Introduction to bioinformatics

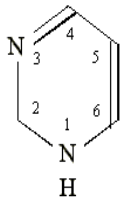
Lecture 2

Genes and Genomes

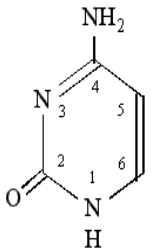
DNA sequence

```
.....acctc ctgtgcaaga acatgaaaca cctgtggttc ttccttctcc
tggtggcagc tcccagatgg gtcctgtccc aggtgcacct gcaggagtcg
ggcccaggac tggggaagcc tccagagctc aaaaccccac ttggtgacac
aactcacaca tgcccacggt gcccagagcc caaatcttgt gacacacctc
ccccgtgccc acggtgccca gagcccaaat cttgtgacac acctccccca
tgcccacggt gcccagagcc caaatcttgt gacacacctc ccccgtgccc
ccggtgccca gcacctgaac tcttgggagg accgtcagtc ttcctcttcc
ccccaaaacc caaggatacc cttatgattt cccggacccc tgaggtcacg
tgcgtggtgg tggacgtgag ccacgaagac cccgaggtcc agttcaagtg
gtacgtggac ggcgtggagg tgcataatgc caagacaaag ctgcgggagg
agcagtacaa cagcacgttc cgtgtggtca gcgtcctcac cgtcctgcac
caggactggc tgaacggcaa ggagtacaag tgcaaggtct ccaacaaagc
aaccaagtca gcctgacctg cctggtcaaa ggcttctacc ccagcgacat
cgccgtggag tgggagagca atgggcagcc ggagaacaac tacaacacca
cgcctcccat gctggactcc gacggctcct tcttcctcta cagcaagctc
accgtggaca agagcaggtg gcagcagggg aacatcttct catgctccgt
gatgcatgag gctctgcaca accgctacac gcagaagagc ctctc.....
```

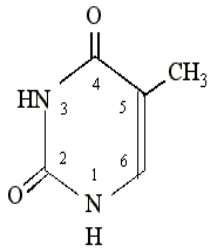
Four DNA nucleotide building blocks



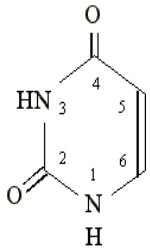
Pyrimidine



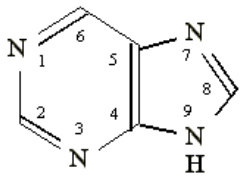
Cytosine
111.1 gm/mol



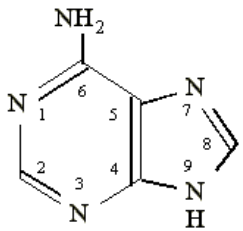
Thymine
126.1 gm/mol



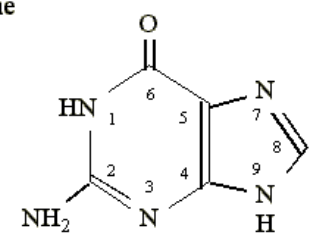
Uracil
(RNA)



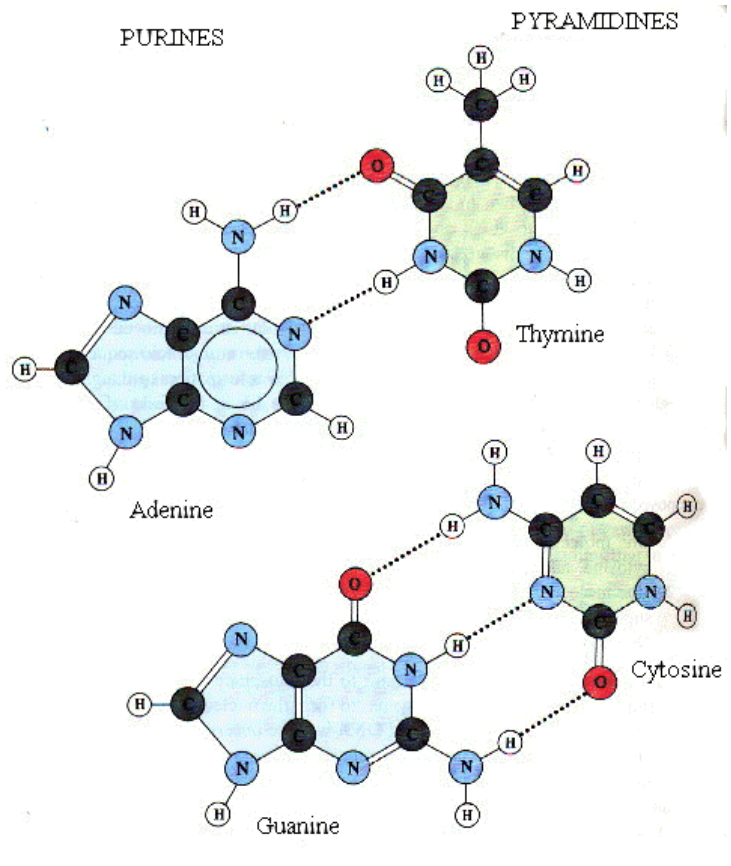
Purine



Adenine
135.1 gm/mol



Guanine
151.1 gm/mol



DNA compositional biases

- **Base composition of genomes**:
- *E. coli*: 25% A, 25% C, 25% G, 25% T
- *P. falciparum* (Malaria parasite): 82% A+T

- **Translation initiation**:
- ATG (AUG) is the near universal **motif** indicating the start of translation in DNA coding sequence.

First Letter	Second Letter				Third Letter
	U	C	A	G	
U	phenylalanine PHE	serine SER	tyrosine TYR	cysteine CYS	U
	phenylalanine PHE	serine SER	tyrosine TYR	cysteine CYS	C
	leucine LEU	serine SER	stop	stop	A
	leucine LEU	serine SER	stop	tryptophan TRP	G

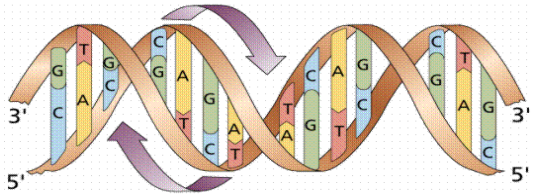
First Letter	Second Letter				Third Letter
	U	C	A	G	
C	leucine LEU	proline PRO	histidine HIS	arginine ARG	U
	leucine LEU	proline PRO	histidine HIS	arginine ARG	C
	leucine LEU	proline PRO	glutamine GLN	arginine ARG	A
	leucine LEU	proline PRO	glutamine GLN	arginine ARG	G

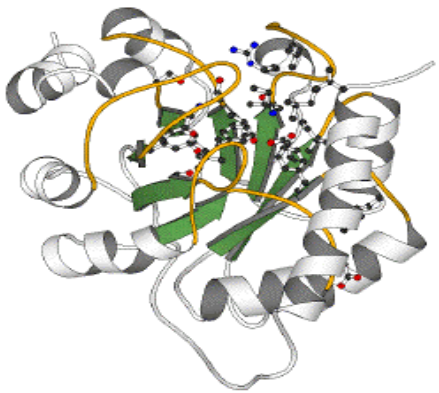
First Letter	Second Letter				Third Letter
	U	C	A	G	
A	isoleucine ILE	threonine THR	asparagine ASN	serine SER	U
	isoleucine ILE	threonine THR	asparagine ASN	serine SER	C
	isoleucine ILE	threonine THR	lysine LYS	arginine ARG	A
	methionine MET	threonine THR	lysine LYS	arginine ARG	G

First Letter	Second Letter				Third Letter
	U	C	A	G	
G	valine VAL	alanine ALA	aspartate ASP	glycine GLY	U
	valine VAL	alanine ALA	aspartate ASP	glycine GLY	C
	valine VAL	alanine ALA	glutamate GLU	glycine GLY	A
	valine VAL	alanine ALA	glutamate GLU	glycine GLY	G

Amino Acid	SLC	DNA codons
Isoleucine	I	ATT, ATC, ATA
Leucine	L	CTT, CTC, CTA, CTG, TTA, TTG
Valine	V	GTT, GTC, GTA, GTG
Phenylalanine	F	TTT, TTC
Methionine	M	ATG
Cysteine	c	TGT, TGC
Alanine	A	GCT, GCC, GCA, GCG
Glycine	G	GGT, GGC, GGA, GGG
Proline	P	CCT, CCC, CCA, CCG
Threonine	T	ACT, ACC, ACA, ACG
Serine	S	TCT, TCC, TCA, TCG, AGT, AGC
Tyrosine	Y	TAT, TAC
Tryptophan	W	TGG
Glutamine	Q	CAA, CAG
Asparagine	N	AAT, AAC
Histidine	H	CAT, CAC
Glutamic acid	E	GAA, GAG
Aspartic acid	D	GAT, GAC
Lysine	K	AAA, AAG
Arginine	R	CGT, CGC, CGA, CGG, AGA, AGG
Stop codons	Stop	TAA, TAG, TGA

A gene codes for a protein





DNA

transcription

mRNA

translation

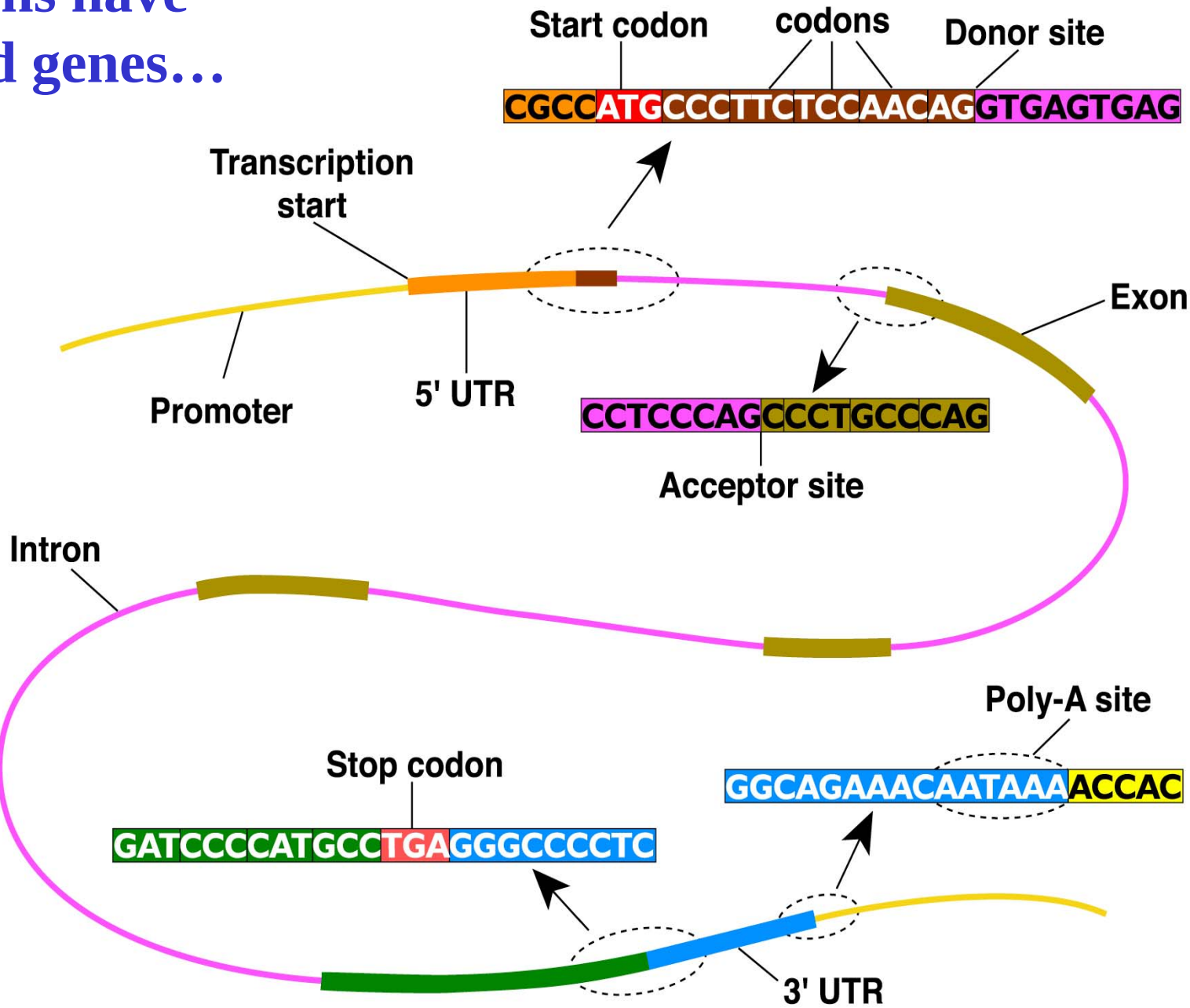
Protein

CCTGAGCCAACTATTGATGAA

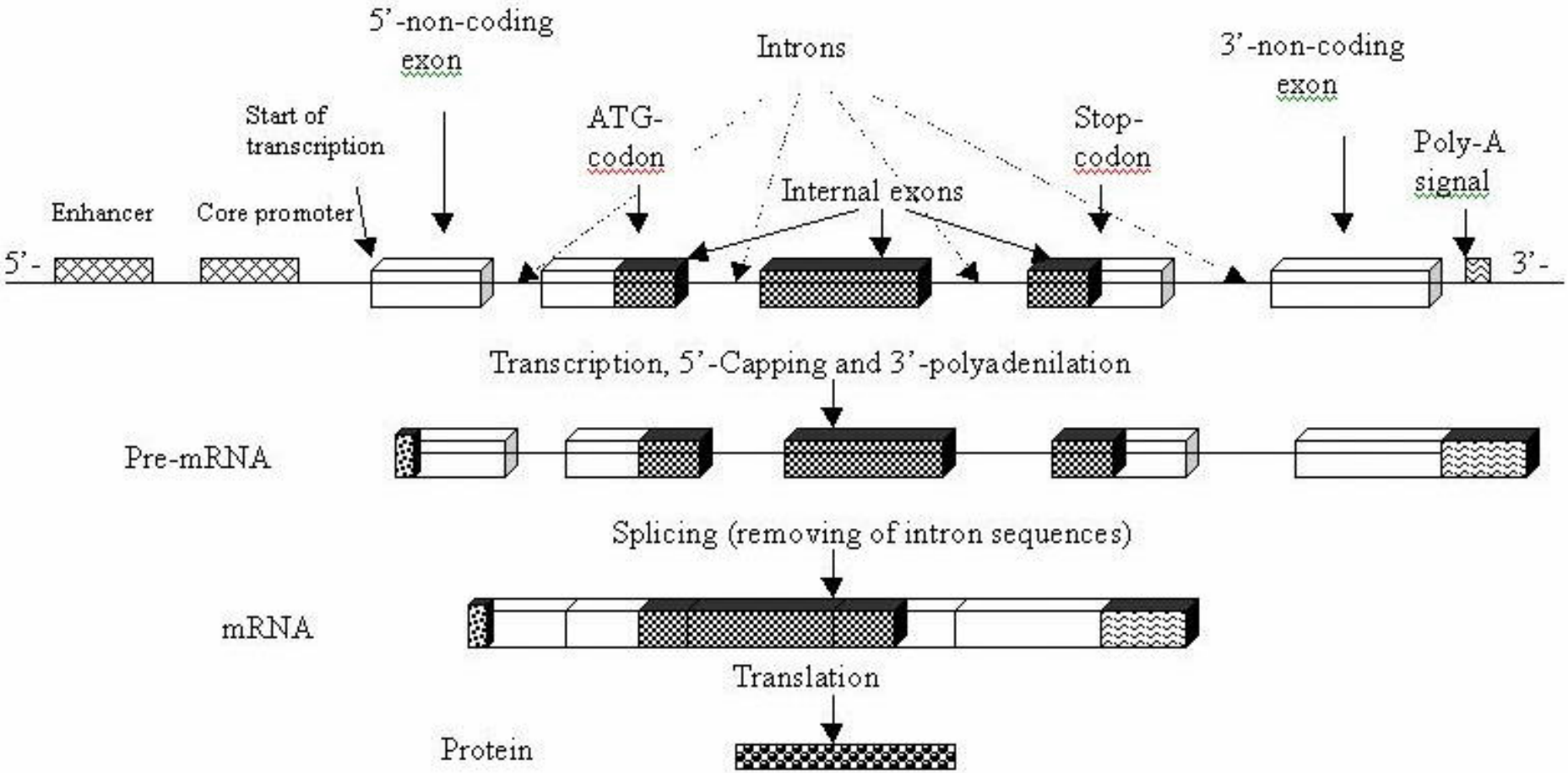
CCUGAGCCAACUAUUGAUGAA

PEPTIDE

Humans have spliced genes…



DNA makes RNA makes Protein



Some facts about human genes

- Comprise about 3% of the genome
- Average gene length: ~ 8,000 bp
- Average of 5-6 exons/gene
- Average exon length: ~200 bp
- Average intron length: ~2,000 bp
- ~8% genes have a single exon

- Some exons can be as small as 1 or 3 bp.
- HUMFMR1S is not atypical: 17 exons 40-60 bp long comprising 3% of a 67,000 bp gene

Genetic diseases

- Many diseases run in families and are a result of genes which predispose such family members to these illnesses
- Examples are Alzheimer's disease, cystic fibrosis (CF), breast or colon cancer, or heart diseases.
- Some of these diseases can be caused by a problem within a single gene, such as with CF.

Genetic diseases (Cont.)

- For other illnesses, like heart disease, at least 20-30 genes are thought to play a part, and it is still unknown which combination of problems within which genes are responsible.
- With a “problem” within a gene is meant that a single nucleotide or a combination of those within the gene are causing the disease (or make that the body is not sufficiently fighting the disease).
- Persons with different combinations of these nucleotides could then be unaffected by these diseases.

Genetic diseases (Cont.)
Cystic Fibrosis

- Known since very early on ("Celtic gene"). One in 10,000 people displays disease, 1 in 20 is an *unaffected carrier* of an abnormal CF gene. These people usually are unaware that they are carriers. About 30,000 Americans, 3000 Canadians, and 20,000 Europeans have CF.
- Inherited autosomal recessive condition (Chr. 7)
- Symptoms:
  - Clogging and infection of lungs (early death)
  - Intestinal obstruction
  - Reduced fertility and (male) anatomical anomalies

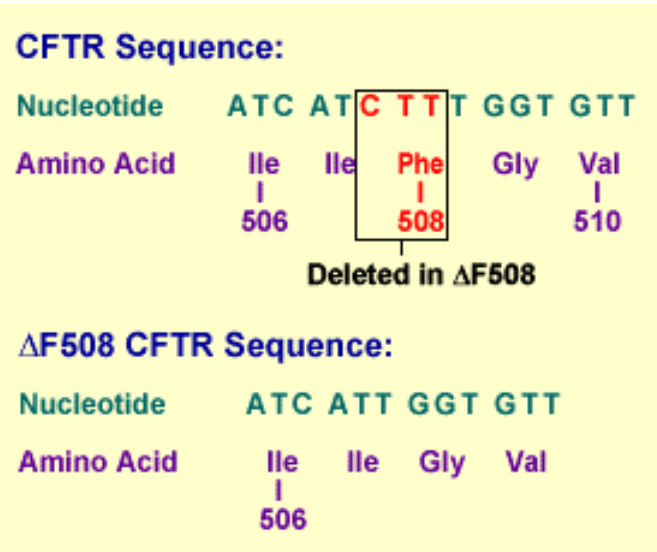
Genetic diseases (Cont.)
Cystic Fibrosis

- **Name of Gene Product:** cystic fibrosis transmembrane conductance regulator (CFTR)
- CFTR is an ABC (ATP-binding cassette) transporter or traffic ATPase. These proteins transport molecules such as sugars, peptides, inorganic phosphate, chloride, and metal cations across the cellular membrane. CFTR transports chloride ions (Cl^-) ions across the membranes of cells in the lungs, liver, pancreas, digestive tract, reproductive tract, and skin.

Genetic diseases (Cont.)

Cystic Fibrosis

- CF gene CFTR has **3-bp deletion** leading to **Del508 (Phe)** in 1480 aa protein (epithelial Cl^- channel) – the protein is degraded in the Endoplasmatic Reticulum (ER) instead of being inserted into cell membrane



The deltaF508 deletion is the most common cause of cystic fibrosis. The isoleucine (Ile) at amino acid position 507 remains unchanged because both ATC and ATT code for isoleucine

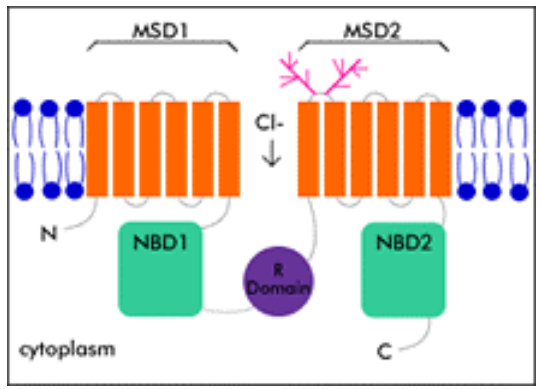
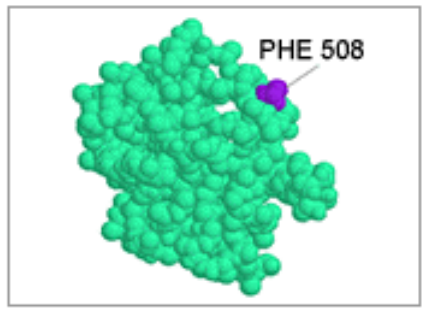


Diagram depicting the five domains of the CFTR membrane protein (Sheppard 1999).



Theoretical Model of NBD1. PDB identifier 1NBD as viewed in Protein Explorer http://proteinexplorer.org

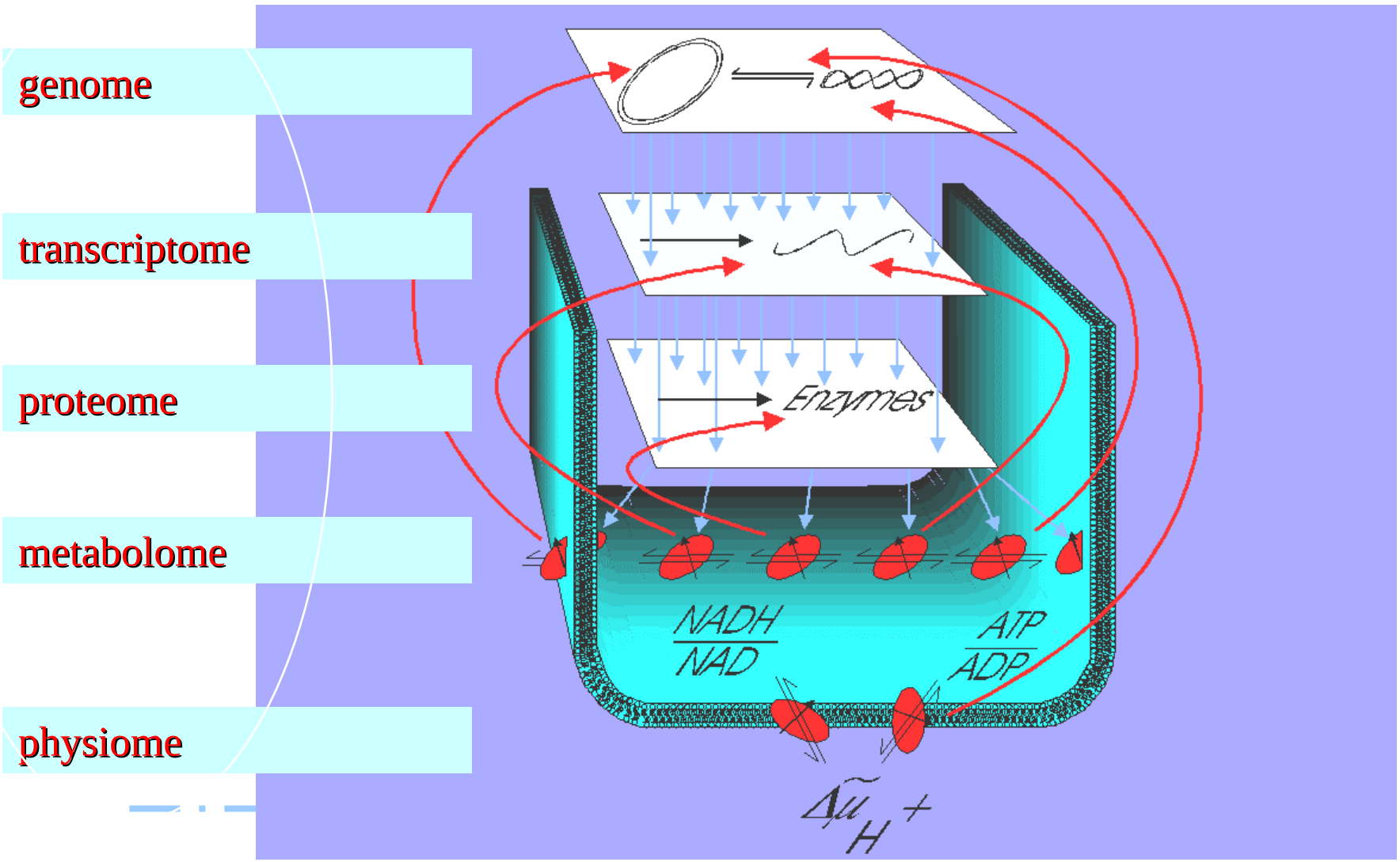
Genomic Data Sources

- DNA/protein sequence
- Expression (microarray)
- Proteome (xray, NMR, mass spectrometry)
- Metabolome
- Physiome *(spatial, temporal)*

} *Integrative bioinformatics*

Genomic Data Sources

Vertical Genomics



Remark

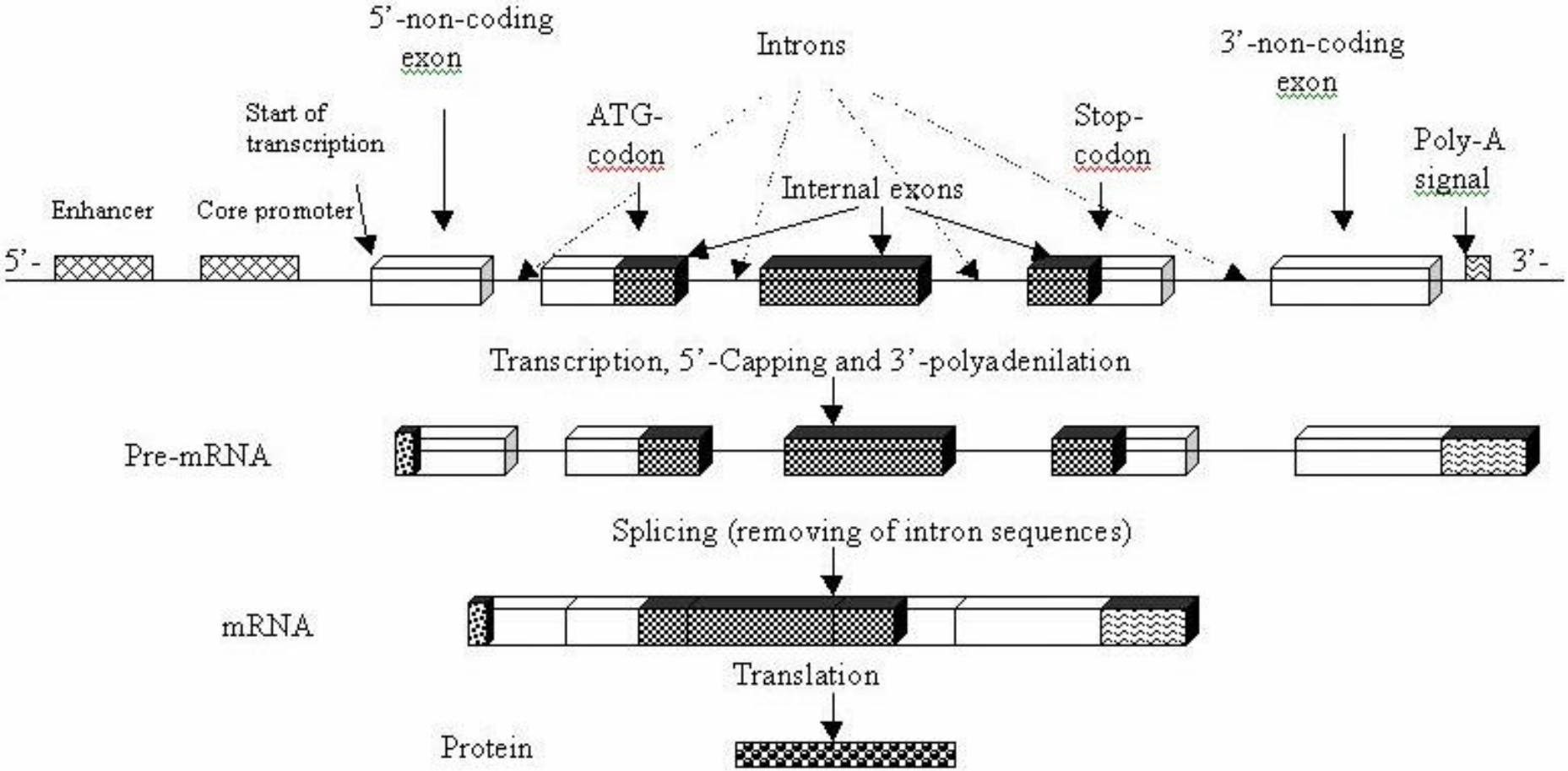
- Identifying (annotating) human genes, i.e. finding what they are and what they do, is a difficult problem. It is considerably harder than the early success story for ß-globin might suggest (see Lesk's "Introduction to bioinf").

- **The human factor VIII gene (whose mutations cause hemophilia A) is spread over ~186,000 bp. It consists of 26 exons ranging in size from 69 to 3,106 bp, and its 25 introns range in size from 207 to 32,400 bp. The complete gene comprises ~9 kb of exon and ~177 kb of intron.**

- The biggest human gene yet is for **dystrophin**. It has >30 exons and is spread over 2.4 million bp.

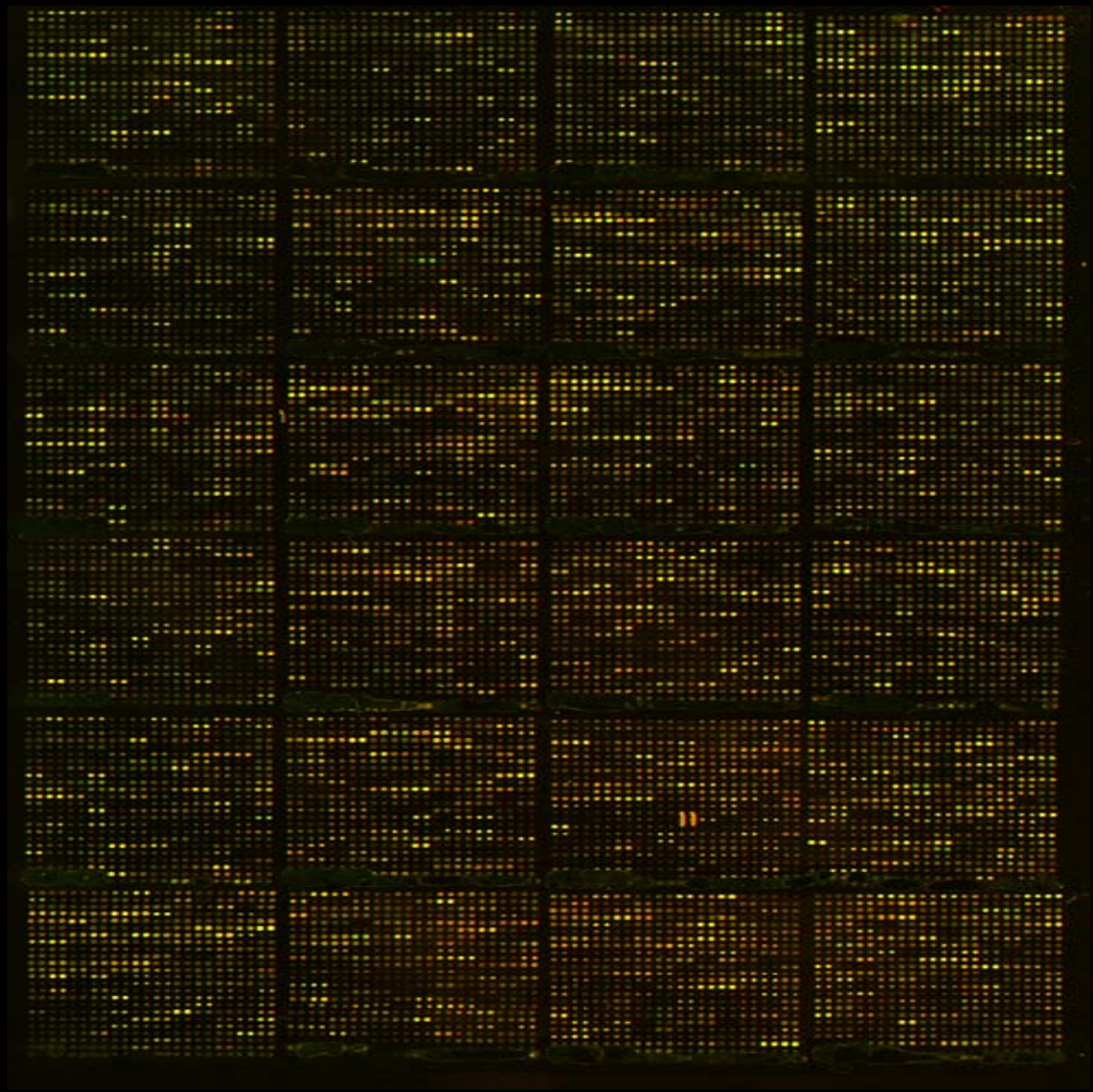
DNA makes RNA makes Protein

(reminder)



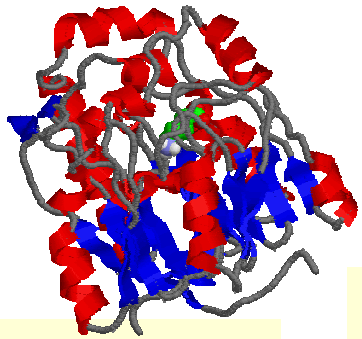
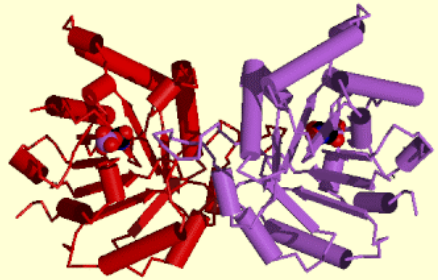
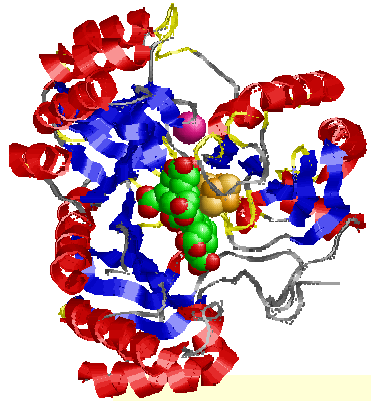
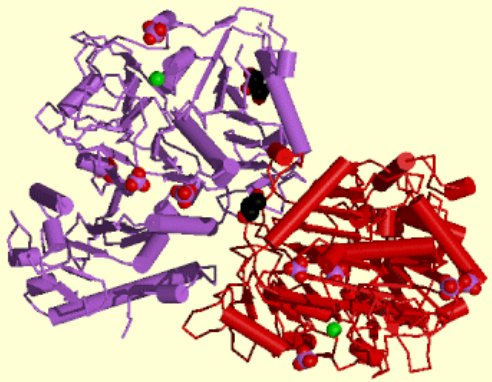
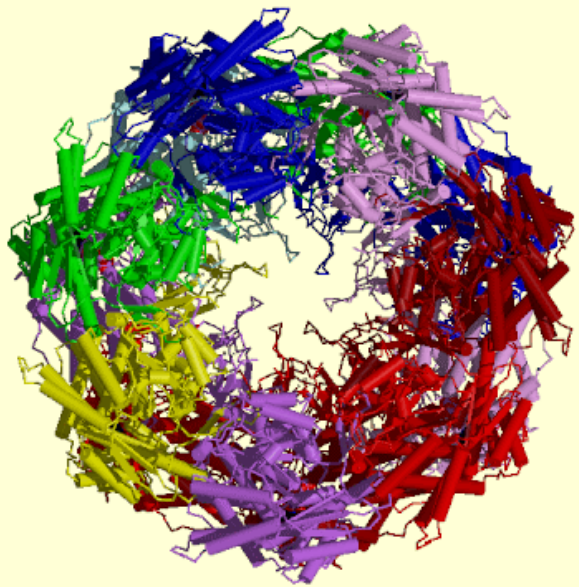
DNA makes RNA makes Protein: Expression data

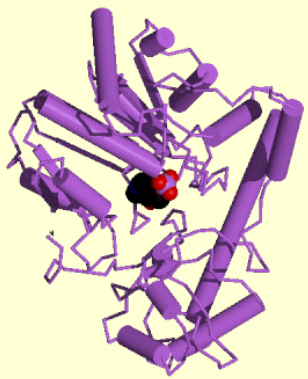
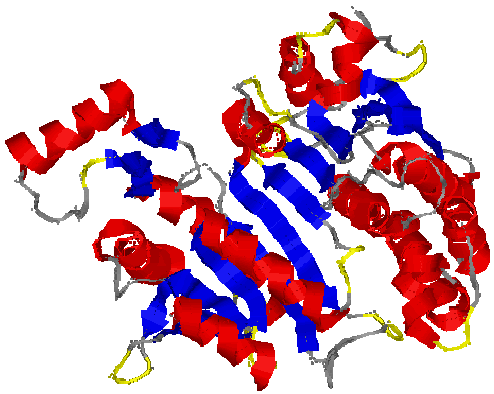
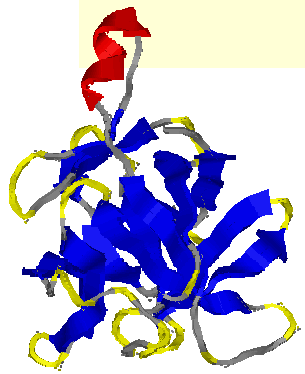
- More copies of mRNA for a gene leads to more protein
- mRNA can now be measured for all the genes in a cell at ones through microarray technology
- Can have 60,000 spots (genes) on a single gene chip
- Colour change gives intensity of gene expression (over- or under-expression)



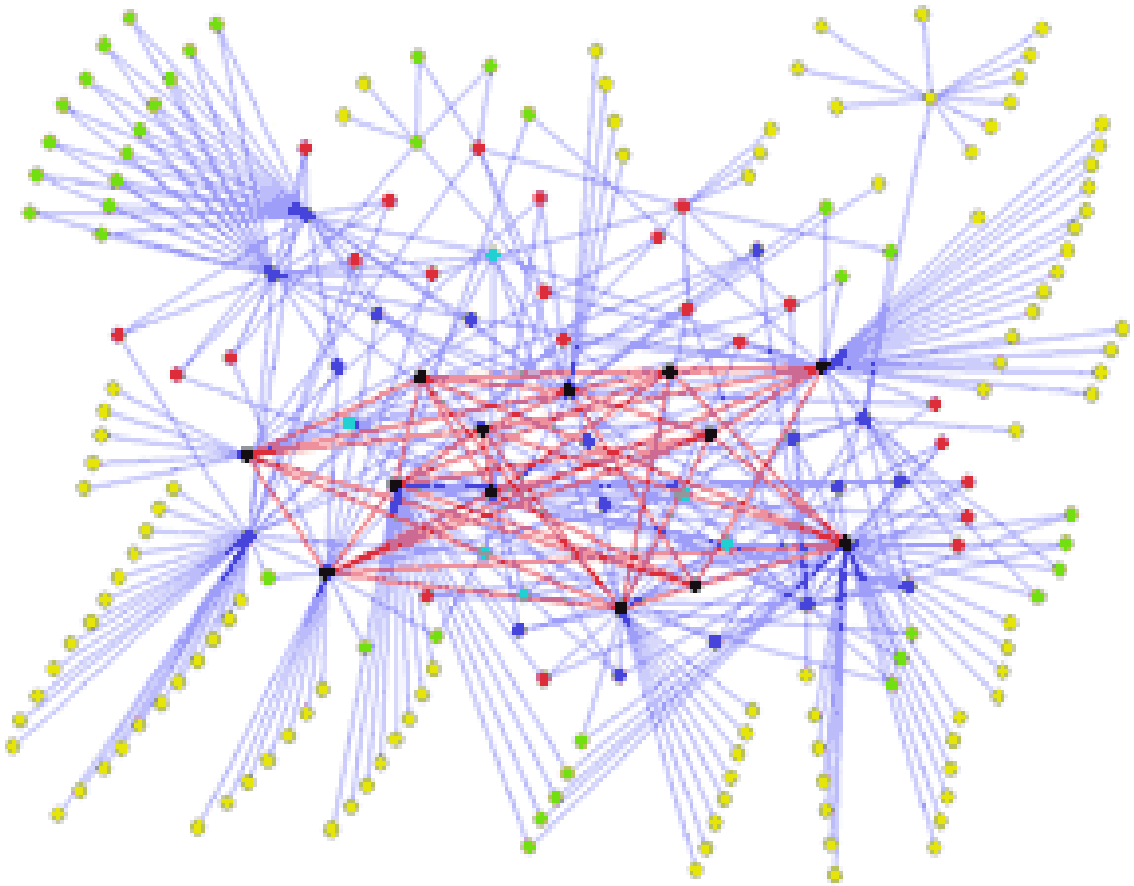
Proteomics

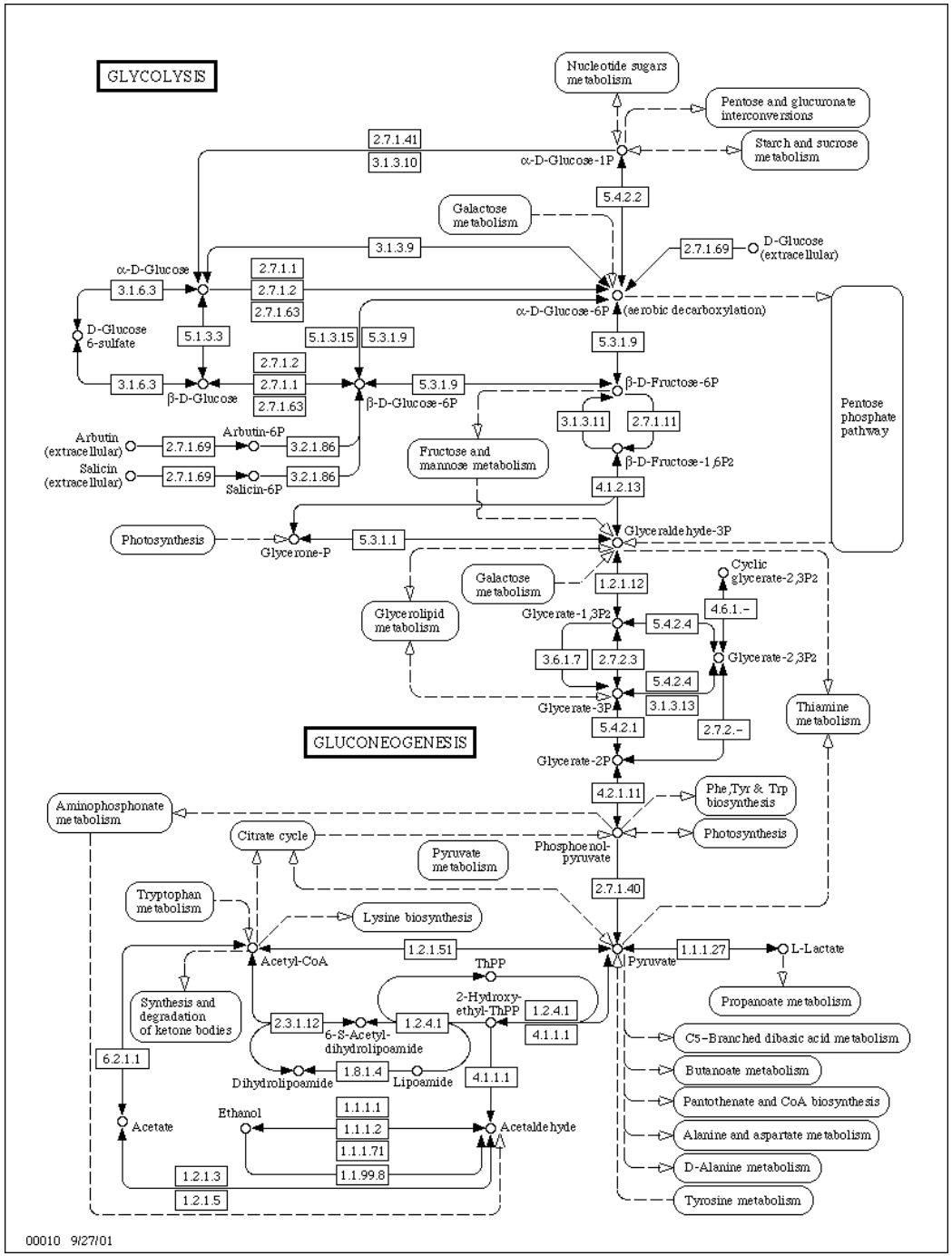
- Elucidating all 3D structures of proteins in the cell
- This is also called *Structural Genomics*
- Finding out what these proteins do
- This is also called *Functional Genomics*

Protein-protein interaction networks





Metabolic networks

Glycolysis and Gluconeogenesis

Kegg database (Japan)

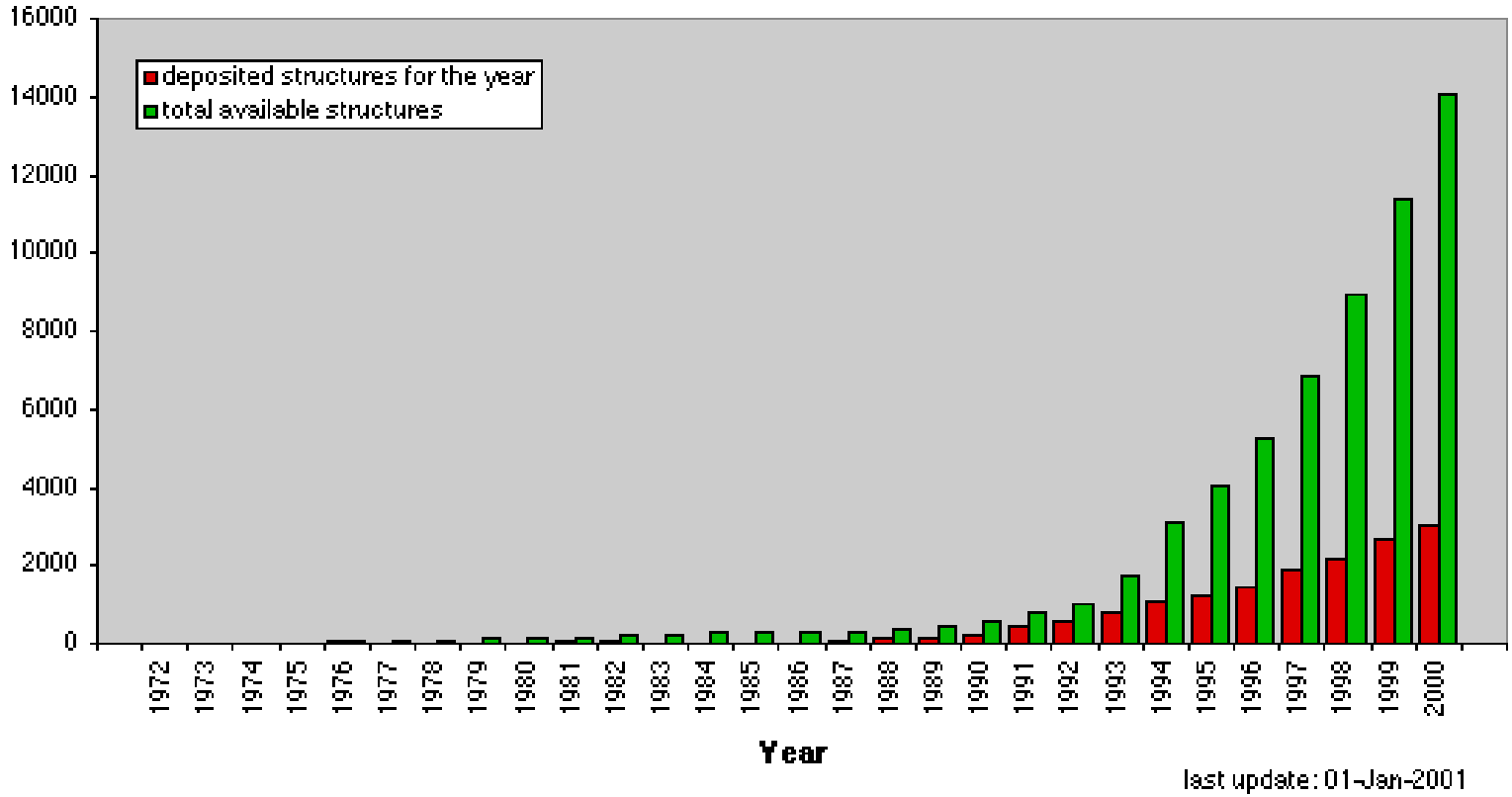
High-throughput Biological Data

- Enormous amounts of biological data are being generated by high-throughput capabilities; even more are coming
  - genomic sequences
  - gene expression data
  - mass spec. data
  - protein-protein interaction
  - protein structures
  - .....

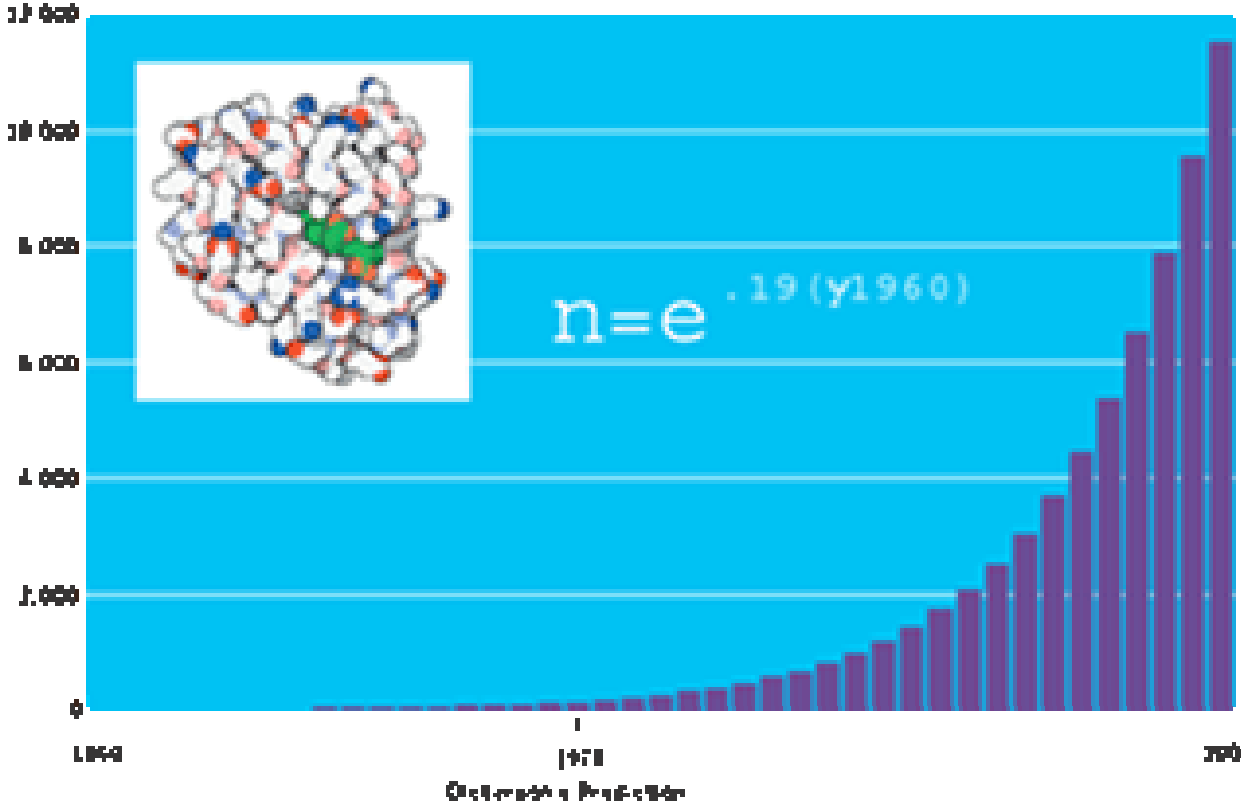
Protein structural data explosion

Protein Data Bank (PDB): 14500 Structures (6 March 2001)

10900 x-ray crystallography, 1810 NMR, 278 theoretical models, others...



Dickerson's formula: equivalent to Moore's law



$n = e^{0.19(y-1960)}$ with y the year.

On 27 March 2001 there were 12,123 3D protein structures in the PDB: Dickerson's formula predicts 12,066 (within 0.5%)!

Sequence versus structural data

- Structural genomics initiatives are now in full swing and growth is still exponential.

- However, growth of sequence data is even more rapidly. There are now more than 300 completely sequenced genomes publicly available.

Increasing gap between structural and sequence data (“Mind the gap”)

Bioinformatics

Large - external
(*integrative*)

Small – internal (*individual*)

Science

Planetary Science
Population Biology
Sociobiology
Systems Biology
Biology

Human

Cultural Anthropology
Sociology
Psychology

Medicine

Molecular Biology
Chemistry
Physics

Bioinformatics

- Offers an ever more essential input to
  - Molecular Biology
  - Pharmacology (drug design)
  - Agriculture
  - Biotechnology
  - Clinical medicine
  - Anthropology
  - Forensic science
  - Chemical industries (detergent industries, etc.)