Computational Genomics

10-810/02-710, Spring 2009

Biology Background II

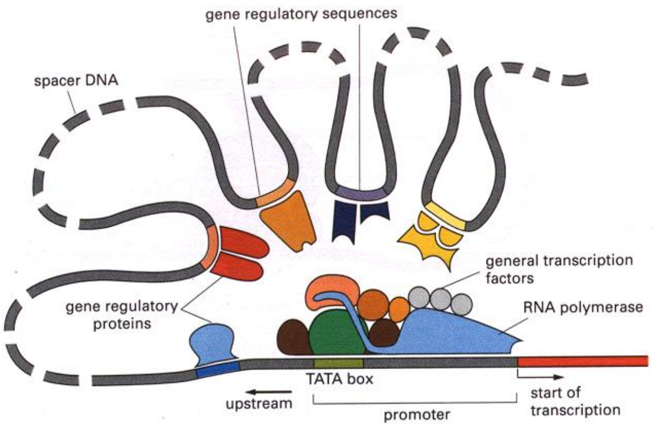
Molecular Biology, Development, and Genetics

Eric Xing

Lecture 2, January 14, 2009

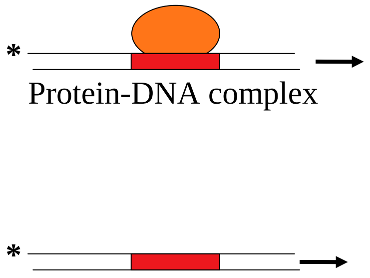
Reading: Chap. 1, DTM book

Regulatory Machinery of Gene Expression



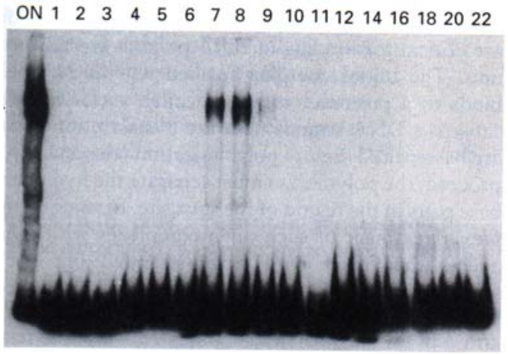
Classical Analysis of Transcription Regulation Interactions

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



Protein-DNA complex

Free DNA probe



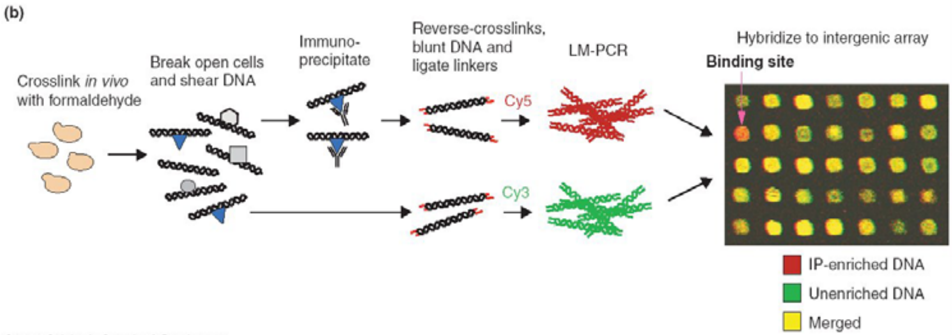
Advantage: sensitive

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

---

Modern Analysis of Transcription Regulation Interactions

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

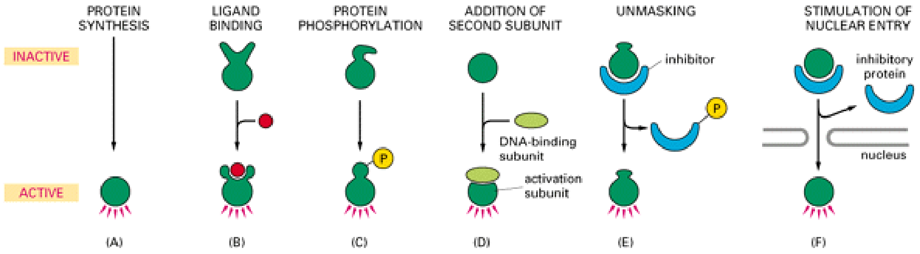


Current Opinion in Genetics & Development

Advantage: High throughput

Disadvantage: Inaccurate

Ways to activate a gene regulation protein

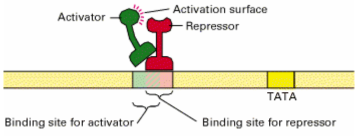


Examples:
GATA-1 CAP NtrC Adenovirus E1A + CBP/p300 NF-KB/ glucocorticoid receptor

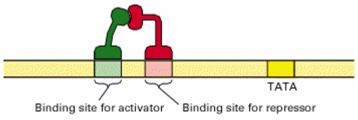
---

Inhibition by steric mechanisms

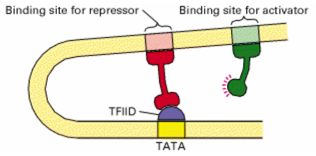
- Competitive binding with activator



- Interaction with activation domain of bound activator

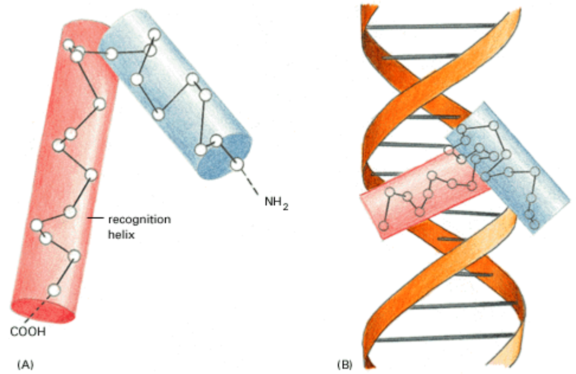


- Interaction with general TF



Binding Site Recognition

- The DNA-binding helix-turn-helix motif.



The motif is shown in (A), where each *white* circle denotes the central carbon of an amino acid. The carboxyl-terminal a helix (*red*) is called the recognition helix because it participates in sequence-specific recognition of DNA. As shown in (B), this helix fits into the major groove of DNA, where it contacts the edges of the base pairs.

Insertion of an α-helix into the major groove

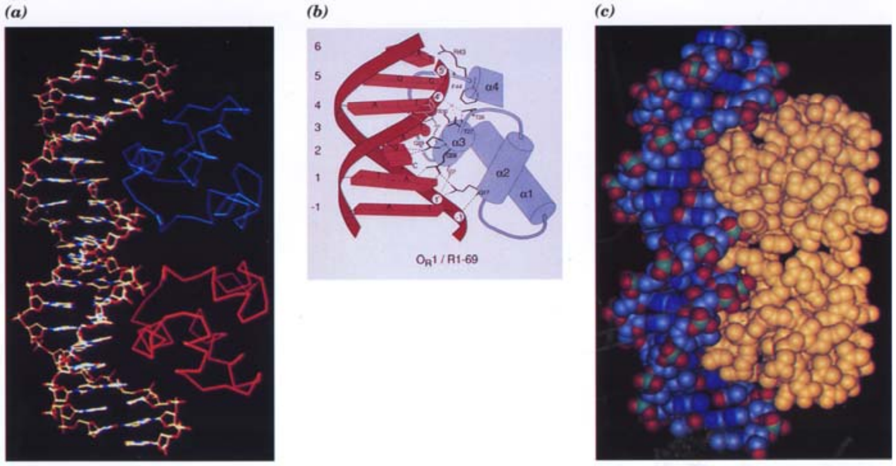
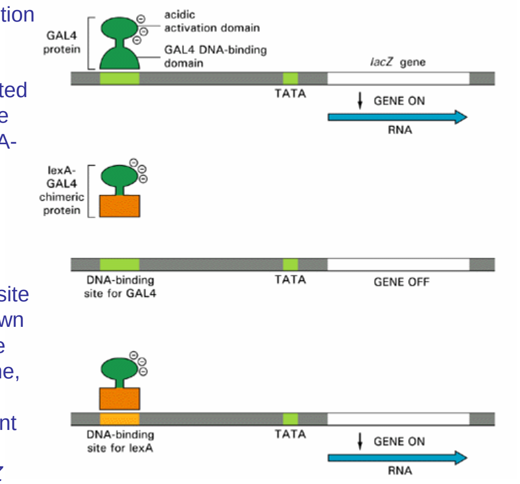


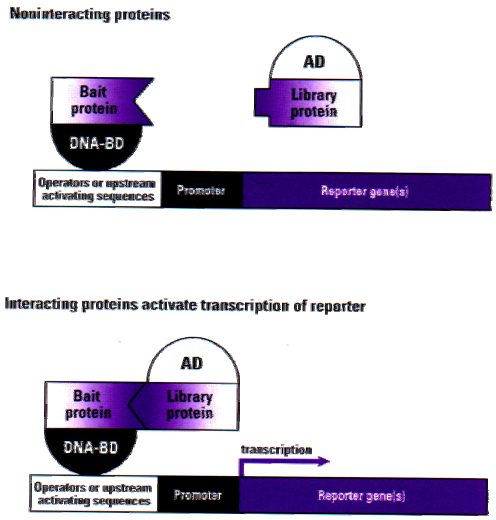
Figure 23-32. **The X-ray structure of a portion of the 434 phage repressor in complex with its target DNA.** [Courtesy of Aneel Aggarwal, John Anderson, and Stephen Harrison, Harvard University.]

The modular structure of a gene activator protein.

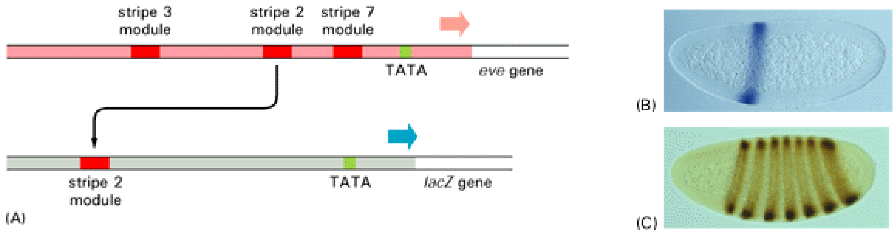
- The normal activation of gene transcription produced by the GAL4 protein.
- A functional activator can be reconstituted from the carboxyl-terminal portion of the GAL4 protein if it is attached to the DNA-binding domain of a bacterial gene regulatory protein (the lexA protein) by gene fusion techniques.
- The chimeric gene regulatory protein requires the lexA-protein DNA-binding site for its activity. For the experiments shown here, the control region for one of these genes was fused to the *E. coli lacZ* gene, which codes for the enzyme b-galactosidase and provides a convenient way to monitor the expression level specified by a gene control region; *lacZ* thus serves as a *reporter gene*.



"Two-hybrid" system for *in vivo* analysis of protein-protein interaction



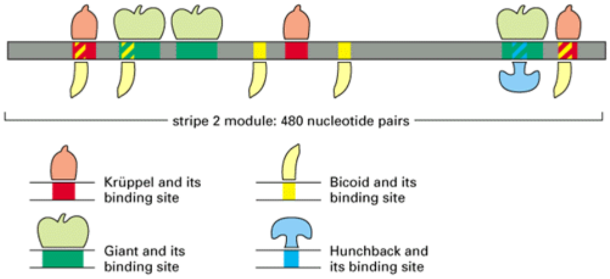
Cis-Regulatory Modules (enhancers)



Experiment demonstrating the modular construction of the *eve* gene regulatory region.
(A) A 480-nucleotide-pair piece of the *eve* regulatory region was removed and inserted upstream of a test promoter that directs the synthesis of the enzyme β-galactosidase (the product of the *E. coli lacZ* gene). (B) When this artificial construct was reintroduced into the genome of *Drosophila* embryos, the embryos expressed β-galactosidase (detectable by histo-chemical staining) precisely in the position of the second of the seven *eve* stripes (C). (B and C, courtesy of Stephen Small and Michael Levine.)

Enhancers

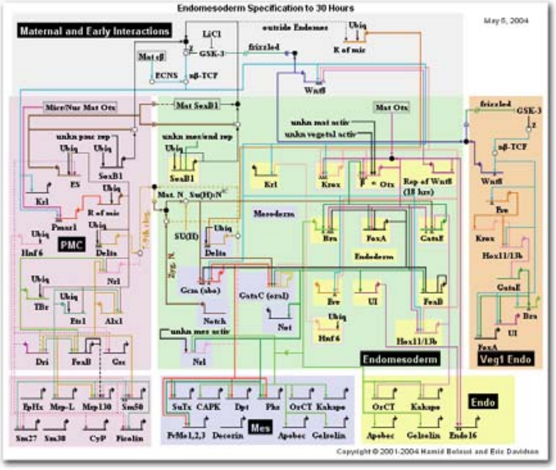
- Enhancers are comprised of binding sites for multiple regulatory proteins.



Close-up view of the *eve* stripe 2 unit. The segment of the *eve* gene control region contains regulatory sequences, each of which binds one or another of four gene regulatory proteins. It is known from genetic experiments that these four regulatory proteins are responsible for the proper expression of stripe 2 of *eve*. In some cases, the binding sites for the gene regulatory proteins overlap and the proteins can compete for binding to the DNA.

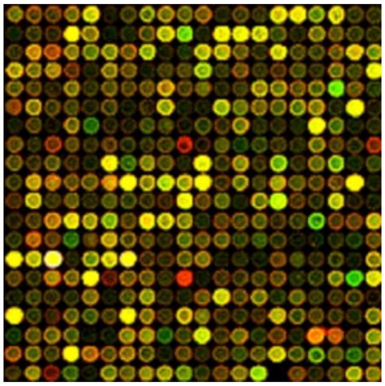
Gene Regulatory Network

- Direct experimental determination?

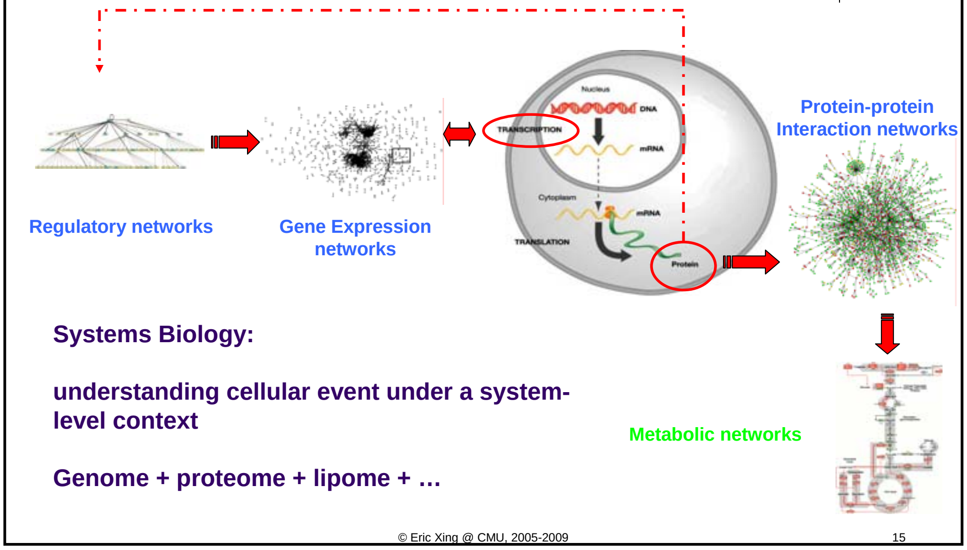


Microarray or Whole-body ISH?

- Inference from gene expression profiles



Biological Networks and Systems Biology



Regulatory networks

Gene Expression networks

Protein-protein Interaction networks

Metabolic networks

Systems Biology:

understanding cellular event under a system-level context

Genome + proteome + lipome + …

Gene Regulatory Functions in Development

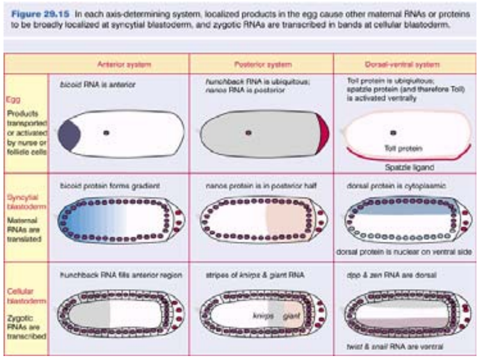
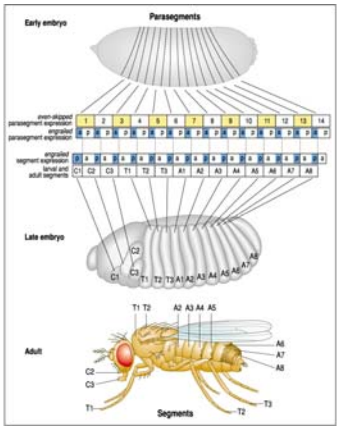
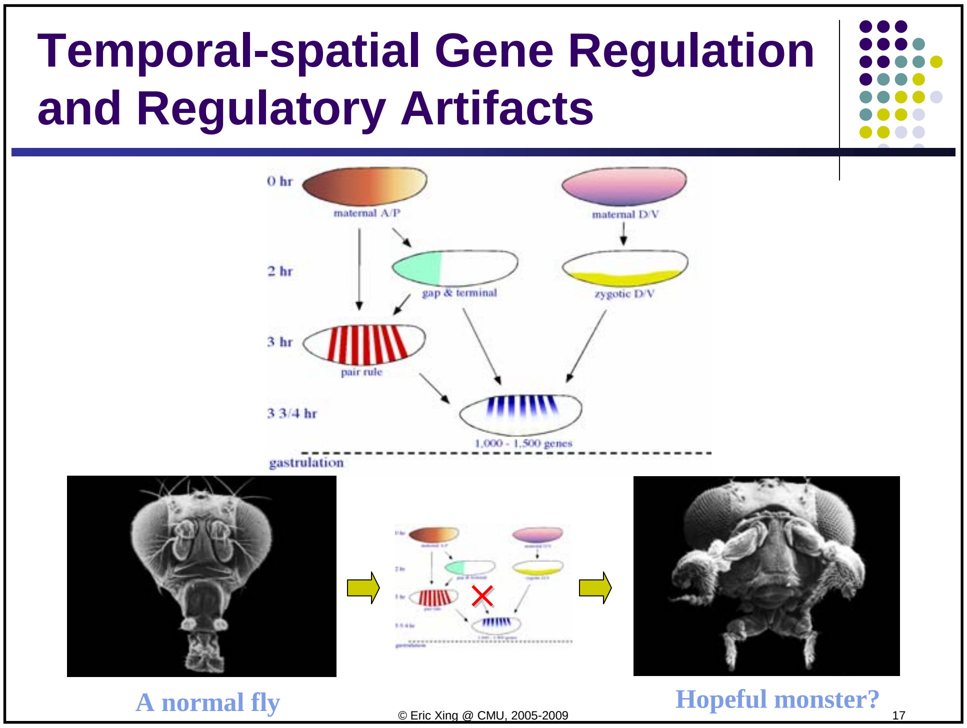


Figure 29.15 In each axis-determining system, localized products in the egg cause other maternal RNAs or proteins to be broadly localized at syncytial blastoderm, and zygotic RNAs are transcribed in bands at cellular blastoderm.

	Anterior system	Posterior system	Dorsal-ventral system
Egg Products transported or activated by nurse or follicle cells	*bicoid* RNA is anterior	*hunchback* RNA is ubiquitous; *nanos* RNA is posterior	Toll protein is ubiquitous; spatzle protein (and therefore Toll) is activated ventrally Toll protein Spatzle ligand
Syncytial blastoderm Maternal RNAs are translated	bicoid protein forms gradient	nanos protein is in posterior half	dorsal protein is cytoplasmic dorsal protein is nuclear on ventral side
Cellular blastoderm Zygotic RNAs are transcribed	*hunchback* RNA fills anterior region	stripes of *knirps* & *giant* RNA knirps giant	*dpp* & *zen* RNA are dorsal *twist* & *snail* RNA are ventral



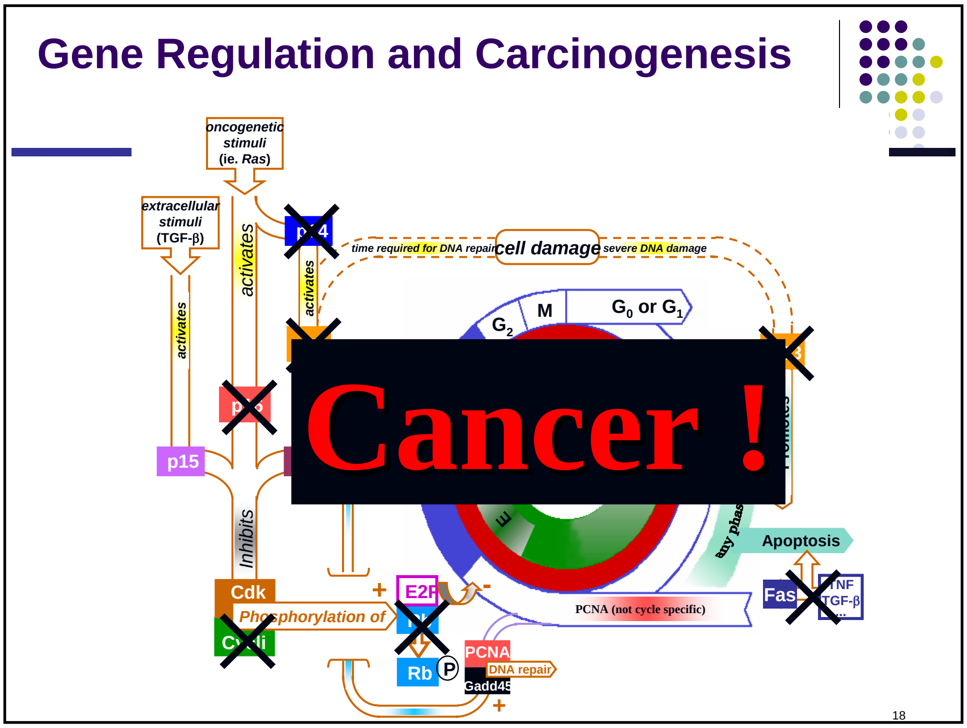


Temporal-spatial Gene Regulation and Regulatory Artifacts

A normal fly

Hopeful monster?

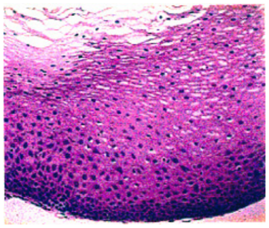
© Eric Xing @ CMU, 2005-2009



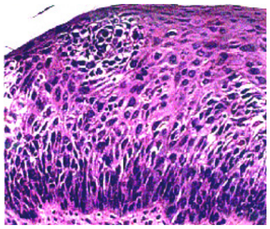
Gene Regulation and Carcinogenesis

Cancer !

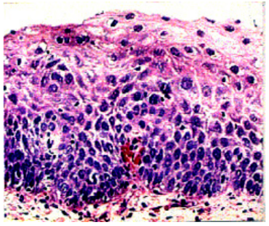
The Pathogenesis of Cancer



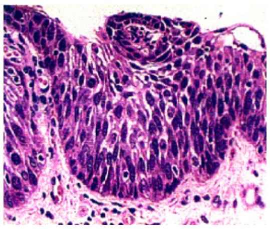
Normal



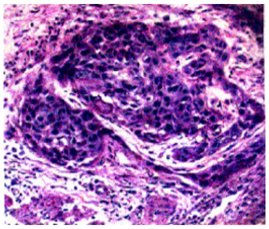
BCH



DYS

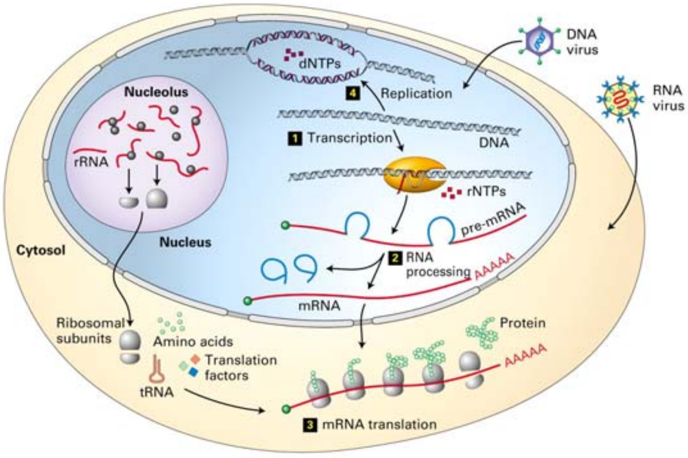


CIS

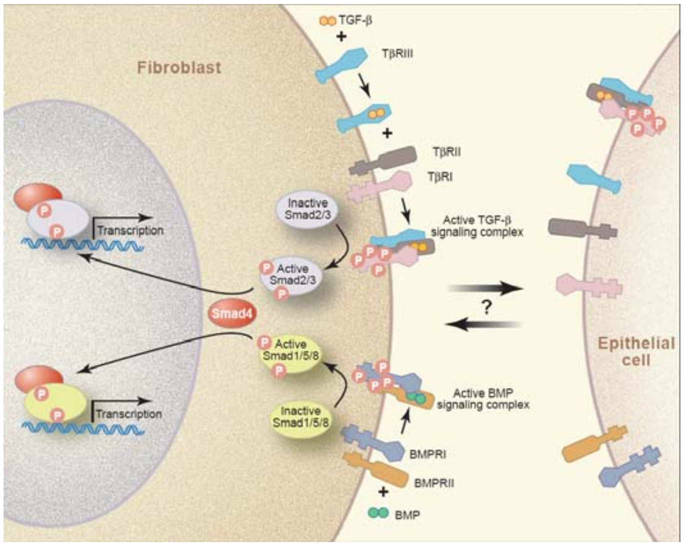


SCC

The Information Flow in Cell

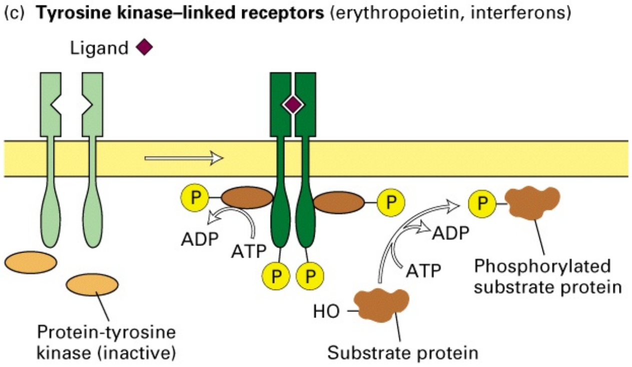


Genome-environment interactions

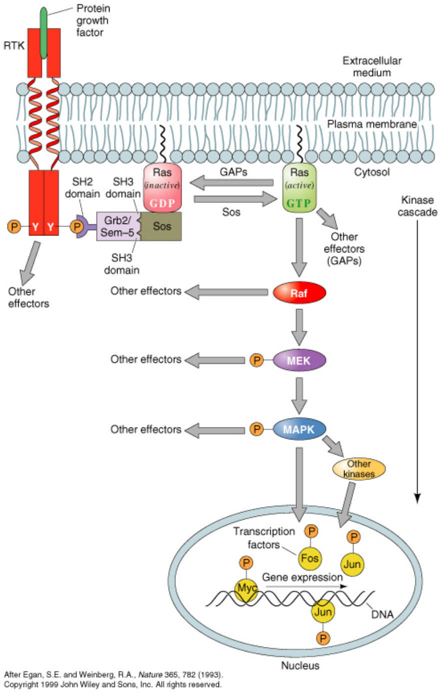


Signal Transduction

- A variety of plasma membrane receptor proteins bind extracellular signaling molecules and transmit signals across the membrane to the cell interior


(c) **Tyrosine kinase–linked receptors** (erythropoietin, interferons)

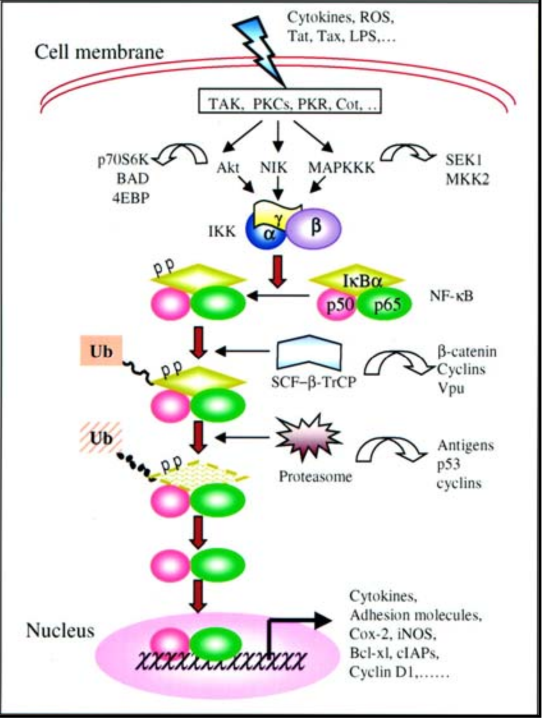
Signal Transduction Pathway



After Egan, S.E. and Weinberg, R.A., *Nature* 365, 782 (1993).

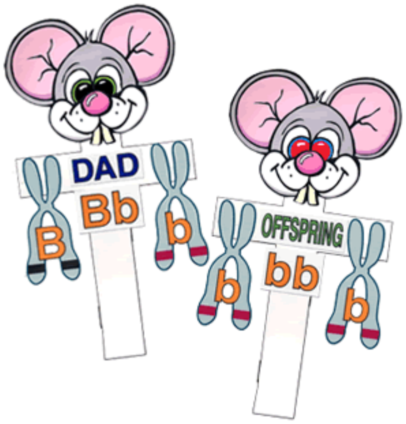
Signaling pathways of NF-κB activation.

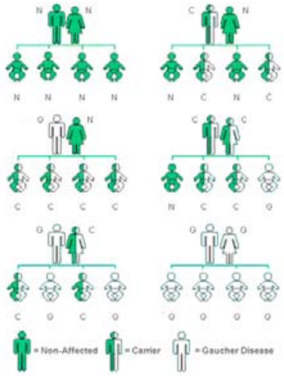
Extracellular inducers, including cytokines, reactive oxygen species (ROS), and viral and bacterial products, activate IKK through upstream kinases directly or indirectly. Activated IKK phosphorylates N-terminal S32 and S36 residues of IB- that is associated with NF-B p50 and p65 heterodimer. The SCF-ß-TrCP complex recognizes phosphorylated IκB- and modifies IκB- with polyubiquitin chains. This is followed by proteasome-mediated degradation of IκB-. After degradation of IκB-, the activated NF-κB translocates into the nucleus where it binds to the κB-sites of gene promoters or enhancers to up-regulate target gene expression. **Line arrows** and **filled arrows** denote the NF-κB signaling pathways; **open arrows** denote the connections with the by-standing signaling pathways.



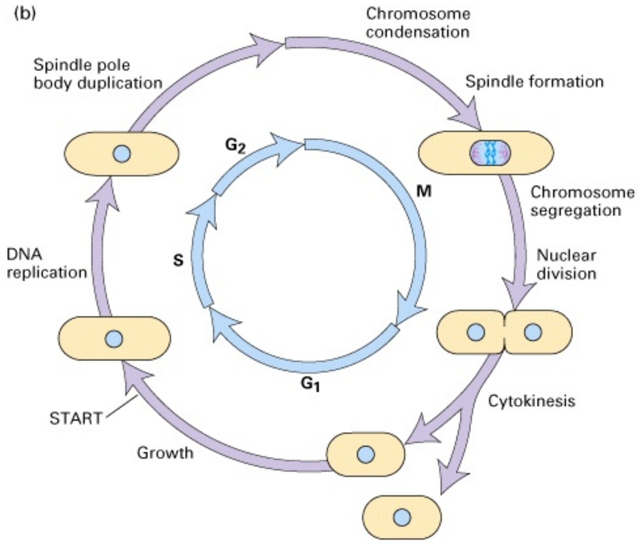
Chen F, Castranova V, Shi X.
New insights into the role of nuclear factor-kappaB in cell growth regulation.
Am J Pathol. 2001 Aug;159(2):387-97.

The Information Flow through Generations

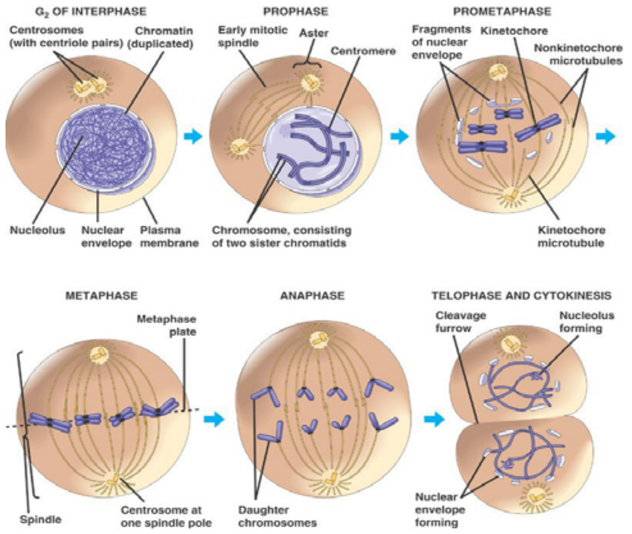




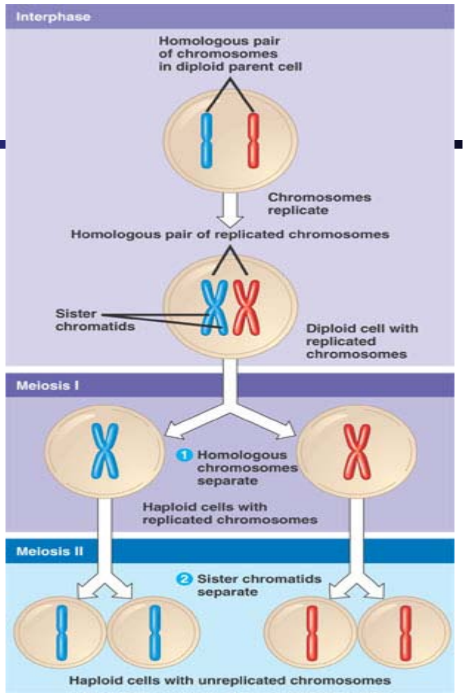
Cell Cycle



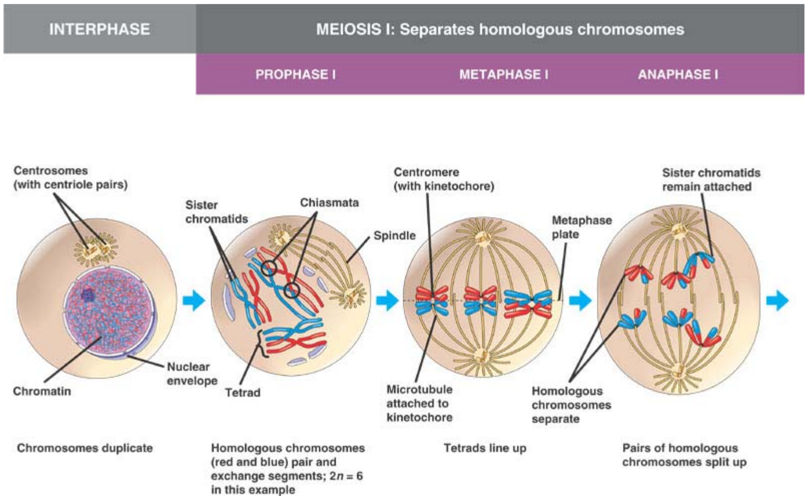
Mitosis



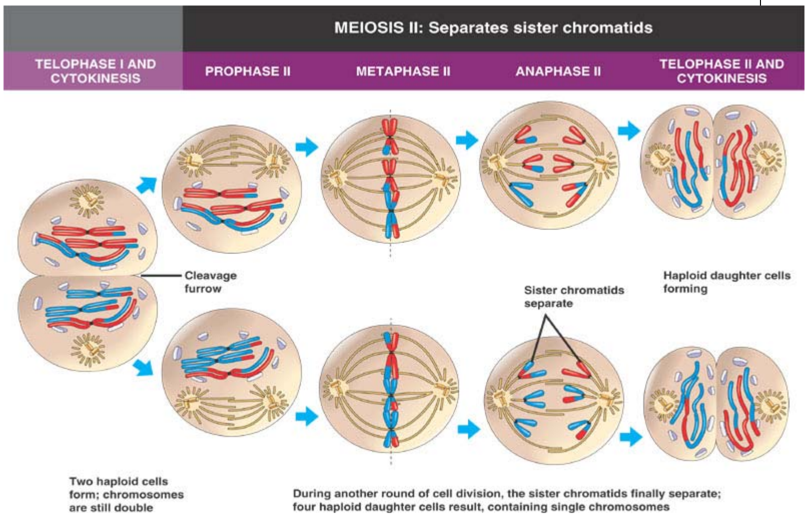
Meiosis



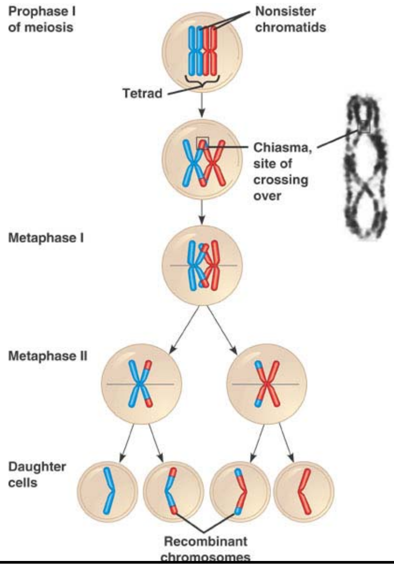
Meiosis I



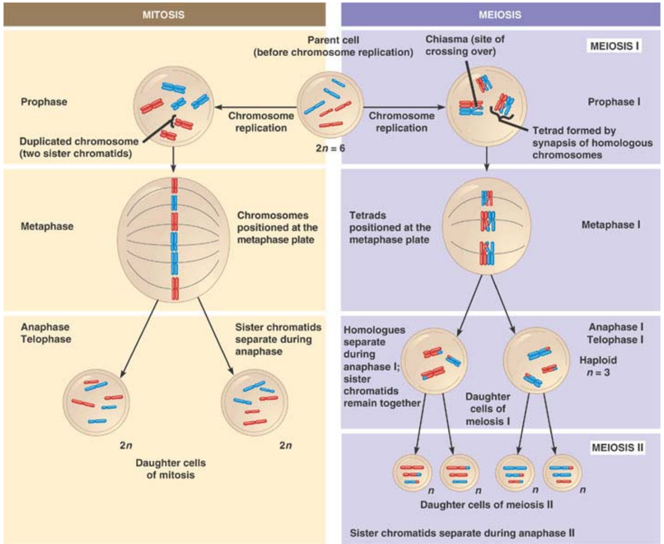
Meiosis II



Crossover



Mitosis and Meiosis Compared



Haploid, Diploid Cycle

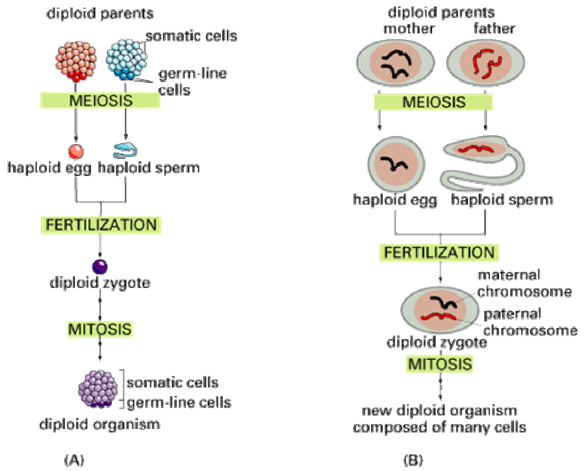
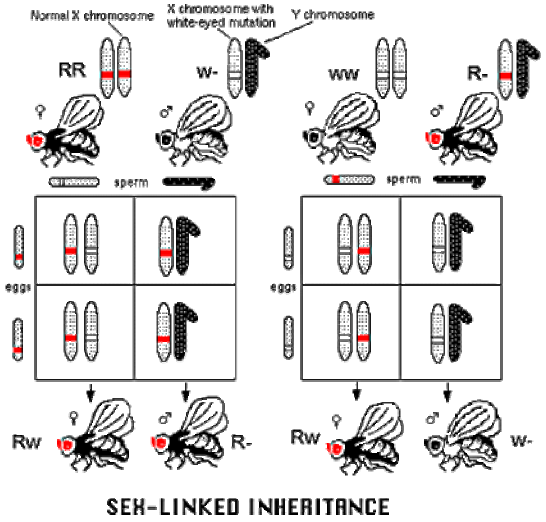


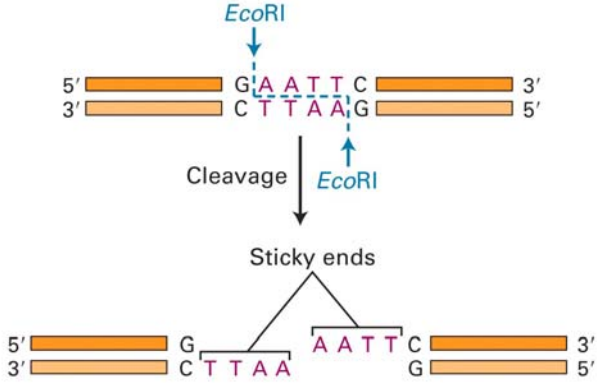
Figure 20-4 Essential Cell Biology, 2/e (© 2004 Garland Science)

Sex-Linked Inheritance



Genetic Engineering: Manipulating the Genome

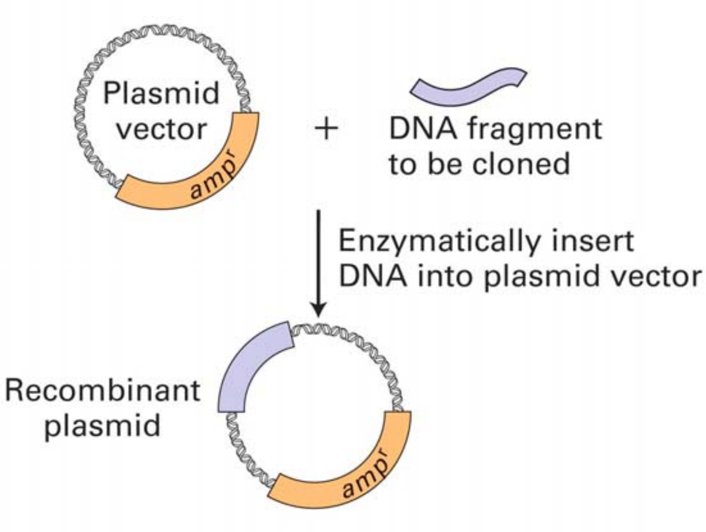
- Restriction Enzymes, naturally occurring in bacteria, that cut DNA at very specific places.



Definitions

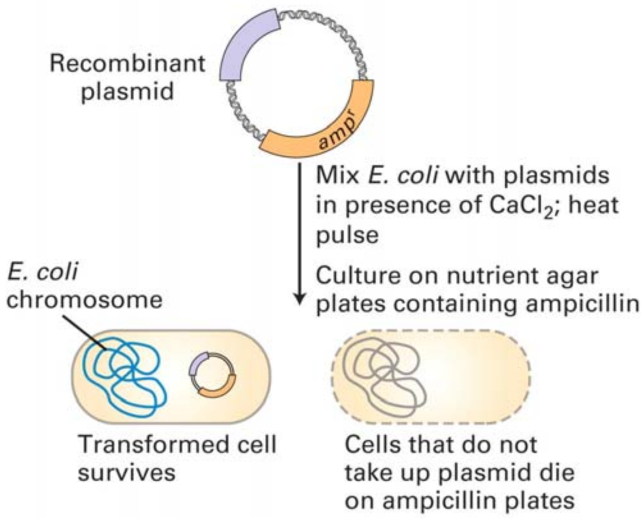
- Recombinant DNA: Two or more segments of DNA that have been combined by humans into a sequence that does not exist in nature.
- Cloning: Making an exact genetic copy. A **clone** is one of the exact genetic copies.
- Cloning vector: Self-replicating agents that serve as vehicles to transfer and replicate genetic material.

Recombinant DNA



Enzymatically insert DNA into plasmid vector

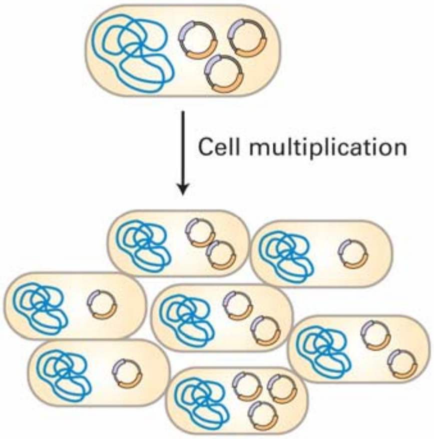
Transformation



Mix *E. coli* with plasmids in presence of $CaCl_2$; heat pulse

Culture on nutrient agar plates containing ampicillin

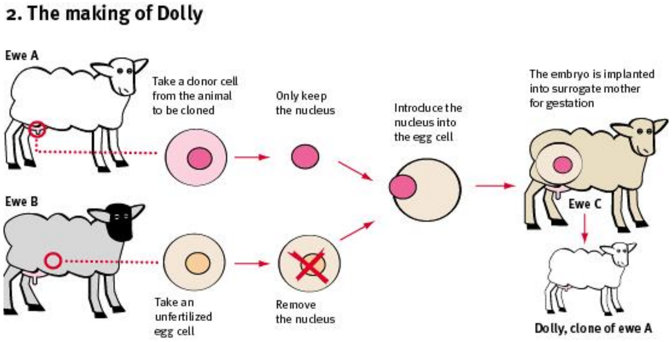
Formation of Cell Colony


Cell multiplication

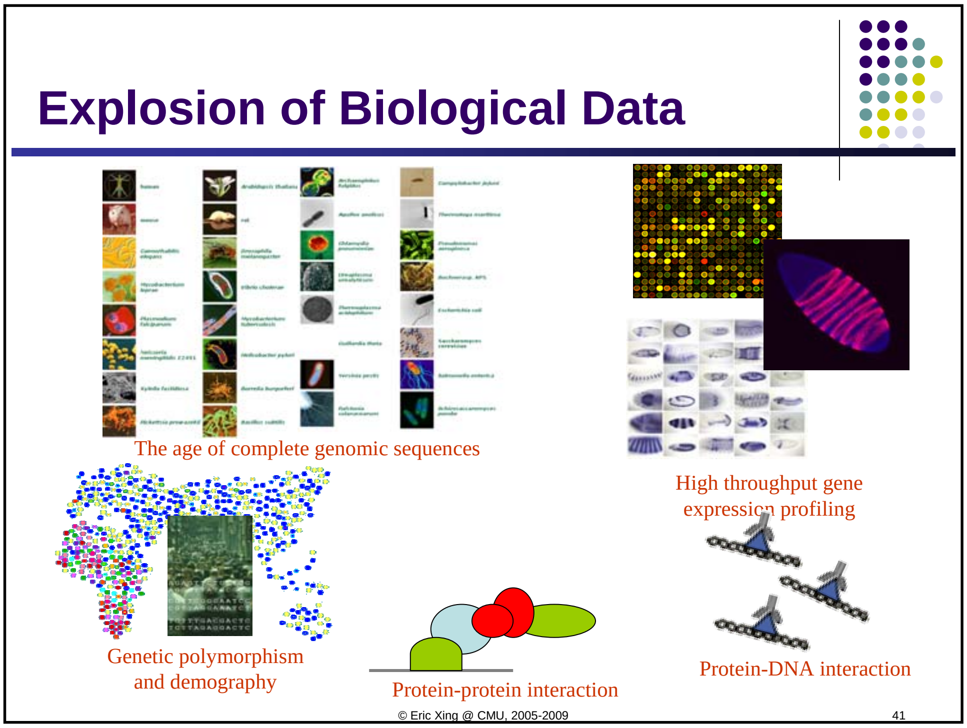
Colony of cells, each containing copies of the same recombinant plasmid

How was Dolly cloned?

- Dolly is claimed to be an exact genetic replica of another sheep.



- Is it exactly "exact"?



Explosion of Biological Data

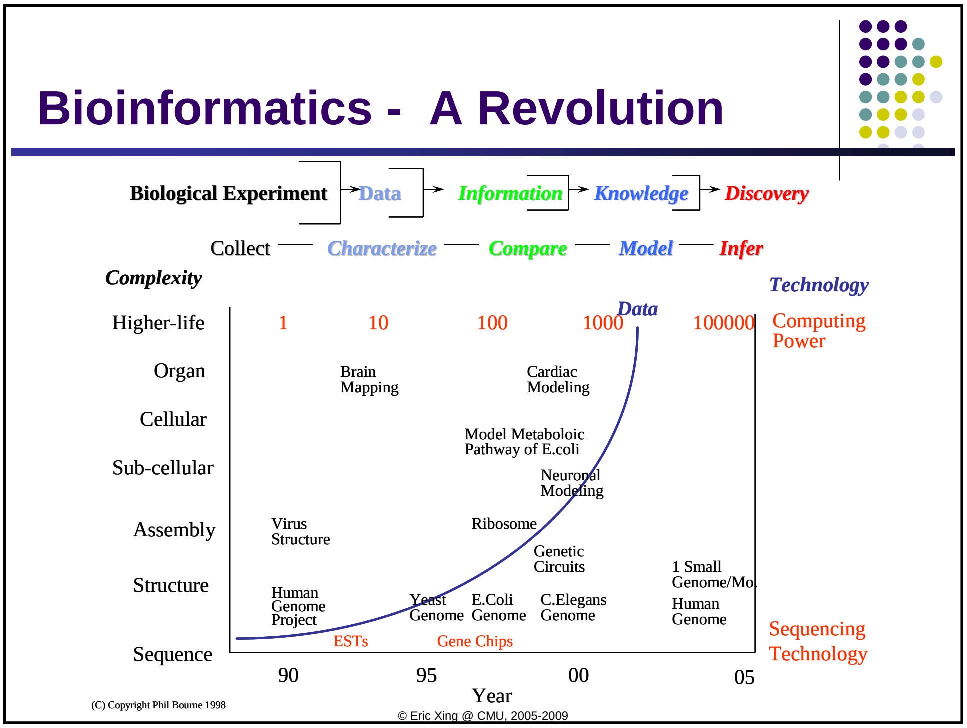
The age of complete genomic sequences

High throughput gene expression profiling

Genetic polymorphism and demography

Protein-protein interaction

Protein-DNA interaction



Bioinformatics - A Revolution

Biological Experiment → *Data* → *Information* → *Knowledge* → *Discovery*

Collect — *Characterize* — *Compare* — *Model* — *Infer*

Complexity

Technology

Higher-life

Organ

Cellular

Sub-cellular

Assembly

Structure

Sequence

1 10 100 1000 100000

Data

Computing Power

Brain Mapping

Cardiac Modeling

Model Metaboloic Pathway of E.coli

Neuronal Modeling

Virus Structure

Ribosome

Genetic Circuits

1 Small Genome/Mo.

Human Genome Project

Yeast Genome

E.Coli Genome

C.Elegans Genome

Human Genome

ESTs

Gene Chips

Sequencing Technology

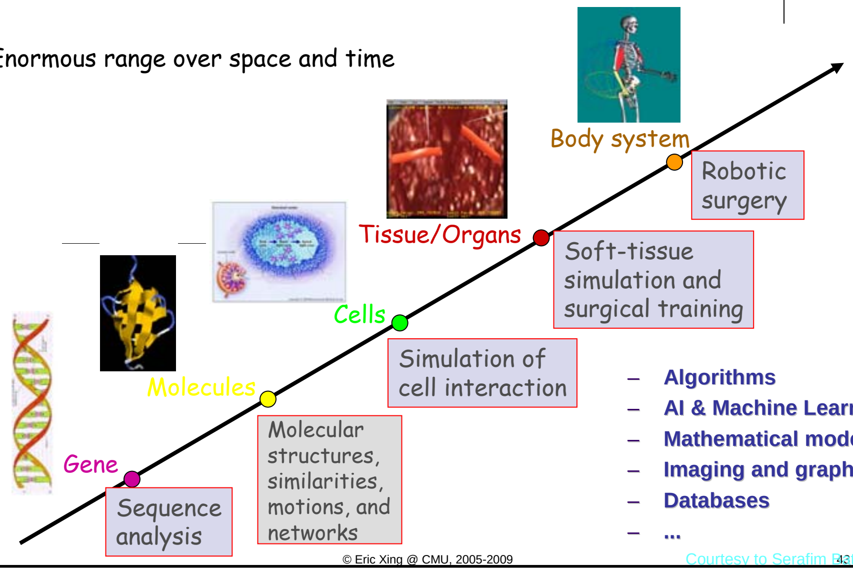
90 95 00 05

Year

(C) Copyright Phil Bourne 1998

Bio-CS Interaction As Of Now

Enormous range over space and time



- Algorithms
- AI & Machine Learning
- Mathematical modeling
- Imaging and graphics
- Databases
- ...

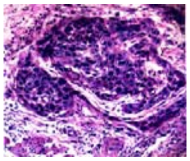
Courtesy to Serafim Batzoglou

Grand Challenge I

Understanding and Modeling

Morphogenesis and Tumorigenesis

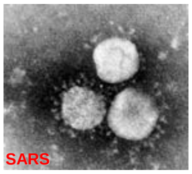
Create a monster !?



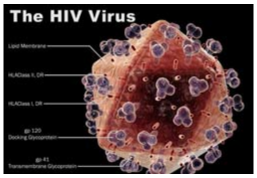
Grand Challenge II

Containing
and/or
Out-Pacing/Smarting

Fast Evolving Systems


SARS

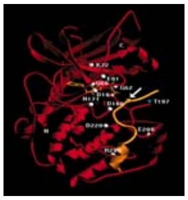

Influenza Virus

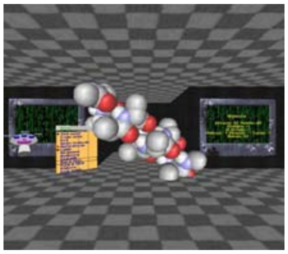

The HIV Virus

Grand Challenge III

Computer Aided

Molecular Drug Design

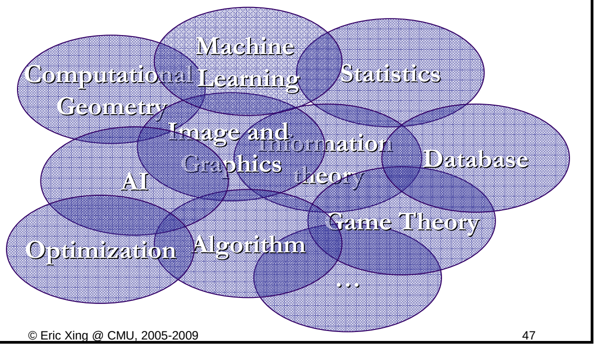

Structure determination


Molecular Simulation


Drug design and trial

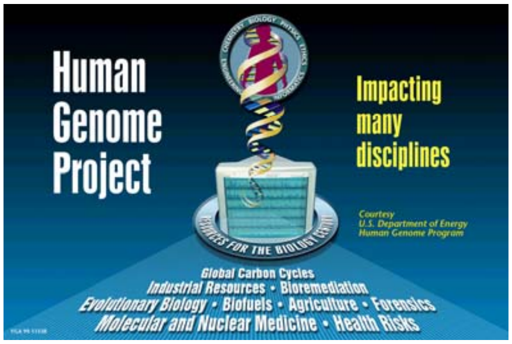
Embedded CS/Math Challenges

- Modeling, interpreting, visualizing and predicting high-dimensional, relational, and often unobservable data
  - The control program (network inference)
  - Predicting temporal-spatial behavior (multi-variant dynamic system)
  - Control and robustness (perturbation analysis)
  - Communication protocol
  - Decentralized computing
  - Tissue geometry
  - Games
  - …



Summary

- Biology is becoming an **information science**

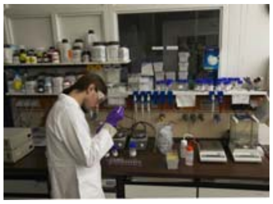


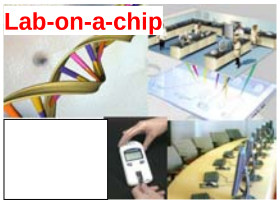
Courtesy U.S. Department of Energy Human Genome Program

- The real heroes are the computer scientists !!!
  - Reading-off the characters are now lab-routine, but assembling countless reads and interpreting the reads needs computers, but more importantly, **algorithms and models**

Summary

- Need new and powerful Computer **Systems** and **Theories** to:
  - Help **unlock** the **laws** and **trends** of living systems from the deluge of experimental data
  - **Reshape** basic biological research, make it more quantitative, rigorous and efficient

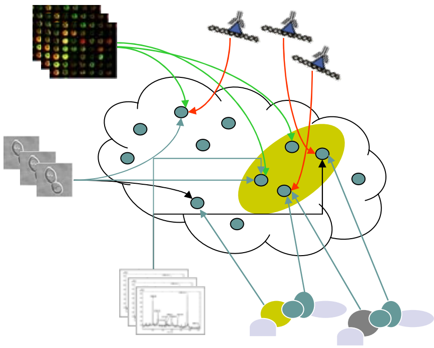




  - **Combat** diseases and prevent dysfunctions

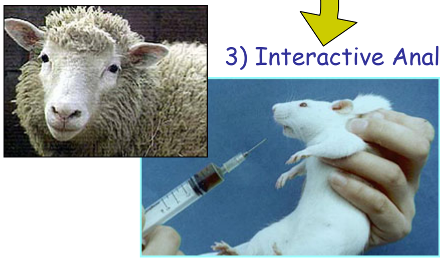
Conclusion

1) Extendable Models



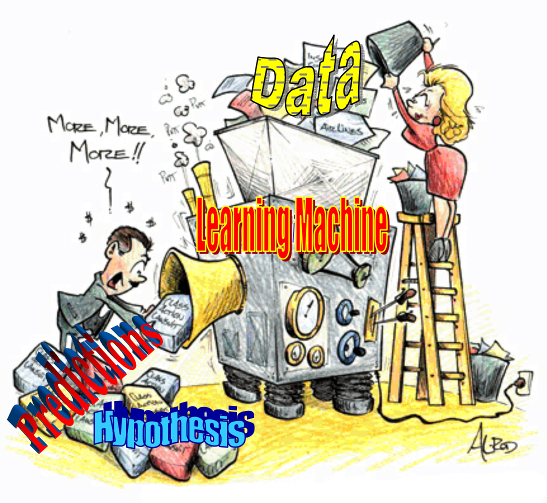
2) Effective Algorithm and Simulators

3) Interactive Analysis



4) Better medicine and experiments

A Machine Learning Approach to Computational Molecular Biology



Using
mathematical models
and
computational reasoning
to pursue
predictive understanding
of life