Computational Genomics

http://www.cs.cmu.edu/~02710

Ziv Bar-Joseph
zivbj@cs.cmu.edu
GHC 8006

Chakra Chennubhotla
chakracs@pitt.edu
Suite 3064, BST3

Topics

- Introduction (1 Week)
- Sequence analysis(4 weeks)
- Gene expression (3 weeks)
- RNA and epigenetics (3 weeks)
- Systems biology (3 weeks)

Class overview

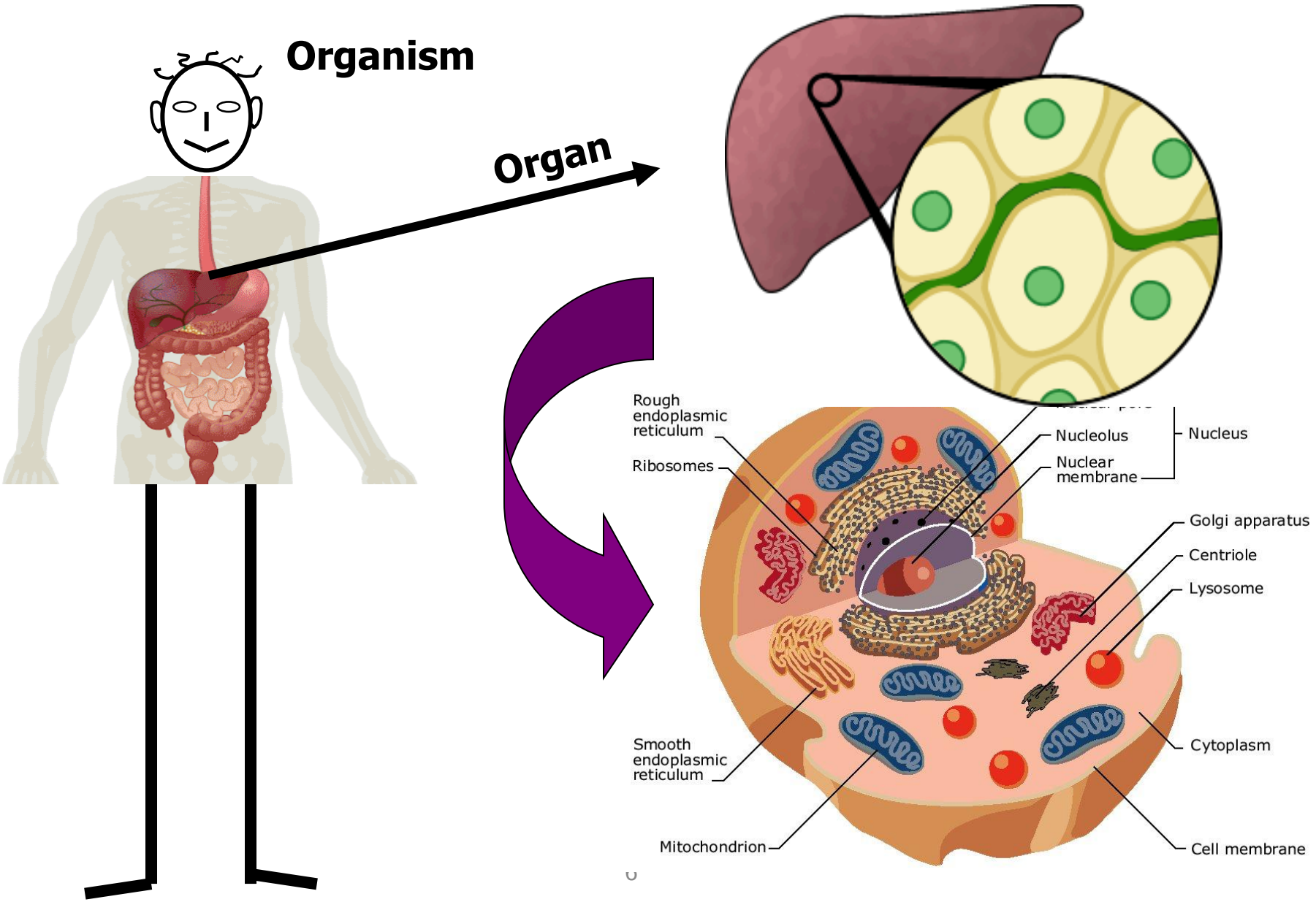
- 4 problem sets
- Midterm
- Project (and poster)
- Class attendance and participation

Class grades

- Problem sets (40%)
- Midterm (30%)
- Project (25%)
- Class participation (5%)

High level and brief intro to molecular biology and genomics

Organism, Organ, Cell



Types of Cells

- Eukaryots:
  - Plants, animals, humans
  - DNA resides in the nucleus
  - Contain also other compartments
- Prokaryots:
  - Bacteria
  - Do not contain compartments

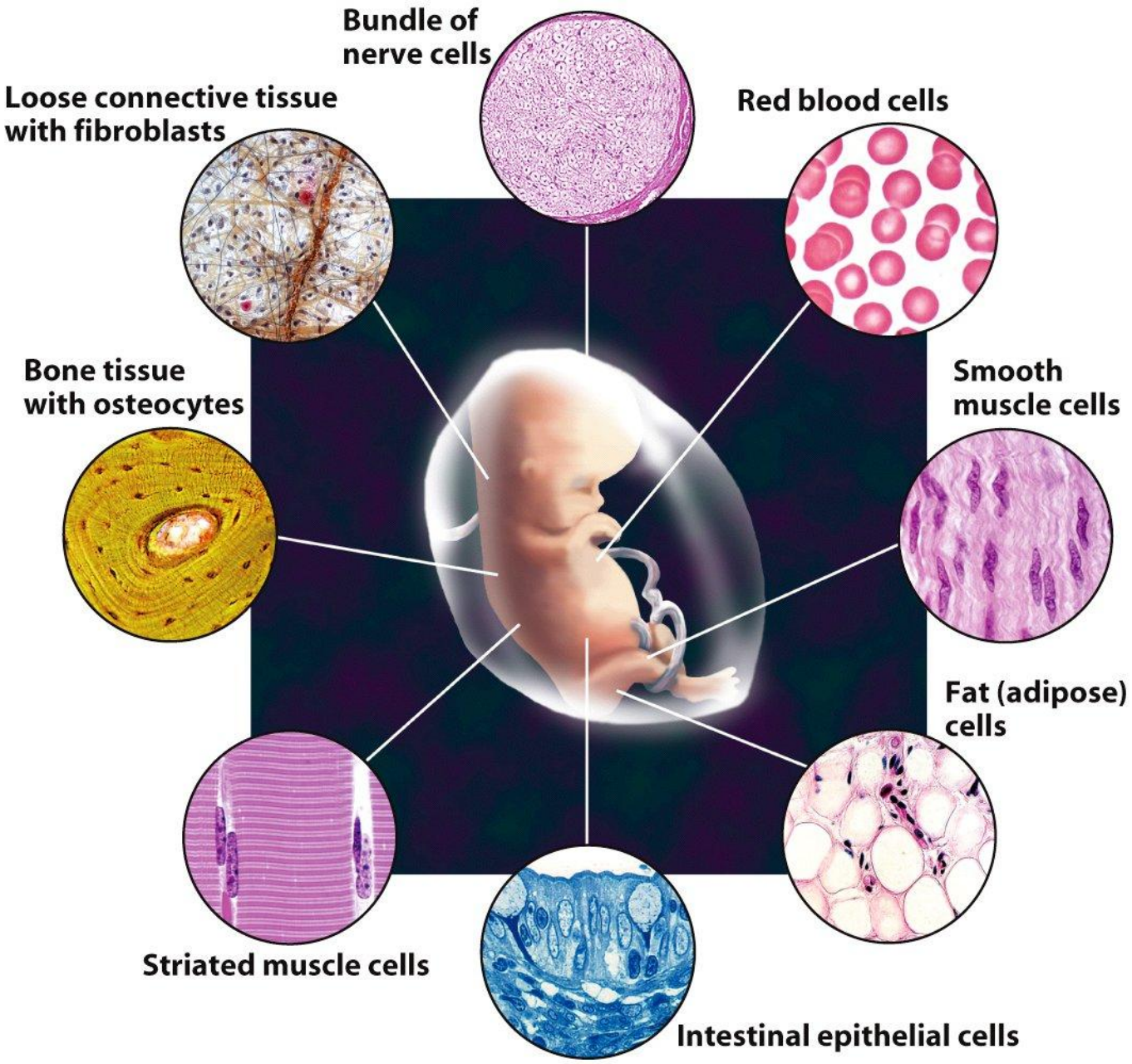


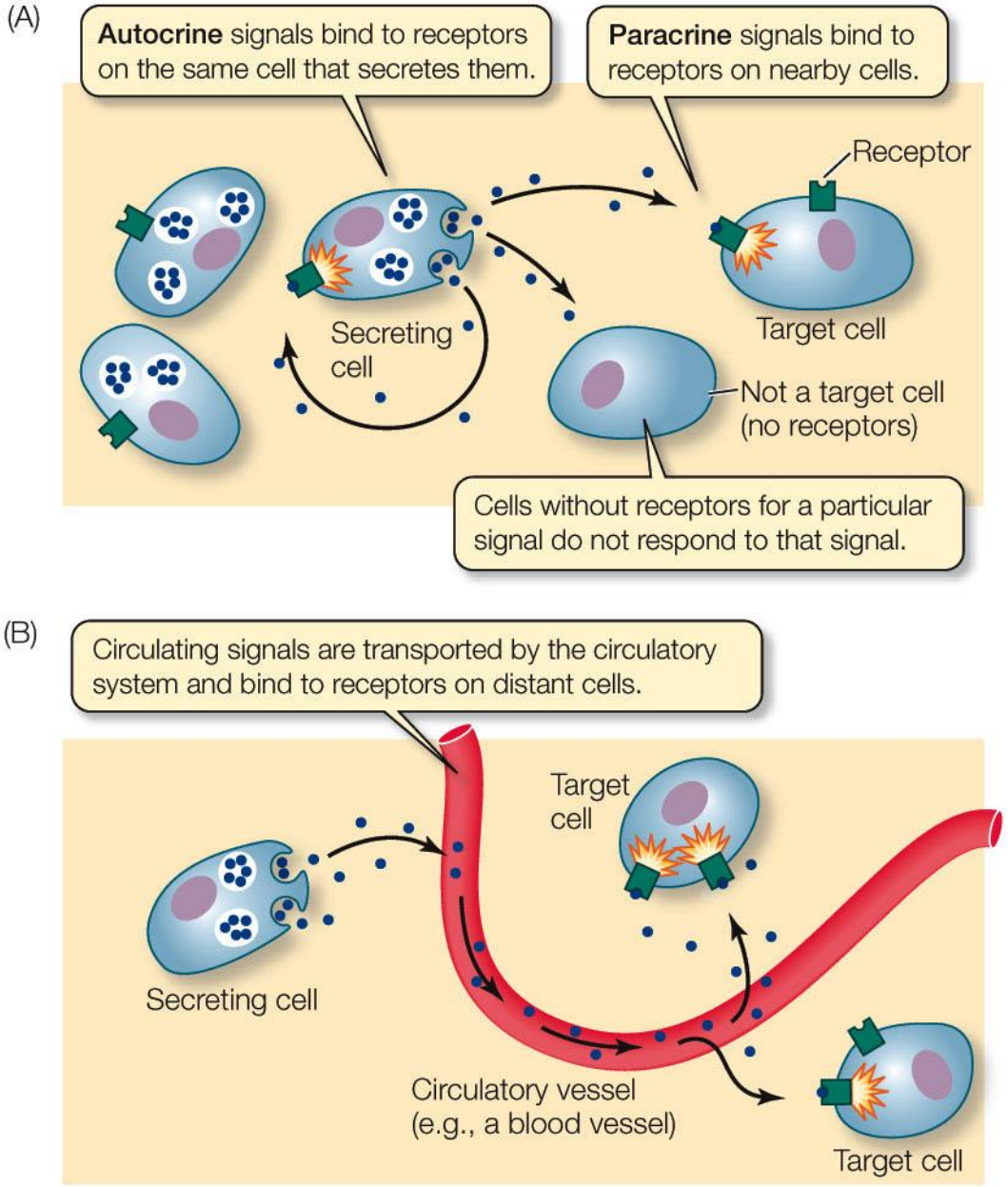
Figure 1-17 Cell and Molecular Biology, 4/e (© 2005 John Wiley & Sons)

Cell signaling

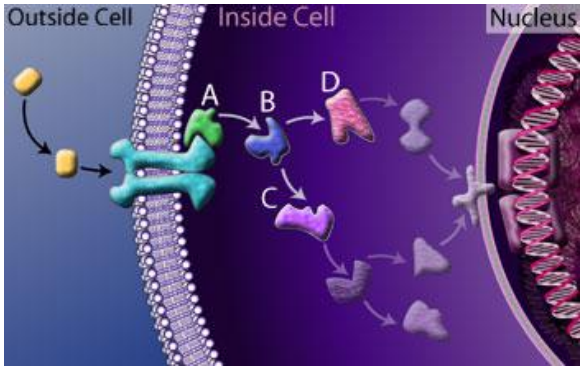
- Cells communication is based on chemical *signals* & *receptors*
  - If you have the correct receptor, respond to signal; no receptor = no response
  - Single-celled organisms receive cues about the environment, status of other individuals
- Process termed the **signal transduction pathway**
  - From signal interacting with receptor to cellular response

Types of Signals

- Local signaling: short-distance
  - affect the cells that produce them
  - affect nearby cells (diffuse)
- Hormonal signaling: long-distance
  - Typically found in multicellular organisms & use circulatory system for distribution



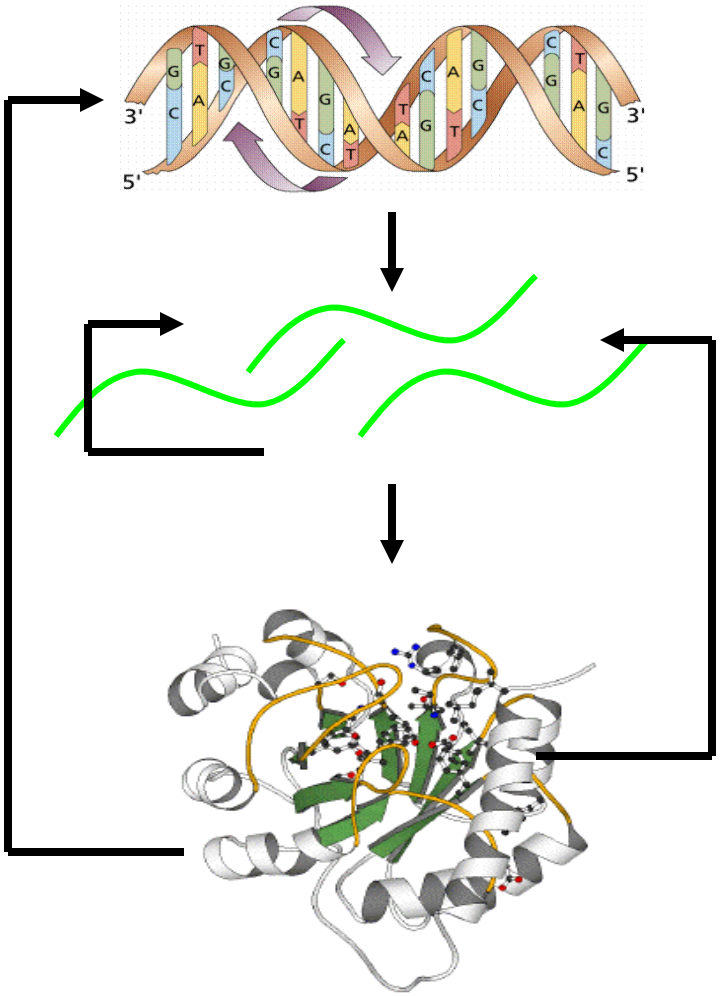
***LIFE 8e*, Figure 15.1**



Cell Signaling Stages

1. **Reception**: **signal** molecule interacts with **receptor**
2. **Transduction** typically several steps that involve changes to **responder** molecules and downstream targets
3. **Outcome:** often triggers a cellular response (*effect*)

Central dogma



DNA

transcription

mRNA

translation

Protein

CCTGAGCCAACTATTGATGAA

CCUGAGCCAACUAUUGAUGAA

PEPTIDE

Genome

- A genome is an organism's complete set of DNA (including its genes).
- In humans, less than 3% of the genome actually encodes for genes.
- However, a much larger % of the genome is transcribed (miRNAs, lincRNAs, …)
- And a large part of the rest of the genome serves as a control regions.

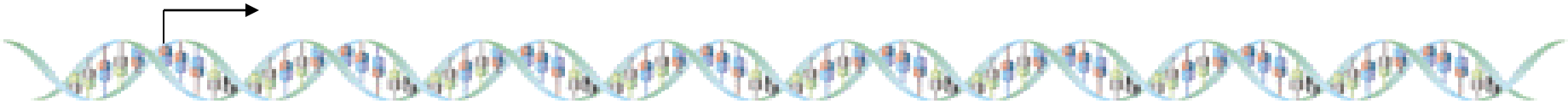
Comparison of Different Organisms

	Genome size	Num. of genes
E. coli	$.05*10^8$	4,200
Yeast	$.15*10^8$	6,000
Worm	$1*10^8$	18,400
Fly	$1.8*10^8$	13,600
Human	$30*10^8$	25,000
Plant	$1.3*10^8$	25,000

Genes

What is a gene?

Promoter | Protein coding sequence | Terminator



Genomic DNA

Example of a Gene: Gal4 DNA

ATGAAGCTACTGTCTTCTATCGAACAAGCATGCGATATTTGCCGACTTAAAAAGCTCAAG
TGCTCCAAAGAAAAACCGAAGTGCGCCAAGTGTCTGAAGAACAACTGGGAGTGTCGCTAC
TCTCCCAAAACCAAAAGGTCTCCGCTGACTAGGGCACATCTGACAGAAGTGGAATCAAGG
CTAGAAAGACTGGAACAGCTATTTCTACTGATTTTTCCTCGAGAAGACCTTGACATGATT
TTGAAAATGGATTCTTTACAGGATATAAAAGCATTGTTAACAGGATTATTTGTACAAGAT
AATGTGAATAAAGATGCCGTCACAGATAGATTGGCTTCAGTGGAGACTGATATGCCTCTA
ACATTGAGACAGCATAGAATAAGTGCGACATCATCATCGGAAGAGAGTAGTAACAAAGGT
CAAAGACAGTTGACTGTATCGATTGACTCGGCAGCTCATCATGATAACTCCACAATTCCG
TTGGATTTTATGCCCAGGGATGCTCTTCATGGATTTGATTGGTCTGAAGAGGATGACATG
TCGGATGGCTTGCCCTTCCTGAAAACGGACCCCAACAATAATGGGTTCTTTGGCGACGGT
TCTCTCTTATGTATTCTTCGATCTATTGGCTTTAAACCGGAAAATTACACGAACTCTAAC
GTTAACAGGCTCCCGACCATGATTACGGATAGATACACGTTGGCTTCTAGATCCACAACA
TCCCGTTTACTTCAAAGTTATCTCAATAATTTTCACCCCTACTGCCCTATCGTGCACTCA
CCGACGCTAATGATGTTGTATAATAACCAGATTGAAATCGCGTCGAAGGATCAATGGCAA
ATCCTTTTTAACTGCATATTAGCCATTGGAGCCTGGTGTATAGAGGGGGAATCTACTGAT
ATAGATGTTTTTTACTATCAAAATGCTAAATCTCATTTGACGAGCAAGGTCTTCGAGTCA

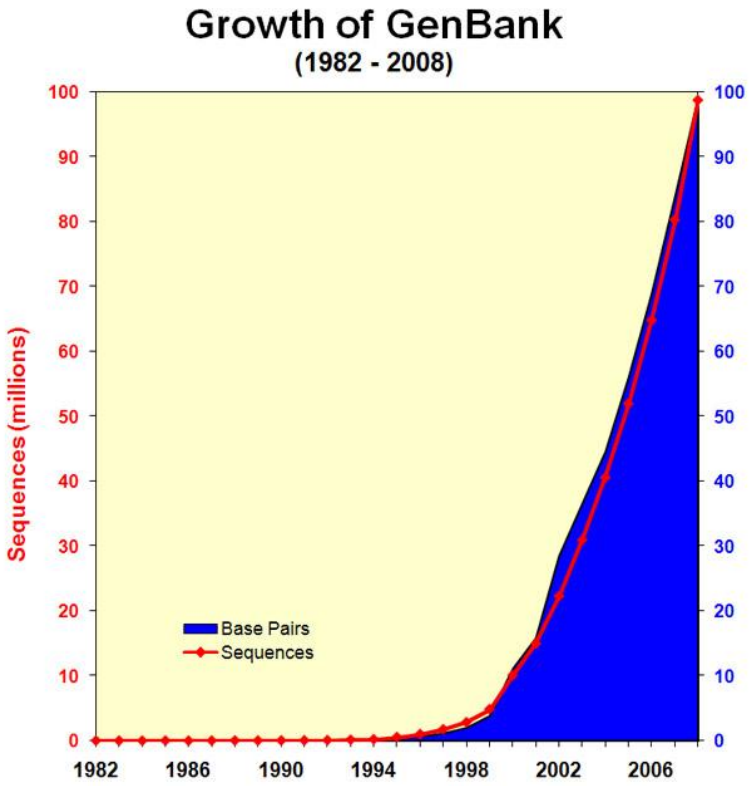
Genes Encode for Proteins

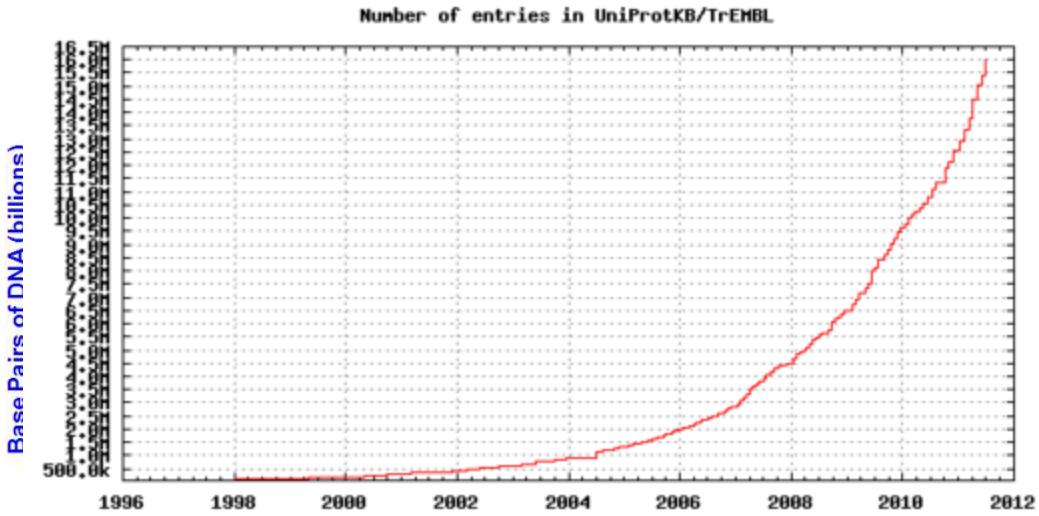
1st letter	Second Letter: U	Second Letter: C	Second Letter: A	Second Letter: G	3rd letter
U	UUU Phe UUC UUA Leu UUG	UCU UCC Ser UCA UCG	UAU Tyr UAC UAA Stop UAG Stop	UGU Cys UGC UGA Stop UGG Trp	U C A G
C	CUU CUC Leu CUA CUG	CCU CCC Pro CCA CCG	CAU His CAC CAA Gln CAG	CGU CGC Arg CGA CGG	U C A G
A	AUU AUC Ile AUA AUG Met	ACU ACC Thr ACA ACG	AAU Asn AAC AAA Lys AAG	AGU Ser AGC AGA Arg AGG	U C A G
G	GUU GUC Val GUA GUG	GCU GCC Ala GCA GCG	GAU Asp GAC GAA Glu GAG	GGU GGC Gly GGA GGG	U C A G

Example of a Gene: Gal4 AA

MKLLSSIEQACDICRLKKLKCSKEKPKCAKCLKNNWECRYSPKTKRSPLTRAHLTEVESR
LERLEQLFLLIFPREDLDMILKMDSLQDIKALLTGLFVQDNVNKDAVTDRLASVETDMPL
TLRQHRISATSSSEESSNKGQRQLTVSIDSAAHHDNSTIPLDFMPRDALHGFDWSEEDDM
SDGLPFLKTDPNNNGFFGDGSLLCILRSIGFKPENYTNSNVNRLPTMITDRYTLASRSTT
SRLLQSYLNNFHPYCPIVHSPTLMMLYNNQIEIASKDQWQILFNCILAIGAWCIEGESTD
IDVFYYQNAKSHLTSKVFESGSIILVTALHLLSRYTQWRQKTNTSYNFHSFSIRMAISLG
LNRDLPSSFSDSSILEQRRRIWWSVYSWEIQLSLLYGRSIQLSQNTISFPSSVDDVQRTT
TGPTIYHGIIETARLLQVFTKIYELDKTVTAEKSPICAKKCLMICNEIEEVSRQAPKFLQ
MDISTTALTNLLKEHPWLSFTRFELKWKQLSLIIYVLRDFFTNFTQKKSQLEQDQNDHQS
YEVKRCSIMLSDAAQRTVMSVSSYMDNHNVTPYFAWNCSYYLFNAVLVPIKTLLSNSKSN
AENNETAQLLQQINTVLMLLKKLATFKIQTCEKYIQVLEEVCAPFLLSQCAIPLPHISYN
NSNGSAIKNIVGSATIAQYPTLPEENVNNISVKYVSPGSVGPSPVPLKSGASFSDLVKLL
SNRPPSRNSPVTIPRSTPSHRSVTPFLGQQQQLQSLVPLTPSALFGGANFNQSGNIADSS

Number of Genes in Public Databases

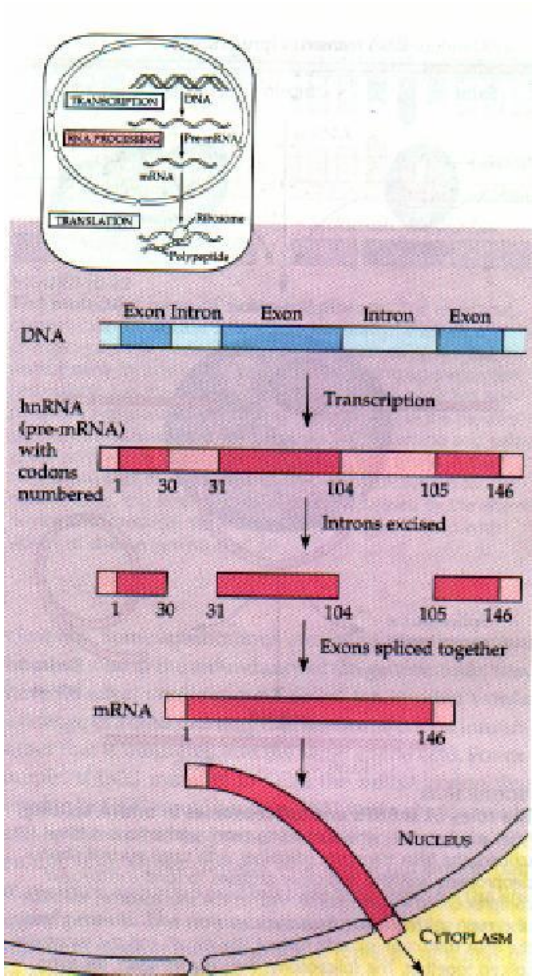




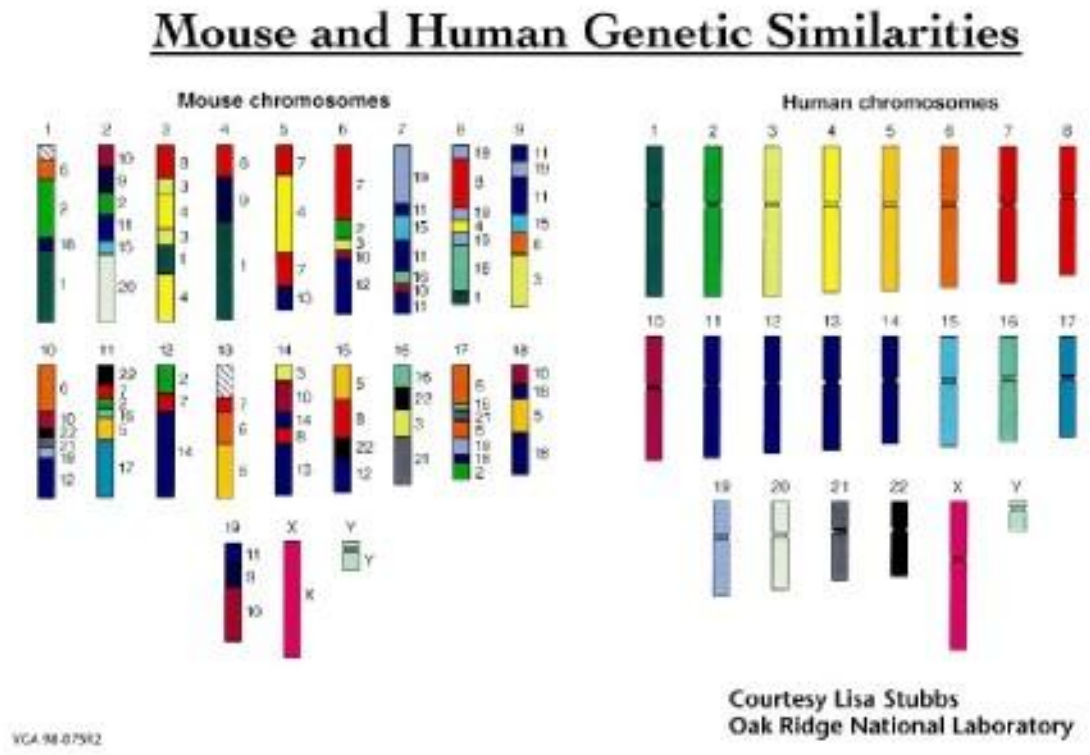
Structure of Genes in Mammalian Cells

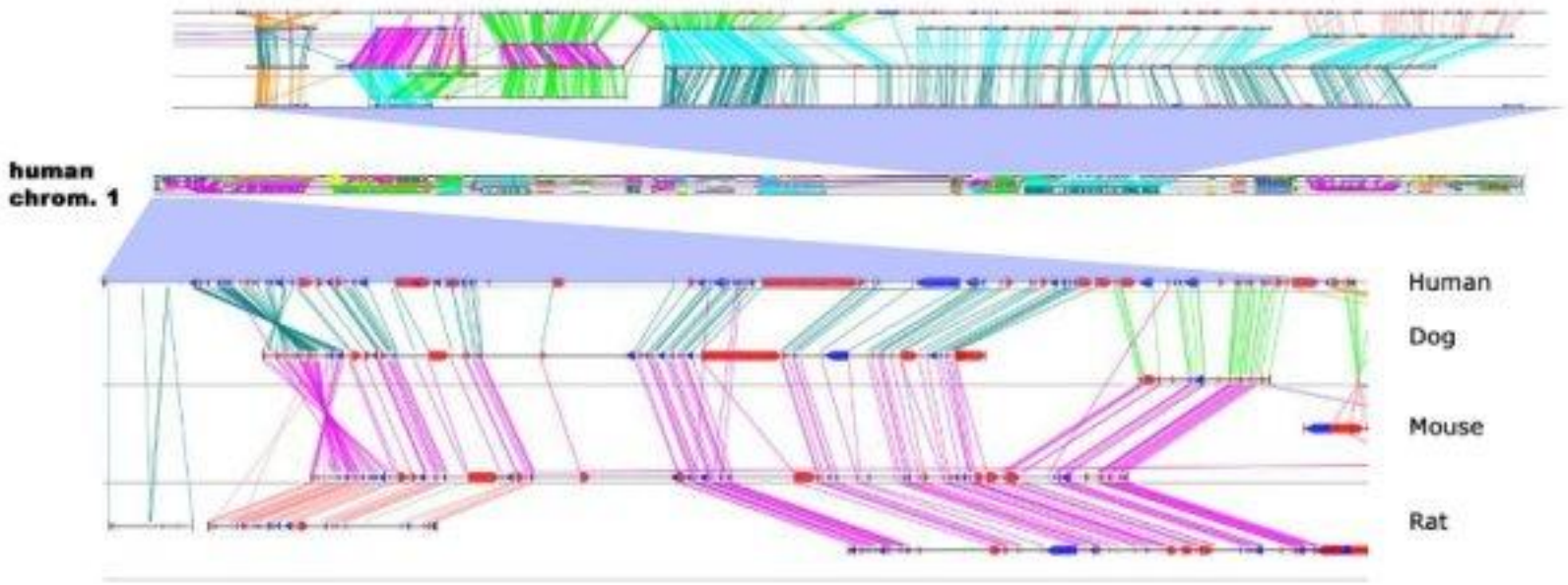
- Within coding DNA genes there can be un-translated regions (Introns)
- Exons are segments of DNA that contain the gene's information coding for a protein
- Need to cut Introns out of RNA and splice together Exons before protein can be made

- Alternative splicing increases the potential number of different proteins, allowing the generation of millions of proteins from a small number of genes.


TRANSCRIPTION
DNA
RNA PROCESSING
Pre-mRNA
mRNA
TRANSLATION
Ribosome
Polypeptide
DNA
Exon
Intron
Exon
Intron
Exon
hnRNA
(pre-mRNA)
with
codons
numbered
Transcription
1
30
31
104
105
146
Introns excised
1
30
31
104
105
146
Exons spliced together
mRNA
1
146
NUCLEUS
CYTOPLASM

Comparative genomics

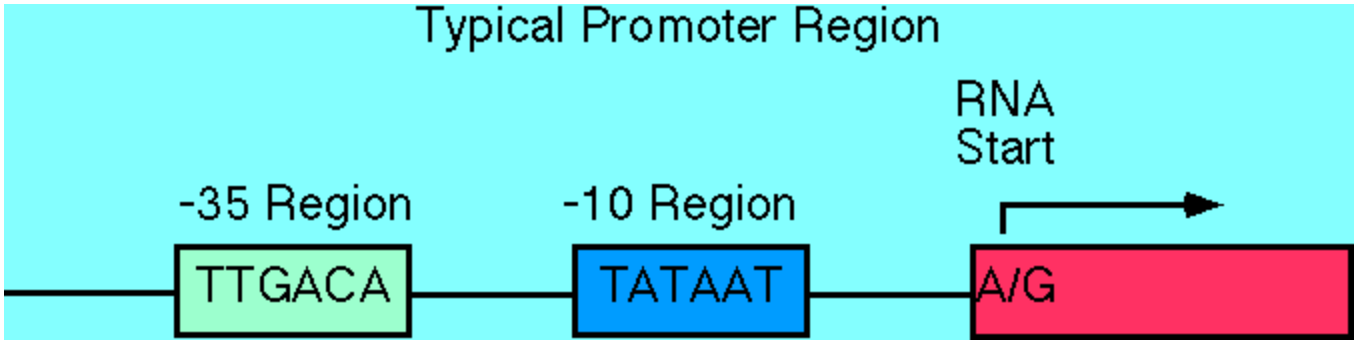



human
chrom. 1
Human
Dog
Mouse
Rat

Regulatory Regions

Promoter

The promoter is the place where RNA polymerase binds to start transcription. This is what determines which strand is the coding strand.



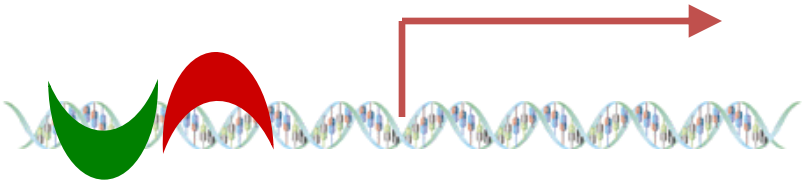
DNA Binding Motifs

• In order to recruit the transcriptional machinery, a transcription factor (TF) needs to bind the DNA in front of the gene.

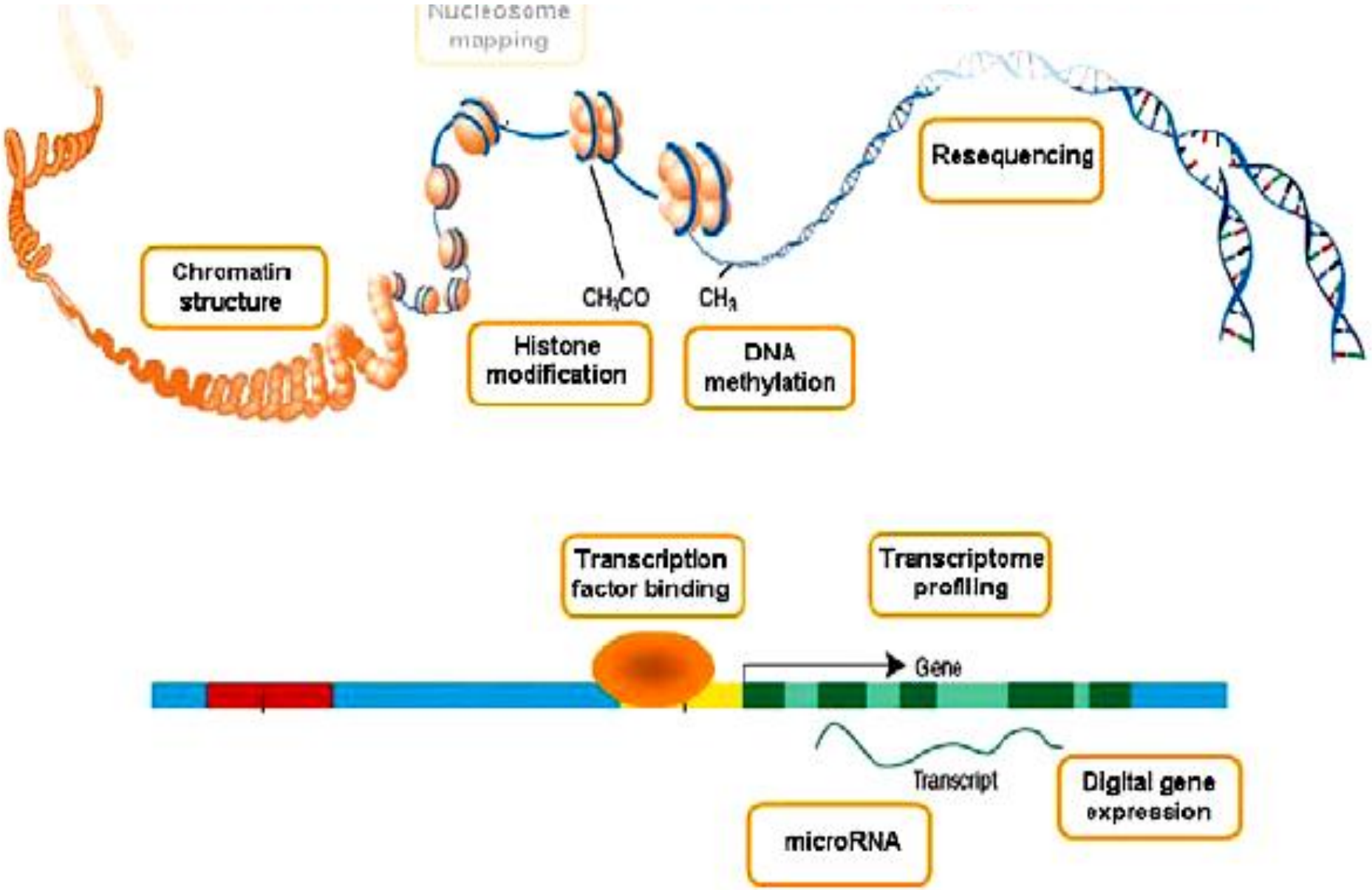
•TFs bind in to short segments which are known as DNA binding motifs.

• Usually consists 6 – 8 letters, and in many cases these letters generate palindromes.

• Note however that TF binding requires an open chromatin (a set of proteins that pack the DNA). Several factors are general chromatin modifiers. ‘Chicken and egg’ problem.



Epigenetics



Messenger RNAs (mRNAs)

RNA

Four major types (one recently discovered regulatory RNA).

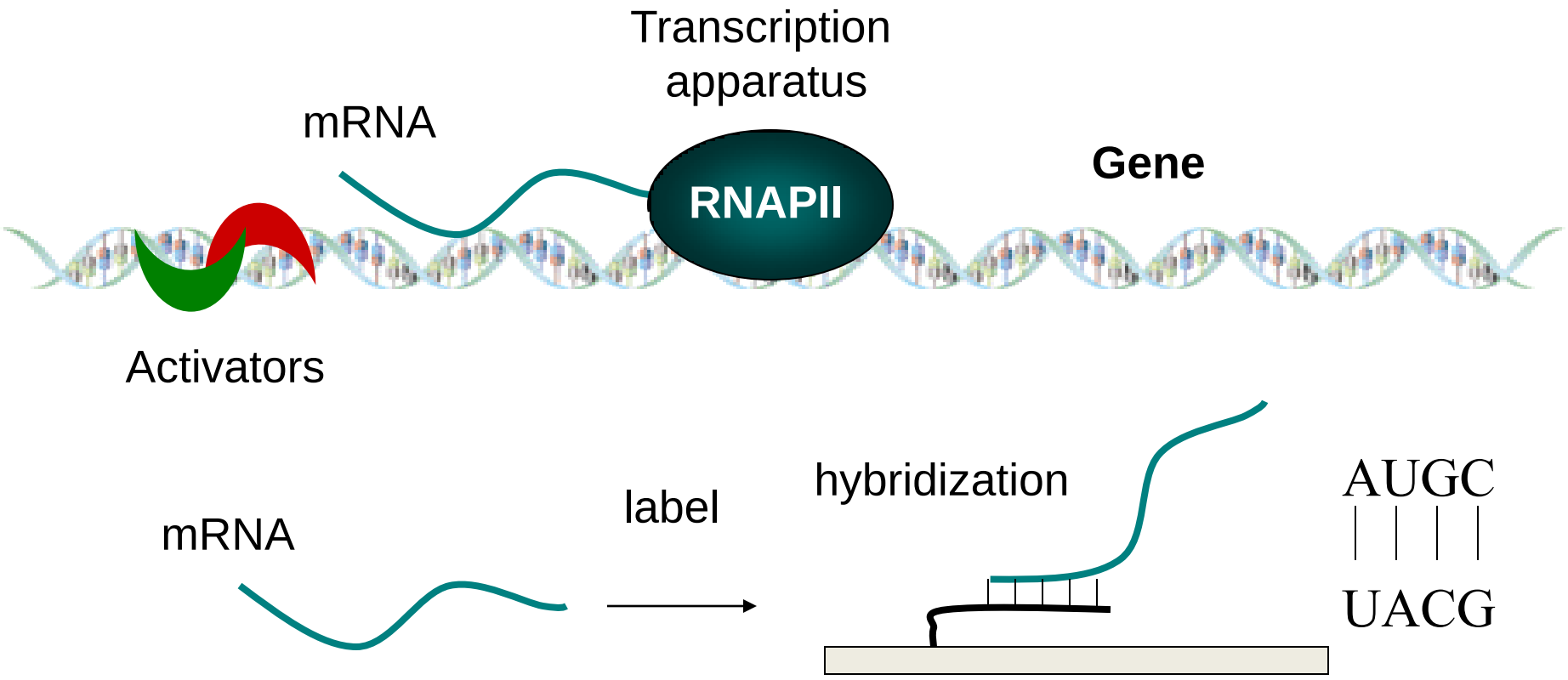
- mRNA – messenger RNA
- tRNA – Transfer RNA
- rRNA – ribosomal RNA
- shRNA, microRNA – RNA interference

Messenger RNA

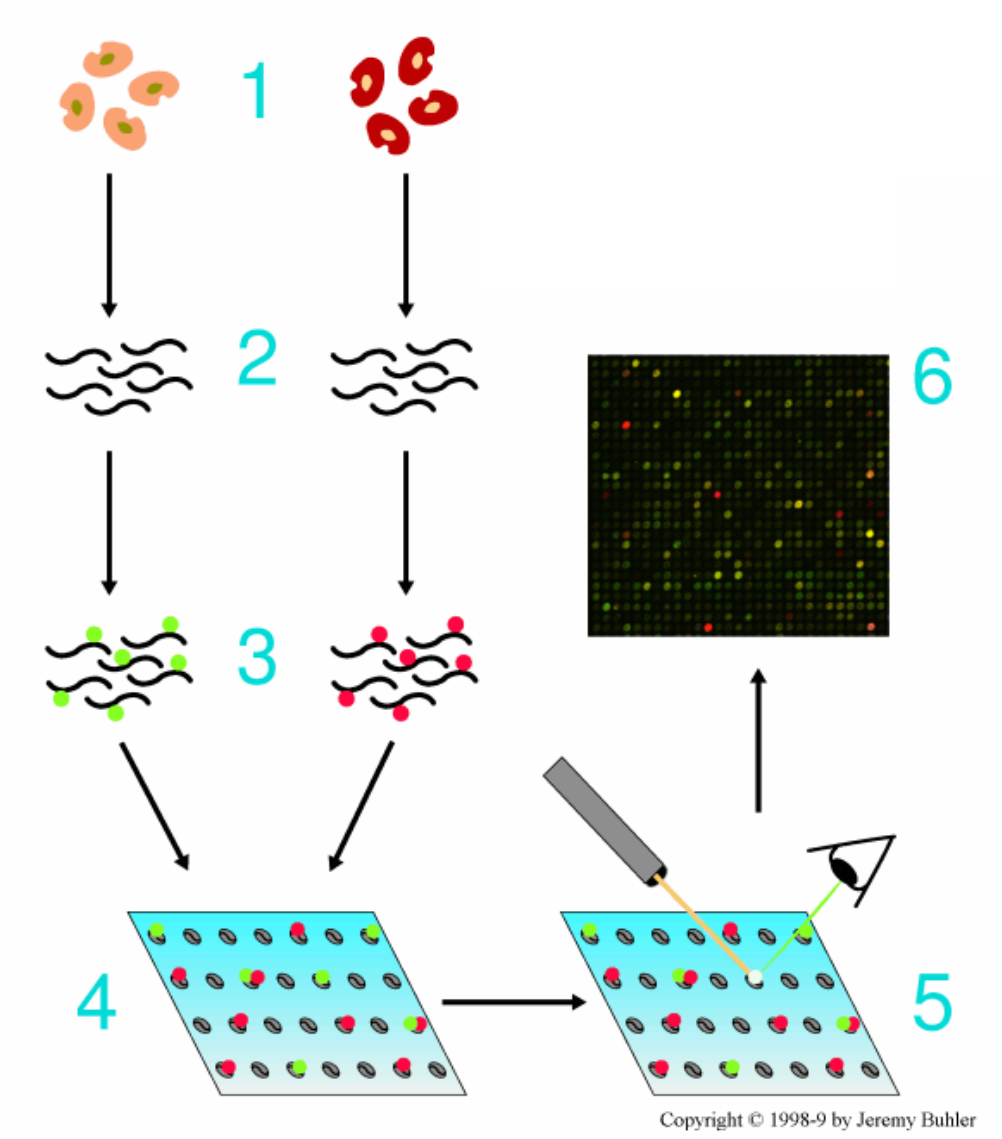
- Basically, an intermediate product
- Transcribed from the genome and translated into protein
- Number of copies correlates well with number of proteins for the gene.
- Unlike DNA, the amount of messenger RNA (as well as the number of proteins) differs between different cell types and under different conditions.

Complementary base-pairing

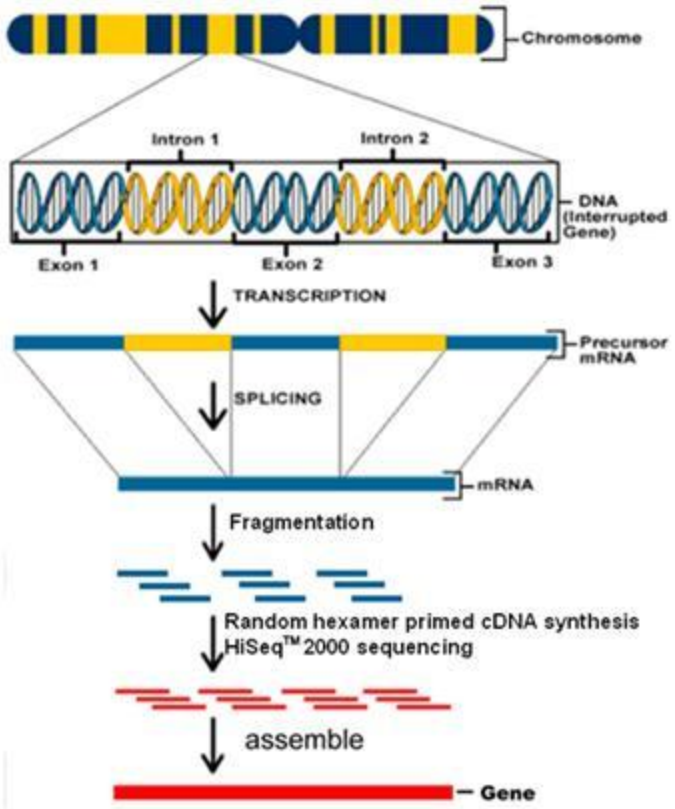
- mRNA is transcribed from the DNA
- mRNA (like DNA, but unlike proteins) binds to its complement



Hybridization and Scanning— Microarrays



RNASeq using next generation sequencing methods



Perturbation

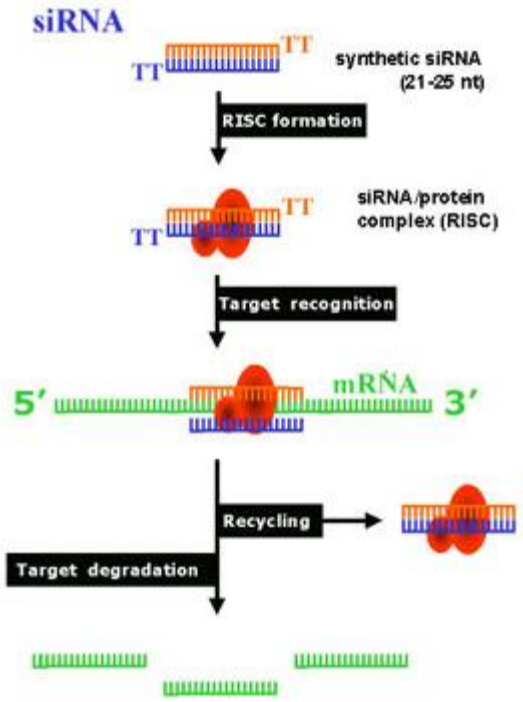
- In many cases we would like to perturb the systems to study the impacts of individual components (genes).
- This can be done in the sequence level by removing (knocking out) the gene of interest.
- Not always possible:

  - higher organisms

  - genes that are required during development but not later

  - genes that are required in certain cell types but not in others

Perturbations: RNAi

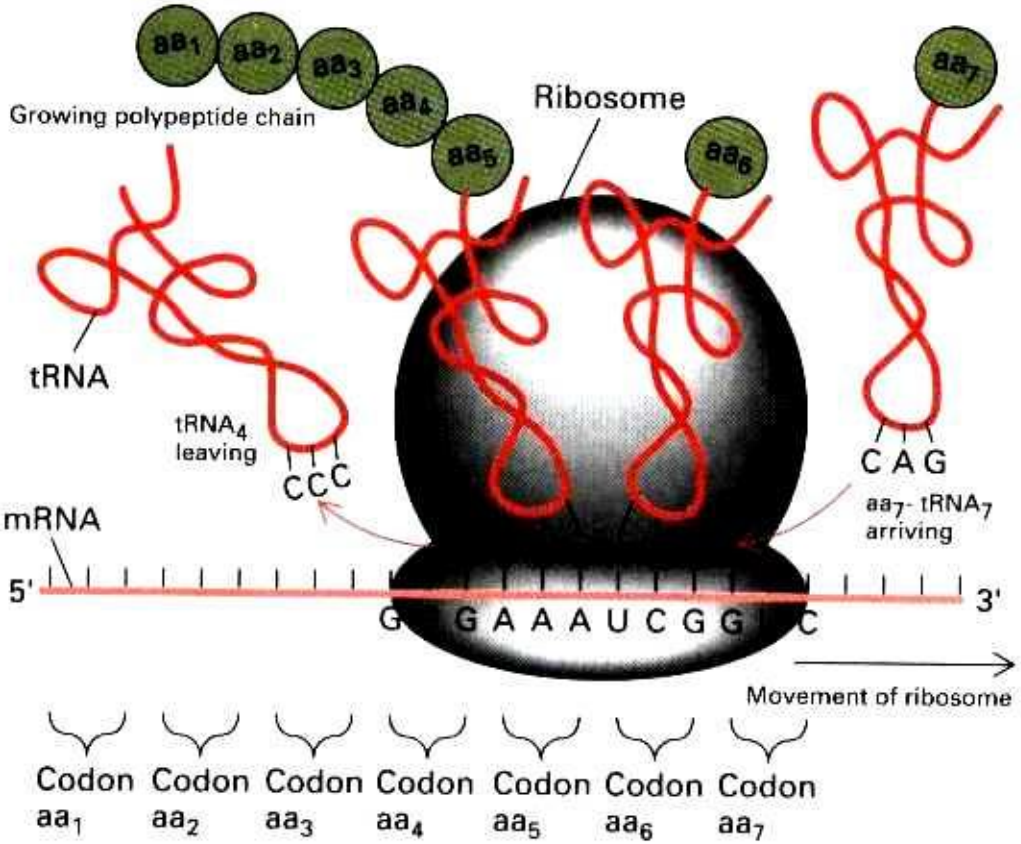


Proteins

From RNA to proteins: The Ribosome

- Decoding machine.
- Input: mRNA, output: protein
- Built from a large number of proteins and a number of RNAs.
- Several ribosomes can work on one mRNA

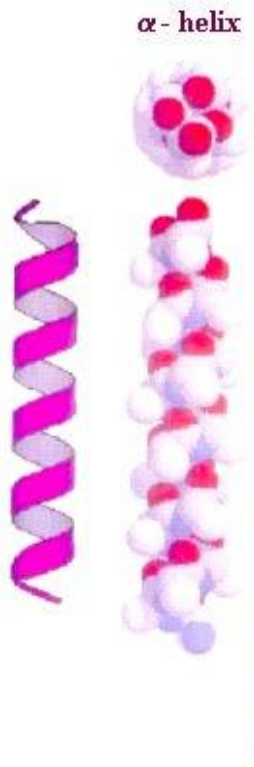
The Ribosome



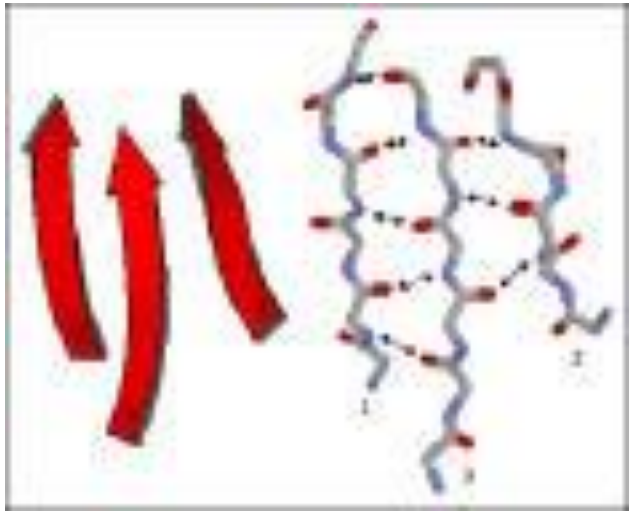
Proteins

- Proteins are polypeptide chains of amino acids.
- Four levels of structure:
  - Primary Structure: The sequence of the protein
  - Secondary structure: Local structure in regions of the chain
  - Tertiary Structure: Three dimensional structure
  - Quaternary Structure: multiple subunits

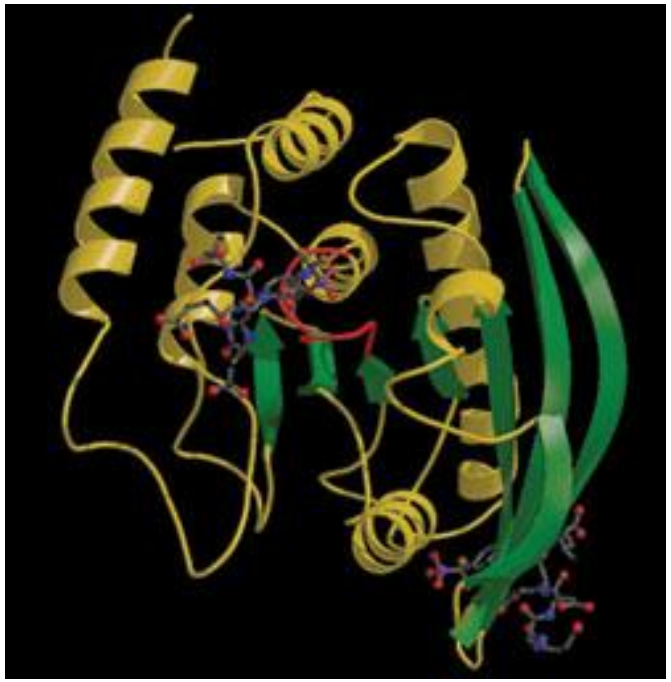
Secondary Structure: Alpha Helix



Secondary Structure: Beta Sheet



Protein Structure



Domains of a Protein

- While predicting the structure from the sequence is still an open problem, we can identify several domains within the protein.
- Domains are compactly folded structures.
- In many cases these domains are associated with specific biological function.

Protein Interaction

In order to fulfill their function, proteins interact with other proteins in a number of ways including:

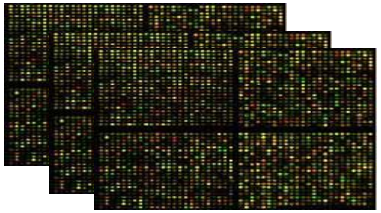
- Regulation
- Pathways, for example A -> B -> C
- Post translational modifications
- Forming protein complexes

Putting it all together: Systems biology

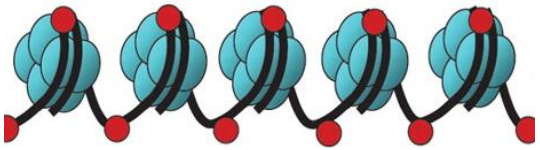
High throughput data

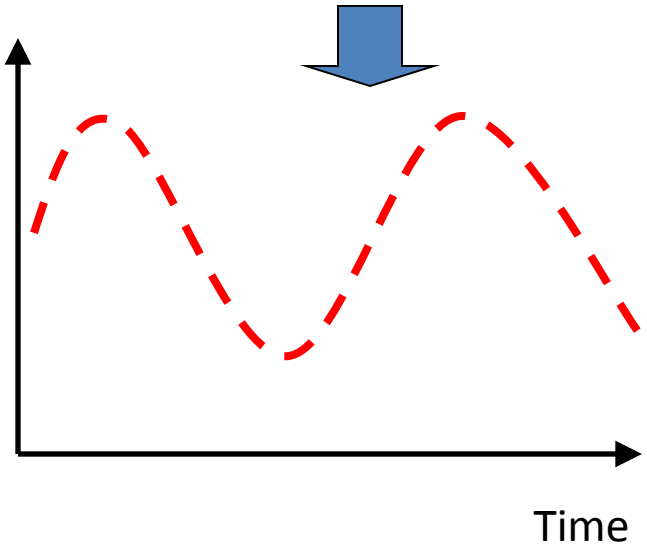
Time-series measurements

gene expression



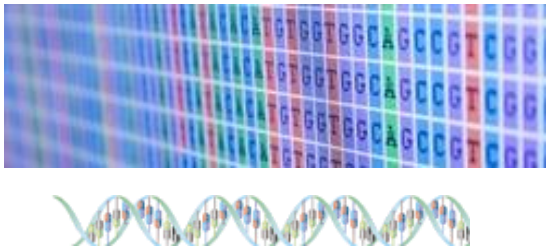
epigenetics





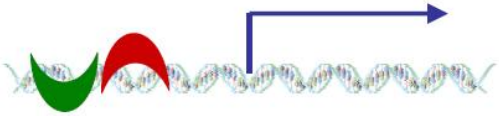
Static data sources

sequence

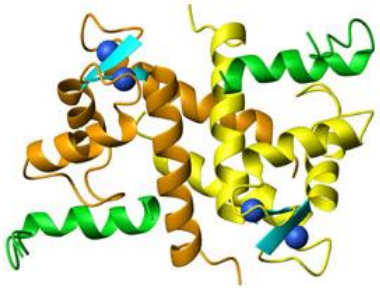


motif

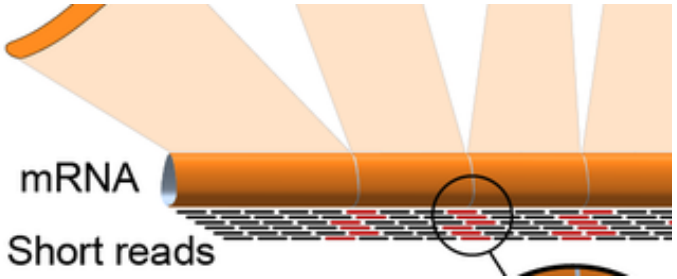
CHIP-Seq



PPI



RNA-Seq

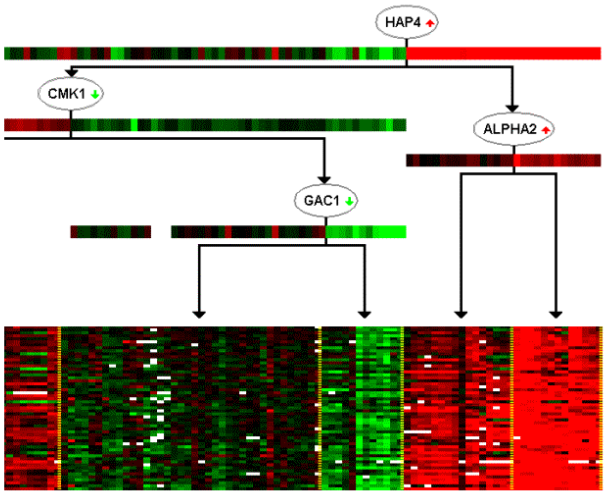


High throughput data

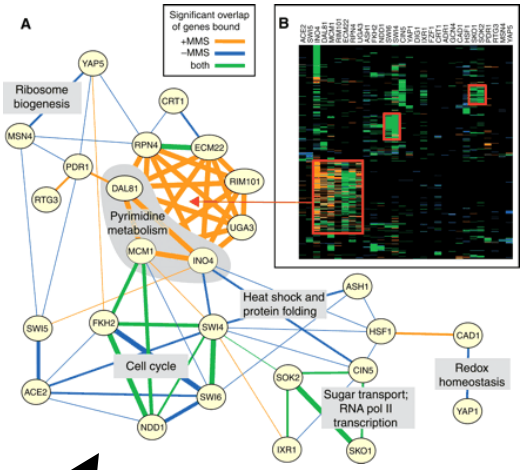
- We now have many sources of data, each providing a different view on the activity in the cell
  - Sequence (genes)

How to combine these different data types together to obtain a unified view of the activity in the cell is one of the major challenges of systems biology

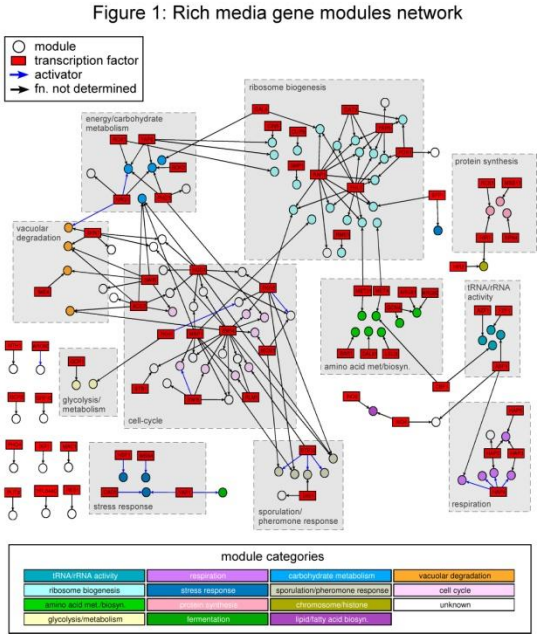
Reverse engineering of regulatory networks



Segal et al *Nature Genetics* 2003



Workman et al *Science* 2006

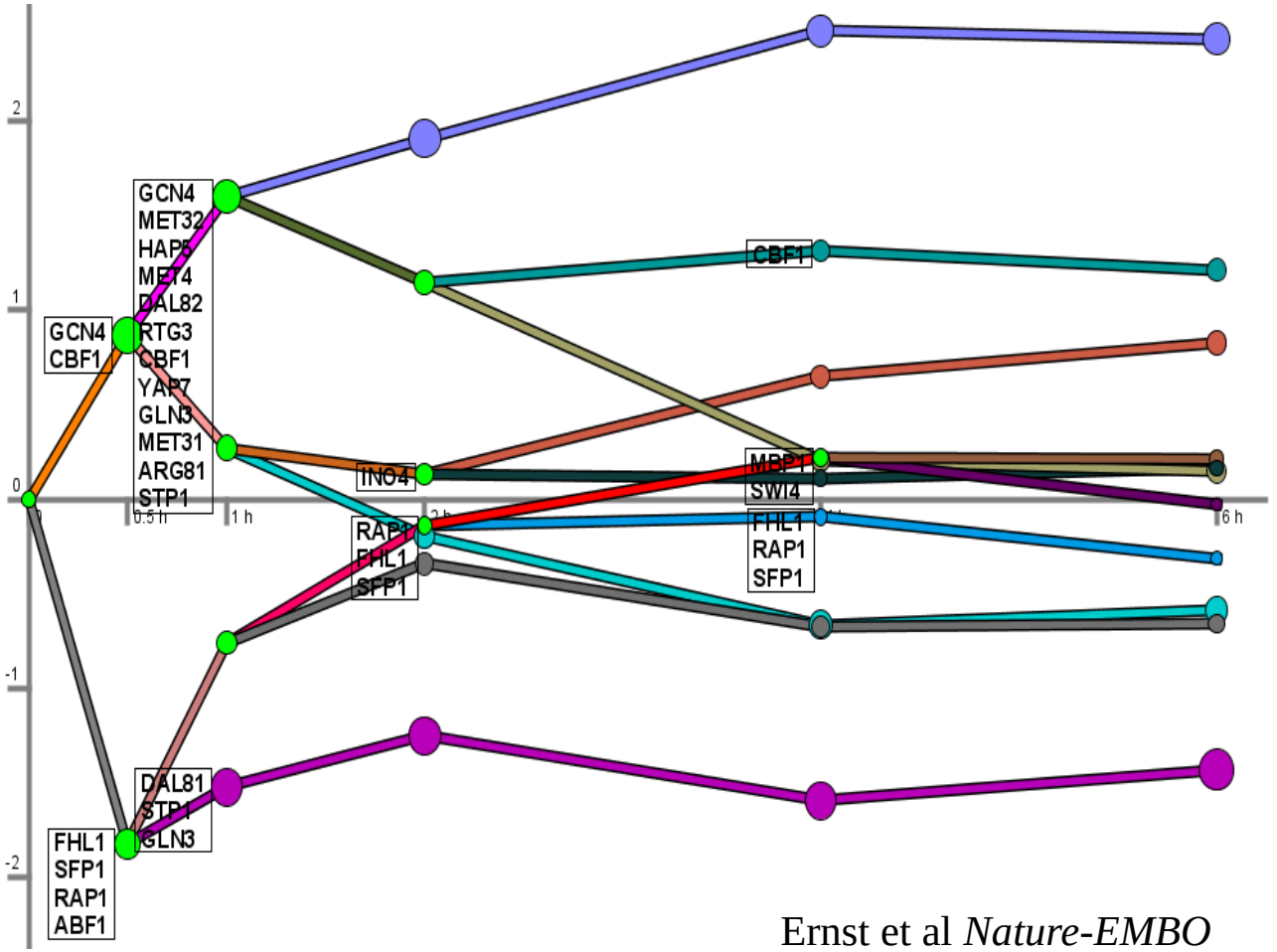


Bar-Joseph et al *Nature Biotechnology* 2003

- Gene expression
- Protein-DNA and gene expression

Dynamic regulatory networks

Protein-DNA, motif and time series gene expression data



Ernst et al *Nature-EMBO Mol. & Systems Bio.* 2007, PNAS 2013