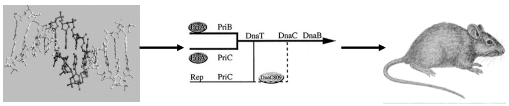
Genomic Signal Processing: The Sequel

From an EE perspective

Professor Gail L. Rosen



Outline of Class

- ✓Paper Discussions
- ✓Progress Reports
- ✓Final Projects

Paper Review

- ✓Everyone must read paper and turn in notes about each paper
- ✓Discussion Leader: 30 minute presentation on topic (slides or no slides) – turn in presentation if do one
- ✓Reviewer will begin "discussion"
- ✓The stage will open up for people to ask questions, chime in with further derivations
- ✓All to help understanding of paper

Deadlines

- ✓Progress Reports due 5/12
- ✓Final Project Presentations 6/11 or 6/12 (Wednesday or Thursday evening)
- ✓Everyone should have their date… if they do not , email me.

Syllabus Highlights

- ✓Project Topics– 4/14
  - ✓Start thinking about a topic (please feel free to meet with me).
  - ✓Schedule meetings with me to check feasibility of topic.
- ✓Final Project -- Exploratory research project for YOU to learn about the State-of-the-Art in the field (Due. June 11th).

Classic SP for Biology Applications

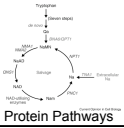
- ✓Most Popular: Speech Signal Processing
- ✓Pattern Recognition / Hidden Markov Models: Aligning sequences, classifying similar genes, gene prediction
- ✓Boolean Networks: Modeling Genetic Regulatory Networks

How have we historically looked at Biology?

Historical Understanding of Biology

Function

DNA



Protein Pathways

Beginnings of Medicine: 2000 B.C. (Asia), 500 B.C. (Hippocrates)

Discovery of DNA: 1950 (Wilkins and Franklin), 1953 (Watson and Crick)

Feedback Regulation in Metabolism: 1957 (Umbarger, Brown) (Yates, Pardee)
1970's: major breakthroughs

DNA Composition

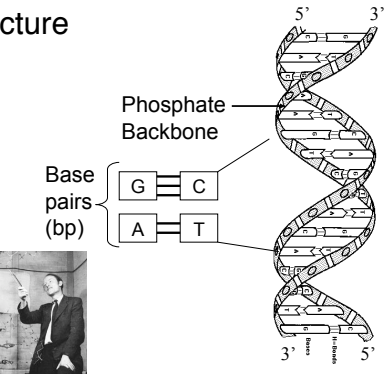
- ✓A – Adenine
- ✓T – Thymine
- ✓C – Cytosine
- ✓G – Guanine

- ✓ 4 Nucleotides (bases)
- ✓ 3 Bonds for G-C
- ✓ 2 Bonds for A-T
- ✓ Helical twist

DNA Structure

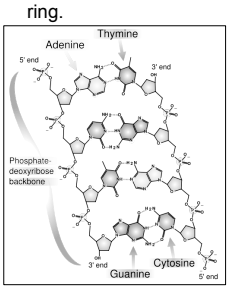
In mRNA
T ->U

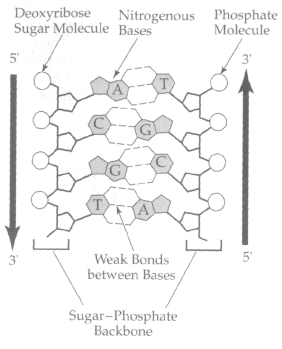
1953



Directional Reading

Third (3') and Fifth(5') Carbon Atoms in Sugar ring.





We have the bases -- now what?

- ✓What is a gene?

Genetic Code

- ✓Marshall Nirnberg (60's) discovers the genetic code
- ✓3 nucleotides produc one amino acid

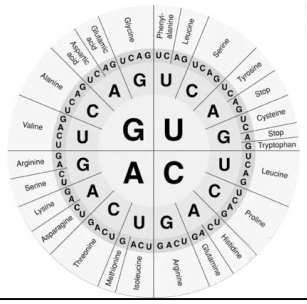
Standard Genetic Code

64 Codons map to:

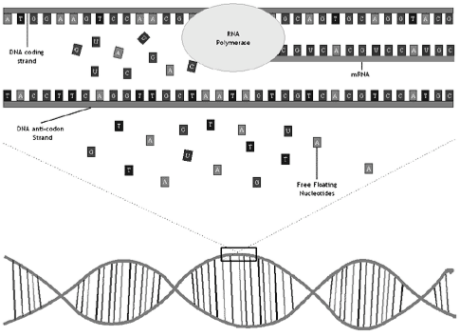
- 20 amino acids and
- start/stop codons

Genetic codes can vary among species

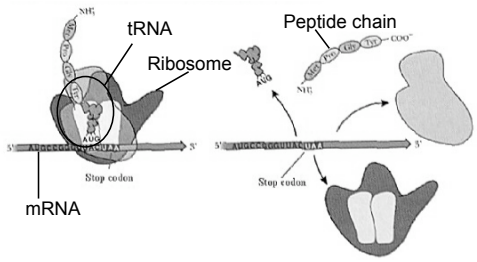
DNA RNA
T ≈ U



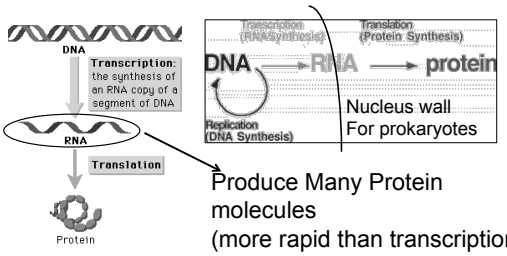
Transcription



Translation



Transcription / Translation



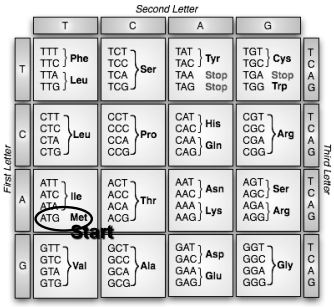
Produce Many Protein molecules
(more rapid than transcription)

Animation: http://www.johnkyrk.com/DNAtranscription.html
http://www.johnkyrk.com/DNAtranslation.html

DNA/RNA

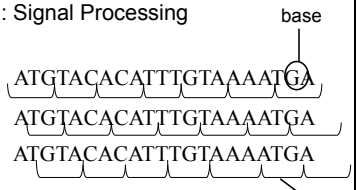
DNA	RNA
Stable: Double-stand, helix	Stable: Single-stranded, many shapes
Function: Stores genetic information	Function: Stores information, catalysts, larger structures (tRNA, Ribosomal)
Replication catalyst: DNA Polymerase -- conducts Proofreading	Replication catalyst: RNA Polymerase – no proofreading

Genetic Code



Codon Positions

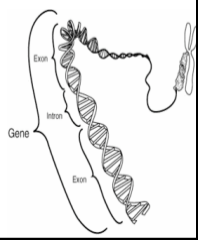
- ✓ Positions in Open Reading Frames (ORFs) : Biology Windows/Frames : Signal Processing



- ✓ Ribosome "slippage" in gene coding region could mean that a gene may be:
  1) Misinterpreted
  2) Not stopped
  3) Truncated early

Prokaryotes (without nucleus) vs. Eukaryotes (with nucleus)

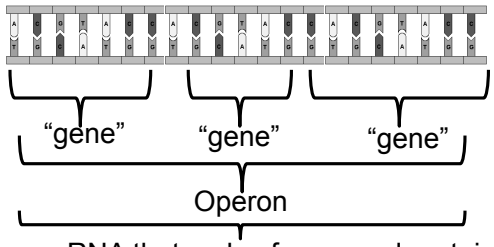
- ✓ Transcription and Translation different
  - ✓ http://highered.mcgraw-hill.com/olc/dl/120077/bio25.swf
- ✓ Motifs are different
- ✓ Eukaryotes have nucleus -- transcription occurs inside and translation outside
- ✓ Eukaryotes:Introns spliced out of mRNA
- ✓ Prokaryotes: Exons only



mRNA differences –Prokaryote vs. Eukaryote

Prokaryote

Coding region = gene



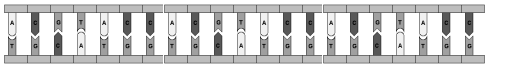
"gene" "gene" "gene"

Operon

Long mRNA that codes for several proteins

mRNA differences –Prokaryote vs. Eukaryote

Eukaryote



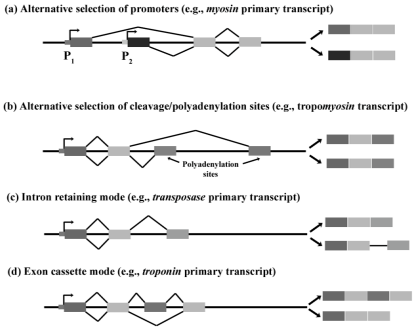
"coding" "coding" "coding" "coding"

gene gene

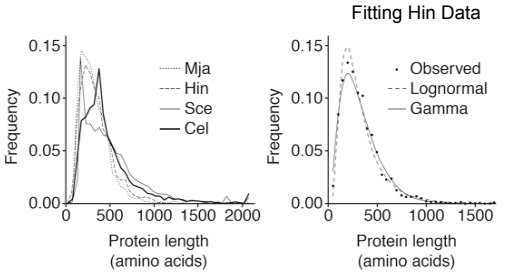
Single mRNA-> Single Protein

Single mRNA-> Single Protein

Alternative Splicing (Eukaryotes)

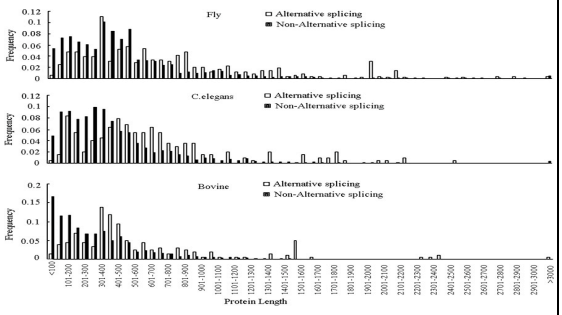


Histogram of Prokaryotic protein-lengths

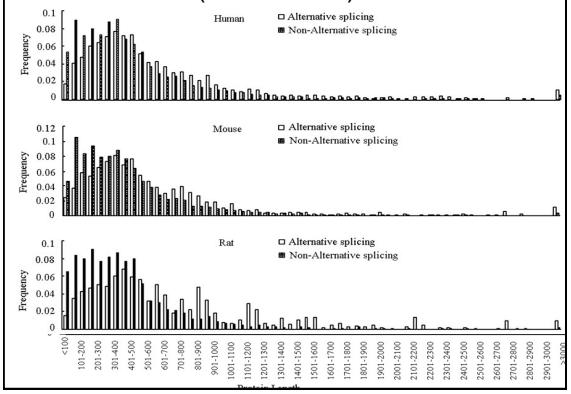


Mja, Methanococcus jannaschii; Hin, Haemophilus influenzae; Sce, Saccharomyces cerevisiae; Cel, Caenorhabditis elegans

Gene Length Distributions (Eukaryotes)



(Mammalian)



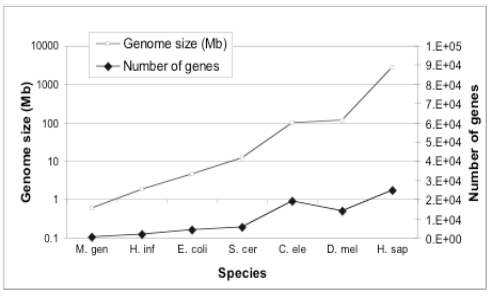
What are the comparative genome sizes of humans and other organisms being studied?

organism	estimated size	estimated gene number	average gene density	chromosc numbe
Homo sapiens (human)	2900 million bases	~30,000	1 gene per 100,000 bases	46
Rattus norvegicus (rat)	2,750 million bases	~30,000	1 gene per 100,000 bases	42
Mus musculus (mouse)	2500 million bases	~30,000	1 gene per 100,000 bases	40
Drosophila melanogaster (fruit fly)	180 million bases	13,600	1 gene per 9,000 bases	8
Arabidopsis thaliana (plant)	125 million bases	25,500	1 gene per 4000 bases	10
Caenorhabditis elegans (roundworm)	97 million bases	19,100	1 gene per 5000 bases	12
Saccharomyces cerevisiae (yeast)	12 million bases	6300	1 gene per 2000 bases	32
Escherichia coli (bacteria)	4.7 million bases	3200	1 gene per 1400 bases	1
H. influenzae (bacteria)	1.8 million bases	1700	1 gene per 1000 bases	1

Genome size does not correlate with evolutionary status, nor is the number of genes proportionate with genome size. (Largest Genome size - Protopterus aethiopicus, Marbled lungfish @ 130 Gbp)

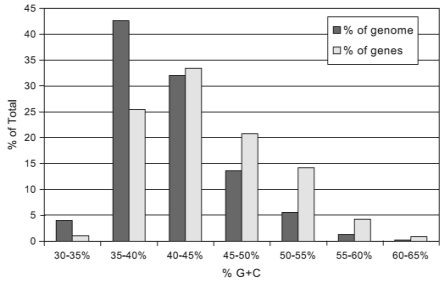
Organism	Year	Millions bases sequenced	Predicted genes	Genes per million bases
H. Influenza	1995	1.8	1,850	1030
E. coli	1997	4.6	4,200	900
M. genitalium	1995	0.58	468	806
S. cerevisiae	1996	12	5,800	483
C. elegans	1998	97	19,000	195
D. melanogaster	2005	116	14,100	122
A. thaliana	2000	115	25,500	220
Human Chr. 21	2000	34	225	7
Human Chr. 22	1999	34	545	16
Human (draft)	2001	2,693	31,780	12
Human (finished)	2004	2,850	~25,000	9
Mouse (draft)	2002	2,372	~30,000	13

Genome size increases exponentially, but not the number of genes



C+G Content Distribution in Humans

50 kbp Windows



Why are eukaryotic genomes larger?

- ✓Gene structure
  - ✓ No operons (genetic elements to produce mRNA)
  - ✓ Introns
  - ✓ Extensive 5' and 3' regulatory regions
- ✓Spacer regions
- ✓Repetitive DNA
- ✓Non-coding functional DNA

Genes - only ~3% of Human Genome

- ✓ What is in between ? Signal sites for ribosomes but currently speculated as "junk" (insertions from viruses and leftover elements that got weeded out due to mutation.
- ✓ Where are the Genes?
  - ✓ Start/Stop codons are a clue (not always)
  - ✓ Transition probabilities (HMM's successful)
  - ✓ Use genes from other species to find genes in a new species?

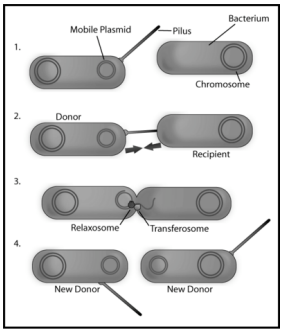
Haploid vs. Diploid genomes

- ✓E. Coli: Only one copy of each gene (Haploid)
- ✓Humans: Two copies of each gene (Diploid) … Pairs of Chromosomes

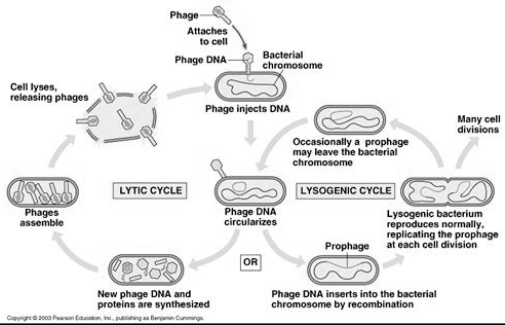
E. Coli shows new phenotypes sooner (Adapts faster to its environment)

Bacterial Mating

- ✓Plasmids

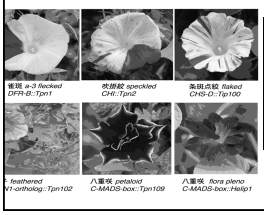


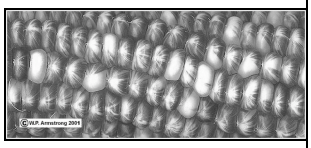
Insertion of DNA by Virii



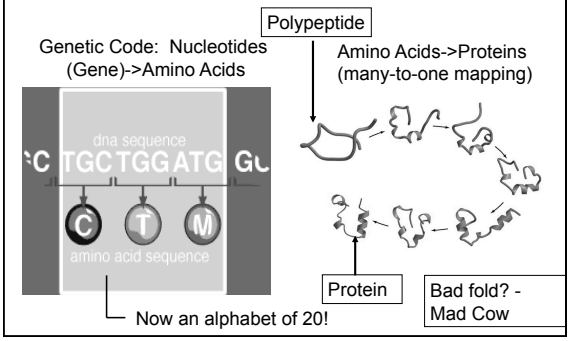
Transposons

- ✓Move from place to place within DNA
- ✓"Transposable Elements"
- ✓"Jumping Genes" or Mobile Elements



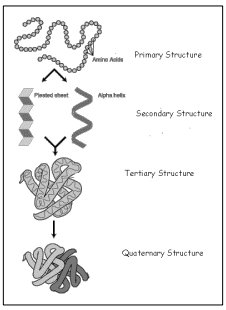


Nucleotides to Proteins



Four Structures towards Proteins

- ✓Primary Structure: Polypeptides
- ✓Secondary Structure: Alpha helices and Beta sheets
- ✓Tertiary Structure: 3D
- ✓Quaternary:
  - ✓ Multiple polypeptides



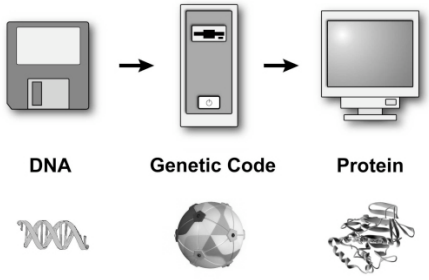
Challenges in Protein Structure Prediction

- ✓ The number of possible structures is extremely large
- ✓ The physical basis of protein structural stability is not fully understood
- ✓ The amino acid sequence itself may not fully specify the tertiary structure. (Environment)
- ✓ Simulating protein folding is computationally intensive (thus the Folding@Home project)

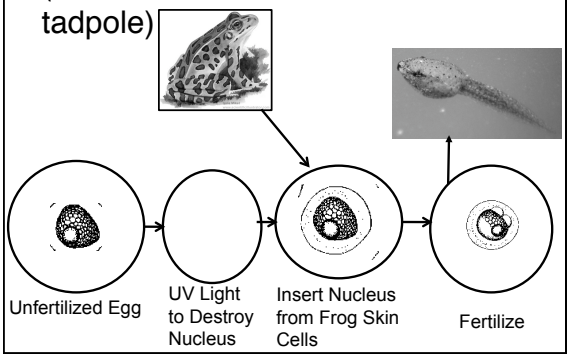
OK - we have a protein. Now what?

- ✓ Protein-Protein Interactions (Gene Regulatory Networks)
  - ✓ Enzymes, proteins, etc. trigger the production of proteins
  - ✓ How do they know when?!
- ✓ Protein-Protein Interactions (Metabolic Pathways)
  - ✓ Energy Production/ Glycolysis
  - ✓ Pathways building more complex functions
- ✓ Gene Expression Analysis (how much protein generated from a gene)

The Genetic Machinery (Central Dogma of Genomics)



The Context for Gene Regulation (Nucleus from Skin Cell creates tadpole)



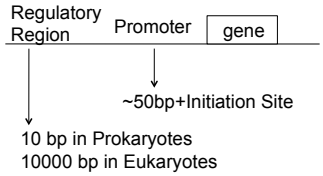
Two types of gene classes

Genes Common to all Cells	Genes characteristic to cell
Housekeeping Genes DNA Polymerase RNA Polymerase Ribosomes	E.g. Red Blood Cell Hemoglobin E.g. Insulin Pancreas

Cell Control of Protein Generation

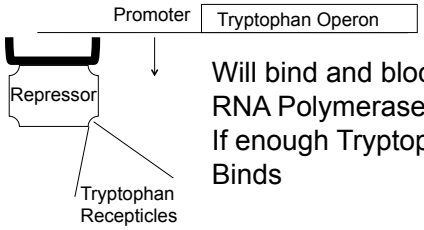
- ✓ Controlling when and how often a given gene is transcribed
- ✓ Controlling how the primary transcript is spliced or otherwise processed
- ✓ Selecting which mRNAs are activated by the proteins
- ✓ Activating or Inactivating Proteins

What we know about the regulation



These sites take several "input" gene regulatory proteins
The rate of transcription
Repressors turn off transcription
Activators turn on transcription

Constitutive Gene Expression (Tryptophan example)



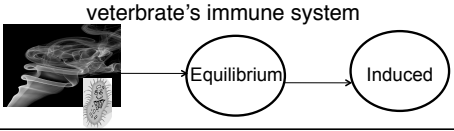
Will bind and block RNA Polymerase If enough Tryptophan Binds

Tryptophan is always present in low-levels (continual expression) – Constitutive

Tryptophan is an essential amino acid for the human diet

Induced/Facultative Gene Expression

- ✓ Induced Gene Expression
  - ✓ Responding to Stimulus Examples:
  - ✓ SIEGE: Smoking Induced Epithelial Gene Expression Database (on airways)
  - ✓ Interferon-induced gene expression (Interferon released only upon "intruder" to veterbrate's immune system



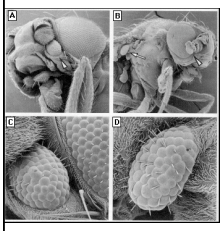
Prokaryotic vs. Eukaryotic Transcription

- ✓Prokaryotes – only need to generate RNA Polymerase to initiate transcripion
- ✓Eukaryotes – very complicated
  - Need to generate 3 RNA polymerases AND Transcription Factors
  - Transcription factors help place RNA polymerase at promoter, pull apart DNA strands, and release RNA polymerase when done
  - Because of "bends", regulatory regions may be 1000 of bases away and control transcription initiation

P vs. E Transcription Initiation

- ✓Prokaryotes:
  - ✓ Bacterial genes controlled by a single activator or repressor protein
  - ✓ Expression level coordinated by single protein
- ✓Eukaryotes:
  - ✓ Committee of regulatory proteins – thought of as combinatorial control or Boolean logic
  - ✓ Expression level controlled by single protein (if protein is last element needed in combinatorial control)

Eyeless (or too many), Ey, Gene

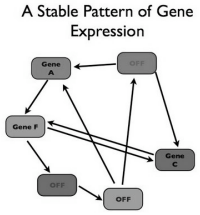


Expression of Ey in Development of wings, Legs, etc.

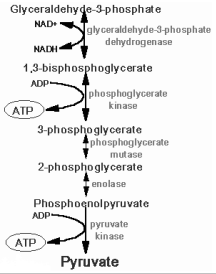
A) Scanning electron micrograph of an ectopic eye (arrowhead) in the head region formed by the antennal disc.
(B) Overview of a fly with an ectopic eye under the wing (arrow) and on the antenna (arrowhead).
(C) Higher magnification of (A). The ectopic eye (to the left) contains hexagonal ommatidia and interommatidial bristles. The organization of the facets in the ectopic eye is very similar to the pattern in the normal eye (to the right). Some facets, however, are fused and some irregularities are observed.
(D) Higher magnification of the ectopic eye under the wing shown in (B) (arrow). The ectopic eye protrudes out of the thoracic body wall (ventral pleura). The organization of the facets and interommatidial bristles are similar to that of the ectopic eye shown in (C).

Pathway Diagrams: Qualitative

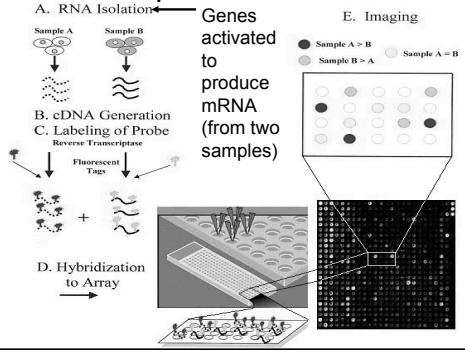
Genetic Regulatory Network



Metabolic Pathways (Glycolysis)



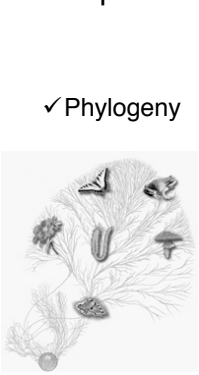
Microarrays: Quantitative attempts for Gene Expression

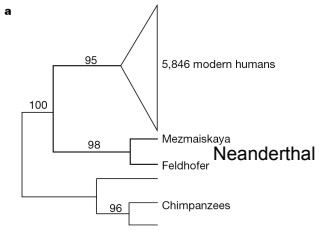


Genes activated to produce mRNA (from two samples)

Comparitive Genomics

- ✓Phylogeny

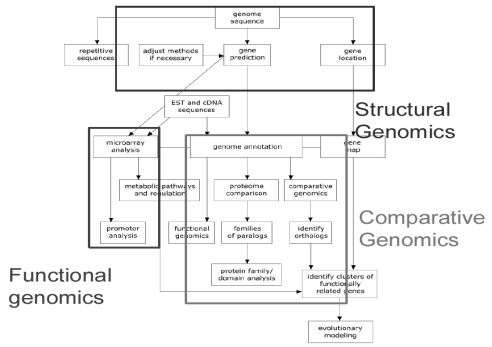




Tying it all together

Structure analysis	Sequence analysis	Function analysis
nucleic acid structure prediction	genome comparison	metabolic pathway modeling
protein structure prediction	phylogeny	gene expression profiling
protein structure classification	gene & promoter prediction	protein interaction prediction
protein structure comparison	motif discovery	protein subcellular localization prediction
	sequence database searching	
	sequence alignment	

Application Areas for Projects



Part II: Databases

- ✓ Nucleotide Sequences (Genbank, EMBL,etc.)
- ✓ Protein sequences (Uniprot)
- ✓ Human Sequence patterns (STRBase)
- ✓ Macromolecular 3D structure (MMDB)
- ✓ Gene Expression data (Omnibus, NCI, Stanford, etc.)
- ✓ Metabolic Pathways (KEGG, MetaCyc, etc.)

A brief history of biological databases

1965 M. O. Dayhoff *et al.* publish "Atlas of Protein Sequences and Structures"

1982 EMBL initiates DNA sequence database, followed within a year by GenBank (then at LANL) and in 1984 by DNA Database of Japan

1988 EMBL/GenBank/DDBJ agree on common format for data elements

2002 Gene Expression Omnibus

NCBI National Center for Biotechnology Information
National Library of Medicine National Institutes of Health

www.ncbi.nlm.nih.gov

- ✓Created in 1988 as part of the National Library of Medicine at NIH
  - ✓Establish public databases
  - ✓Research in computational biology
  - ✓Develop software tools for sequence analysis
  - ✓Disseminate biomedical information

Growth of GenBank database