Computational Genomics

10-810/02-710, Spring 2009

Class Overview

and

Introduction to Molecular Biology

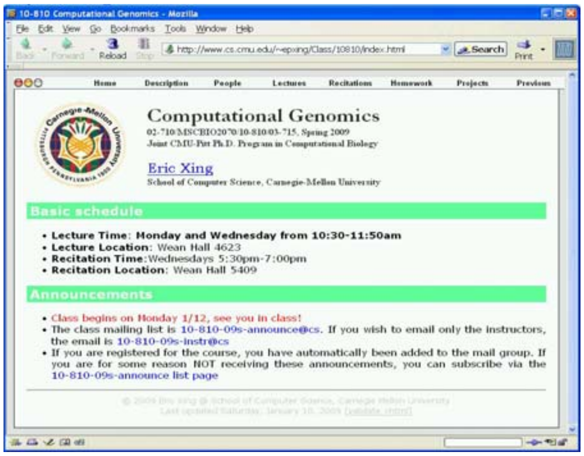
Eric Xing

Lecture 1, January 12, 2009

Reading: Chap. 1, DTM book

Logistics

- Class webpage:
  - http://www.cs.cmu.edu/~epxing/Class/10810/



Computational Genomics

02-710/MSCBIO2070/10-810/03-715, Spring 2009

Joint CMU-Pitt Ph.D. Program in Computational Biology

Eric Xing

School of Computer Science, Carnegie-Mellon University

Basic schedule

- **Lecture Time: Monday and Wednesday from 10:30-11:50am**
- **Lecture Location**: Wean Hall 4623
- **Recitation Time**:Wednesdays 5:30pm-7:00pm
- **Recitation Location**: Wean Hall 5409

Announcements

- Class begins on Monday 1/12, see you in class!
- The class mailing list is 10-810-09s-announce@cs. If you wish to email only the instructors, the email is 10-810-09s-instr@cs
- If you are registered for the course, you have automatically been added to the mail group. If you are for some reason NOT receiving these announcements, you can subscribe via the 10-810-09s-announce list page

Logistics

- **Instructors:**
  - **Eric Xing**, Wean Hall 4127, x8-2559, Office hours: TBA Wed 1-2
  - **Bino John**, 3080 Biomedical Science Tower 3, 3501 Fifth Avenue, Pittsburgh PA 15260, 412 648 8607, Office hours: Mon 4-5
- **Class Assistant:**
  - Michelle Martin, 3080 Biomedical Science Tower 3, 3501 Fifth Avenue, Pittsburgh PA 15260, 412-648-8607
- **Teaching Assistant:**
  - Pradipta Ray, Newell Simon Hall 4505, x8-3559, Office hours: Thurs 5pm - 6pm,
  - Grace Huang, Biological Sciences Tower 3, Suite 3064, N/A, Office hours: Thu 4pm - 5pm
  - Guy Zinman, Mellon Institute 409J, x8-9111, Office hours: Fri 12pm - 1 pm
  - Sabah Kadri, Mellon Institute 409F, x8-8101, Office hours: Mon 4:30 pm - 5:30 pm

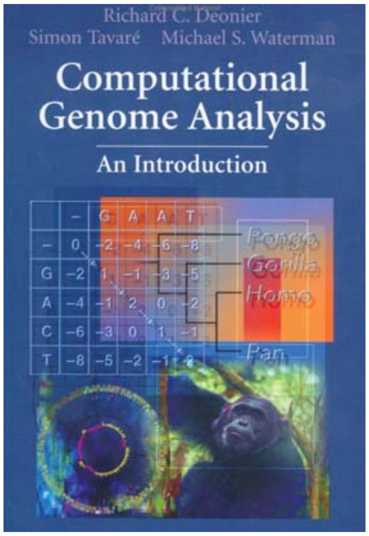
Logistics

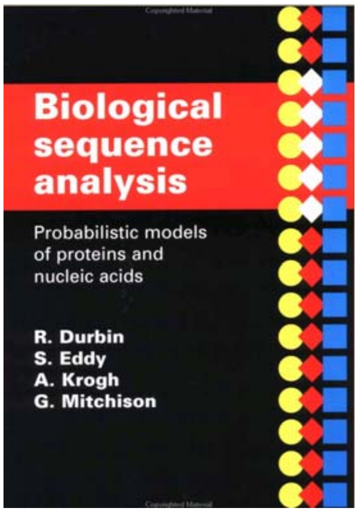
- 5 homeworks: 35% of grade
  - Theory exercises
  - Implementation exercises
- Final project: 30% of grade
  - Applying what you learned in the class to a realistic, non-trivial CompBio problem
    - Sequence analysis, network modeling, microarray mining, genetic polymorphism, evolution …
    - If on your own current research, higher expectation, talk to us first …
  - Expected outcome
    - A deliverable software based on an new/extended model or algorithm
    - Addressing a biological problem via a thorough analysis of a realistic dataset
    - A mini-paper …
  - Collaboration policies …
    - Tow persons per team at most
- Exam: 25% of grade
  - Midterm only
- Class participation, reading, and final quiz: 10% of grade

Logistics

- No required text books, but suggested reading will be announced prior every class from:
  - Durbin et al, **Biological Sequence Analysis.**
  - Deonier, Tavare and Waterman, **Computational Genome Analysis.**
  - Selected papers
- Mailing Lists:
  - Send email with full name, department, register/audit
  - To contact the instructors: 10810-09s-instr@cs.cmu.edu
  - Class announcements list: 10810-09s-announce@cs.cmu.edu

Books

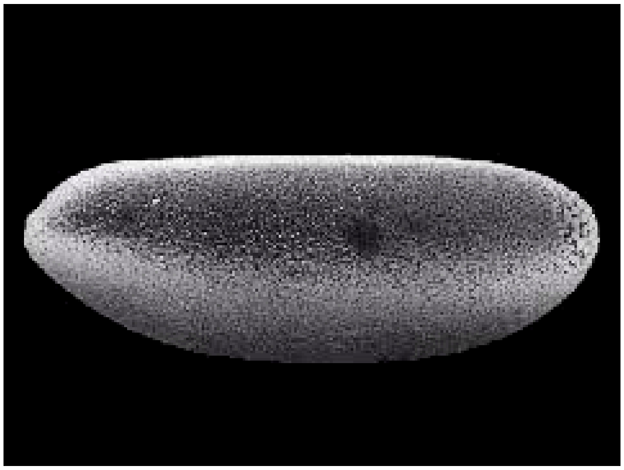




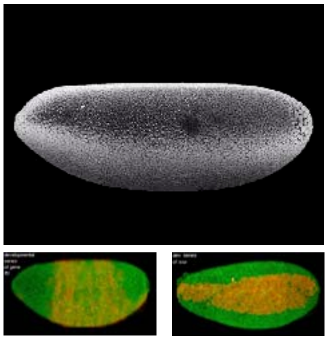
Class Plan

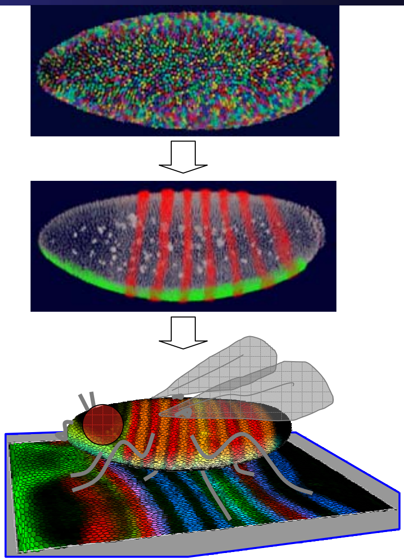
- Introduction (1week)
- Molecular Genomics: biological sequence analysis(3 weeks)
- Genome Evolution and Comparative Genomics(2.5 weeks)
- Gene Expression and Functional Genomics (2 weeks)
- Statistical Genetics and Genome-Phenome Associations (3 weeks)
- Systems Biology (3 weeks)

Biological Systems

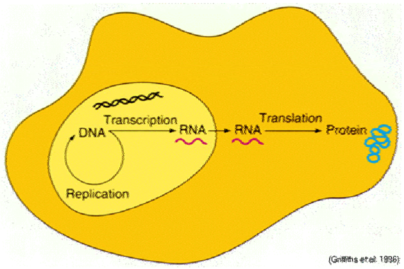


It is a developing system under tight spatial-temporal controls …





Understanding The Design, Regulation, and Evolution of Biological Systems

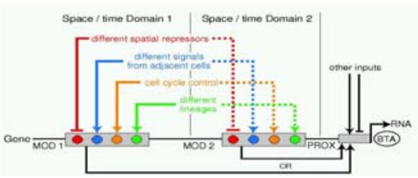


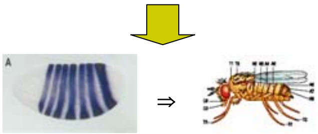
Essential bio-molecules are **digital**!

DNA

nucleotides

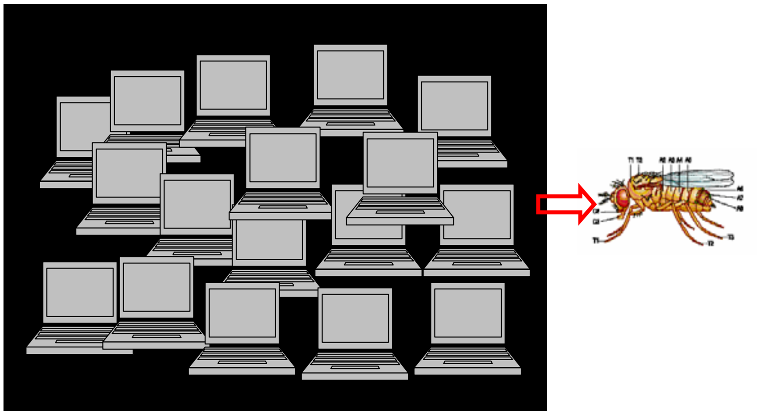
Gene expressions are controlled via **combinatorial logic**!



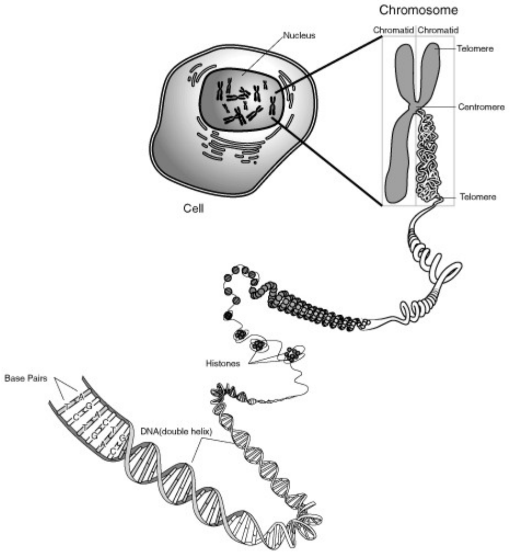


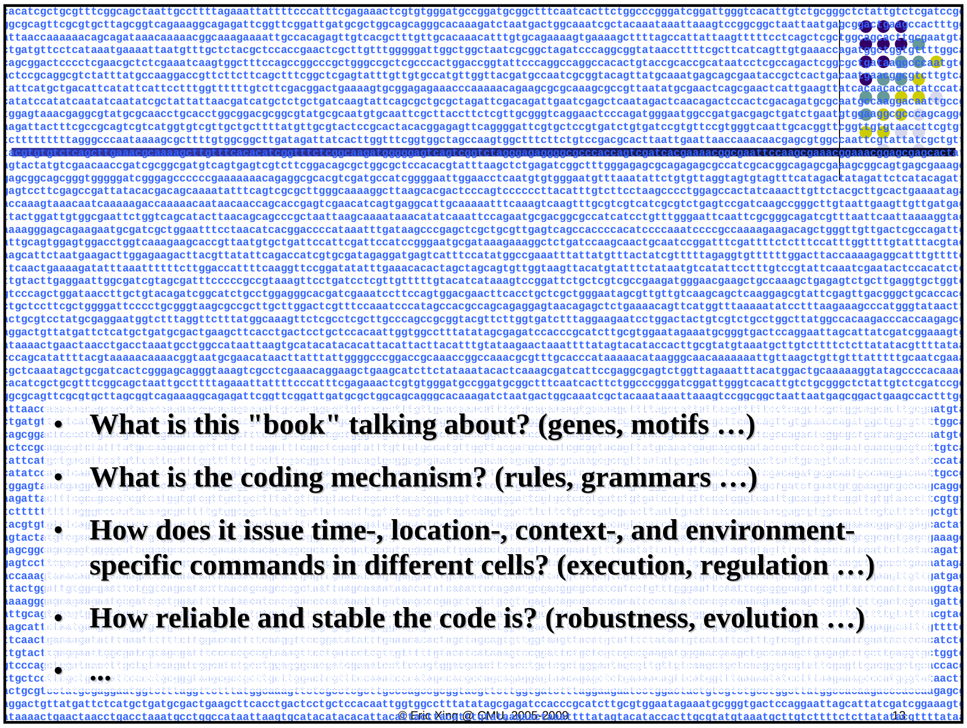
Multi-cellular organisms are parallel computers that **differentiate** in some **coordinated** way!

A living system is a perfect parallel and evolving computer system that we still know very little about …

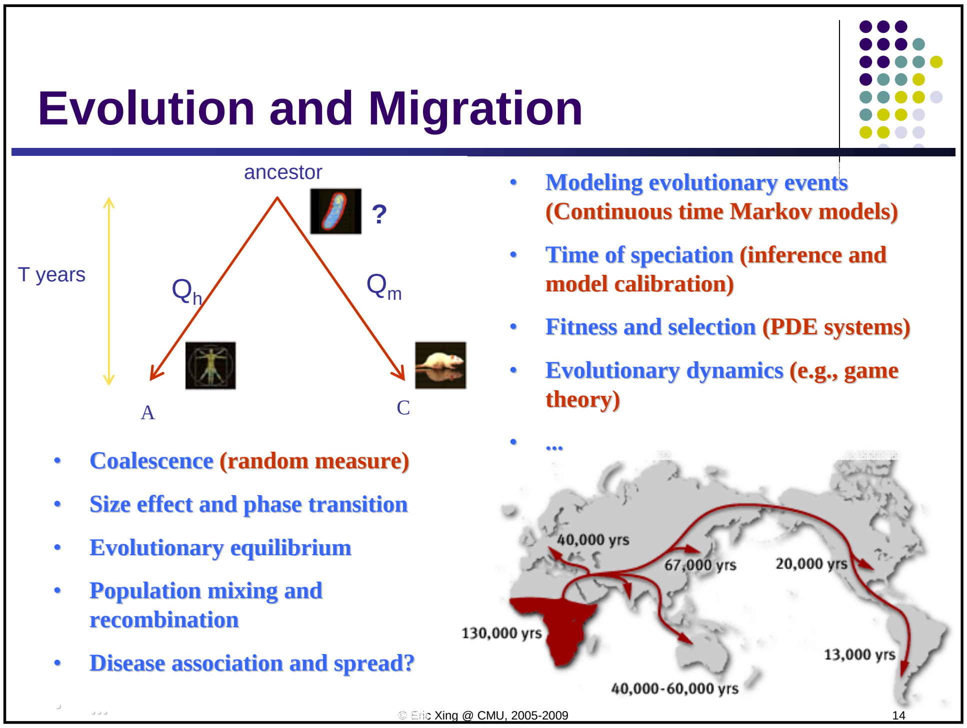


DECODING





- What is this "book" talking about? (genes, motifs …)
- What is the coding mechanism? (rules, grammars …)
- How does it issue time-, location-, context-, and environment-specific commands in different cells? (execution, regulation …)
- How reliable and stable the code is? (robustness, evolution …)
- …

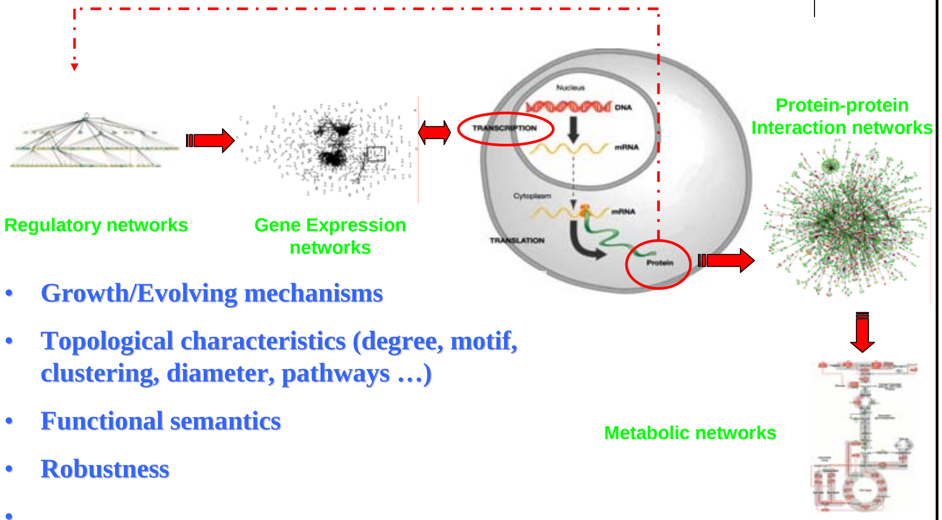


Evolution and Migration

- Modeling evolutionary events (Continuous time Markov models)
- Time of speciation (inference and model calibration)
- Fitness and selection (PDE systems)
- Evolutionary dynamics (e.g., game theory)
- …

- Coalescence (random measure)
- Size effect and phase transition
- Evolutionary equilibrium
- Population mixing and recombination
- Disease association and spread?
- …

Network Inference



Regulatory networks

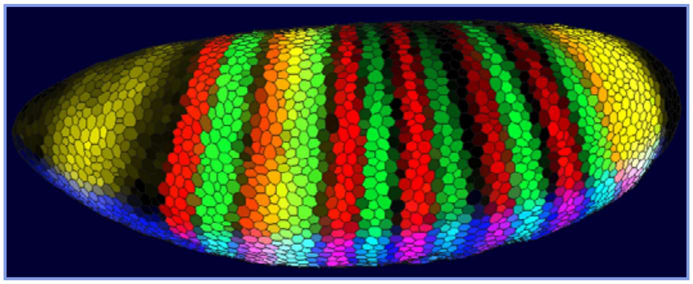
Gene Expression networks

Protein-protein Interaction networks

Metabolic networks

- Growth/Evolving mechanisms
- Topological characteristics (degree, motif, clustering, diameter, pathways …)
- Functional semantics
- Robustness
- …

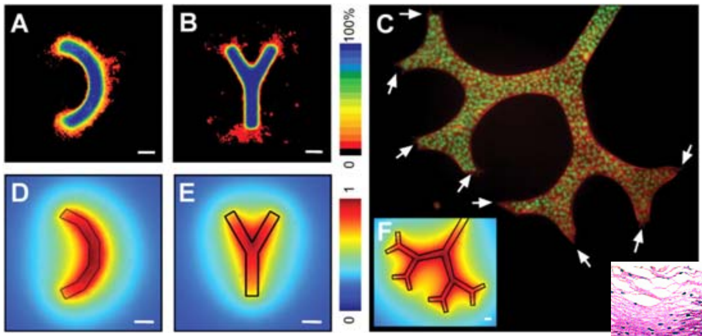
Bio-Imaging

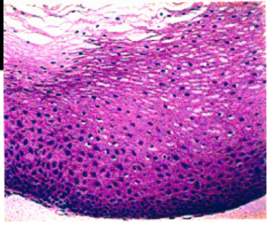


- Temporal-spatial gene expression pattern?
- Tissue/Organ formation?
- Morphogenesis?
- …

Courtesy to BDTNP group

Geometry

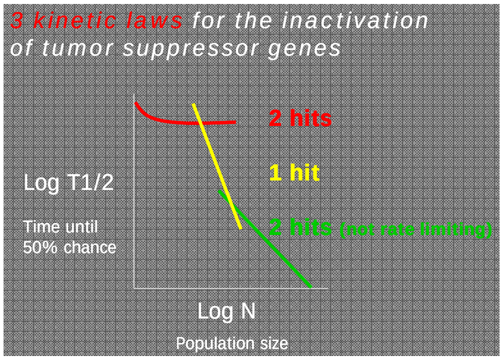




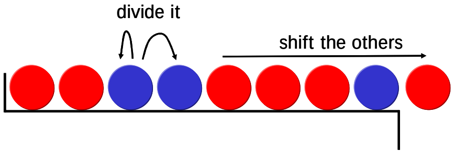
- Random fields and diffusion processes
- Tissue architecture
- Cell signaling
- Differentiation
- …

Nelson et al Science 2006

Population dynamics

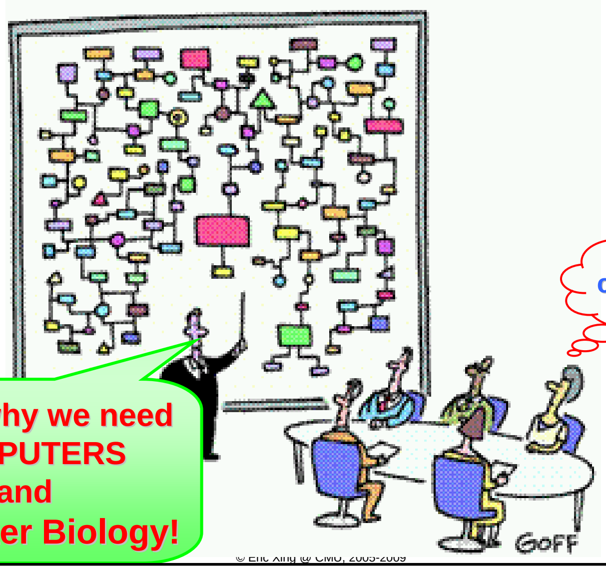


- Random discrete processes (jump Markov processes, Fokker-Planck model with different boundary conditions, Moran processes …)
- Size and geometry effects
- Tumorigenesis
- …



Courtesy to Martin Novak

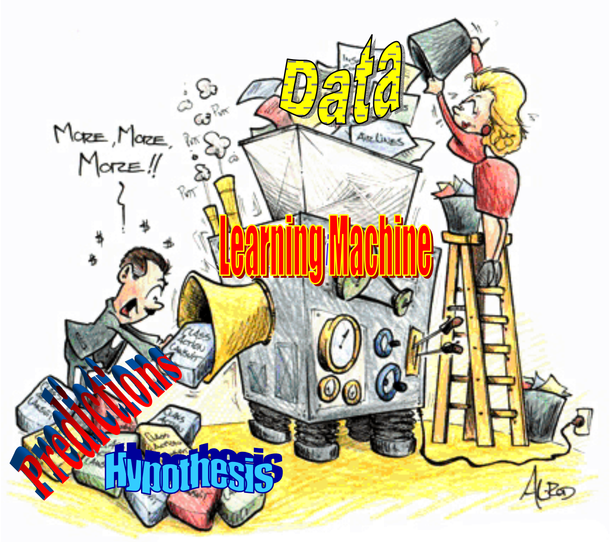
"Understanding a biological system as a whole is … "



So darn complicated !!

That's why we need COMPUTERS and Computer Biology!

Our Goal !!!



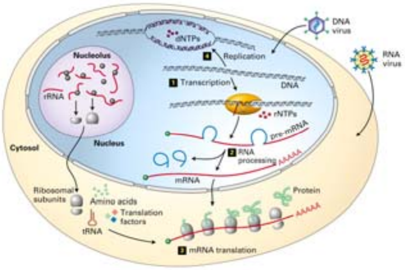
Data

Learning Machine

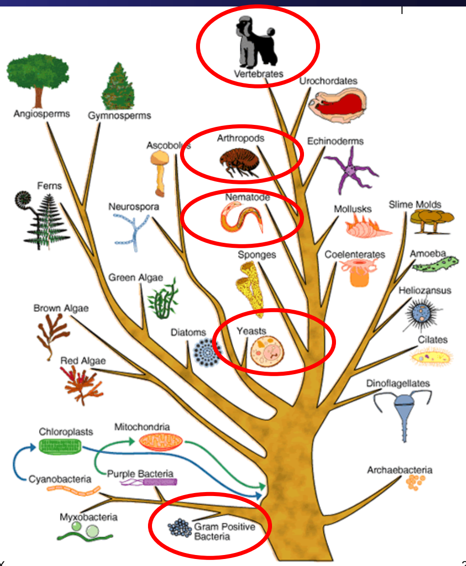
Predictions

Hypothesis

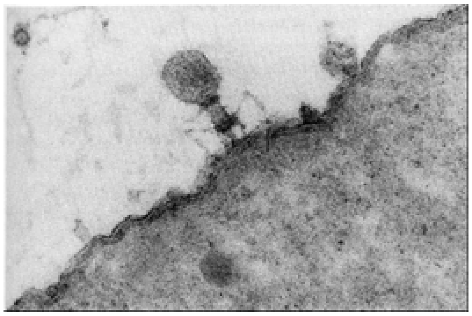
Introduction to Molecular Biology, Cell Biology, Development, Genetics, Probability, etc.

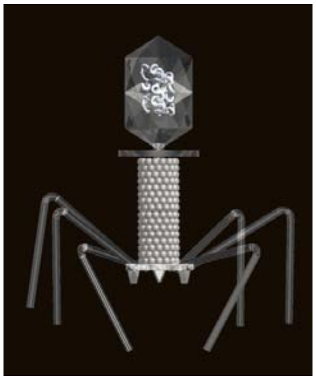


Model Organisms

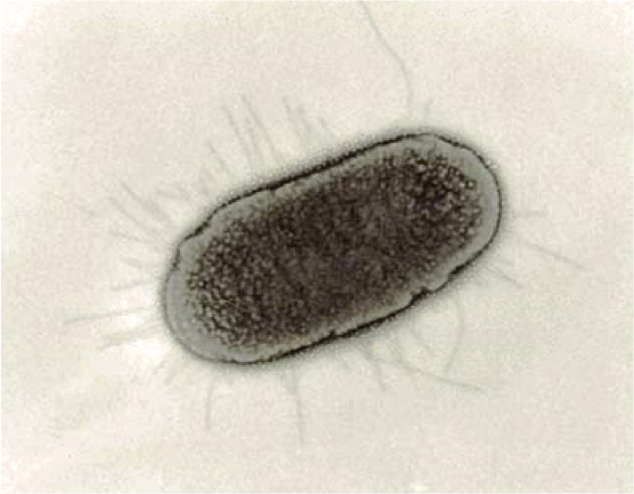


Bacterial Phage: T4

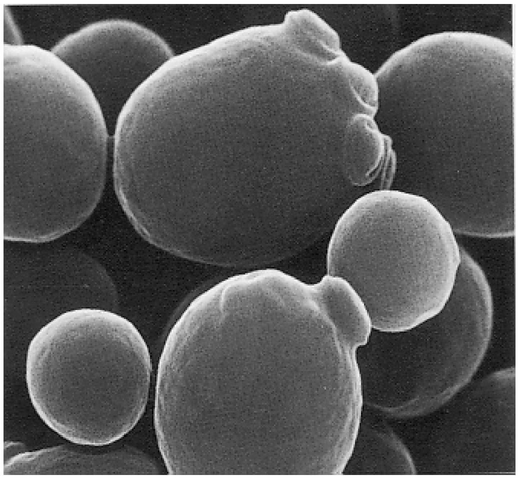




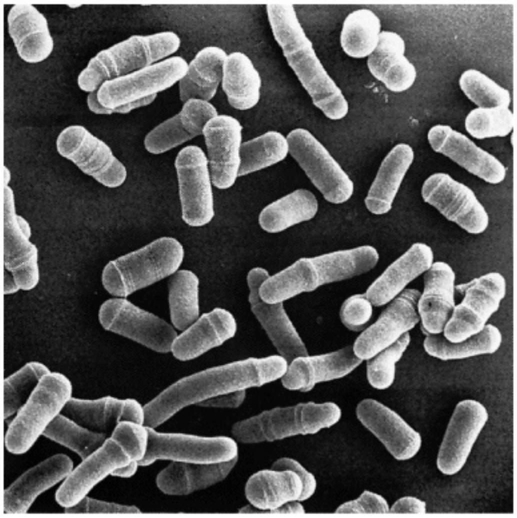
Bacteria: E. Coli



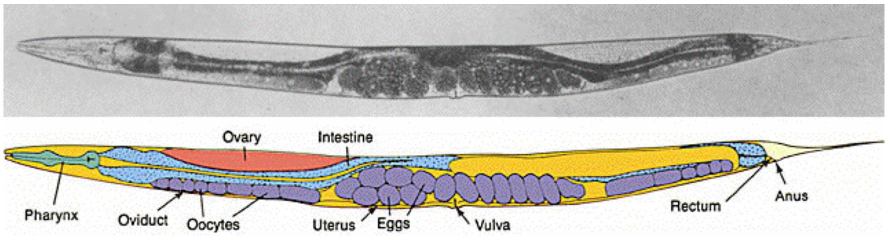
The Budding Yeast:
Saccharomyces cerevisiae



The Fission Yeast:
Schizosaccharomyces pombe



The Nematode: *Caenorhabditis elegans*



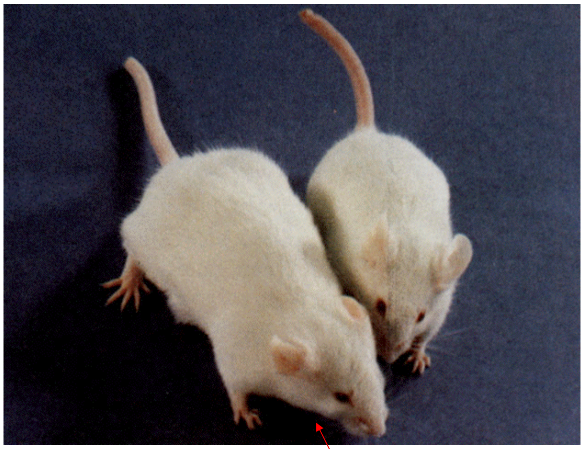
- SMALL: ~ 250 μm
- TRANSPARENT
- 959 CELLS
- 300 NEURONS
- SHORT GENERATION TIME
- SIMPLE GROWTH MEDIUM
- SELF- FERTILIZING HERMAPHRODITE
- RAPID ISOLATION AND CLONING OF MULTIPLE TYPES OF MUTANT ORGANISMS

The Fruit Fly: Drosophila Melanogaster

Normal

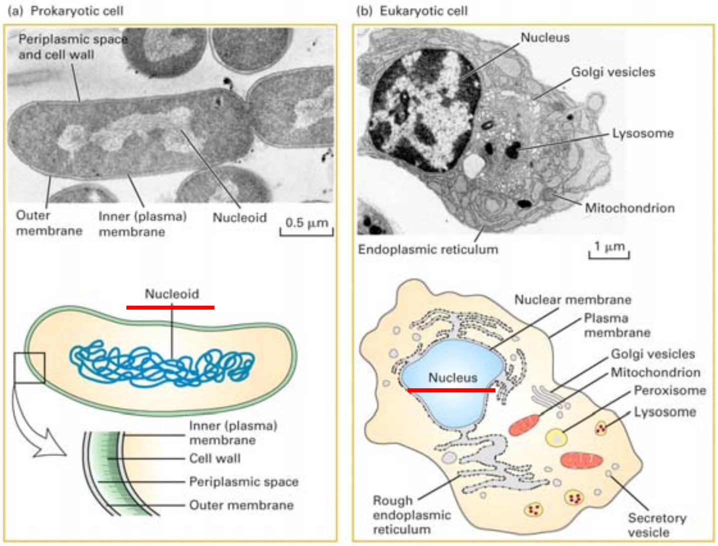
Ubx mutant

The Mouse



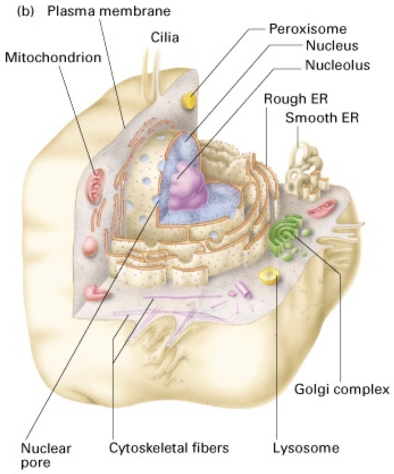
transgenic for human growth hormone

Prokaryotic and Eukaryotic Cells

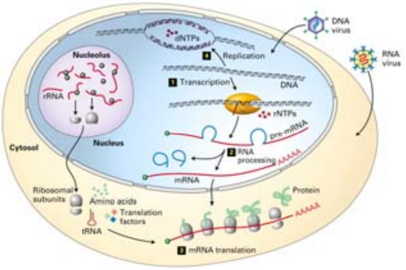


A Close Look of a Eukaryotic Cell

The structure:



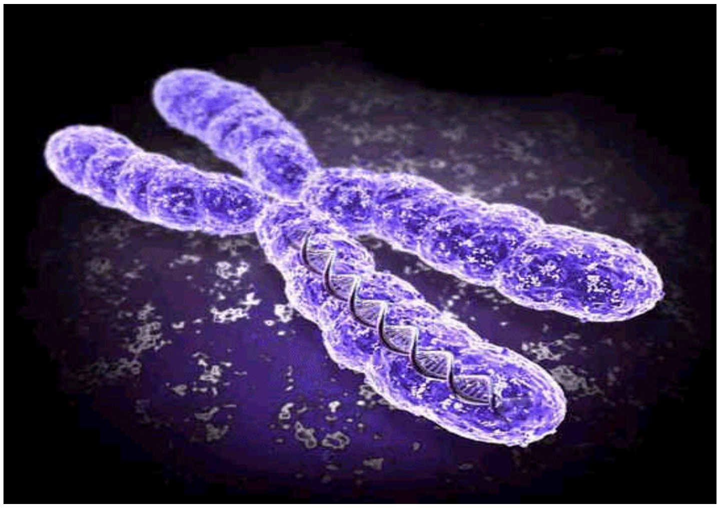
The information flow:



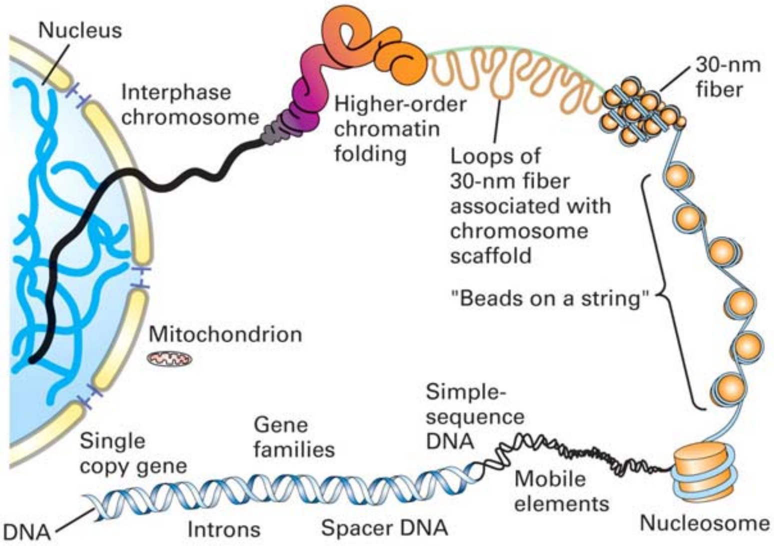
"Central Dogma" of Molecular Biology: Information Flow within the Cell

$$\text{DNA} \xrightarrow{\text{transcription}} \text{RNA} \xrightarrow{\text{translation}} \text{PROTEIN}$$

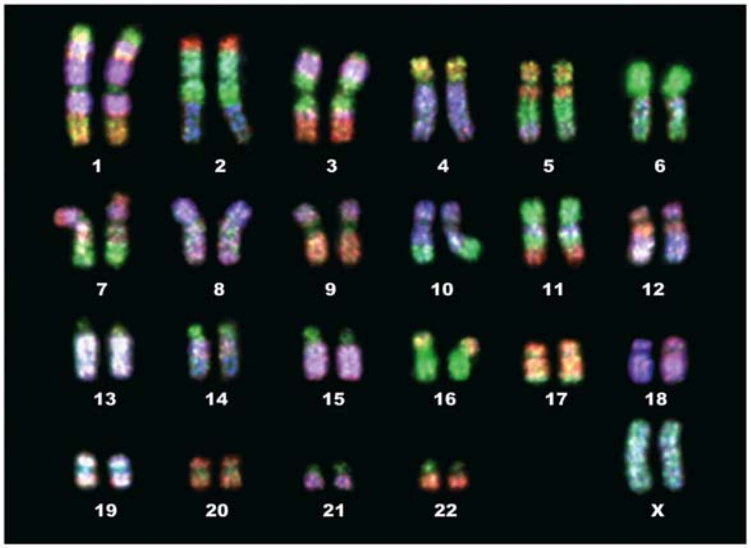
The Chromosome



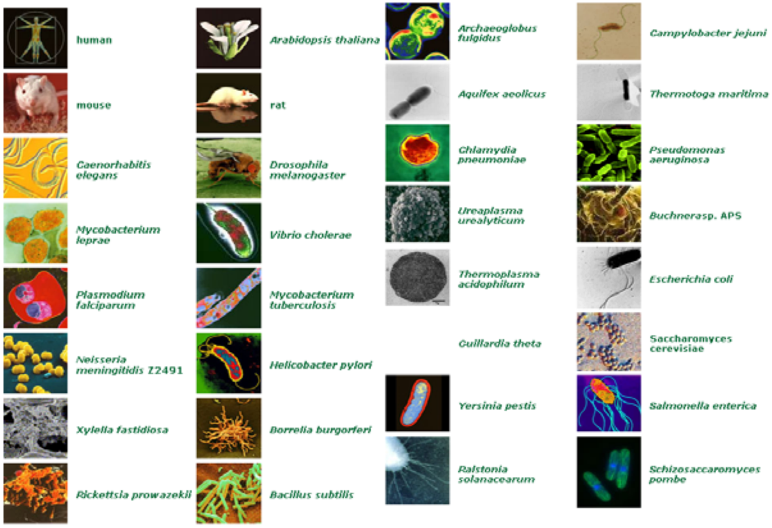
A Multi-resolution View of the Chromosome



The Human Genome



Genomes of many organisms are now "sequenced"



DNA Content of Representative Types of Cells

Organism	Number of base pairs (millions)	Number of encoded proteins	Number of chromosomes
PROKARYOTIC			
Mycoplasma genitalum **(Bacterium)**	0.58	470	1
Helicobacter pylori **(Bacterium)**	1.67	1590	1
Haemophilus influenza **(Bacterium)**	1.83	1743	1
EUKARYOTIC			
Saccharomyces cerevisiae **(yeast)**	12	5885	17
Drosophila melanogaster **(insect)**	165	13,601	4
Caenorhabditis elegans **(worm)**	97	19,099	6
Homo sapiens **(human)**	2900	30,000 TO 40,000	23
Arabidopsis thaliana (plant)	125	25,498	10

Functional Genomics

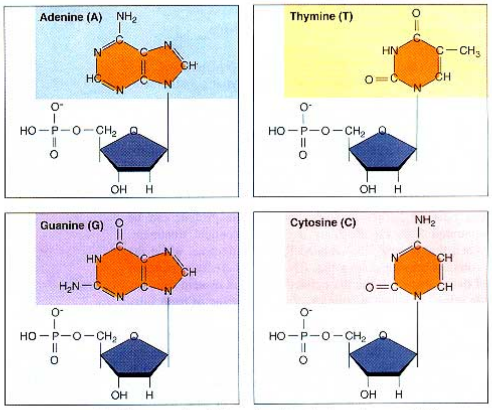
- The various **genome projects** have yielded the complete DNA sequences of many organisms.
  - E.g. human, mouse, yeast, fruitfly, etc.
  - Human: 3 billion base-pairs, 30-40 thousand genes.
- Challenge: **go from sequence to function**,
  - i.e., define the role of each gene and understand how the genome functions as a whole.

DNA

- DNA is made of 4 types of nucleotide subunit, where each nucleotide consists of
  - A phosphate group
  - A sugar group (deoxyribose)
  - One of the 4 bases: adenine (A), cytosine (C), guanine (G), and thymine (T)
- A DNA molecule consists of two strands witch form a “double helix”; each strand has a direction (5’ to 3’), and the strands are antiparallel.
- Watson-Crick basepairing
  - A in one strand always pairs with T in the other
  - And C always pairs with G

DNA

- DNA is made of 4 types of nucleotide subunit, where each nucleotide consists of
  - A phosphate group
  - A sugar group (deoxyribose)
  - One of the 4 bases:
    adenine (A),
    cytosine (C),
    guanine (G),
    and thymine (T)

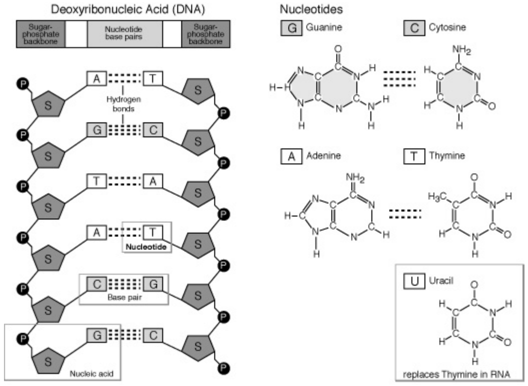


The four nucleotide subunits that make up DNA

The nucleotide subunits of DNA are composed of three elements: a central five-carbon sugar, a phosphate group, and an organic, nitrogen-containing base.

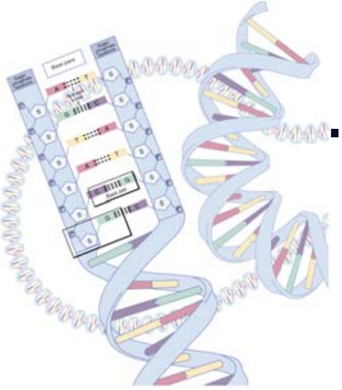
DNA

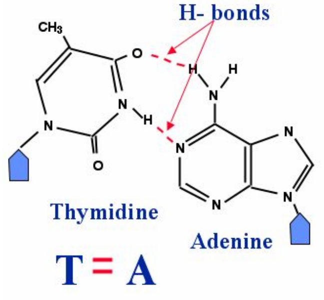
- A DNA molecule consists of two strands witch form a “double helix”; each strand has a direction (5’ to 3’), and the strands are antiparallel.
- Watson-Crick basepairing
  - A in one strand always pairs with T in the other
  - And C always pairs with G

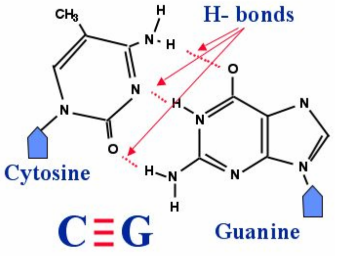


DNA

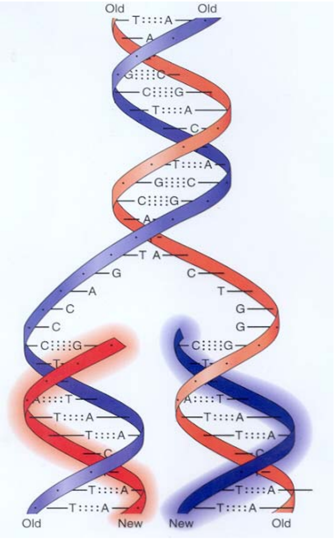
- Watson-Crick basepairing
  - A in one strand always pairs with T in the other
  - And C always pairs with G



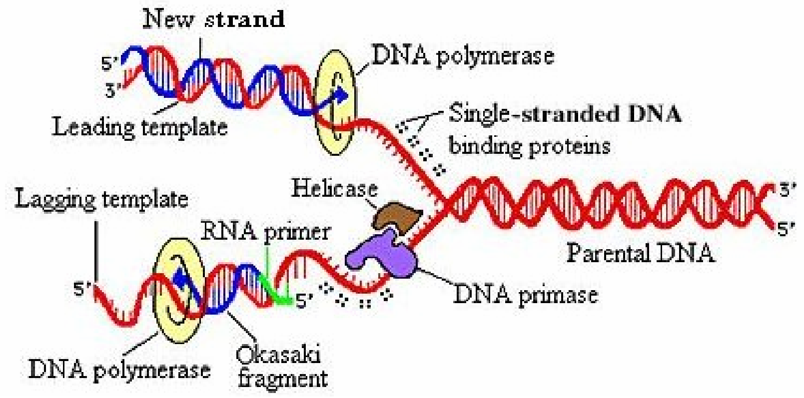




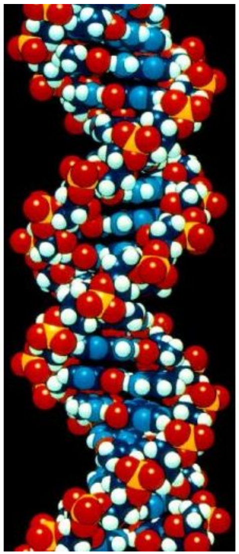
DNA is a self-replicating molecule



Collaborations of Proteins at the replication folk



Three-D structure of DNA



from http://www.dna-dna.net/

Writing DNA sequence

- One strand is written by listing its bases in 5’ to 3’ order

5’ACCGTTACT3’

- Each strand uniquely determines the complementary strand, witch runs in the opposite direction:

5’ACCGTTACT3’

- So the *reverse complementary* of ACCGTTACT is written AGTAACGGT

Some definitions

- Palindromes:

  A palindromic sequence is one which equals its reverse complement:

  e.g. GAATTA

- Length unit of DNAs
  - bp: basepair
  - Kb: 1000bp
  - Mb: 1000Kb
  - Gb: 1000Mb

"Central Dogma" of Molecular Biology: Information Flow within the Cell

$$\text{DNA} \xrightarrow{\text{transcription}} \text{RNA} \xrightarrow{\text{translation}} \text{PROTEIN}$$

RNA

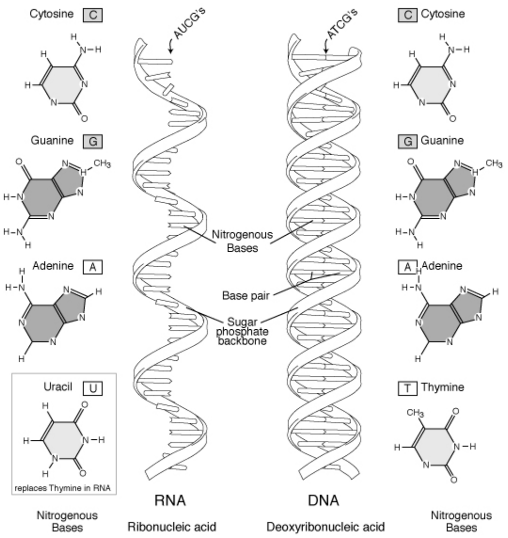
- RNA is similar to DNA, differing in the use of a different sugar (ribose) and in having the base *uracil* (U) in place of thymine

- RNA molecules are usually single stranded, but the strand typically base-pair with itself over parts of its length. As a result, it can adopt complex 3 dimensional structures

- There are three major types of RNAs
  - Messenger RNA (tRNA)
  - Transfer RNA (mRNA)
  - Ribosome RNA (rRNA)
  - Others:
    - Small interfering RNA

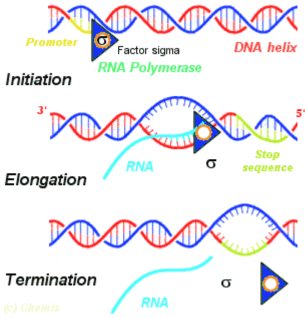
RNA

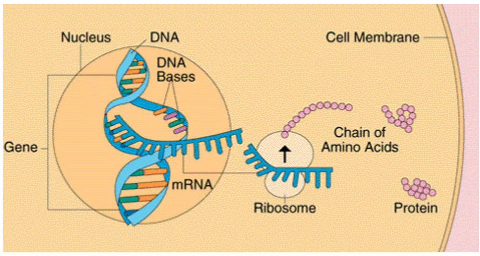
- Structure



RNA

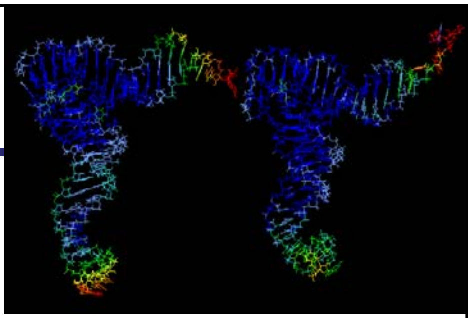
- mRNA

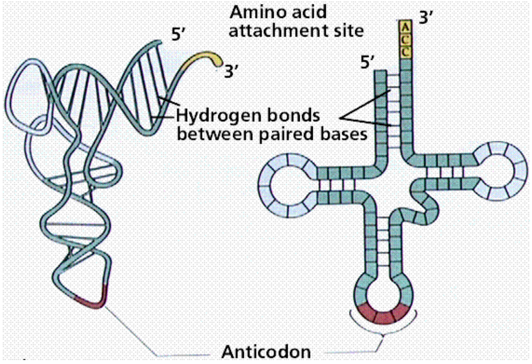




RNA

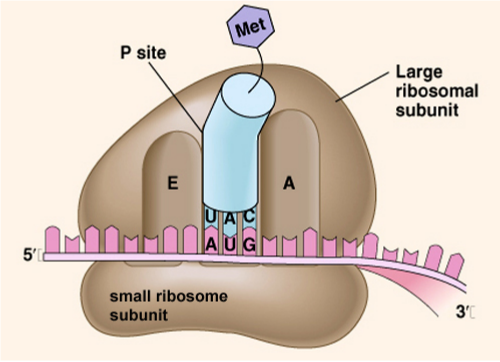
- tRNA

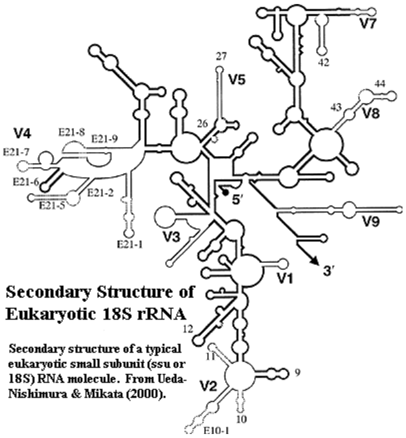




RNA

- rRNA





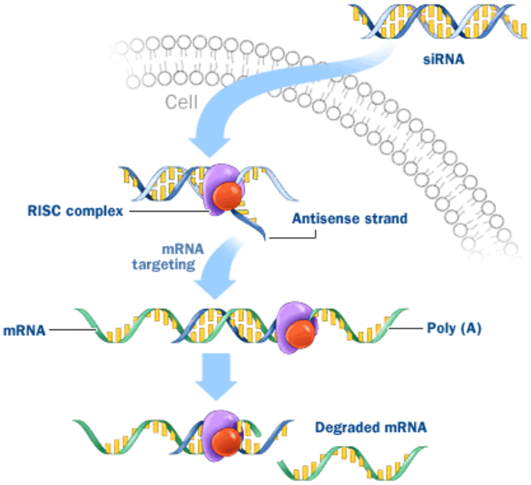
Secondary Structure of Eukaryotic 18S rRNA

Secondary structure of a typical eukaryotic small subunit (ssu or 18S) RNA molecule. From Ueda-Nishimura & Mikata (2000).

The Ribosome:

RNA

- siRNA



Transcription

- Transcription of DNA into RNA by RNA polymerase.

- **Also know as**

  “gene expression”

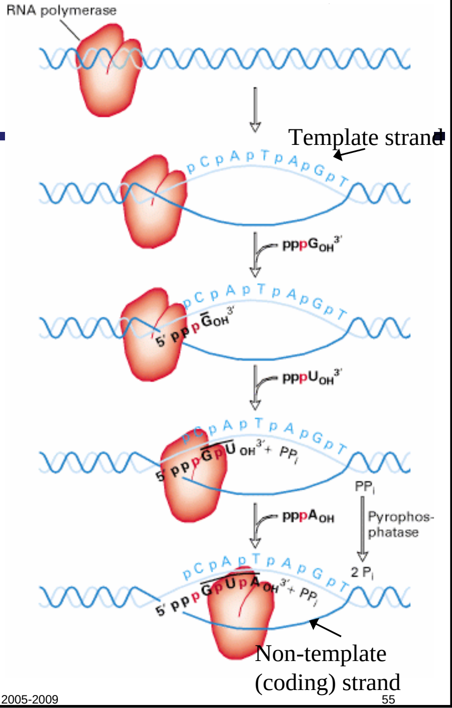
-

  1. Requires DNA template,

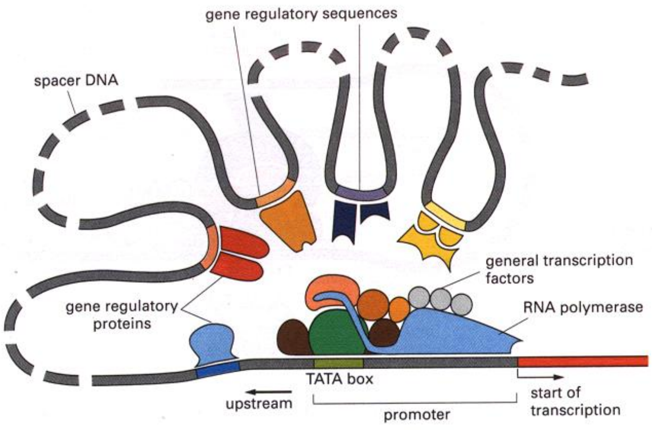
  2. De novo synthesis: does not require a primer.
  Low fidelity compared to DNA polymerase: errors 1/10^4-10^5. RNA polymerase incorporates ~30 nt/s (much slower than DNA polymerase).

  4. In most cases, only one DNA strand is transcribed into RNA.

  3. Activity highly regulated *in vivo*: at initiation, elongation and termination.

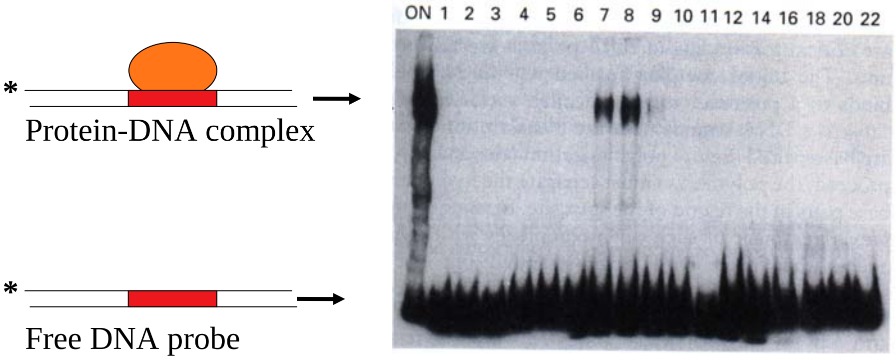


Regulatory Machinery of Gene Expression



Classical Analysis of Transcription Regulation Interactions

"Gel shift": electorphoretic mobility shift assay ("EMSA") for DNA-binding proteins

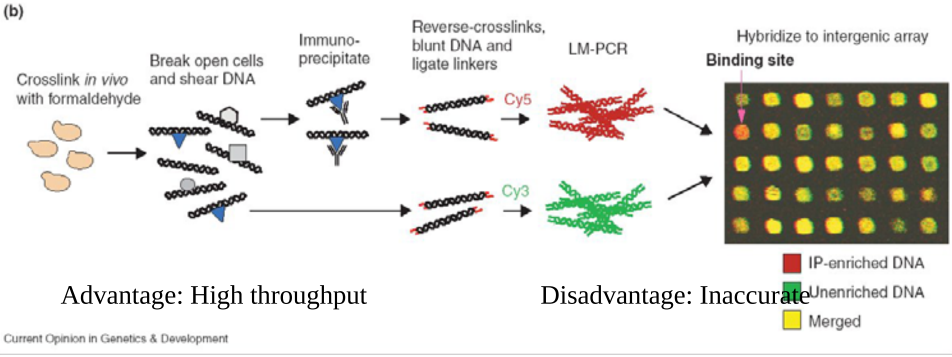


Advantage: sensitive

Disadvantage: requires stable complex; little "structural" information about which protein is binding

Modern Analysis of Transcription Regulation Interactions

- **Genome-wide Location Analysis (ChIP-chip)**



Advantage: High throughput

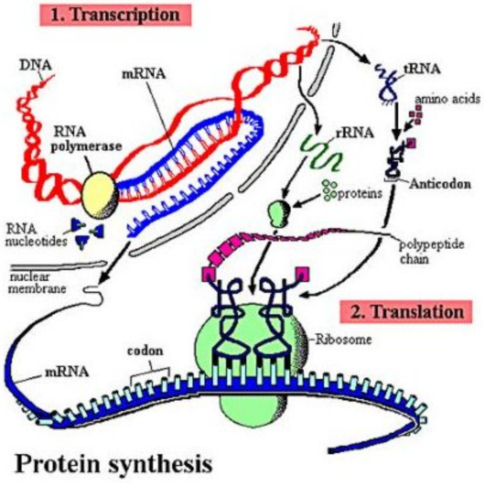
Disadvantage: Inaccurate

Current Opinion in Genetics & Development

“Central Dogma” of Molecular Biology: Information Flow within the Cell

$$\text{DNA} \xrightarrow{\text{transcription}} \text{RNA} \xrightarrow{\text{translation}} \text{PROTEIN}$$

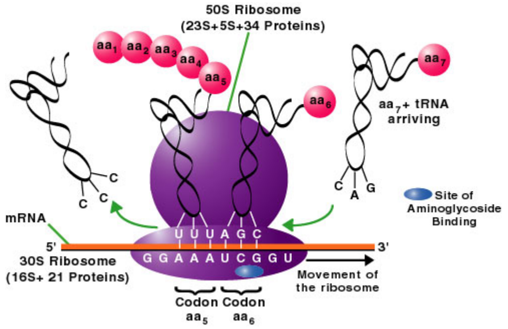
Protein Synthesis: Translation

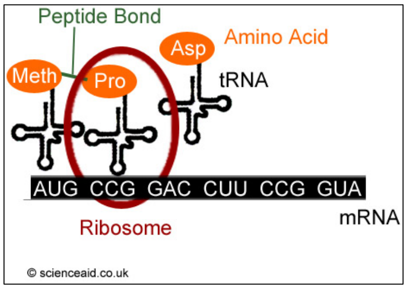


Protein synthesis

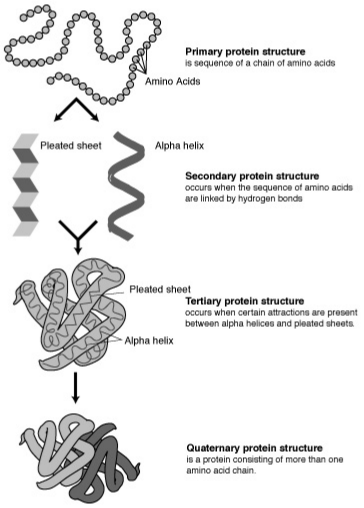
Copyright 1999 Access Excellence @ the National Health Museum. All rights reserved.

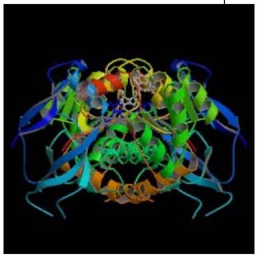
Translation

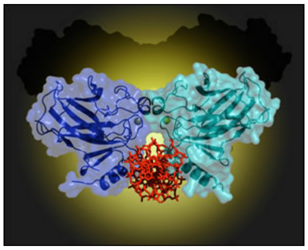




Protein

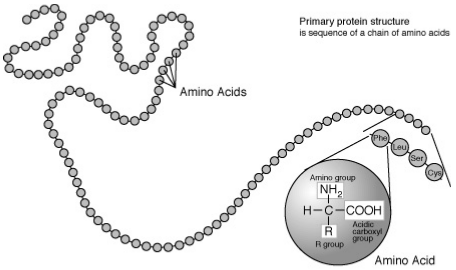


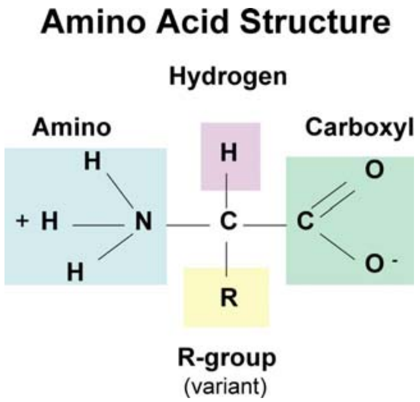




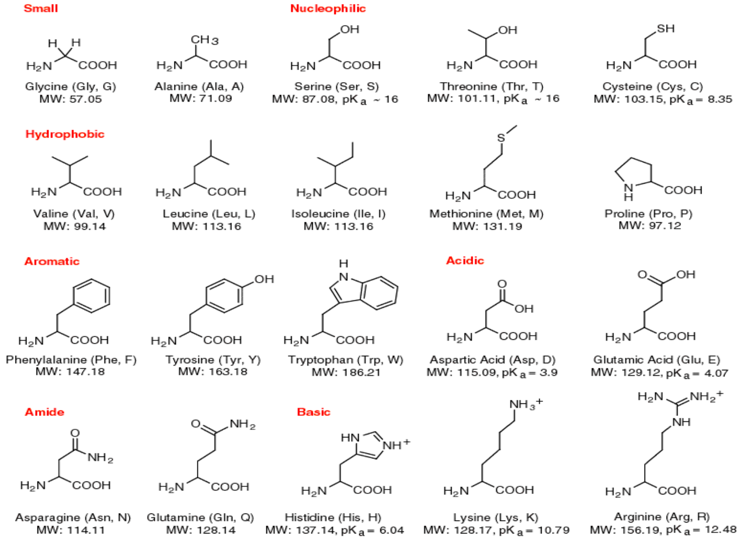
The structure of p53

Amino Acid





Amino Acid Alphabet



Polypeptide

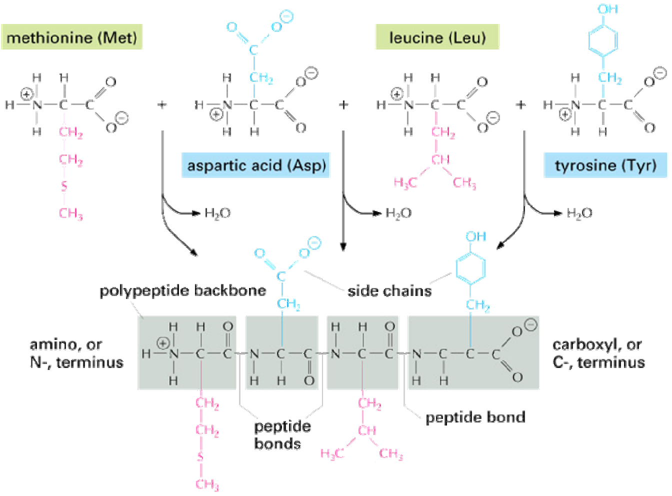
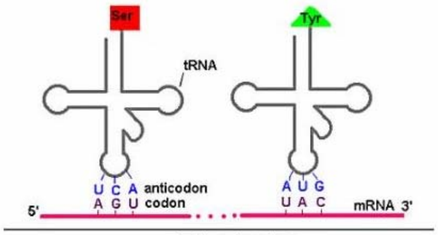
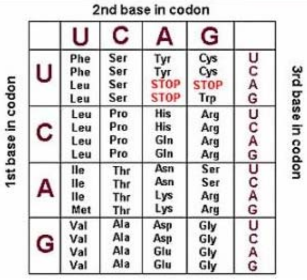


Figure 4-2 Essential Cell Biology, 2/e. (© 2004 Garland Science)

The Genetic Codon





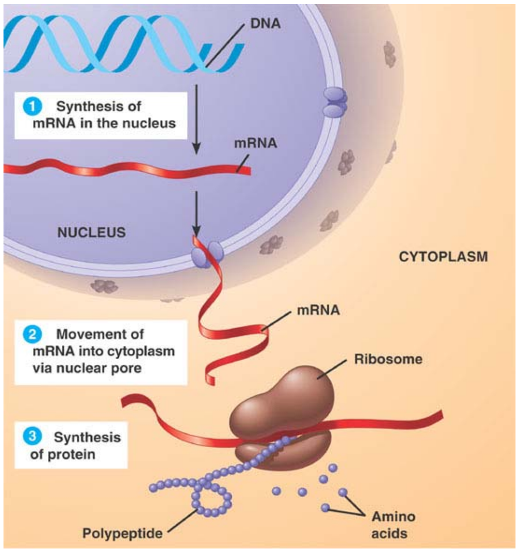
2nd base in codon

1st base in codon

3rd base in codon

	U	C	A	G	
U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G
C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G

Summary: Cell Information Flow



Summary: Control of Gene Exp.