...    43   0.002  L
```

- **Sequence alignments and database searching are key to all of bioinformatics**
- **There are four different methods for doing sequence comparisons 1) Dot Plots; 2) Dynamic Programming; 3) Fast Alignment; and 4) Multiple Alignment**
- **Understanding the significance of alignments requires an understanding of statistics and distributions**

NCBI Entrez Genome Projects



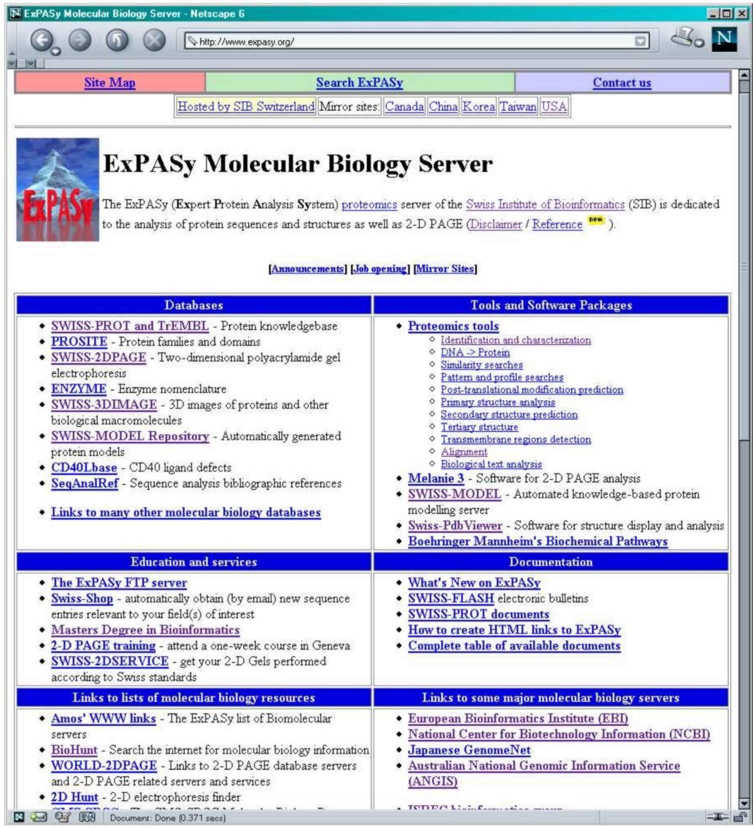
Welcome to the NCBI Entrez Genome Project database.
This searchable database is a collection of complete and incomplete large-scale sequencing, assembly, annotation, and mapping projects for cellular organisms. The database is organized into organism-specific overviews that function as portals from which all projects in the database pertaining to that organism can be browsed and retrieved. Read more...

NCBI Resources

Entrez Gene	gene-related information
Entrez Genome	sequence and map data from whole genomes
Eukaryotic Projects	eukaryotic-specific genome projects
Genomic Biology	organism-specific links
Prokaryotic Projects	prokaryotic-specific genome projects
Organellar Genomes	organellar reference sequences and tools
Plant Genomes	major plant genome projects
RefSeq	the reference sequence project
Viral Genomes	viral reference sequences and tools
WGS Sequences	whole genome shotgun sequences

SWISS-PROT/TrEMBL

- Collaboration between the SIB (CH) and EMBL/EBI (UK)
- SWISS-PROT: Fully annotated (manually), non-redundant, cross-referenced, documented protein sequence database
- TrEMBL: is automatically generated (from annotated EMBL coding sequences (CDS)) and annotated using software tools



ExPASy Molecular Biology Server

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 (Disclaimer / Reference).

[Announcements] [Job opening] [Mirror Sites]

Databases

- SWISS-PROT and TrEMBL - Protein knowledgebase
- **PROSITE** - Protein families and domains
- SWISS-2DPAGE - Two-dimensional polyacrylamide gel electrophoresis
- **ENZYME** - Enzyme nomenclature
- SWISS-3DIMAGE - 3D images of proteins and other biological macromolecules
- **SWISS-MODEL Repository** - Automatically generated protein models
- **CD40Lbase** - CD40 ligand defects
- **SeqAnalRef** - Sequence analysis bibliographic references
- **Links to many other molecular biology databases**

Tools and Software Packages

- **Proteomics tools**
  - Identification and characterization
  - DNA -> Protein
  - Similarity searches
  - Pattern and profile searches
  - Post-translational modification prediction
  - Primary structure analysis
  - Secondary structure prediction
  - Tertiary structure
  - Transmembrane regions detection
  - Alignment
  - Biological text analysis
- **Melanie 3** - Software for 2-D PAGE analysis
- SWISS-MODEL - Automated knowledge-based protein modelling server
- **Swiss-PdbViewer** - Software for structure display and analysis
- **Boehringer Mannheim's Biochemical Pathways**

Education and services

- **The ExPASy FTP server**
- **Swiss-Shop** - automatically obtain (by email) new sequence entries relevant to your field(s) of interest
- **Masters Degree in Bioinformatics**
- **2-D PAGE training** - attend a one-week course in Geneva
- SWISS-2DSERVICE - get your 2-D Gels performed according to Swiss standards

Documentation

- **What's New on ExPASy**
- **SWISS-FLASH** electronic bulletins
- **SWISS-PROT documents**
- **How to create HTML links to ExPASy**
- **Complete table of available documents**

Links to lists of molecular biology resources

- **Amos' WWW links** - The ExPASy list of Biomolecular servers
- **BioHunt** - Search the internet for molecular biology information
- **WORLD-2DPAGE** - Links to 2-D PAGE database servers and 2-D PAGE related servers and services
- **2D Hunt** - 2-D electrophoresis finder

Links to some major molecular biology servers

- **European Bioinformatics Institute (EBI)**
- **National Center for Biotechnology Information (NCBI)**
- **Japanese GenomeNet**
- **Australian National Genomic Information Service (ANGIS)**

Protein (amino acid) databases

- They are big databases too:

  -Swiss-Prot (very high level of annotation)

  http://au.expasy.org/

  -PIR (protein identification resource) the world's most comprehensive catalog of information on proteins

  http://www.pir.uniprot.org/

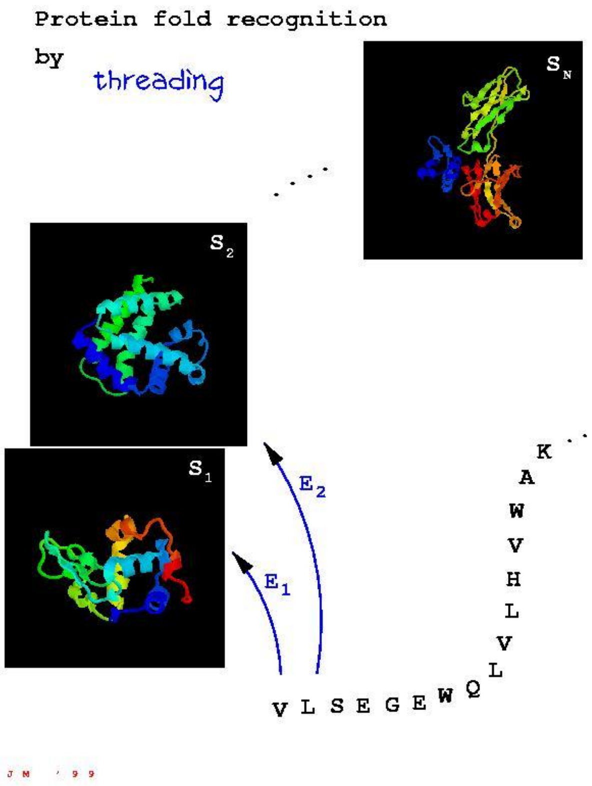
- Trans atabases:

  -TREMBL (translated EMBL): *includes entries that have not been annotated yet into Swiss-Prot.*

  http://www.ebi.ac.uk/trembl/access.html

  -GenPept (translation of coding regions in GenBank)

Assigning function by similarity

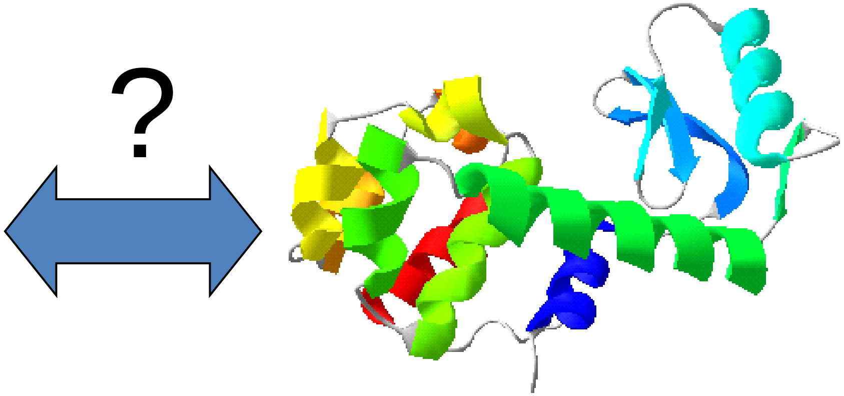


- Sequence similarity: BLAST and others
- Beyond sequence similarity: matching sequences and shapes (threading)

Protein Structure Modeling

- *Ab initio* modeling
- Threading & Fold Recognition
- Homology Modeling

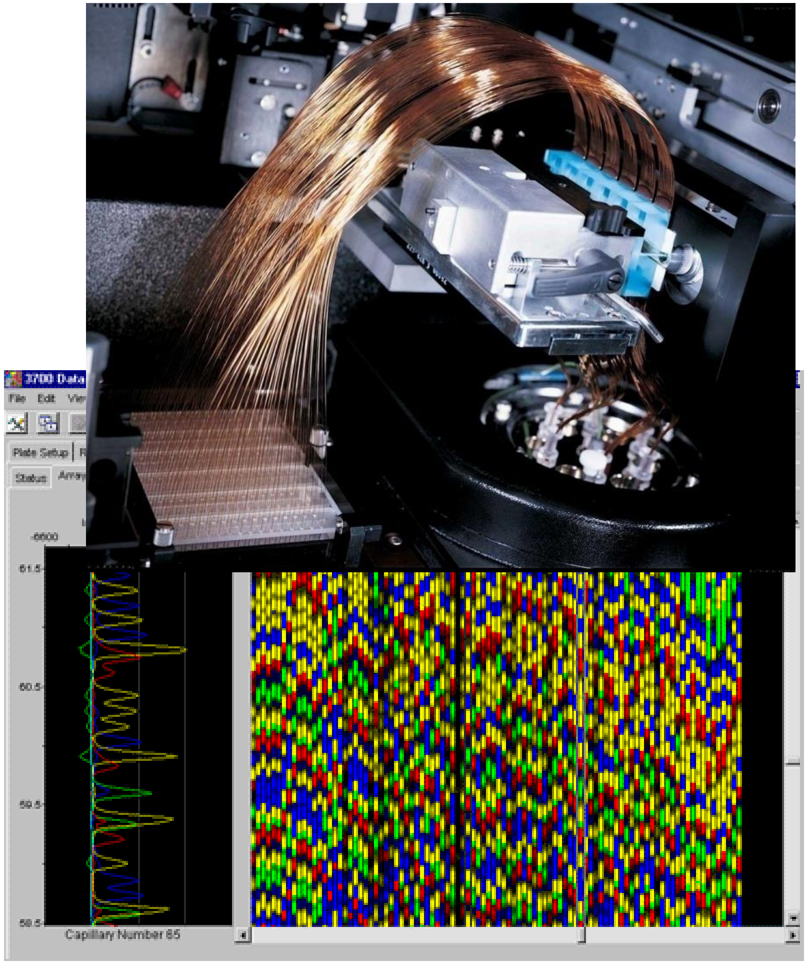
```
MNIFEMLRID EGLRLKIYKD TEGYYTIGIG
HLLTKSPSLN AAKSELDKAI GRNCNGVITK
DEAEKLFNQD VDAAVRGILR NAKLKPVYDS
LDAVRRCALI NMVFQMGETG VAGFTNSLRM
LQQKRWDEAA VNLAKSRWYN QTPNRAKRVI
TTFRTGTWDA YKNL
```



High-throughput DNA sequencing

- Top image: confocal detection by the MegaBACE sequencer of fluorescently labeled DNA

- Bottom image: computer image of sequence read by automated sequencer

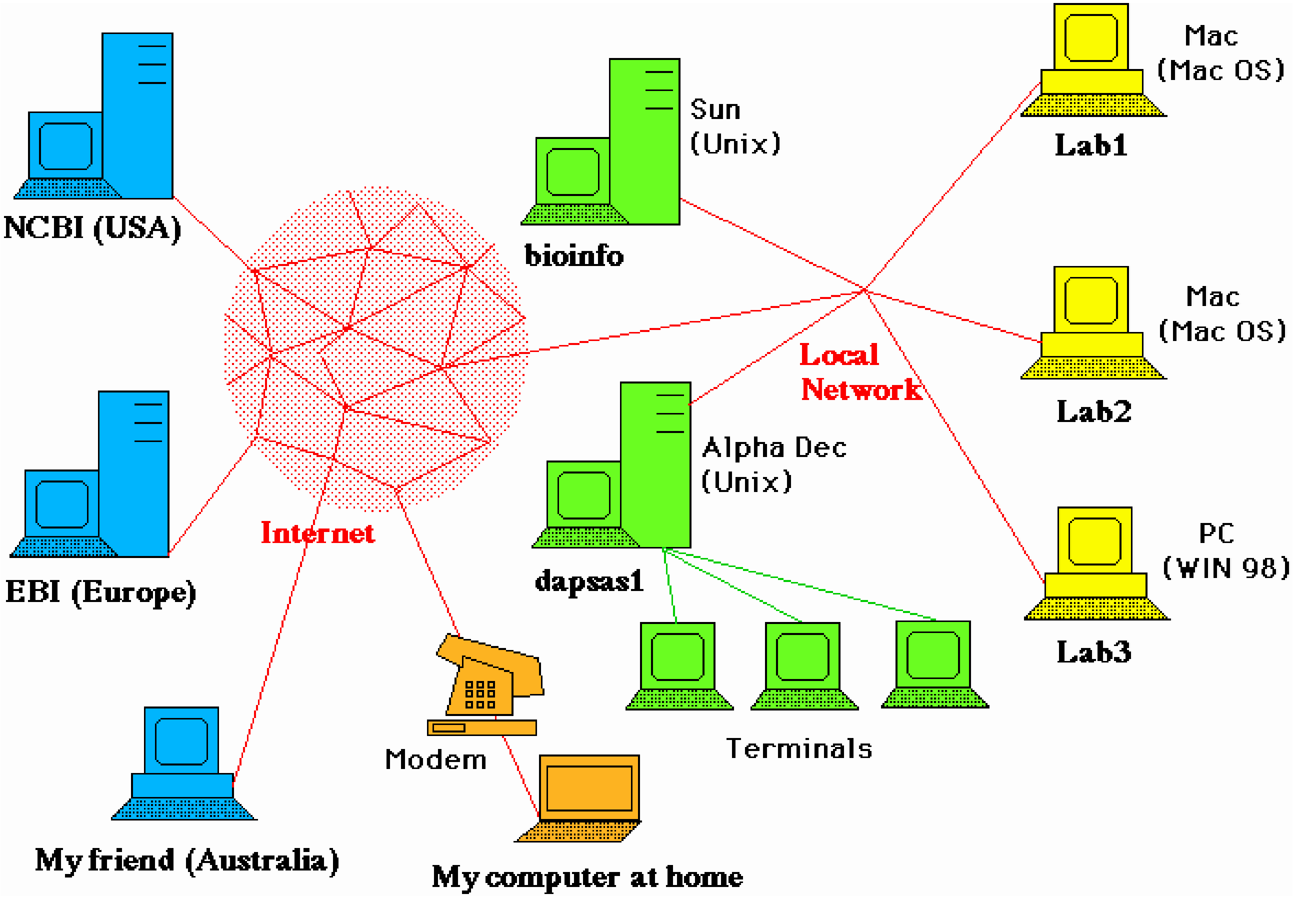


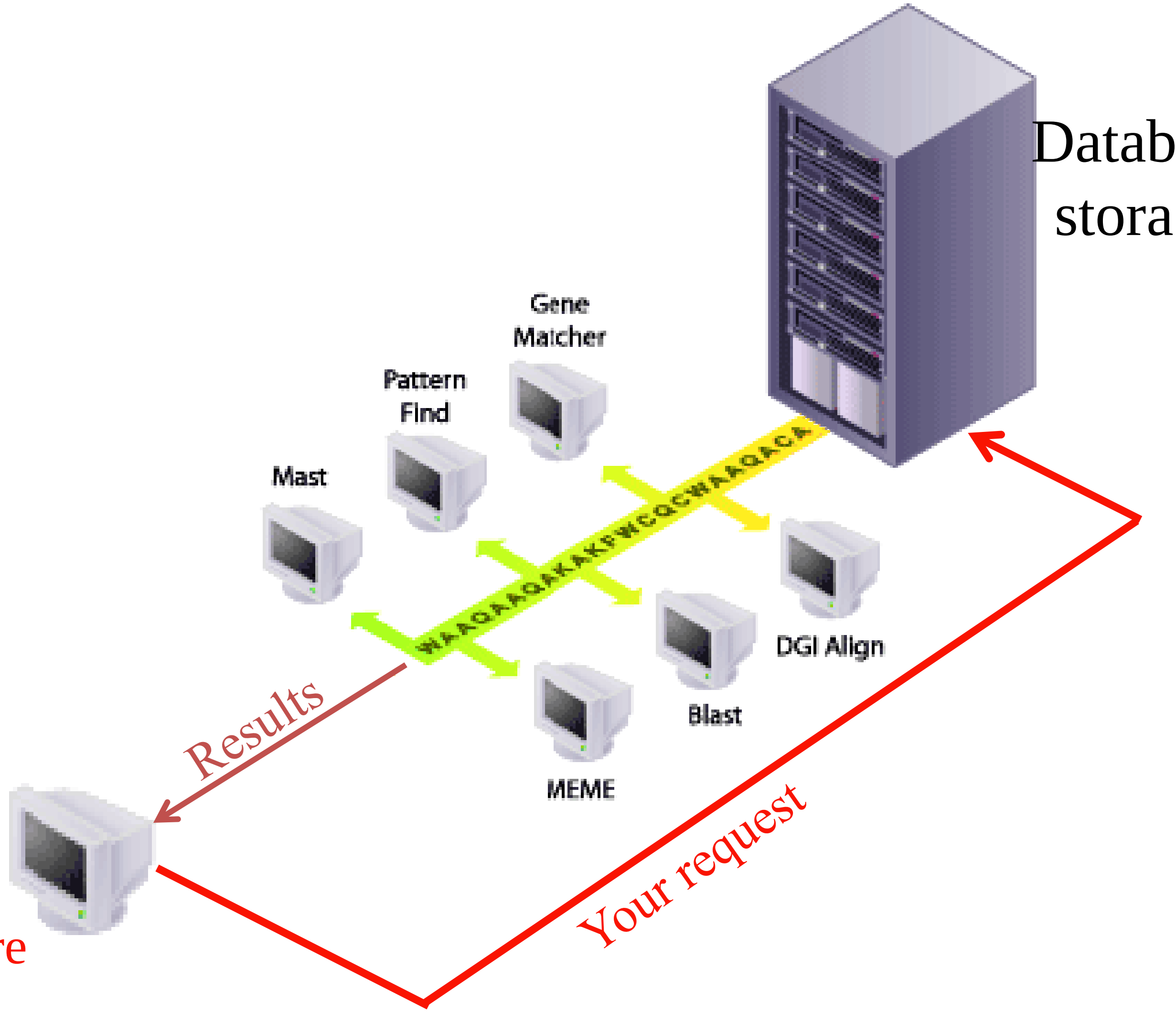
The Challenges of Molecular Biology Computing

- The big dataset problem
- DNA sequencing
- Pairwise and Multiple Alignments
- Similarity searching the databanks
- Structure-function relationships; Can sequence patterns predict function?
- Phylogenetic analysis: Sequence conservation across evolution
- Genomics

Why do we need the Internet?

- "omics" projects and the information associated with involve a huge amount of data that is stored on computers all over the world.
- Because it is impossible to maintain up-to-date copies of all relevant databases within the lab. Access to the data is via the internet.





Database storage

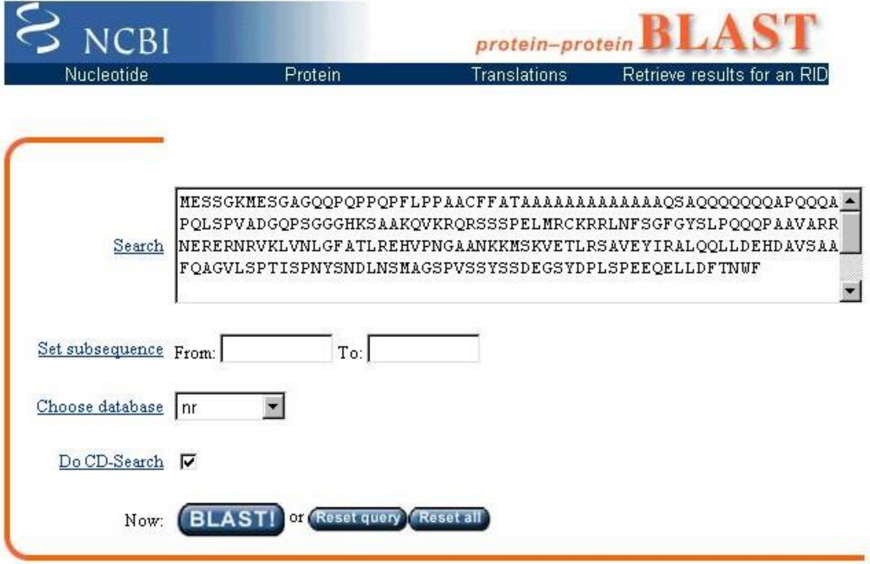
Results

Your request

You are here

A practical example of sequence alignment

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



Search
MESSGKMESGAGQQPQPPQPFLPPAACFFATAAAAAAAAAAAAAQSAQQQQQQQAPQQQA
PQLSPVADGQPSGGGHKSAAKQVKRQRSSSPELMRCKRRLNFSGFGYSLPQQQPAAVARR
NERERNRVKLVNLGFATLREHVPNGAANKKMSKVETLRSAVEYIRALQQLLDEHDAVSAA
FQAGVLSPTISPNYSNDLNSMAGSPVSSYSSDEGSYDPLSPEEQELLDFTNWF

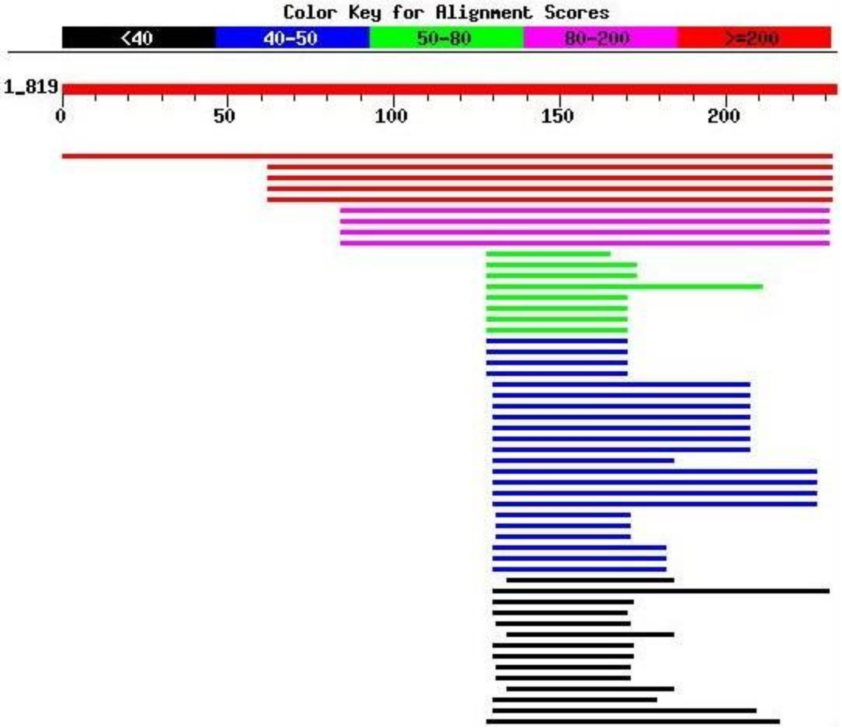
Set subsequence From: To:

Choose database nr

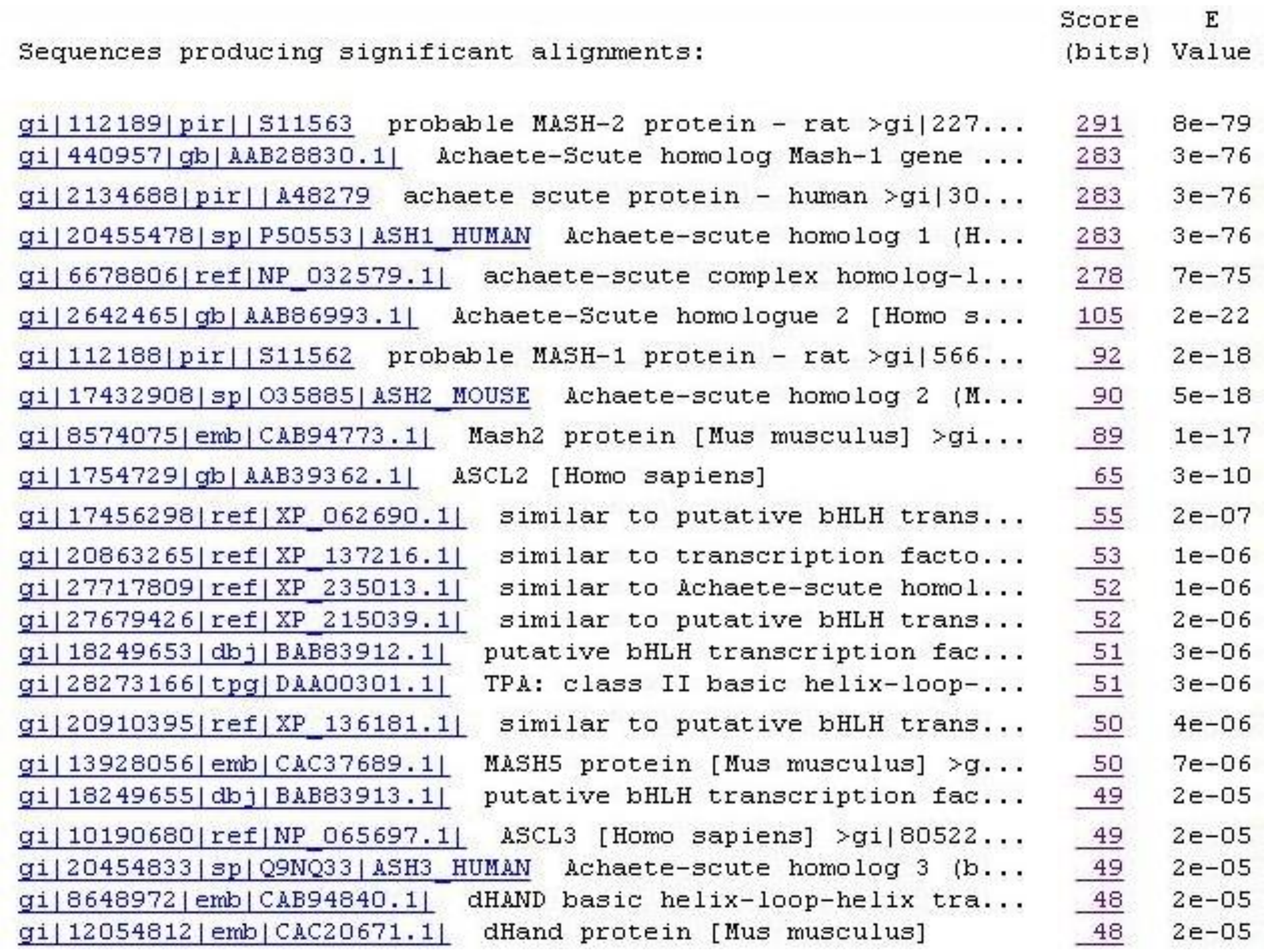
Do CD-Search ☑

Now: BLAST! or Reset query Reset all

BLAST results



Detailed BLAST results



```
                                                                   Score    E
Sequences producing significant alignments:                        (bits) Value

gi|112189|pir||S11563  probable MASH-2 protein - rat >gi|227...     291   8e-79
gi|440957|gb|AAB28830.1|  Achaete-Scute homolog Mash-1 gene ...     283   3e-76
gi|2134688|pir||A48279  achaete scute protein - human >gi|30...     283   3e-76
gi|20455478|sp|P50553|ASH1_HUMAN  Achaete-scute homolog 1 (H...     283   3e-76
gi|6678806|ref|NP_032579.1|  achaete-scute complex homolog-l...     278   7e-75
gi|2642465|gb|AAB86993.1|  Achaete-Scute homologue 2 [Homo s...     105   2e-22
gi|112188|pir||S11562  probable MASH-1 protein - rat >gi|566...      92   2e-18
gi|17432908|sp|O35885|ASH2_MOUSE  Achaete-scute homolog 2 (M...      90   5e-18
gi|8574075|emb|CAB94773.1|  Mash2 protein [Mus musculus] >gi...      89   1e-17
gi|1754729|gb|AAB39362.1|  ASCL2 [Homo sapiens]                      65   3e-10
gi|17456298|ref|XP_062690.1|  similar to putative bHLH trans...      55   2e-07
gi|20863265|ref|XP_137216.1|  similar to transcription facto...      53   1e-06
gi|27717809|ref|XP_235013.1|  similar to Achaete-scute homol...      52   1e-06
gi|27679426|ref|XP_215039.1|  similar to putative bHLH trans...      52   2e-06
gi|18249653|dbj|BAB83912.1|  putative bHLH transcription fac...      51   3e-06
gi|28273166|tpg|DAA00301.1|  TPA: class II basic helix-loop-...      51   3e-06
gi|20910395|ref|XP_136181.1|  similar to putative bHLH trans...      50   4e-06
gi|13928056|emb|CAC37689.1|  MASH5 protein [Mus musculus] >g...      50   7e-06
gi|18249655|dbj|BAB83913.1|  putative bHLH transcription fac...      49   2e-05
gi|10190680|ref|NP_065697.1|  ASCL3 [Homo sapiens] >gi|80522...      49   2e-05
gi|20454833|sp|Q9NQ33|ASH3_HUMAN  Achaete-scute homolog 3 (b...      49   2e-05
gi|8648972|emb|CAB94840.1|  dHAND basic helix-loop-helix tra...      48   2e-05
gi|12054812|emb|CAC20671.1|  dHand protein [Mus musculus]            48   2e-05
```

- E value: is the expectation value or probability to find by chance hits similar to your sequence. The lower the E, the more significant the score.

Most widely used sites for sequence analysis

- Sites for alignment of 2 sequences:
  - T-COFFEE (http://www.ch.embnet.org/software/TCoffee.html): *more accurate than ClustalW for sequences with less than 30% identity.*
  - ClustalW (http://www.ch.embnet.org/software/ClustalW.html; http://align.genome.jp)
  - bl2sequ (http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi)
  - LALIGN (http://www.ch.embnet.org/software/LALIGN_form.html)
  - MultiALIGN (http://prodes.toulouse.inra.fr/multalin/multalin.html)
- Sites for DNA to protein translation:

These algorithms can translate DNA sequences in any of the 3 forward or three reverse sense frames.

  - Translate (http://au.expasy.org/tools/dna.html)
  - Translate a DNA sequence: (http://www.vivo.colostate.edu/molkit/translate/index.html)
  - Transeq (http://www.ebi.ac.uk/emboss/transeq)

Methods of Analysis of mARRAY Data

- **Supervised:** Class prediction, from set of labeled genes/data or experiment to the models that define relationship between gene expression and classes.
- **Unsupervised**: Class discovery, from set of unlabelled genes/ data to the discovery of the underplaying classes by clustering genes or experiments with similar pattern of expression

Unsupervised

- *Clustering*
- *Gene trees (hierarchical dendrograms)*
- *K-means (non-hierarchical)*
- *Similar to a known gene or average of genes*
- *Similar to a specific pattern*
- *Selections of genes*
- *Genes with high statistical confidence*
- *Genes with relative expression in certain ranges*

Unsupervised

- *Comparison with existing knowledge*
- *Comparison with results of old analyses*
- *Comparison with results of analyses by colleagues*
- *Pathway analysis finding genes that fit in a certain place in a pathway.*
- *Sequence analysis to automatically find potential regulatory sequences.*
- *Find the most significant patterns in experiments (can be used for principal component analysis)*

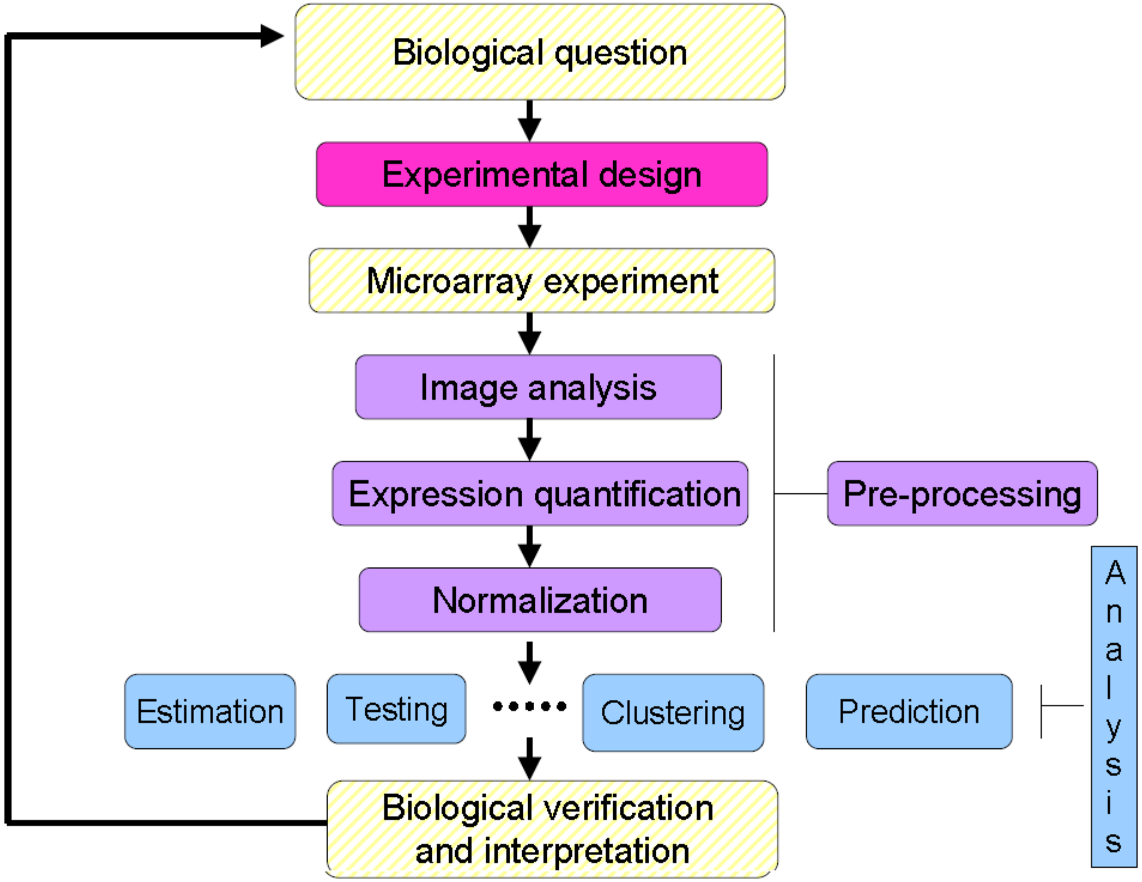
Visualization and Graphics packages

- *DISLIN - Scientific Data Plotting Software*
- *EasyPlot Scientific Graphing*
- *GGobi Data Visualization System - high interaction dynamic graphics*
- *SigmaPlot*
- *STATISTICA from StatSoft*

Tools of Visualisation:

- ✓ **The Box-Whisker Plot**
- ✓ **The Dendrogram**
- ✓ **The Multi-dimensional Scaling Plot**

Design for Micoarray Expt.



Biological question

Experimental design

Microarray experiment

Image analysis

Expression quantification

Normalization

Pre-processing

Estimation

Testing

Clustering

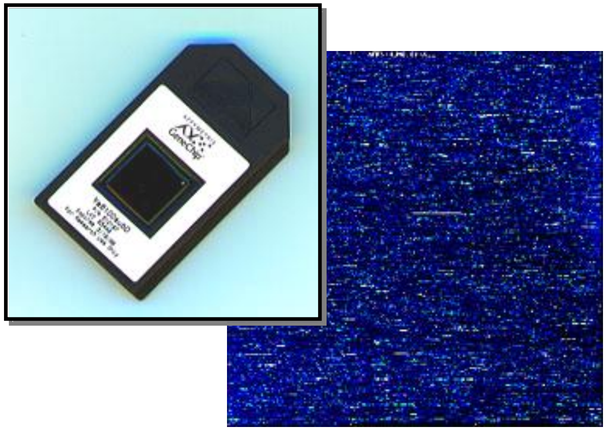
Prediction

Analysis

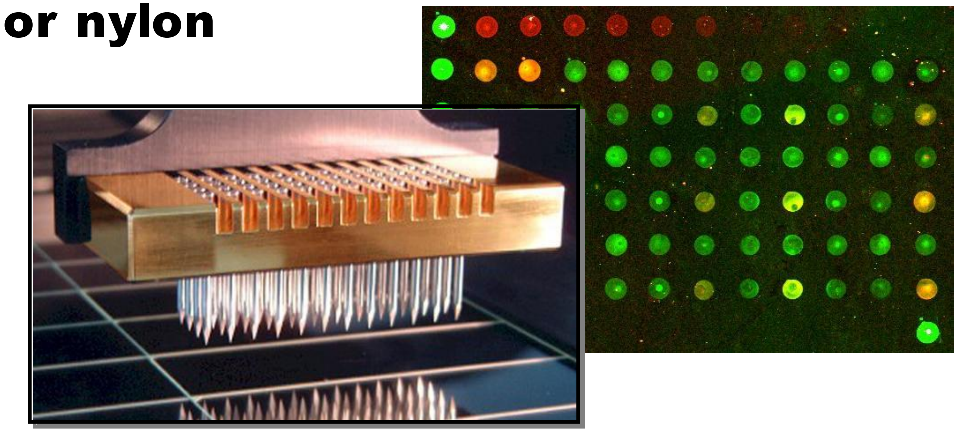
Biological verification and interpretation

Microarray technologies

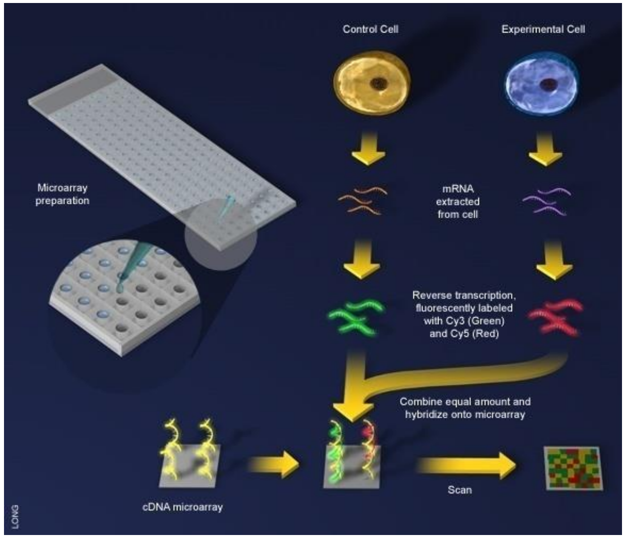
- Affymetrix Genechip
  - **Oligonucleotides (25mers) synthesized on a glass slide (photolitography).**



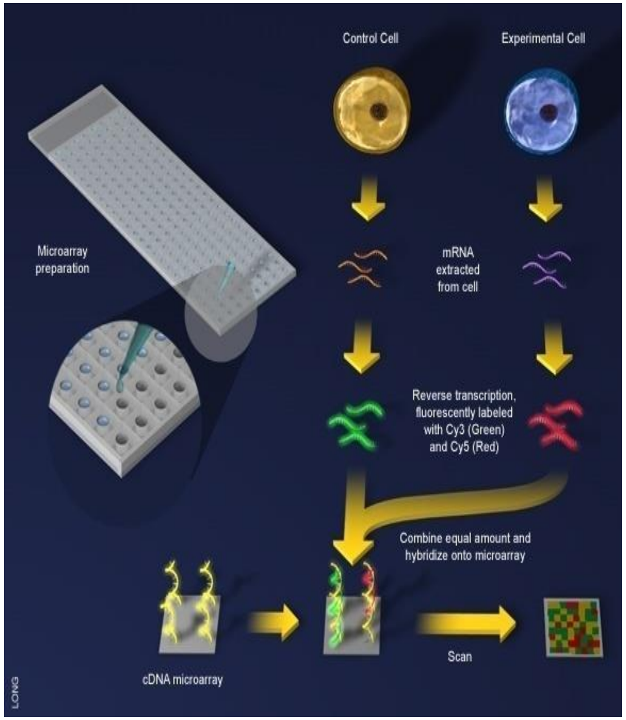
- Two-color chips
  - **cDNA or oligonucleotides spotted on glass or nylon membranes**



Kits are available that are designed to be used with dual-color microarray analysis in which differential expression of genes is being identified and quantified. Two different cDNA probes are prepared independently for this analysis.

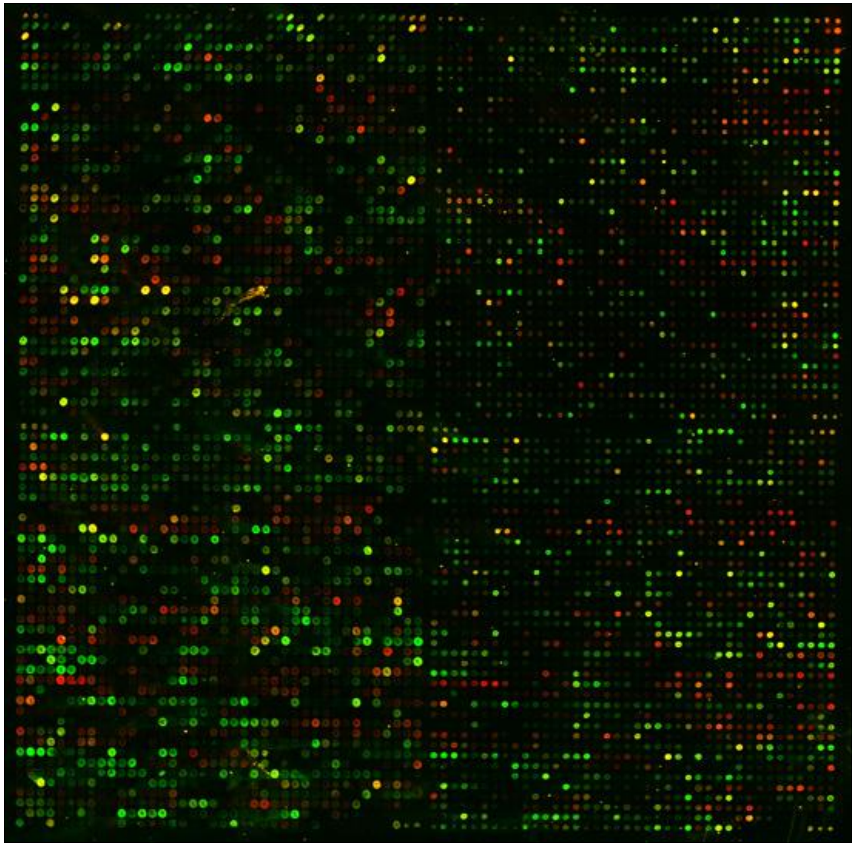


Source:**http://scq.ubc.ca**, artist: Jiang Long)



- Always compares two samples
- Uses two different colors (red/green)
- Slides have probes printed on them

Source:**http://scq.ubc.ca**, artist: Jiang Long)



- Microarray Labeling Kits provide the solution for high sensitivity dual-color differential gene expression analysis.
- Traditionally the probes have been labeled by direct incorporation of Cy3- or Cy5-labeled deoxynucleotides into

  first-strand cDNA during synthesis

Standard Statistical Tests for Differential Expression of Individual Genes

Two Groups:

- Fold change
- Student’s t-test: gene-specific, global, or regularized.

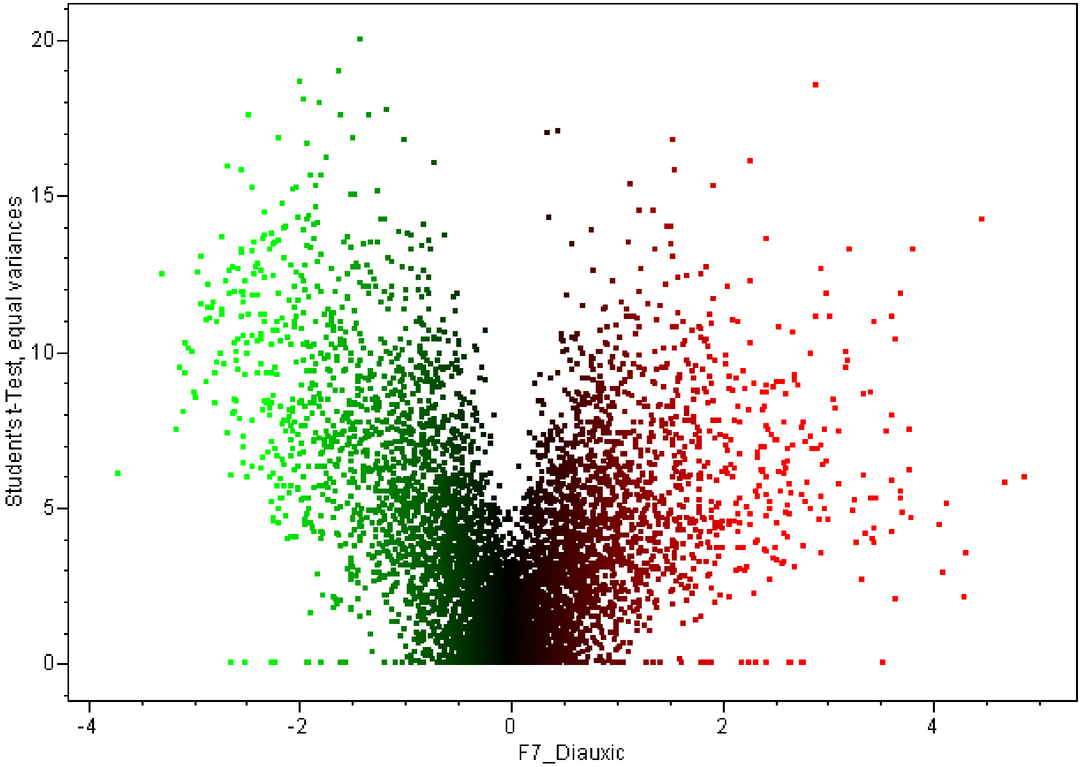
More than two groups:

- ANOVA : gene-specific, global, or regularized F-test.

Significance and multiple testing:

- Nominal p-values
- Family-wise error rate: the probability of accumulating one or more false-positive errors over a number of statistical tests.
- False discovery rate: the proportion of false positives among all of the genes initially identified as being differentially expressed.

Volcano plot



Scatter plot of -log(p-value) from a t-test *vs.* log ratio. Visualises fold-change and statistical significance at the same time: Find genes that are significant and have large fold change, **and** genes that are significant but have small fold change.

expression measures: MAS 4.0

Affymetrix GeneChip MAS 4.0 software uses AvDiff, a trimmed mean:

$$AvDiff = \frac{1}{\# J} \sum_{j \in J} (PM_j - MM_j)$$

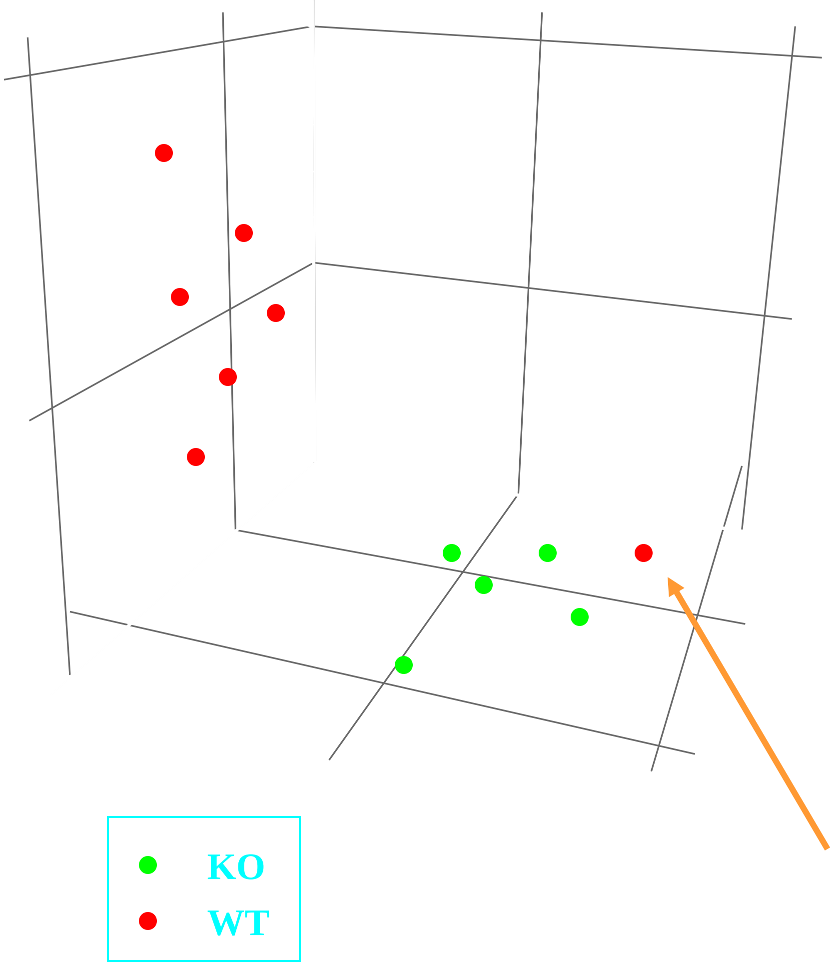
- sort $d_j = PM_j - MM_j$
- exclude highest and lowest value
- J := those pairs within 3 standard deviations of the average

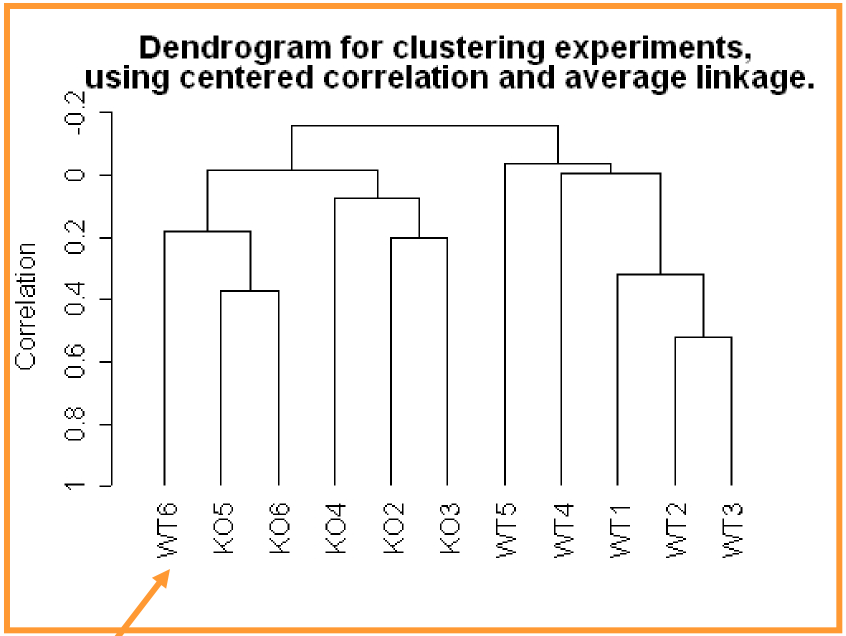
Quality control

Outlier detection

Principal components analysis, PCA

Hierarchical clustering

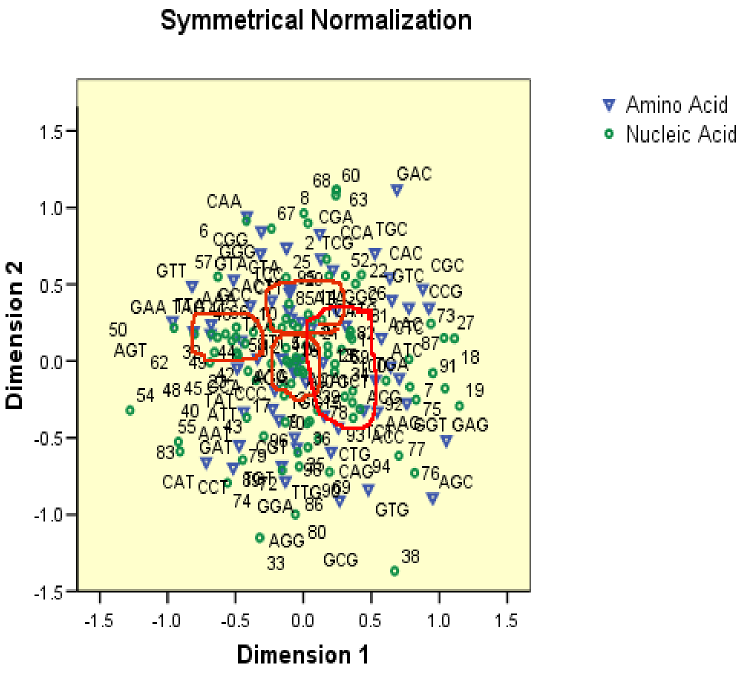


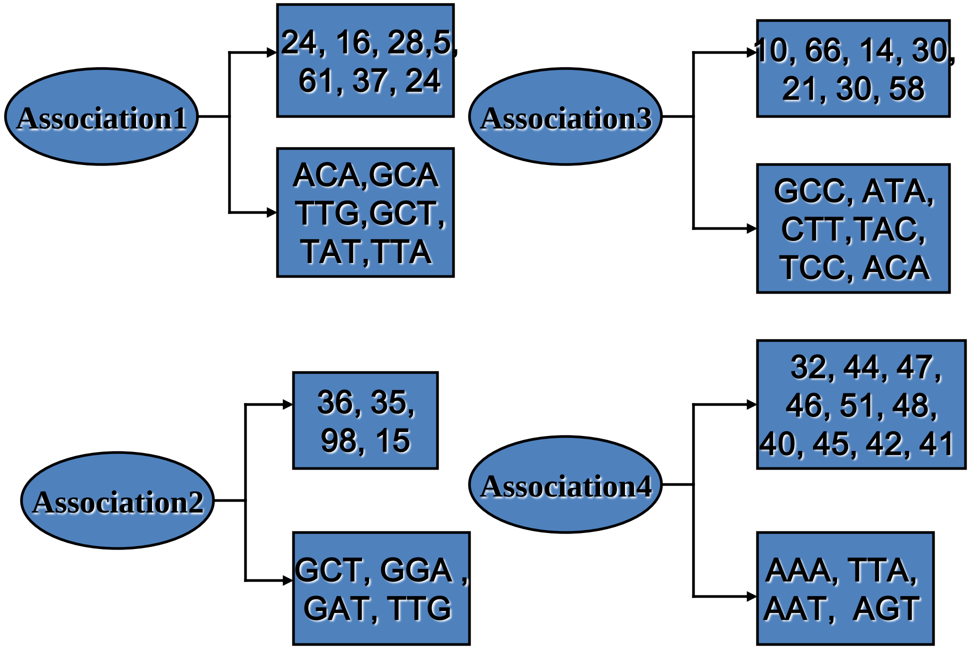


outlier

Results & Discussion

Biplot between Row and column points:





Association1

24, 16, 28,5, 61, 37, 24

ACA,GCA TTG,GCT, TAT,TTA

Association3

10, 66, 14, 30, 21, 30, 58

GCC, ATA, CTT,TAC, TCC, ACA

Association2

36, 35, 98, 15

GCT, GGA , GAT, TTG

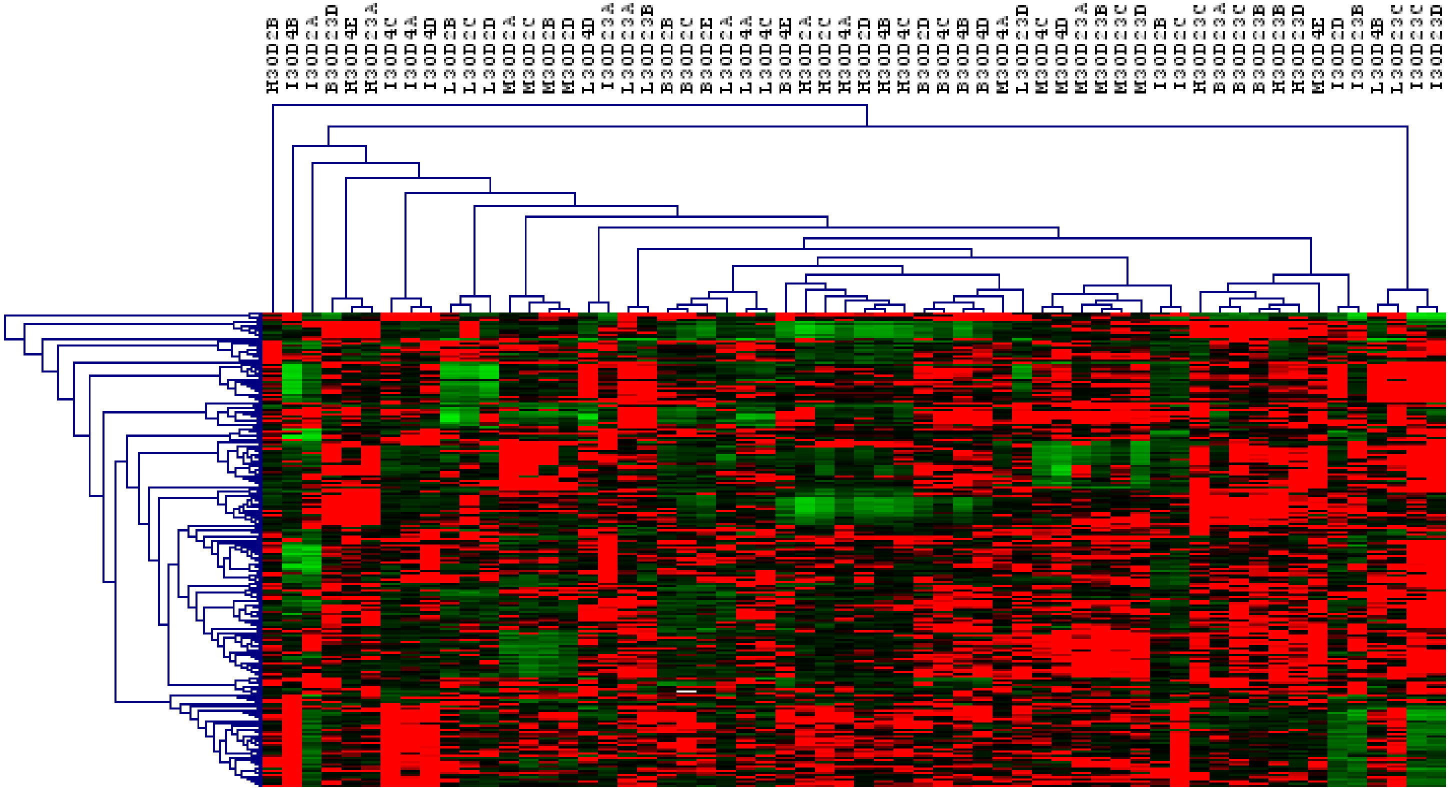
Association4

32, 44, 47, 46, 51, 48, 40, 45, 42, 41

AAA, TTA, AAT, AGT

- The associated nucleotide sequences and codons can be identified.
- AT ending codons are frequent in ntd seqs on the left side of the plot.
- GC ending codons fall on the right side of the plot.
- Genes with higher level of expression fall around the origin or centroid.

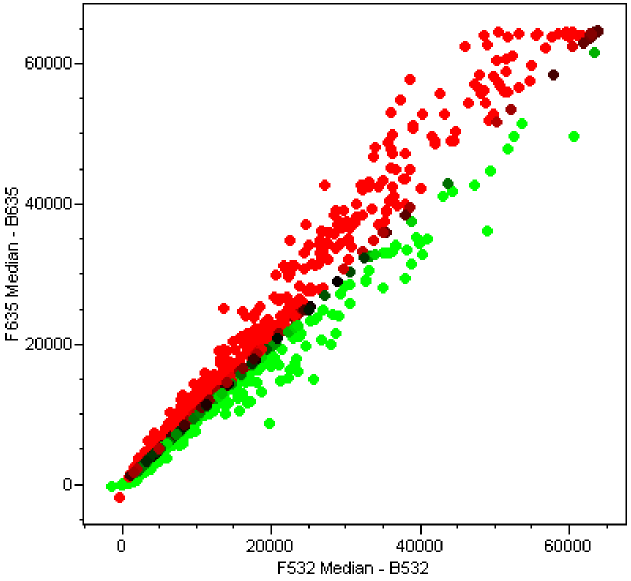
Analysis of Microarray data



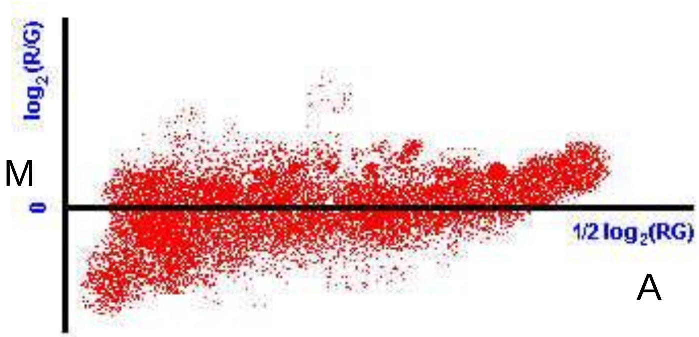
K means Hierarchical clustering of *C. carpio* obtained using MeV Software

Scatter plots

classical scatter plot



M-A plot for microarray analysis



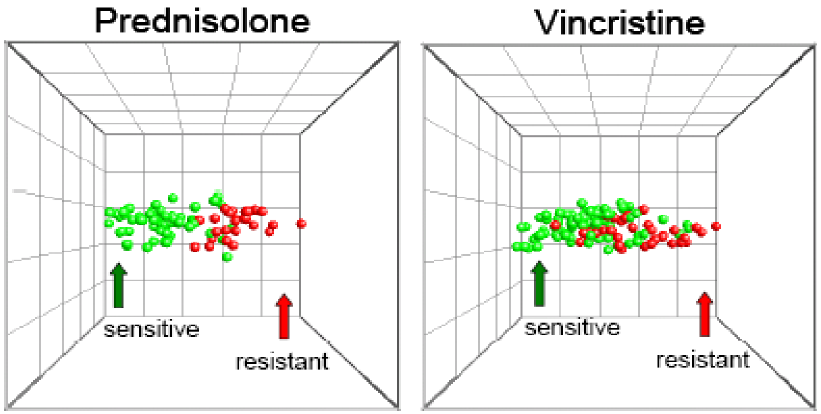
$$M = \log_2\left(R/_G\right)$$

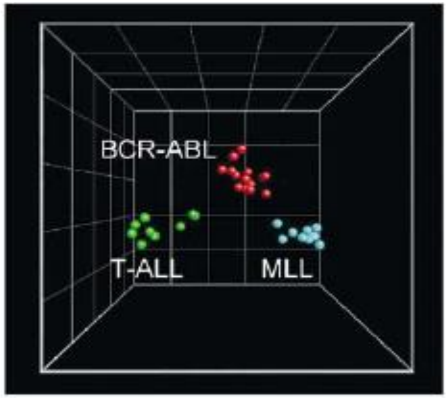
$$A = \log_2\sqrt{RG} \quad \text{OR} \quad 1/_2 \log_2 RG$$

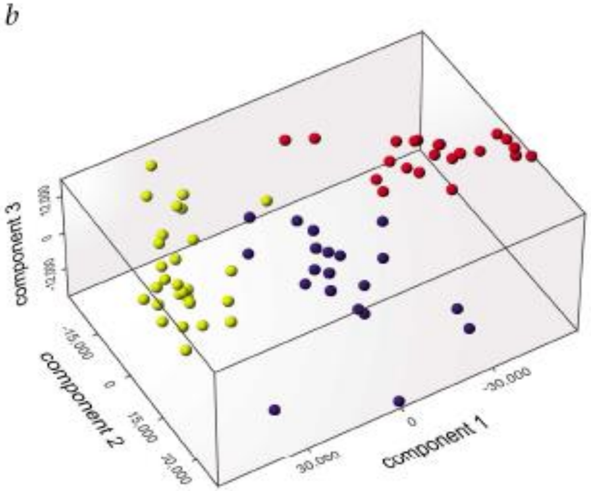
Differentially expressed genes are higher (or lower) in one of the samples

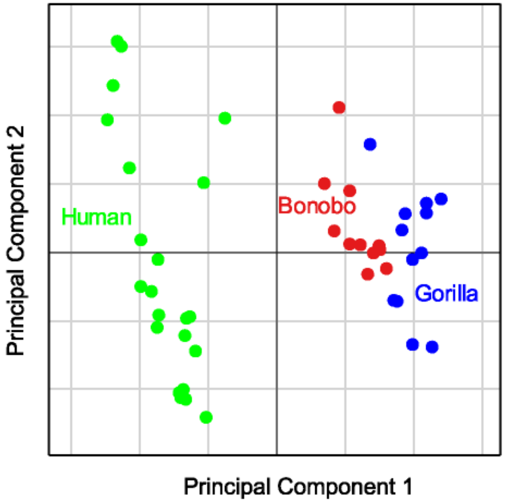
Use an appropriate cut-off (‘distance’ from diagonal) to select relevant genes → **highly arbitrary!**

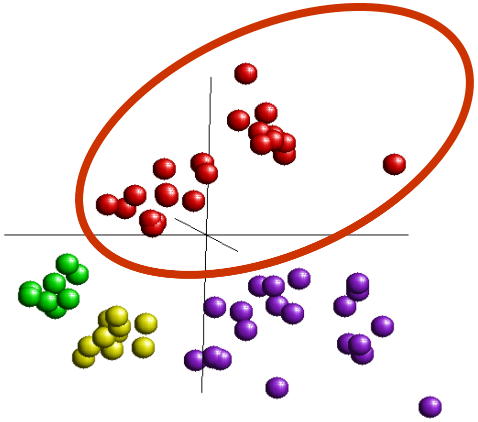
Ordination of Gene Expression Data



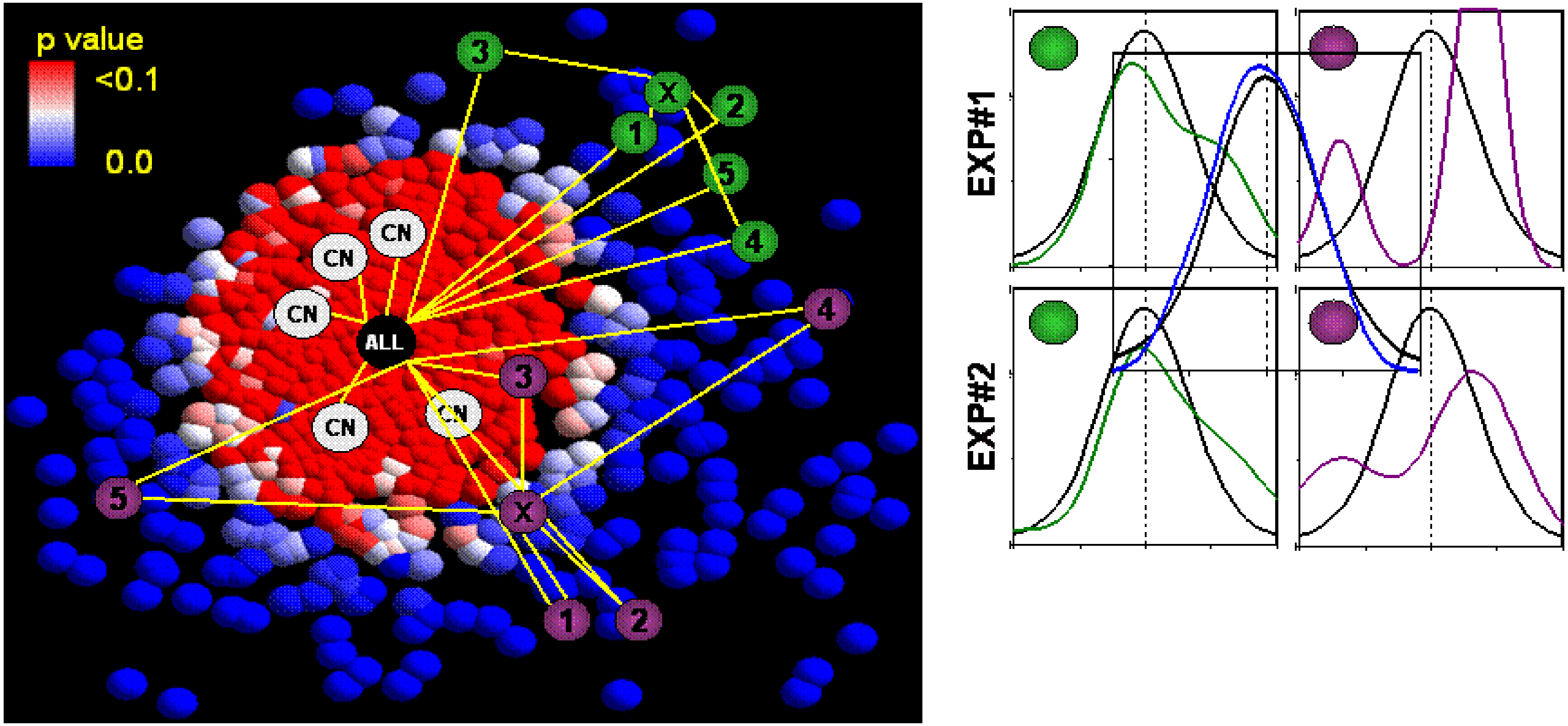




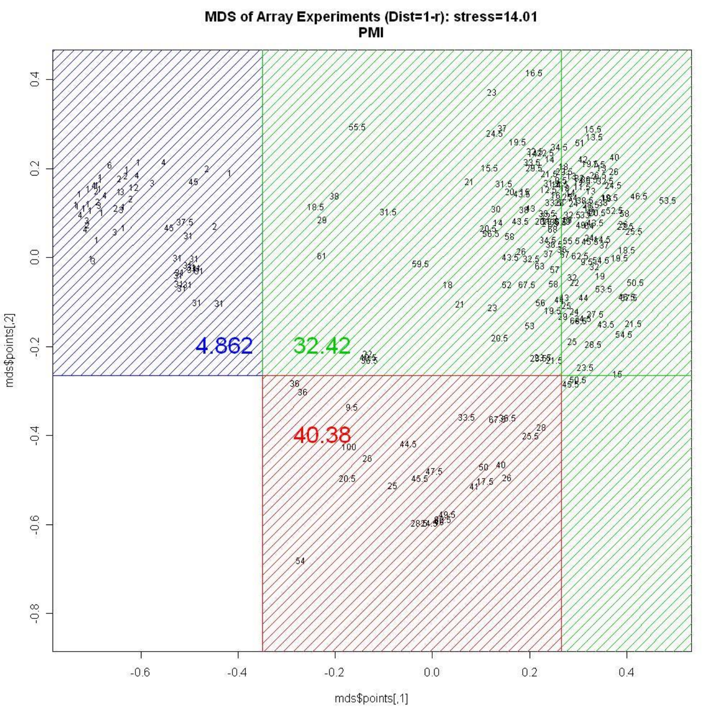




Similar Differential Expression Patterns Across Multiple Experiments



The distribution of gene expression values for each gene group in each sample is plotted as a single point in low dimensional space. This is achieved using Principal Components Analysis along with Non-Metric Multi-Dimensional Scaling.


MDS of Array Experiments (Dist=1-r): stress=14.01
PMI
4.862
32.42
40.38
mds$points[,2]
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
mds$points[,1]
-0.6
-0.4
-0.2
0.0
0.2
0.4

Important Points in Drug Design based on Bioinformatics Tools

- Application of Genome
  - 3 billion bases pair
  - 30,000 unique genes
  - Any gene may be a potential drug target
  - ~500 unique target
  - Their may be 10 to 100 variants at each target gene
  - 1.4 million SNP
  - 10^{200} potential small molecules

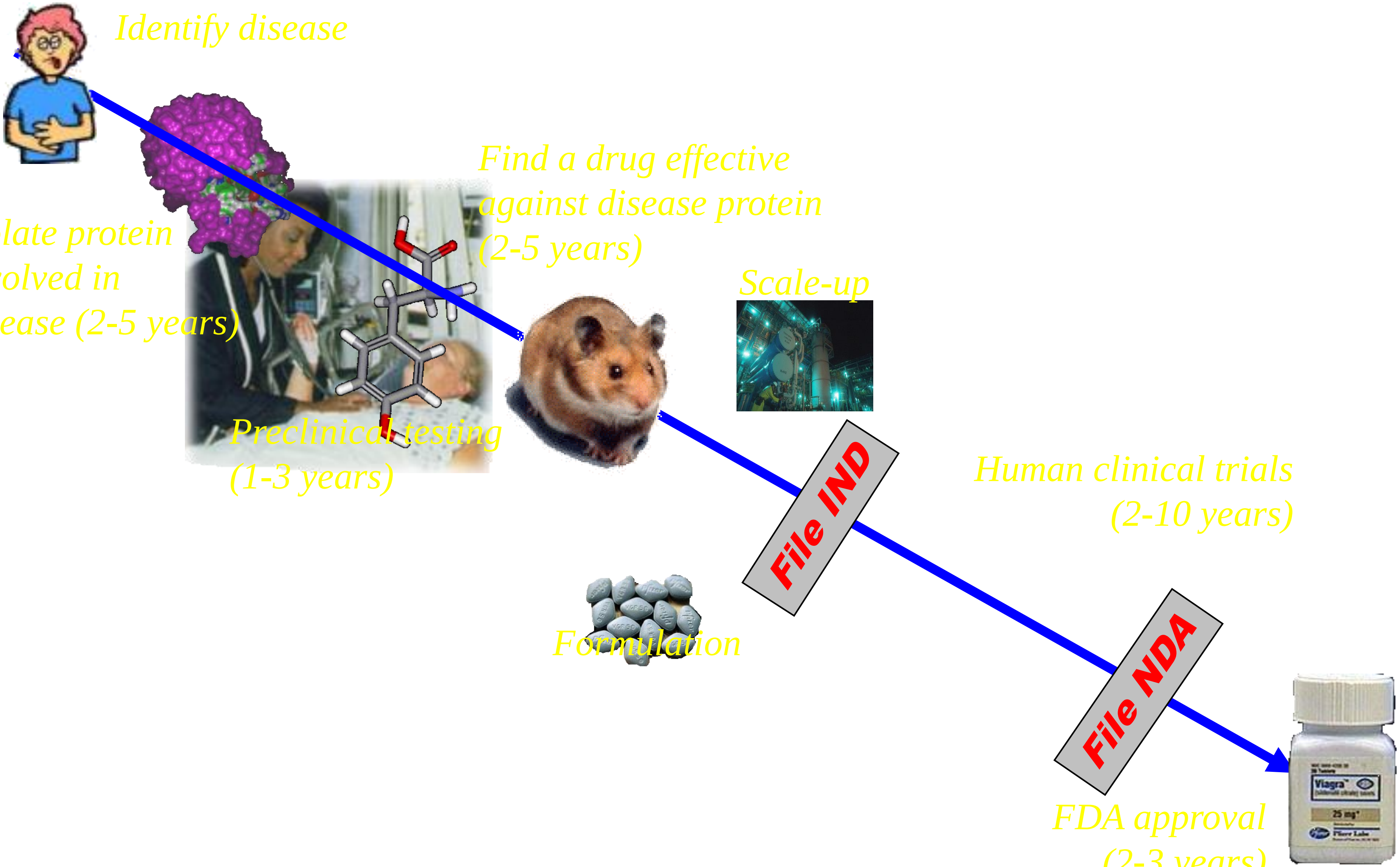
Important Points in Drug Design based on Bioinformatics Tools

- **Detect the Molecular Bases for Disease**
  - Detection of drug binding site
  - Tailor drug to bind at that site
  - Protein modeling techniques
  - Traditional Method (brute force testing)
- **Rational drug design techniques**
  - Screen likely compounds built
  - Modeling large number of compounds (automated)
  - Application of Artificial intelligence
  - Limitation of known structures

Important Points in Drug Design based on Bioinformatics Tools

- **Refinement of compounds**
  - Refine lead compounds using laboratory techniques
  - Greater drug activity and fewer side effects
  - Compute change required to design better drug
- **Quantitative Structure Activity Relationships (QSAR)**
  - Compute functional group in compound
  - QSAR compute every possible number
  - Enormous curve fitting to identify drug activity
  - chemical modifications for synthesis and testing.
- **Solubility of Molecule**
- **Drug Testing**

Drug Discovery & Development



Identify disease

Isolate protein involved in disease (2-5 years)

Find a drug effective against disease protein (2-5 years)

Preclinical testing (1-3 years)

Scale-up

Formulation

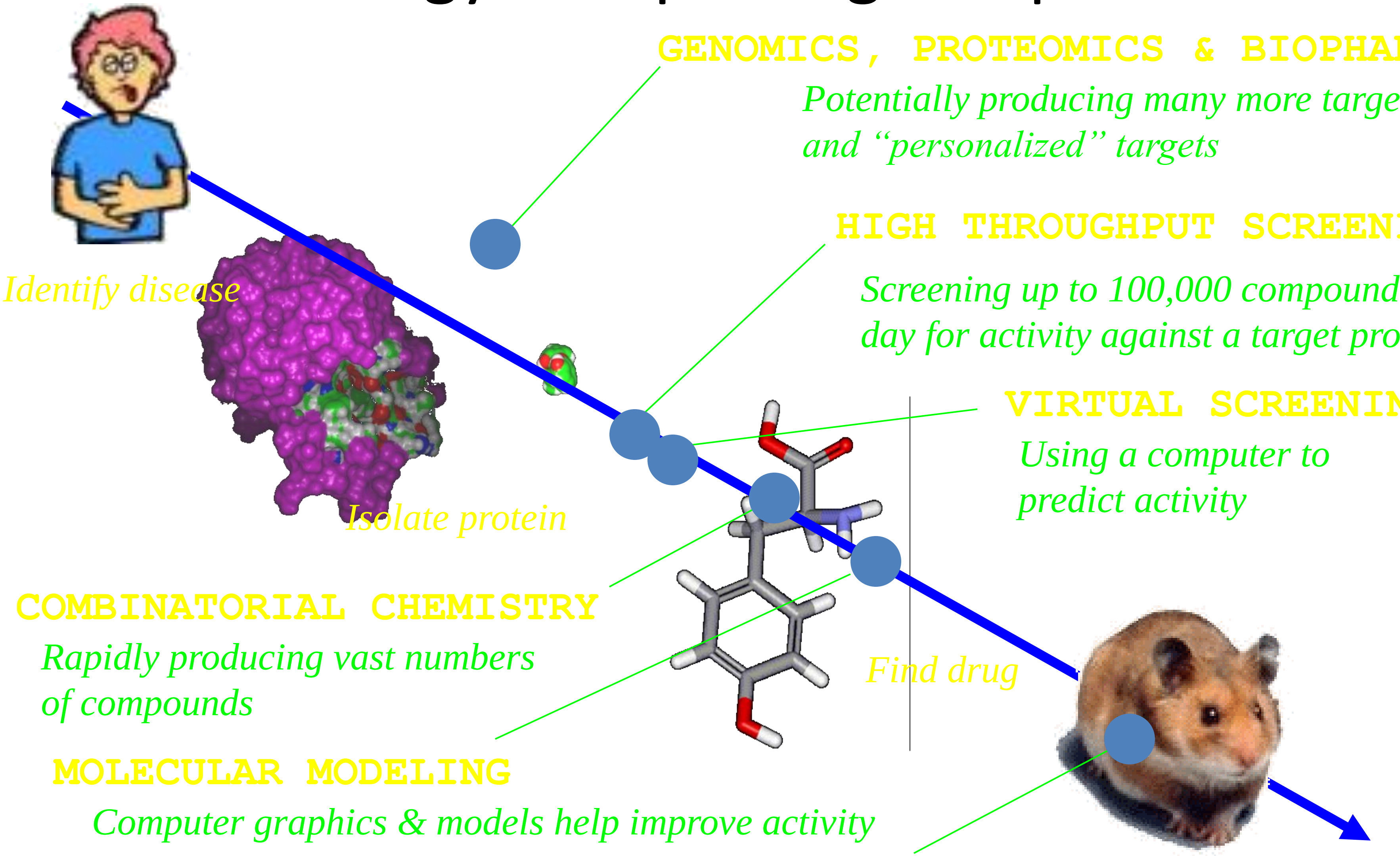
File IND

Human clinical trials (2-10 years)

File NDA

FDA approval (2-3 years)

Techology is impacting this process



Identify disease

Isolate protein

Find drug

Preclinical testing

GENOMICS, PROTEOMICS & BIOPHARM.

Potentially producing many more targets and "personalized" targets

HIGH THROUGHPUT SCREENING

Screening up to 100,000 compounds a day for activity against a target protein

VIRTUAL SCREENING

Using a computer to predict activity

COMBINATORIAL CHEMISTRY

Rapidly producing vast numbers of compounds

MOLECULAR MODELING

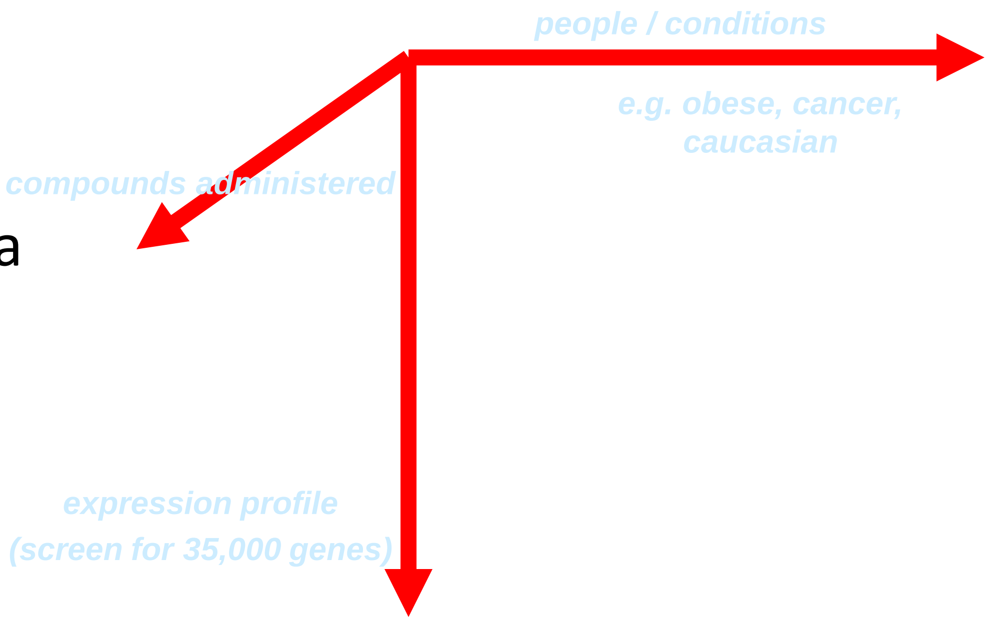
Computer graphics & models help improve activity

IN VITRO & IN SILICO ADME MODELS

Tissue and computer models begin to replace animal testing

1. Gene Chips

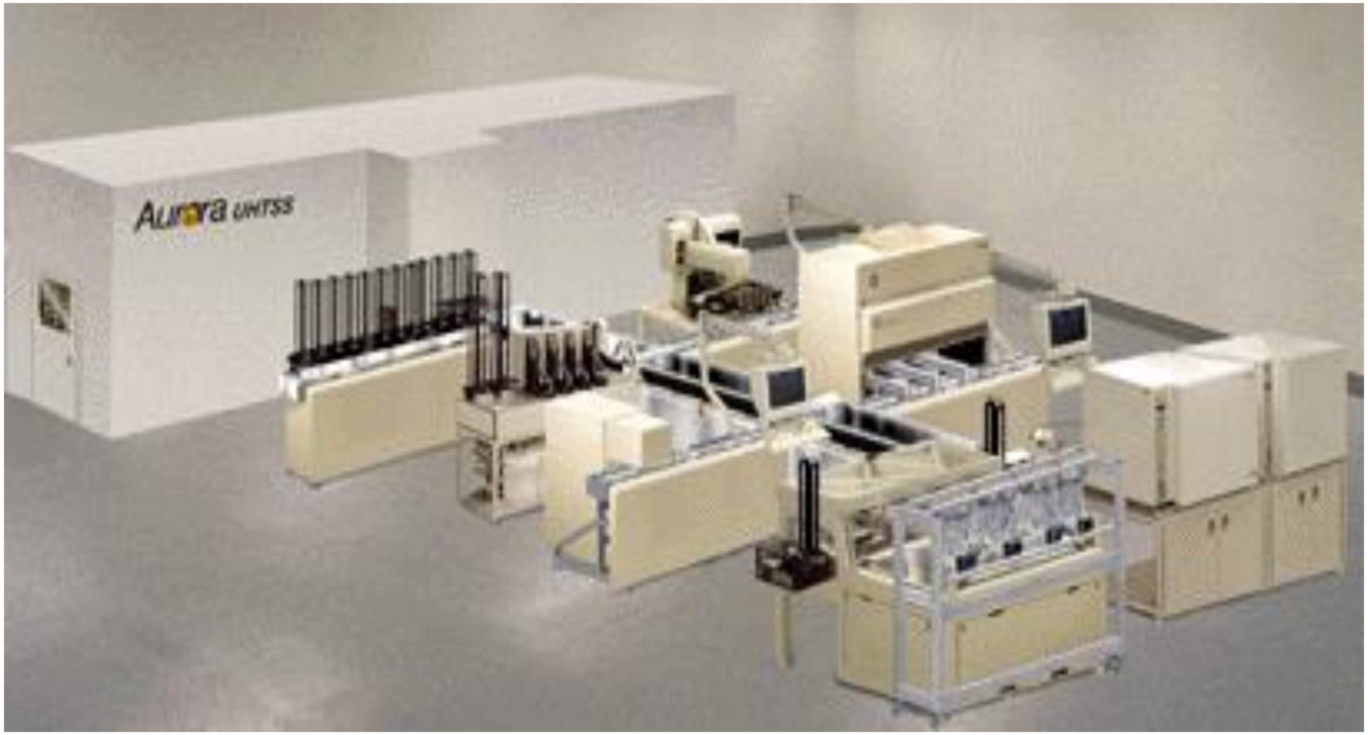
- "Gene chips" allow us to look for changes in protein expression for different people with a variety of conditions, and to see if the presence of drugs changes that expression
- Makes possible the design of drugs to target different phenotypes



Biopharmaceuticals

- Drugs based on proteins, peptides or natural products instead of small molecules (chemistry)
- Pioneered by biotechnology companies
- Biopharmaceuticals can be quicker to discover than traditional small-molecule therapies
- Biotechs now paring up with major pharmaceutical companies

2. High-Throughput Screening



Screening perhaps millions of compounds in a corporate collection to see if any show activity against a certain disease protein

High-Throughput Screening

- Drug companies now have millions of samples of chemical compounds
- High-throughput screening can test 100,000 compounds a day for activity against a protein target
- Maybe tens of thousands of these compounds will show some activity for the protei
- The chemist needs to intelligently select the 2 - 3 classes of compounds that show the most promise for being drugs to follow-up

Application

- **Agroinformatics / Agricultural Informatics:** Agroinformatics concentrates on the aspects of bioinformatics dealing with plant genomes.
- **Systems Biology:** Systems biology is the coordinated study of biological systems by investigating the components of cellular networks and their interactions, by applying experimental high-throughput and whole-genome techniques, and integrating computational methods with experimental efforts.
- **Bioprogramming:** Bioprogramming identifies and assembles the main technical and biological criteria which are used to define the physical design of the facility.
- **Drug Design & Development: Drug Design** (often represented as *molecular modeling*, in protein structure prediction) is the approach of finding drugs by design, based on their biological targets.

Web Resources in Bioinformatics

Bioinformatics Web Sites-Freely Accessible

GeneAtlas: http://genatlas.medecine.univ-paris5.fr/fiche.php?symbol=ESR1

General biocomputing services

- ReNaBi (France)
- GenomeNet WWW Server (Kyoto, Japan)
- Center for Information Biology (Mishima, Japan)
- Pôle Bio-Informatique Lyonnais (Lyon, France)
- European Bioinformatics Institute (Hinxton, UK)
- DKFZ Theoretical Bioinformatics (Heidelberg, Germany)
- ExPASy Molecular Biology Server (Geneva, Switzerland)
- National Center for Genome Resources (Santa Fe, USA)
- National Center for Biotechnology Information (Bethesda, USA)
- Swiss Institute for Experimental Cancer Research (Lausanne, Switzerland)
- Mobyle Portal at Pasteur Institute (Paris, France

Molecular databases

General nucleotide databases

- EMBL (Hinxton, UK)
- Genbank (Bethesda, USA)
- DDBJ (Mishima, Japan)

General protein databases

- SWISS-PROT (Geneva, Switzerland)
- PROSITE (Geneva, Switzerland)
- PRODOM (Lyon, France)
- PIR (Washington DC, USA)
- PDB (San Diego, USA)

Specialized databases

- **Prokaryotes**
  - GenoList (Paris, France)
  - HGT-DB (Tarragona, Spain)
  - Genome division at NCBI (Bethesda, USA)
  - Genome Information Broker (Mishima, Japan)
- **Eukaryotes**
  - GDB (Baltimore, USA)
  - MGI (Bar Harbor, USA)
  - IMGT (Montpellier, France)
  - OMIM (Bethesda, USA)
  - Ensembl (Hinxton, UK)
- **Complete genomes projects**
  - Genomes OnLine Database (Chicago, USA)
- **Gene expression**
- Sage GENIE (Bethesda, USA)

Retrieval systems

- DBGET (Kyoto, Japan)
- WWW-Query (Lyon, France)
- Public SRS servers (Cambridge, UK)

Alignments and similarity searches on-line

- **Similarity searches**
- BLAST at NCBI (Bethesda, USA)
- BLAST at PBIL (Lyon, France)
- BLAST at ISREC (Lausanne, Switzerland)
- WU-BLAST at EMBL (Hinxton, UK)
- FASTA and SSEARCH at EBI (Hinxton, UK)

Pairwise alignments

- LFASTA at PBIL (Lyon, France) *Local alignments*
- SIM4 at PBIL (Lyon, France) *Aligns cDNA and genomic DNA*
- WISE2 at Pasteur (Paris, France) *Aligns protein and genomic DNA*
- SIM at ExPASy (Geneva, Switzerland)
- BLAST two sequences at NCBI (Bethesda, USA)
- SIM, GAP, NAP and LAP at Michigan Tech. Univ. (Houghton, USA)
- JAligner (Alexandria, Egypt) *Java implementation of the Smith-Waterman algorithm*

Multiple alignments

- BLAST+ClustalW at EMBL (Heidelberg, Germany) *Multiple alignment of homologous sequences detected by BLAST*
- ClustalW2 at EBI (Hinxton, UK) *Display and edit alignments with JalView*
- ClustalW at DDBJ (Mishima, Japan)
- ClustalW at GenomeNet (Kyoto, Japan)
- ClustalW at Pasteur (Paris, France)
- ClustalW at PBIL (Lyon, France) *Colored alignments and secondary structure predictions*
- Multalin at PBIL (Lyon, France) *Colored alignments and secondary structure predictions*
- Multalin at INRA (Toulouse, France) *Colored alignments*
- MAFFT at MiB (Fukuoka, Japan)

Multiple Allignment

- MUSCLE at EBI (Hinxton, UK).
- DIALIGN2 at Pasteur (Paris, France)
- DIALIGN2 at BiBiServ (Bielefeld, Germany)
- DCA at BiBiServ (Bielefeld, Germany)
- MAP at Michigan Tech. Univ. (Houghton, USA)
- T-COFFEE at SIB (Lausanne, Switzerland)
- MATCH-BOX at MBR (Namur, Belgium)
- MAVID/AMAP at UCB (Berkeley, USA)
- BlastAlign at BioAfrica (Pretoria, South Africa)

Aligns nucleotide sequences that have large indels

- Kalign at the Karolinska Institute (Stockholm, Sweden)
- **Motif/Pattern/Profile searches**
- PIMA II at BMERC (Boston, USA)

Methods
Gene recognition

- GeneMark at GeorgiaTech (Atlanta, UK)
- GrailEXP at ORNL (Oak Ridge, USA)
- GenScan at MIT (Boston, USA)
- GenScan at Pasteur (Paris, France)

- **DNA analysis**
- Repeat Masker (Seattle, USA)
- Bend.it at ICGEB (Trieste, Italy)
- Signal scan service (Bethesda, USA)
- Promoter scan service (Bethesda, USA)
- Dragon promoter finder (Singapore)
- Elements of transcription (Neuherberg, Germany)
- **RNA analysis**
- ESSA RNA software (Toulouse, France)
- The RNA world at IMB (Jena, Germany)
- **Protein analysis**
- SignalP server (Lyngby, Denmark)
- ExPASy protein analysis tools (Geneva, Switzerland)

- **RNA analysis**
- ESSA RNA software (Toulouse, France)
- The RNA world at IMB (Jena, Germany)
- **Protein analysis**
- SignalP server (Lyngby, Denmark)
- ExPASy protein analysis tools (Geneva, Switzerland)

Downloadable software

- CodonW (Oxford, UK)

Index of services in molecular biology

- The BioCatalog (Hinxton, UK)
- Institut Pasteur (Paris, France)
- Bioinformatics links (Lyngby, Denmark)
- Atelier de BioInformatique (Marseille, France)
- Bioinformatics software on the Web (Baltimore, USA)
- Amos Bairoch's WWW links page (Geneva, Switzerland)

Journals

- Gene
- Nature
- Science
- Bioinformatics
- Nature Genetics
- The BioInformer
- Nucleic Acids Research
- Journal of Biological Chemistry
- Molecular Biology and Evolution

Web Resources for Bioinformatics

- http://www.faseb.org/opar/protfold/protein.html
- http://bioinformatics.org/faq/
- http://www.hhmi.org/news/baker2.html
- http://bioinfo.mshri.on.ca/trades/
- http://www.ncbi.nlm.nih.gov/Education/
- http://bioinformatics.org/faq/
- http://www.toplab.de/proteomics.htm
- http://www.wiley.co.uk/wileychi/genomics/proteomics.html
- http://everything2.com/?node=proteome
- http://us.expasy.org/proteomics_def.html
- http://www.sdu.dk/Nat/CPA/proteomics.html
- http://www.accessexcellence.org/AB/BC/Gregor_Mendel.html
- http://www.laskerfoundation.org/news/gnn/timeline/1888.html
- http://www.webref.org/scientists/
- http://dmoz.org/Science/Biology/Genetics/History/
- http://www.cshl.org/
- http://bioinformatics.org/faq/
- http://www.netsci.org/Science/Bioinform/feature06.html
- http://www.emc.maricopa.edu/faculty/farabee/BIOBK/BioBookDNAMOLGEN.html
- http://www.accessexcellence.org/AE/AEPC/WWC/1994/geneticstln.html
- http://www.mun.ca/biology/scarr/4241/TKAMgenetics.html
- http://www.cs.iastate.edu/jva/jva-archive.shtml
- http://www-sop.inria.fr/acacia/personnel/Fabien.Gandon/lecture/uk1999/history/
- http://inventors.about.com/library/inventors/blsoftware.htm
- http://www.nature.com/genomics/

Foundation Tutorials for Bioinformatics Aspirants: Computer Tutorials for Science Stream Students

- **Introduction to computer Concepts:**
- (www.compume.com)
- (www.grassrootsdesign.com)
- (www.fayette.k12.il.us)
- (www.glencoe.com)
- (www.comedition.com)
- (www.hitmill.com)
- (www.pstcc.cc.tn.us)

Java/Perl Tutorials

- **Java Tutorials:**
- (java.sun.com/docs/books/tutorial/) (http://docs.oracle.com/javase/tutorial/)
- (http://www.oracle.com/technetwork/java/index.html)
- (javaboutique.internet.com/tutorials/)
- (www.freewarejava.com/tutorials/index.shtml)
- (www.javacoffeebreak.com/tutorials/)
- (www.apl.jhu.edu/~hall/java/FAQs-and-Tutorials.html)
- (http://www.javaworld.com/)
- **Perl Tutorials:**
- (www.devdaily.com/perl/)
- (www.pageresource.com/cgirec/index2.htm)
- (www.perl.com/pub/q/resources)
- (www.perlmonks.org)
- Perl for Biologists (Weizmann Institute) (http://www.uni-hohenheim.de/~rebhan/perl/)
- (www.webknowhow.net/dir/Perl/Tutorials/)
- Perl for Biologists (http://www.mrc-lmb.cam.ac.uk/genomes/jong/perl_bio_book.html)
- (savage.net.au/Perl-tutorials.html)
- Welcome to the Bioperl Project ! (http://www.techfak.uni-bielefeld.de/bcd/ Perl/Bio/welcome.html)

Science Tutorials for Computer Stream Students

- **Introduction to Biology:**
- (scidiv.bcc.ctc.edu)
- (library.thinkquest.org/12413/)
- (www.biology-online.org/tutorials/home.htm)
- (www.lsic.ucla.edu/ls3/tutorials/)
- (biology-online.org/)
- Cartoon Guide to Genetics (http://www.amazon.com/exec/obidos/ASIN/0062730991/o/qid=957735943/sr=2-1/103-6986286-3259052)
- (genomebiology.com/tutorials/)
- (http://biomed.nus.sg/HIS/txt/menu/tacmenu.html)
- Tutorials in Molecular Biology (http://locutus.lsic.ucla.edu/ls3/tutorials/)
- (http://www.iacr.bbsrc.ac.uk)
- (http://gened.emc.maricopa.edu/Bio/BIO181/BIOBK/BioBookgloss.html) - BioBook Glossary
- (http://highveld.com) - Internet Directory of Biology and Biotechnology
- (http://esg-www.mit.edu:8001/esgbio/7001main.html) - ESG Biology Hyper-textbook Home Page)
- (http://web.ornl.gov/sci/techresources/Human_Genome/redirect.shtml) - DOE Primer on Molecular Genetics)

About DNA

- **About RNA:**
- (http://biog-101-104.bio.cornell.edu/BioG101_104/ tutorials/recomb_DNA. html)
- (http://avery.rutgers.edu/WSSP/Tutorials/) (chime plug-in required)
- (www.umass.edu/molvis/freichsman/)
- (www.tutorgig.com/showurls.jsp?group=6732&index=0)
- (www.scientific.org/tutorials/articles/riley/riley.html)
- (www.tulane.edu/~biochem/nolan/lectures/rna/intro.htm)
- (http://lenti.med.umn.edu/recombinant_dna/recombinant_flowchart.html)
- DNA tutorial (http://www.101science.com/dna.html)
- DNA from the beginning (http://www.dnaftb.org/)
- Central Dogma Glossary (http://homepage.smc.edu/ hodson_kent/Dictionary/ Glossary.htm)
- (www.imsb.au.dk/~raybrown/)
- (http://zombie.imsb.au.dk/~raybrown/)
- (ndbserver.rutgers.edu/NDB/structure-finder/tutorials/full_ndb.dna.rna.res.html)
- **About Genome:**
- (http://genomebiology.com/tutorials/)
- (http://www.genomeweb.com/)
- (http://anatomy.med.unsw.edu.au/cbl/GENOME/tutorials.htm)
- (http://rsat.ulb.ac.be/rsat/tutorials/tut_genome-scale-patser.html)
- (home.uchicago.edu/~ebetran/guides.html)
- Basic Genome Glossary - (http://www.nytimes.com/library/national/ science/062600sci-genome-glossary.html)
- Limited Genome Glossary - (http://homepage.smc.edu/ hodson_kent/Dictionary/Glossary.htm)
- Genome Glossary - (https://gitso-outage.oracle.com/ thinkquest)
- The Gene-School Glossary (https://gitso-outage.oracle.com/thinkquest)
- Glossary of Genetic Terms (http://www.nhgri.nih.gov/DIR/VIP/Glossary/)

TUTORIALS-online

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

- Tom Smith and Don Emmeluth have produced a nice little exploration (http://www.angelfire.com/ga2/nestsite2/bioinform.html) of bioinformatics using NCBI resources and tools.
- A promising set of online lecture notes (http://www.genzentrum.lmu.de/) currently under construction by B. Steipe at the Genzentrum (Gene Center) (http://www.genzentrum.lmu.de/) at the Ludwig-Maximilians-Universität München (http://www.uni-muenchen.de/index.html) (University of Munich).

General introduction to biology for computer scientists

- Estrella Mountain Community College (http://www. emc.maricopa.edu/) in the States offers this excellent short introduction (http://www2.estrellamountain.edu/ faculty/farabee/BIOBK/BioBookintro.html)

outstanding interactive tutorial

- The Dolan DNA Learning Center at Cold Spring Harbor has an outstanding interactive tutorial (http://www.dnaftb.org/1/) introducing genetics.
- The Institute of Arable Crop Research Beginner's Guide to Molecular Biology (http://www.rothamsted.bbsrc.ac.uk/notebook/courses/guide/)
- Unilever Education Advanced Series tutorial on proteins (http://www.schoolscience.co.uk/content/5/chemistry/proteins/ index.html

- The University of Arizona (http://www.arizona.edu/) has made available a high-quality tutorial in cell biology (http://www.biology.arizona.edu/cell_bio/cell_bio.html).

Practical Bioinformatics: Lists of Bioinformatics Tutorials Websites

- (http://spdbv.vital-it.ch/TheMolecularLevel/Matics/)
- (http://www.mrc-lmb.cam.ac.uk/rlw/text/bioinfo_tuto/introduction.html)
- (https://genome.ucsc.edu/training.html)
- (http://mendel.informatics.indiana.edu/~yye/lab/teaching/fall2011-I519.php)
- (https://www.bits.vib.be/index.php/training/122-basic-bioinformatics)
- (imp basic concept: https://www.bits.vib.be/index.php/training/ 122-basic-bioinformatics)

Tutorials---

- **Imp-Bioinfo-Solutions:** (http://www.thebioinformatica.com/onlinetutorial.htm)
- **Dummies:** (http://spdbv.vital-it.ch/TheMolecularLevel/Matics/)
- **BioMedNet Research Tools**: (http://research.bmn.com/)
- **CMS Molecular Biology Resource**: (http://mbcf.dfci.harvard.edu/cmsmbr/)
- **Genetics Tutorials**: (http://science.nhmccd.edu/biol/genetics.html)
- **Health Web/Genetics**: (http://healthweb.org/browse.cfm?subjectid=42)
- **Morgan--GeneticTutorial**: (http://morgan.rutgers.edu/morganWebFrames/How_to_use/HTU_Frameset.html)
- **Online Mendelian Inheritance in Men (NCBI):** (http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucleotide)
- **Protein Information Resource**: (http://pir.georgetown.edu/)
- **Talking Glossary of Genetic Terms**: (http://www.genome.gov/glossary.cfm)
- **WWW Virtual Library-Genetics**: (http://www.ornl.gov/sci/techresources/Human_Genome/genetics.shtml)

Bioinformatics Tutorials Useful Resource to Self-Study

- **Bioinformatics(**https://www.biostars.org/p/56401/
- **Introduction to Bioinformatics**
- Bioinformatics (Genomics) (http://post.queensu.ca/~forsdyke/bioinfor.htm)
- Biocomputing in a Nutshell. (http://www.techfak.uni-bielefeld.de/bcd/ ForAll/Basics)
- Biologist's Guide to Internet Resources
- Computational Molecular Biology Course (http://cmb.washington.edu/)
- Course on Bioinformatics (http://www.cbi.pku.edu.cn/Doc/)
- EMBNet Biocomputing Tutorials (http://www.hgmp.mrc. ac.uk/Embnetut/Universi/embnettu.html)
- Finding the genes in the genomic sequences (http://helpdesk.ugent.be/ webhosting/rugac.php)
- The Genetic Programming Tutorial (http://geneticprogramming.com/)
- Jose R. Valverde's training course documents (http://www.es.embnet.org/ Doc/)
- Principles of Computational Biology, Steven Salzberg. (http://www.cs.jhu.edu/~salzberg/cs439.html)
- Principles of Protein Structure Using the Internet (http://www.cryst.bbk.ac.uk/PPS2/course/)
- Practical Course "Bioinformatics: Computer Methods in Molecular Biology" (http://www.icgeb.trieste.it/net/courses/bioinfo98.html)
- Sequence analysis course (José R, Valverde, EMBNet/CNB) (http://www.es.embnet.org/Doc/ECJ/ ECJ-1999-01/coursecnb/)
- Bioinformatics (http://post.queensu.ca/~forsdyke/ bioinfor.htm) - An excellent review on genetic code and information processing
- Molecular Sequence Analysis (http://www.sequenceanalysis.com/) - Introductory sequence analysis by Andrew S Louka
- Homology Modelling (http://swift.embl-heidelberg.de/course/) - Protein and homology modeling for beginners
- Biocompanion (http://www.doelz.com/) - Tutorial for sequence analysis
- Bioinformatics and Genomic Analysis (http://www.blc.arizona.edu/courses/bioinformatics/) - Link to graduate student course at the university of Arizona
- EMBnet Biocomputing Tutorials - Introduction (http://www.hgmp.mrc.ac.uk/Embnetut/Universl/index. html)
- Integrative Bioinformatics: Practical Kinetic Modeling of Biological Systems (http://www.bioinformaticsservices.com/bis/resources/cybertext/IBcont.html)

Bioinformatics Tutorials Useful Resource to Self-Study

- Biocomputing For Everyone ! (http://www.techfak.uni-bielefeld.de/bcd/ForAll/welcome.html)
- The Biocomputing Glossary (http://www.cryst.bbk.ac.uk/BCD/ bcdgloss.html)
- Computational Biology Course, Martin Tompa (http://courses.cs. washington.edu/courses/cse527/00wi/)
- Course Distance Learning in Bioinformatics (http://130.88.90.2:8900/)
- Functional genomics glossaries (http://ihome.cuhk.edu.hk/~b400559/ glossaries.html)
- How to become a bioinformatics expert (http://www.techfak.uni-bielefeld.de/bcd/ForAll/Econom/study.html)
- Internet for biologists (http://biobase.dk/Embnetut/Ifb/ifb_intr.html)
- Jose R. Valverde's 'dirty' training course documents (http://www.es.embnet.org/Doc/Training/)
- Algorithms in Molecular Biology (University of Washington) (http://www.cs.washington.edu/education/cources/590bi/98w/)
- Protein Sequence Analysis in the Genomic Era (http://lipid.biocomp.unibo.it/school/)
- Protein sequence and structure analysis : A practical guide. (http://www.biochem.ucl.ac.uk/bsm/dbbrowser/jj/)
- Topics of Evolutionary Computation (http://www.evalife.dk/index.php?lefturl=/eacourse2000/ topicsofEC2000.php)
- VSNS BioComputing Division (http://www.techfak.uni-bielefeld.de/bcd/welcome.html)
- Bioinformatics (http://twod.med.harvard.edu/seqanal/ index.html) - Primer on bio-sequence comparisons
- Algorithms in Molecular Biolgy (http://www.math. tac.ac.il/~rshamir/algmb/algmb98.html) - Excellent for learning basics about many bioinfo tools
- Biocomputing (http://www.hgmp.mrc.ac.uk/Embnetut/ Universi/embnettu.html) - Bio-computing tutorial at EBI
- Bioinforamtics Training Resources (http://www.med.nyu.edu/customerror/404) - Links to an excellent selection of bioinformatics tools training at NYU
- DNA composition and Exon prediction (http://www. pdg.cnb.uam.es/cursos/FVi2001/GenomAna/ GeneIdentification/SearchContent/main.html) - Sequence based measures indicative of protein-coding function in genomic DNA
- BCD BioComputing Tutorial (http://www.techfak.uni-bielefeld.de/bcd/Curric/welcome.html

More Bioinformatics Tutorials

- Bioinformatics (http://beckerinfo.net/bioinformatics/category/bioinformatics/)
- BLAST (http://beckerinfo.net/bioinformatics/category/blast/)
- database (http://beckerinfo.net/bioinformatics/category/database/)
- EBI (http://beckerinfo.net/bioinformatics/category/ebi/)
- Education (http://beckerinfo.net/bioinformatics/category/education/)
- Ensembl (http://beckerinfo.net/bioinformatics/category/ensembl/)
- genome (http://beckerinfo.net/bioinformatics/category/genome/)
- genome browser (http://beckerinfo.net/bioinformatics/category/ genome-browser/)
- helpful (http://beckerinfo.net/bioinformatics/category/helpful/)
- informatics (http://beckerinfo.net/bioinformatics/category/ informatics/)
- Information (http://beckerinfo.net/bioinformatics/category/ information/)
- NCBI (http://beckerinfo.net/bioinformatics/category/ncbi/)
- resources (http://beckerinfo.net/bioinformatics/category/resources/)
- sequence (http://beckerinfo.net/bioinformatics/category/sequence/)
- tool (http://beckerinfo.net/bioinformatics/category/tool/)
- tutorial (http://beckerinfo.net/bioinformatics/category/tutorial/)
- webinar (http://beckerinfo.net/bioinformatics/category/webinar/)

Virtual Online Tutorials

- Virtual Institute of Bioinformatics (http://www.bioinf.org/ vibe/index.html) - **National University of Ireland**
- UNIX, GCG, SEQLAB and STADEN Tutorials (http://www.molbiol.ox.ac.uk/tutorials.shtml) - Oxford University, UK
- BIOTOOLS96 (http://www.vsms.nottingham.ac.uk/vsms/biotools/index.html) - (University of Nottingham, UK, Virtual school of molecular sciences
- the principles of protein structure, using the internet (http://www.cryst. bbk.ac.uk/PPS2/top.html) - Birkback College (University of London), UK
- Free online bioinformatics courses! s-star.org (http://s-star.org/main.htm)
- Science and technology directory (http://www.technolgy-resource.co.uk/)
- Weizmann Institute of Science Genome and Bioinformatics (http://www.bioinfo.wizmann.ac.il/bioinfo.html)
- Algorithms for Molecular Biology (http://www.cs.tau. ac.il/~shamir/algmb.html) - Bioinformatics course notes, Tel Aviv University (TAU, Israel)
- Certificate Program in Bioinformatics (http://scpd. stanford.edu/home) - Stanford
- Courses Offered by BU Bioinformatics Program (http://engpub1.bu.edu/bioinfo/course.html)
- ISCB Training information (http://www.iscb.org/training.html)
- Penn Database Research Group- Classes (http://www.db.cis.upenn.edu/ Classes/)
- VSNS Biocomputing Division (http://www.techfak.uni-bielefeld.de/bcd/ welcome.html)
- Yale Bioinformatics -- Courses and Lectures (http://bioinfo.mbb.yale.edu/ lectures/)
- Bioinformtics Online lecture (I) (http://lectures.molgen.mpg.de/)
- Bioinformtics Online lecture (II) (http://www. genzentrum.lmu.de/) or (http://www.Imb.uni-muenchen.de/groups/bioinformatics/bioifo.html)
- MRes Biomolecular Sciences Lecture Notes: 1. The Gene and Bioinformatics (http://www.hgmp.mrc.ac.uk/ ~dcounsel/MRes/MRes.html)
- MRes Biomolecular Sciences Lecture Notes: 2. The Gene and Bioinformatics
- biocomputing on internet (http://www.techfak.uni-bielefeld.de/bcd/ welcome.html) (University of) Bielefeld, Germany Virtual School of Natural Sciences
- Sequence comparison (http://www.dir.univ-rouen.fr/~charras/seqcomp/) Universite de Rouen, France
- A Guide to Molecular Sequence Analysis (http://www.sequenceanalysis.com/) National Hospital University of Oslo, Norway
- Distant homologies: motifs, patterns, profiles (http://www.icgeb.trieste.it/net/courseware/Tiotle.htm) International Centre for Genetic Engineering and Biotechnology , Trieste, Italy
- Virtual School of Natural Sciences BioComputing Division (http://merlin. mbcr.bcn.tme.edu:8001/bcdusa/ welcome.html) - Virtual bio-computing course
- Algorithms for Computational Biology (Advanced Topics #6, 236606) (http://bioinfo.cs.technion.ac.il/) - Israel Institute of Technology
- CSE 590BI (http://www.cs.washington.edu/education/ courses/590bi/) - Computational Biology, University of Washington
- MBB 447b3 (747b3) Classes (http://bioinfo.mbb.yale. edu/course/classes/) - Yale
- UCSC School of Engineering- Class Home Pages (https://courses.soe.ucsc.edu/classes/) - University of California at Santa Cruz
- Virtual Bioinformatics Distance Learning (http://protein.uta.fi/bioinfo_courses) - Bioinformatics and Functional genomics courses offered by IMC Bioinformatics, University of Tampere
- Tutorials using NCBI Bioinformtics Tools (http://www.angelfire.com/ga2/nestsite2/bioinform.html)

Free online training from GeneGo

- (http://beckerinfo.net/bioinformatics/free-online-training-from-genego-july-dates/): Knowledge mining GeneGo content, EZ Search and MetaSearch Tired of spending hours searching the public domain building your research objectives? GeneGo now provides a new Google-like interface to search your favorite gene, protein, disease or compound with just one click. In this you will find details about expression (http://beckerinfo.net/bioinformatics/category/ expression/), genomics (http://beckerinfo.net/bioinformatics/category/genomics/), Information (http://beckerinfo.net/bioinformatics/category/information/), tutorial (http://beckerinfo.net/bioinformaticscategory/tutorial/), webinar (http://beckerinfo.net/ bioinformatics/category/webinar/).
- Train online with EMBL-EBI(http://www.ebi.ac.uk/training/online/)

Online Courses:

The educational institutions listed below have submitted information on their bioinformatics related online courses

Focus	Course Title	University/Institution
Bioinformatics	Bioinformatics for transcriptomics	The University of Manchester
Computational Biology	Bioinformatics for Systems Biology	University of Manchester
Math/Statistics	Mathematics for metabolic modelling	University of Manchester
Math/Statistics	Statistics and R for the Life Sciences	HarvardX
Bioinformatics	Advanced sequence analysis	The University of Manchester
Bioinformatics	Network Analysis in Systems Biology	Mount Sinai
Bioinformatics	Bioinformatics Algorithms (Part 1)	University of California, San Diego
Bioinformatics	Perl and Unix for Bioinformatics	PerlSource Informatics
Bioinformatics	Perl for Biologists, Level 1	Bioinformatics.Org

www.openhelix.com

- **Washington University, USA**

(http://www.openhelix.com/cgi/createAccount.cgi)

List of all Open Helix Tutorials: This is the full list of tutorials offered by OpenHelix. click on the category name or display all tutorials grouped by category (http://www.openhelix.com/cgi/tutorialCategory.cgi)

BIOINFORMATICS, INTERNET AND LINKS

- 1. (www.cats.ucsc.edu)
- 2. (www.cato.com/biotech/)
- 3. (www.the-scientist.com)
- 4. (www.nbif.org/links/1.20.php)
- 5. (www.genomics.phrma.org/today/)
- 6. (www.gnn.tigr.org/main.shtml)
- 7. (www.ebi.ac.uk/ismb-97/papers2.html)
- 8. (www.cgl.ucsf.edu/psb/)
- 9. (www.Biosis.org.zrdocs/zoolinfo/biol_inf.htm)
- 10. (www.ars-genome.cornell.edu/)
- 11. (www.molbio.info.nih.gov/molbio/)
- 12. (www.golgi.harvard.edu/Biolinks.html)
- 13. (www.molbio.info.nih.gov/cgi-bin/pdb)
- 14. (www.oup.co.uk/bioinformatics/contents/)
- 15. (www.cgb.indiana.edu/bioinformatics/resources)
- 16. (www.genome.ucsc.edu/)
- 17. (www.ebi.ac.uk/index.html)
- 18. (www.atcc.org)
- 19. (www.bio.org/welcome.html)
- 20. (www.biotech-register.com/)
- 21. (www.nbif.org)
- 22. (www.proteome.com/services/index.html)
- 23. (www.rcsb.org/pdb/)
- 24. (www.sciewb.com)
- 25. (www.bioinfo.com/fbdhome.html)
- 26. (www.bioinformatics.org)
- 27. (www.yahoo.com)
- 28. (www.cbt.org)
- 29. (www.unipune.ernet.in)
- 30. (www.mcrcr0.med.nyu.edu/rcr/molbio/syllabus-98.html)
- 31. (www.ncbi.nlm.nih.gov/Entrez/index.html)
- 32. (www.ncbi.lm.nih.gov/BLAST)

Internet educational resources for Bioinformatics

- **NCBI:** sequence data repository, US **Bioinformatics** center. (http://www.ncbi.nlm.nih.gov/)
- **EBI:** sequence data repository, European **Bioinformatics** center. (http://www.ebi.ac.uk/)
- **Pasteur:** France bioinfo. center, Bio Netbook is an excellent database of Internet information for biosciences, **Bioinformatics**. (http://bioweb.pasteur.fr/intro-uk.html) Bio Netbook
- **ExPASy/SWISSPROT:** protein sequence data center. (http://www.expasy.ch/) good list of bioinfo resources
- **Sanger:** European sequencing, **Bioinformatics** center. (http://www.sanger.ac.uk/)

- **Weizmann:** Israel **Bioinformatics** center. (http://bioinformatics.weizmann.ac.il/)
- **GenomeWeb: Bioinformatics** resources. (http://www.hgmp.mrc.ac.uk/ GenomeWeb/)
- **CSHL:** US sequencing, **Bioinformatics** center. (http://www.cshl.org/)
- **WUSTL:** US sequencing, **Bioinformatics** center. (http://www.ibc.wustl.edu/)
- **Stanford genome center** US sequencing, **Bioinformatics** center. (http://genome-www.stanford.edu/)
- **TIGR:** US sequencing, **Bioinformatics** center. (http://www.tigr.org/)
- **Celera:** US commercial sequencing, **Bioinformatics** center. (http://www.celera.com/)
- **GenomeNet** : Japan **Bioinformatics** center (http://www.genome.ad.jp/)
- **Bionet:** Usenet network news for biology. (http://www.bio.net/)
- **BioMedNet: Bioinformatics** resources including HMS Beagle, (http://www.bmn.com/)
- **BioInform** mostly commercial news, services - good list of companies in **Bioinformatics**. (http://www.bioinform.com/)

Sequence Analysis Tools

- HOME (http://phobos.ramapo.edu/~pbagga/index.html)
- Comparison & Alignments (http://phobos.ramapo.edu/~pbagga/binf/binf_res/ bioinfo_sat_comp.htm)
- Functional Analysis (http://phobos.ramapo.edu/~pbagga/binf/binf_res/bioinfo_sat_func.htm)
- Structural Analysis (http://phobos.ramapo.edu/~pbagga/binf/binf_res/ bioinfo_sat_struc.htm)
- Other Analytical Tools (http://phobos.ramapo.edu/~pbagga/binf/binf_res/ bioinfo_sat_other.htm)

Important Tutorial Videos

- Bioinformatics Tutorials & Articles are here. Tutorials classified laying foundation course for both science and computer students for those aspiring bioinformatics. List also includes Bioinformatics Tutorials & Articles related to various tools.
- **(http://digitalworldbiology.com/dwb/bioinformatics-tutorials)**
- (http://www.youtube.com/watch?v=UohaqFb8_ME)
- (http://nihlibrary.nih.gov/Services/Bioinformatics/Pages/Biotutorials.aspx)
- (https://www.countway.harvard.edu/menuNavigation/libraryServices/classes/ videoTutorials.html)
- (http://www.youtube.com/user/bimaticsblog)
- (http://bioinformaticsonline.com/bookmarks/view/3868/next-generation-sequencing-ngs-tutorials)
- (http://mybio.wikia.com/wiki/Tutorials_in_bioinformatics)
- (http://lectures.molgen.mpg.de/online_lectures.html)
- (http://www.bioperl.org/wiki/HOWTO:Beginners#Abstract)
- (http://www.ccg.unam.mx/~vinuesa/Bioinformatics_resources_web.html)
- **Video Tutorials**
- Imp: (http://digitalworldbiology.com/dwb/bioinformatics-tutorials)
- (http://www.youtube.com/watch?v=UohaqFb8_ME)
- (http://nihlibrary.nih.gov/Services/Bioinformatics/Pages/Biotutorials.aspx)
- (https://www.countway.harvard.edu/menuNavigation/libraryServices/classes/videoTutorials.html)
- (http://www.youtube.com/user/bimaticsblog)
- (http://bioinformaticsonline.com/bookmarks/view/3868/next-generation-sequencing-ngs-tutorials
- (http://mybio.wikia.com/wiki/Tutorials_in_bioinformatics)
- (http://lectures.molgen.mpg.de/online_lectures.html)
- (http://www.bioperl.org/wiki/HOWTO:Beginners#Abstract)
- (http://www.ccg.unam.mx/~vinuesa/Bioinformatics_resources_web.html)

UTUBE VIDEOs

Tutorial videos of bioinformatics resources: online distribution trial in Japan named Togo TV

- Togo TV (http://togotv.dbcls.jp/en/)

BIOINFORMATICS BLOGS

Site	Rating	Description
123 bioinformatics (http://forum.123bioinformatics.com/)	*	A blog site with over 1167 biotechnology help topics but these often relate to non-scientific issues.
All-wow videos- Mass Spectrometry (http://www.wowtube.ru/ index.php?key=Spectrometry)	**	Over 50 short videos related to mass spectrometry. The Wow sites also provide instructional videos on a wide range of other topics.
American Biotechnologist (http://www. americanbiotechnologist.com/blog/)	**	This blog from Bio-Rad (Hercules, CA) has been created as a place where PIs, Graduate Students, Technicians and Science Educators can network, post and view articles, videos, seminars, techniques etc of interest and generally find subject matter relevant to them.
Aetiology (http://scienceblogs.com/aetiology/)	**	A personal blog site from Dr. Tara Smith, who is an assistant professor of Epidemiology in Iowa, which focuses on the causes, origins, evolution and implication of disease and other phenomena.
Bad Science (http://www.badscience.net/)	**	A personal blog site from Dr. Ben Goldacre, who is a medical doctor, writer and broadcaster. Bad Science covers a wide range of health and science topics.
Becker (http://beaker.sanfordburnham.org/)	**	A blog site produced by Sanford-Burnham Medical Research Institute (La Jolla, CA) that

NIH Video casts

(http://beckerinfo.net/bioinformatics/nih-videocasts)

- National Library of Medicine update(http://beckerinfo.net/bioinformatics/nlmnews/
- bioinformatics (http://beckerinfo.net/bioinformatics/category/bioinformatics/) database (http://beckerinfo.net/bioinformatics/category/ database/), event (http://beckerinfo.net/bioinformatics/category/event/), Information (http://beckerinfo.net/bioinformatics/category/information/), NCBI (http://beckerinfo.net/bioinformatics/category/ncbi/), NLM (http://beckerinfo.net/bioinformatics/category/nlm/), resources (http://beckerinfo.net/bioinformatics/category/resources/), science (http://beckerinfo.net/bioinformatics/category/science/), tutorial (http://beckerinfo.net/bioinformatics/category/tutorial/), webinar

  (http://beckerinfo.net/bioinformatics/category/webinar/).

- Genetic Alliance webinars and resources

(http://beckerinfo.net/ bioinformatics/genetic-alliance/
)

- NCBI Short Read Archive of NexGen Sequencing Data
- (http://jeansong.wordpress.com/2009/08/19/ncbi-short-read-archive-of-nexgen-sequencing-data/): They are now maintaining the Short Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/Traces/sra/)

- Search and display SRA project data through Entrez (http://www.ncbi.nlm.nih.gov/sra?term=all%5Bsb%5D)
- Download data through Aspera Connect (http://www.aspera.com/en/software-license-management/)
- BLAST service (http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&BLAST_PROGRAMS=megaBlast&PAGE_TYPE=BlastSearch&BLAST_SPEC=SRA) for sequence similarity searching of 454 sequencing reads for transcriptome studies
- NCBI 3D Structure Help
- (https://jeansong.wordpress.com/2009/08/ 05/ncbi-3d-structure-help/) : So the folks at NCBI have posted a whole bunch of useful help files on the web for their 3D structure resources, 3D Macromolecular Structure (http://www.ncbi.nlm.nih.gov/Structure/MMDB/docs/mmdb_how_to.html) and Conserved Domains (http://www.ncbi.nlm.nih.gov/Structure/cdd/docs/cdd_how_to.html).

BIOINFORMATICS IN SOCIAL MEDIA

- (http://www.webicina.com/bioinformatics/news-and-information-on-bioinformatics/#package_container)

News

- (http://www.webicina.com/ bioinformatics/news-and-information-on-bioinformatics# package container)
- blog (http://www.webicina.com/bioinformatics/ bioinformatics-in-the-blogosphere#package_container)
- podcast (http://www.webicina.com/bioinformatics/ bioinformatics-podcasts-and-interviews# package container)
- community (http://www.webicina.com/bioinformatics/ bioinformatics-community-sites-facebook-groups-and-forums#package_container)
- twitter (http://www.webicina.com/bioinformatics/ microblogging-twitter-and-friendfeed#package_container)
- wiki (http://www.webicina.com/bioinformatics/ bioinformatics-wikis#package_container)
- video (http://www.webicina.com/bioinformatics/ bioinformatics-videos-animations-and-videocasts#package_container)
- slideshow (http://www.webicina.com/bioinformatics/ slideshows-about-bioinformatics#package_container

Thank you