10-810 /02-710
Computational Genomics

Ziv Bar-Joseph
zivbj@cs.cmu.edu
WeH 4107

Takis Benos
benos@pitt.edu
3078 BST3 (Pitt)

Eric Xing
epxing@cs.cmu.edu
WeH 4127

http://www.cs.cmu.edu/~epxing/Class/10810-07/

Topics

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

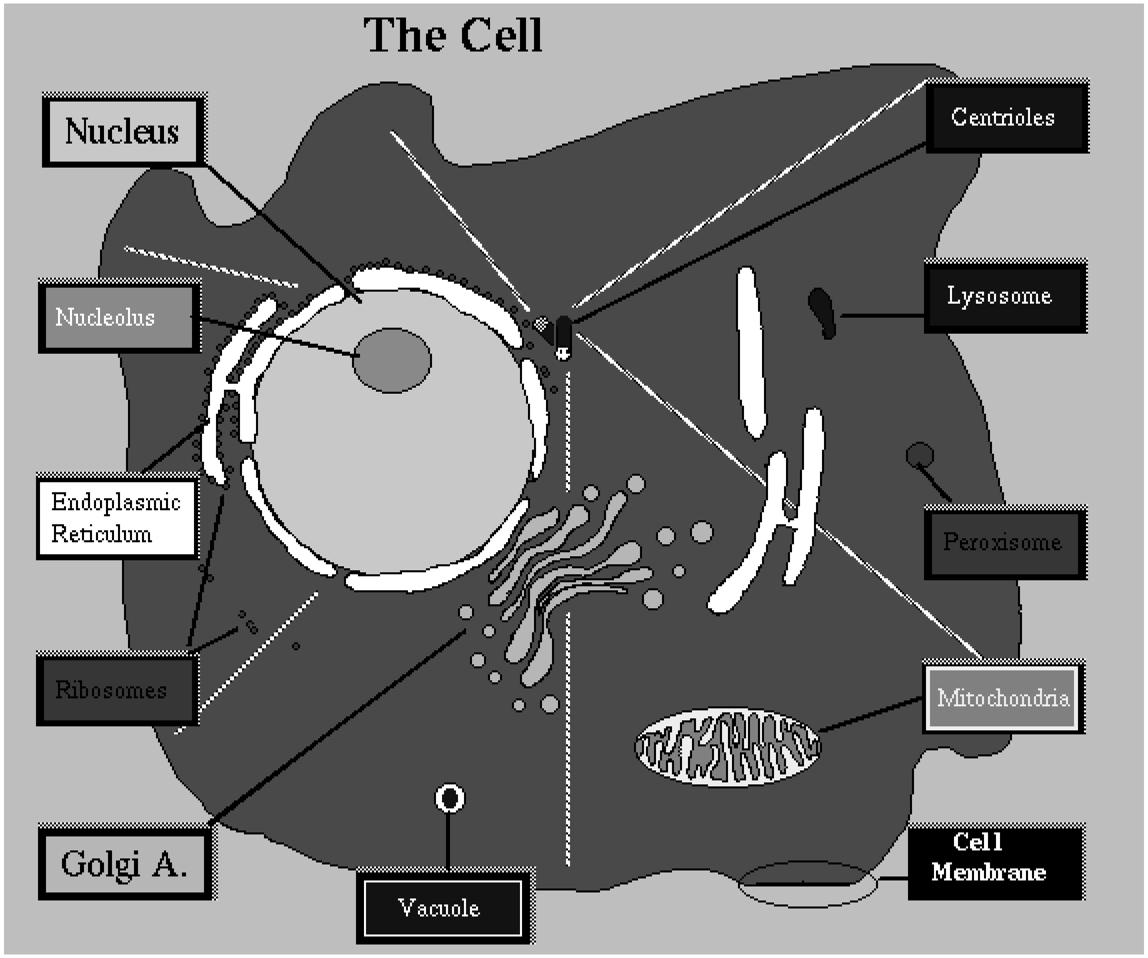
Grades

- 4 Problem sets: 36%
- Midterm: 24%
- Projects: 30%
- Class participation and reading: 10%

Introduction to Molecular Biology

- Genomes
- Genes
- Regulation
- mRNAs
- Proteins
- Systems

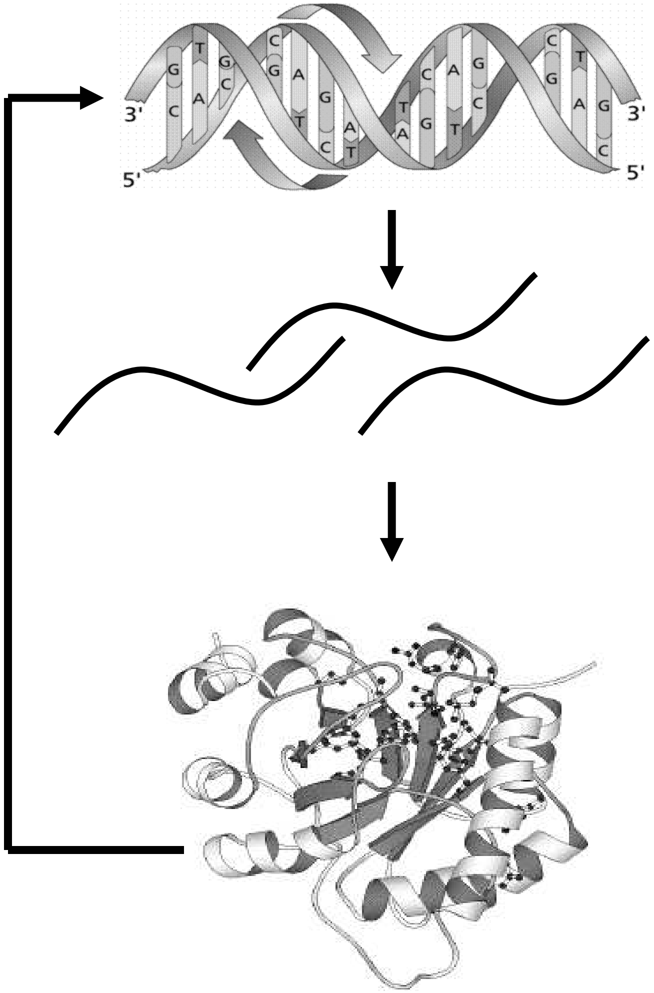
The Eukaryotic Cell

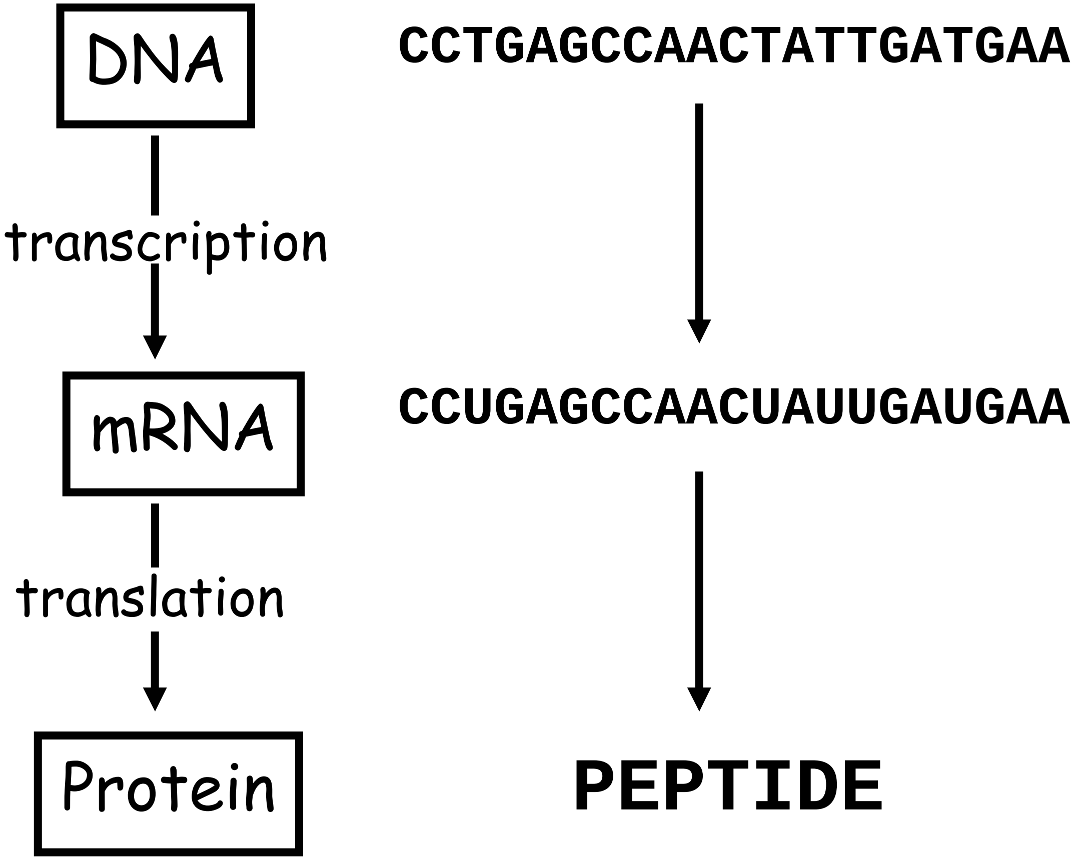


Cells Type

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

Central dogma





DNA

transcription

mRNA

translation

Protein

CCTGAGCCAACTATTGATGAA

CCUGAGCCAACUAUUGAUGAA

PEPTIDE

Genome

- A genome is an organism’s complete set of DNA (including its genes).
- However, in humans less than 3% of the genome actually encodes for genes.
- A part of the rest of the genome serves as a control regions (though that’s also a small part).
- The goal of the rest of the genome is unknown (a possible project …).

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

Assigning function to genes / proteins

- One of the main goals of molecular (and computational) biology.
- There are 25000 human genes and the vast majority of their functions is still unknown
- Several ways to determine function

  - Direct experiments (knockout, overexpression)
  - Interacting partners
  - 3D structures

  Hard

  - Sequence homology

  Easier

Function from sequence homology

- We have a query gene: ACTGGTGTACCGAT
- Given a database with genes with a known function, our goal is to find another gene with similar sequence (possibly in another organism)
- When we find such gene we predict the function of the query gene to be similar to the resulting database gene
- Problems

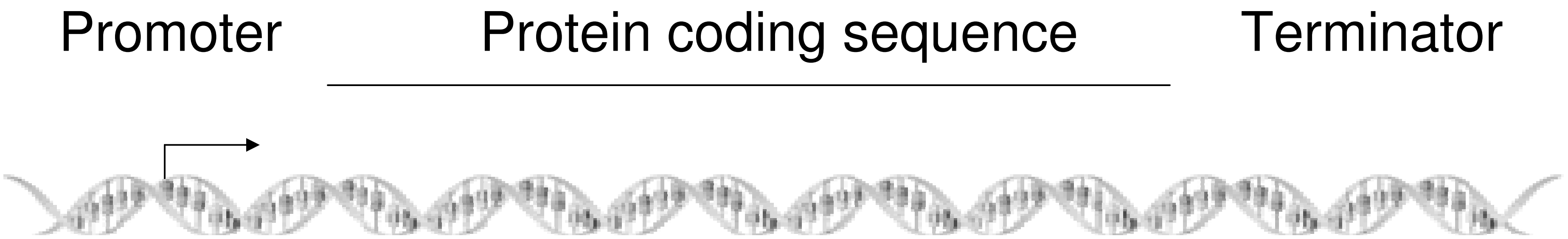
  - How do we determine similarity?

Sequence analysis techniques

- A major area of research within computational biology.
- Initially, based on deterministic (dynamic programming) or heuristic (Blast) alignment methods
- More recently, based on probabilistic inference methods (HMMs).

Genes

What is a gene?



Promoter

Protein coding sequence

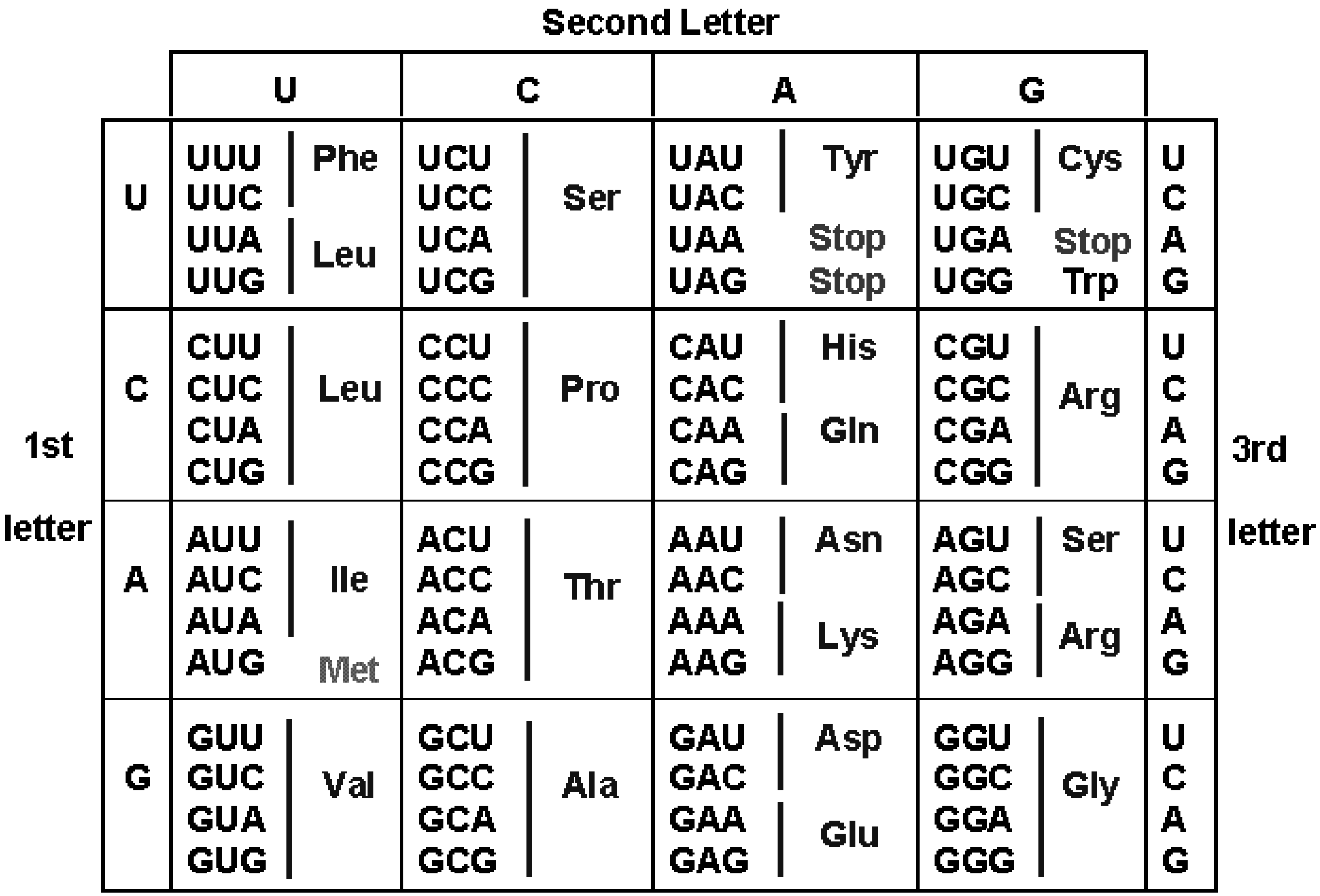
Terminator

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

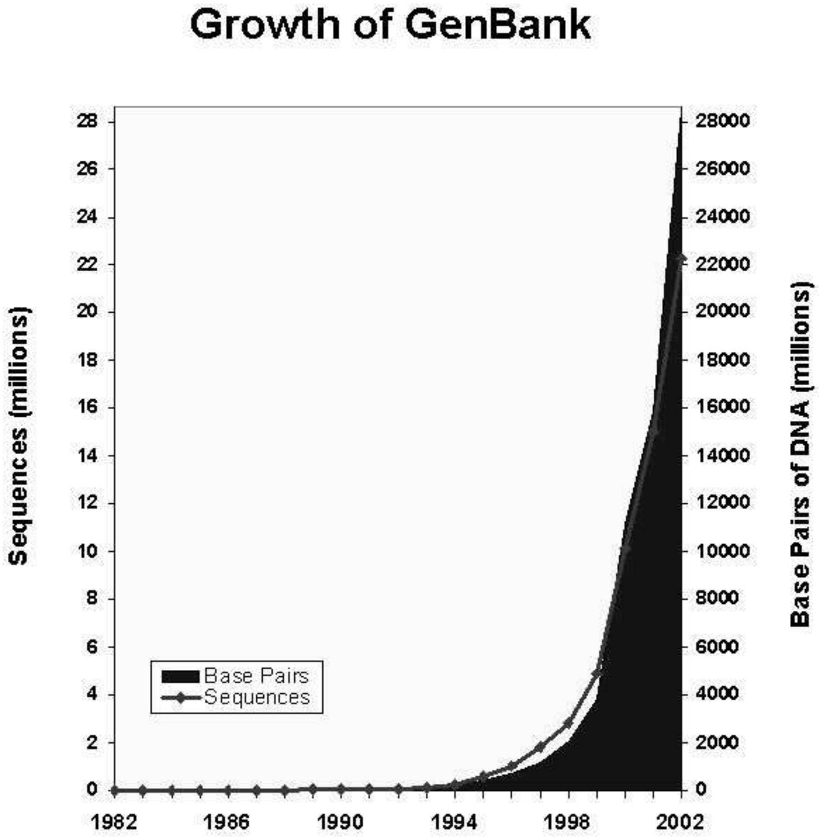


	Second Letter								
1st letter	U		C		A		G		3rd letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	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

Example of a Gene: Gal4 AA

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

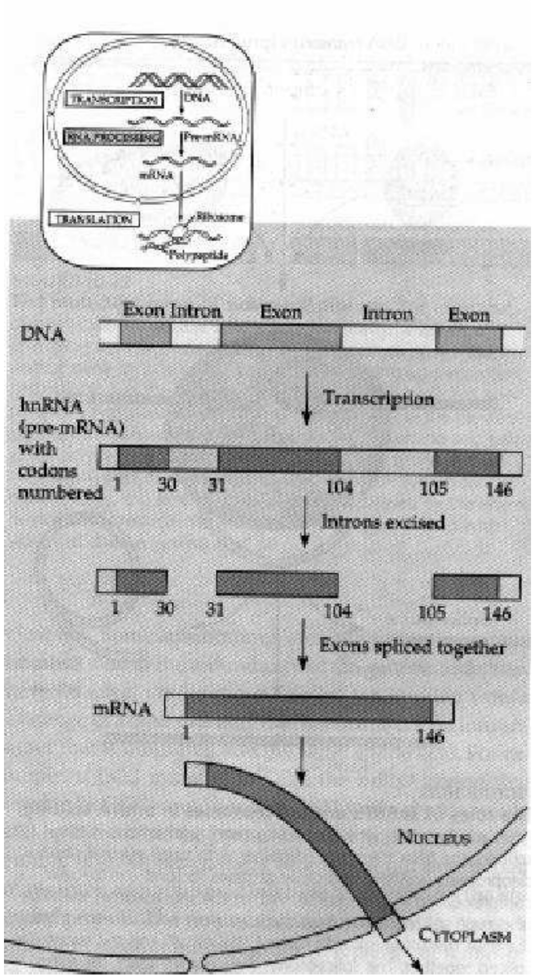
Number of Genes in Public Databases



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
Exon
Intron
Exon
Intron
Exon
DNA
Transcription
hnRNA
(pre-mRNA)
with
codons
numbered
1
30
31
104
105
146
Introns excised
1
30
31
104
105
146
Exons spliced together
mRNA
1
146
NUCLEUS
CYTOPLASM

Identifying Genes in Sequence Data

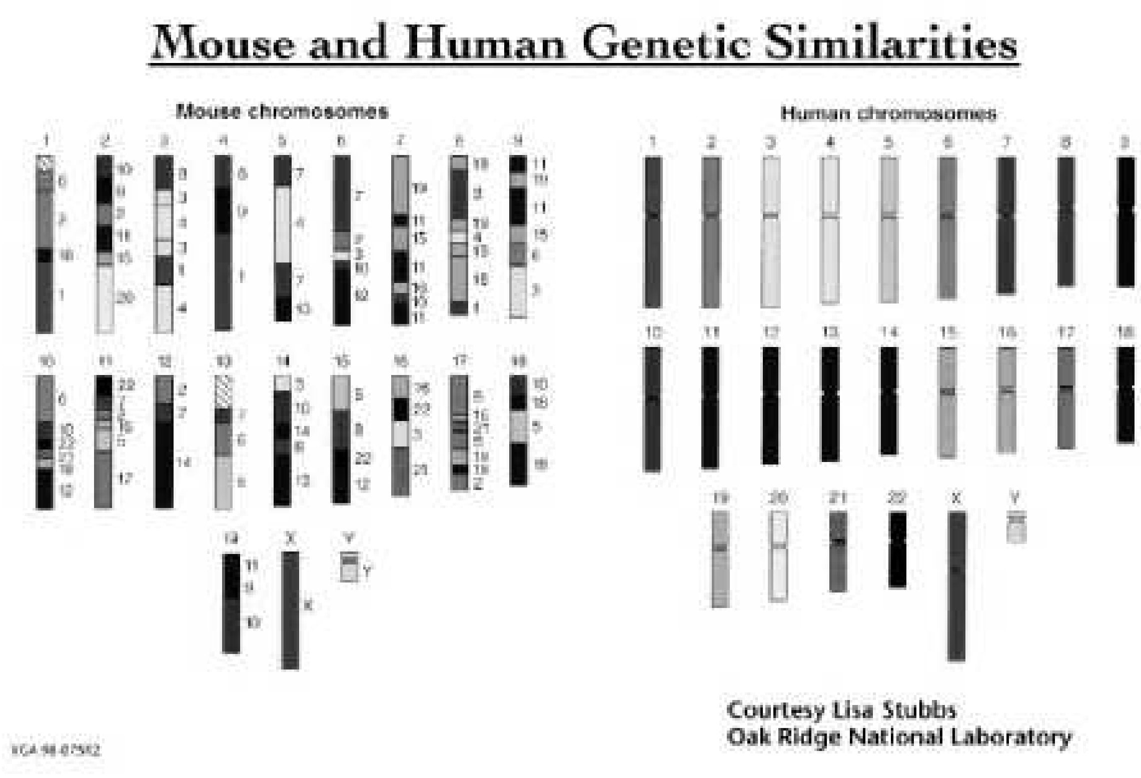
- Predicting the start and end of genes as well as the introns and exons in each gene is one of the basic problems in computational biology.

- Gene prediction methods look for *ORFs* (Open Reading Frame).

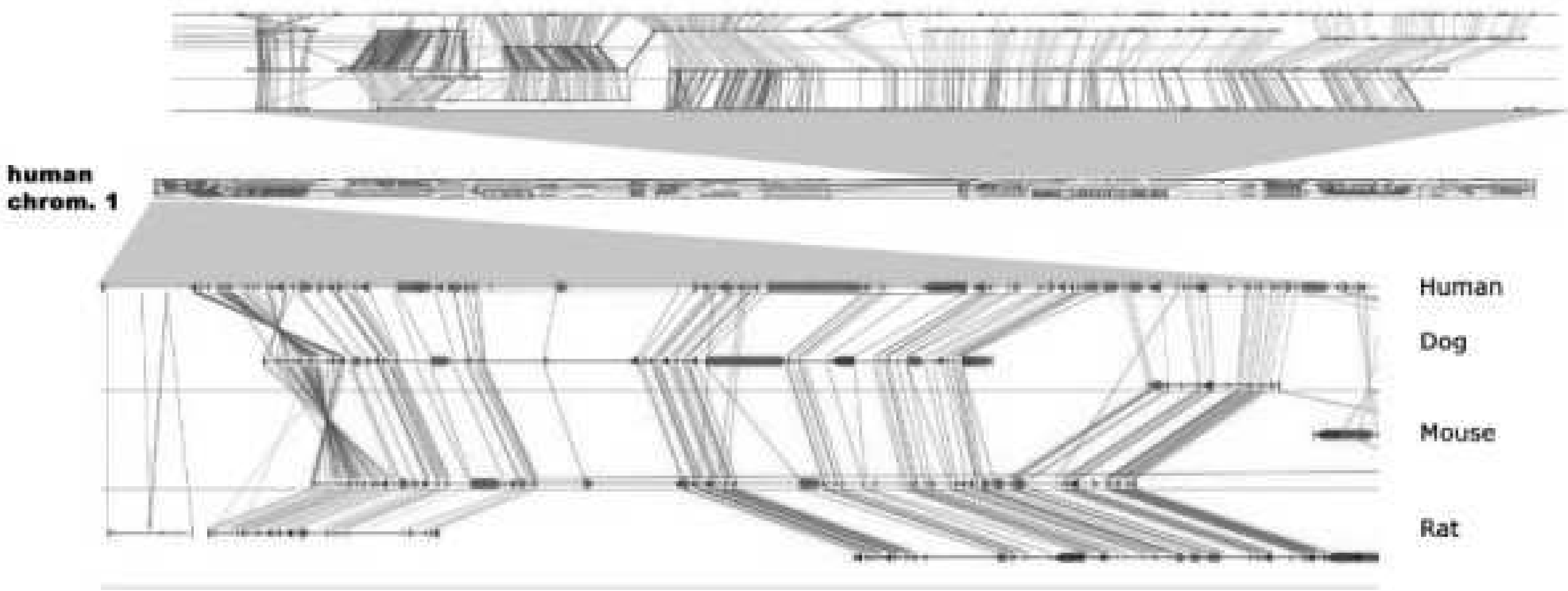
- These are (relatively long) DNA segments that start with the start codon, end with one of the end codons, and do not contain any other end codon in between.

- Splice site prediction has received a lot of attention in the literature.

Comparative genomics



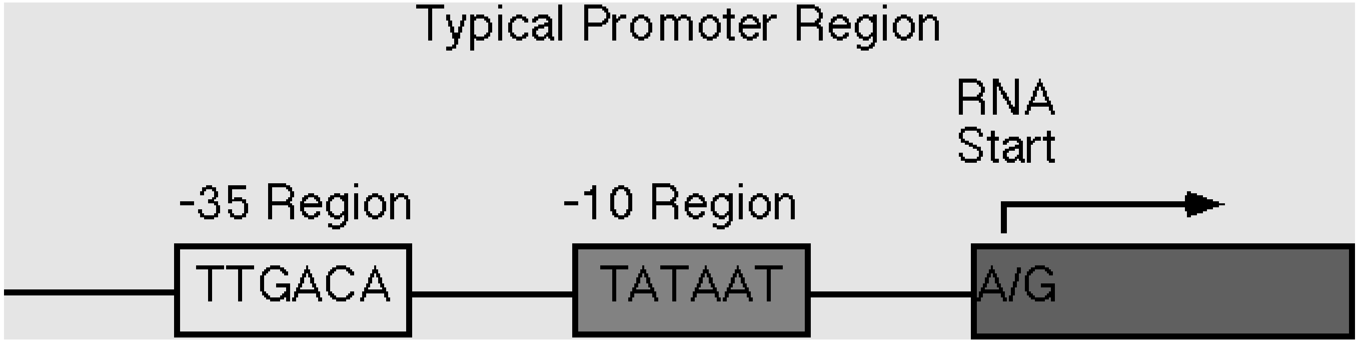
Courtesy Lisa Stubbs
Oak Ridge National Laboratory


human
chrom. 1
Human
Dog
Mouse
Rat

Regulatory Regions

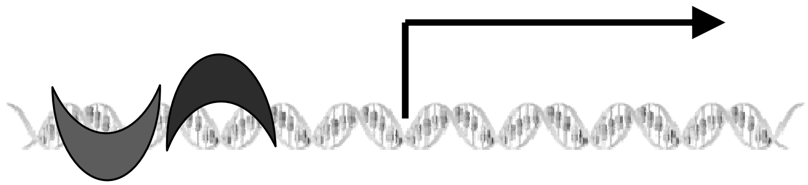
Promoter

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

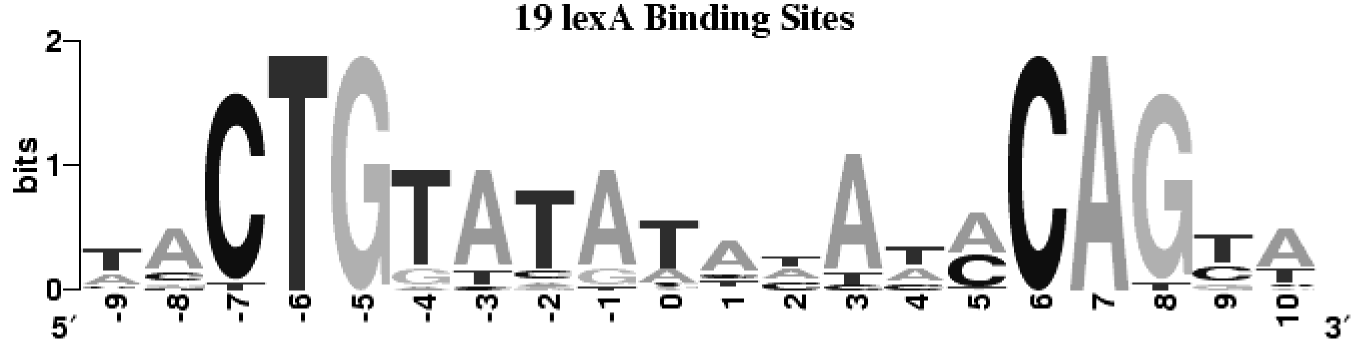


DNA Binding Motifs

- 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.



Example of Motifs



Messenger RNAs (mRNAs)

RNA

Four major types (one recently discovered regulatory RNA).

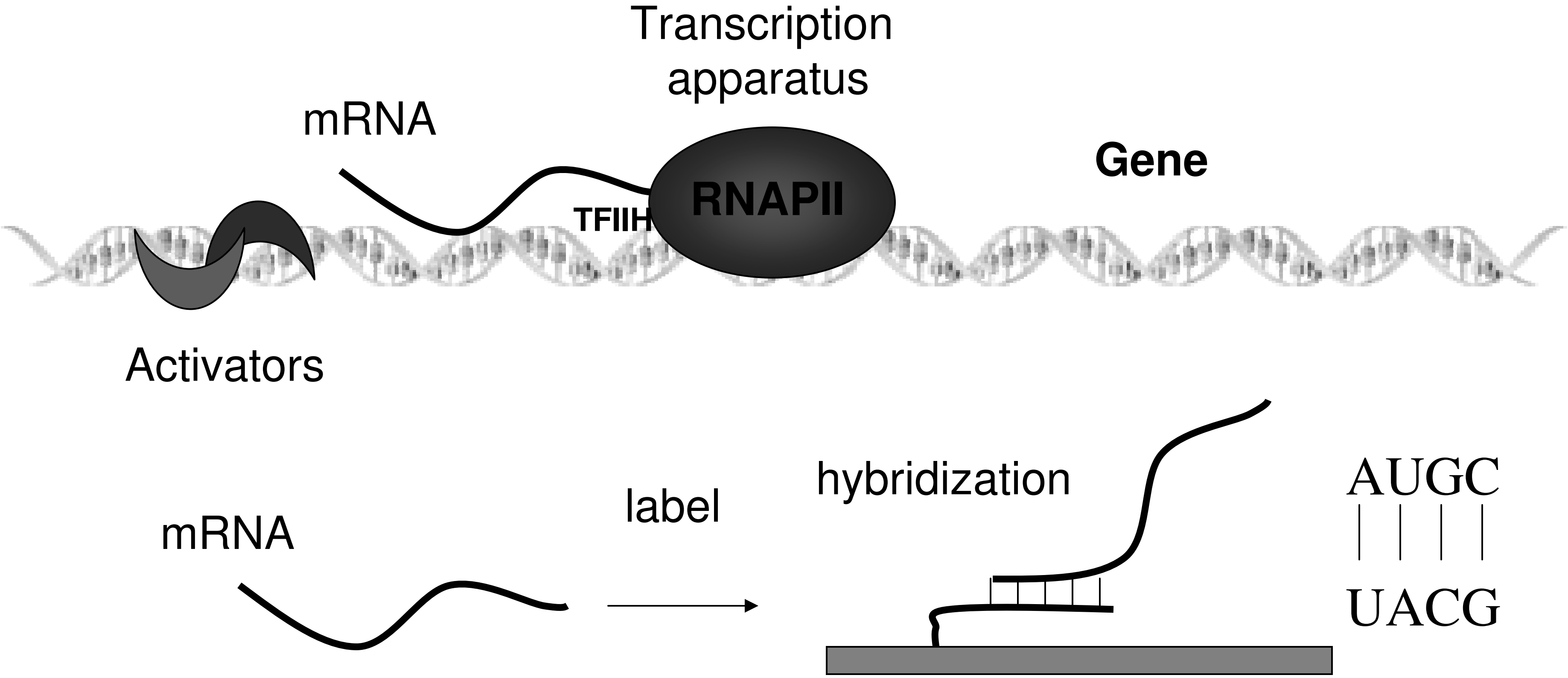
- mRNA – messenger RNA
- tRNA – Transfer RNA
- rRNA – ribosomal RNA
- RNAi, microRNA – RNA interference

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

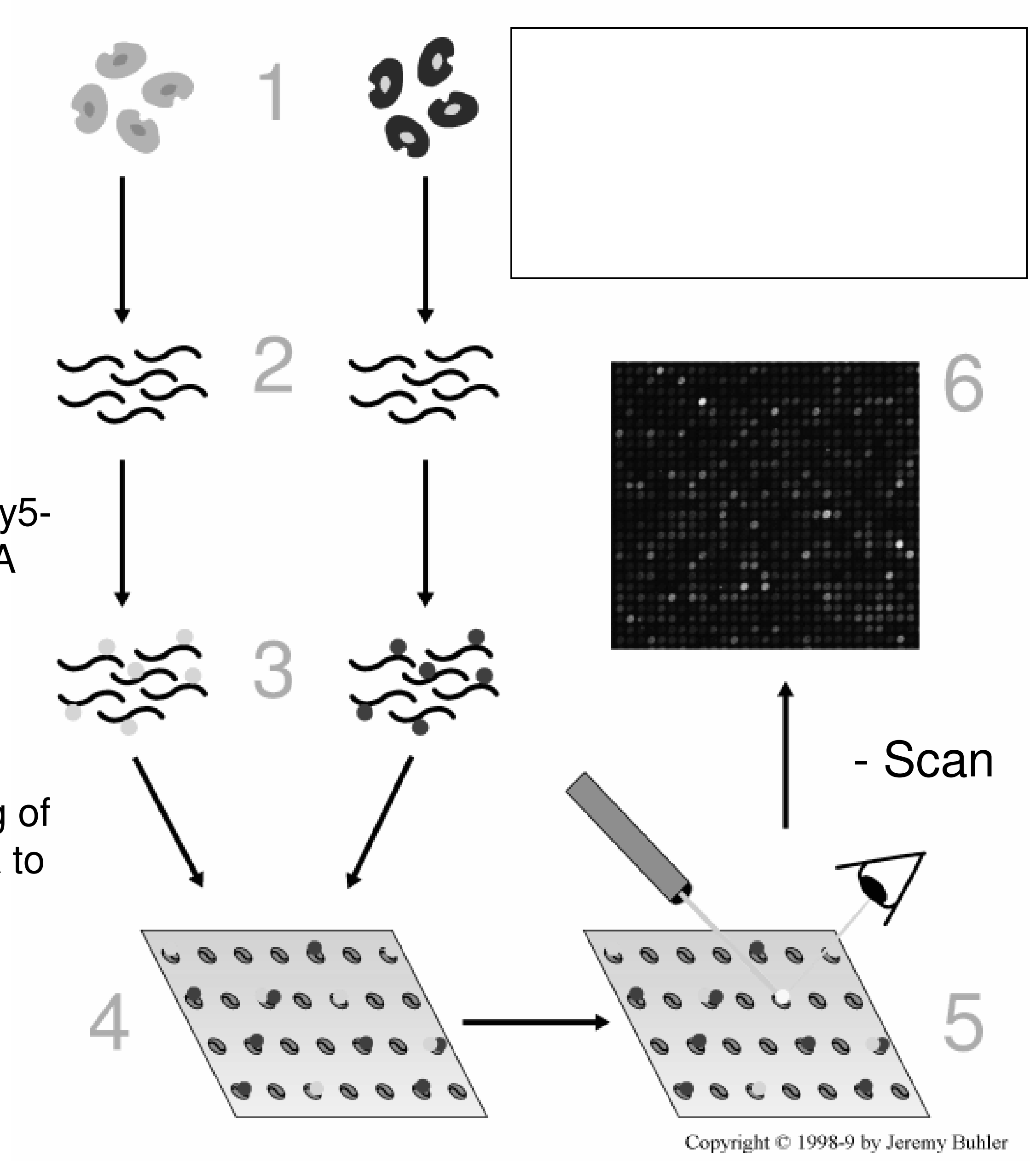
- mRNA is transcribed from the DNA
- mRNA (like DNA, but unlike proteins) binds to its complement



Hybridization and Scanning—Glass slide arrays

- Prepare Cy3, Cy5-labeled ss cDNA

- Hybridize 600 ng of labeled ss cDNA to glass slide array

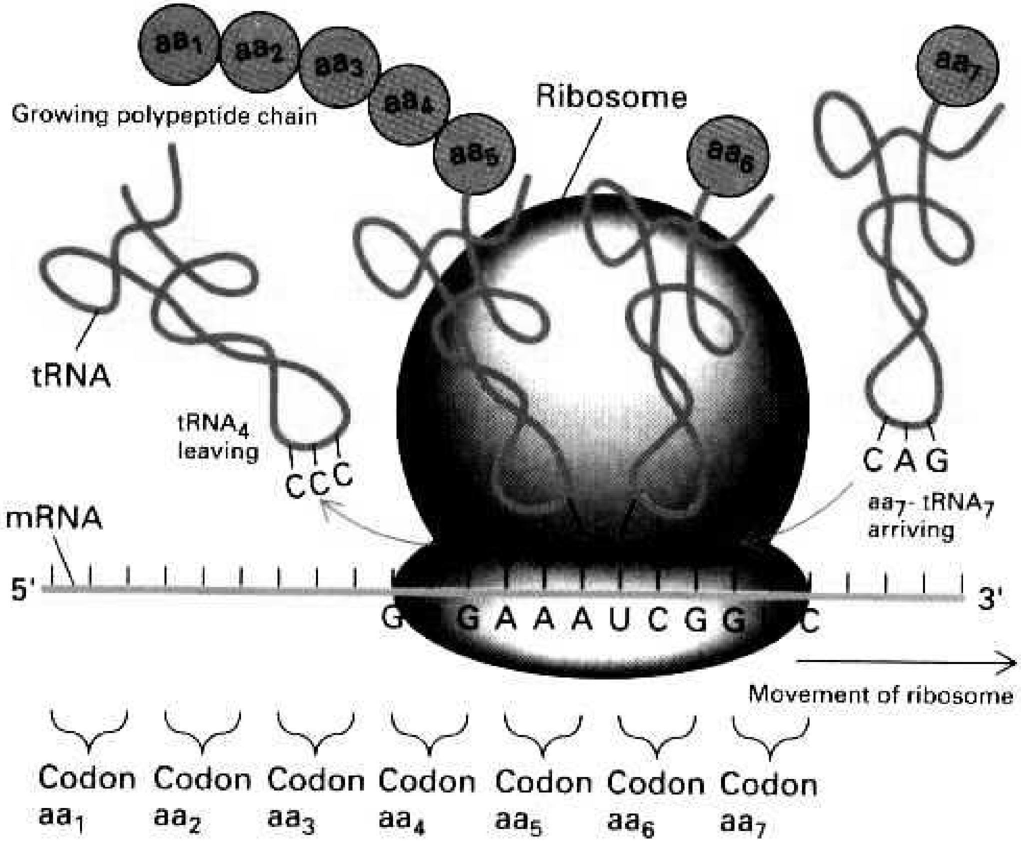


- Scan

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



Perturbation

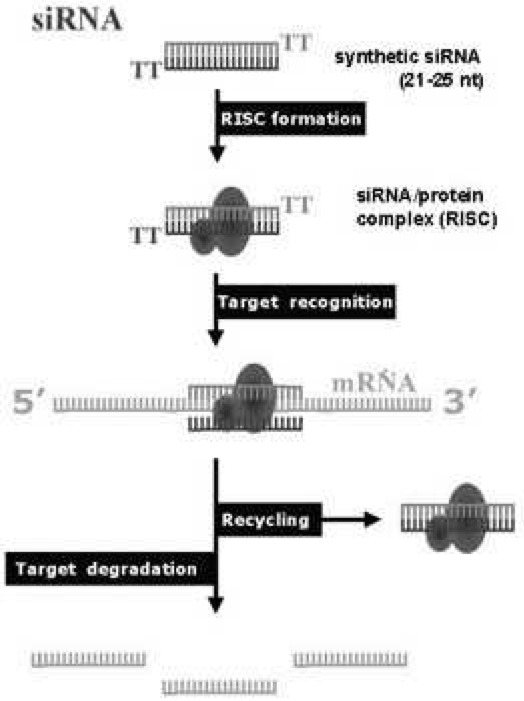
- In many cases we would like to perturb the systems to study the impacts of individual components (genes).
- This can be done in the sequence level by removing (knocking out) the gene of interest.
- Not always possible:

  - 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

Proteins

- Proteins are polypeptide chains of amino acids.
- Four levels of structure:

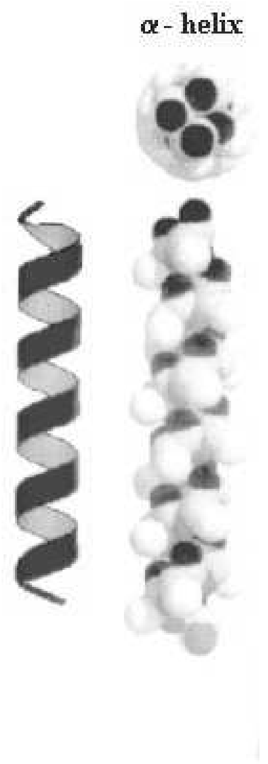
  - Primary Structure: The sequence of the protein

  - Secondary structure: Local structure in regions of the chain

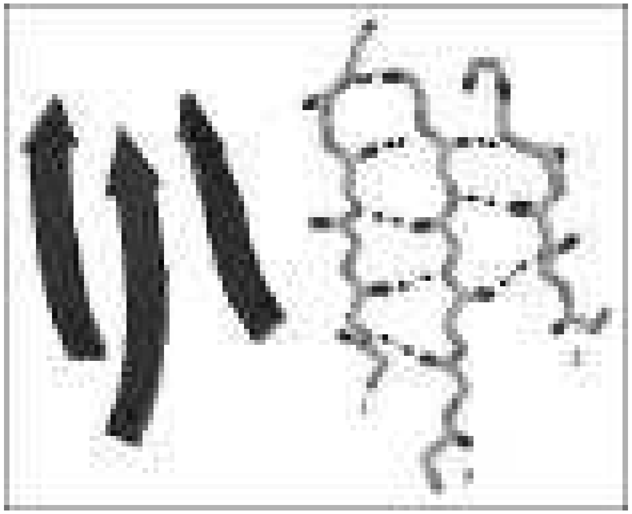
  - Tertiary Structure: Three dimensional structure

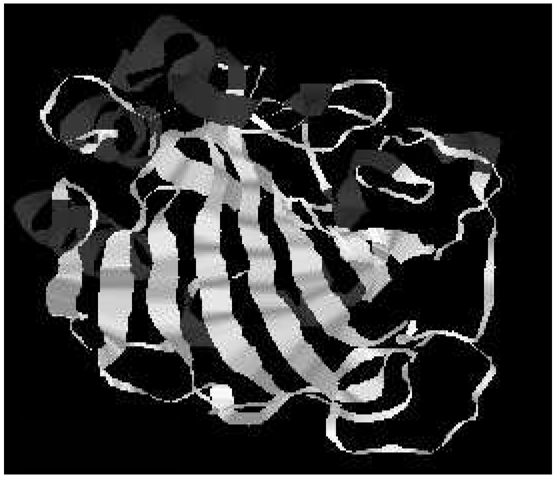
  - Quaternary Structure: multiple subunits

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.

Assigning Function to Proteins

- While almost 30000 genes have been identified in the human genome, relatively few have known functional annotation.

- Determining the function of the protein can be done in several ways.

  - Sequence similarity to other (known) proteins

  - Using domain information

  - Using three dimensional structure

  - Based on high throughput experiments (when does it functions and who it interacts with)

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

- We now have many sources of data, each providing a different view on the activity in the cell
  - Sequence (genes)
  - DNA motifs
  - Gene expression
  - Protein interactions
  - Image data
  - Protein-DNA interaction
  - Etc.

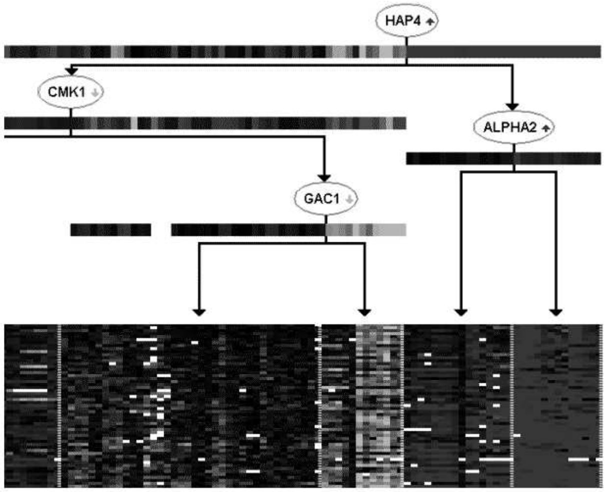
High throughput data

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

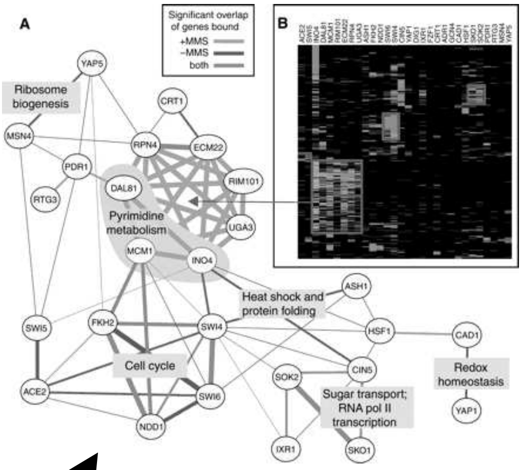
- Sequence (genes)

How to combine these different data types together to obtain a unified view of the activity in the cell is one of the focuses of this class

Reverse engineering of regulatory networks



Segal et al *Nature Genetics* 2003



Workman et al *Science* 2006

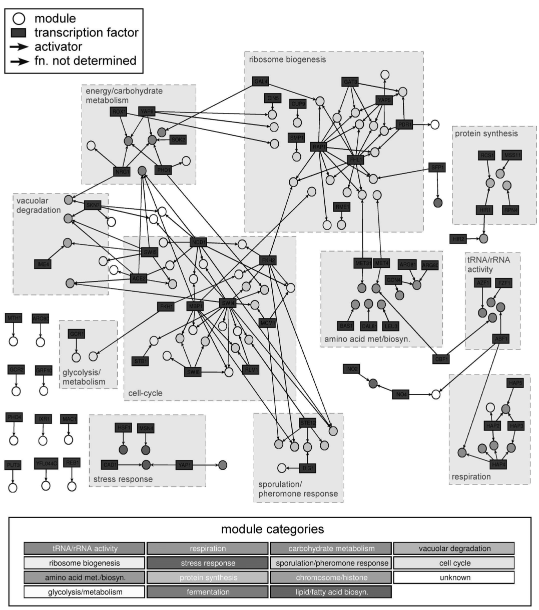


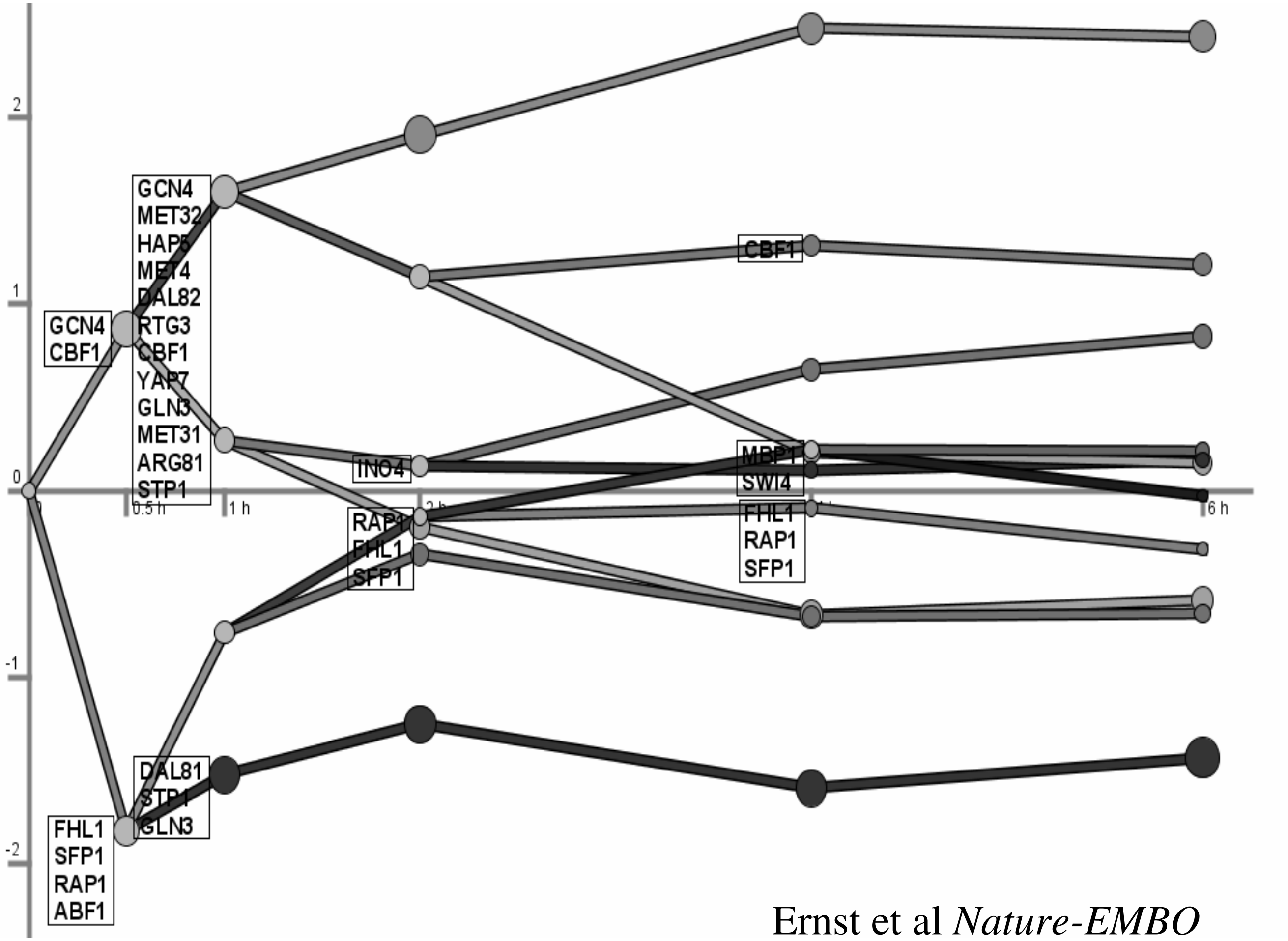
Figure 1: Rich media gene modules network

Bar-Joseph et al *Nature Biotechnology* 2003

- Gene expression
- Protein-DNA and gene expression

Dynamic regulatory networks

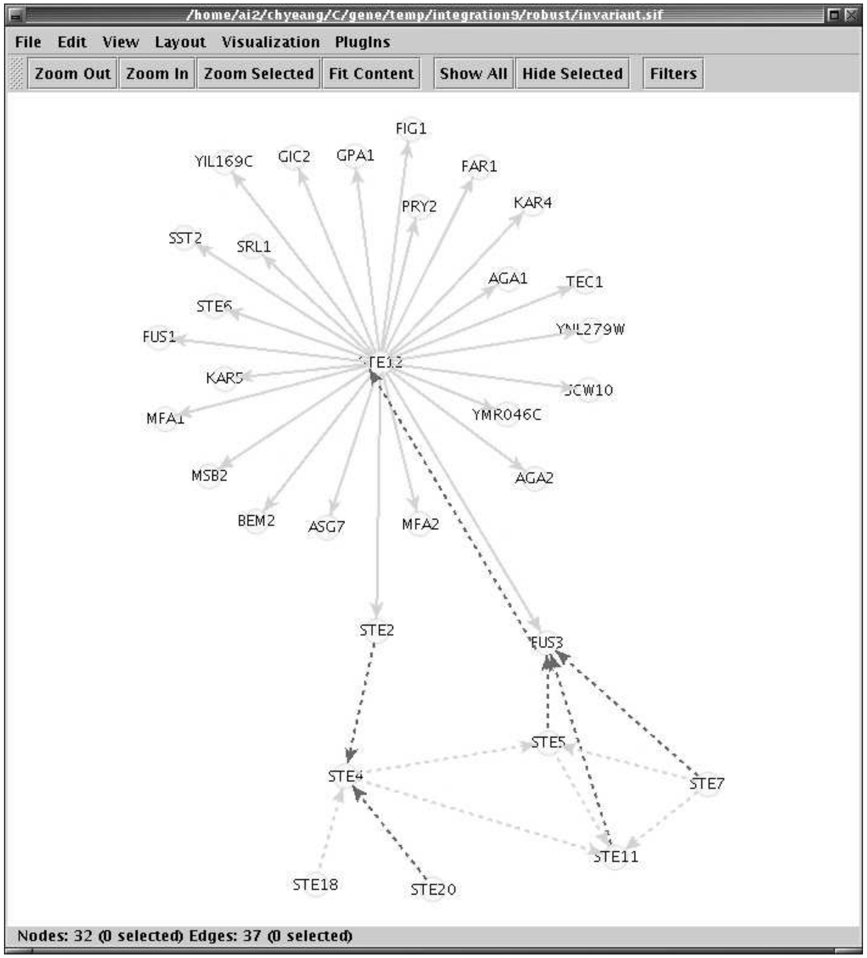
Protein-DNA, motif and time series gene expression data



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

Physical networks

Protein-DNA, protein-protein and gene expression data



Nodes: 32 (0 selected) Edges: 37 (0 selected)

Yeang *et al*, Genome Bio. 2005

What you should remember

- Course structure:

  - Genomes (genetics)

  - Genes and regulatory regions (sequence analysis)

  - mRNA and high throughput methods (microarrays)

  - Systems biology