Lecture 1
Introduction to Bioinformatics

Burr Settles
IBS Summer Research Program 2008
bsettles@cs.wisc.edu
www.cs.wisc.edu/~bsettles/ibs08/

About Me

- instructor: Burr Settles
  - 7th year graduate student in Computer Sciences
  - thesis topic: “Active Learning”
  - advisor: Dr. Mark Craven

- office: 6775 Medical Sciences Center (upstairs)
- email: **bsettles@cs.wisc.edu**
- course webpage: **http://www.cs.wisc.edu/~bsettles/ibs08/**

What About You?

- school, major, year?
  - what is your background in biology and/or computer science & statistics?
- what attracted you to the ISB program and the CBB track in particular?
- what do you think “bioinformatics” is?

"Bioinformatics"

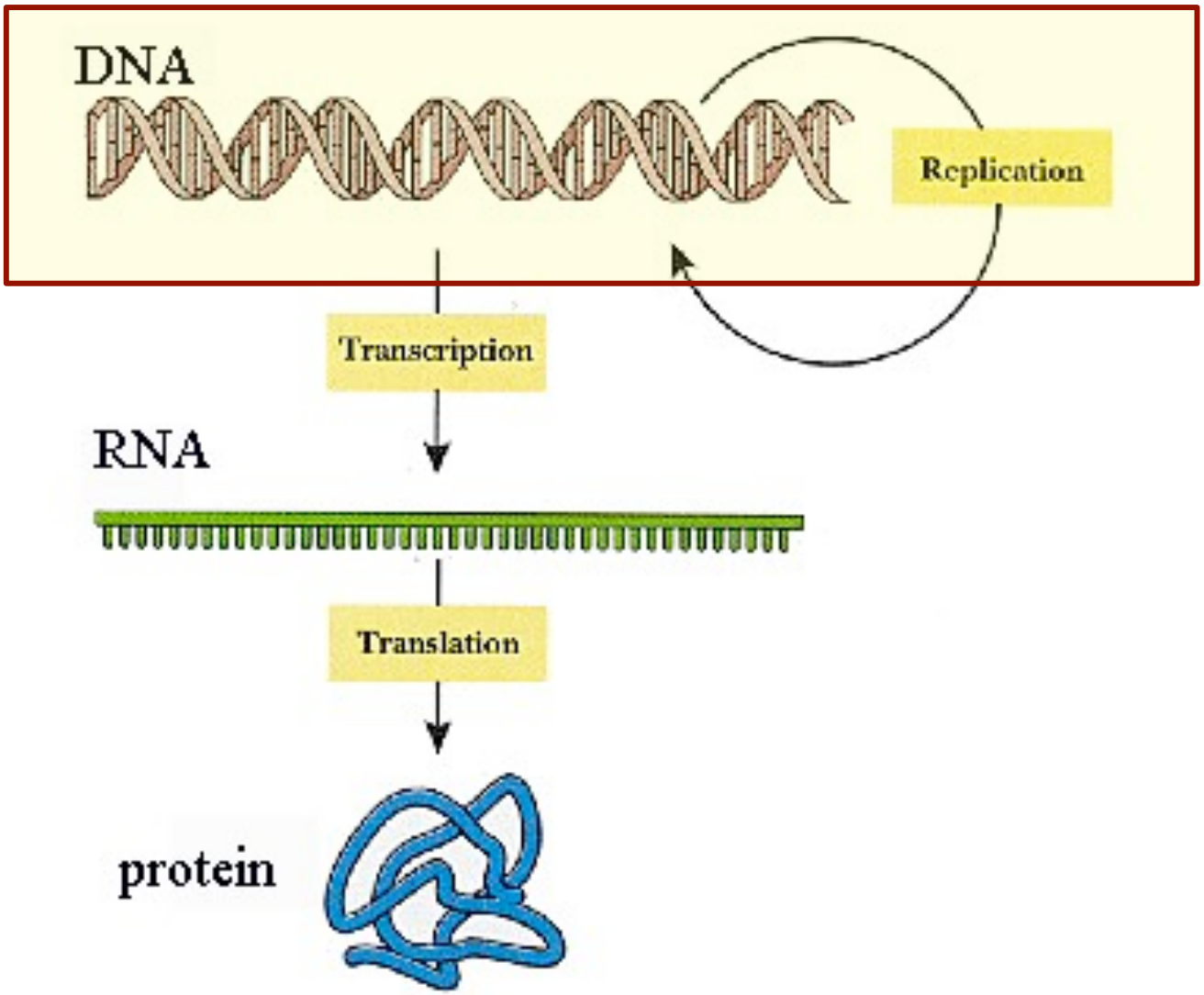
- general definition: computational techniques for solving biological problems
  - data problems: representation (graphics), storage and retrieval (databases), analysis (statistics, artificial intelligence, optimization, etc.)
  - biology problems: sequence analysis, structure or function prediction, data mining, etc.

- also called *computational biology*

Course Overview

- **basic molecular biology**
- sequence alignment
- probabilistic sequence models
- gene expression analysis
- protein structure prediction
  - by Ameet Soni

The Central Dogma



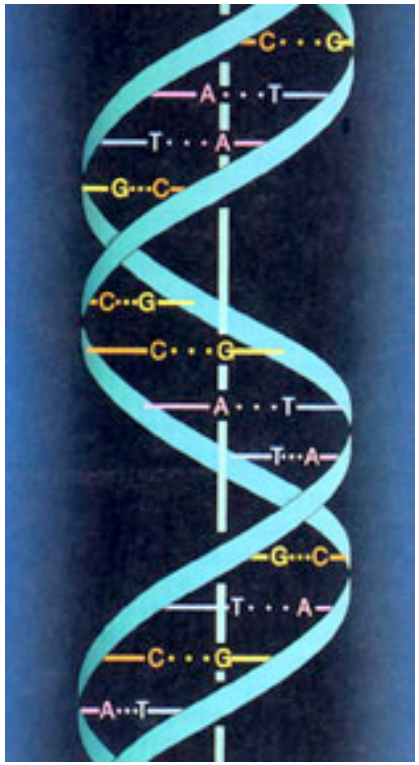
DNA

- can be thought of as the "recipe" for an organism
- composed of small molecules called *nucleotides*
  - four different nucleotides distinguished by the four *bases*: adenine (A), cytosine (C), guanine (G) and thymine (T)
- is a *polymer*: large molecule consisting of similar units (nucleotides in this case)
- a single strand of DNA can be thought of as a string composed of the four letters: A, C, G, T

  ctgctggaccgggtgctaggaccctgactgcccggg
  gccgggggtgcggggcccgctgag...

The Double Helix

- DNA molecules usually consist of two strands arranged in the famous double helix



Watson-Crick Base Pairs

- in double-stranded DNA
  - A always bonds to T
  - C always bonds to G

The Double Helix

- each strand of DNA has a "direction"
  - at one end, the terminal carbon atom in the backbone is the 5' carbon atom of the terminal sugar
  - at the other end, the terminal carbon atom is the 3' carbon atom of the terminal sugar

- therefore we can talk about the *5'* and the *3'* ends of a DNA strand

- in a double helix, the strands are *antiparallel* (arrows drawn from the 5' end to the 3' end go in opposite directions)

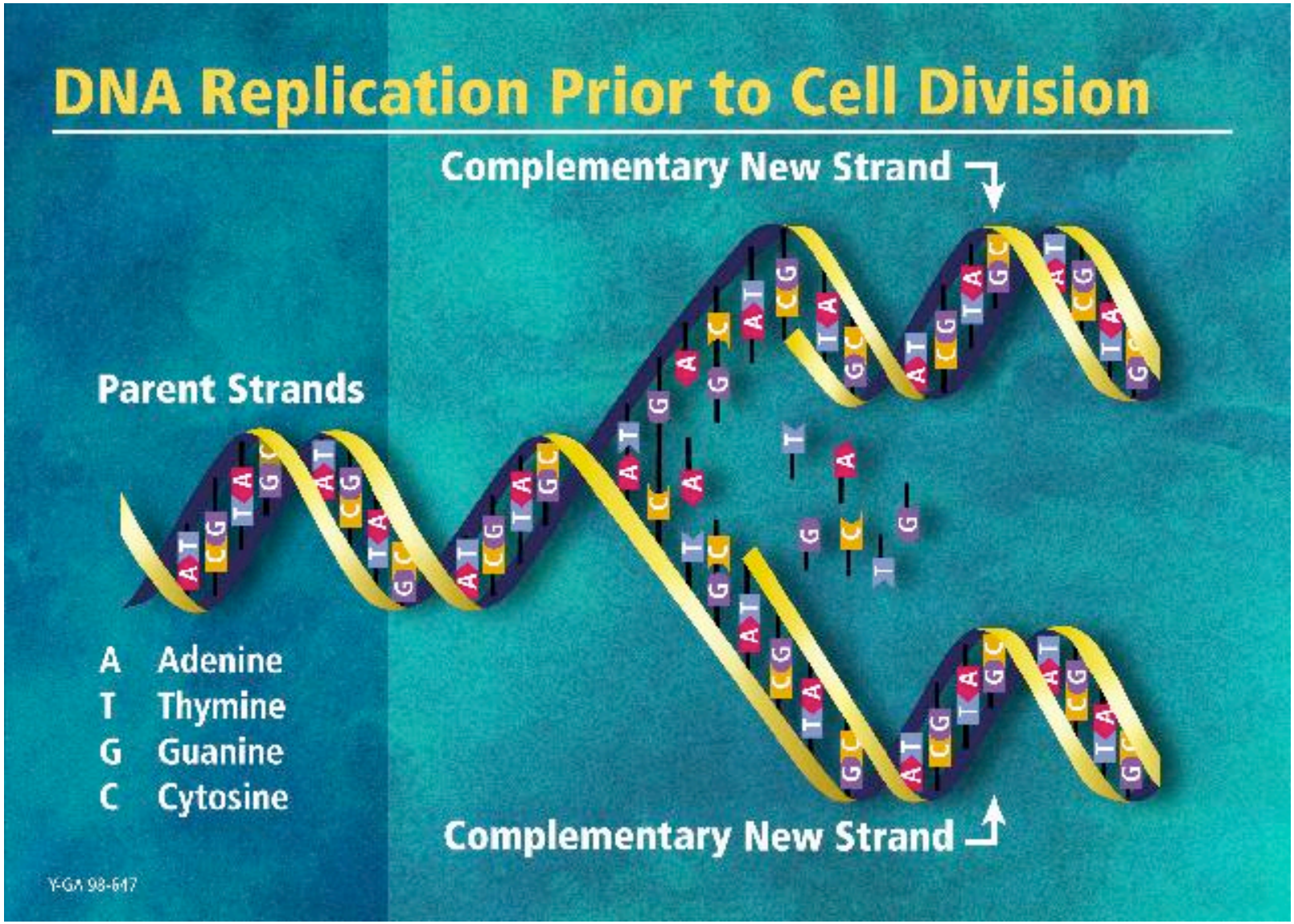
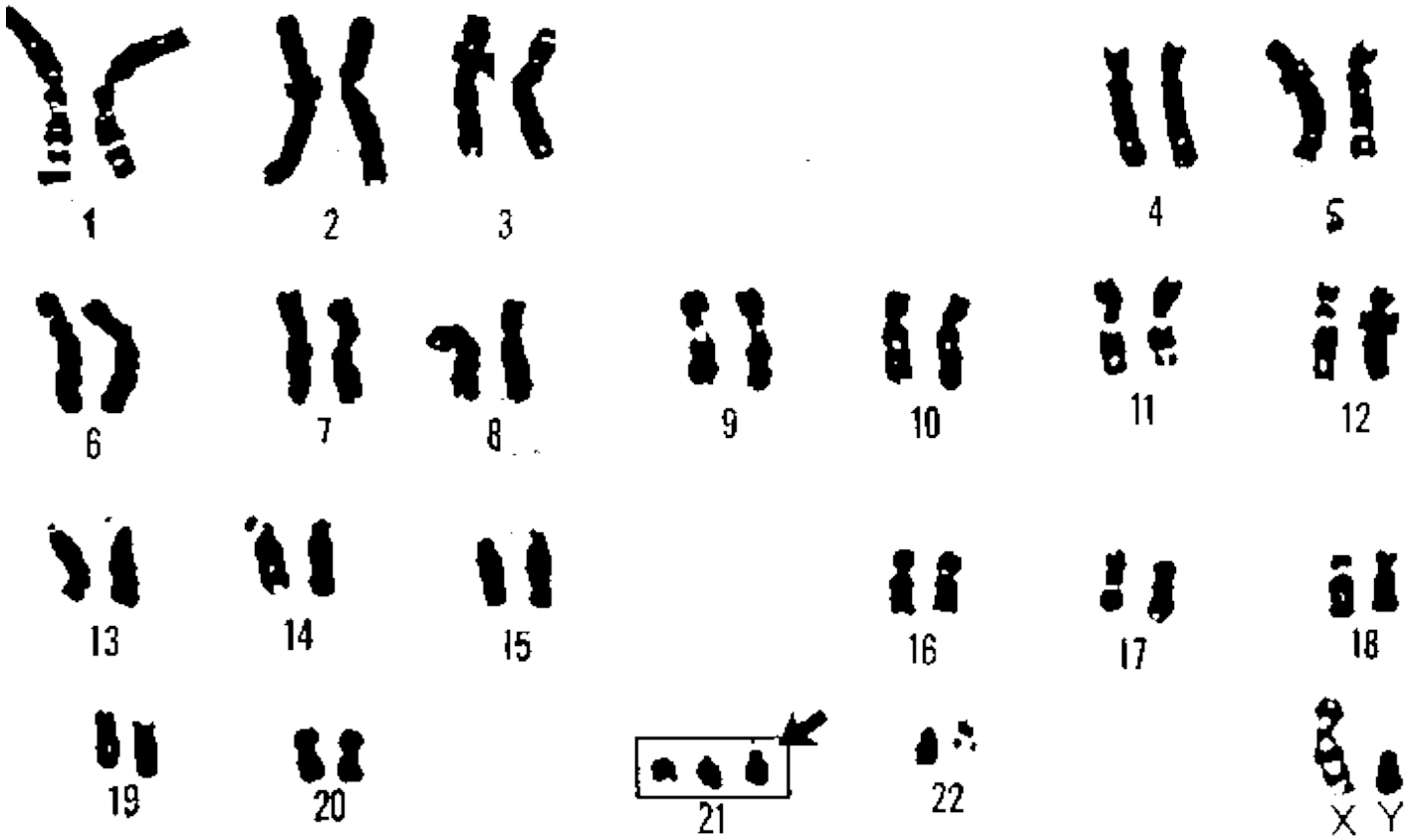


image from the DOE Human Genome Program
http://www.ornl.gov/hgmis

Chromosomes

- DNA is packaged into individual *chromosomes*
- *prokaryotes* (single-celled organisms lacking nuclei) typically have a single circular chromosome
  - examples: bacteria, archea
- *eukaryotes* (organisms with nuclei) have a species-specific number of linear chromosomes
  - examples: animals, plants, fungi

Human Chromosomes



Genomes

- the term *genome* refers to the complete complement of DNA for a given species

- the human genome consists of 23 pairs of chromosomes
  - mosquitos have 3 pairs
  - camels have 35 pairs!

- every cell (except sex cells and mature red blood cells) contains the complete genome of an organism

Genes

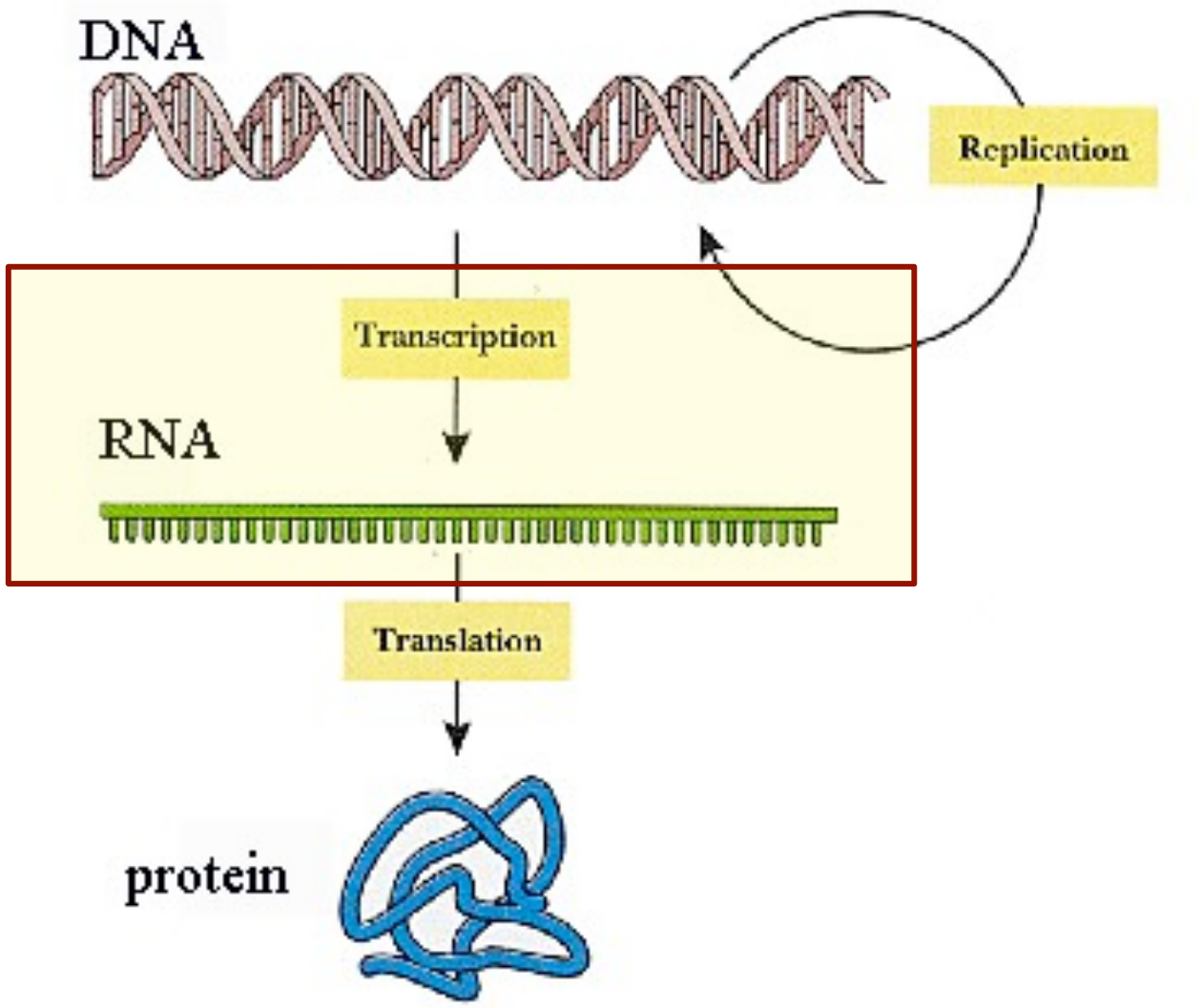
- genes are the basic units of heredity
- a gene is a sequence of bases that carries the information required for constructing a particular protein (more accurately, polypeptide)
- such a gene is said to *encode* a protein
- the human genome comprises ~ 25,000 protein-coding genes

Gene Density

- not all of the DNA in a genome encodes protein:

bacteria	~90% coding	gene/kb
human	~1.5% coding	gene/35kb

The Central Dogma

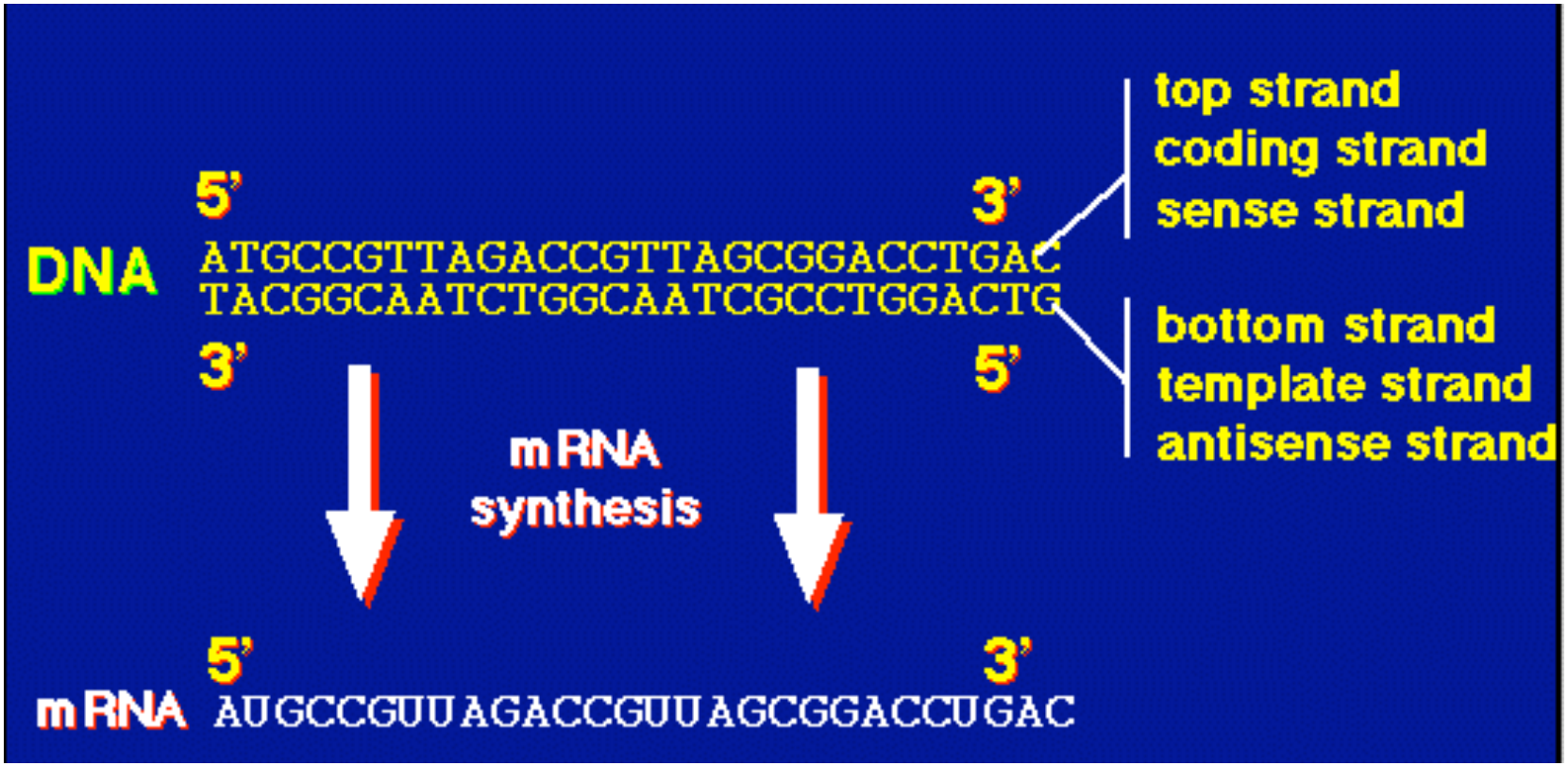


RNA

- RNA is like DNA except:
  - backbone is a little different
  - often single stranded
  - the base uracil (U) is used in place of thymine (T)

- a strand of RNA can be thought of as a string composed of the four letters: A, C, G, U

Transcription

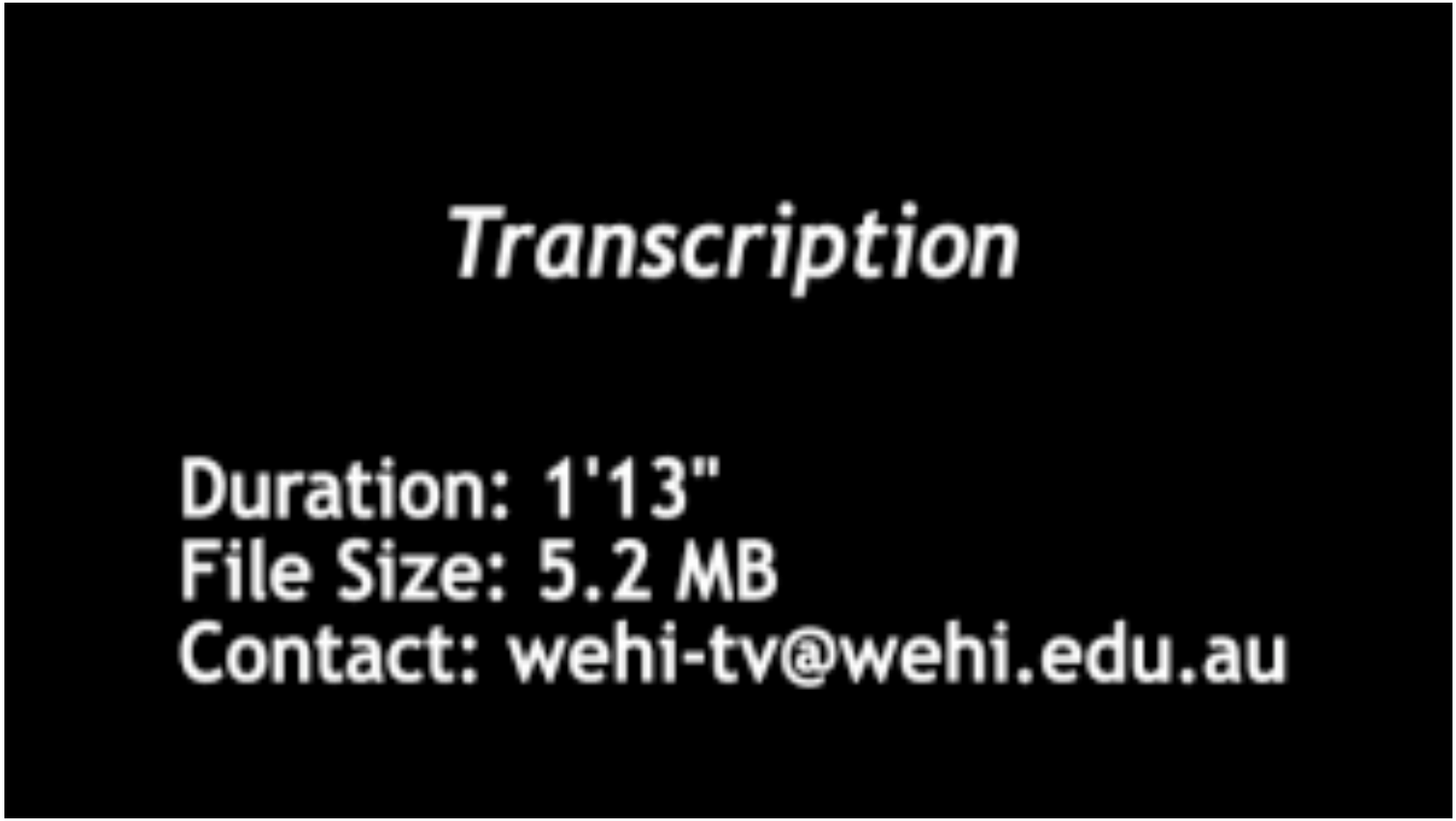


Transcription

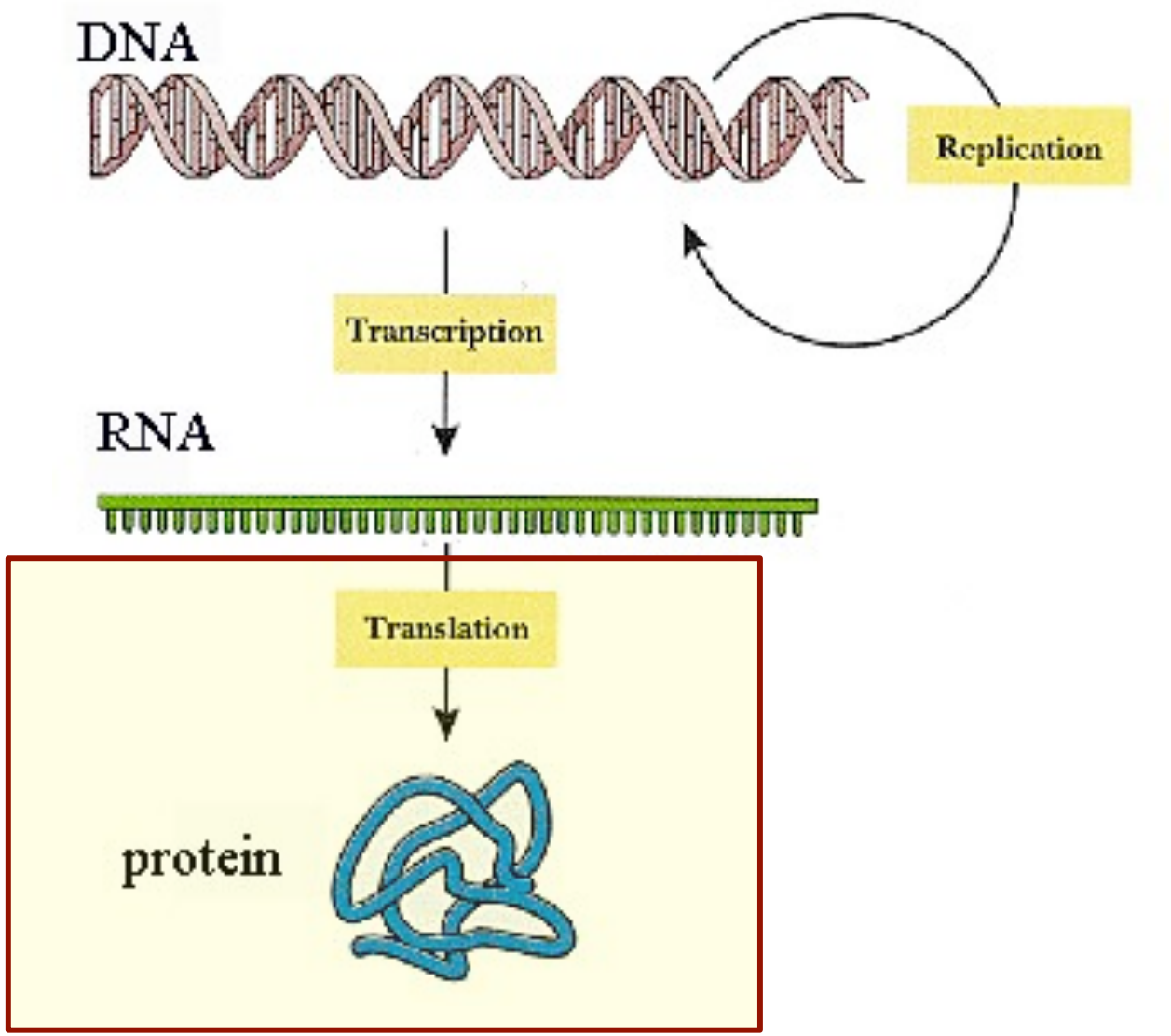
- *RNA polymerase* is the enzyme that builds an RNA strand from a gene

- RNA that is transcribed from a gene is called *messenger RNA* (mRNA)

Transcription Movie



The Central Dogma



Proteins

- proteins are molecules composed of one or more *polypeptides*
- a polypeptide is a polymer composed of *amino acids*
- cells build their proteins from 20 different amino acids
- a polypeptide can be thought of as a string composed from a 20-character alphabet

Protein Functions

- structural support
- storage of amino acids
- transport of other substances
- coordination of an organism's activities
- response of cell to chemical stimuli
- movement
- protection against disease
- selective acceleration of chemical reactions

Amino Acids

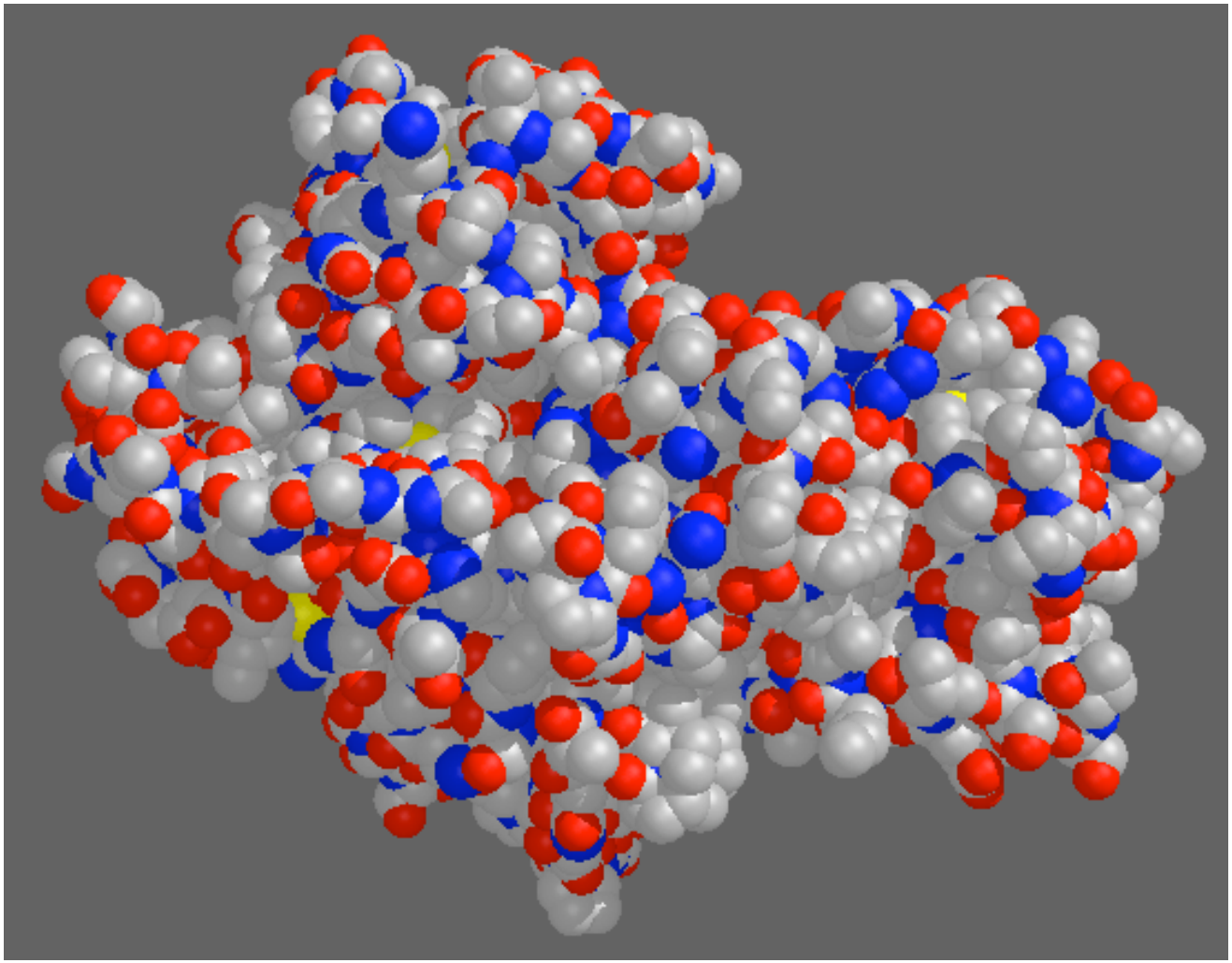
Alanine	**Ala**	**A**
Arginine	**Arg**	**R**
Aspartic Acid	**Asp**	**D**
Asparagine	**Asn**	**N**
Cysteine	**Cys**	**C**
Glutamic Acid	**Glu**	**E**
Glutamine	**Gln**	**Q**
Glycine	**Gly**	**G**
Histidine	**His**	**H**
Isoleucine	**Ile**	**I**
Leucine	**Leu**	**L**
Lysine	**Lys**	**K**
Methionine	**Met**	**M**
Phenylalanine	**Phe**	**F**
Proline	**Pro**	**P**
Serine	**Ser**	**S**
Threonine	**Thr**	**T**
Tryptophan	**Trp**	**W**
Tyrosine	**Tyr**	**Y**
Valine	**Val**	**V**

Amino Acid Sequence: Hexokinase

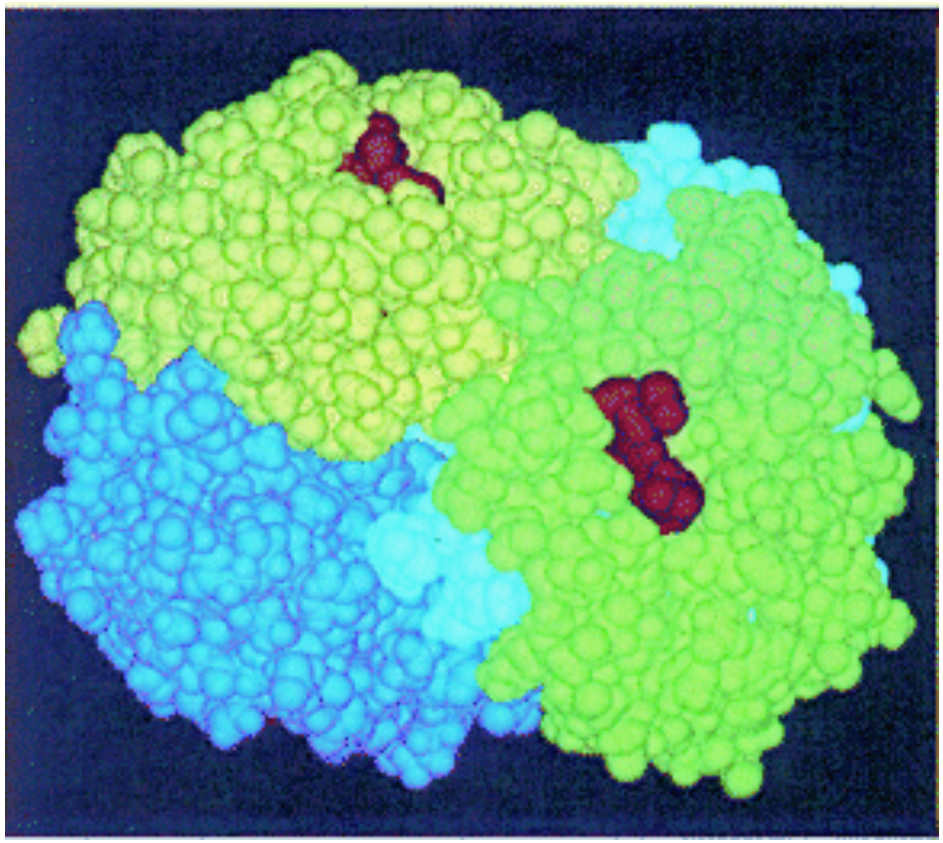
```
          5        10        15        20        25        30
  1 A A S X D X S L V E V H X X V F I V P P X I L Q A V V S I A
 31 T T R X D D X D S A A A S I P M V P G W V L K Q V X G S Q A
 61 G S F L A I V M G G G D L E V I L I X L A G Y Q E S S I X A
 91 S R S L A A S M X T T A I P S D L W G N X A X S N A A F S S
121 X E F S S X A G S V P L G F T F X E A G A K E X V I K G Q I
151 T X Q A X A F S L A X L X K L I S A M X N A X F P A G D X X
181 X X V A D I X D S H G I L X X V N Y T D A X I K M G I I F G
211 S G V N A A Y W C D S T X I A D A A D A G X X G G A G X M X
241 V C C X Q D S F R K A F P S L P Q I X Y X X T L N X X S P X
271 A X K T F E K N S X A K N X G Q S L R D V L M X Y K X X G Q
301 X H X X X A X D F X A A N V E N S S Y P A K I Q K L P H F D
331 L R X X X D L F X G D Q G I A X K T X M K X V V R R X L F L
361 I A A Y A F R L V V C X I X A I C Q K K G Y S S G H I A A X
391 G S X R D Y S G F S X N S A T X N X N I Y G W P Q S A X X S
421 K P I X I T P A I D G E G A A X X V I X S I A S S Q X X X A
451 X X S A X X A
```

- enzyme involved in glycolysis
- in every organism known from bacteria to humans

Space-Filling Model of Hexokinase



Hemoglobin

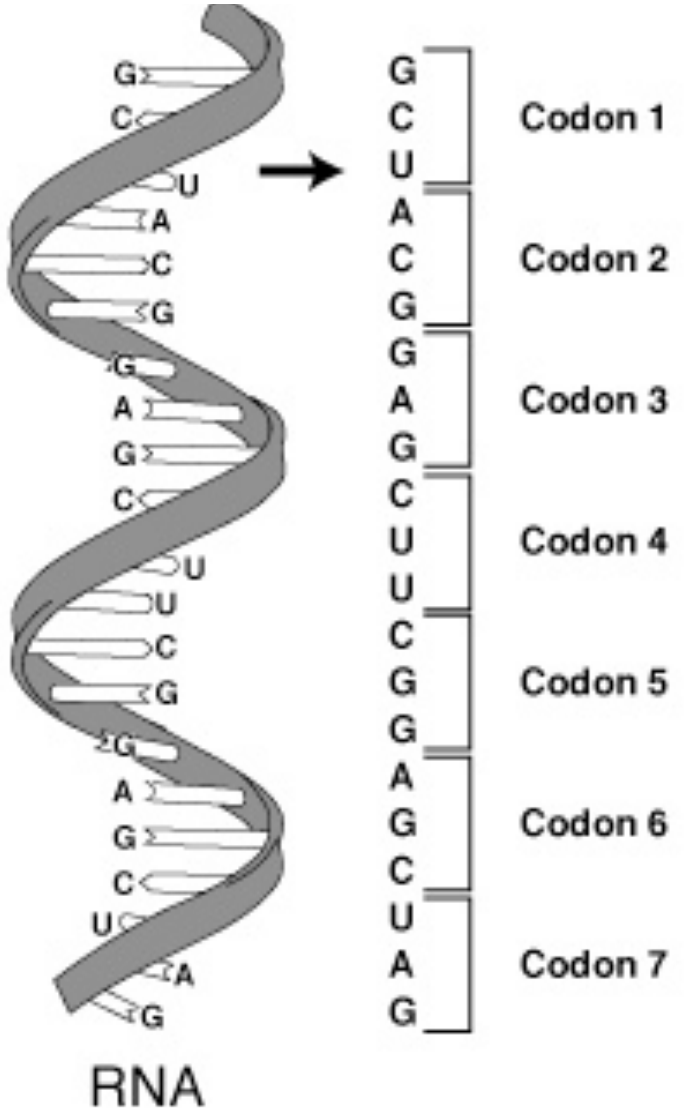


- protein built from 4 polypeptides
- responsible for carrying oxygen in red blood cells

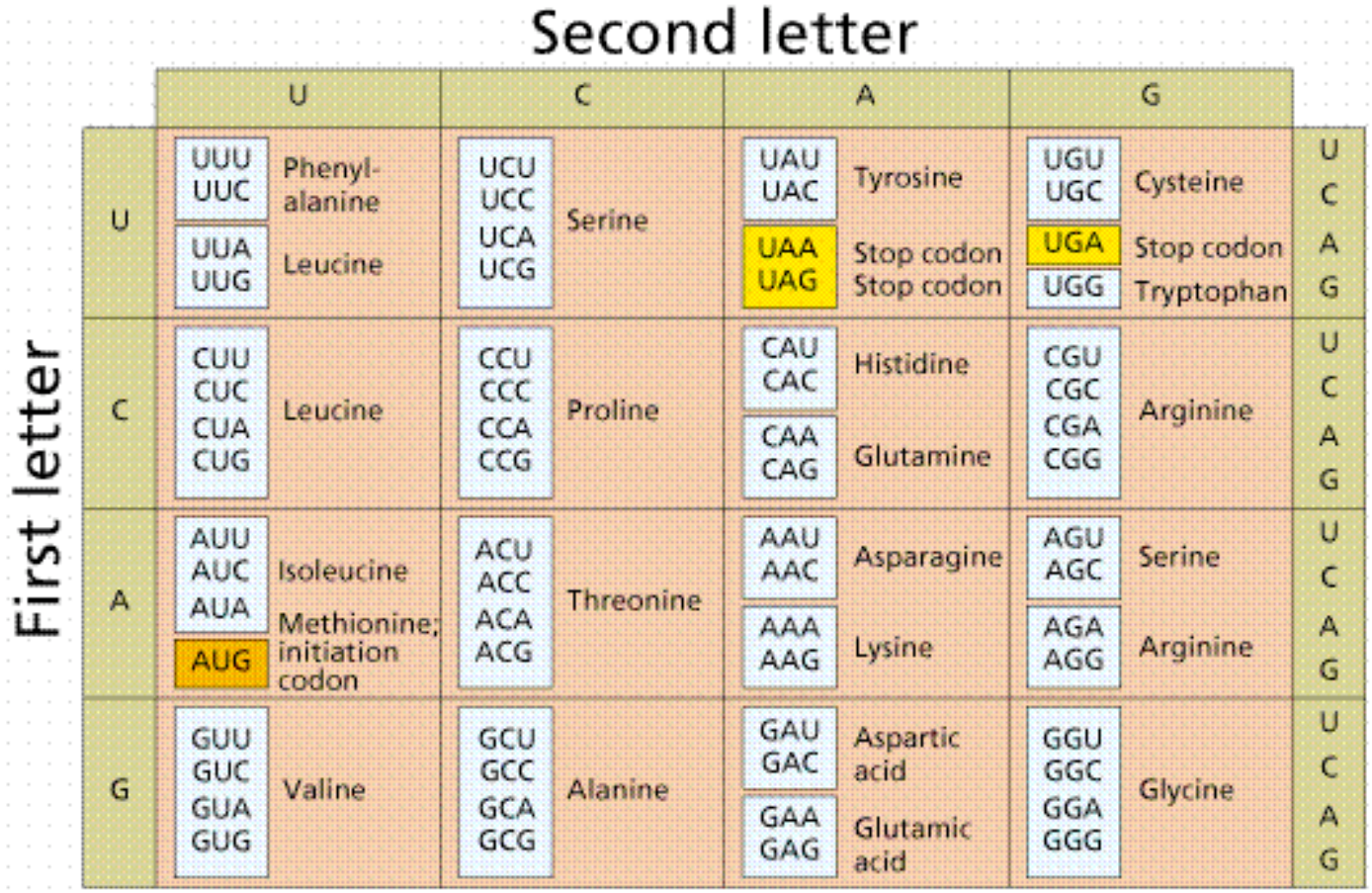
Translation

- *ribosomes* are the machines that synthesize proteins from mRNA
- the grouping of codons is called the *reading frame*
- translation begins with the *start codon*
- translation ends with the *stop codon*

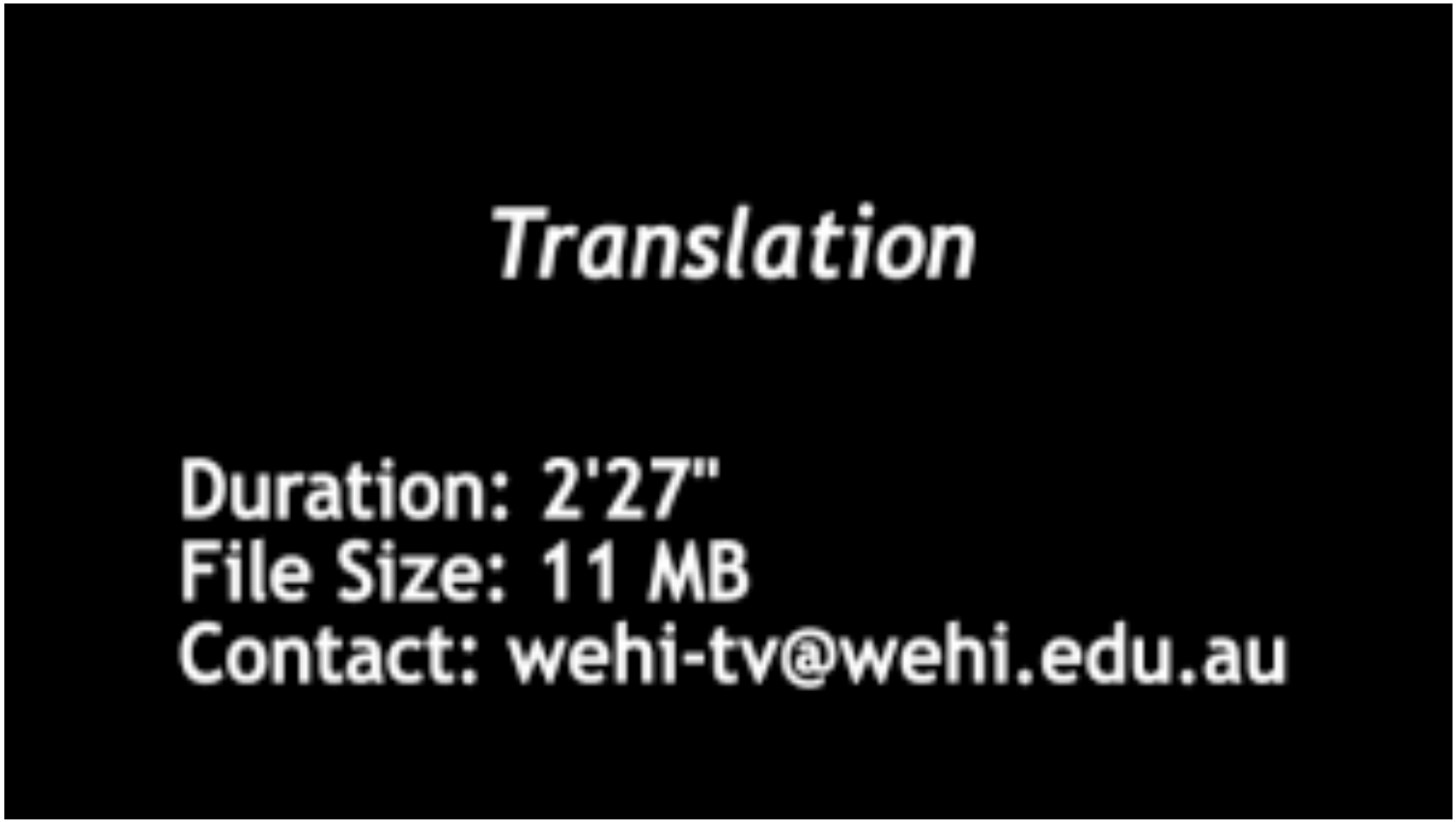
Codons and Reading Frames



The Genetic Code



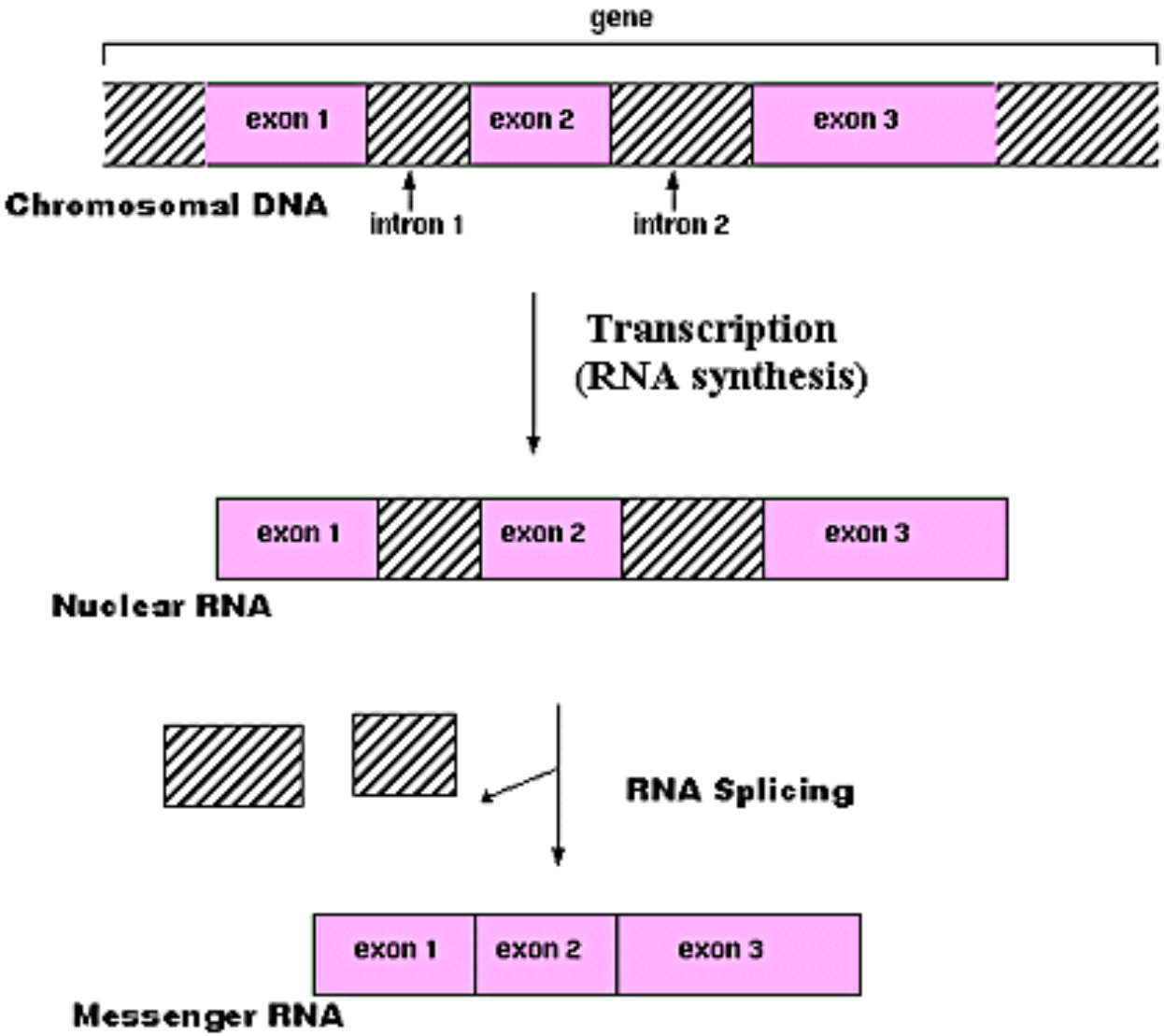
Translation Movie



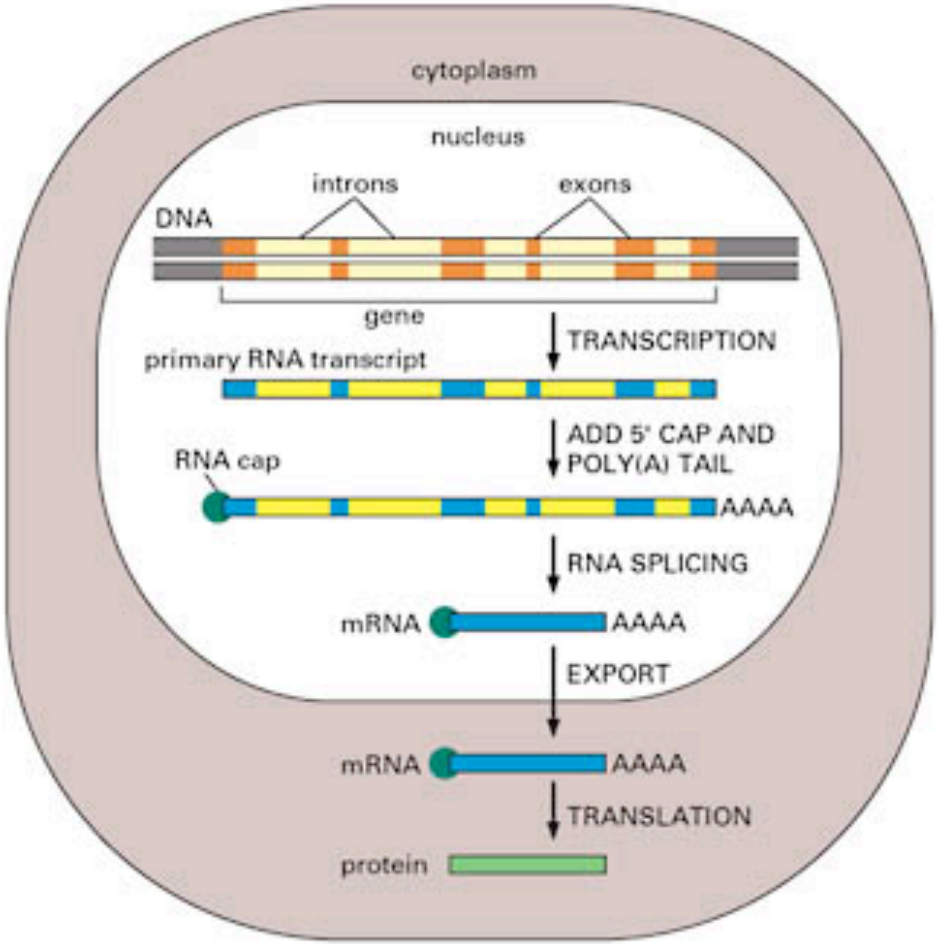
RNA Processing in Eukaryotes

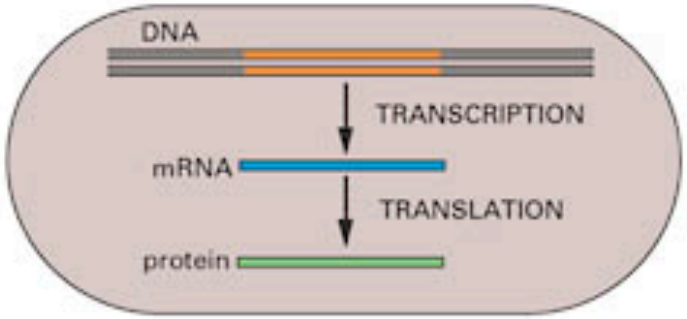
- *eukaryotes* (animals, plants, fungi, etc.) are organisms that have enclosed nuclei in their cells
- in many eukaryotes, genes/mRNAs consist of alternating *exon/intron* segments
- *exons* are the coding parts
- *introns* are spliced out before translation

RNA Splicing



Protein Synthesis in Eukaryotes vs. Prokaryotes





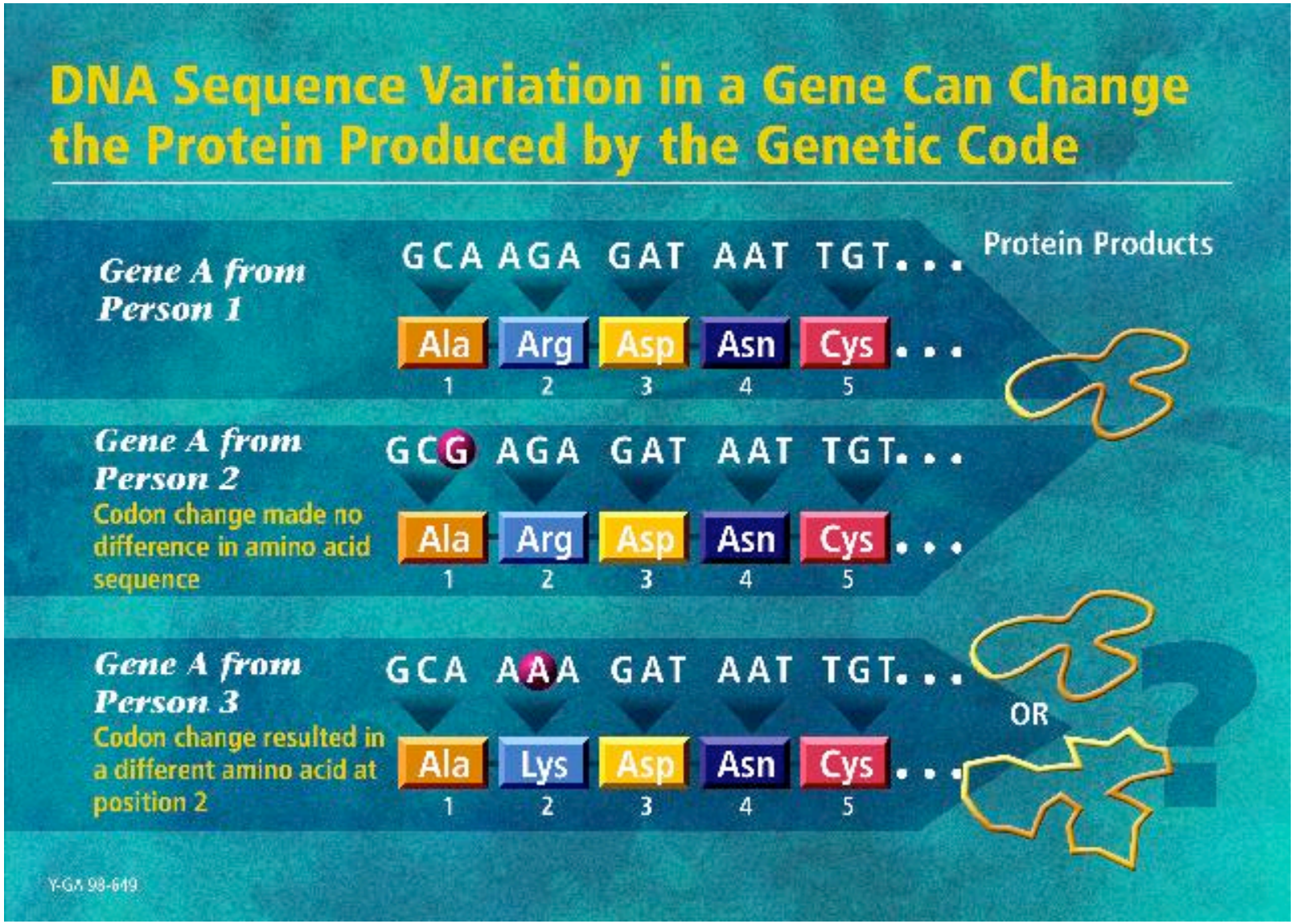


image from the DOE Human Genome Program
http://www.ornl.gov/hgmis

RNA Genes

- not all genes encode proteins
- for some genes the end product is RNA
  - *ribosomal RNA* (rRNA), which includes major constituents of ribosomes
  - *transfer RNAs* (tRNAs), which carry amino acids to ribosomes
  - *micro RNAs* (miRNAs), which play an important regulatory role in various plants and animals
  - etc.

The Dynamics of Cells

- all cells in an organism have the same genomic data, but the genes expressed in each vary according to cell type, time, and environmental factors
- there are networks of interactions among various biochemical entities in a cell (DNA, RNA, protein, small molecules) that carry out processes such as
  - metabolism
  - intra-cellular and inter-cellular signaling
  - regulation of gene expression

Overview of the E. coli Metabolic Pathway Map

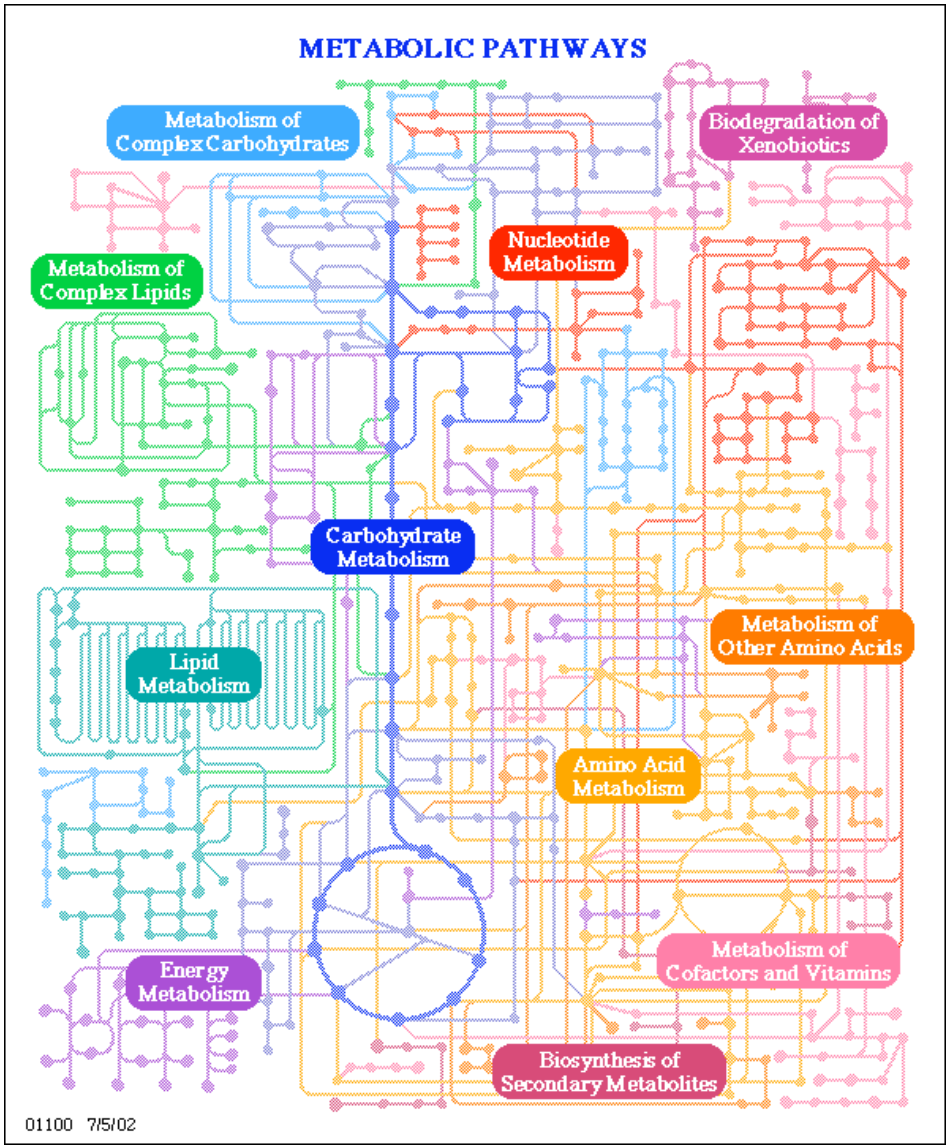
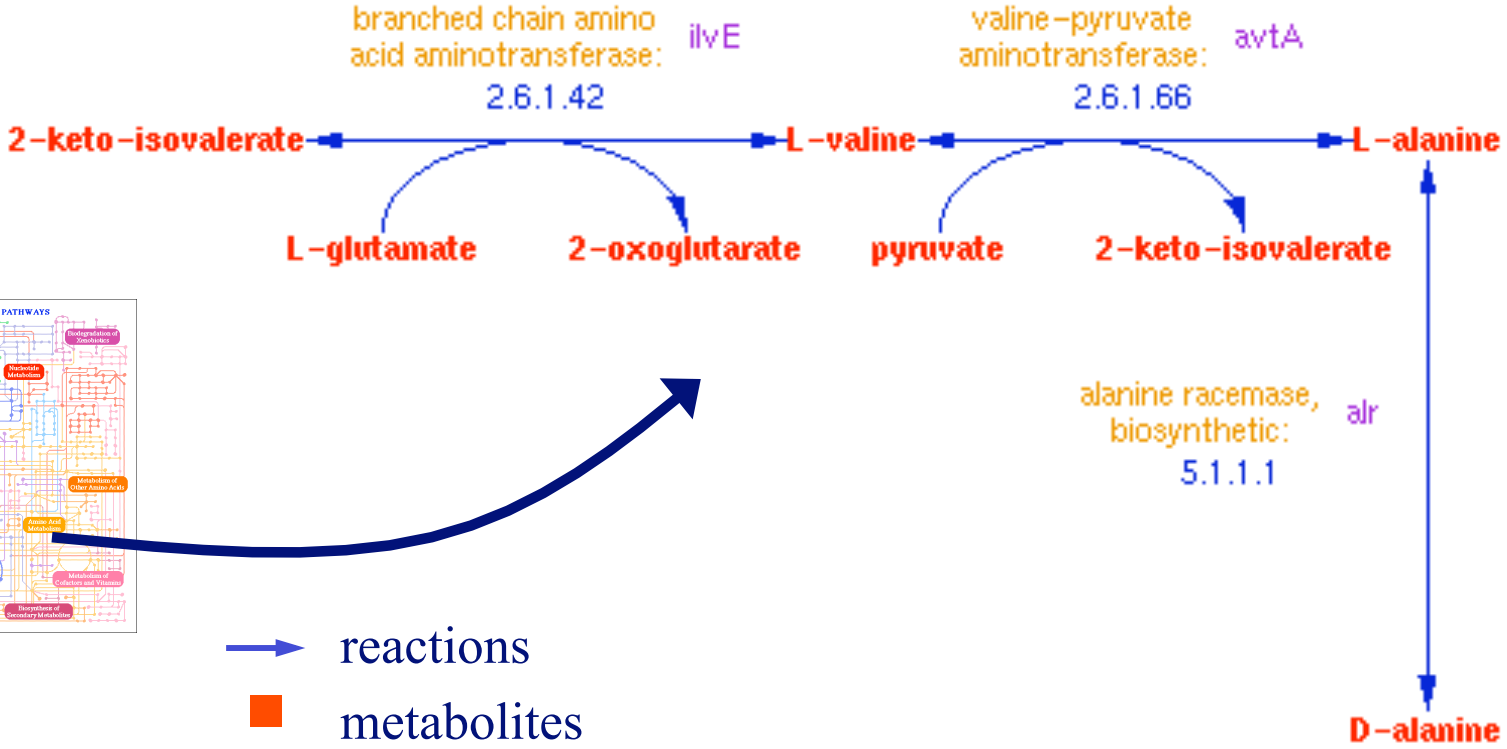
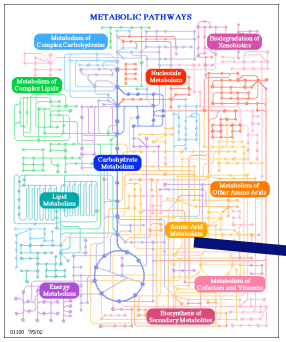


image from the KEGG database

The Metabolic Pathway for Synthesizing the Amino Acid Alanine

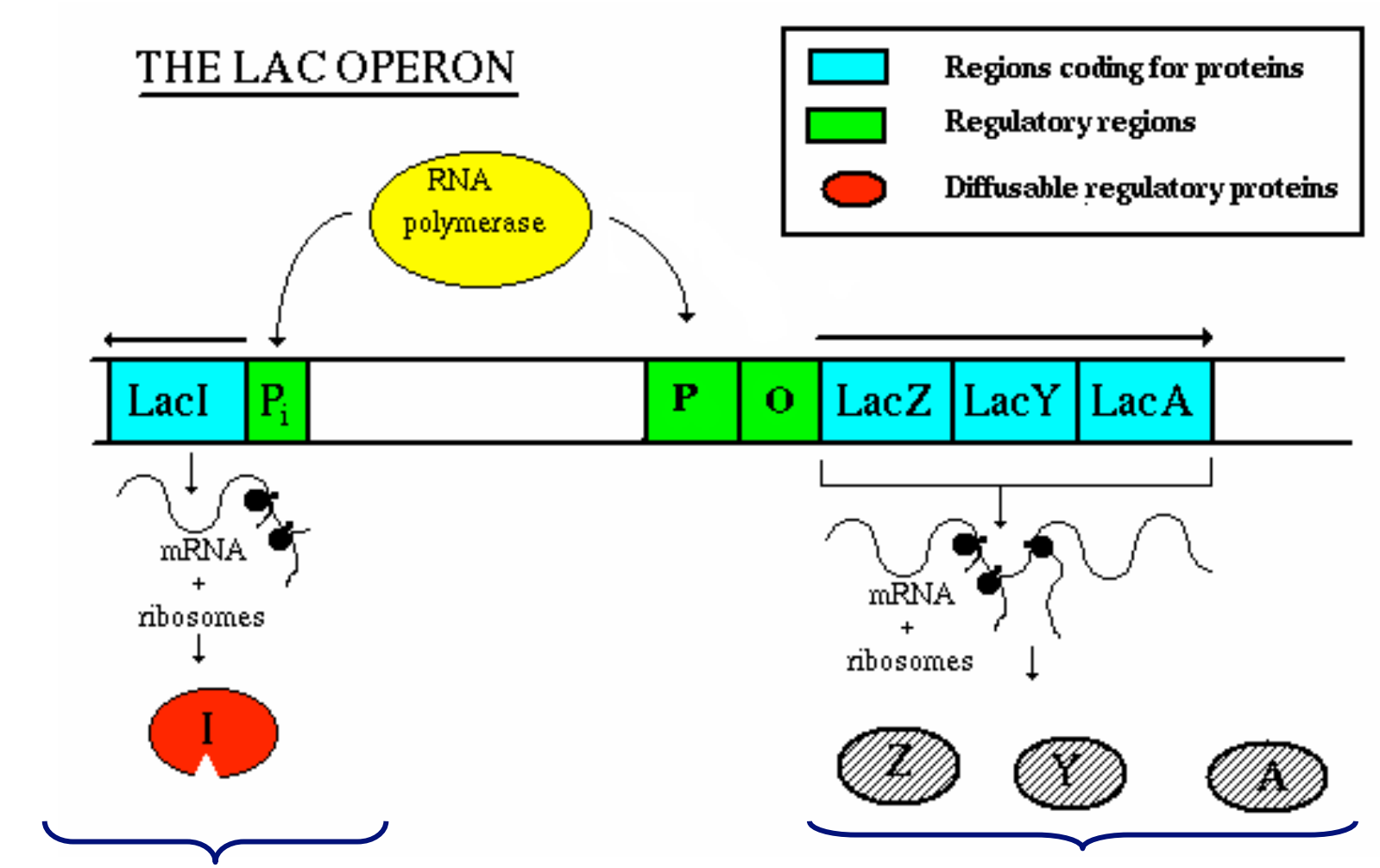




- reactions
- metabolites
- enzymes (proteins that catalyze reactions)
- genes encoding the enzymes

image from the Ecocyc database
www.biocyc.org

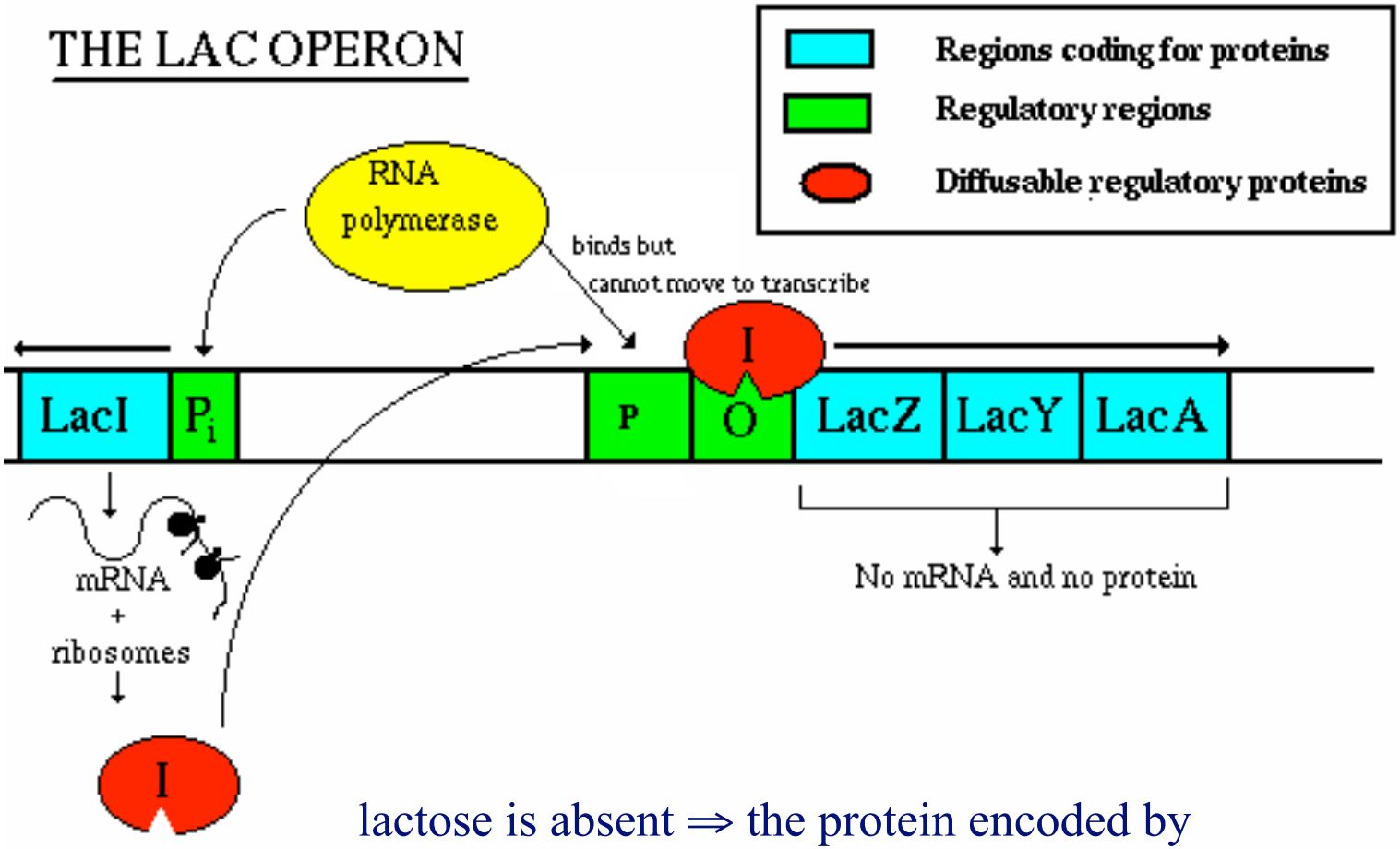
Gene Regulation Example: the lac Operon



this protein regulates the transcription of LacZ, LacY, LacA

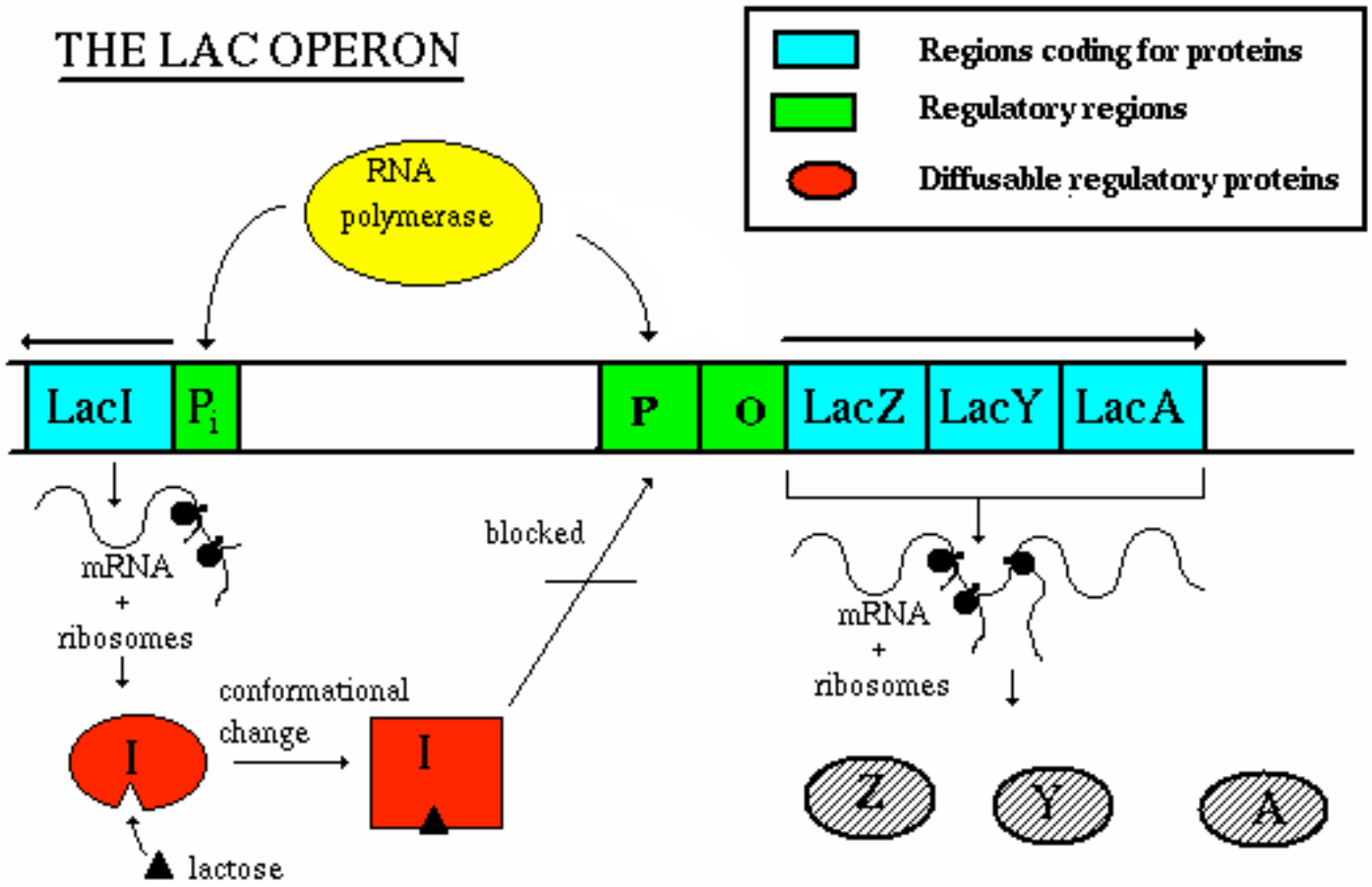
these proteins metabolize lactose

Gene Regulation Example: the lac Operon



lactose is absent ⇒ the protein encoded by lacI represses transcription of the lac operon

Gene Regulation Example: the lac Operon



lactose is present ⇒ it binds to the protein encoded by lacI changing its shape; in this state, the protein doesn't bind upstream from the lac operon; therefore the lac operon can be transcribed

Gene Regulation Example: the lac Operon

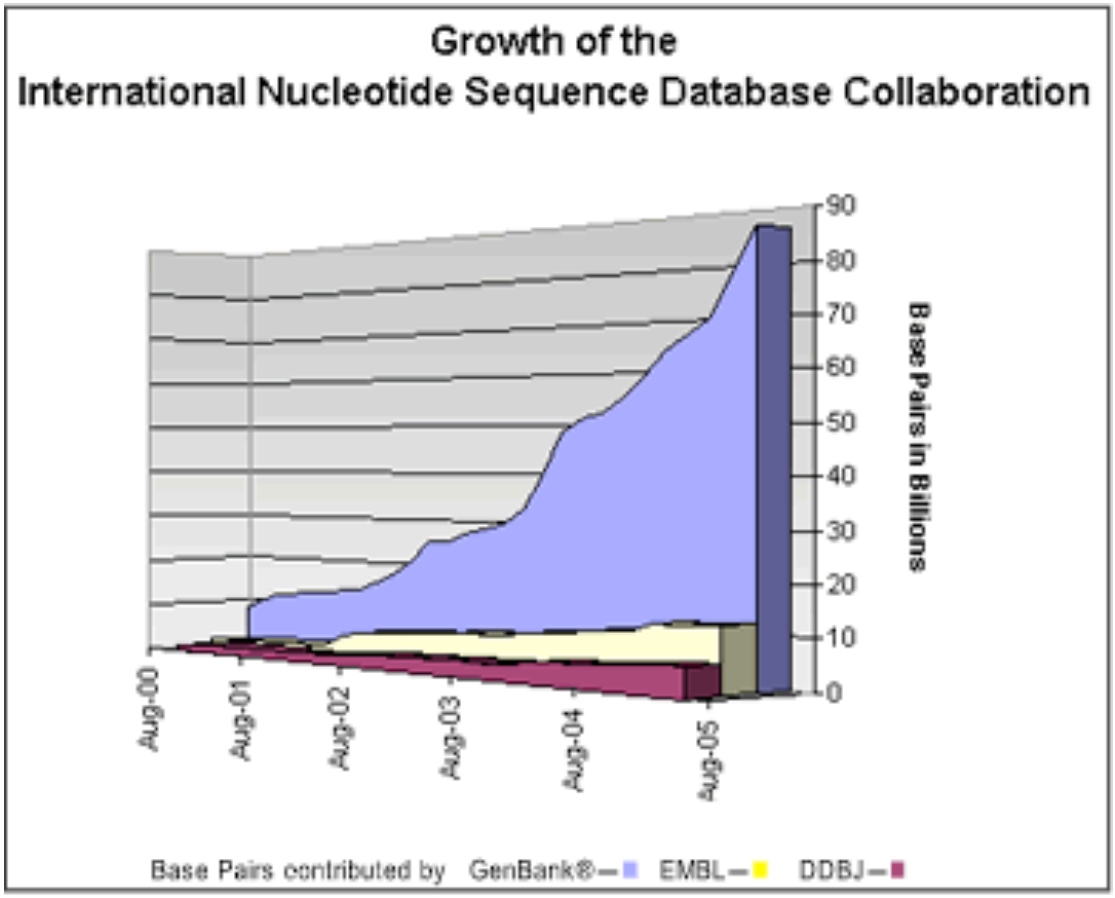
- this example provides a simple illustration of how a cell can regulate (turn on/off) certain genes in response to the state of its environment
  - an *operon* is a sequence of genes transcribed as a unit
  - the lac operon is involved in metabolizing lactose
    - it is "turned on" when lactose is present in the cell
    - the lac operon is regulated at the transcription level
- the depiction here is incomplete; for example, the level of glucose (not just lactose) in the cell influences transcription of the lac operon as well

Completed Genomes

Type	Approx # Completed
Archaea	46
Bacteria	524
Eukaryota	65
metagenomes	108
Organelles, Phages, Plasmids, Viroids, Viruses	too many to keep track

* Genomes OnLine Database (9/07)

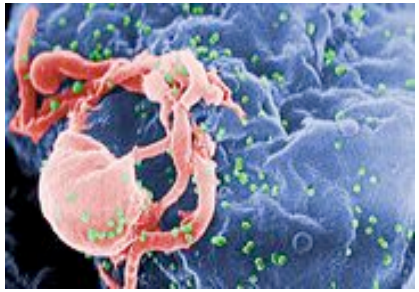
Completed Base Pairs



Some Greatest Hits

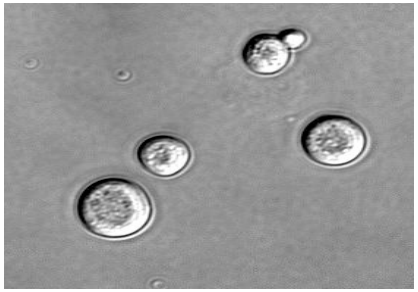
Genome	Where	Year
H. Influenza	TIGR	1995
E. Coli K -12	Wisconsin	1997
S. cerevisiae (yeast)	internat. collab.	1997
C. elegans (worm)	Washington U./Sanger	1998
Drosophila M. (fruit fly)	multiple groups	2000
E. Coli 0157:H7 (pathogen)	Wisconsin	2000
H. Sapiens (that's us)	internat. collab./Celera	2001
Mus musculus (mouse)	internat. collaboration	2002
Rattus norvegicus (rat)	internat. Collaboration	2004

Some Genome Sizes



HIV: 9.8k

E. coli: 4.6m



S. cerevisiae: 12m

D. melanogaster: 137m

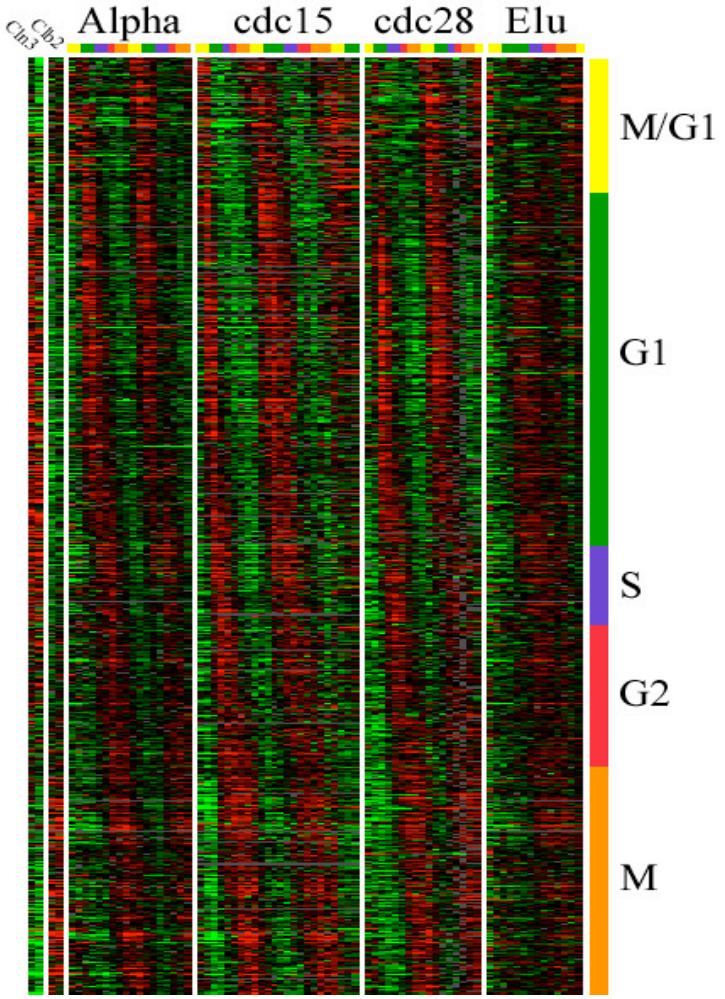
H. sapiens, R. norvegicus: 3.1b

More Than Just Genomes…

- > 300 other publicly available databases pertaining to molecular biology (see pointer to *Nucleic Acids Research* directory on course home page)
- GenBank
  - > 61 million sequence entries
  - > 65 billion bases
- UnitProtKB / Swis-Prot
  - > 277 thousand protein sequence entries
  - > 100 million amino acids
- Protein Data Bank
  - 45,632 protein (and related) structures

* all numbers current about 9/07

Even More Data: High-Throughput Experiments



- this figure depicts one yeast gene -expression data set
  - each row represents a gene
  - each column represents a measurement of gene expression (mRNA levels) at some time point
  - red indicates that a gene is being expressed more than usual; green means less

Figure from Spellman et al., *Molecular Biology of the Cell,* 9:3273-3297, 1998

Significance of the Genomics Revolution

- data driven biology
  - functional genomics
  - comparative genomics
  - systems biology
- molecular medicine
  - identification of genetic components of various maladies
  - diagnosis/prognosis from sequence/expression
  - gene therapy
- pharmacogenomics
  - developing highly targeted drugs
- toxicogenomics
  - elucidating which genes are affected by various chemicals

Bioinformatics Revisited

representation/storage/retrieval/analysis of biological data concerning:

- sequences (DNA, protein, RNA)
- structures (protein, RNA)
- functions (protein, sequence *signals*)
- activity levels (mRNA, protein, metabolites)
- networks of interactions (metabolic pathways, regulatory pathways, signaling pathways) of/among biomolecules
- even textual information from biomedical *literature*!

Next Time…

- basic molecular biology
- **sequence alignment**
- probabilistic sequence models
- gene expression analysis
- protein structure prediction
  - by Ameet Soni