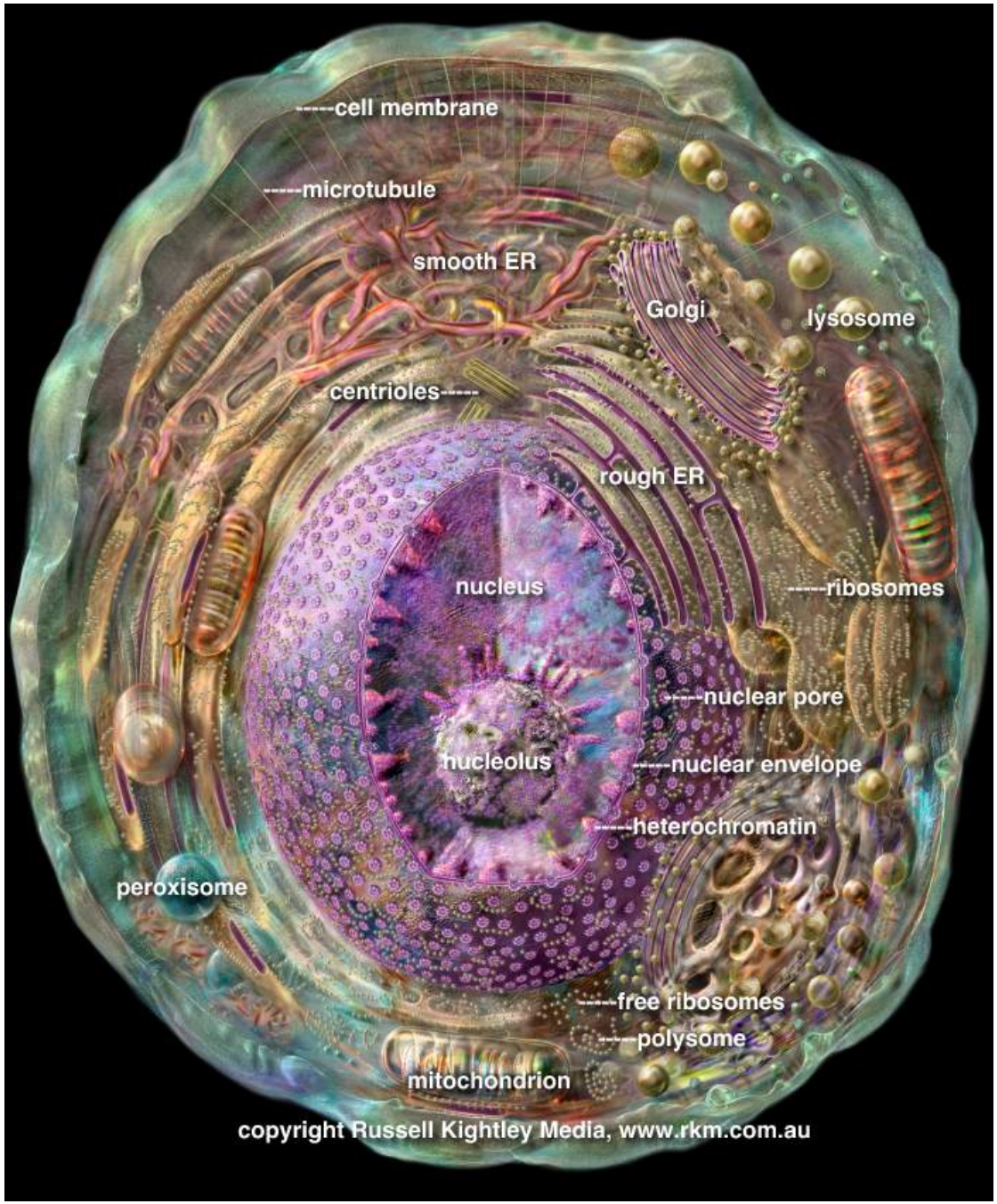
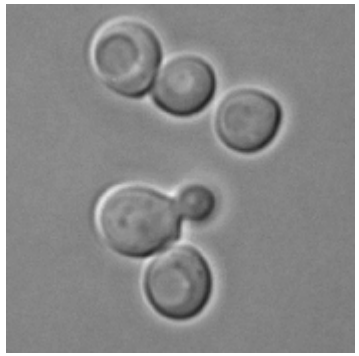
Molecular biology 101
or
“why bother?”

Cells are fundamental working units of all organisms



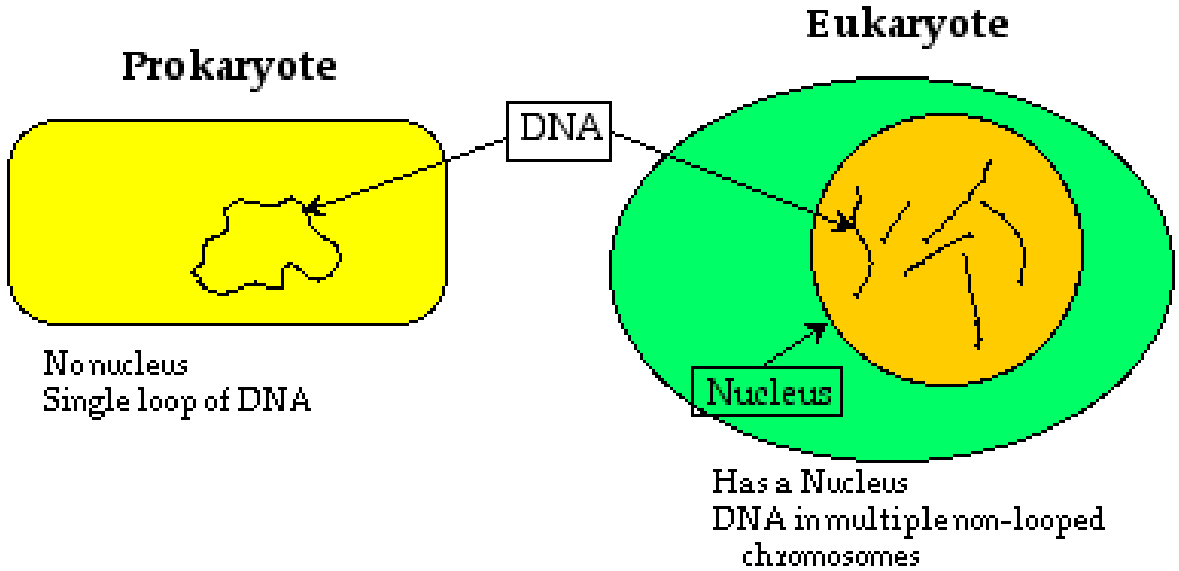


Yeast are unicellular organisms

Humans are multi-cellular organisms

Understanding **how a cell works** is critical to understanding how the organism functions

Prokaryotes vs. Eukaryotes



Yeast is a eukaryote just like humans. Fundamental biological processes are very similar.

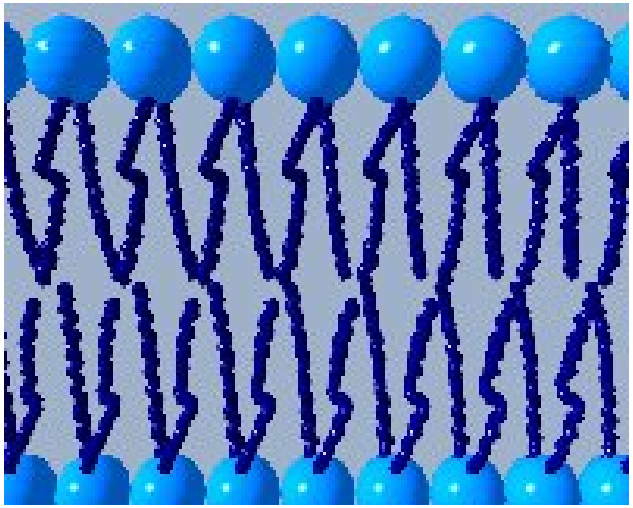
Biological macromolecules

What are the main players in molecular biology?

What is DNA, RNA, protein, lipid?

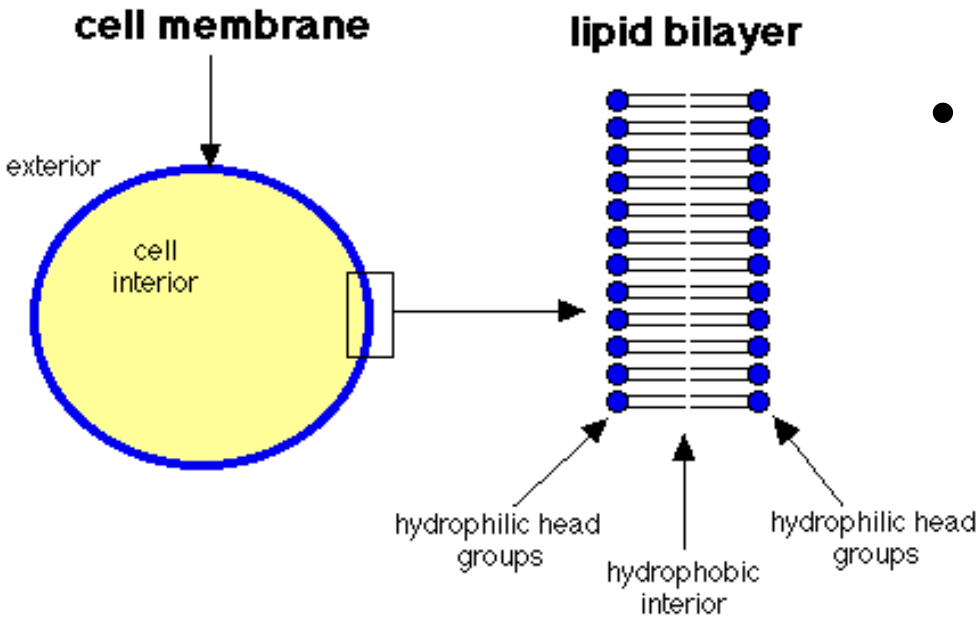
Key biological macromolecules

- Lipids:
  - mostly structural function
  - Construct compartments that separate inside from outside
- DNA
  - Encodes hereditary information
- Proteins
  - Do most of the work in the cell
  - Form 3D structure and complexes critical for function

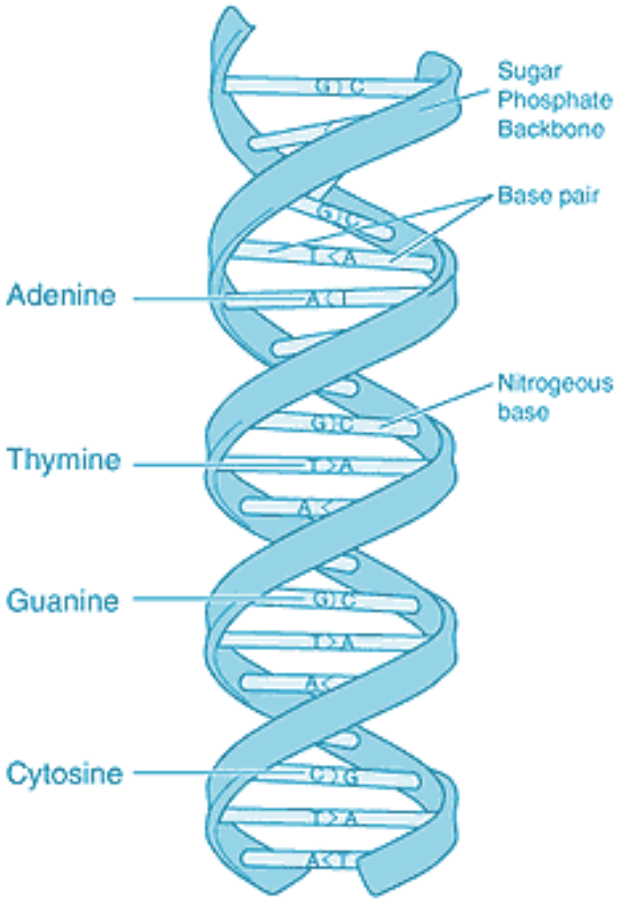


Lipids

- Each lipid consists of a hydrophilic (water loving) and hydrophobic fragment
- Spontaneously form lipid bilayers => membranes

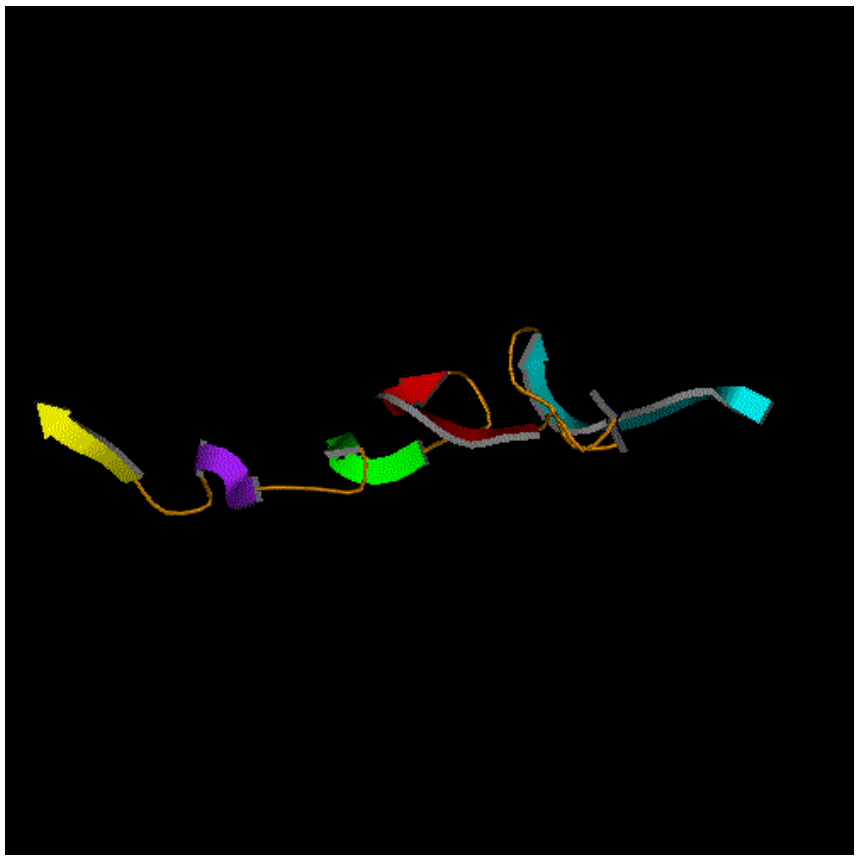


DNA



- Uses alphabet of 4 letters {ATCG}, called bases
- Encodes genetic information in triplet code
- Structure: a double helix

Proteins



Courtesy of the Zhou Laboratory, The State University of New York at Buffalo

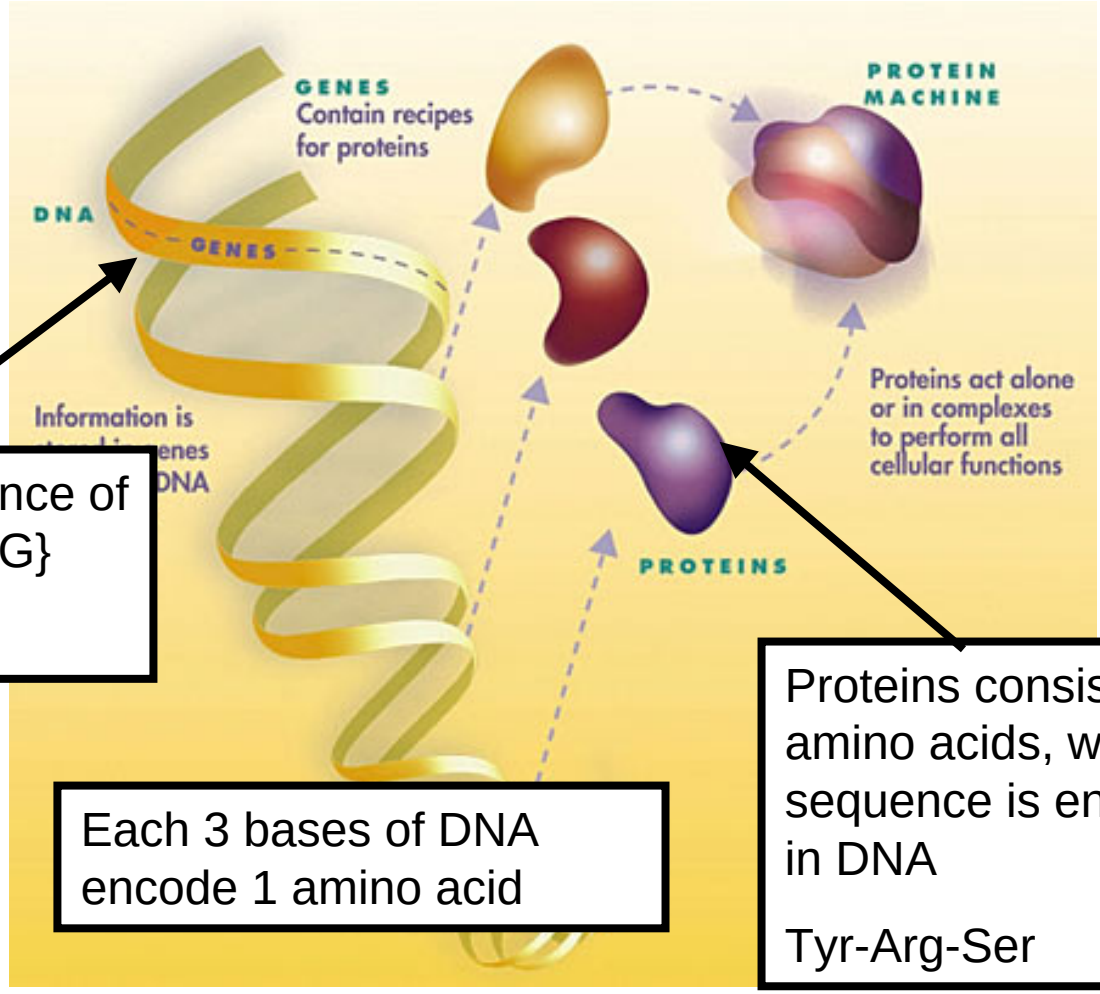
- A sequence of amino acids (alphabet of 20)
- Each amino acid encoded by 3 DNA bases
- Perform most of the actual work in the cell
- Fold into complex 3D structure

How does a cell function?
The “Central Dogma” of biology

How are proteins made?

What are translation & transcription?

How does a cell function?



DNA is a sequence of bases {A, T, C, G}

TAT-CGT-AGT

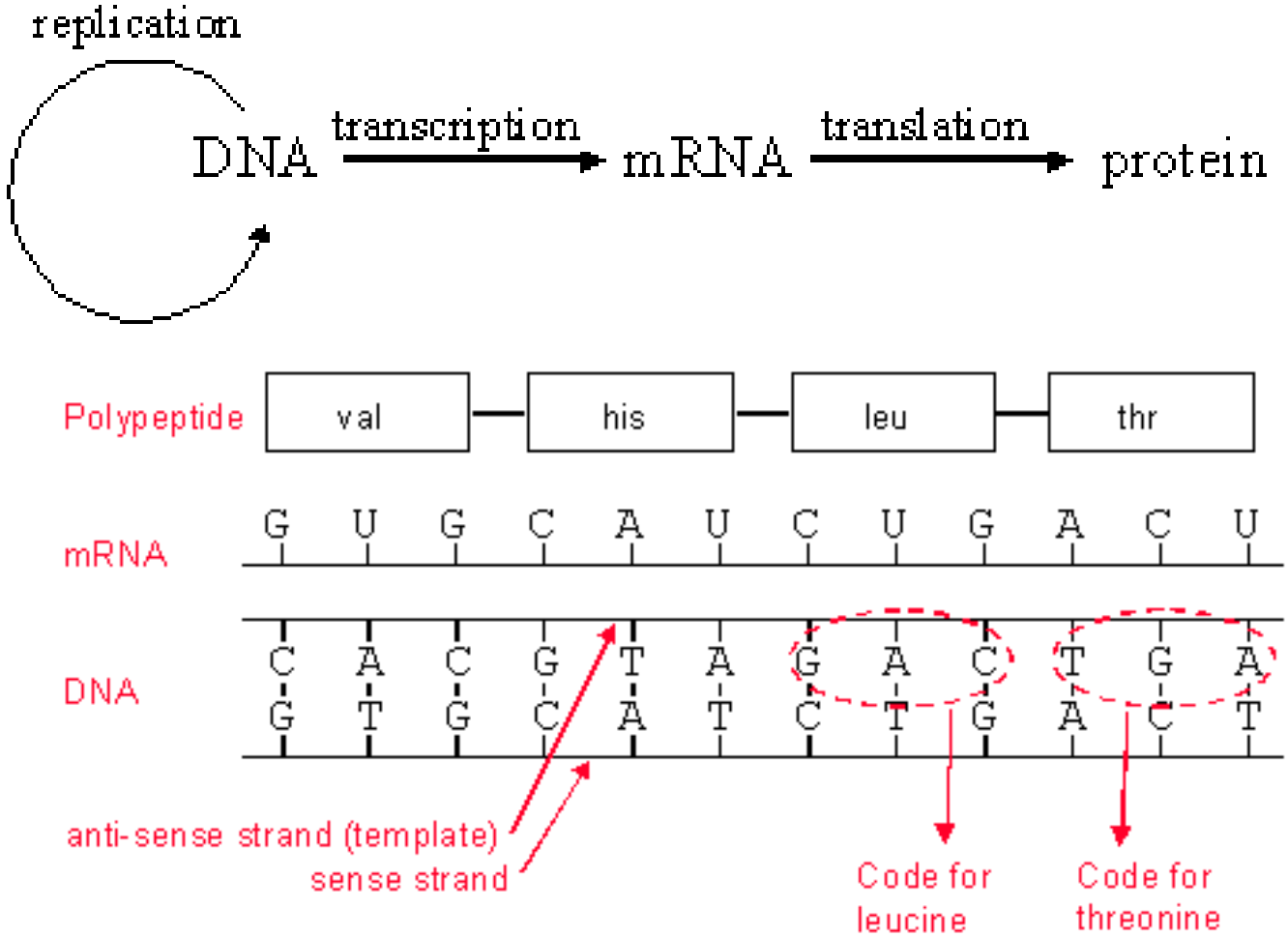
Each 3 bases of DNA encode 1 amino acid

Proteins consist of amino acids, whose sequence is encoded in DNA

Tyr-Arg-Ser

Courtesy U.S. Department of Energy Genomes to Life program

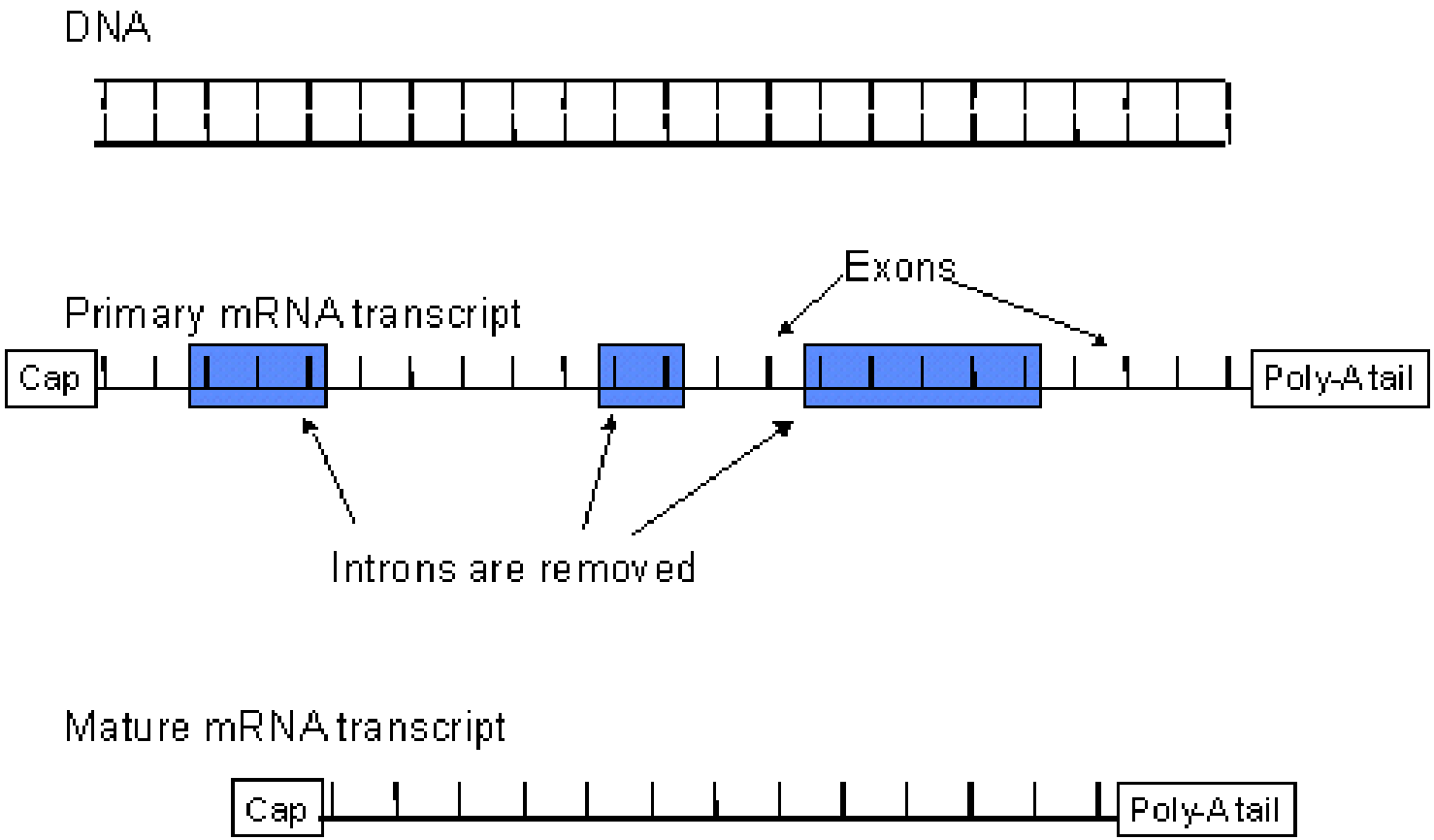
DNA-RNA-protein



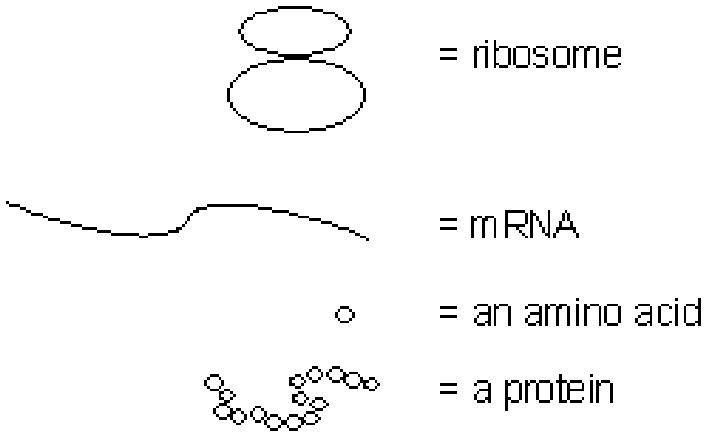
Transcription (DNA->RNA)

- DNA unwinds.
- ***RNA polymerase*** recognizes a specific base sequence in the DNA called a ***promoter*** and binds to it. The promoter identifies the start of a gene, which strand is to be copied, and the direction that it is to be copied.
- Complementary bases are assembled (U instead of T).
- A ***termination code*** in the DNA indicates where transcription will stop.
- The mRNA produced is called a ***mRNA transcript***.

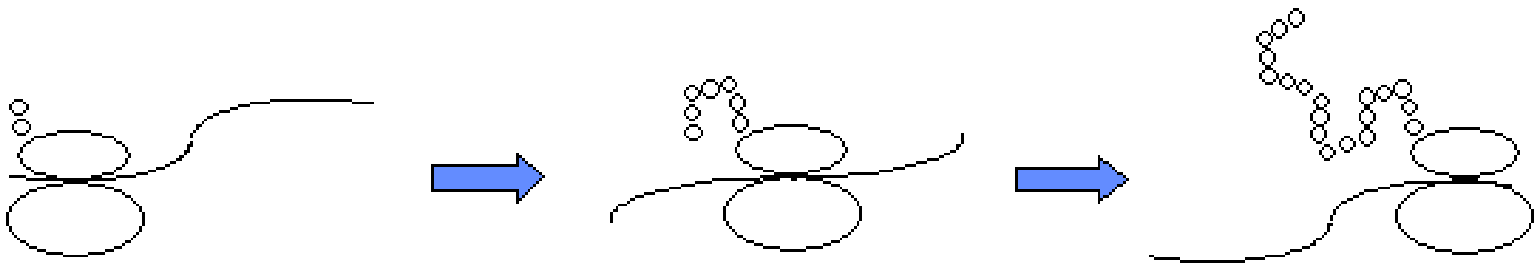
mRNA processing (Euk)



Translation (mRNA->protein)



Ribosomes attach to mRNA and move along it to make protein by adding amino acids in order based on mRNA sequence



Gene regulation: from circuits to networks

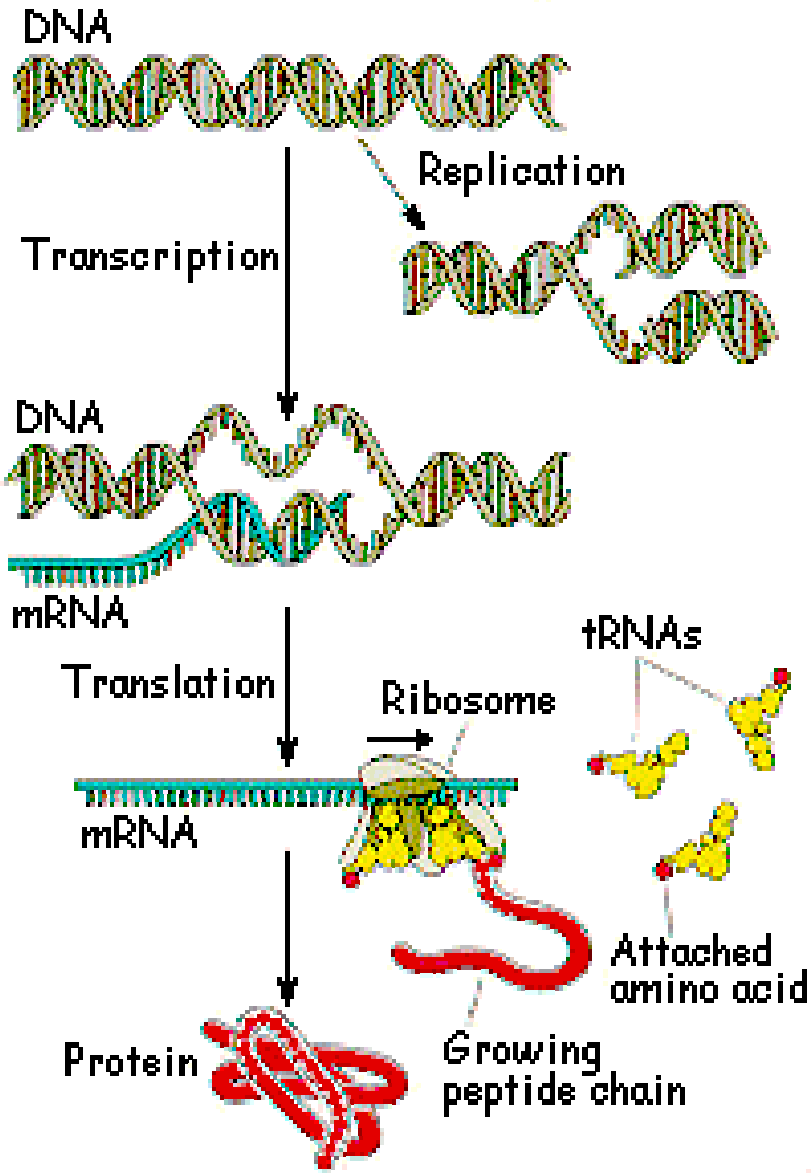
How are genes regulated?

How are biological circuits formed?

What are biological networks?

Genes vs. proteins

- Genes are units of inheritance
- They are static blueprints
- It’s proteins (dynamic) that do most of the work
- The process of making mRNA, and then protein from a gene (or genes) is called GENE EXPRESSION
- It’s the control of gene expression that causes most phenotypic differences in organisms



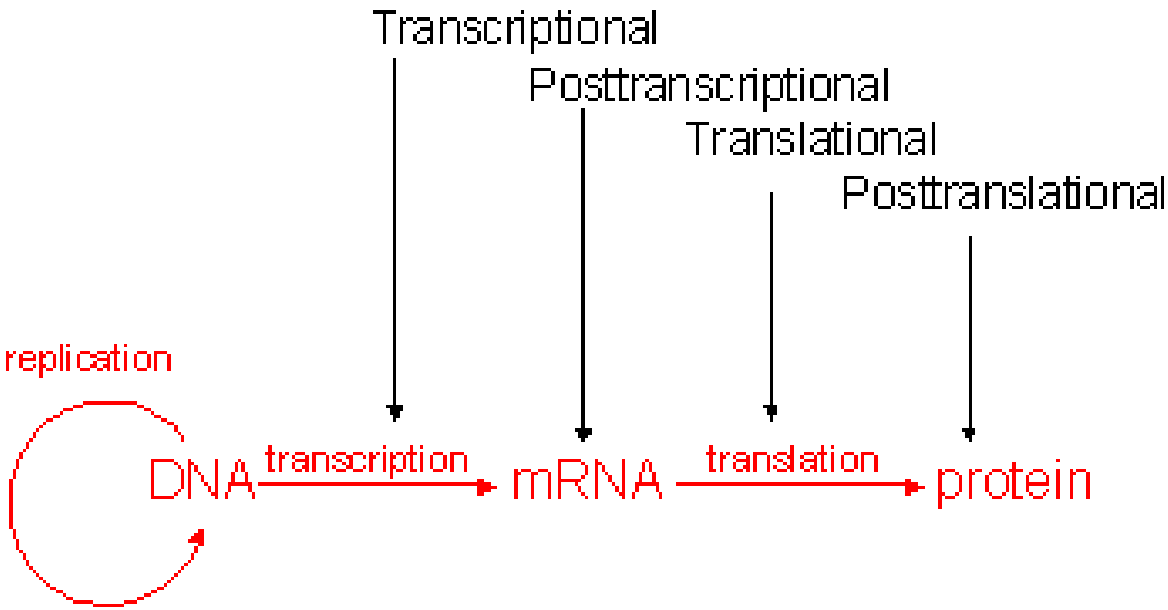
Opportunities for gene regulation

- Opening of DNA duplex
- Transcription
- mRNA stability
- Translation
- Protein stability
- Protein modification

Control of gene expression

- GE is controlled on many levels (in eukaryotes)

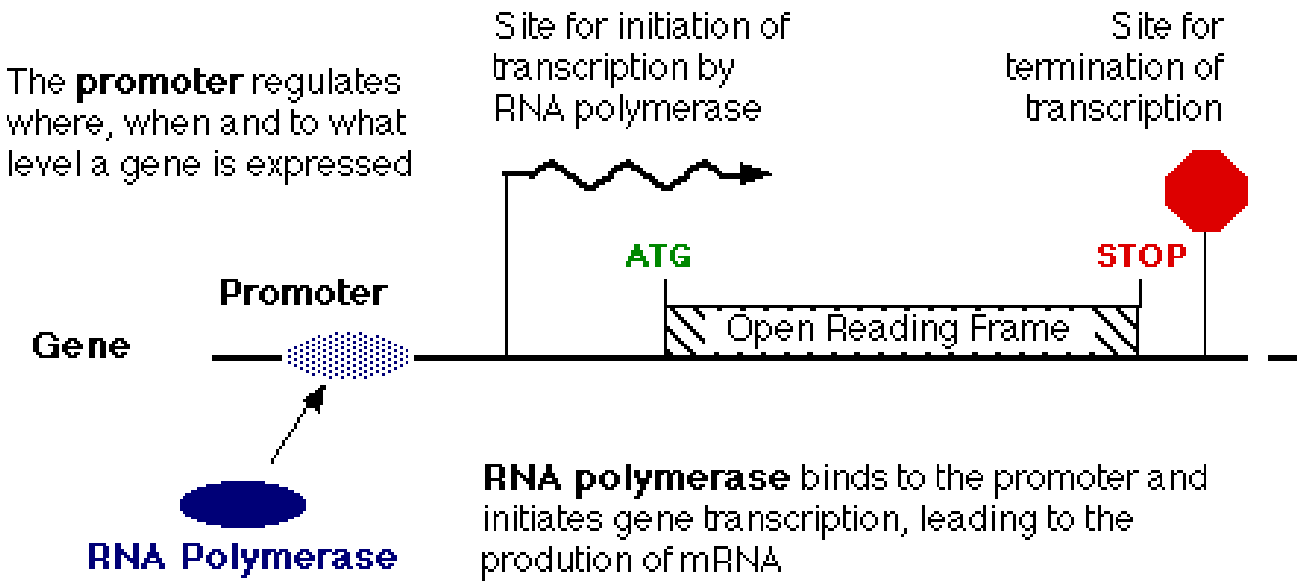
Transcriptional - prevent transcription.
Posttranscriptional - control or regulate mRNA after it has been produced.
Translational - prevent translation, often involve protein factors needed for translation.
Posttranslational - act after the protein has been produced.

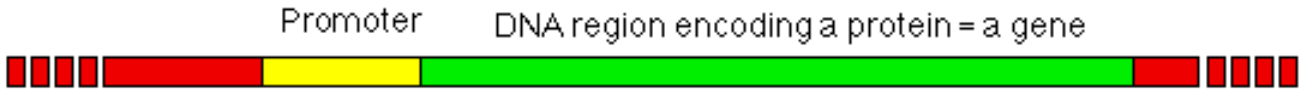


Transcriptional regulation

- Thought to be the most used
- Does not waste intermediate products (mRNA, protein, etc)
- But transcriptional regulation is slow, and thus may not be used in cases when fast, transient regulation is necessarily

Promoters are important elements for gene expression

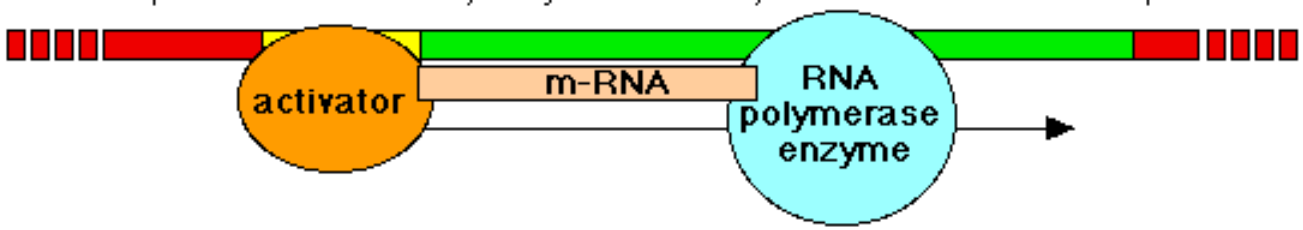




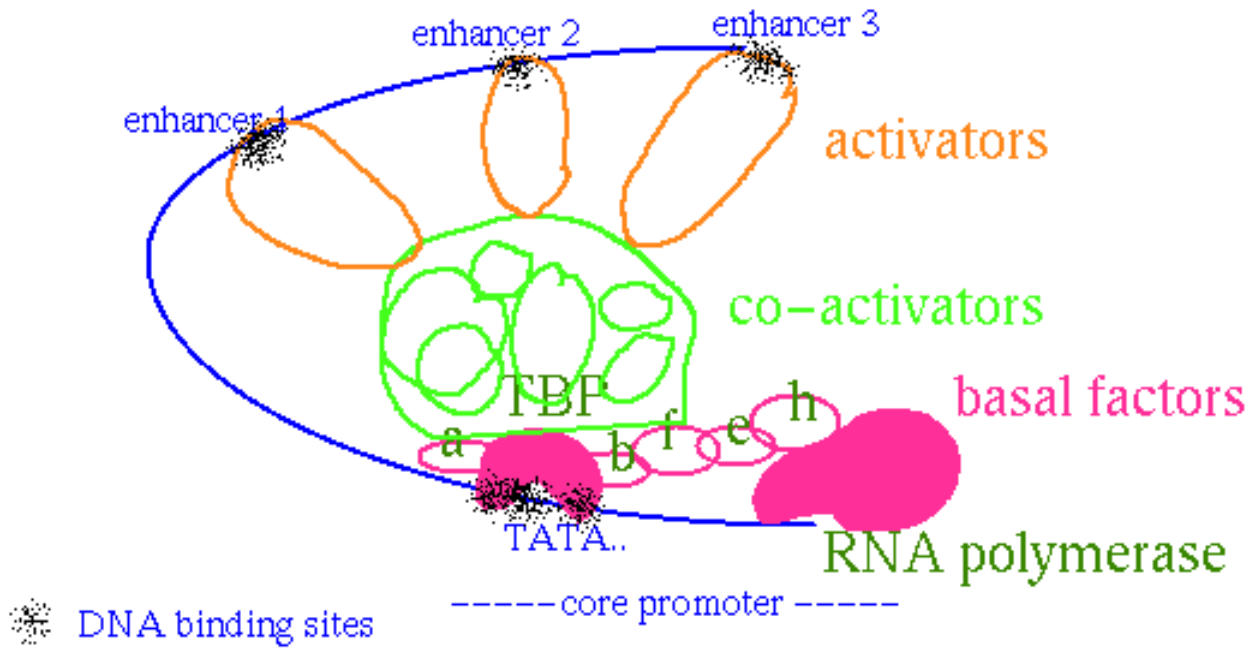
Will this gene be expressed? Depends on
whether specific activators bind to the promoter

State 1: no activation, enzyme can't bind, so no RNA is made. Gene is not expressed.

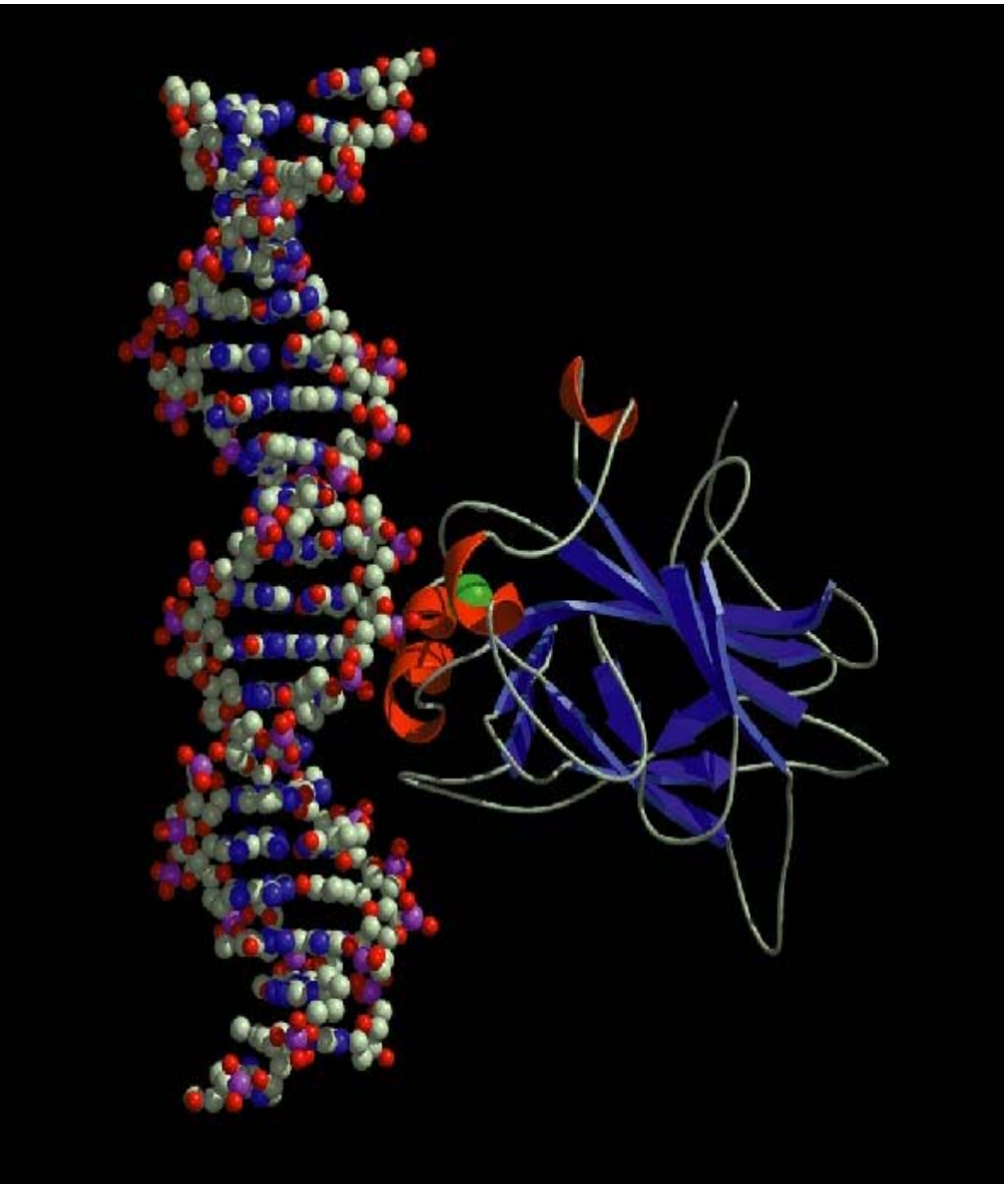
State 2: promoter is activated, enzyme cant bind, RNA is made. Gene is expressed.



Transcriptional activation



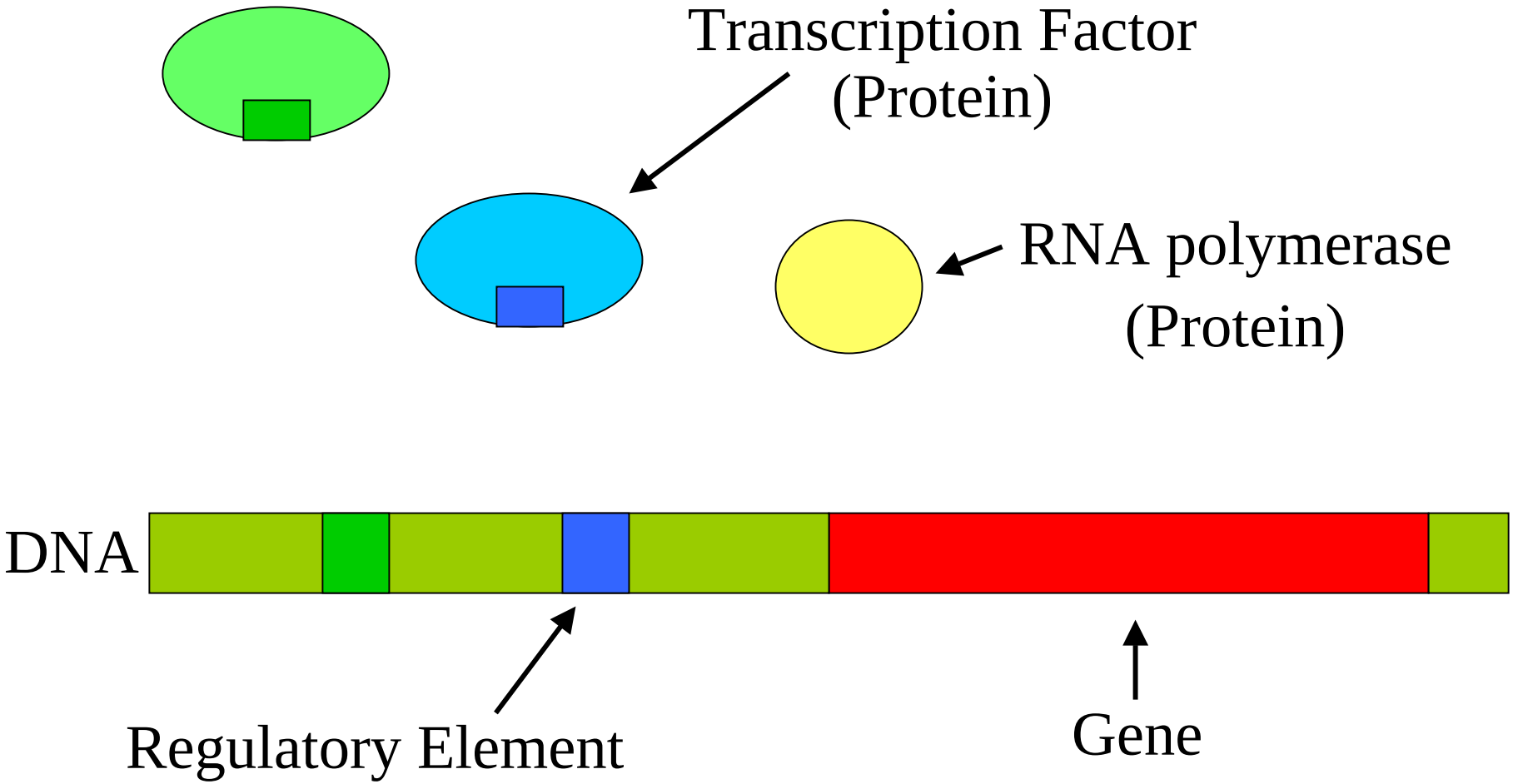
Transcription Factors Bind DNA



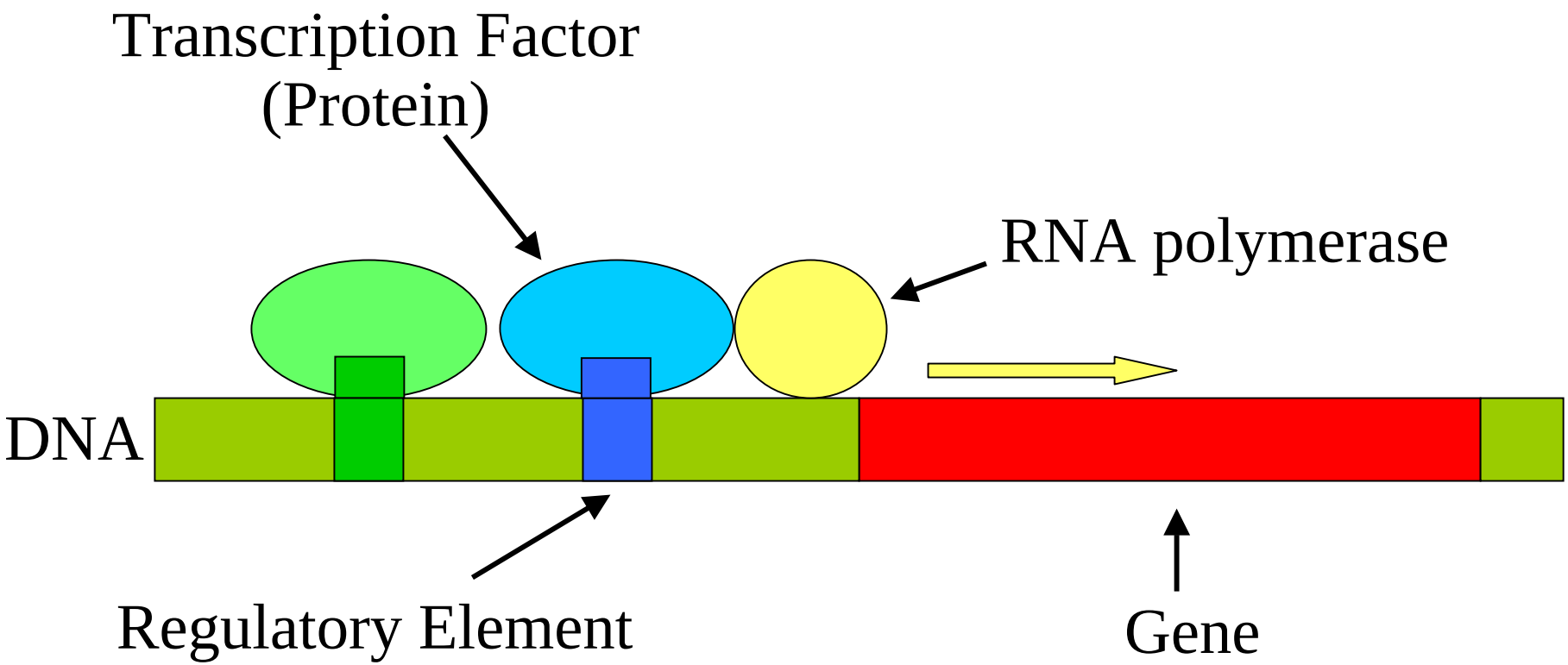
Transcription factors bind DNA in a specific manner.

Binding recognizes DNA substrings called **regulatory motifs**

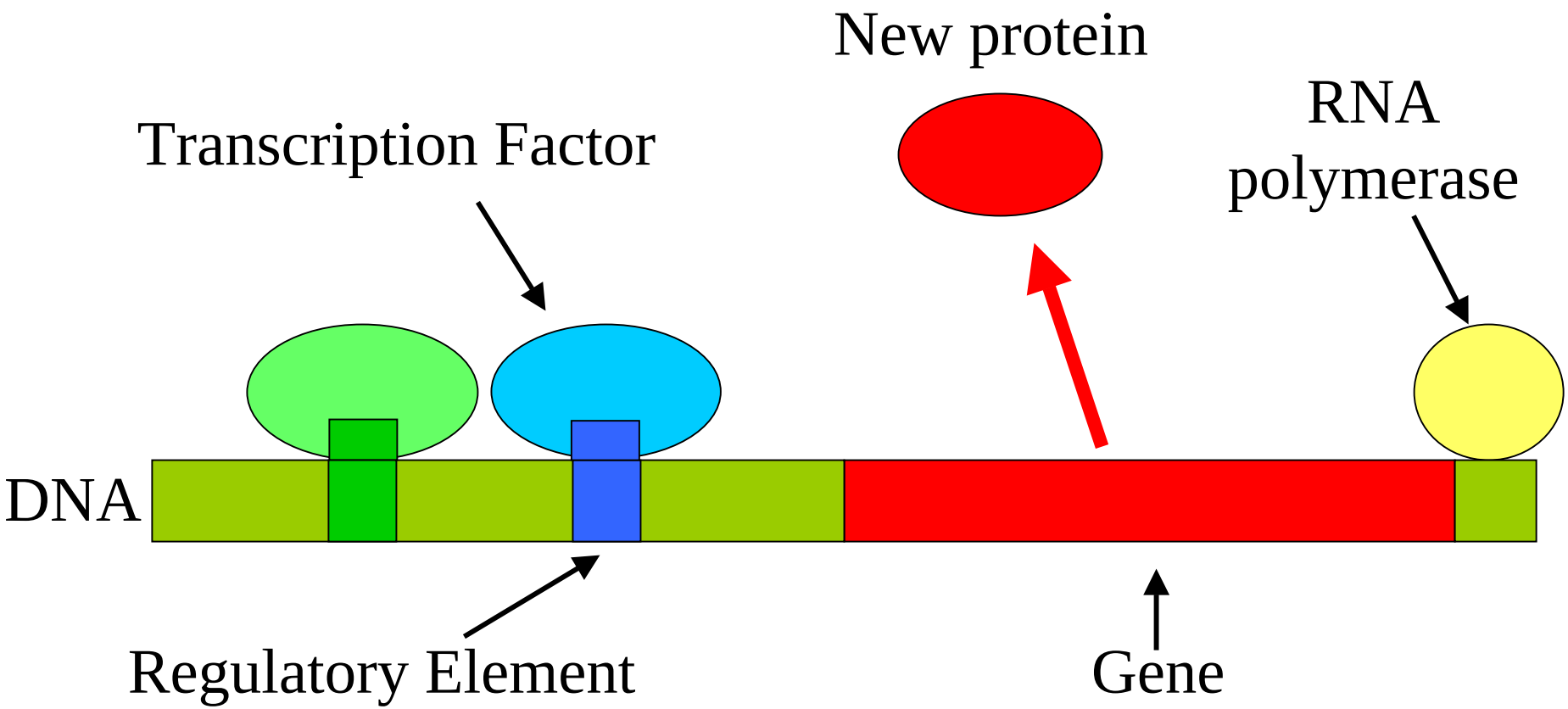
Regulation of Genes



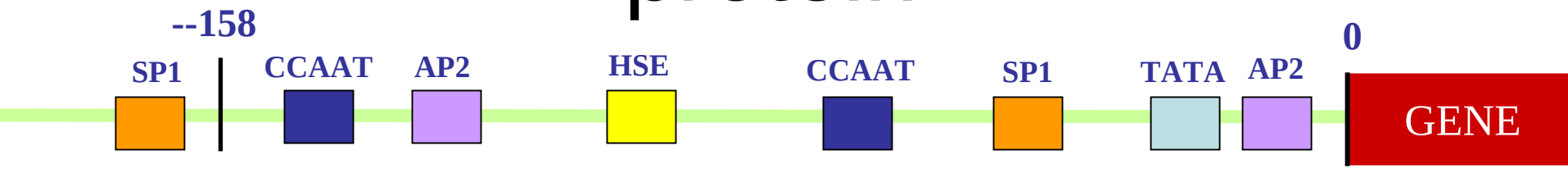
Regulation of Genes



Regulation of Genes



Example: A Human heat shock protein

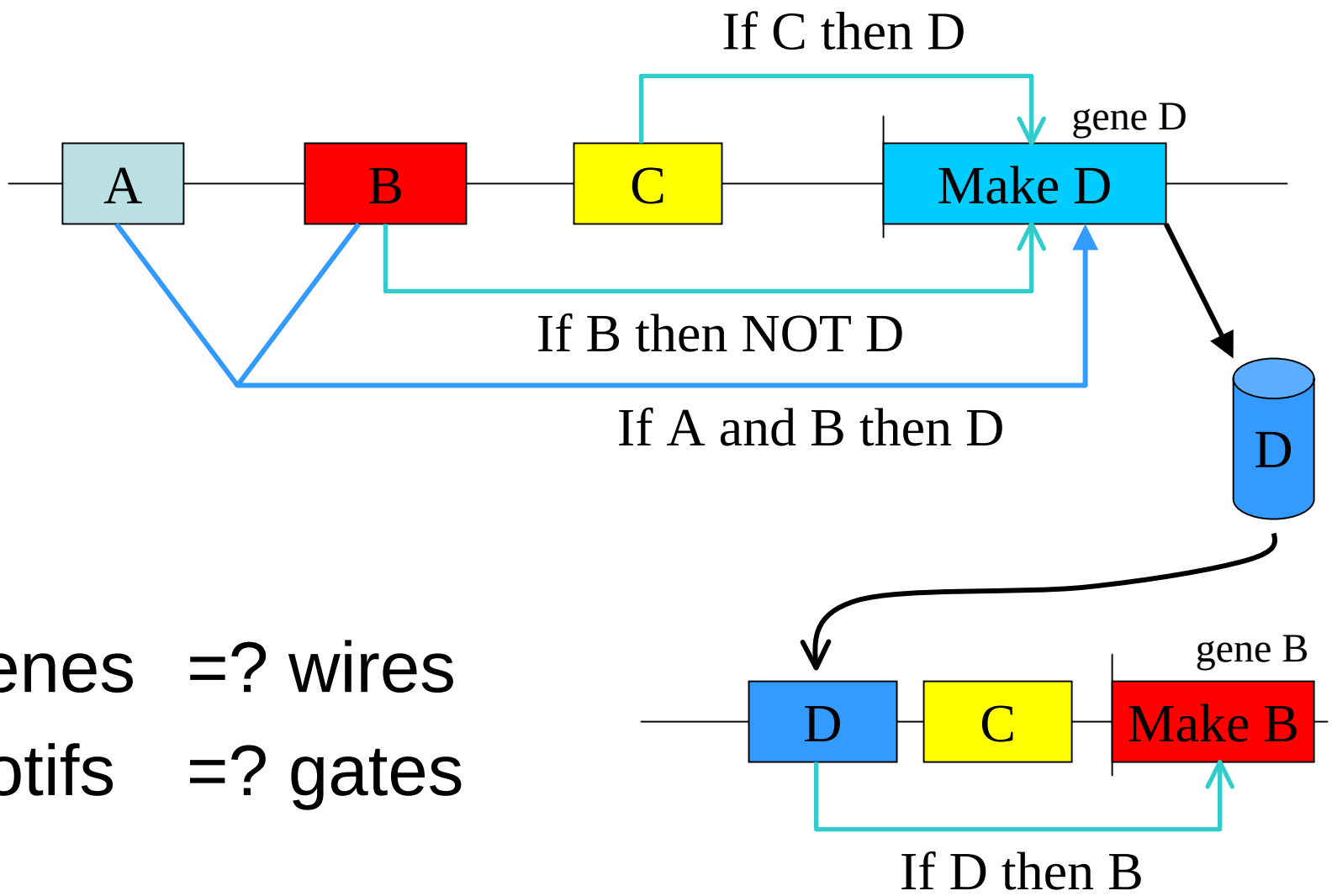


promoter of heat shock hsp70

- TATA box: positioning transcription start

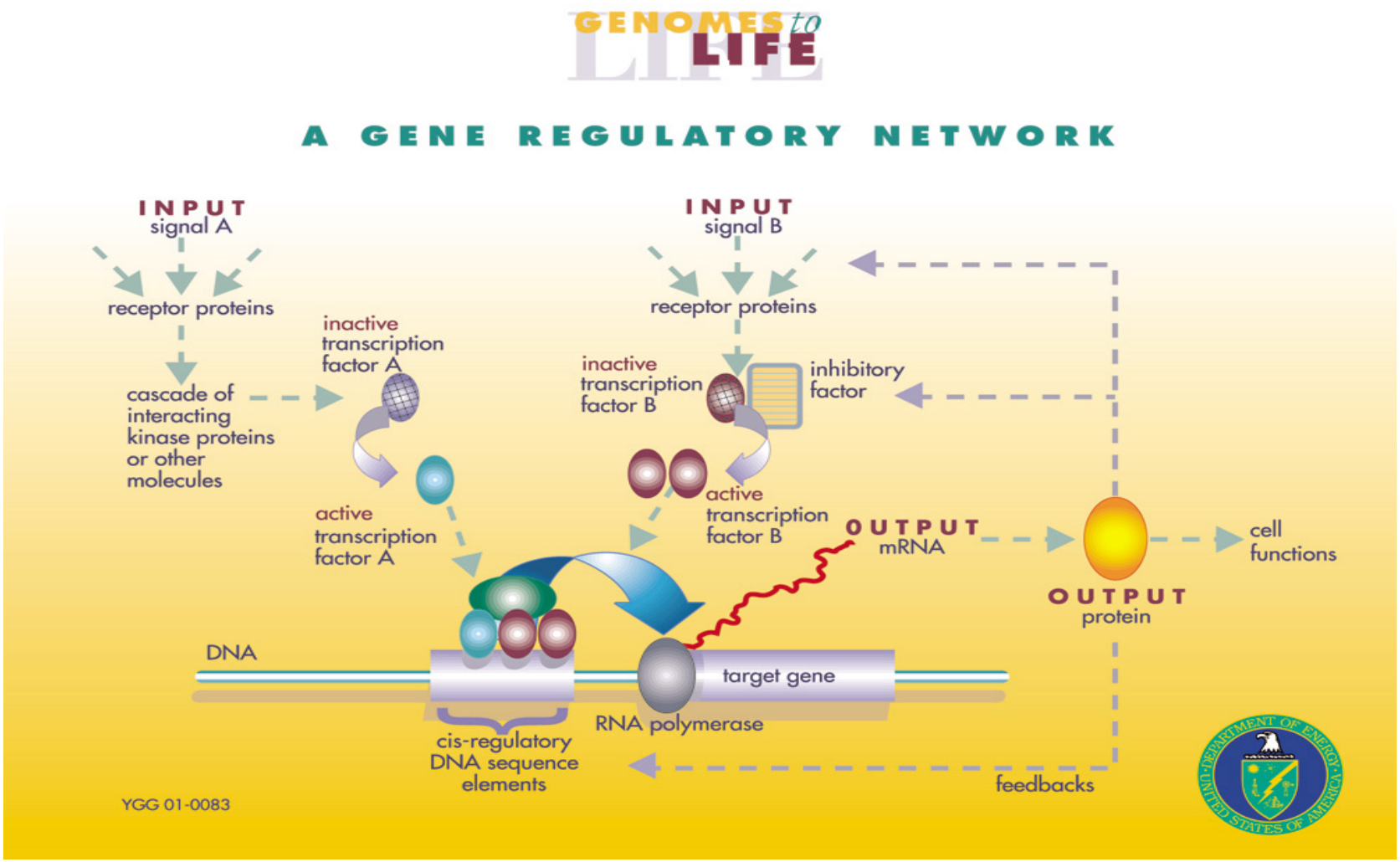
- TATA, CCAAT: constitutive transcription
- GRE: glucocorticoid response
- MRE: metal response
- HSE: heat shock element

Gene Regulatory Circuit



- Genes =? wires
- Motifs =? gates

Regulatory Networks



What is genomics all about?

The “omes” in biology.

Why bioinformatics?

What is “systems biology”?

The “omes”

- Genome – organism’s complete set of DNA
  - Relatively stable through an organism’s lifetime
  - Size: from 600,000 to several billion bases
  - Gene is a basic unit of heredity (only 2% of the human genome)
- Proteome – organism’s complete set of proteins
  - Dynamic – changes minute to minute
  - Proteins actually perform most cellular functions, they are encoded by genes (not a 1-to-1 relationship)
  - Protein function and structure form molecular basis for disease

Beyond the “omes” – systems biology

- Understanding the function and regulation of cellular machinery, as well as cell-to-cell communication on the molecular level
- Why? Because most important biological problems are fundamentally systems-level problems
  - Systems-level understanding of disease (e.g. cancer)
  - Molecular medicine
  - Gene therapy

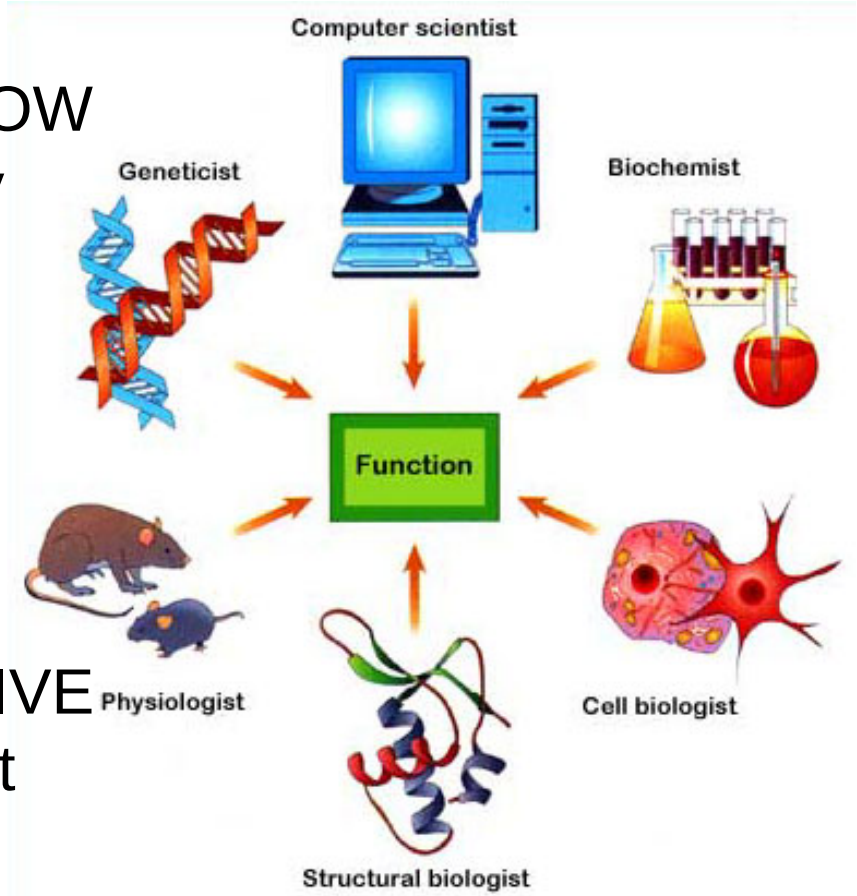
Systems-level challenges

- **Gene function annotation – what does a gene do**
  - ~30,000 genes in the human genome => systems-level approaches necessary
  - A modern human microarray experiment produces ~500,000 data points => computational analysis & visualization necessary
  - Many high-throughput functional technologies => computational methods necessary to integrate the data
- **Biological networks – how do proteins interact**
  - Large amounts of high-throughput data => computation necessary to store and analyze it
  - Data has variable specificity => computational approaches necessary to separate reliable conclusions from random coincidences
- **Comparative genomics – comparing data between organisms**
  - Need to map concepts across organisms on a large scale => practically impossible to do by hand
  - High amount of variable quality data => computational methods needed for integration, visualization, and analysis
  - Data often distributed in databases across the globe, with variable schemas etc => data storage and consolidation methods needed

Function

- To study WHAT proteins DO, HOW they INTERACT, and HOW they are REGULATED, need data beyond genomic sequence

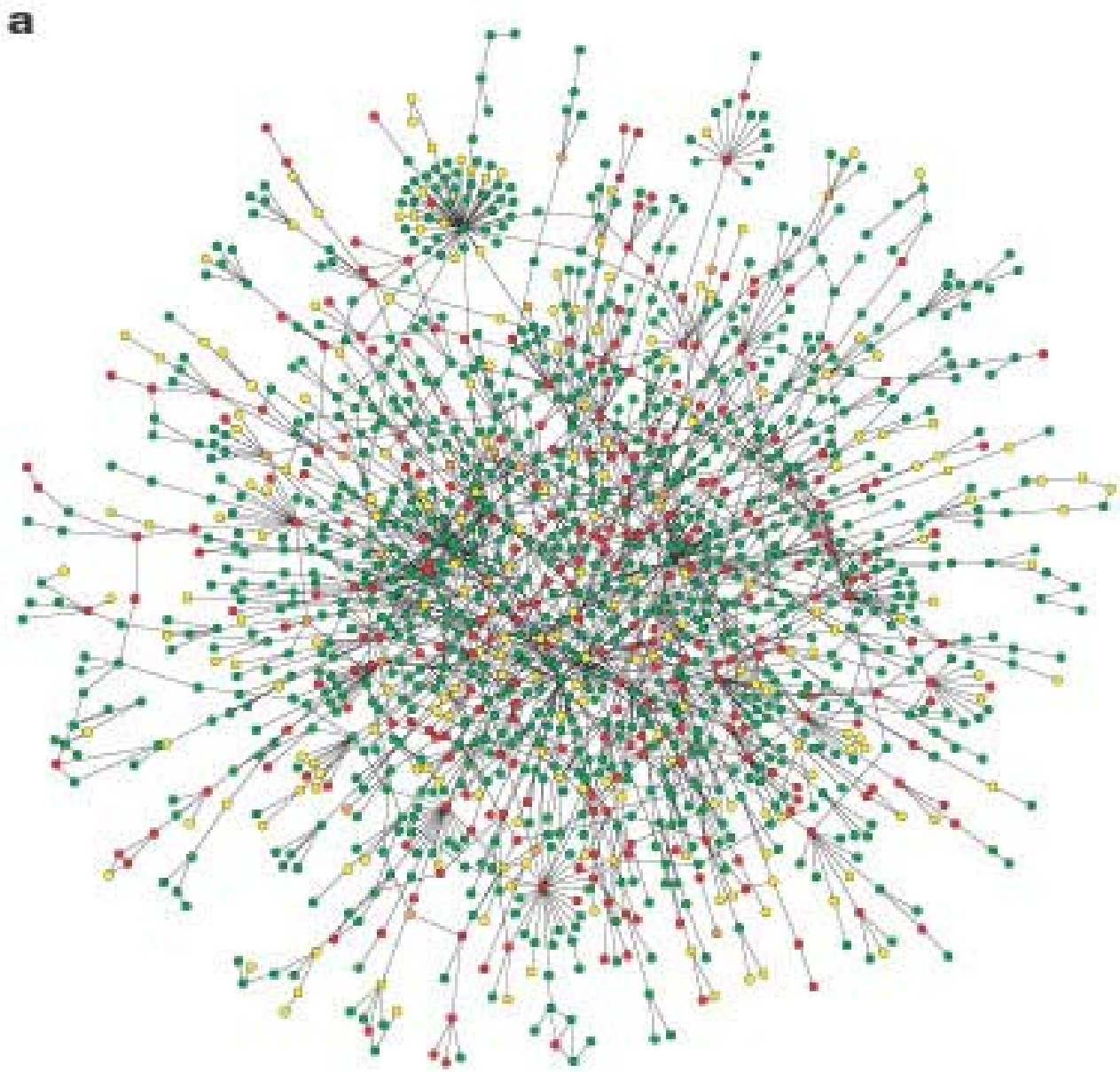
- Genomics/Bioinformatics is fundamentally a COLLABORATIVE and MULTIDISCIPLINARY effort

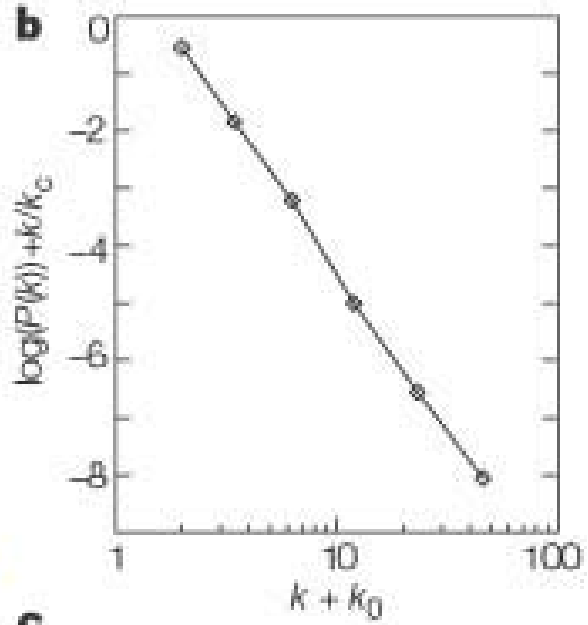


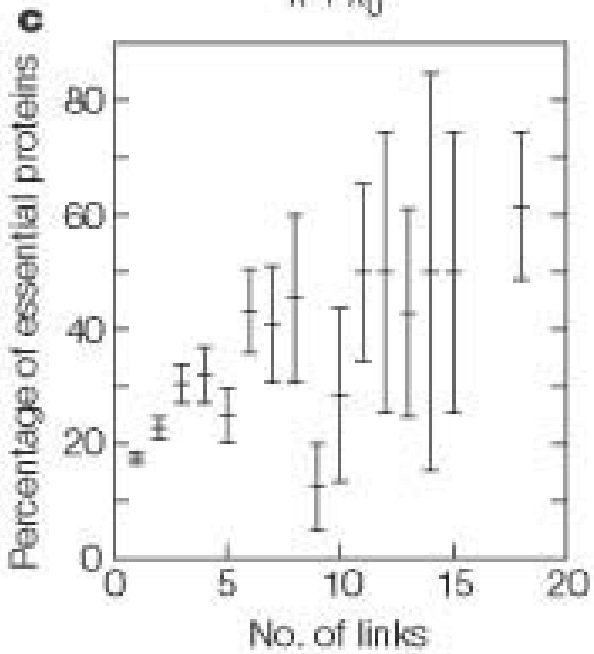
Biological networks

- Interaction maps (no directions)
- Pathway models (dynamic or static)
- Metabolic networks
- Genetic regulatory networks

Yeast interaction network



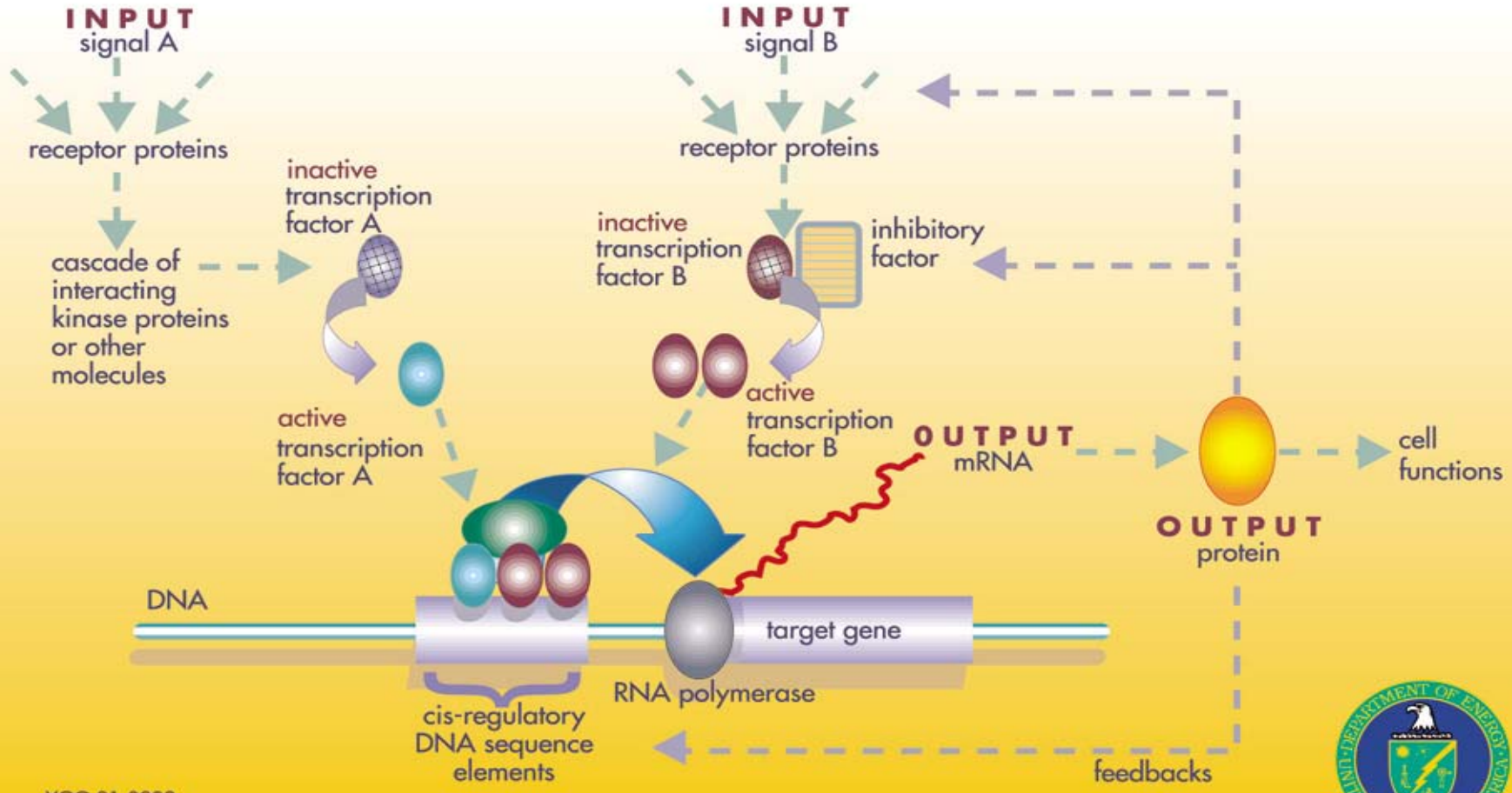




Hawoong Jeong et al. Oltvai Centrality and lethality of protein networks. Nature 411, 41-42 (2001)

GENOMES to LIFE

A GENE REGULATORY NETWORK



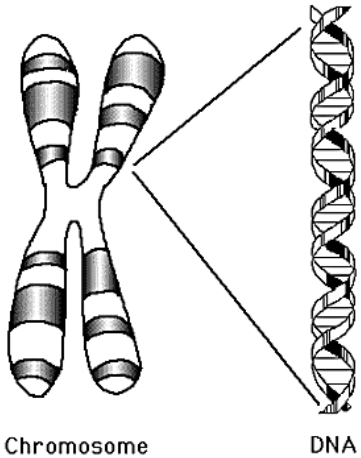
YGG 01-0083

Gene expression microarrays – one type of high-throughput functional data

Why microarray analysis: the questions

- Large-scale study of biological processes
- What is going on in the cell at a certain point in time?
- On the large-scale genetic level, what accounts for differences between phenotypes?
- Sequence important, but genes have effect through expression

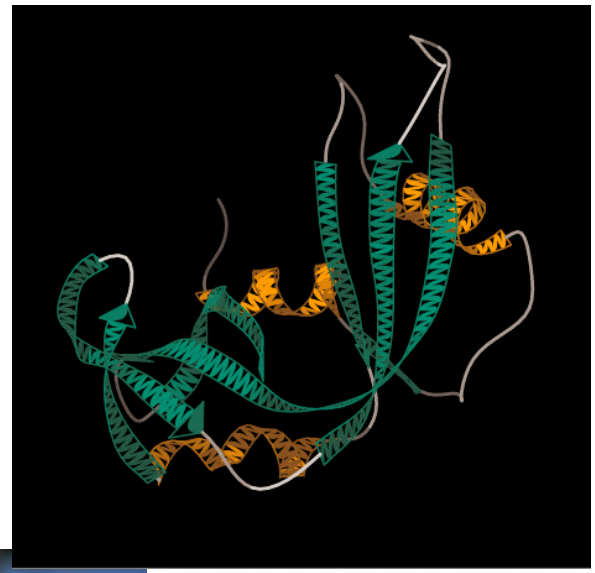
Why study gene expression



DNA

Gene Expression

Proteins



People

Microarray technology - example of high-throughput data

What is it?

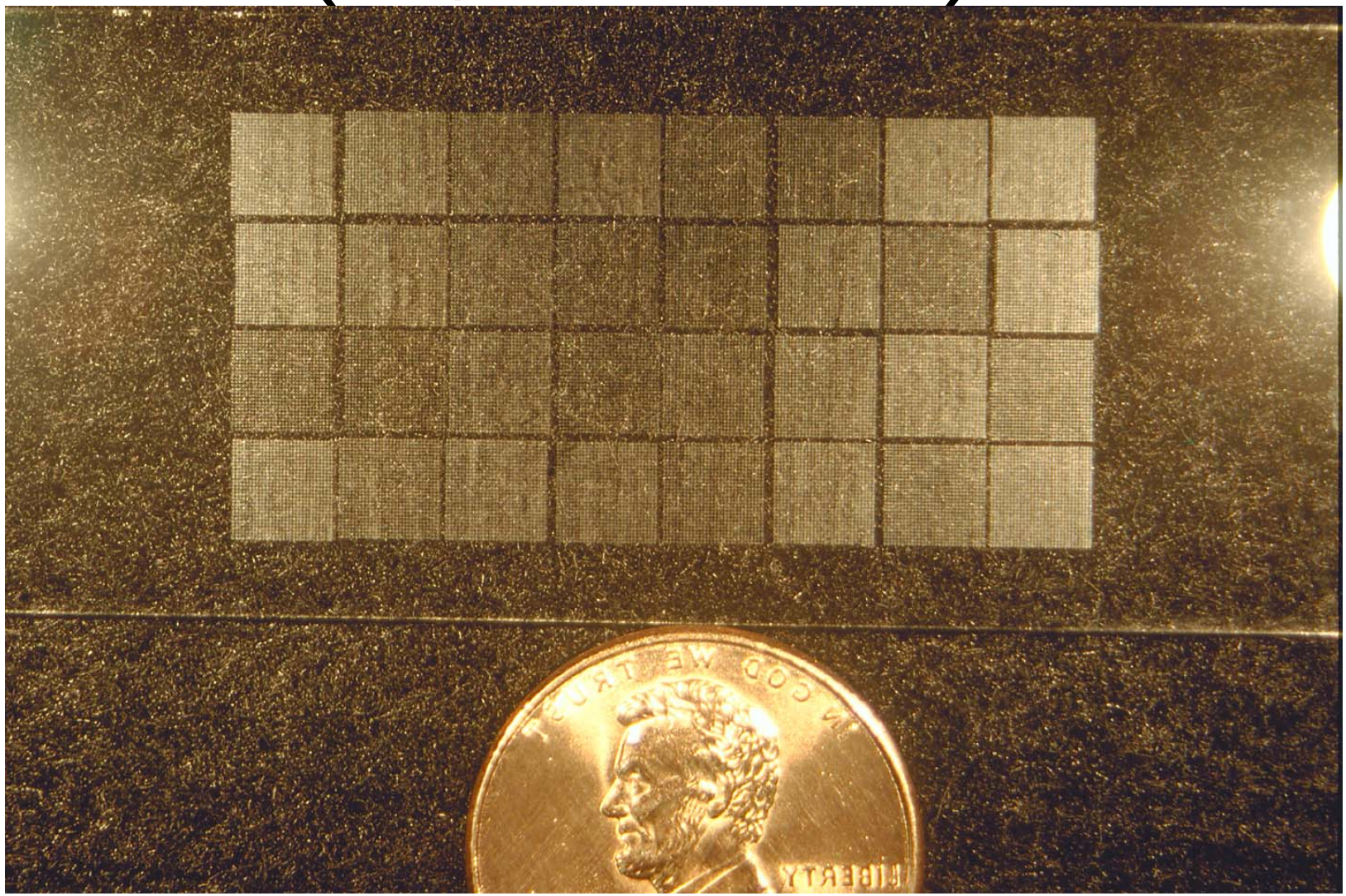
Why do we care?

And how can it help cure cancer?

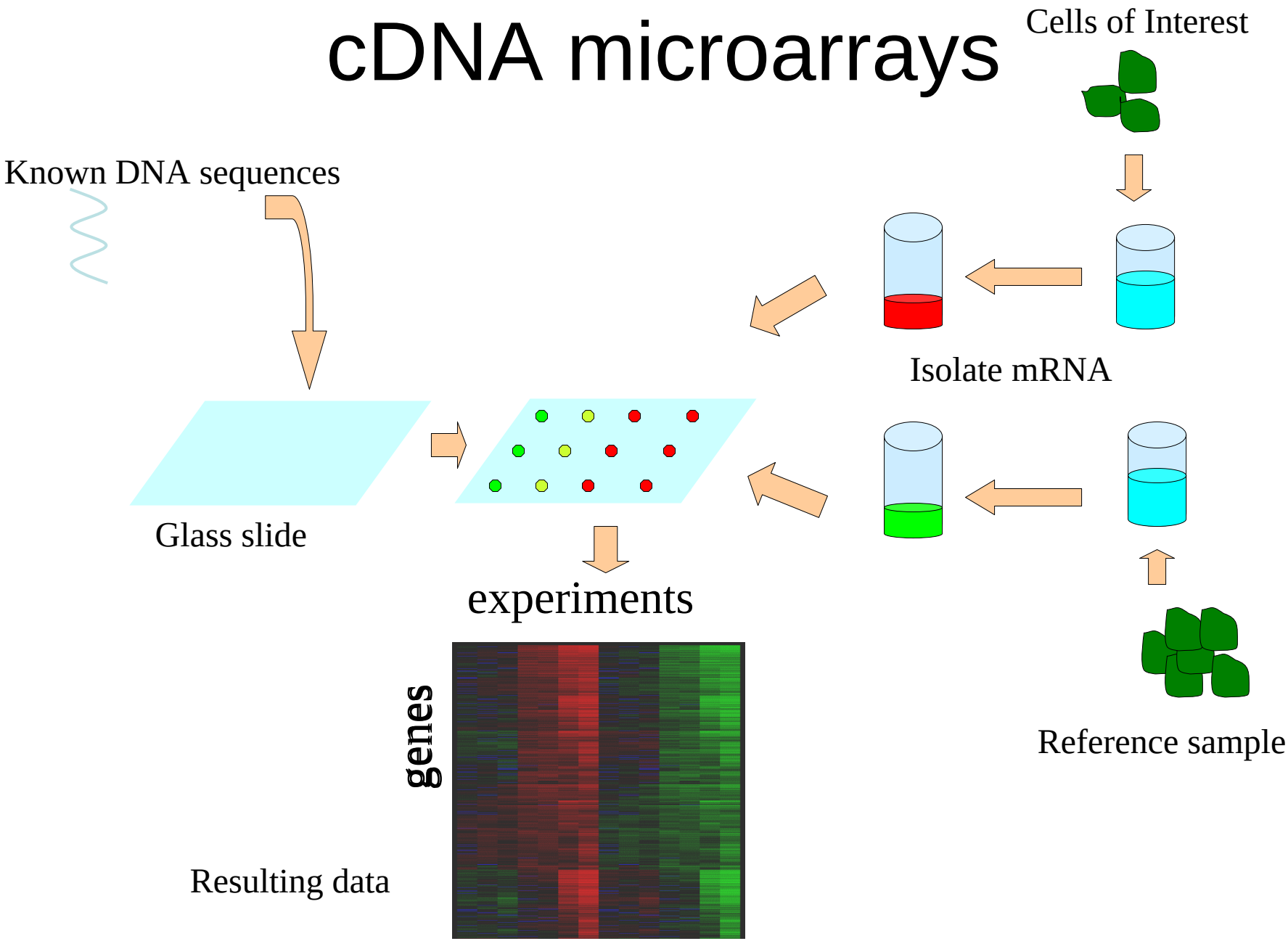
Microarray technologies

- Spotted cDNA arrays
  - Developed by Pat Brown (Stanford U)
  - Robotic microspotting
  - PCR products of full-length genes (>100nts)
- Affymetrix GeneChips
  - Photolithography (from computer industry)
  - Each gene represented by many n-mers
- Bubble jet / Ink jet arrays
  - Oligos (25-60 nts) built directly on arrays (in situ synthesis)
  - Highly uniform spots, very expensive

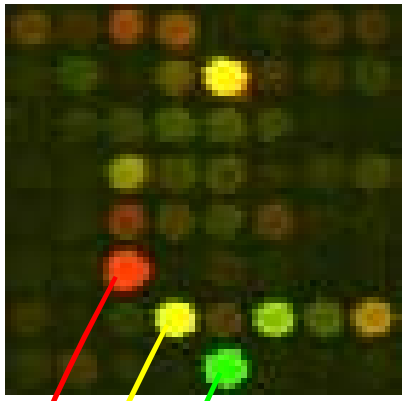
Early cDNA microarray (18,000 clones)



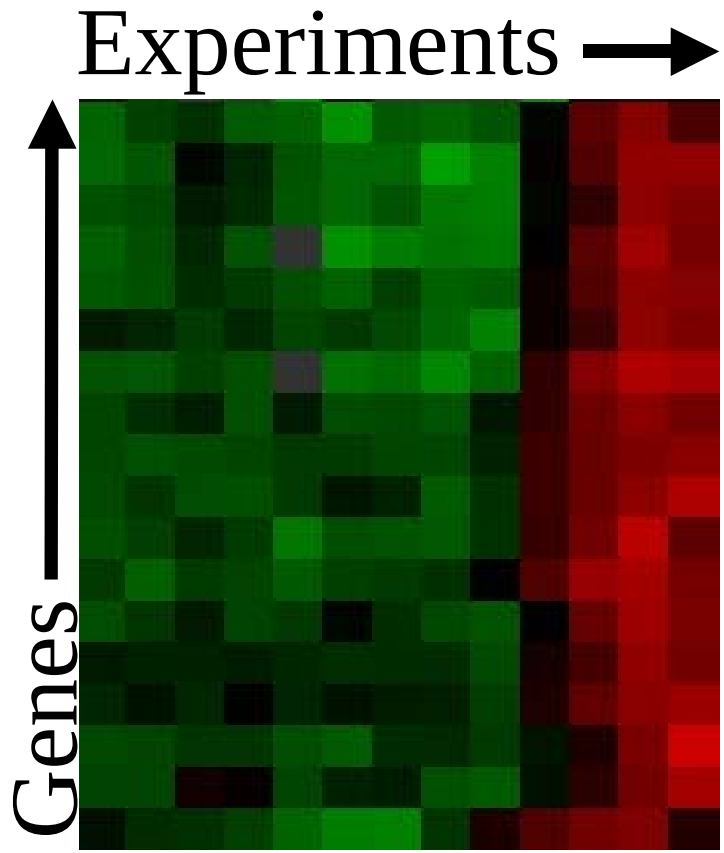
cDNA microarrays



Extracting Data



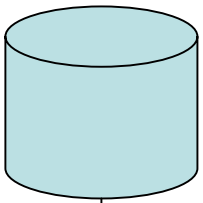
Cy3	Cy5	$\frac{Cy5}{Cy3}$	$\log_2\left(\frac{Cy5}{Cy3}\right)$
200	10000	50.00	5.64
4800	4800	1.00	0.00
9000	300	0.03	-4.91



Microarray Data Flow

Microarray experiment

Image Analysis



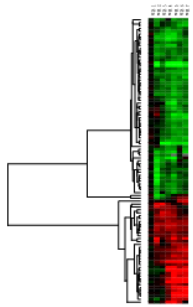
Database

Data Selection & Missing value estimation

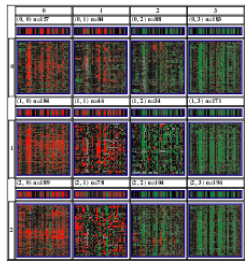
UID	NAME	GWEIGHT	spo0	spo30	spo2	spo5	spo7	spo9	spo11
EWEIGHT			1	1	1	1	1	1	1
YAL003W	EFB1	1	0.23	-1.79	-1.29	-1.56		-0.27	
YAL004W		1	0.41	-0.38	-0.89	-1.06	-1.6	-1.84	-1.6
YAL005C	SSA1	1	0.61	-0.07	-1.29	-1.29	-2	-1.84	-2.25
YAL010C	MDM10	1	0.16	-0.15	-0.76	-1.25	-1.89	-1.74	-1.6
YAL012W	CYS3	1	0.03	1.39	-0.84	-1.64	-2.84	-2.47	-2.4
YAL015C	NTG1	1	-0.18	-0.18	-0.62	-1.32	-1.69	-1.43	-1.79
YAL018C	YAL018C	1	-0.51	-0.62	-0.76	3.74	4.54	3.22	4.33
YAL025C	MAK16	1	-0.14	-3.32	-1.84	-1.12	-2.4	-1.03	-0.6
YAL034C	FUN19	1	0.19	-0.03	-1.03	-1.29	-1.84	-1.94	-1.74
YAL035W	FUN12	1	0.01	-1.47	-1.15	-0.69	-1.36	-1.64	-1.29
YAL036C	FUN11	1	-0.15	-2.74	-1.79	-1.32	-2.12	0.3	-0.89
YAL038W	CDC19	1	-0.06	-1.89	-1.69	-2.32	-2.4	-0.81	-1.6
YAL040C	CLN3	1	-0.17	-2.25	-1.69	-2.25	-2.56	-0.3	-2.4
YAL054C	ACS1	1	0.51	2.6	1.9	1.7	1.35	-0.03	-0.23
YAL055W	YAL055W	1	-0.32	0.83	0.58	0.82	1.4	2.05	2.24
YAL062W	GDH3	1	0.3	2.59	3	1.44	0.31	0.34	1.36
YAL067C	SEO1	1	-0.17	3.44	0.58	1.55	3.26	1.61	2.8
YAR003W	YAR003W	1	-0.29	0.54	0.6	1.08	1.42	1.86	1.42
YAR007C	RFA1	1	-0.14	1.74	2.41	2.1	2.04	0.57	0.84
YAR015W	ADE1	1	0.11	-1.51	-1.4	-1.36	-1.84	-1.89	-2
YAR027W	YAR027W	1	0.24	-1.06	-1.36	-1.56	-1.25	-0.94	-1.36
YBL009W	YBL009W	1	-0.01	0.62	1.04	1.3	2.52	2.15	2.24
YBL010C	YBL010C	1	0.01	0.21	0.7	1.45	2.25	1.77	1.24
YBL015W	ACH1	1	0.52	1.01	1.49	1.75	1.49	0.58	0.19

Data Matrix

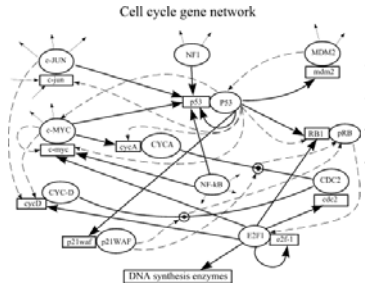
Normalization & Centering



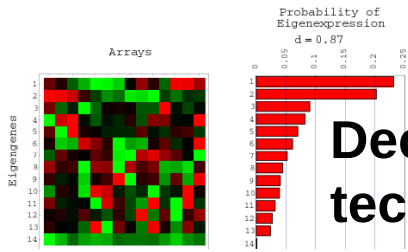
Unsupervised Analysis – clustering



Supervised Analysis

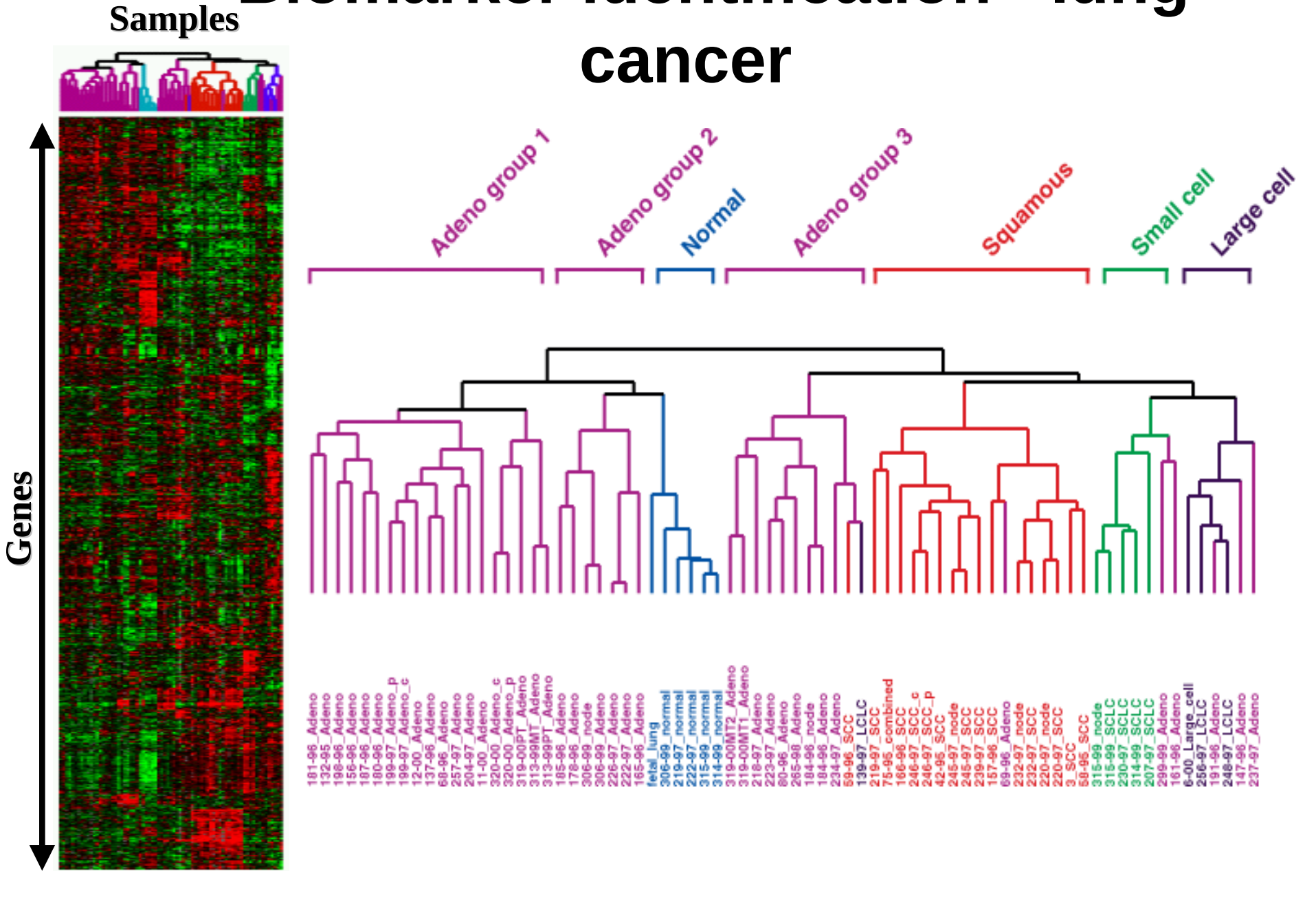


Networks & Data Integration

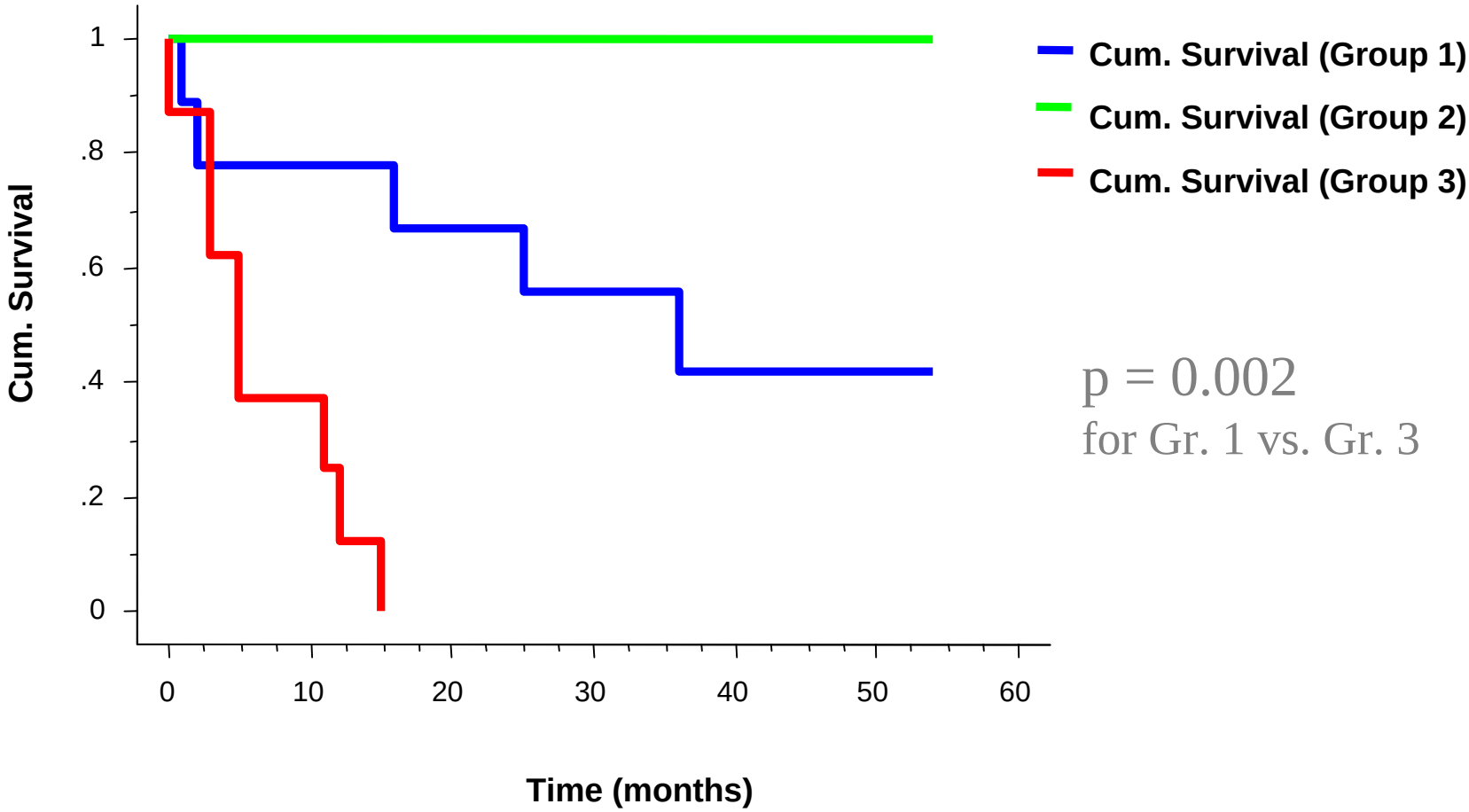


Decomposition techniques

Biomarker identification - lung cancer

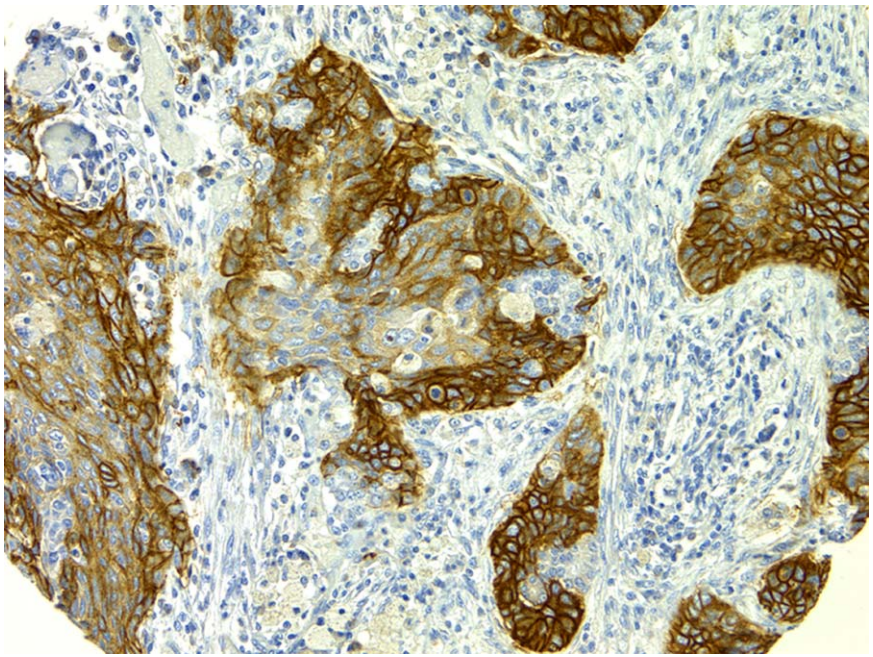


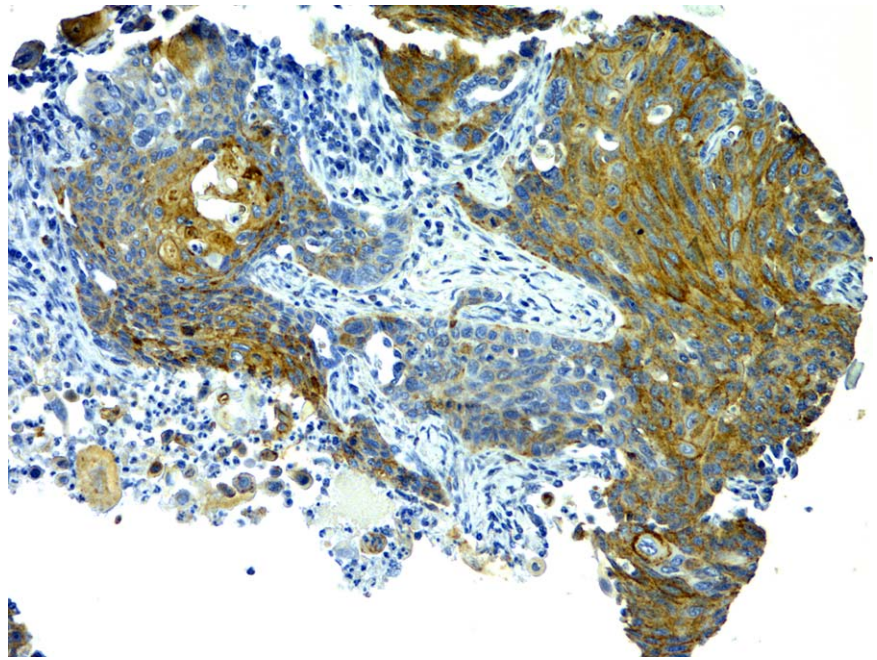
Data partitioning clinically important: Patient survival for lung cancer subgroups



p = 0.002
for Gr. 1 vs. Gr. 3

Genes overexpressed in low-survival lung adenocarcinomas





for details see: Garber, Troyanskaya et al. Diversity of gene expression in adenocarcinoma of the lung. PNAS 2001, 98(24):13784-9.

Computational biology/bioinformatics

What does it study?

Where do we get the data?

Computational Molecular Biology

- In order to gather insight into the ways in which genes and gene products (proteins) function, we:

- 1. Analyze DNA and protein sequences, searching for clues about structure, function, and control.

  SEQUENCE ANALYSIS

- 2. Analyze biological structures, searching for clues about sequence, function and control.

  STRUCTURE ANALYSIS

- 3. Understand how cellular components function in living systems.

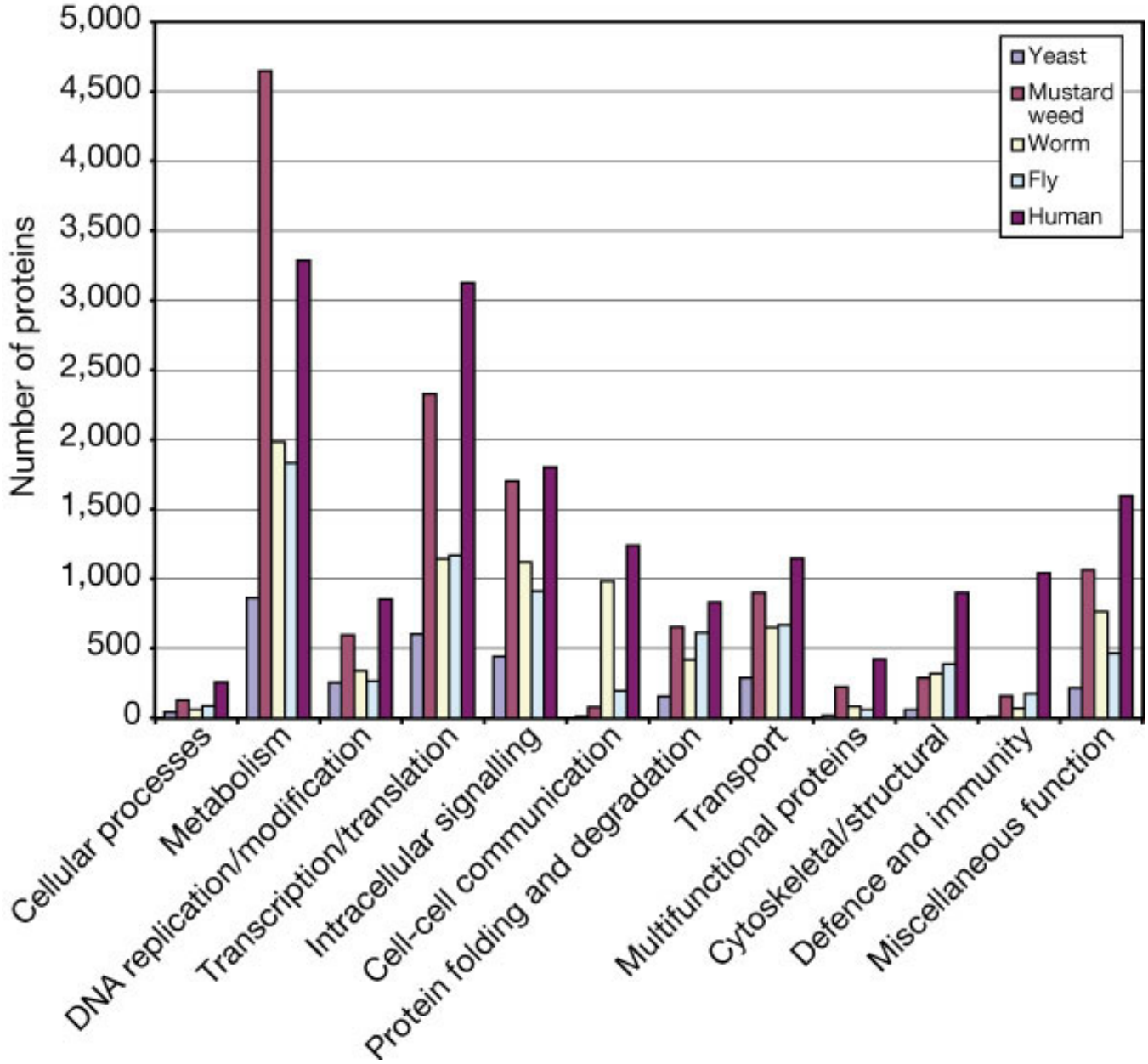
  FUNCTION ANALYSIS

What are functions of genes?

- Signal transduction: sensing a physical signal and turning into a chemical signal
- Structural support: creating the shape and pliability of a cell or set of cells
- Enzymatic catalysis: accelerating chemical transformations otherwise too slow.
- Transport: getting things into and out of separated compartments

What are the functions of genes?

- Movement: contracting in order to pull things together or push things apart.
- Transcription control: deciding when other genes should be turned ON/OFF
- Trafficking: affecting where different elements end up inside the cell


Number of proteins
5,000
4,500
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
Yeast
Mustard weed
Worm
Fly
Human
Cellular processes
Metabolism
DNA replication/modification
Transcription/translation
Intracellular signalling
Cell-cell communication
Protein folding and degradation
Transport
Multifunctional proteins
Cytoskeletal/structural
Defence and immunity
Miscellaneous function

Evolution is key.

1. Common descent of organisms implies that they will share many “basic technologies.”
2. Development of new phenotypes in response to environmental pressure can lead to “specialized technologies.”
3. More recent divergence implies more shared technologies between species.
4. All of biology is about two things: understanding shared or unshared features.

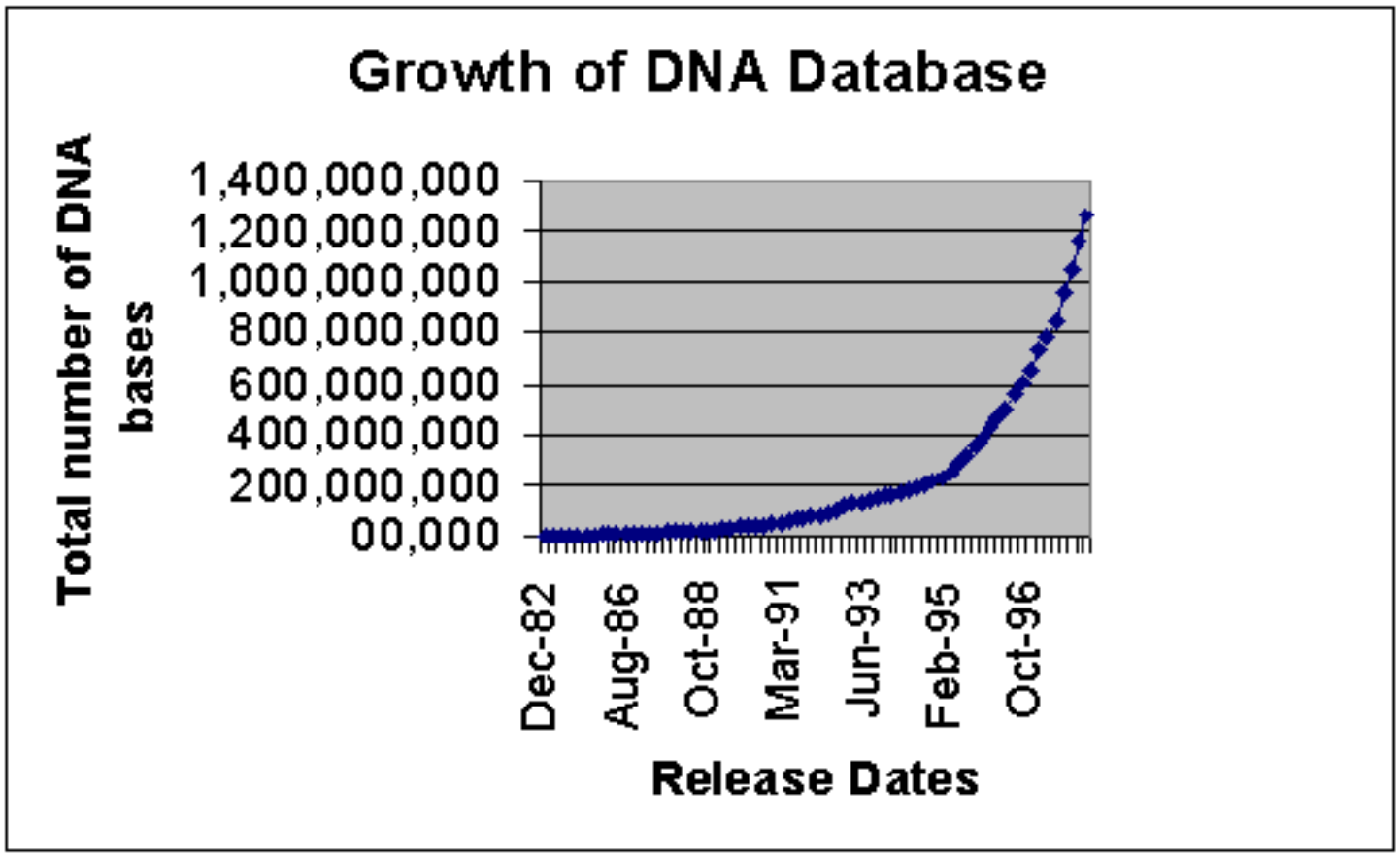
Where is the information?

- DNA sequence information

- GENBANK release 128 (2/02) contains
- 17,089,143,893 bases in 1546532 sequences
- (2001: 11,720,120,326 bases in 10,896,781 sequences)
- (2000: 5,805,414,935 bases in 5,691,170 sequences)
- (1999: 2,162,067,871 bases in 3,043,729)
- (1998: 1,622,041,465 bases in 2,355,928 sequences)
- (1997: 786,898,138 bases in 1,192,505 sequences)
- (1996: 463,800,000 bases in 686,000 sequences)

Biology and Medicine are fundamentally information sciences.

http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html



Species represented in GENBANK

Entries	Bases	Species
5230974	8242537293	Homo sapiens
3535771	2619042515	Mus musculus
402487	1058062734	Rattus norvegicus
328165	666672794	Drosophila melanogaster
209574	367443679	Oryza sativa
224766	279095020	Arabidopsis thaliana
196851	221097337	Caenorhabditis elegans
189066	165746672	Tetraodon nigroviridis
160327	150223453	Pan troglodytes
199339	132154949	Brassica oleracea
237369	120533407	Bos taurus
195458	114913763	Danio rerio
233551	106834331	Glycine max
200940	102692364	Xenopus laevis
208338	93920641	Zea mays
160455	83402180	Lycopersicon esculentum
140819	72490418	Medicago truncatula
80590	72104802	Entamoeba histolytica
106882	66276523	Hordeum vulgare

Complete Genomes Known (900 currently available publically)

- Aquifex aeolicus
- Archaeoglobus fulgidus
- Bacillus subtilis
- Borrelia burgdorferi
- Chlamydia trachomatis
- Escherichia coli
- Haemophilus influenzae
-
- Methanobacterium thermoautotrophicum
- Helicobacter pylori
- Methanococcus jannaschii
- Mycobacterium tuberculosis
- Mycoplasma genitalium
- Mycoplasma pneumoniae
- Pyrococcus horikoshii
- Treponema pallidum
- Saccharomyces cerevisiae
- Drosophila melanogaster

http://www.ncbi.nlm.nih.gov:80/PMGifs/Genomes/org.html

Where is the information?

- Protein Sequences

- PIR or Swiss-prot (as of 3/02)

- 106,736 sequences
- Over 39,242,287total amino acids
- From 95, 408 literature references

- `http://us.expasy.org/sprot/relnotes/relstat.html`

Where is the information?

- Protein three-dimensional Structures

- Protein Data Bank (PDB), as of March 26, 2002:

- 17,679 Coordinate Entries
- 15,855 proteins
- 1060 nucleic acids
- 746 protein/nucleic acid complex
- 18 carbohydrates

- `http://www.rcsb.org/pdb/`

Online access to DNA chip data

- http://smd.stanford.edu/
- Many published data sets available from Stanford site, 10,000 to 40,000 genes per chip
- Each set of experiments involves 3 to 40 “conditions”
- Each data set is therefore near 1 million data points.
- People gearing up for these measurements everywhere…

Where’s the information?

- Medical Literature on line.
- Online database of published literature since 1966 = Medline = PubMED resource
- 4,600 journals
- 11,000,000+ articles (most with abstracts)

Human Genome Browsers

- UC Santa Cruz:
- http://genome.ucsc.edu/

- NCBI:
- http://www.ncbi.nlm.nih.gov:80/cgi-bin/Entrez/map_search

- ENSEMBL:
- http://www.ensembl.org/Homo_sapiens/