A biological introduction

For a reasonably short explanatory text read Chapter 1 Molecular Biology for Computer Scientists by *Lawrence Hunter* available at

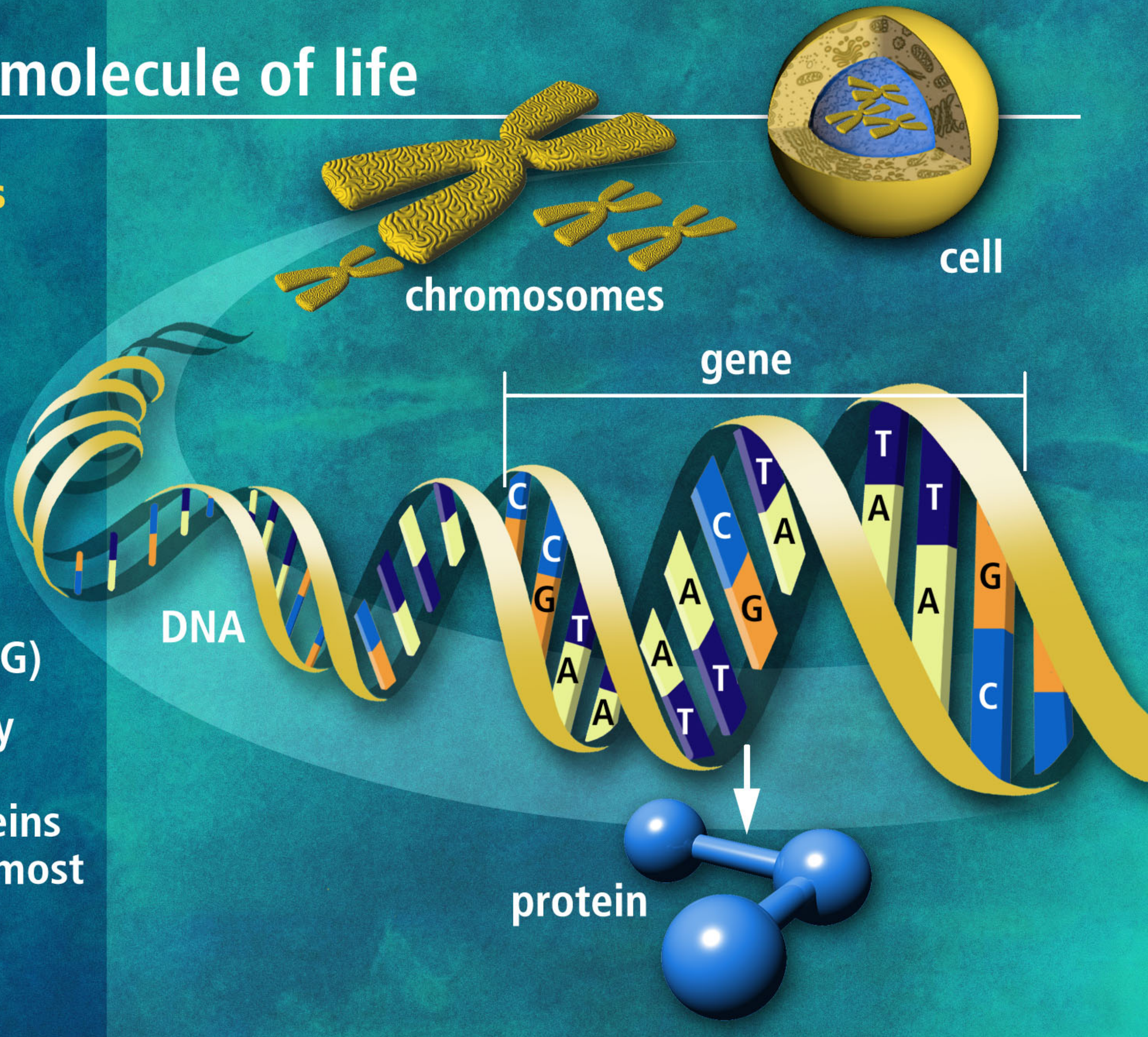
http://www.aaai.org/Library/Books/Hunter/01-Hunter.pdf

DNA the molecule of life

Trillions of cells

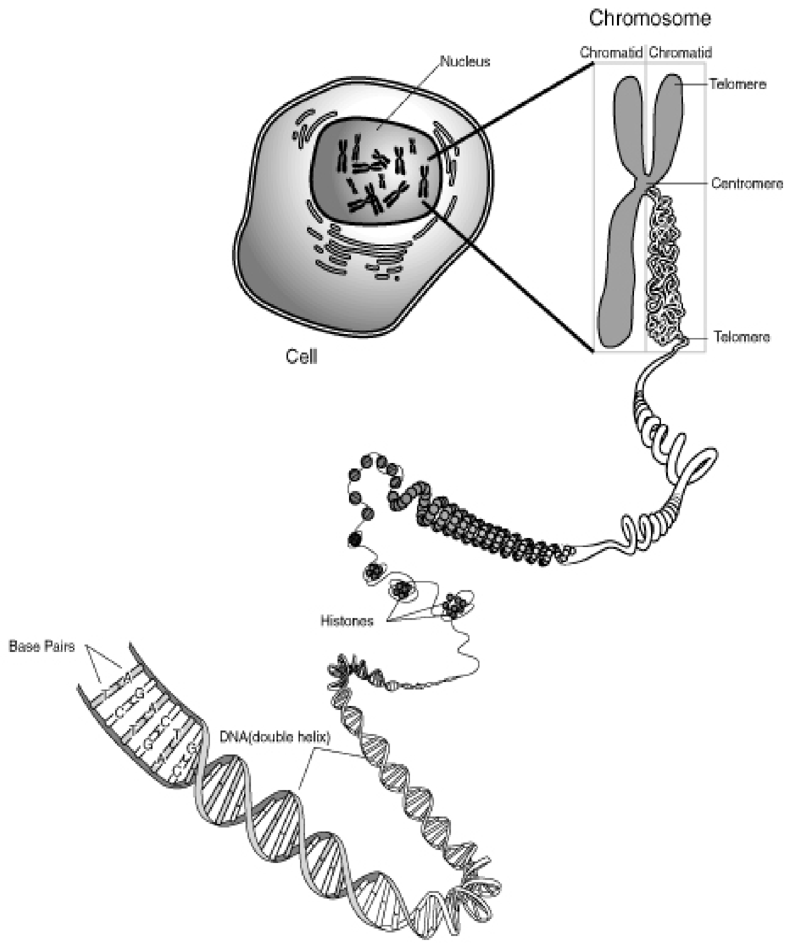
Each cell:

- 46 human chromosomes
- 2 meters of DNA
- 3 billion DNA subunits (the bases: A, T, C, G)
- Approximately 30,000 genes code for proteins that perform most life functions



Y-GG 01-0085

Schematic view of DNA organization in a cell



Genes and Genomes

- A Gene is the fundamental physical and functional unit of heredity. A gene is an ordered sequence of nucleotides located in a particular position on a particular chromosome that encodes a specific functional product (i.e., a protein or RNA molecule).

- A Genome is all the genetic material (DNA) in the chromosomes of a particular organism; its size is generally given as its total number of base pairs.

Genomes: How Many are Sequenced?

- **59 genomes completed as of April 2001**
- **Eukaryotes:**
  - *Saccharomyces*
  - *Caenorhabditis*
  - *Drosophila*
  - *Arabidopsis*
  - *Human - First Draft*
- **Expected year 2001**
  - *Mouse* (april)
- **Human complete Genome 2003**

Composition of the Genome: *Drosophila*

- 180 Mb
  - 1/6 of the size of Human Genome (3 Gb)
- 120 Mb Euchromatin
  - portion of the genome that can be cloned stably in BAC's
- 60 Mb Heterochromatin
  - short simple repeats over many kbs
  - occasionally interrupted by inserted transposable elements
  - tandem repeats of rRNA genes
  - few protein encoding genes

Genomes Content Sequenced

Organism	year	total%	euchromatin %
S. cerevisiae	1986	93	100
C. elegans	1988	99	100
D. melanogaster	2000	64	97
A. thaliana	2000	92	100
H. sapiens (public)	2001	84	90
H. sapiens (Celera)	2001	83	99-93

Modified from: Bork and Copley, Nature 2001

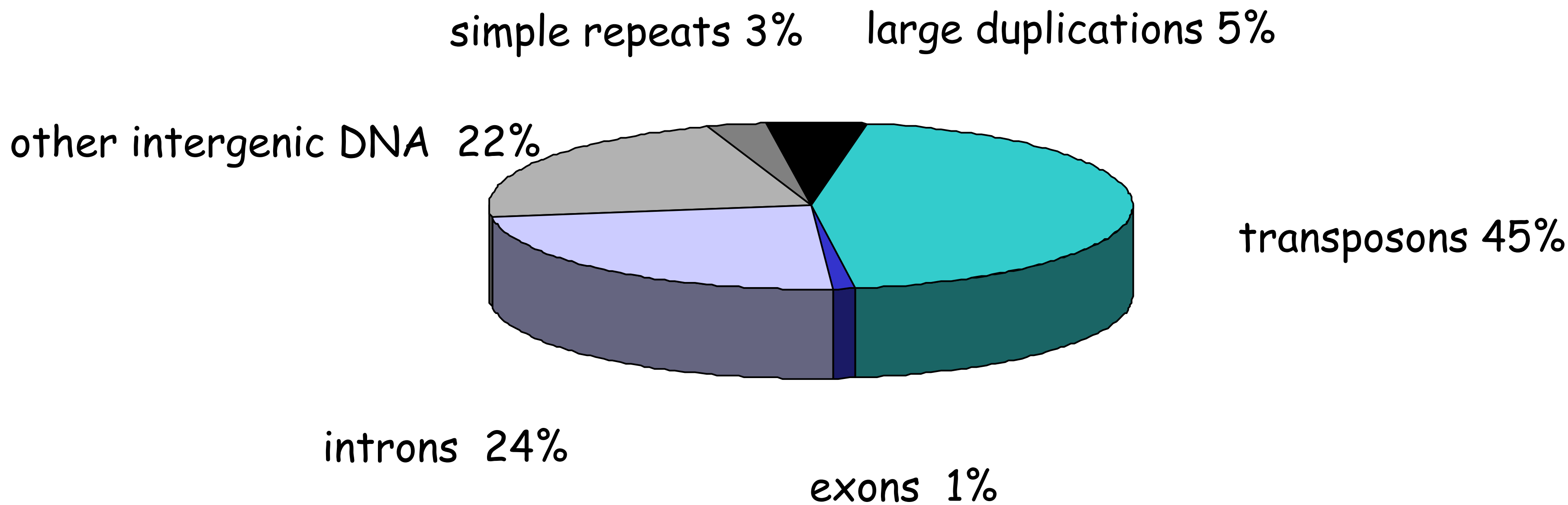
Genome Size and Gene Counts

Species	Genome size (Mb)	Genes	Lethal loci
Mycoplasma genitalium	0.58	470	~300
Rickettsia prowazekii	1.11	834	
Haemophilus influenzae	1.83	1743	
Mthanococcus jannaschi	1.66	1438	
Bacillus subtilis	4.2	4100	
Escherichia coli	4.6	4288	1800
Saccharomyces cerevisiae	13.5	6034	3600
Arabidopsis thaliana	119	25498	
Caenorhabditis elegans	97	18424	
Drosophila melanogaster	165	13601	3100
Homo sapiens	3.3	31000-40000	

Minimal Genome Project (TIGR)

- *Mycoplasma genitalium*
  - 517 genes
  - 480 proteins
  - 265-350 are essential
  - 100 of these with no known function

Human Genome Composition



What does the draft human genome sequence tell us?

By the Numbers

- The human genome contains 3164.7 million chemical nucleotide bases (A, C, T, and G).
- The average gene consists of 3000 bases, but sizes vary greatly, with the largest known human gene being dystrophin at 2.4 million bases.
- The total number of genes is estimated at 30,000 to 35,000 much lower than previous estimates of 80,000 to 140,000 that had been based on extrapolations from gene-rich areas as opposed to a composite of gene-rich and gene-poor areas.
- Almost all (99.9%) nucleotide bases are exactly the same in all people.
- The functions are unknown for over 50% of discovered genes.

Human Genome Program, U.S. Department of Energy, Genomics and Its Impact on Medicine and Society: A 2001 Primer, 2001

What does the draft human genome sequence tell us?

How It's Arranged

• The human genome's gene-dense "urban centers" are predominantly composed of the DNA building blocks G and C.

• In contrast, the gene-poor "deserts" are rich in the DNA building blocks A and T. GC- and AT-rich regions usually can be seen through a microscope as light and dark bands on chromosomes.

• Genes appear to be concentrated in random areas along the genome, with vast expanses of noncoding DNA between.

• Stretches of up to 30,000 C and G bases repeating over and over often occur adjacent to gene-rich areas, forming a barrier between the genes and the "junk DNA." These CpG islands are believed to help regulate gene activity.

• Chromosome 1 has the most genes (2968), and the Y chromosome has the fewest (231).

Human Genome Program, U.S. Department of Energy, Genomics and Its Impact on Medicine and Society: A 2001 Primer, 2001

What does the draft human genome sequence tell us?

The Wheat from the Chaff

- Less than 2% of the genome codes for proteins.

- Repeated sequences that do not code for proteins ("junk DNA") make up at least 50% of the human genome.

- Repetitive sequences are thought to have no direct functions, but they shed light on chromosome structure and dynamics. Over time, these repeats reshape the genome by rearranging it, creating entirely new genes, and modifying and reshuffling existing genes.

- During the past 50 million years, a dramatic decrease seems to have occurred in the rate of accumulation of repeats in the human genome.

Human Genome Program, U.S. Department of Energy, Genomics and Its Impact on Medicine and Society: A 2001 Primer, 2001

What does the draft human genome sequence tell us?

How the Human Compares with Other Organisms

- Unlike the human's seemingly random distribution of gene-rich areas, many other organisms' genomes are more uniform, with genes evenly spaced throughout.

- Humans have on average three times as many kinds of proteins as the fly or worm because of mRNA transcript "alternative splicing" and chemical modifications to the proteins. This process can yield different protein products from the same gene.

- Humans share most of the same protein families with worms, flies, and plants, but the number of gene family members has expanded in humans, especially in proteins involved in development and immunity.

- The human genome has a much greater portion (50%) of repeat sequences than the mustard weed (11%), the worm (7%), and the fly (3%).

- Although humans appear to have stopped accumulating repeated DNA over 50 million years ago, there seems to be no such decline in rodents. This may account for some of the fundamental differences between hominids and rodents, although gene estimates are similar in these species. Scientists have proposed many theories to explain evolutionary contrasts between humans and other organisms, including those of life span, litter sizes, inbreeding, and genetic drift.

Human Genome Program, U.S. Department of Energy, Genomics and Its Impact on Medicine and Society: A 2001 Primer, 2001

What does the draft human genome sequence tell us?

Variations and Mutations

- Scientists have identified about 1.4 million locations where single-base DNA differences (SNPs) occur in humans. This information promises to revolutionize the processes of finding chromosomal locations for disease-associated sequences and tracing human history.

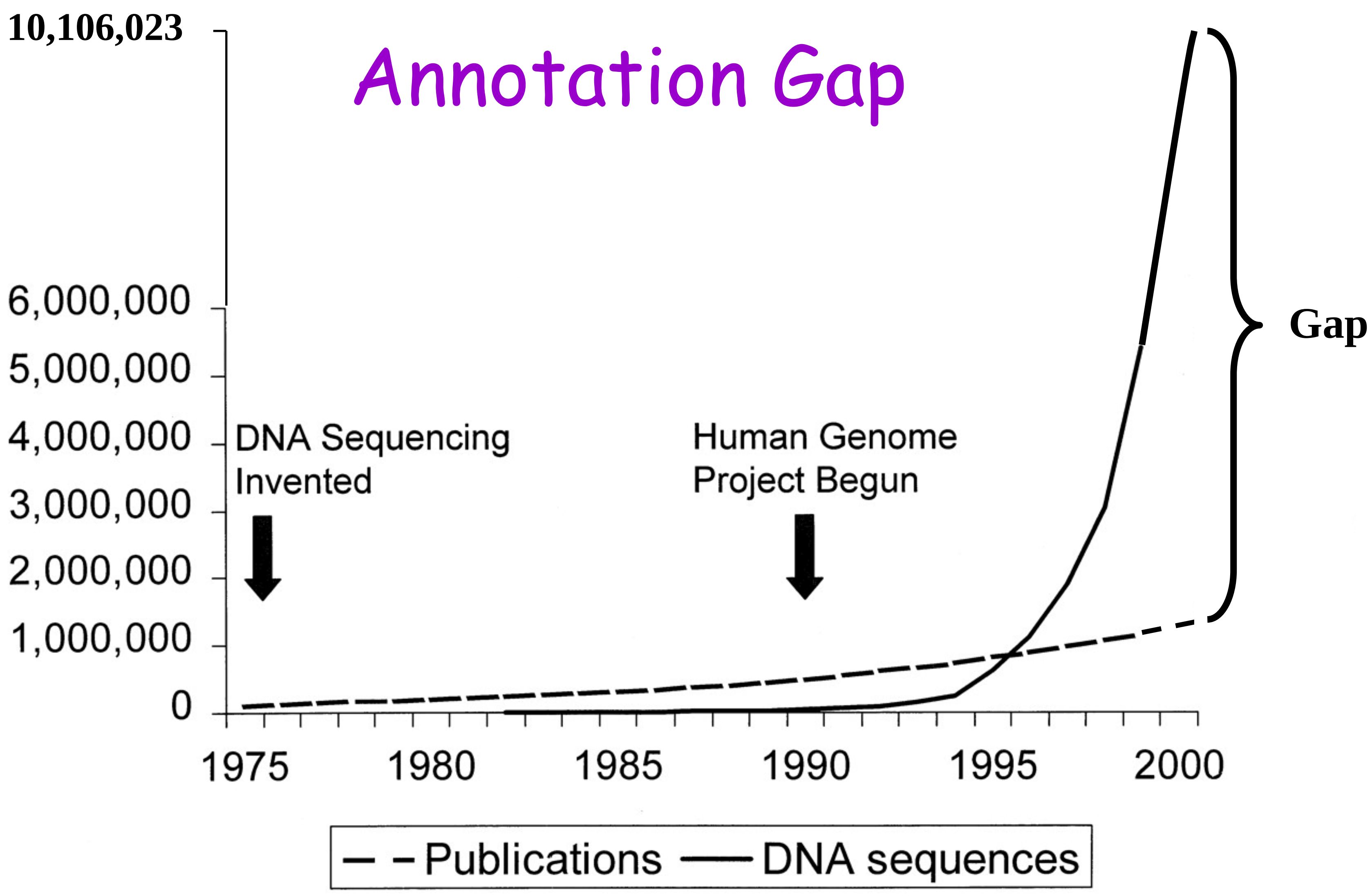
- The ratio of germline (sperm or egg cell) mutations is 2:1 in males vs females. Researchers point to several reasons for the higher mutation rate in the male germline, including the greater number of cell divisions required for sperm formation than for eggs.

Human Genome Program, U.S. Department of Energy, Genomics and Its Impact on Medicine and Society: A 2001 Primer, 2001

What do We Know About Our Genes?

- Human Genome: some **31 000 - 40 000** genes
- GeneCards Database: **18 583** genes
- SwissProt: Human + Any data about the function
  - » **5297 proteins**

Annotation Gap



modified from: Boguski, Science 1999

What does the Genome Looks Like?

GGATCCTTGTTCAACTCAGTATTTAATTGGGATTATTACCTTCAAATTTTCAGCATTACCAACAAGAATC
CTATACTAGCAACATGGTGTCGGTGAGTATTCTACACTTTGGATATAACGTCAATTTAGTGAAAGATTGC
ATAAGAGGAAATTCCCATTTTTCACACTTTTATGGTGTTATATAAGTACAAAGTTTCCTGACTGCGACTT
TCCTTTGCTGAACAGTTCGAGGAAGCCACTGAACTCGCCAAGAAGTTCACCAAGAAGCCCACGGATGCC
G
AGTTCTTGGAATTCTACGGTCTCTTCAAGCAGGCCACCGTTGGCGATGTGAACATCGAGAAGCCCGGCG
C
TCTGGCTCTCAAGGACAAGGCCAAGTACGAGGCCTGGAGCTCCAACAAGGGTCTCTCCAAGGAGGCTG
CC
AAGGAGGCCTACGTAAAGGTGTACGAGAAGTACGCCCCCAAGTACGCCTAAGCCGGCAACCGATCAAAT
C
CGATCCAATCCGATTCCGATTGCGGACCCCCCATGCACCTGCCATCACCACTATAGTACTTAGTTGGACA
ATAAAGATTACCAGATATATGAGCAATAATCAGGTGTTTGTCCGGCTGAGCGGCATTTTATTGGAGCTTT
CGTGCATTATGGAACTAAAATGGAGATGGAGATTTGGGAGCGGTGGGACCCCTCGAACTTTGATCCGAA
C
GAGATTACGTGTCTTGACAGTGGTGTTTGGTGGGGTGAATTATCAGGGAATCGAAATCGGAGCTGTTATC
AGTAGGCCTGCATTATCAGCAACGCGAATGCCTAGCCCCACTTCAGCGAAAGCTTCCCGATCCACTGAG
A
ATCGGAGTGGAGCTTAACCGCTCATTTTGCTCACTGGGGAGAAGCTTTCTCGTTGCGGGCGATCGTGCG
A
TATGGAATGACTTAAATTGAAAGTCACCGCCGGATTGGAATTCAGACACTGTTTAGTTCTGTGGCAAACA
ACACACAACAGTAAGTGTTTAGTTTAAATTCAATTACCCAGCTGCTAATGGGTGGTAATATCAGTGCTGG
TGACACATAAAGTGGACATTTCGTGCGAGTTTATACACCTGCAGAACTCAATTATGGAAATTCGCCAGGC
AACAGGCACTGATGAAAATAAAGTACAAAAAAATCGCTTTTGAAACGCATATTCTGGAACATGCTAAACG
AAACTTGTCACCTTGACCTTTTGCAATTTTCCACGCCCATAAACTAAATGAGTGCCTTTTAAATGATTAC
TTATTTTAGAACTGCAAGATGGCGTGGCCATAACTAAGCATGACCATATGTATGTGCCAGATATTTTGCT
CAGTGCTAGACACGTACAAAGCCCACGTCCGTATCCAAATTGGTCGAAATTCGCCGGCCAAAAGCGGCC
G
CTAACGGTTGCCCGCTCGCTCCCTTTCAAAAATCTCGCAGTCGAGACAAATATTTGCGTCTGCTGTTTCC
GCAGCATAATTTCCCAGTTCTACTCGAATTATGTATTTCAGCCGGGAGAAAGTCGAACAAACATCTATCC
ATTAGCAAACATTCGTAGGGTCTTCAAGATCGCCTCACGGGGCCAAAGGTCATTTCCAGCTGGCTCTGTT
TGGCACACGCCAGCCTCTTCACTCGGAACTGGTTTTACTTATTATGCTAATAAACATTAAGCTGACCACT
TCATAAACATGACTAATTAAACATTTTTTTTTTACAATTCCACTACTTACAGATAAAATCTAACCCAGAA
TGGTTTCCGAGGTAAGTACTTTCCCAAAAGATTATGCACTGATAGCAACAGTCTACCAACATTTAGGGCG
TGATAGCCTTGCTGCTAATCACCTGTTCGATTGCTATACACTGAGCGAAATCTATTCATAAGCGATTGAC
TTTTCAATCTTAAAATTGTAATCGTAATATGAGATAAGATTGTGATTCCGCTGAAGCAAGGACAGCTGTT
CAAATACGATTTTGTCTTGAAGTAATAACCGTTTTCGGTTGAAATCGATAATTATAATTTTTTGTAGTTT
CCAGAGAATATTTTTATCTCGCACAACTATTTTTGTTAGTTTTAATGACAGTTTGTTAGTAGATATTGTT
GCAACTAATATTCCAAACCTTTTTTATAATACCACTTTCAATAGATTTTACGATCCCACTTGATAGATTT
TCCACTAATAAATCATCCCGTCTATCTTCAGCAATTCAACGCCGCCGCCGAGAAGGTGAAGAGCCTAACC
AAGCGTCCCAGTGATGACGAGTTCCTGCAGCTGTACGCCCTGTTCAAGCAGGCCAGCGTTGGTGACAAC
G
ACACCGCCAAGCCGGGTCTCCTGGACCTGAAGGGCAAGGCCAAGTGGGAGGCCTGGAACAAGCAGAA
GGG
CAAGAGCAGCGAGGCCGCCCAGCAGGAGTACATCACCTTTGTGGAGGGCCTGGTGGCCAAGTATGCCT
AA
AAACCCAATCCCAGCAATTAGCGATCTTAAACCAGCTAGAGACTATTGTAATGTTACCTTTAATGCGGAA
TAAACTTCGTATGTTAATTTTGTACTAAAAGATATTGACCCGGTGACTCAAGAGGTGCATACGCCTGGCC
TGGGGCTATTTAACTGGCAAATAGAGGAAGTGGCTGGCTGCACCGCTTTGCAGTACTTTGAACCGCCAG
T
GGCGCTTGTTAGCGGTGCATCGTTAATGGATCTTCGAGTGTGTTTGTCCAAACCGCGCTTCGAACGGGG

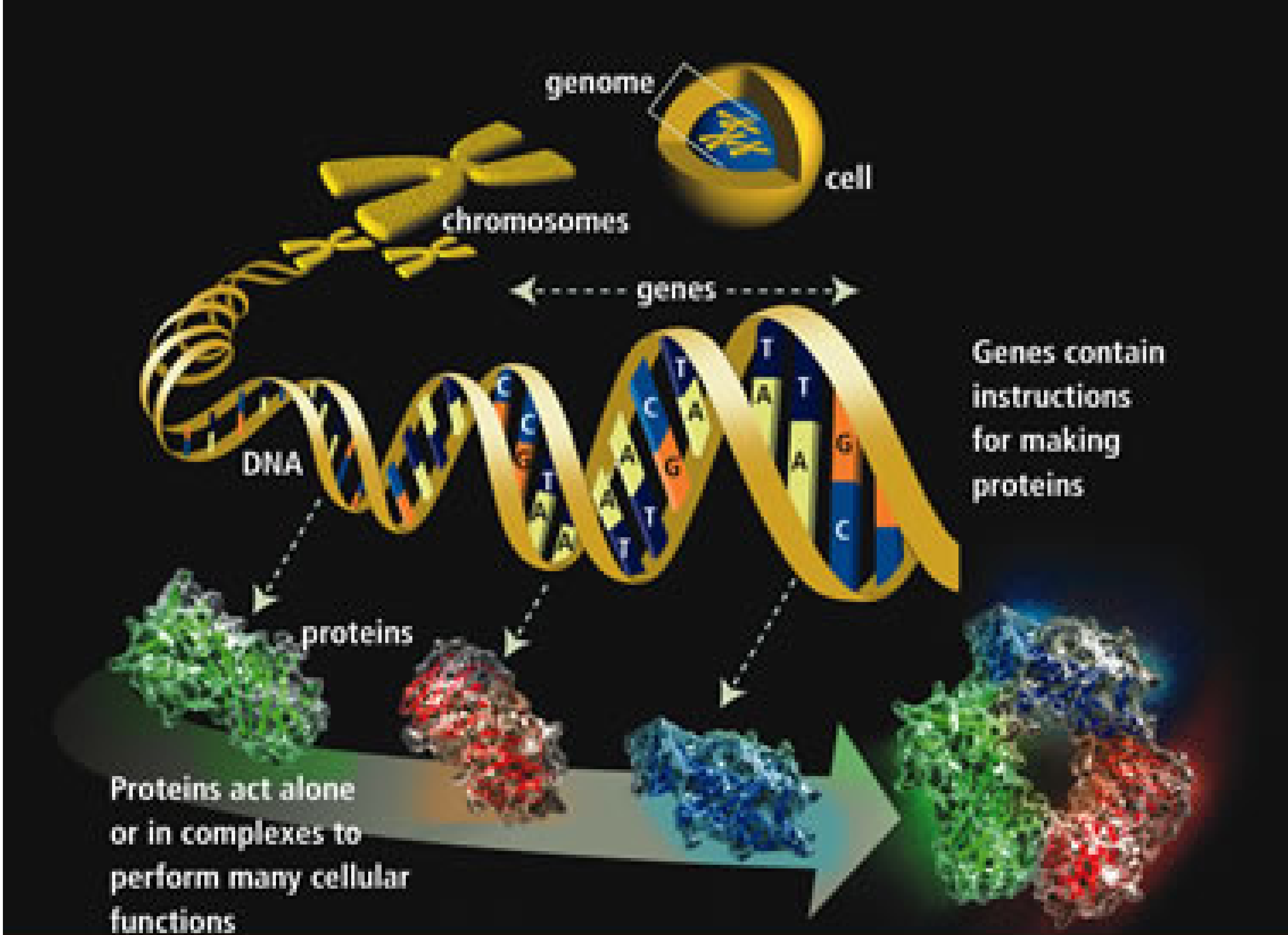
This sequence is:
2821 characters

Human Genome:
about 1.063 milj pages

How does it work: Gene expression

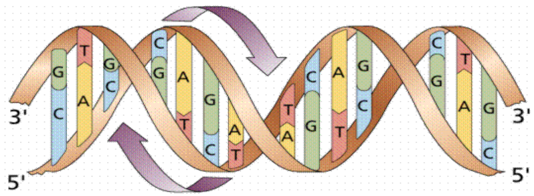
- The process by which a gene's coded information is converted into the structures present and operating in the cell.

- Expressed genes include those that are transcribed into mRNA and then translated into protein and those that are transcribed into RNA but not translated into protein (e.g., transfer and ribosomal RNAs).



From Genes to Proteins

From DNA to Protein



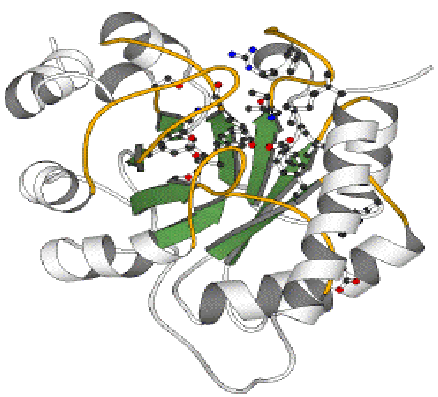
DNA

transcription

mRNA

translation

Protein



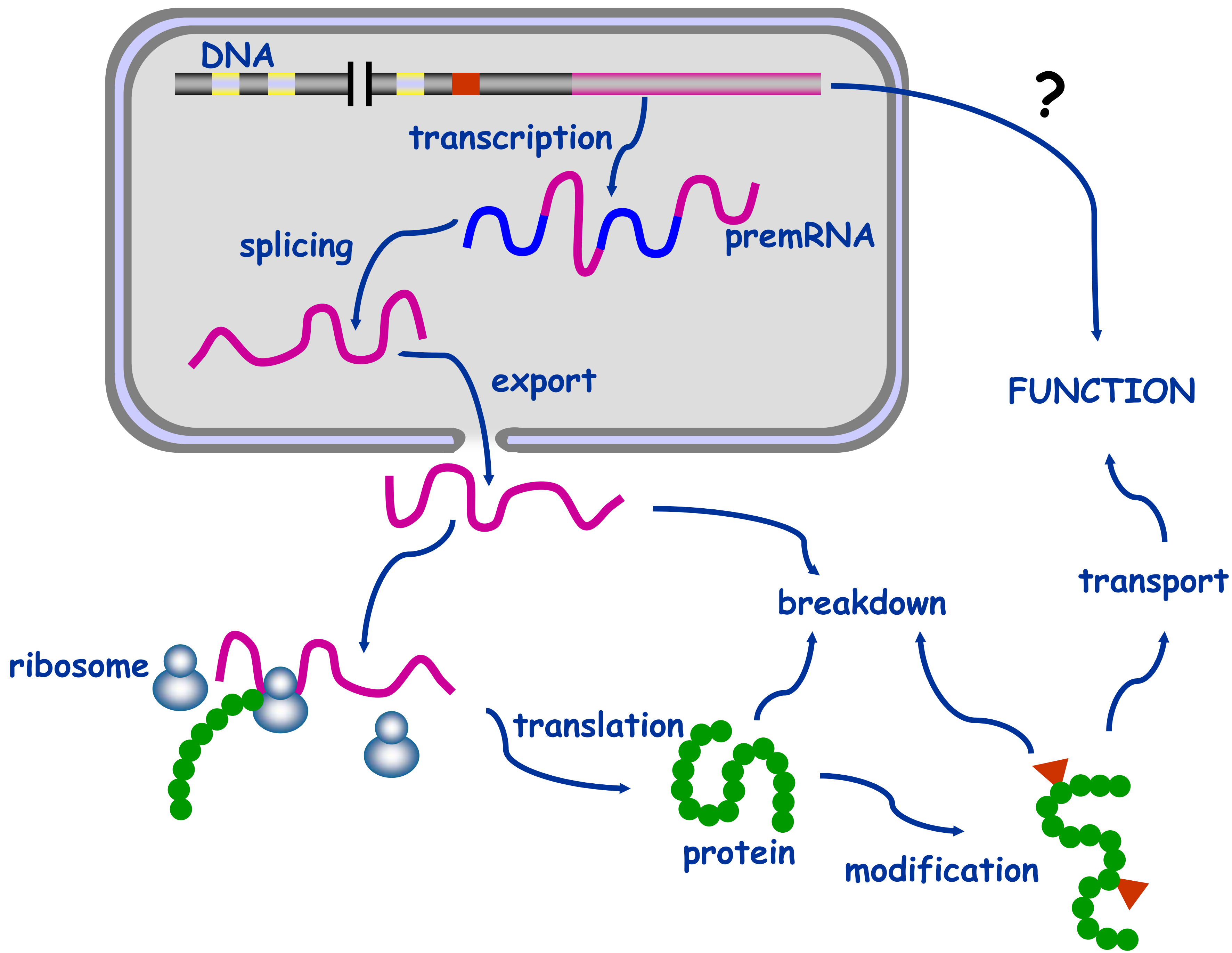
CCTGAGCCAACTATTGATGAA

CCUGAGCCAACUAUUGAUGAA

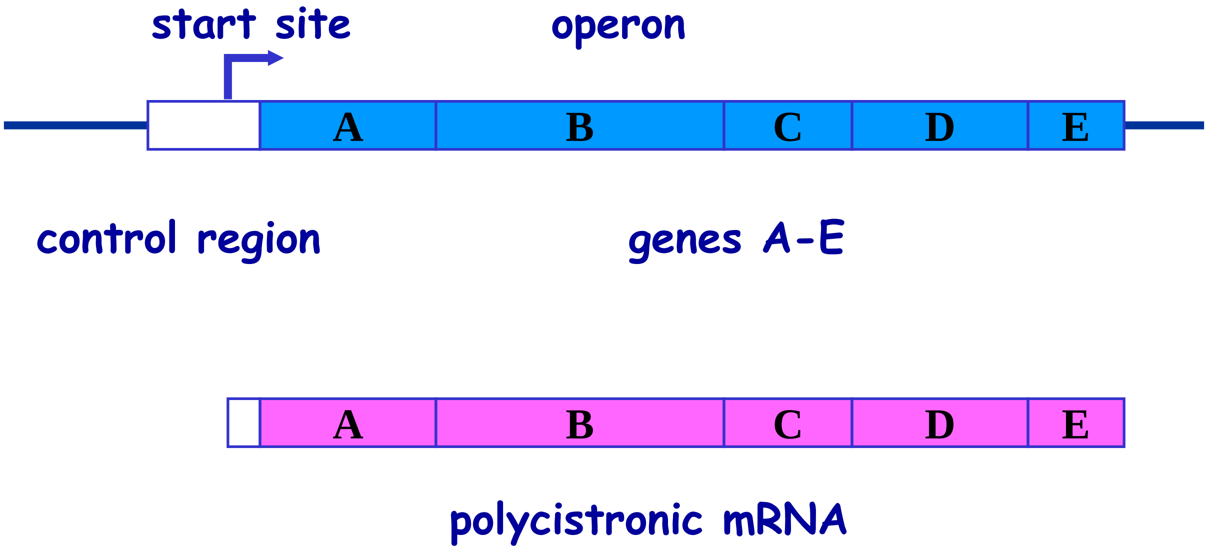
PEPTIDE

Genetic Code

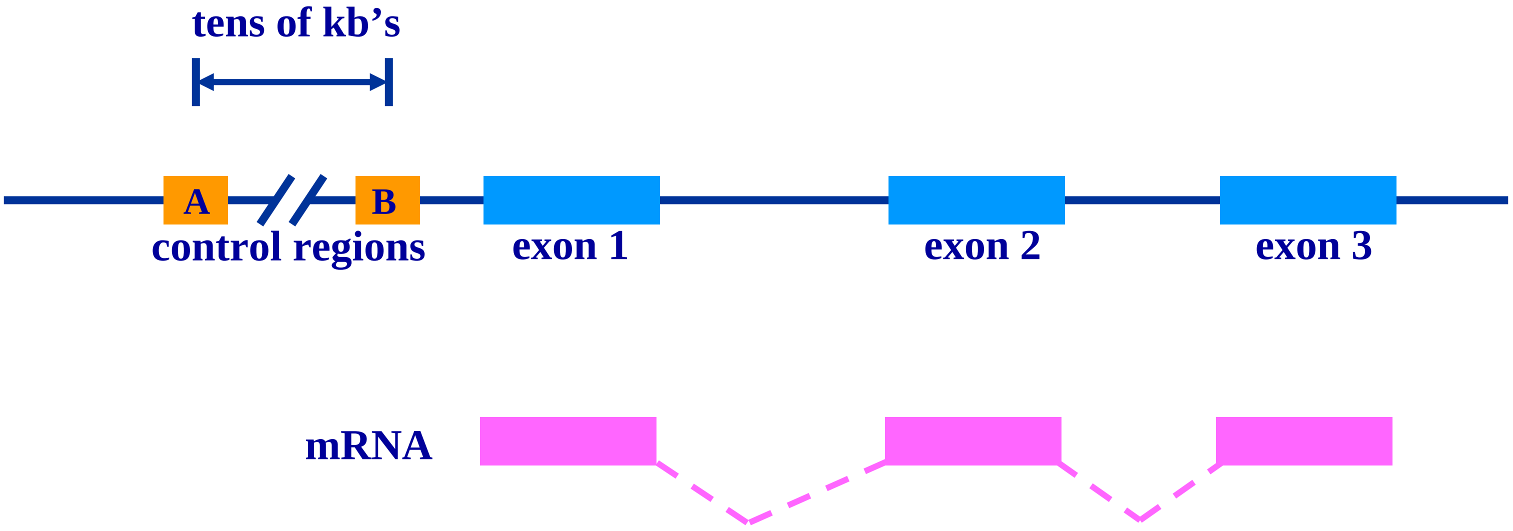
NAME	3 Letter	1 Letter	DNA codons for each Amino Acids
Alanine	Ala	A	GCA,GCC,GCG,GCU
Cysteine	Cys	C	UGC,UGU
Aspartic Acid	Asp	D	GAC,GAU
Glutamic Acid	Glu	E	GAA,GAG
Phenylalanine	Phe	F	UUC,UUU
Glycine	Gly	G	GGA,GGC,GGG,GGU
Histidine	His	H	CAC,CAU
Isoleucine	Ile	I	AUA,AUC,AUU
Lysine	Lys	K	AAA,AAG
Leucine	Leu	L	UUA,UUG,CUA,CUC,CUG,CUU
Methionine	Met	M	AUG
Asparagine	Asn	N	AAC,AAU
Proline	Pro	P	CCA,CCC,CCG,CCU
Glutamine	Gln	Q	CAA,CAG
Arginine	Arg	R	CGA,CGC,CGG,CGU
Serine	Ser	S	UCA,UCC,UCG,UCU,AGC,AGU
Threonine	Thr	T	ACA,ACC,ACG,ACU
Valine	Val	V	GUA,GUC,GUG,GUU
Tryptophan	Trp	W	UGG
Tyrosine	Tyr	Y	UAC,UAU
Stop Codons			UAA,UAG,UGA


DNA
?
transcription
splicing
premRNA
export
FUNCTION
breakdown
transport
ribosome
translation
protein
modification

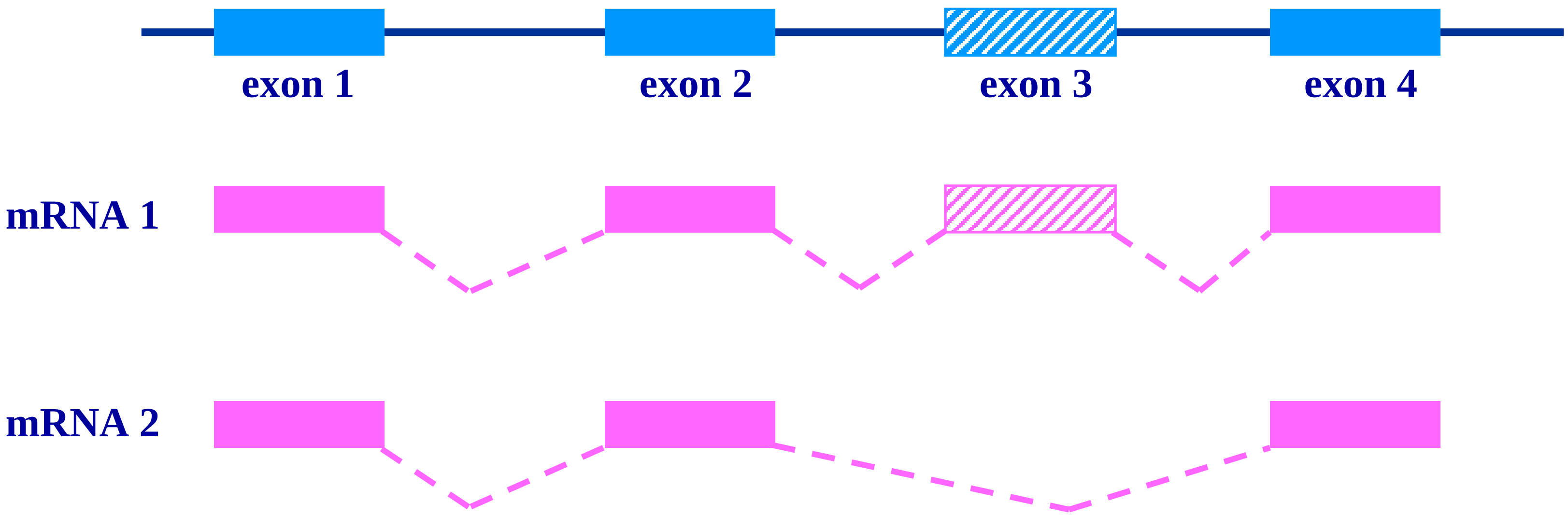
Prokaryotic Genes Come in Operons



Eukaryotic Genes



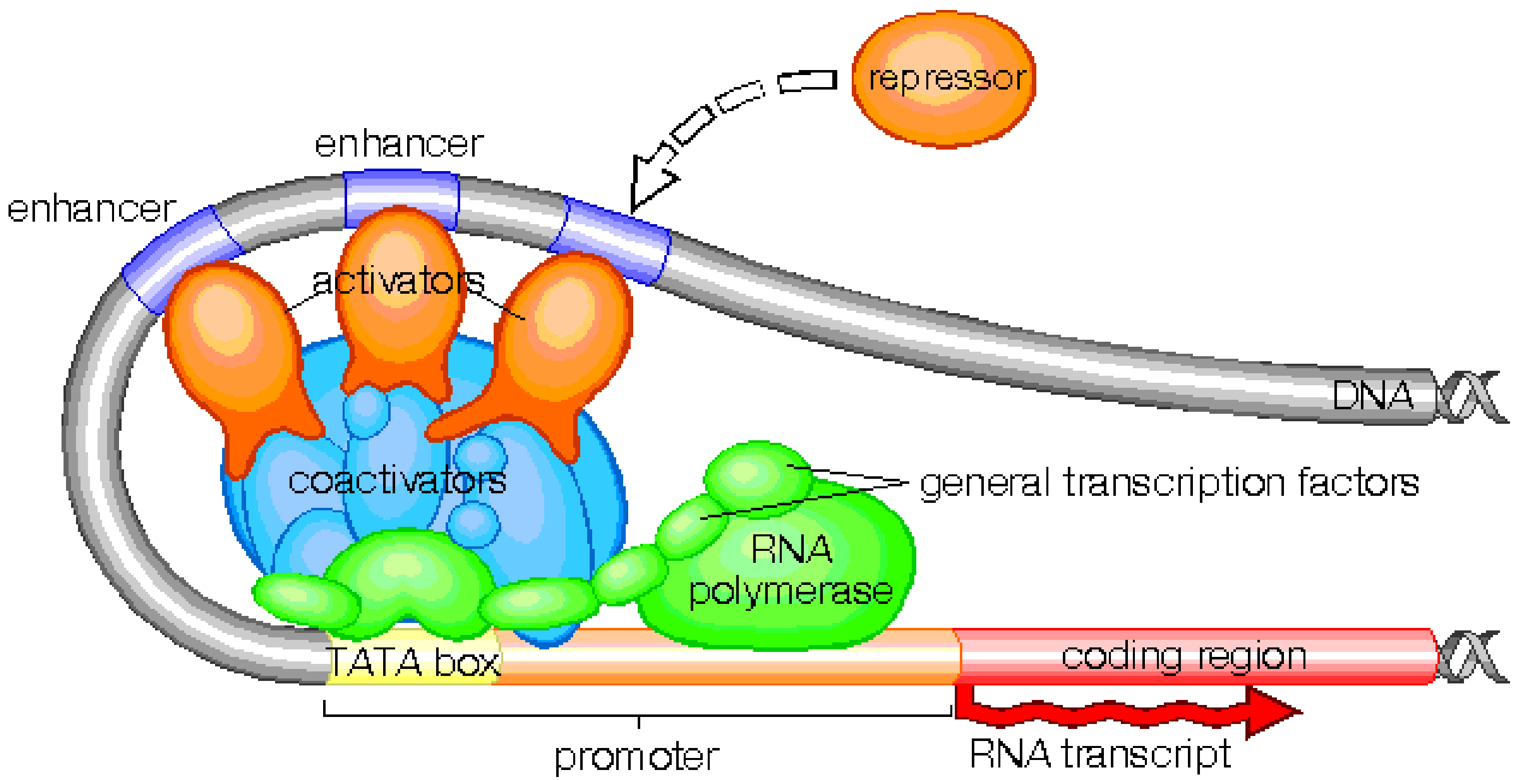
One Gene - Many Products...



Alternative Splicing – How Common?

- Preliminary estimates: 35% of human genes display alternative splicing at 5′ end
  - » Mironov, Genome Res 1999
- Human Genome Draft: ~60% of genes display alternative splicing
  - » International Human Genome Sequencing Consortium, Nature 2001

Promoter, Factors, Coactivators….

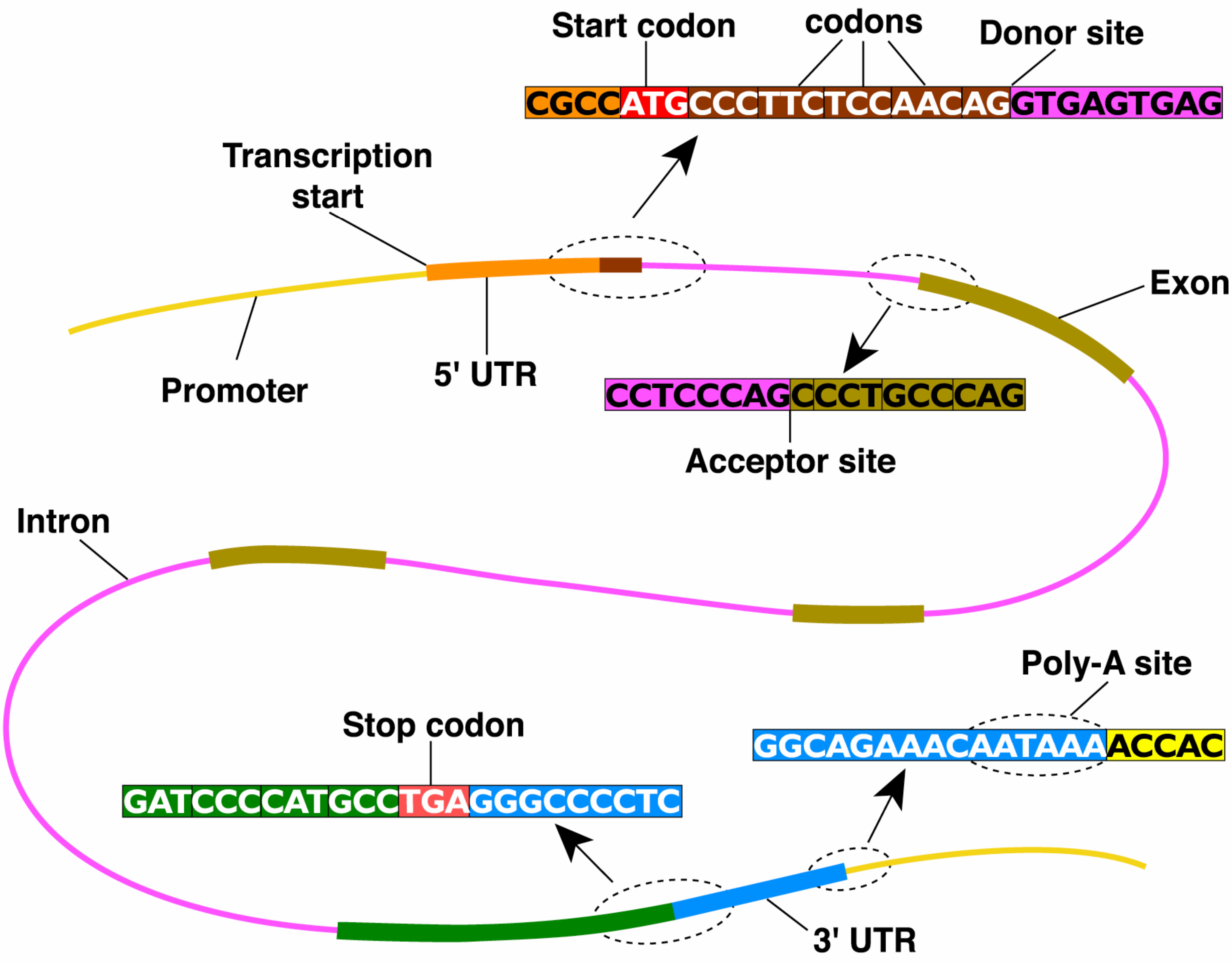


Principles of Development, Wolpert et al 1998

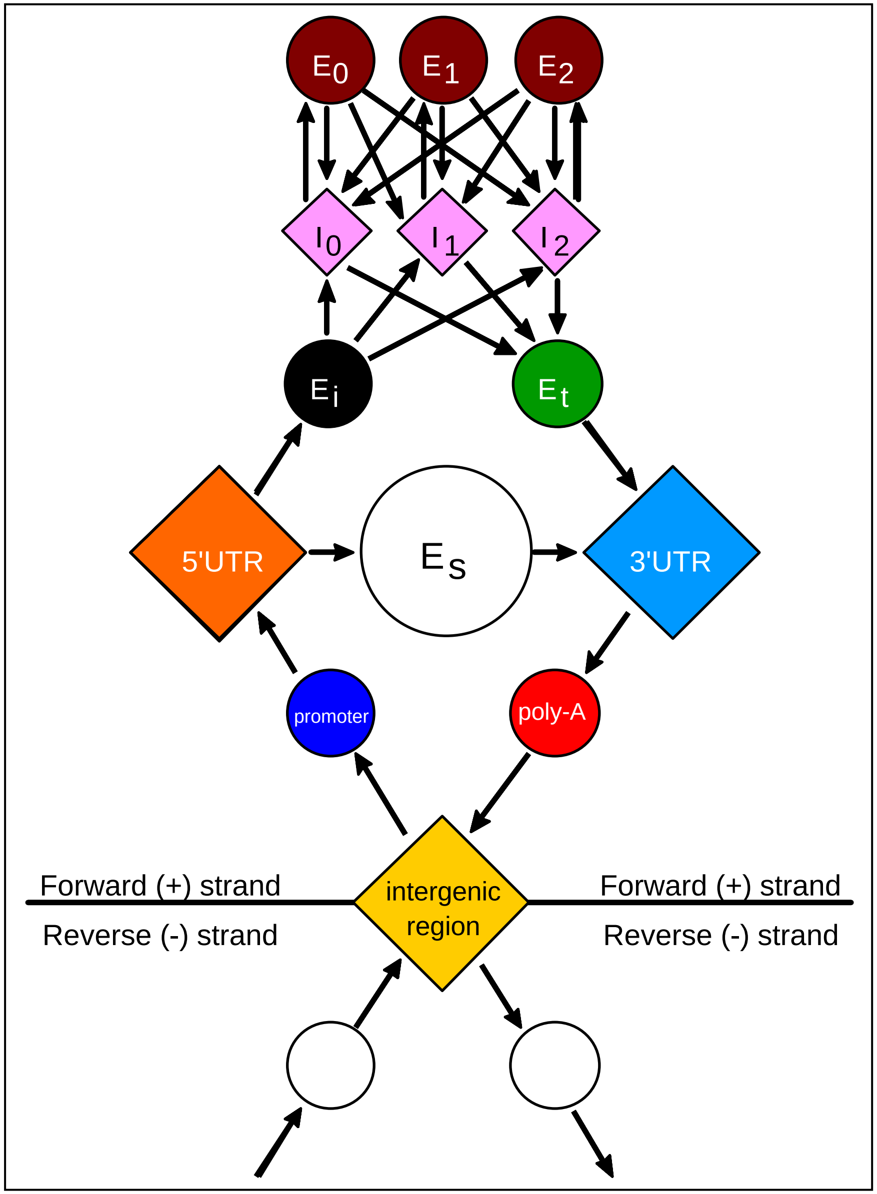
DNA Transcription and Motifs

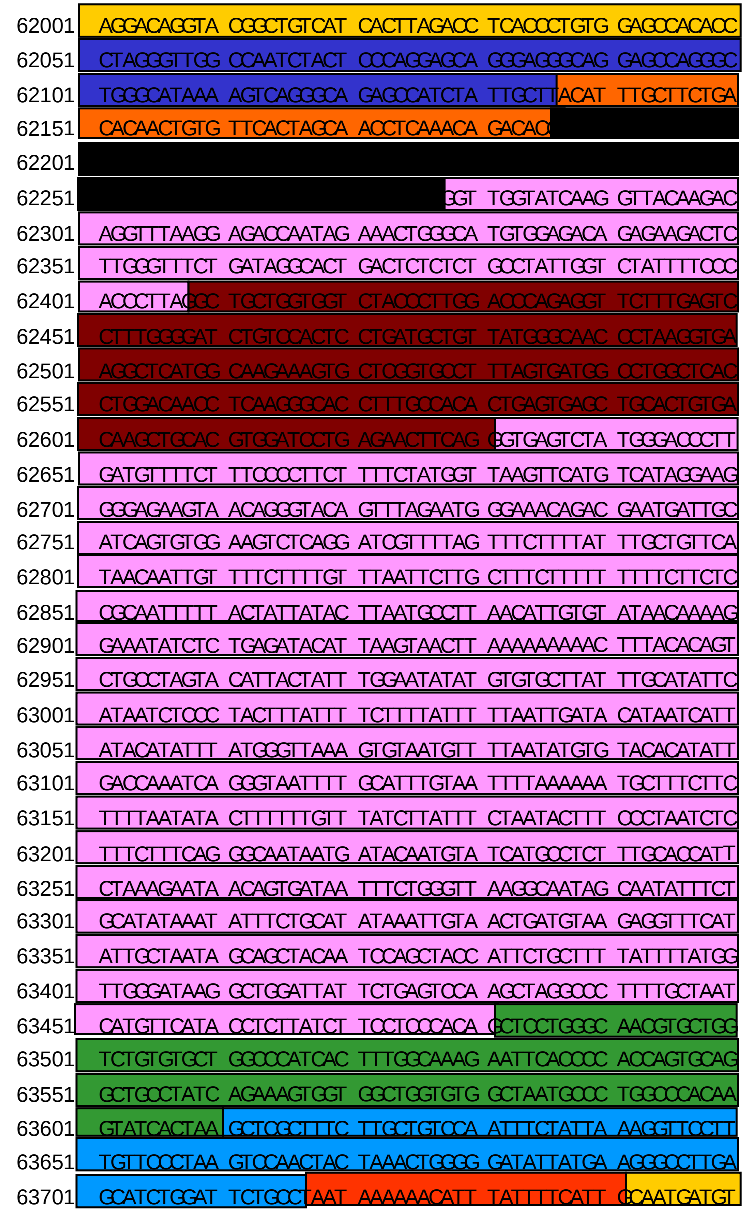
- The DNA sequence to be transcribed is longer than the translated portion. At transcription:
  - the introns are removed
  - the exons (expressed sequence) are concatenated
- The resulted sequence is formed of triples called codons
- The sequence has a unique start (ATG) and one of three stop (TAA, TAG, TGA) codons
- The intron-exon boundaries are called splice donor and acceptor sites
- There is a variety of other motifs: promoters, transcription start sites, polyA sites,branching sites.

Biologicaly significant sites

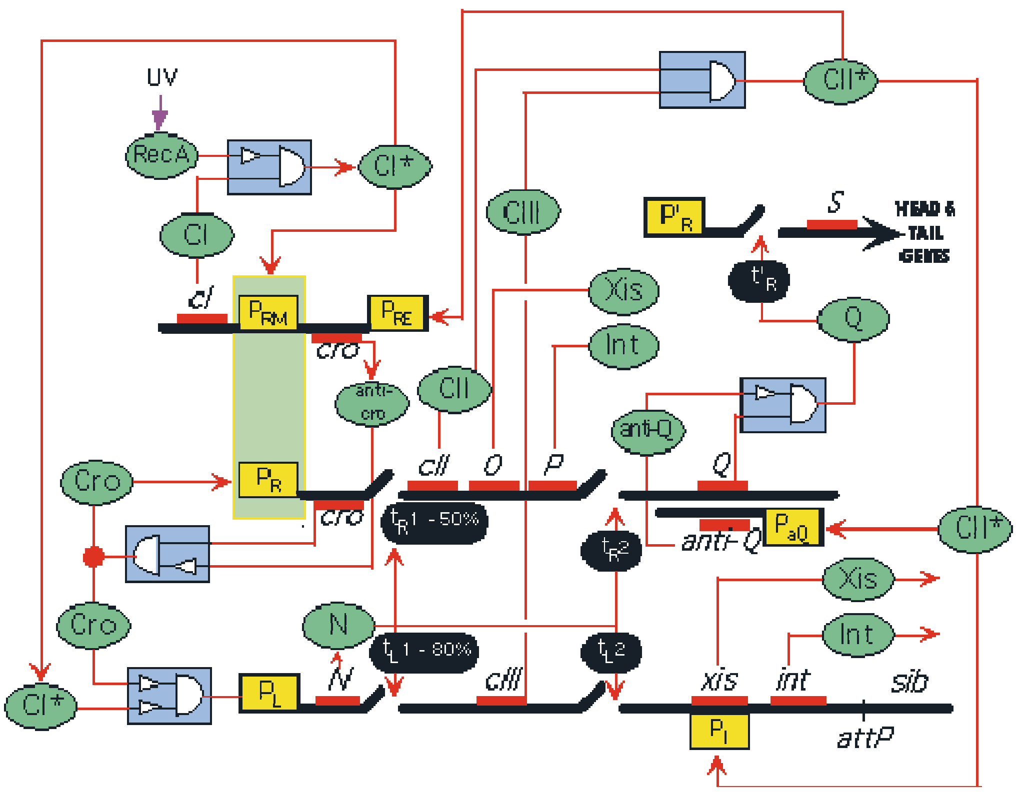


GENSCAN (Burge & Karlin)





Can We Model Regulatory Networks?



Lytic cycle decision λ-phage: 11 genes

Human Genome: ~ 31 000 – 40 000 genes

McAdams and Shapiro Science, 1995, 269, pp.650-656

Future Challenges: What We Still Don't Know

- Gene number, exact locations, and functions
- Gene regulation
- DNA sequence organization
- Chromosomal structure and organization
- Noncoding DNA types, amount, distribution, information content, and functions
- Coordination of gene expression, protein synthesis, and post-translational events
- Interaction of proteins in complex molecular machines
- Predicted vs experimentally determined gene function
- Evolutionary conservation among organisms
- Protein conservation (structure and function)
- Proteomes (total protein content and function) in organisms
- Correlation of SNPs (single-base DNA variations among individuals) with health and disease
- Disease-susceptibility prediction based on gene sequence variation
- Genes involved in complex traits and multigene diseases
- Complex systems biology including microbial consortia useful for environmental restoration
- Developmental genetics, genomics

Human Genome Program, U.S. Department of Energy, Genomics and Its Impact on Medicine and Society: A 2001 Primer, 2001

Next Step in Genomics

• **Transcriptomics** involves large-scale analysis of messenger RNAs (molecules that are transcribed from active genes) to follow when, where, and under what conditions genes are expressed.

• **Proteomics**—the study of protein expression and function—can bring researchers closer than gene expression studies to what's actually happening in the cell.

• **Structural genomics** initiatives are being launched worldwide to generate the 3-D structures of one or more proteins from each protein family, thus offering clues to function and biological targets for drug design.

• **Knockout studies** are one experimental method for understanding the function of DNA sequences and the proteins they encode. Researchers inactivate genes in living organisms and monitor any changes that could reveal the function of specific genes.

• **Comparative genomics**—analyzing DNA sequence patterns of humans and well-studied model organisms side-by-side—has become one of the most powerful strategies for identifying human genes and interpreting their function.

Human Genome Program, U.S. Department of Energy, Genomics and Its Impact on Medicine and Society: A 2001 Primer, 2001

Medicine and the New Genomics

- Gene Testing
- Gene Therapy
- Pharmacogenomics

<u>Anticipated Benefits</u>

•*improved diagnosis of disease*

•*earlier detection of genetic predispositions to disease*

•*rational drug design*

•*gene therapy and control systems for drugs*

•*personalized, custom drugs*

Human Genome Program, U.S. Department of Energy, Genomics and Its Impact on Medicine and Society: A 2001 Primer, 2001

ssues

- **Privacy and confidentiality of genetic information**.

- **Fairness in the use of genetic information** by insurers, employers, courts, schools, adoption agencies, and the military, among others.

- **Psychological impact, stigmatization, and discrimination** due to an individual's genetic differences.

- **Reproductive issues** including adequate and informed consent and use of genetic information in reproductive decision making.

- **Clinical issues** including the education of doctors and other health-service providers, people identified with genetic conditions, and the general public about capabilities, limitations, and social risks; and implementation of standards and quality-control measures.

Human Genome Program, U.S. Department of Energy, Genomics and Its Impact on Medicine and Society: A 2001 Primer, 2001

ssues (cont.)

- **Uncertainties associated with gene tests for susceptibilities and complex conditions** (e.g., heart disease, diabetes, and Alzheimer’s disease).
- **Fairness in access to advanced genomic technologies**.
- **Conceptual and philosophical implications** regarding human responsibility, free will vs genetic determinism, and concepts of health and disease.
- **Health and environmental issues** concerning genetically modified (GM) foods and microbes.
- **Commercialization of products** including property rights (patents, copyrights, and trade secrets) and accessibility of data and materials.

Human Genome Program, U.S. Department of Energy, Genomics and Its Impact on Medicine and Society: A 2001 Primer, 2001

Anticipated Benefits

Molecular Medicine

- improved diagnosis of disease
- earlier detection of genetic predispositions to disease
- rational drug design
- gene therapy and control systems for drugs
- pharmacogenomics "custom drugs"

Microbial Genomics

- rapid detection and treatment of pathogens (disease-causing microbes) in medicine
- new energy sources (biofuels)
- environmental monitoring to detect pollutants
- protection from biological and chemical warfare
- safe, efficient toxic waste cleanup

Human Genome Program, U.S. Department of Energy, Genomics and Its Impact on Medicine and Society: A 2001 Primer, 2001

Anticipated Benefits

Risk Assessment

• assess health damage and risks caused by radiation exposure, including low-dose exposures
• assess health damage and risks caused by exposure to mutagenic chemicals and cancer-causing toxins
• reduce the likelihood of heritable mutations

Bioarchaeology, Anthropology, Evolution, and Human Migration

• study evolution through germline mutations in lineages
• study migration of different population groups based on maternal inheritance
• study mutations on the Y chromosome to trace lineage and migration of males
• compare breakpoints in the evolution of mutations with ages of populations and historical events

Human Genome Program, U.S. Department of Energy, Genomics and Its Impact on Medicine and Society: A 2001 Primer, 2001

Anticipated Benefits

DNA Identification (Forensics)

- identify potential suspects whose DNA may match evidence left at crime scenes
- exonerate persons wrongly accused of crimes
- identify crime and catastrophe victims
- establish paternity and other family relationships
- identify endangered and protected species as an aid to wildlife officials (could be used for prosecuting poachers)
- detect bacteria and other organisms that may pollute air, water, soil, and food
- match organ donors with recipients in transplant programs
- determine pedigree for seed or livestock breeds
- authenticate consumables such as caviar and wine

Human Genome Program, U.S. Department of Energy, Genomics and Its Impact on Medicine and Society: A 2001 Primer, 2001

Anticipated Benefits

Agriculture, Livestock Breeding, and Bioprocessing

- disease-, insect-, and drought-resistant crops
- healthier, more productive, disease-resistant farm animals
- more nutritious produce
- biopesticides
- edible vaccines incorporated into food products
- new environmental cleanup uses for plants like tobacco

Human Genome Program, U.S. Department of Energy, Genomics and Its Impact on Medicine and Society: A 2001 Primer, 2001

Future: Challenges for Bioinformatics

- **Data visualization**
- **Extraction of sensible information form tons of data**
- **Developing of ways to access different data-structures trough common interface: in essence how to retrive all the relevant data from all of the existing databases with one single query**
- **Omics**
- **Holistic Understanding of Life**
- **Models of life**
- **Systems Biology**
- **Better drugs**
- **"Personalized" therapy**

What Makes the Difference?

Caenorhabtitis elegans

- 959 somatic cells
- 300 neurons
- >19 000 genes

Homo Sapiens

- 50×10^{12} cells
- 50×10^{9} brain cells
- 35 – 40 000 genes

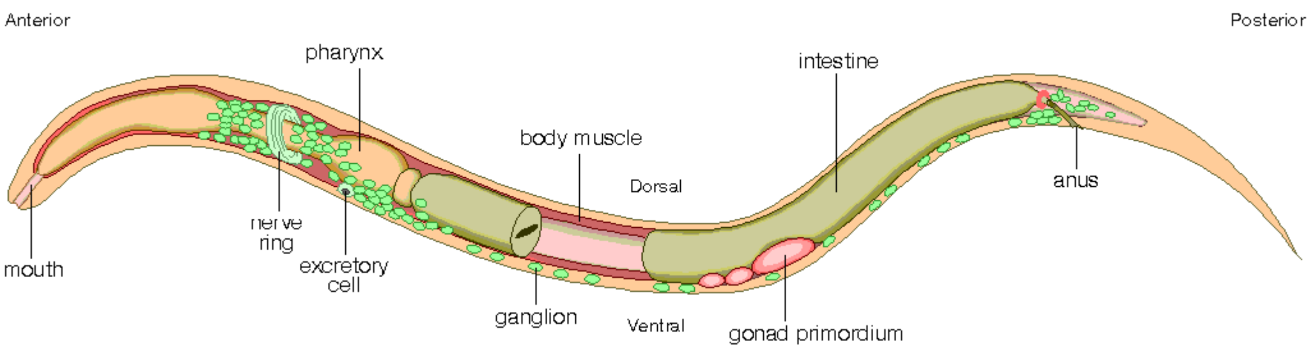


Image from: Principles of Development, Wolpert et al 1998

What Makes Us Human?

- Cognitive skills
  - complex language
  - long-term planning
  - advanced ability to give and receive instructions

- Human and Chimpanzee Genomes share 99% identity

- The only difference at the biochemical level between Humans and other Mammals
  - humans do not express the hydroxylated form of a sialic acid (*N*-glycolyl-neuraminic acid) on the surface of cells and secreted proteins

Analysis of gene expression data

Thesis: the analysis of gene expression data is going to be big in 21st century statistics

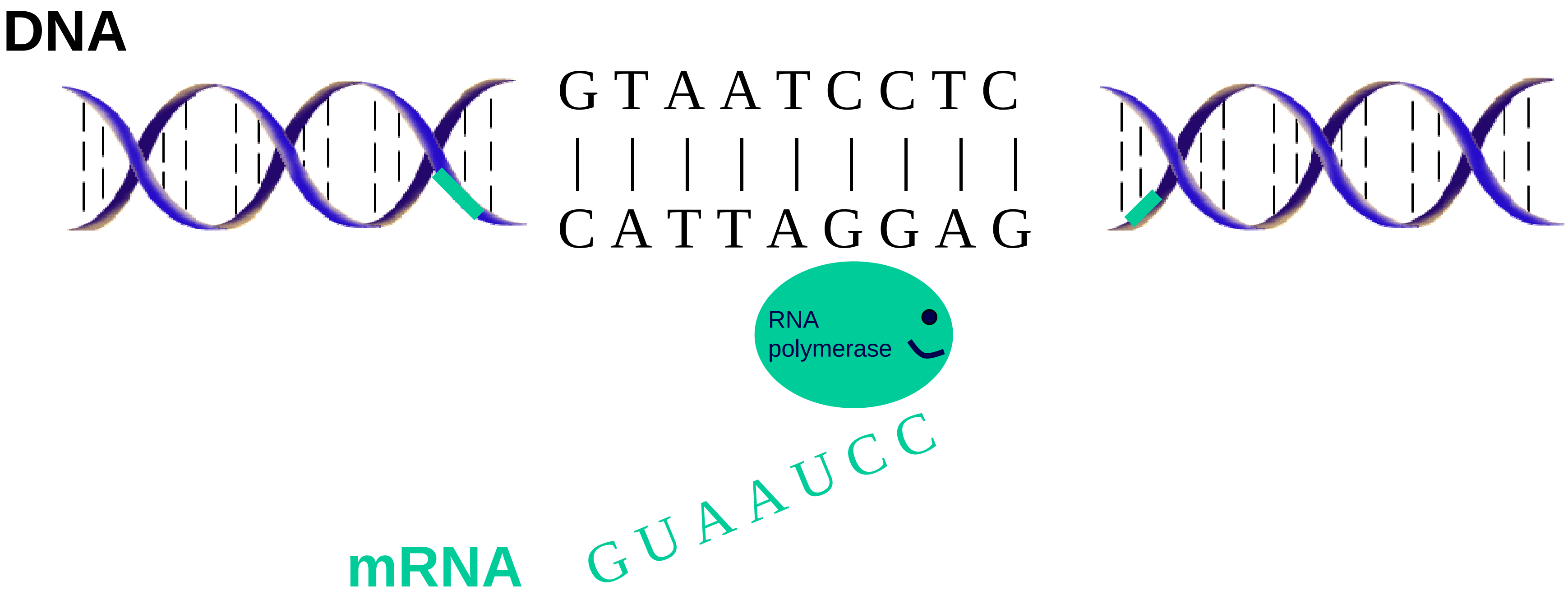
Many different technologies, including

- Serial analysis of gene expression (SAGE)
- Short oligonucleotide arrays (Affymetrix)
- Long oligo arrays (Agilent)
- Fibre optic arrays (Illumina)
- cDNA arrays (Brown/Botstein)*

Common themes

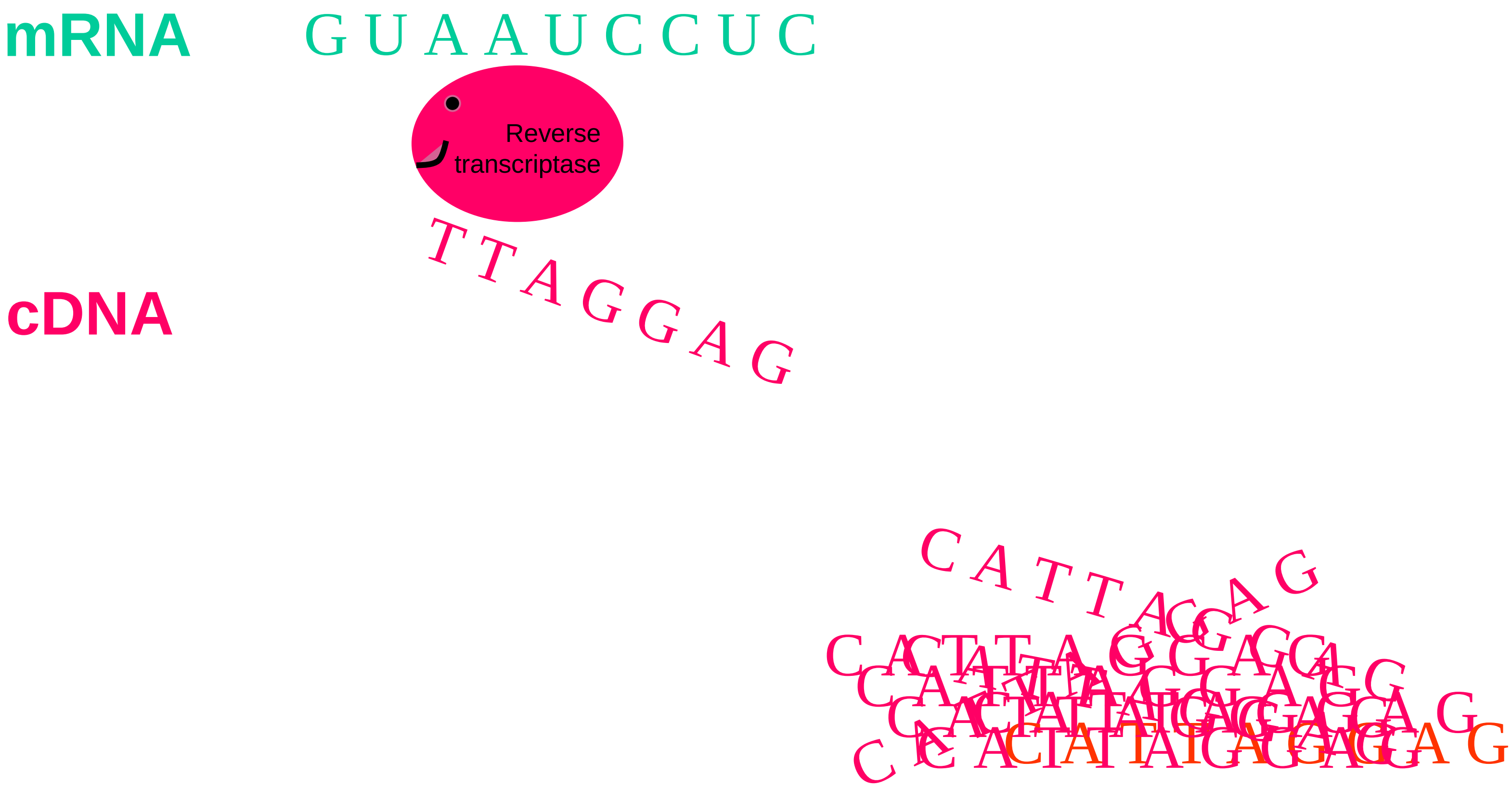
- Parallel approach to collection of very large amounts of data (by biological standards)
- Sophisticated instrumentation, requires some understanding
- Systematic features of the data are at least as important as the random ones
- Often more like industrial process than single investigator lab research
- Integration of many data types: clinical, genetic, molecular.....databases

Transcription



Reverse transcription

Clone cDNA strands, complementary to the mRNA

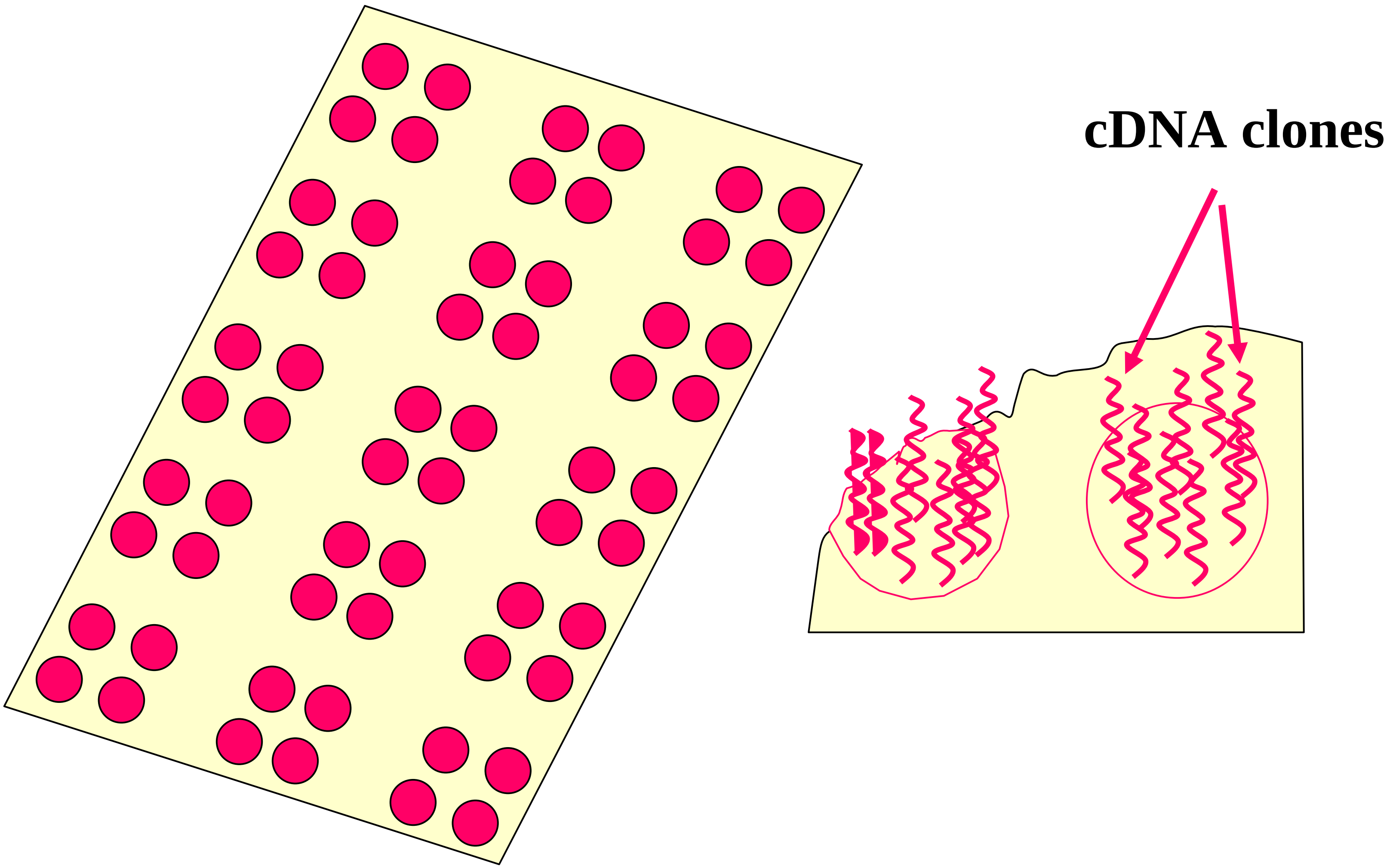


cDNA microarray experiments

mRNA levels compared in many different contexts

- Different tissues, same organism (brain v. liver)
- Same tissue, same organism (ttt v. ctl, tumor v. non-tumor)
- Same tissue, different organisms
- Time course experiments (effect of ttt, development)
- Other special designs (e.g. to detect spatial patterns).

cDNA microarrays

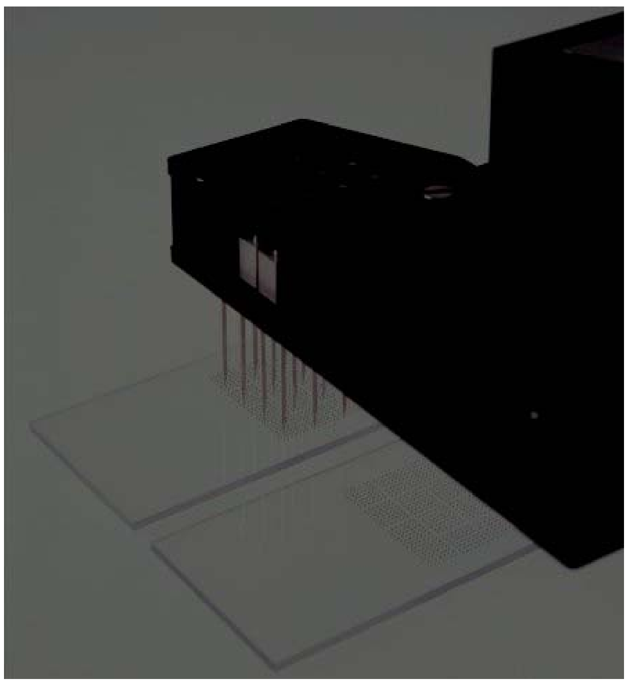


cDNA microarrays

Compare the genetic expression in two samples of cells

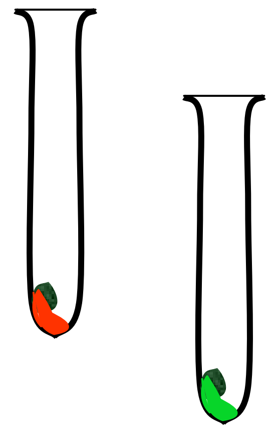
PRINT

cDNA from one gene on each spot



SAMPLES

cDNA labelled red/green

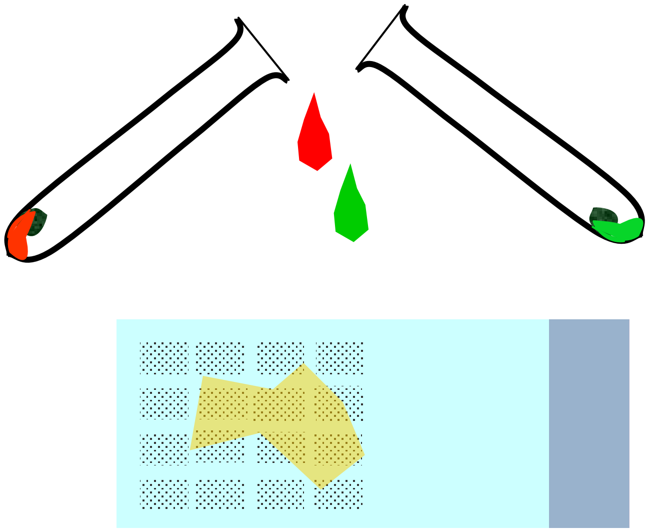


e.g. treatment / control

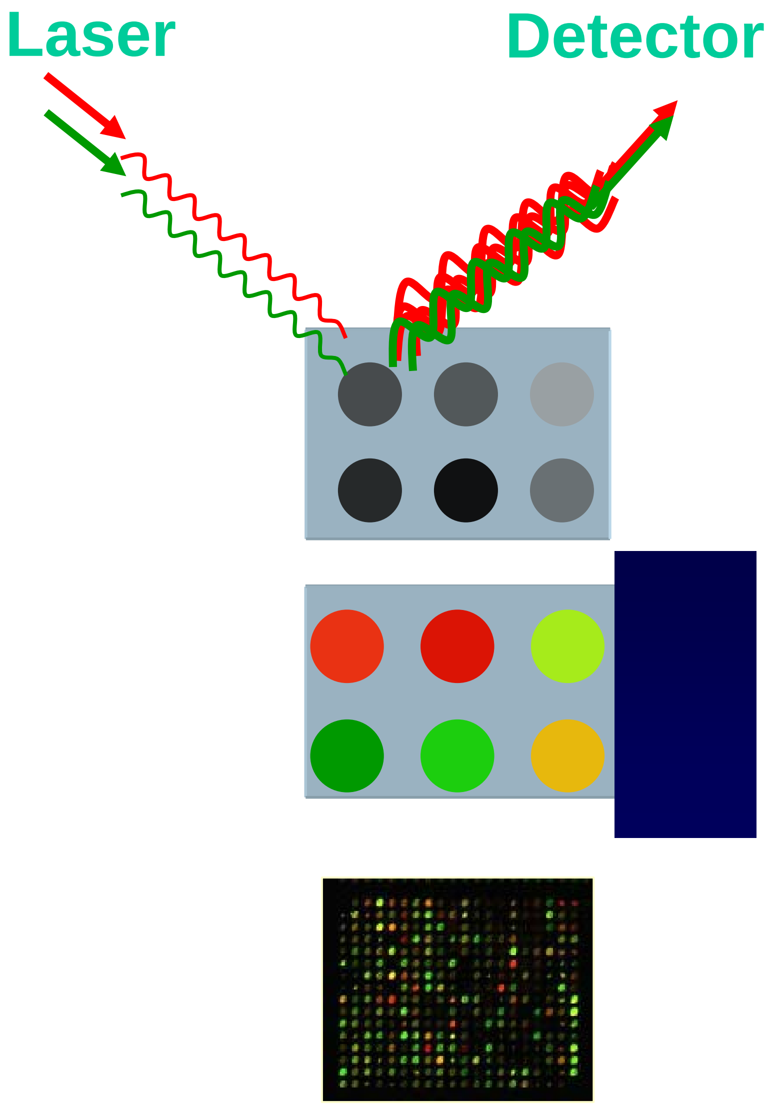
normal / tumor tissue

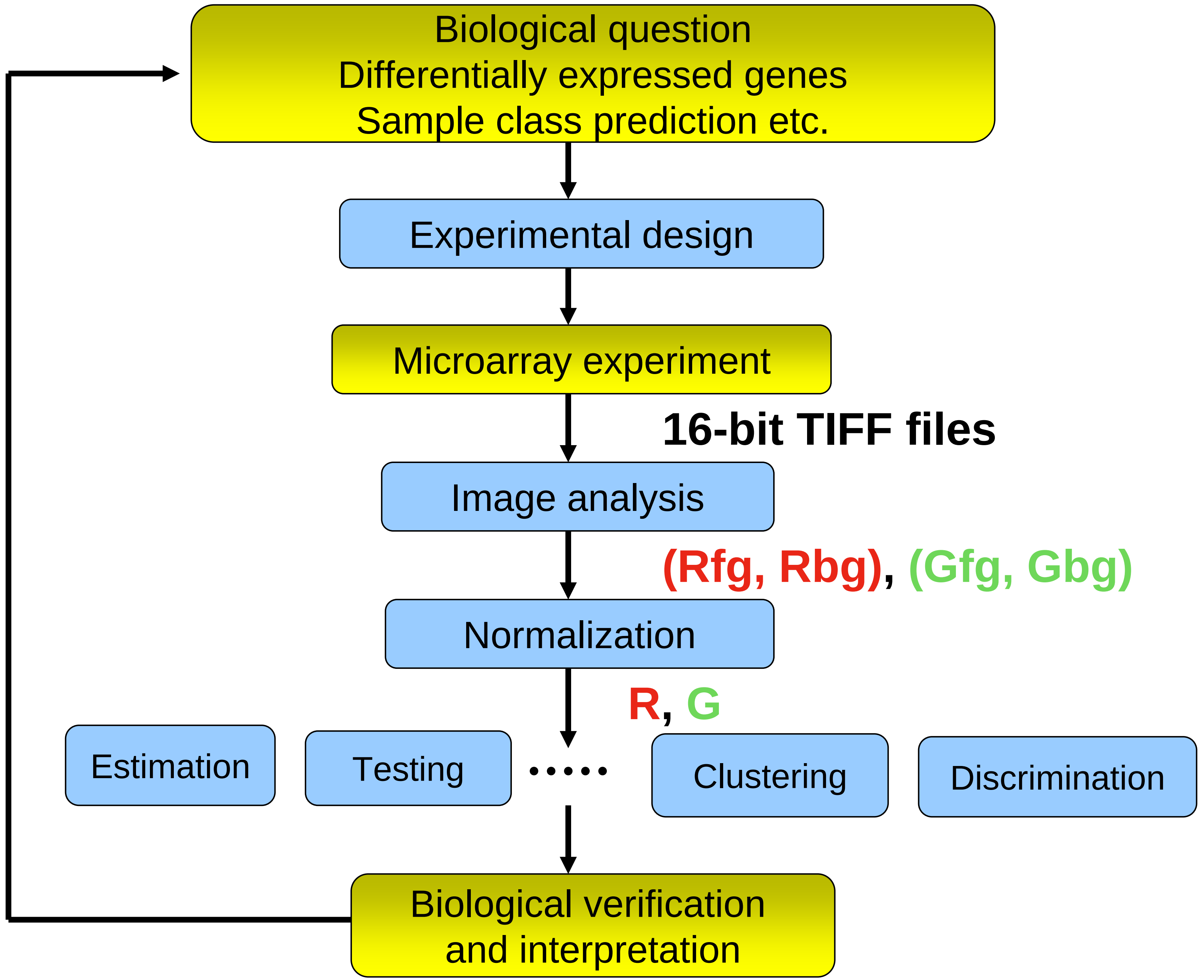
HYBRIDIZE

Add equal amounts of labelled mRNA samples to microarray.



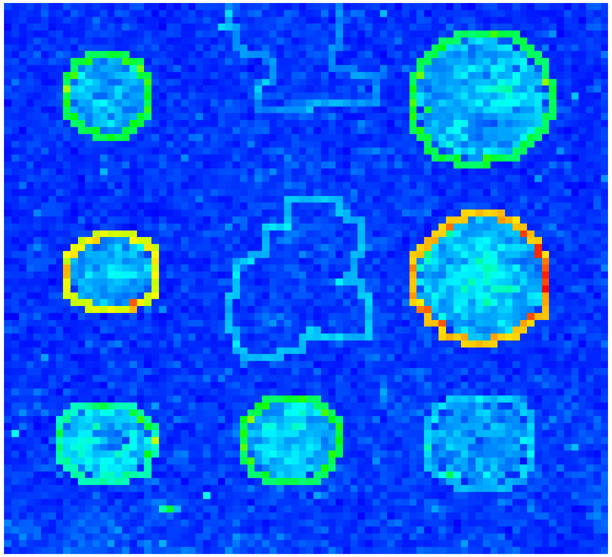
SCAN



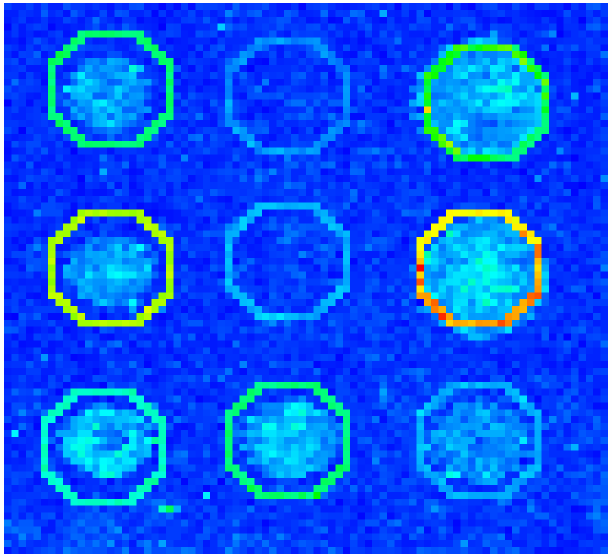


Biological question
Differentially expressed genes
Sample class prediction etc.
Experimental design
Microarray experiment
16-bit TIFF files
Image analysis
(Rfg, Rbg), (Gfg, Gbg)
Normalization
R, G
Estimation
Testing
Clustering
Discrimination
Biological verification
and interpretation

Segmentation: limitation of the fixed circle method



SRG

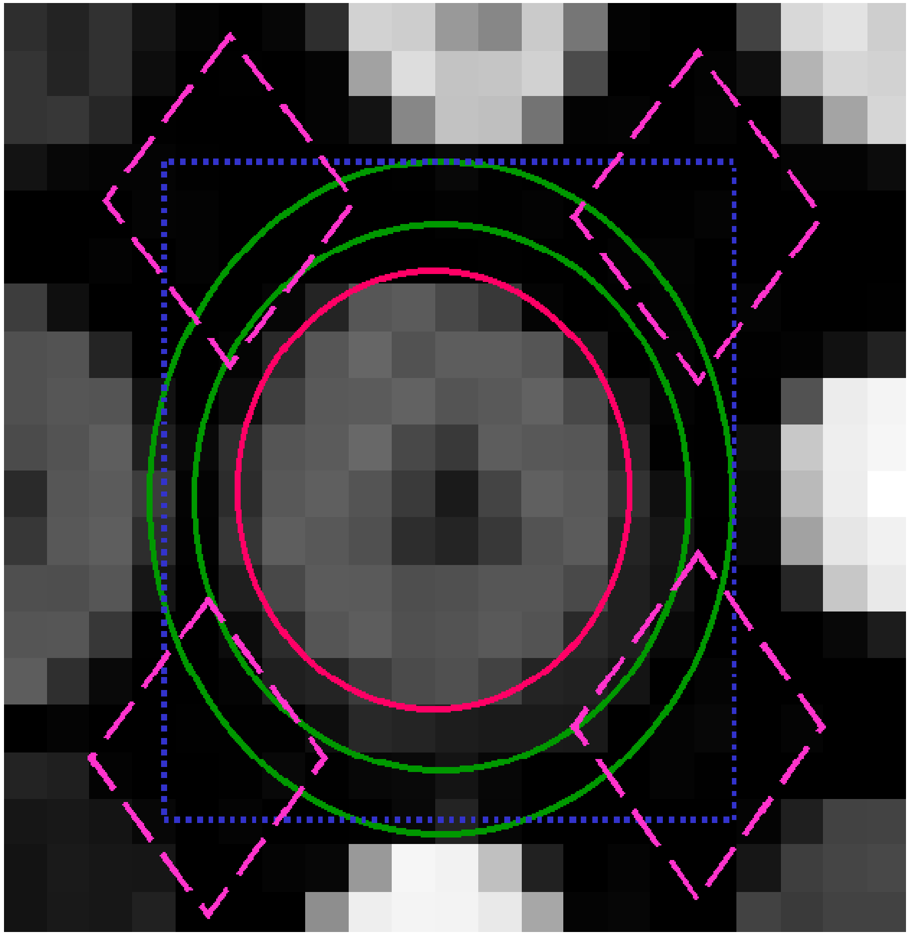


Fixed Circle

Inside the boundary is spot (foreground), outside is not.

Some local backgrounds

Single channel
grey scale



Quantification of expression

For each spot on the slide we calculate

Red intensity = Rfg - Rbg

fg = foreground, bg = background, and

Green intensity = Gfg - Gbg

and combine them in the log (base 2) ratio

$$\mathbf{Log_2(\text{Red intensity} / \text{Green intensity})}$$

Some statistical questions

Image analysis: addressing, segmenting, quantifying

Normalisation: within and between slides

Quality: of images, of spots, of (log) ratios

Which genes are (relatively) up/down regulated?

Assigning p-values to tests/confidence to results.

Some statistical questions, ctd

Planning of experiments: design, sample size

Discrimination and allocation of samples

Clustering, classification: of samples, of genes

Selection of genes relevant to any given analysis

Analysis of time course, factorial and other special experiments…….....& much more.

Some bioinformatic questions

Connecting spots to databases, e.g. to sequence, structure, and pathway databases

Discovering short sequences regulating sets of genes: direct and inverse methods

Relating expression profiles to structure and function, e.g. protein localisation

Identifying novel biochemical or signalling pathways, ………..and much more.