Problem Solving in Bioinformatics

Jianlin Cheng, PhD

Informatics Institute, Computer Science Department

University of Missouri, Columbia

Fall, 2011

Objectives

- Walk students through the complete process of sequencing, assembling and annotating a genome. During the process, students lean key bioinformatics techniques for analyzing a genome and its components (i.e. gene, RNA, protein, and pathway).
- By working on a comprehensive genome annotation project, students develop practical skills to apply bioinformatics methods to solve major problems in genome assembly and annotation.

Instructors

- Jianlin Cheng, PhD (coordinator)
- Dmitry Korkin, PhD
- Chi-Ren Shyu, PhD
- Dong Xu, PhD

Topics

- Introduction to the course and a project (Jianlin Cheng)
- Genome sequencing and assembly (Dong Xu)
- Gene prediction (Dong Xu)
- Protein structure prediction (Jianlin Cheng)
- Protein function prediction (Jianlin Cheng)
- Protein interaction prediction (Dmitry Korkin)
- Biological pathways and networks (Chi-Ren Shyu)

Course Format

- **Theory phase** (one month): lecturing, literature review, mid-term presentation
- **Practice phase** (two and a half months): discussion, planning (group), presentation, programming (group), results (group), assessment (group), and report (group)
- Teamwork & leadership (two groups)
- See syllabus for details

Assignments

- Literature review, topic plan (in presentation style), implementations of genome assembly and annotations (programs and results), topic report, and final report and presentation.
- All the assignments should be posted to the project web site or emailed to me by deadlines. (chengji@missouri.edu)

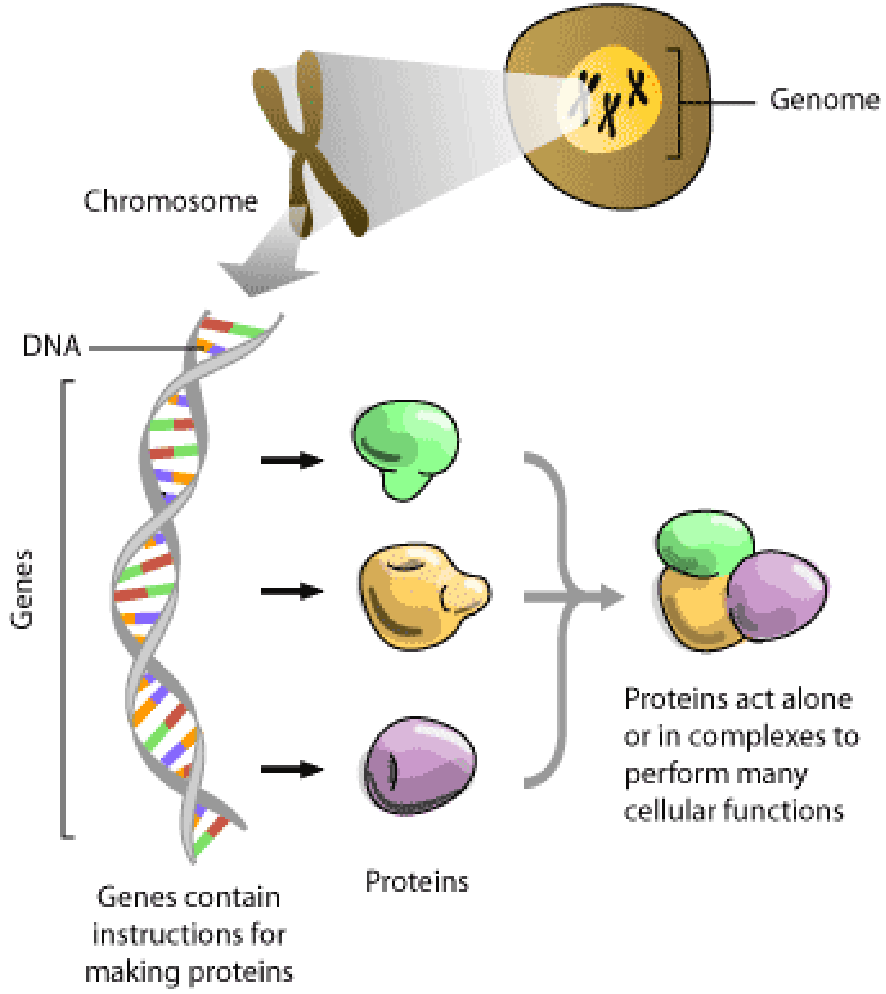
Evaluation and Grading

- literature reviews (individual, 10%), mid-term presentation (individual, 10%), class discussion (individual, 15%), topic presentations (group, 15%), topic plans and reports (i.e. progress and assessment) (group, 15%), topic implementation (group, 20%), a final presentation and report (group, 15%)
- Group components may be graded by both instructors and group peers

Course Web & Class Schedule

- Course web (**demo**): http://www.cs.missouri.edu/~chengji/infoinst8010/
- Class schedule and assignments: http://www.cs.missouri.edu/~chengji/infoinst8010/8010_schedule.htm

Introduction

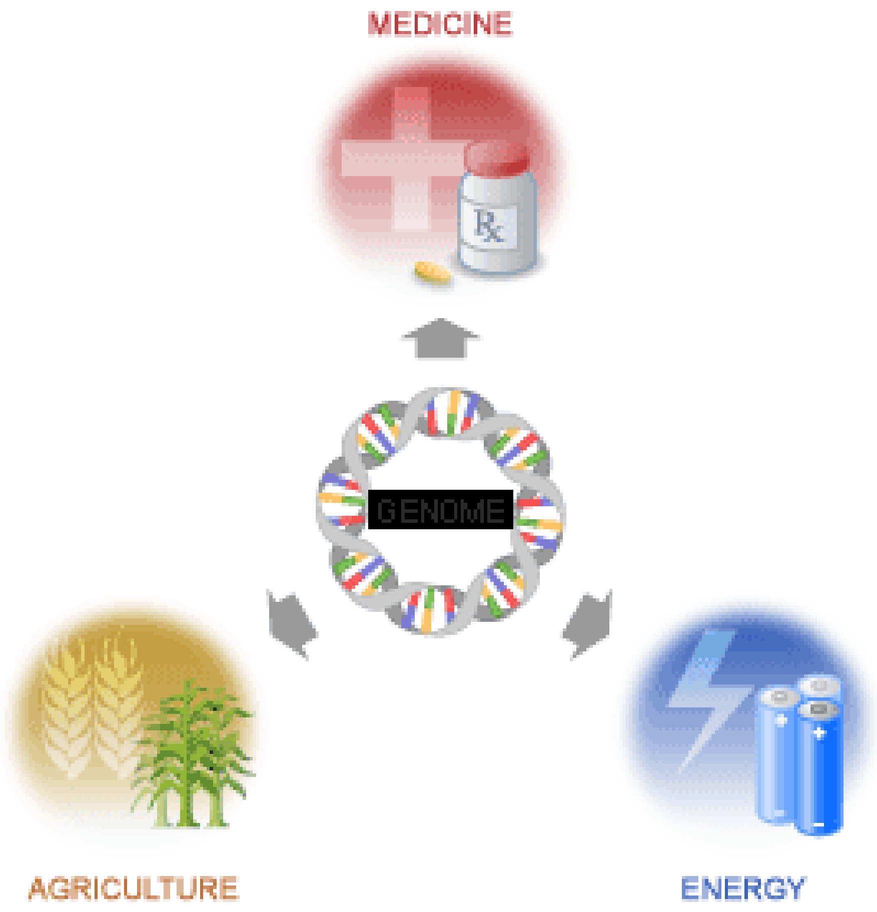

Genome
Chromosome
DNA
Genes
Proteins act alone or in complexes to perform many cellular functions
Proteins
Genes contain instructions for making proteins

- **Grow**
- **Sustain**
- **Adapt**
- **Reproduce**

- **Genome & Components**
- **Environment**

http://www.scq.ubc.ca/wp-content/uploads/2006/08/molecular-machine.gif

Applications of Genome Knowledge



Genome Sequencing – Cracking the Code

- Virus & Bacteria genomes (small)
- Human genome

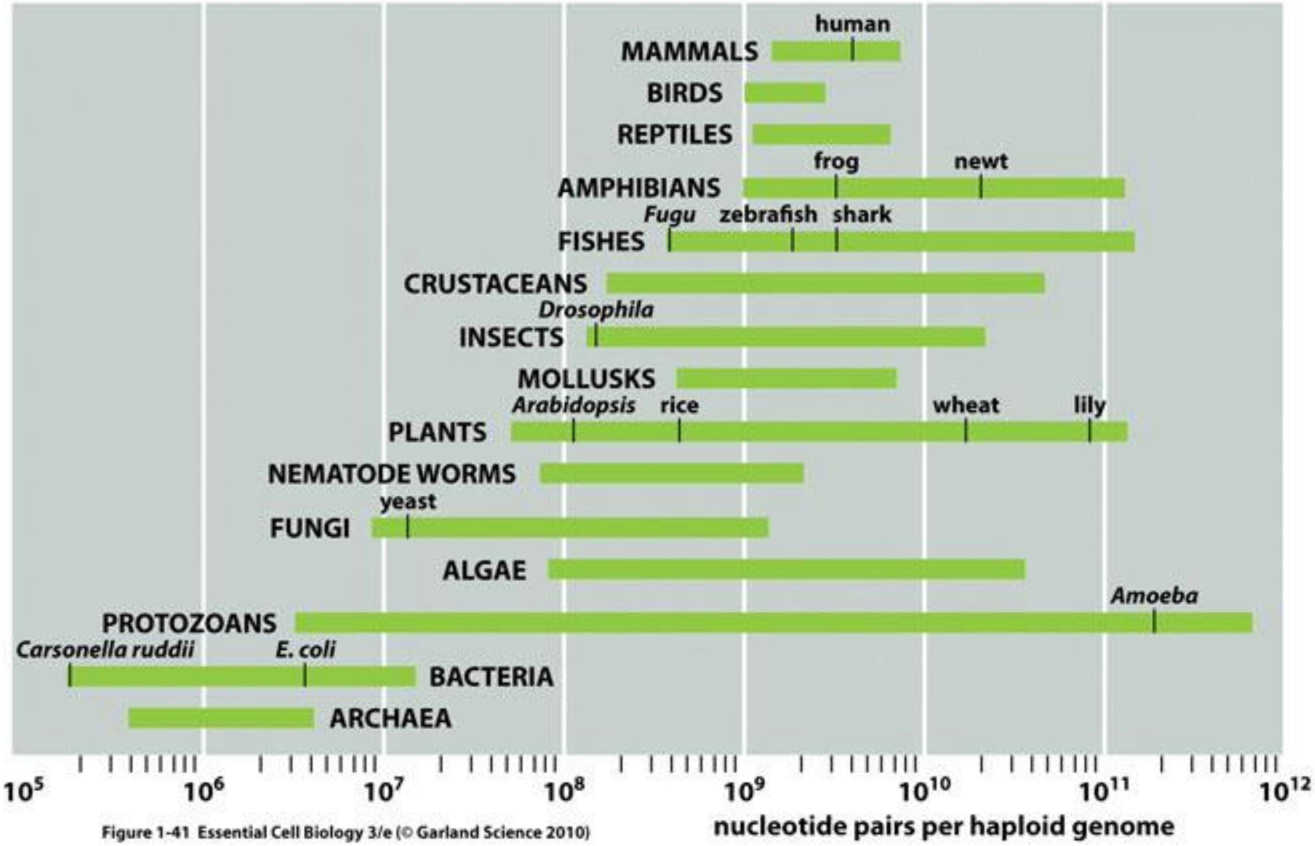
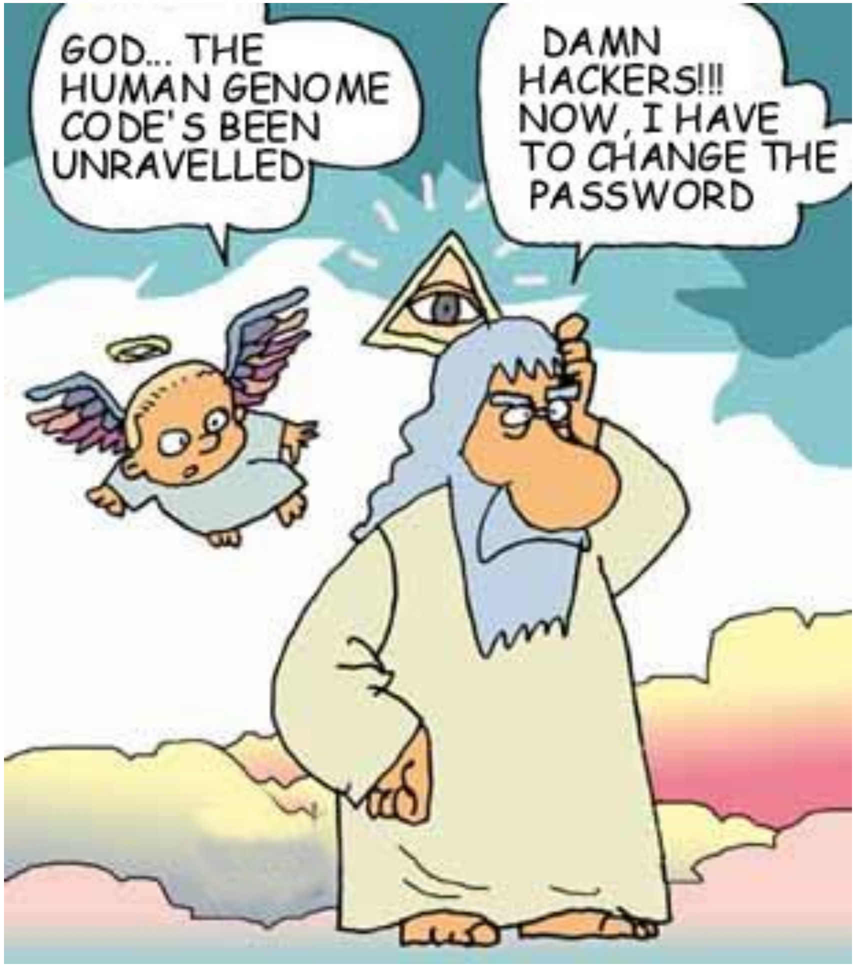


Figure 1-41 Essential Cell Biology 3/e (© Garland Science 2010)

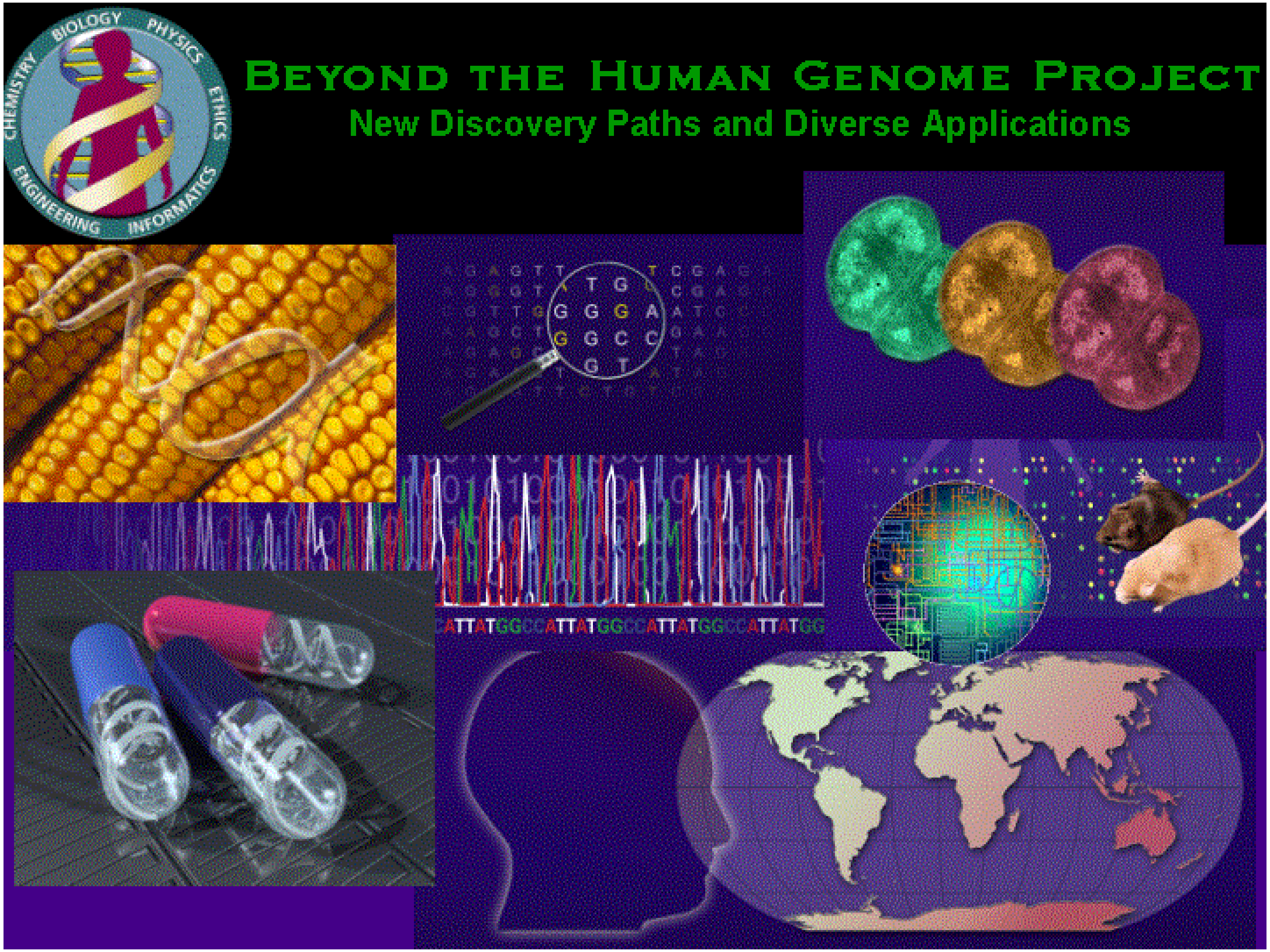
Human Genome Project

The National Human Genome Research Institute
Macintosh / Windows
CD ROM (V. 2.0)
Understanding The Human Genome Project

Fun


GOD... THE HUMAN GENOME CODE'S BEEN UNRAVELLED
DAMN HACKERS!!! NOW, I HAVE TO CHANGE THE PASSWORD

Genome Sequencing Routine

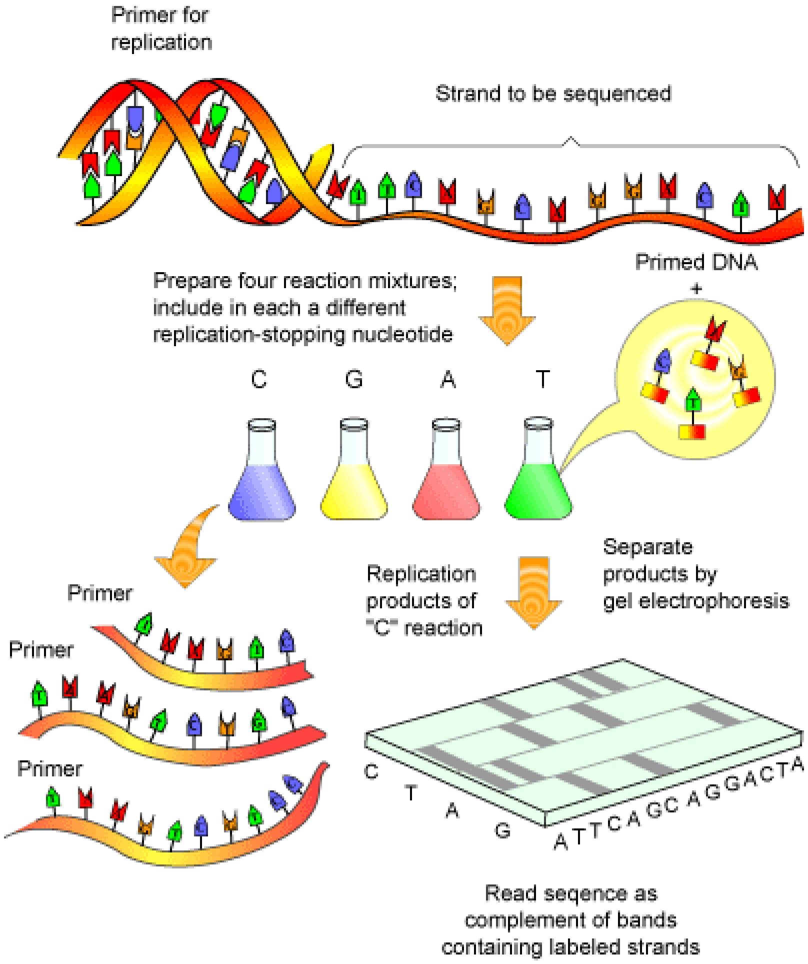

BIOLOGY
PHYSICS
ETHICS
INFORMATICS
ENGINEERING
CHEMISTRY
BEYOND THE HUMAN GENOME PROJECT
New Discovery Paths and Diverse Applications
CATTATGGCCATTATGGCCATTATGGCCATTATGG

MU Genome Sequencing

- Soybean genome: Gary Stacey, Dong Xu, Jay Thelen, Jianlin Cheng, Henry Nguyen, etc (Nature, 2010)
- Chris Pires – plant genomes

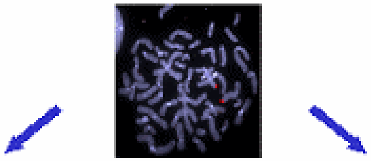
DOE
Bioenergy
Research
Centers

Sequencing process

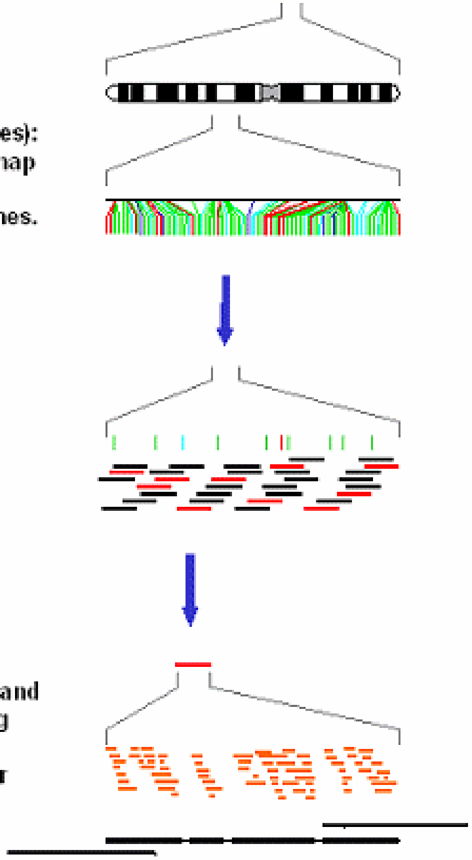

Primer for replication
Strand to be sequenced
Prepare four reaction mixtures; include in each a different replication-stopping nucleotide
Primed DNA +
C
G
A
T
Primer
Primer
Primer
Replication products of "C" reaction
Separate products by gel electrophoresis
C
T
A
G
A T T C A G C A G G A C T A
Read seqence as complement of bands containing labeled strands

Genome Sequencing Machine

STRATEGIES FOR SEQUENCING THE HUMAN GENOME



BY MAPPED CLONES



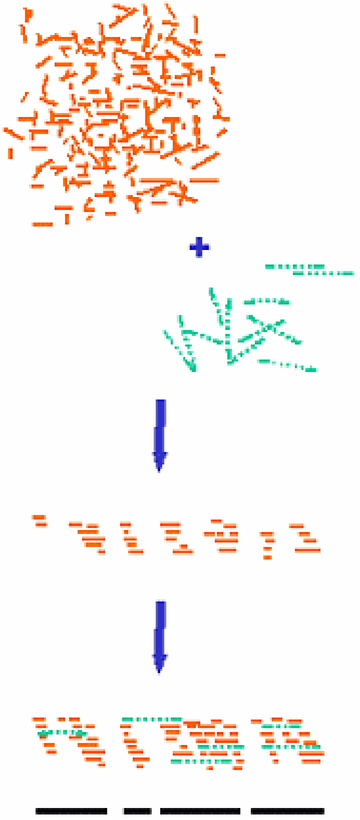
1. Construction of maps of ordered landmarks (genetic markers, genes): provides long-range map and organisation into individual chromosomes.

2. Physical maps of overlapping clones anchored to the landmark maps.

3. Selection of tile path (clones in red)

4. Shotgun sequencing and assembly (for working draft); subsequent directed finishing (for reference sequence).

BY WHOLE GENOME SHOTGUN



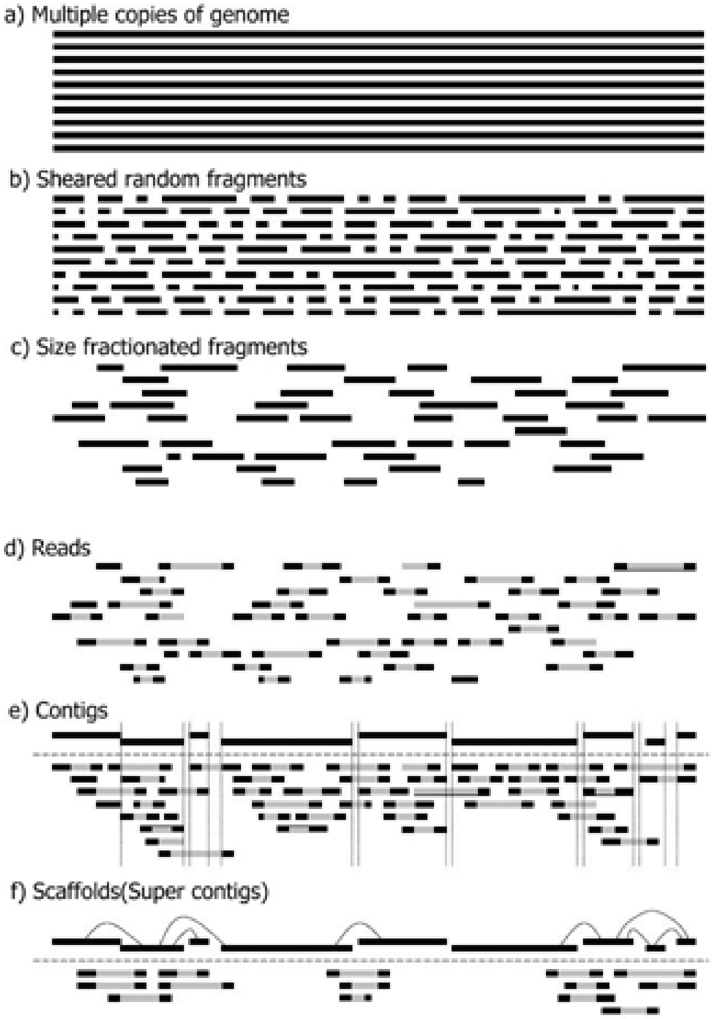
1. Shotgun sequencing of short-insert clones

2. Paired end sequencing of large-insert clones

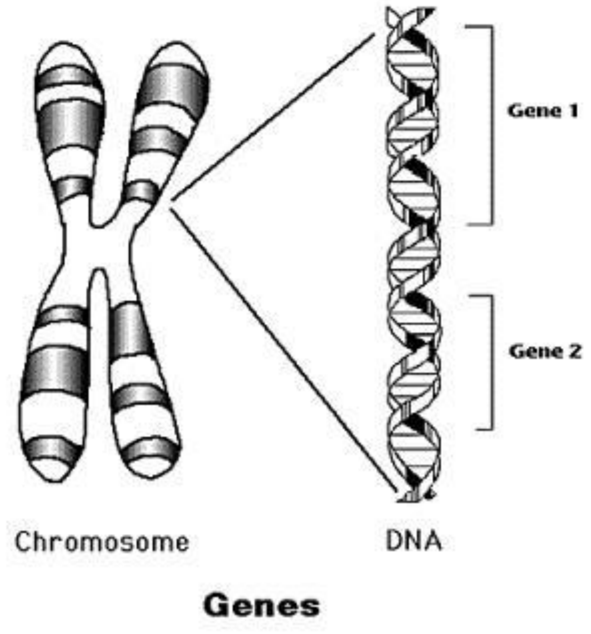
3. Assembly of seed contigs (unitigs)

4. Incorporation of other sequences, and integration of long- range data.

Topic 1: Genome Assembly


a) Multiple copies of genome
b) Sheared random fragments
c) Size fractionated fragments
d) Reads
e) Contigs
f) Scaffolds(Super contigs)

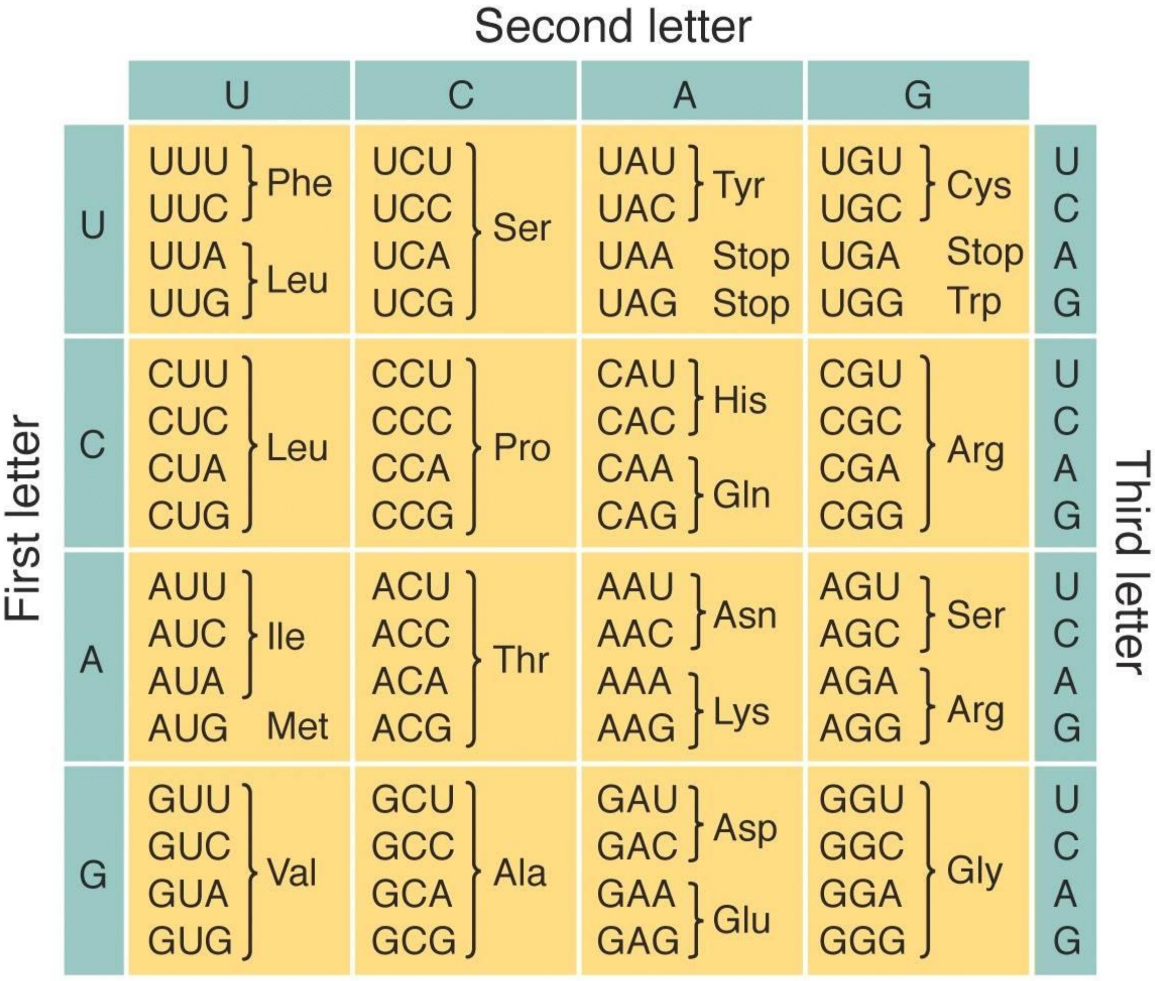
Topic 2. Gene Prediction



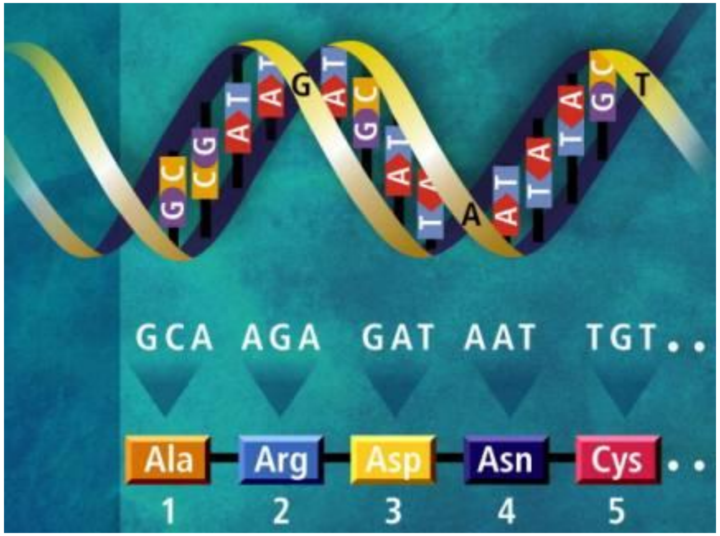
```
AGTGATTAGTGATTACAGCATCATTTTTTAAATTTAGGCATAAAACGCCCTTAAATCAAGGGTTTTTGAG
CGAGCTTTTTGCTCAAAGAATCCAAGATAGCGTTTAAAAATTTAGGGGTGTTAGGCTCAGCGTAGAGTTT
GCCAAGCTCTATGCATTCATTGATGATGATAGGGTTTTGTGTGGGCGTGAAGCCAATTTCATACGCTCCT
AAGCGTAAAATCGCCTTTTCCATGCTCCCTAATCGCTTGAAATCCCAGTCTTTTAAATGCGGTTCGATGA
GAGCGTCAATCTCATTGATTTTTTCTAACACGCCATTAAAAAGGCTTAAAGCGAAAGCGAGCTGGTTGTT
TTTAATCTTTTTTTCTTCTAACATGCTGGAAGCGATTTTTTTAATTTCTTCATTACCGCTCTCAAACGCA
TACAACAATTCAACCACAGCCCCCCTGGCTTGAGTTCGTGTCGCCATTTTAATCCTTGAGAGTTTGGCAC
AAGCTCAATAATTCAATGAGGGTGCTCATCGCTTCAAAGCCCTTATTACCGGCTTTACTGCCCGCTCTTT
CAATCGCTTGTTCAATATTGTCTGTGGTCAGCACGCCAAAGCTTACCGGCATGCTGTATTTTAGCATCGC
ACTAGCAATGCCCTTAGTCGCTTCCGCGCTCACGTAGTCAAAATGCGGAGTCCCCCCTCTAATGATCGCT
CCTAAAACGCACACGCCATCGTATTTTCCGCTCTCTAACAATTTGTCTAAAATCAAAGGCAATTCATAAG
CCCCAGGCACTAGCACGAGATCTAAAAGCTCCTCATCGCCCCCATGCCTTTTAAAGCAATCCATCGCCCC
TTCTTTTAATCTGTCTGTGATGATATGGTTGAAGCGCGATGTTAAAATAGCGATTTTTTCATTCCCTTGT
AATTGCAATTTCCCTTCTATGATTTGCATGAAATTCCTTTAAAATAAATTTTGGATTTTTAACATGTCGG
TTACTAAGTGTTCTAGCTCGTCAGGTTTTAGCATGTTTGCTCCATCGCTTAGGGCGTTTTTAGGATCAAT
ATGCGTTTCAGCGAACAGCCCATCAATCCCCACCGCCGCCGCCGCTCTGGGTAAAATAGGGGGAAAAGAG
```

Pattern Recognition Problem

Gene Product - Protein

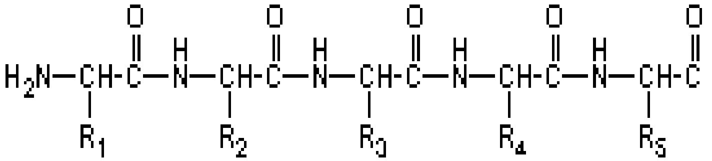


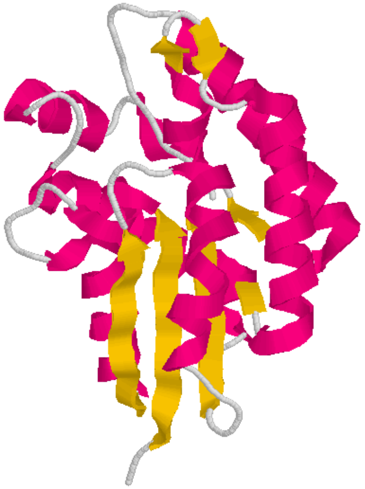
	Second letter								
First letter	U		C		A		G		Third letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

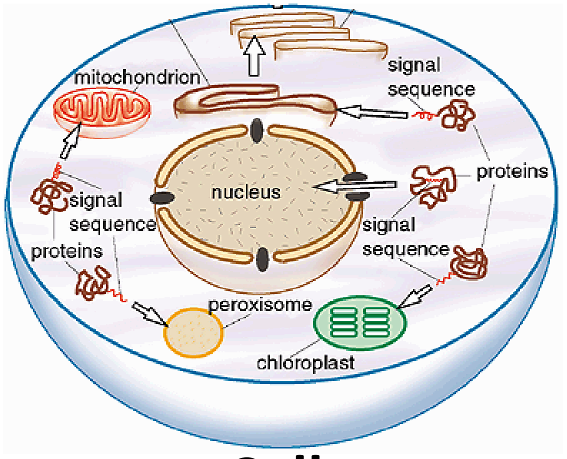


Protein Sequence, Structure, Function

AGCWY……

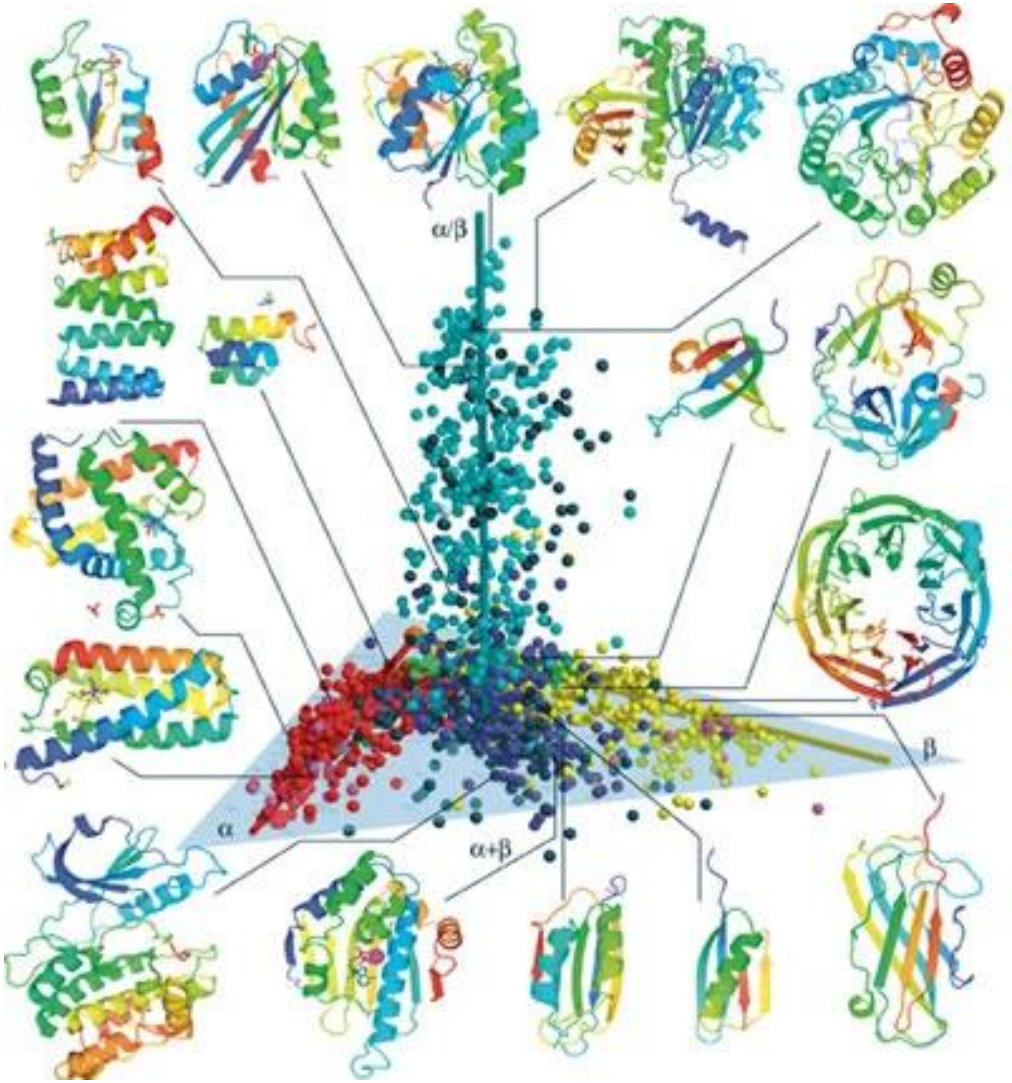




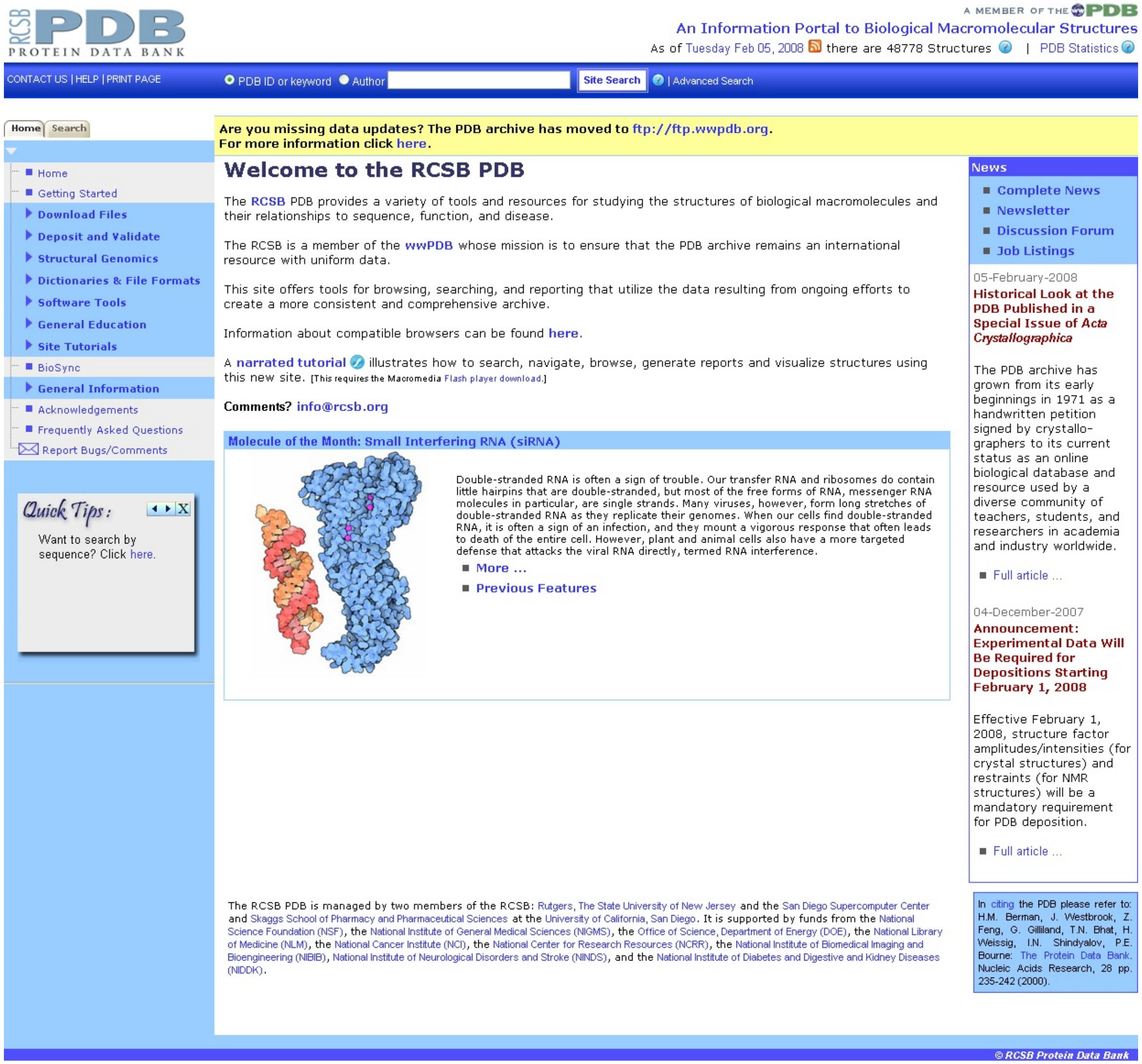


Cell

Protein Structure Space



Protein Data Bank



A MEMBER OF THE PDB

An Information Portal to Biological Macromolecular Structures

As of Tuesday Feb 05, 2008 there are 48778 Structures | PDB Statistics

CONTACT US | HELP | PRINT PAGE

● PDB ID or keyword ○ Author [ ] Site Search | Advanced Search

Home | Search

- Home
- Getting Started
- Download Files
- Deposit and Validate
- Structural Genomics
- Dictionaries & File Formats
- Software Tools
- General Education
- Site Tutorials
- BioSync
- General Information
- Acknowledgements
- Frequently Asked Questions
- Report Bugs/Comments

Quick Tips:

Want to search by sequence? Click here.

Are you missing data updates? The PDB archive has moved to ftp://ftp.wwpdb.org.
For more information click here.

Welcome to the RCSB PDB

The RCSB PDB provides a variety of tools and resources for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

The RCSB is a member of the wwPDB whose mission is to ensure that the PDB archive remains an international resource with uniform data.

This site offers tools for browsing, searching, and reporting that utilize the data resulting from ongoing efforts to create a more consistent and comprehensive archive.

Information about compatible browsers can be found here.

A narrated tutorial illustrates how to search, navigate, browse, generate reports and visualize structures using this new site. [This requires the Macromedia Flash player download.]

Comments? info@rcsb.org

Molecule of the Month: Small Interfering RNA (siRNA)

Double-stranded RNA is often a sign of trouble. Our transfer RNA and ribosomes do contain little hairpins that are double-stranded, but most of the free forms of RNA, messenger RNA molecules in particular, are single strands. Many viruses, however, form long stretches of double-stranded RNA as they replicate their genomes. When our cells find double-stranded RNA, it is often a sign of an infection, and they mount a vigorous response that often leads to death of the entire cell. However, plant and animal cells also have a more targeted defense that attacks the viral RNA directly, termed RNA interference.

- More ...
- Previous Features

The RCSB PDB is managed by two members of the RCSB: Rutgers, The State University of New Jersey and the San Diego Supercomputer Center and Skaggs School of Pharmacy and Pharmaceutical Sciences at the University of California, San Diego. It 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), National Institute of Neurological Disorders and Stroke (NINDS), and the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK).

News

- Complete News
- Newsletter
- Discussion Forum
- Job Listings

05-February-2008

Historical Look at the PDB Published in a Special Issue of *Acta Crystallographica*

The PDB archive has grown from its early beginnings in 1971 as a handwritten petition signed by crystallographers to its current status as an online biological database and resource used by a diverse community of teachers, students, and researchers in academia and industry worldwide.

- Full article ...

04-December-2007

Announcement: Experimental Data Will Be Required for Depositions Starting February 1, 2008

Effective February 1, 2008, structure factor amplitudes/intensities (for crystal structures) and restraints (for NMR structures) will be a mandatory requirement for PDB deposition.

- Full article ...

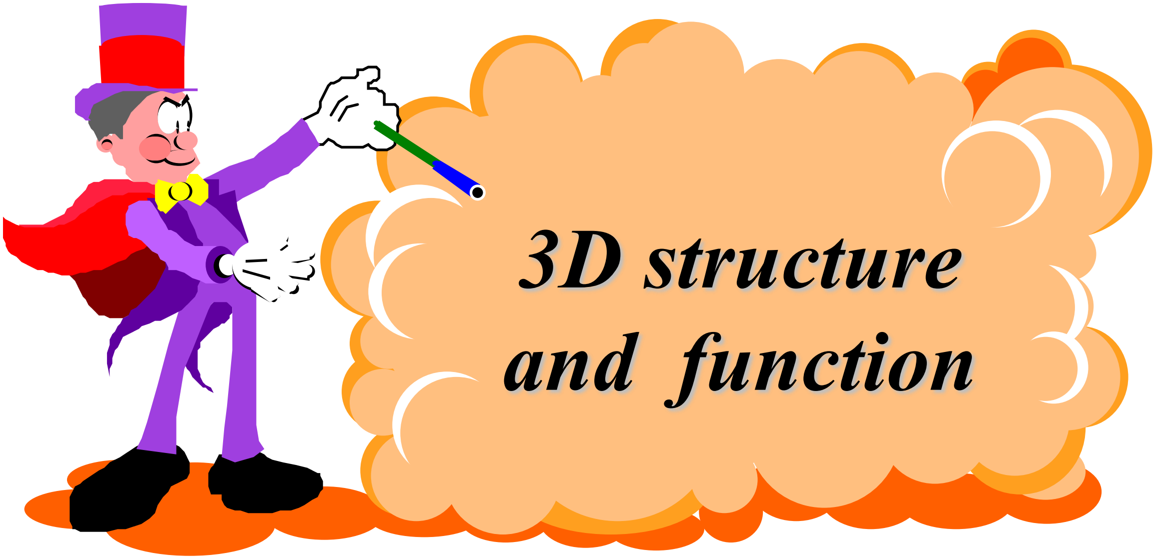
In citing the PDB please refer to: H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne: The Protein Data Bank. Nucleic Acids Research, 28 pp. 235-242 (2000).

© RCSB Protein Data Bank

Topic 3: Protein structure prediction

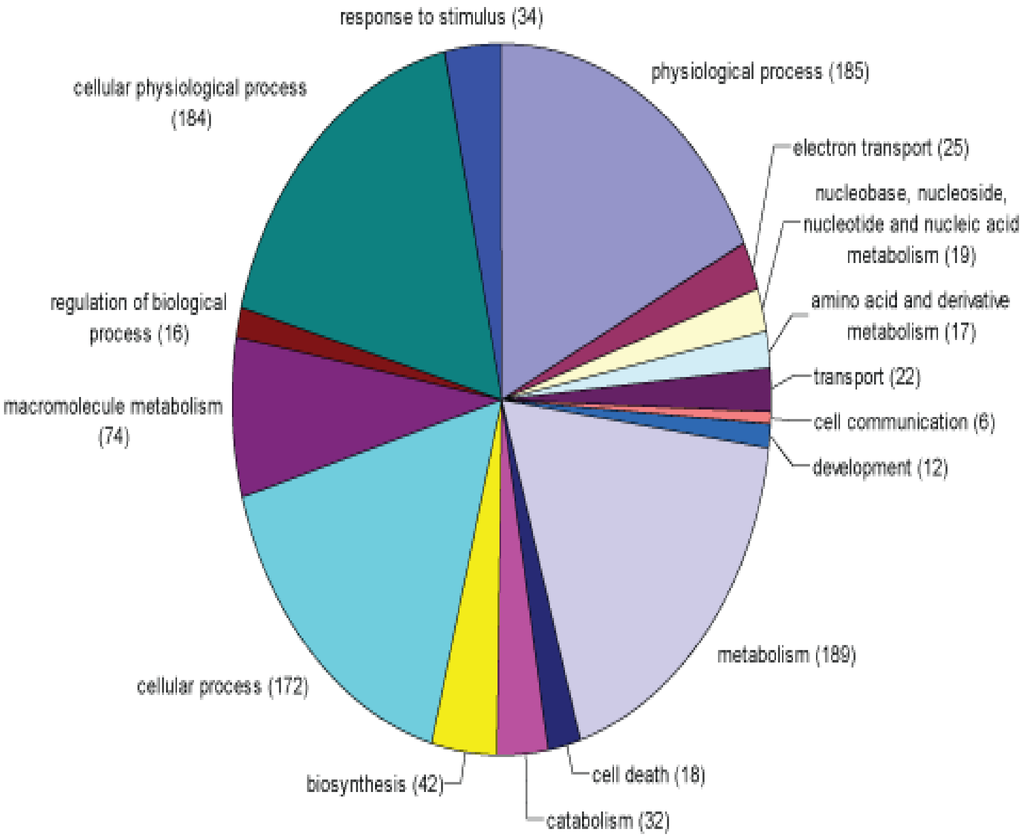
- Epstein & Anfinsen, 1961:
  sequence uniquely determines structure

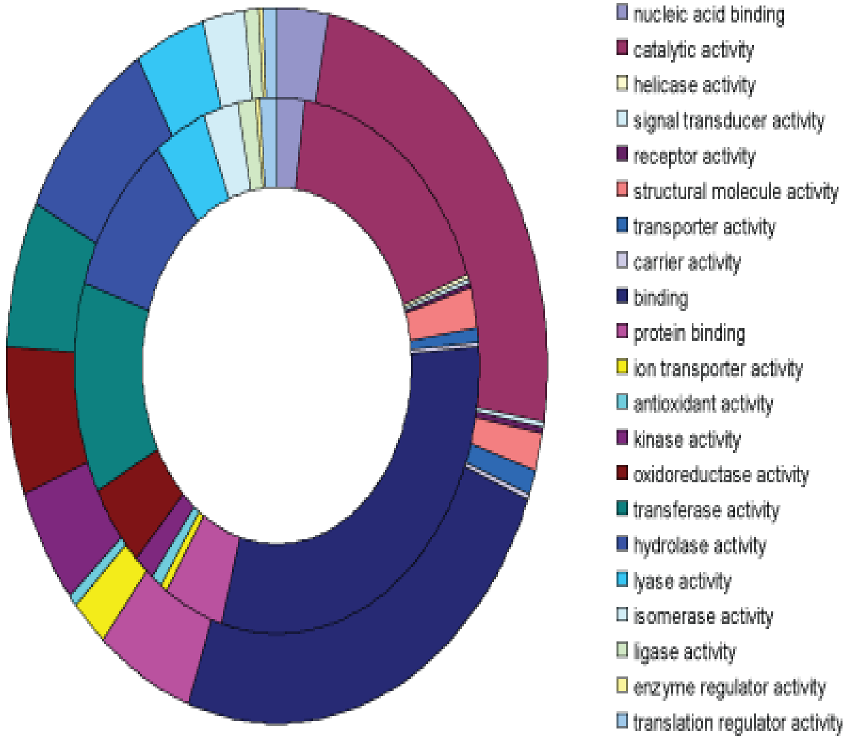
- INPUT: sequence
- OUTPUT:



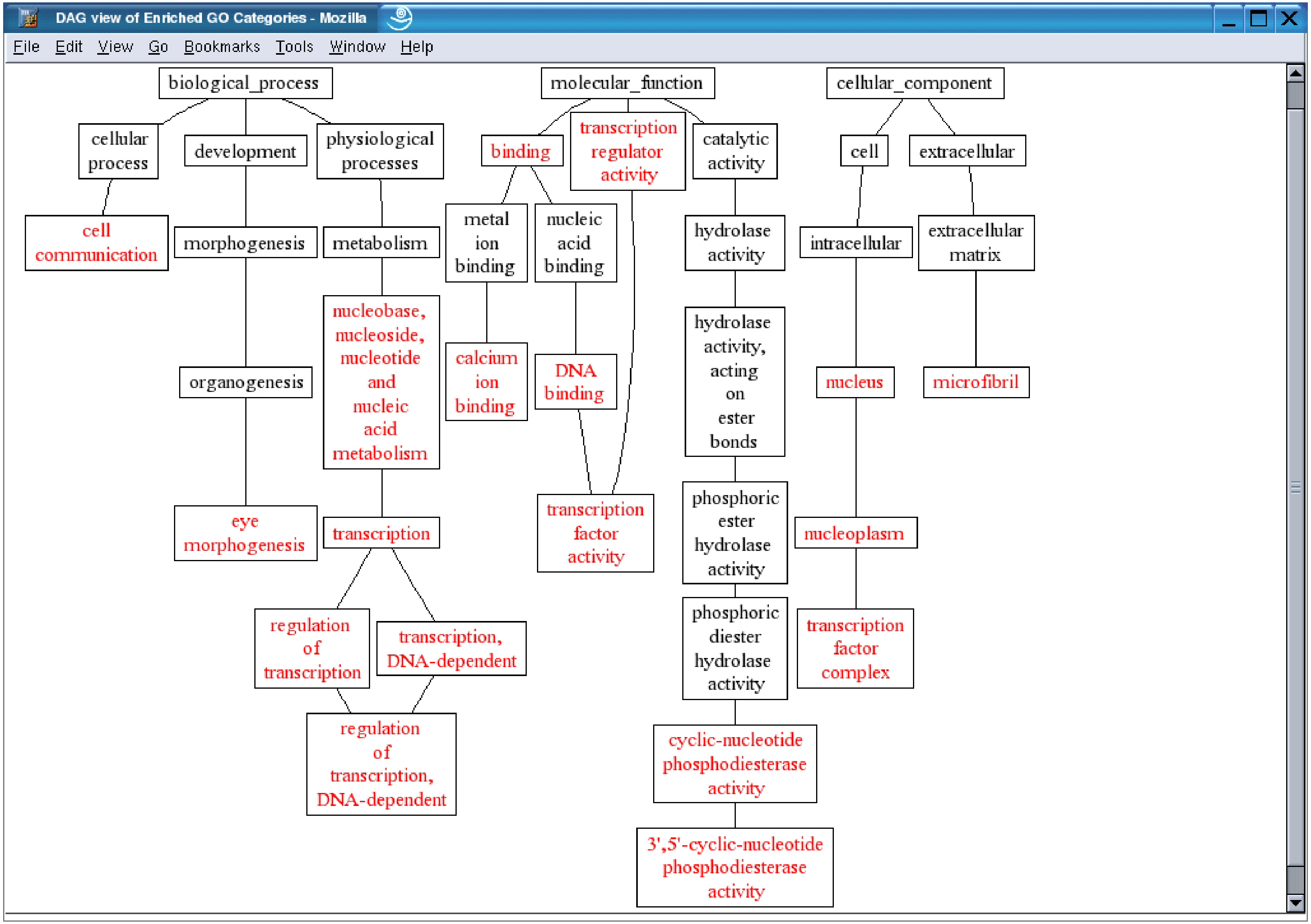
B. Rost, 2005

Topic 4: Protein Function Prediction

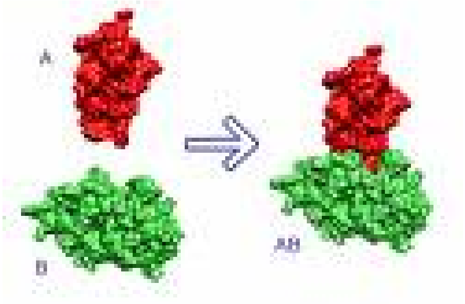


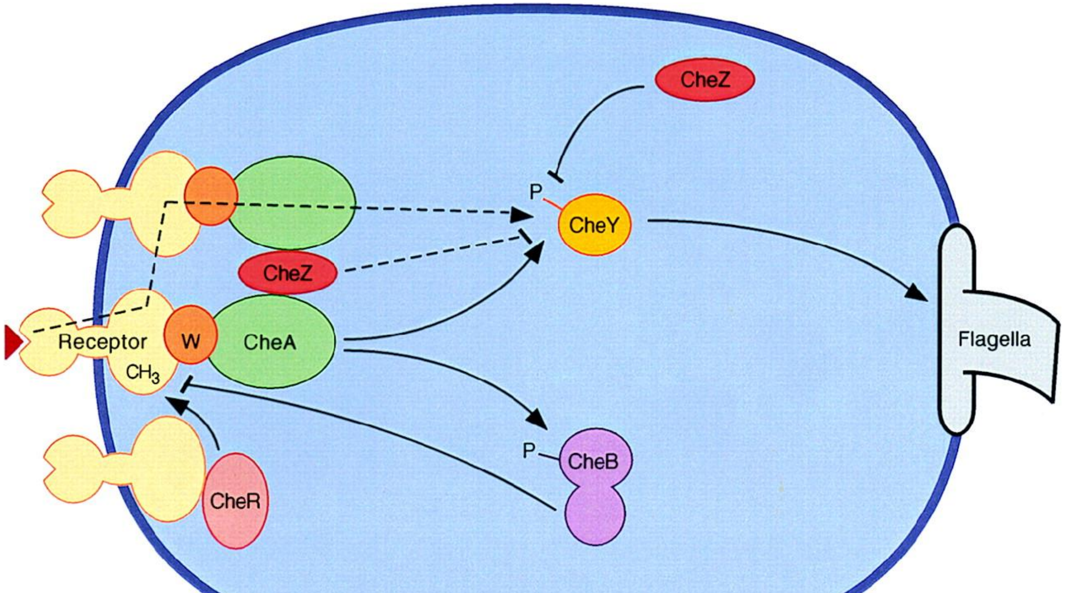


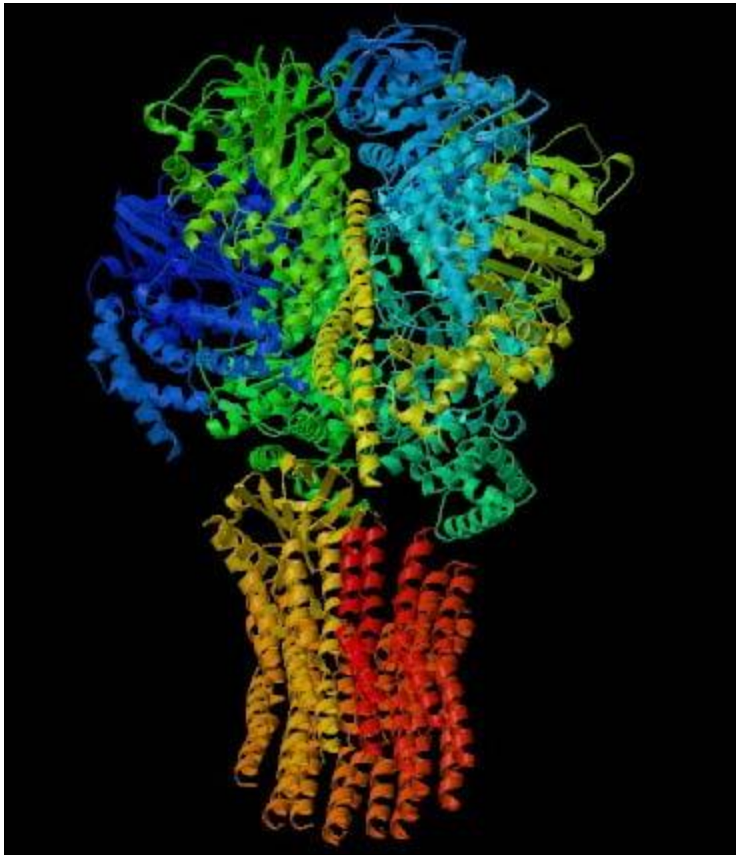
Gene Ontology


DAG view of Enriched GO Categories - Mozilla
File Edit View Go Bookmarks Tools Window Help
biological_process
molecular_function
cellular_component
cellular process
development
physiological processes
binding
transcription regulator activity
catalytic activity
cell
extracellular
cell communication
morphogenesis
metabolism
metal ion binding
nucleic acid binding
hydrolase activity
intracellular
extracellular matrix
nucleobase, nucleoside, nucleotide and nucleic acid metabolism
calcium ion binding
DNA binding
hydrolase activity, acting on ester bonds
organogenesis
nucleus
microfibril
eye morphogenesis
transcription
transcription factor activity
phosphoric ester hydrolase activity
nucleoplasm
regulation of transcription
transcription, DNA-dependent
phosphoric diester hydrolase activity
transcription factor complex
regulation of transcription, DNA-dependent
cyclic-nucleotide phosphodiesterase activity
3',5'-cyclic-nucleotide phosphodiesterase activity

Topic 5: Protein-Protein Interaction Prediction

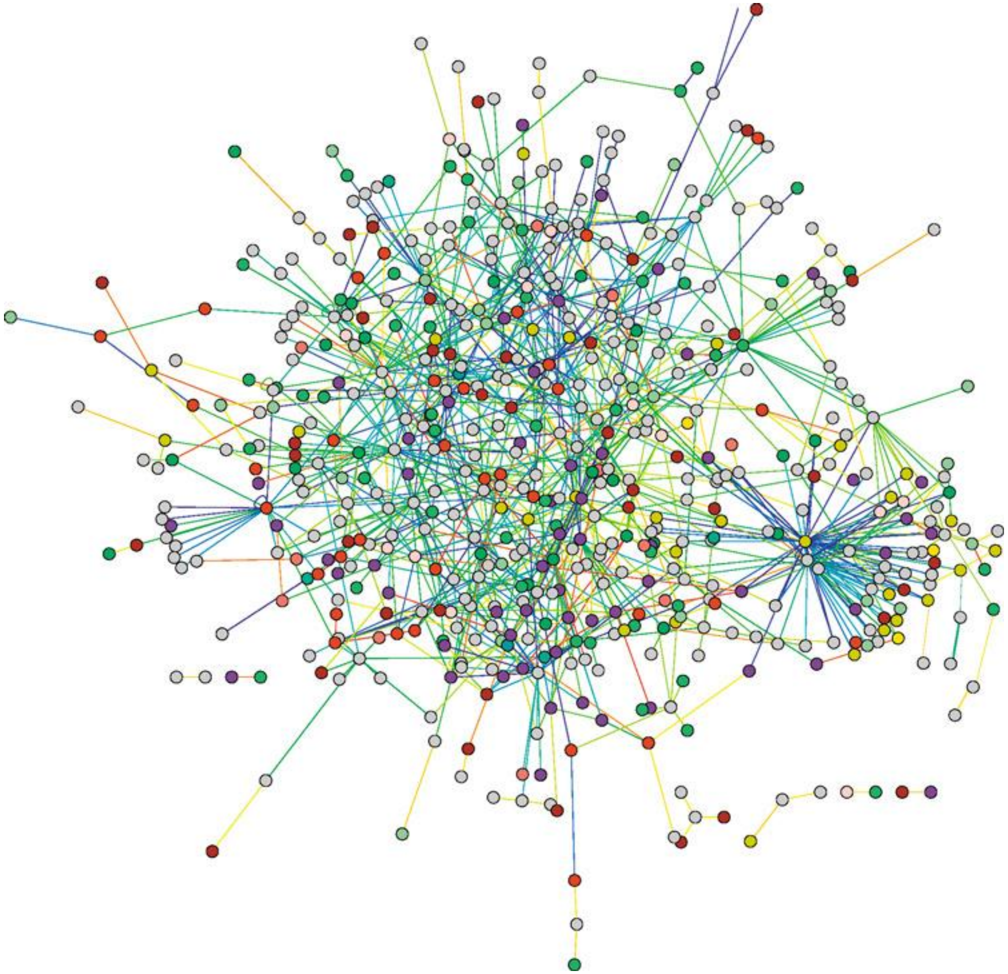






ATP Synthase

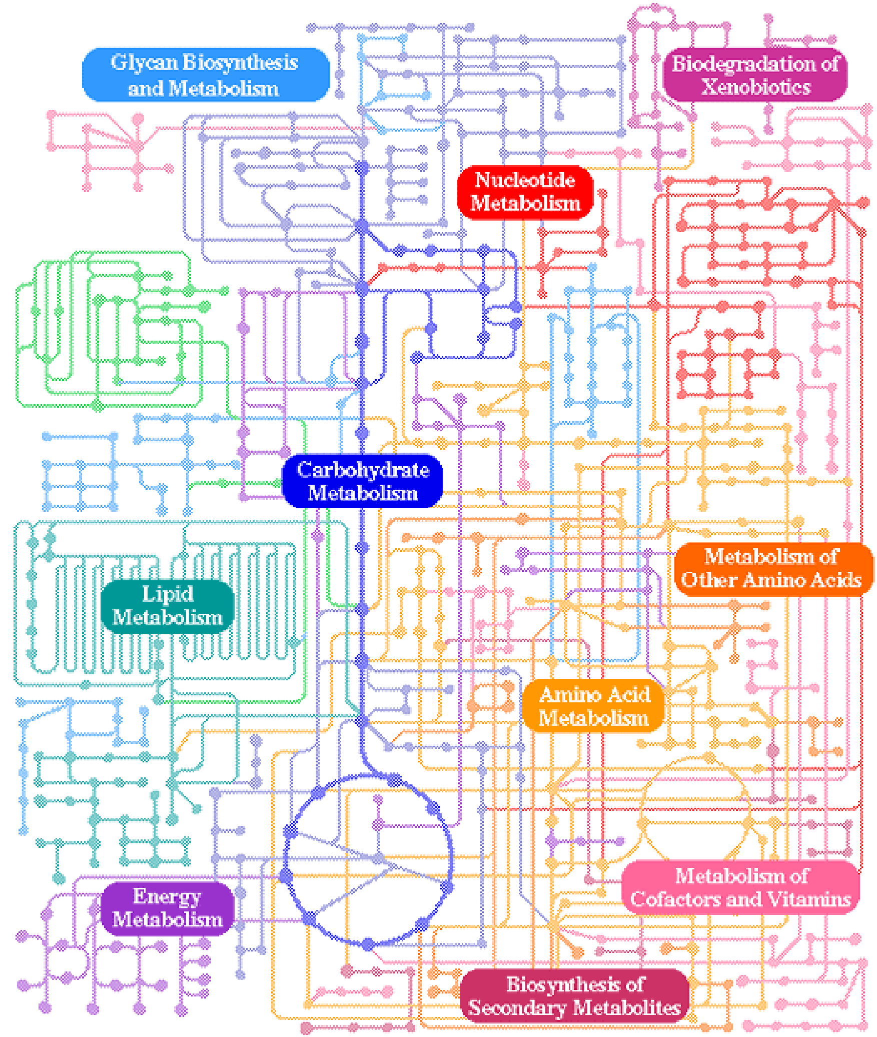
Protein Interaction Network



Topic 6: Reconstruction of Biological Pathway and Networks

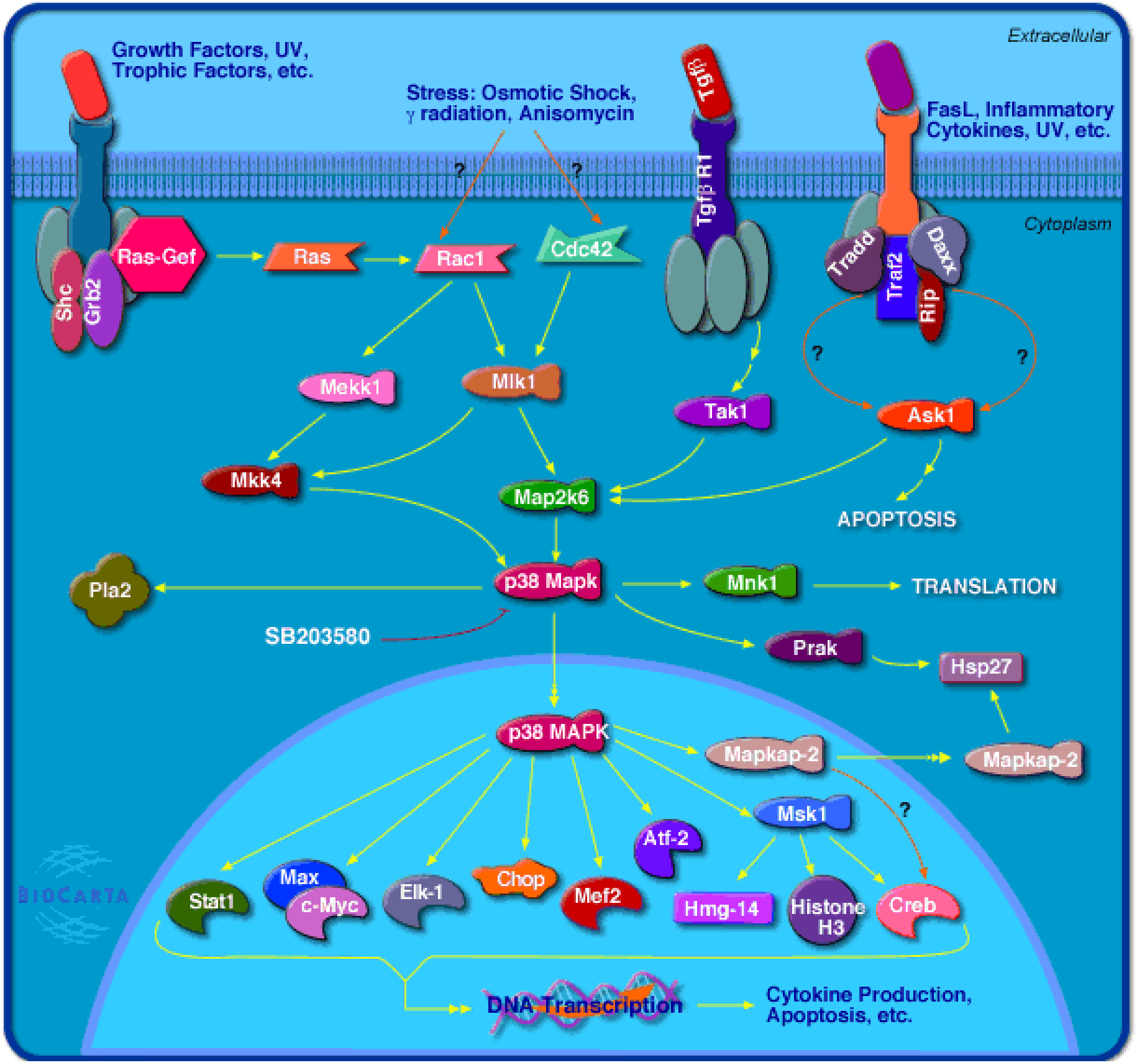
- Metabolic pathway
- Signal transduction pathway
- Gene regulatory pathway

Metabolic Pathway (KEGG)

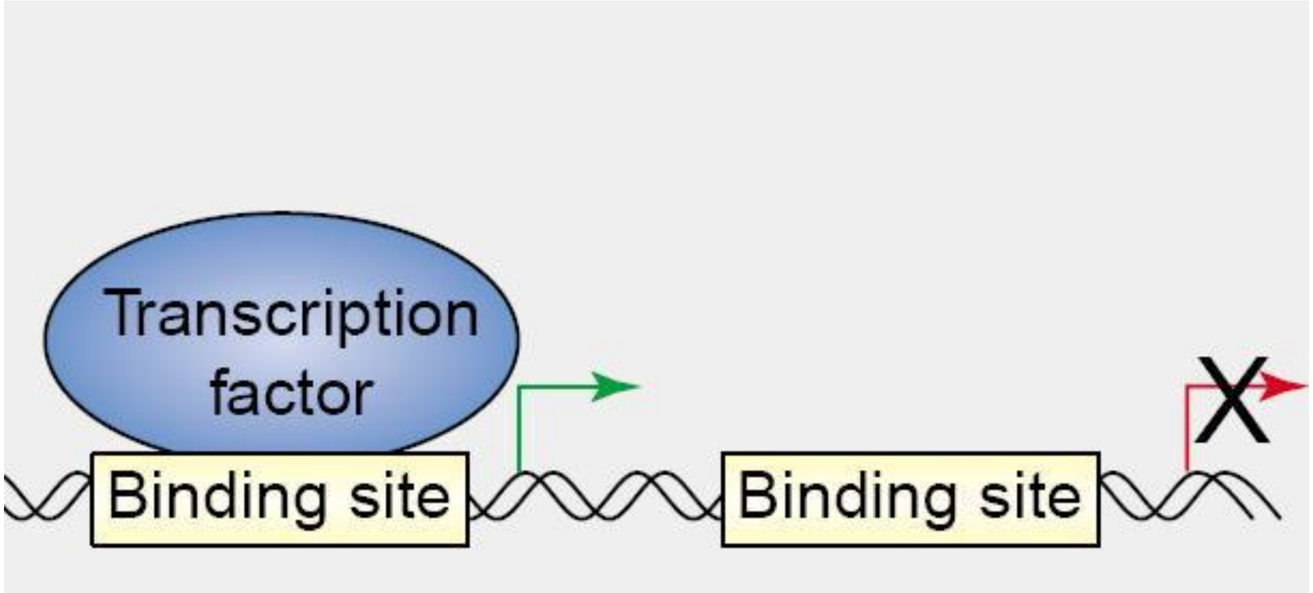

Glycan Biosynthesis and Metabolism
Biodegradation of Xenobiotics
Nucleotide Metabolism
Carbohydrate Metabolism
Metabolism of Other Amino Acids
Lipid Metabolism
Amino Acid Metabolism
Energy Metabolism
Metabolism of Cofactors and Vitamins
Biosynthesis of Secondary Metabolites

01100 5/31/04 Image source from KEGG

Signal Transduction Pathway

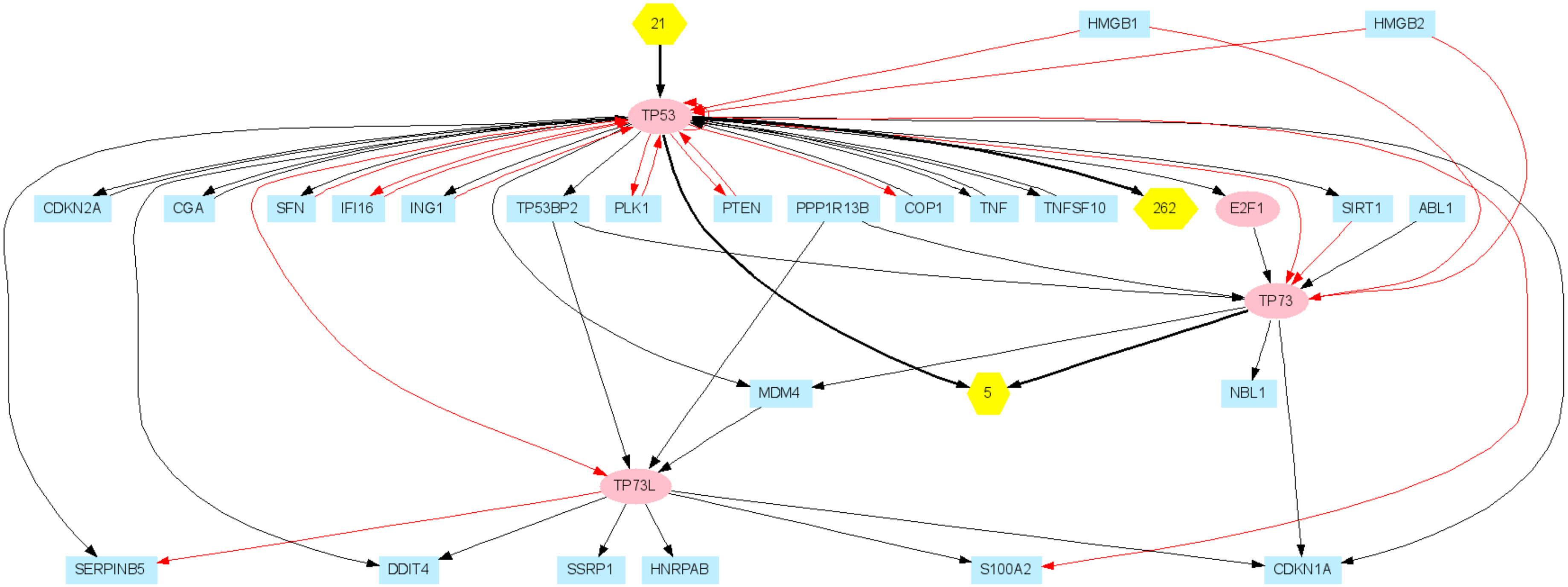


Gene Regulatory Pathway



Gene Regulatory Network

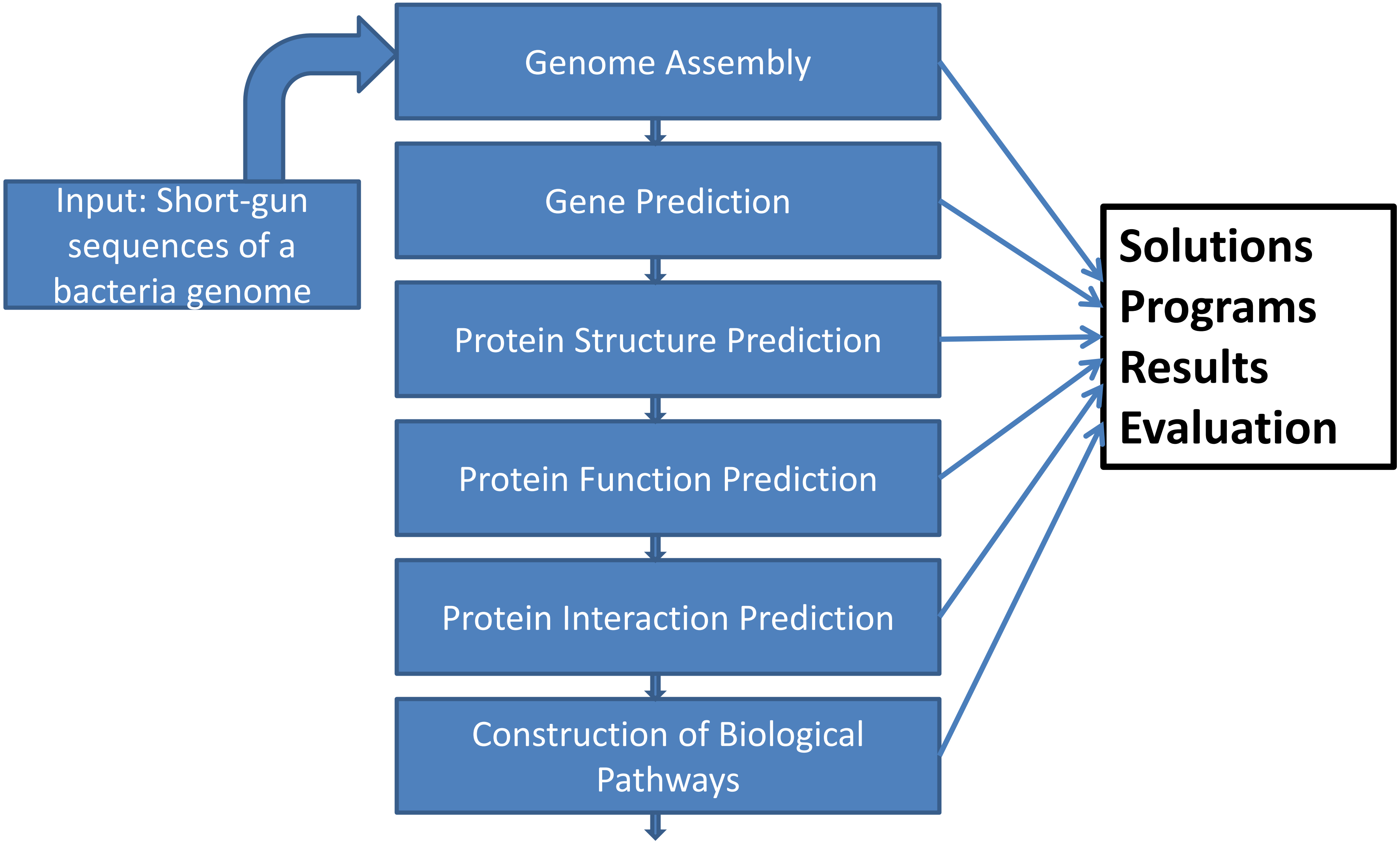
Gene Regulatory Network of TF family p53 in human



A lot of techniques and challenges, how can we get it done in one semester?

Novel learning technique: doing one genome assembly and annotation project in six steps

Group Project



Input: Short-gun sequences of a bacteria genome

Genome Assembly

Gene Prediction

Protein Structure Prediction

Protein Function Prediction

Protein Interaction Prediction

Construction of Biological Pathways

Solutions
Programs
Results
Evaluation

Reading Assignment

J.C. Venter et al.. The sequence of the Huan Genome. Science. 291:1304, 2001

Read: Introduction and first three sections:
http://www.sciencemag.org/cgi/reprint/291/5507/1304.pdf

Write a review (one page) to summarize the main problems, methods and results