CSE 527

Computational Biology

Autumn 2003

Larry Ruzzo

Course Announcement:
CSE 527: Computational Biology

An introduction to the use of computational methods for the understanding of biological systems at the molecular level. Open to biology students interested in learning about algorithms, and to students in computer science, mathematics or statistics interested in applications of those fields to molecular biology.

Time: MW 12:00-1:20

Place: MGH 284

Instructor: Larry Ruzzo (554 Allen, ruzzo@cs.washington.edu)

Course web pages: `http://www.cs.washington.edu/527`

Course mailing list:

Catalog Description (somewhat out of date): CSE 527 Computational Biology (3) Introduces computational methods for understanding biological systems at the molecular level. Problem areas such as mapping and sequencing, sequence analysis, structure prediction, phylogenic inference, regulatory analysis. Techniques such as dynamic programming, Markov models, expectation-maximization, local search. Prerequisite: graduate standing in biological, computer, mathematical or statistical science, or permission of instructor.

Workload: Notes, problem sets and projects. We will encourage projects in which a biologist and a mathematical scientist work together to model and solve a biological problem.

Desired Prerequisites: Ideally, students will have a considerable knowledge of one of computer science, biology, or probability/statistics, together with introductory knowledge of the other two feilds. We'll try to supplement as needed (via lecture, outside reading, project teams, etc.) so that everyone has enough background in the immediately relevant areas to fruitfully proceed.

Rough Course Outline

Essential Background from Molecular Biology

Microarray Analysis Clustering, classification, feature selection for analysis of large scale gene expression data sets generated by microarrays and similar technologies.

Sequence Analysis Statistical modeling of families of DNA or protein sequences: profiles, motif discovery, hidden Markov Models, Expectation - Maximization algorithm. Gene finding.

Molecular Structure Prediction (time permitting) RNA secondary structure prediction; the protein folding problem; protein threading.

He who asks is a fool for five minutes, but he who does not ask remains a fool forever.

-- Chinese Proverb

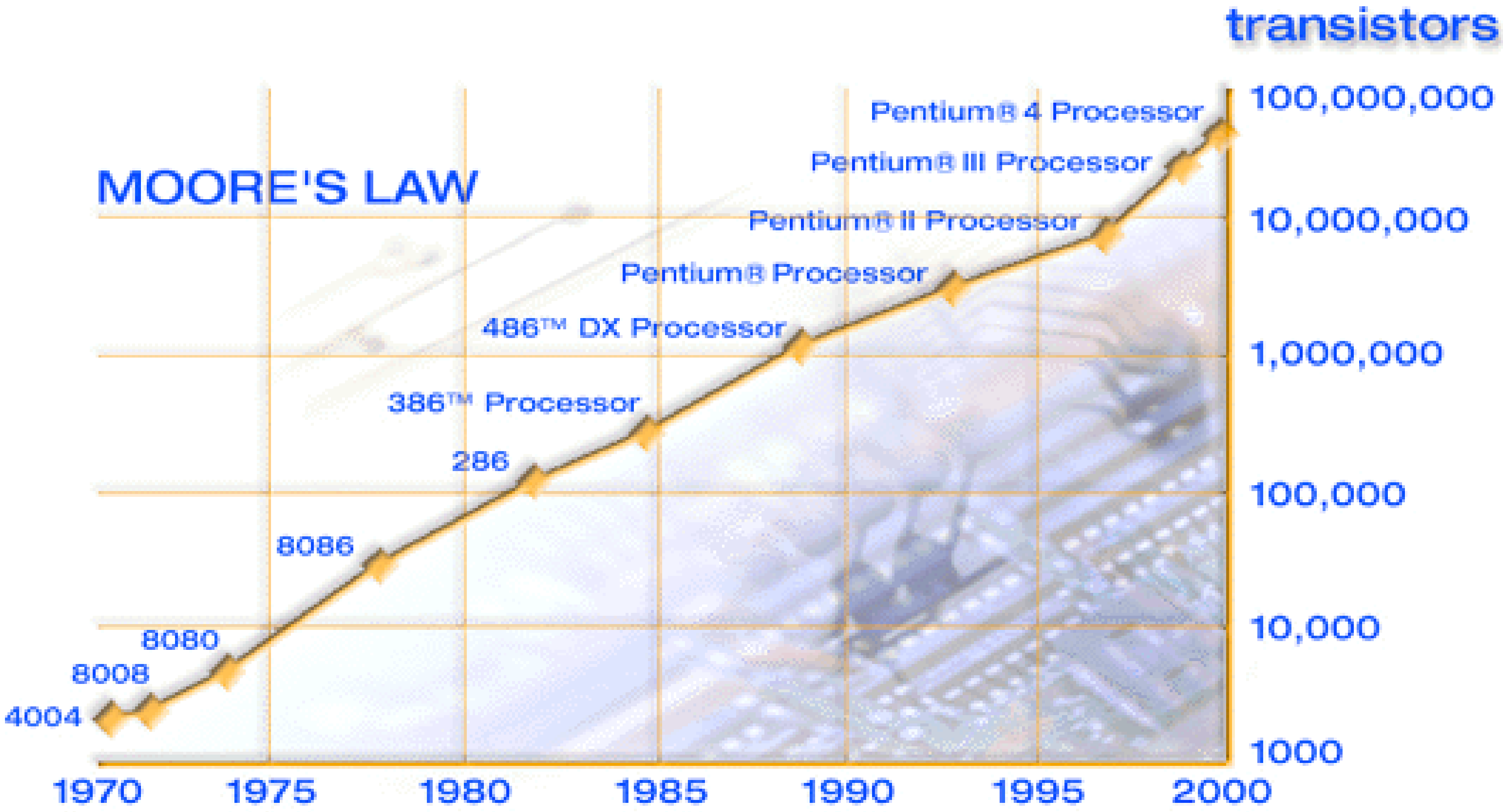
Related Courses

- Genome 540/540 (Winter/Spring)
  - Intro. To Comp. Mol. Bio.
- Genome 590 (A 2003)
  - Population Genetics Seminar

- CSE590CB (AWS)
  - Reading & Research in Comp. Bio.
  - Monday's, 3:30 (MGH 085 this quarter)
- Combi Seminar (Genome 521; AWS)
  - Wednesday's 1:30 K069

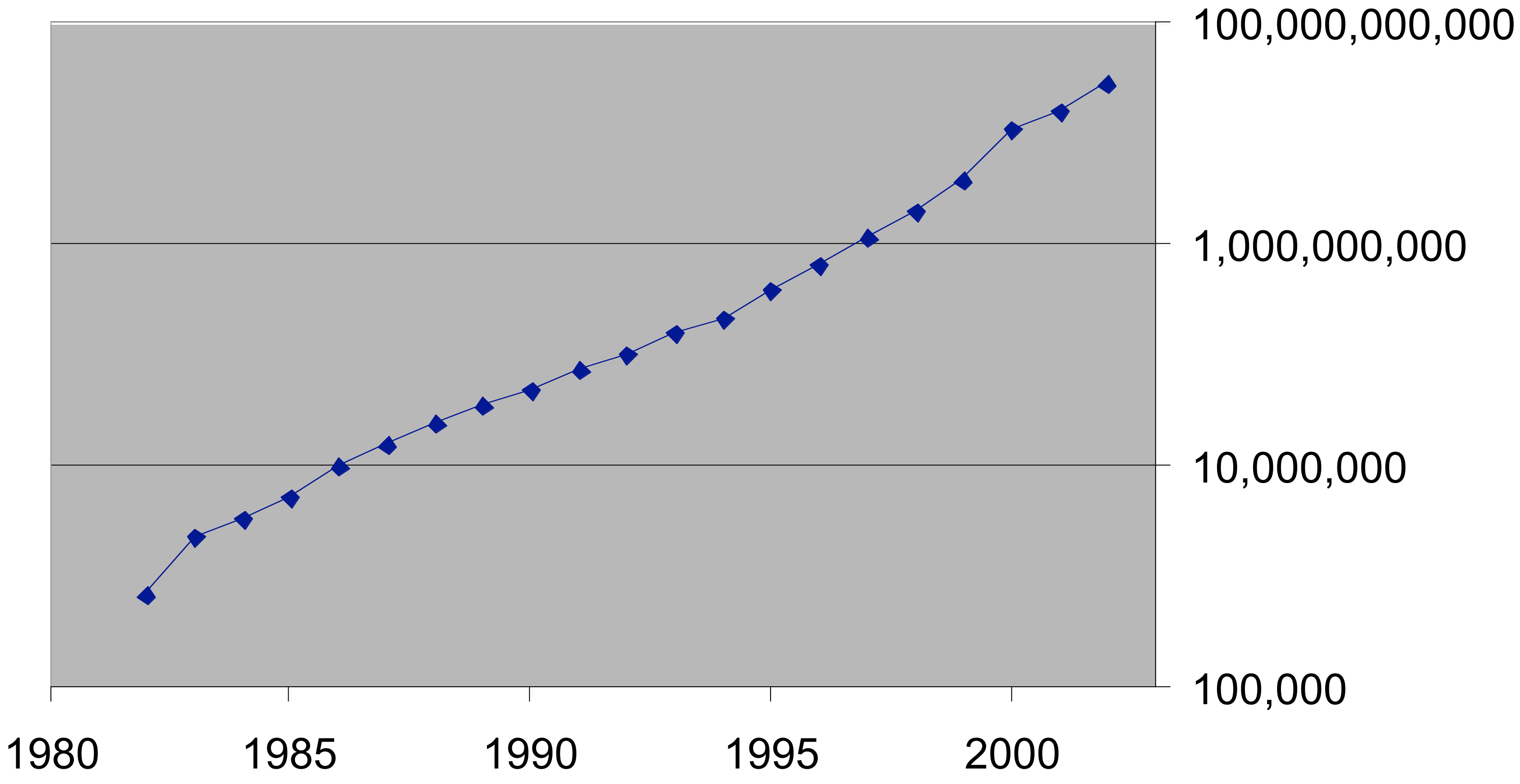
Homework #1

- Find & read a good primer on “bio for cs” (or vice versa, as appropriate)
  e.g., see ones listed on 590cb page
- Email me a few sentences saying
  - What you read (give me a link or citation)
  - Critique it for your meeting your needs
  - Who would it have been good for, if not you



Source: http://www.intel.com/research/silicon/mooreslaw.htm

Growth of GenBank (Nucleotides)



Source: http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.htm (Feb ’03)l

What’s all the fuss?

- The human genome is “finished”…
- Even if it were, that’s only the beginning
- Explosive growth in biological data is revolutionizing biology & medicine

“All pre-genomic lab techniques are obsolete”

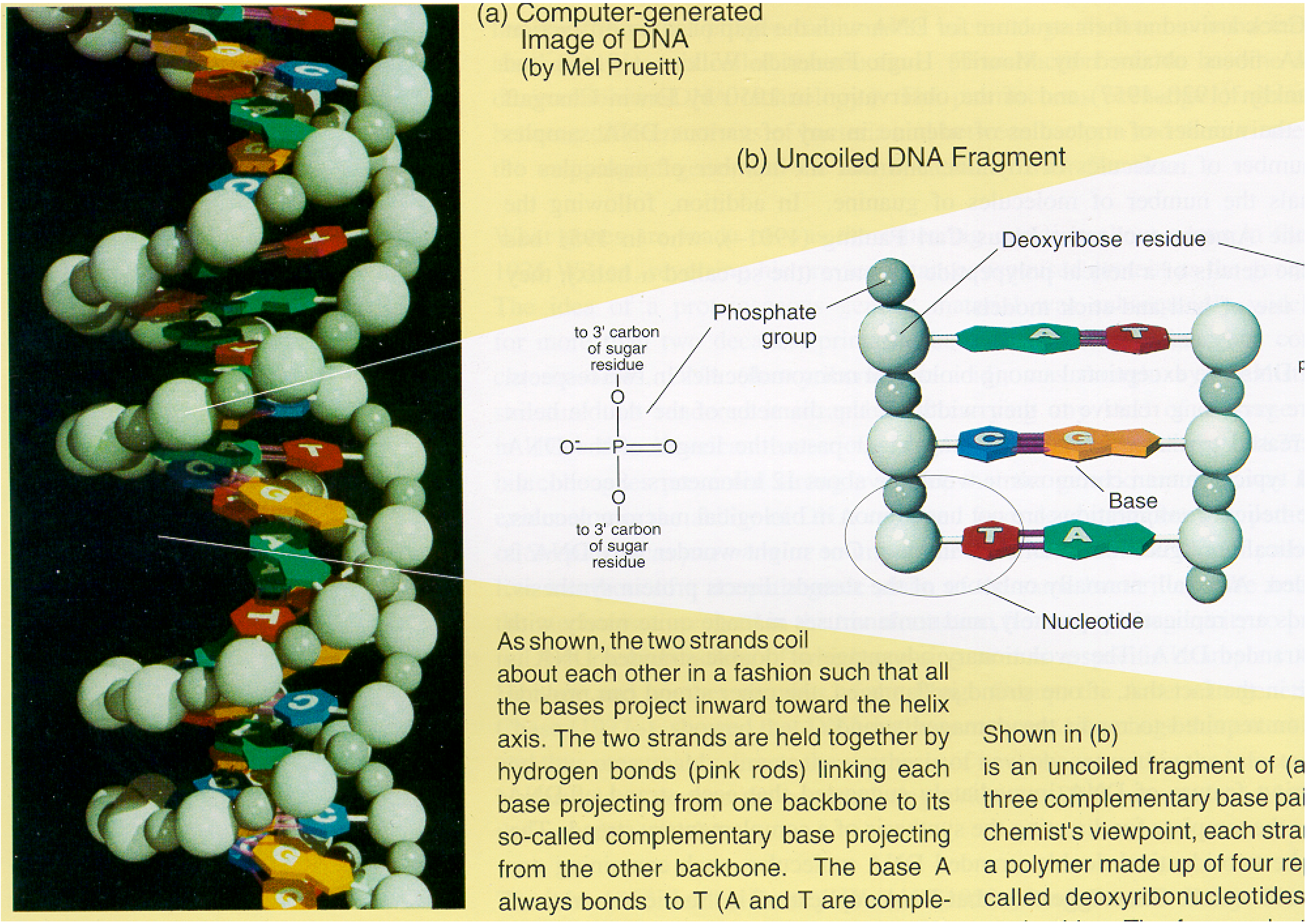
(and computation and mathematics are crucial to post-genomic analysis)

A *VERY* Quick Intro To Molecular Biology

The Genome

- The hereditary info present in every cell
- DNA molecule -- a long sequence of nucleotides (A, C, T, G)
- Human genome -- about 3 x 10^9 nucleotides
- The genome project -- extract & interpret genomic information, apply to genetics of disease, better understand evolution, …

The Double Helix



(a) Computer-generated Image of DNA (by Mel Prueitt)

(b) Uncoiled DNA Fragment

As shown, the two strands coil about each other in a fashion such that all the bases project inward toward the helix axis. The two strands are held together by hydrogen bonds (pink rods) linking each base projecting from one backbone to its so-called complementary base projecting from the other backbone. The base A always bonds to T (A and T are comple-

Shown in (b) is an uncoiled fragment of (a three complementary base pai chemist's viewpoint, each stra a polymer made up of four re called deoxyribonucleotides

Los Alamos Science

Genetics - the study of heredity

- A gene -- classically, an abstract heritable attribute existing in various forms (alleles)
- Genotype vs phenotype
- Mendel
  - Each individual two copies of each gene
  - Each parent contributes one (rfandomly)
  - Independent assortment

Cells

- Chemicals inside a sac - the plasma membrane
- Prokaryotes (e.g., bacteria) - little recognizable substructure
- Eukaryotes (all multicellular organisms, and many single celled ones, like yeast) - genetic material in nucleus, other organelles for other specialized functions

Chromosomes

- 1 pair of DNA molecules (+ protein wrapper)
- Most prokaryotes have just 1 chromosome
- Eukaryotes - all cells have same number of chromosomes, e.g. fruit flies 8, humans & bats 46, …

Mitosis/Meiosis

- Most “higher” eukaryotes are diploid - have homologous pairs of chromosomes, one maternal, other paternal
- Mitosis - cell division, duplicate each chromosome, 1 copy to each daughter cell
- Meiosis - 2 divisions form 4 haploid gametes (egg/sperm)
  - Recombination/crossover -- exchange maternal/paternal segments

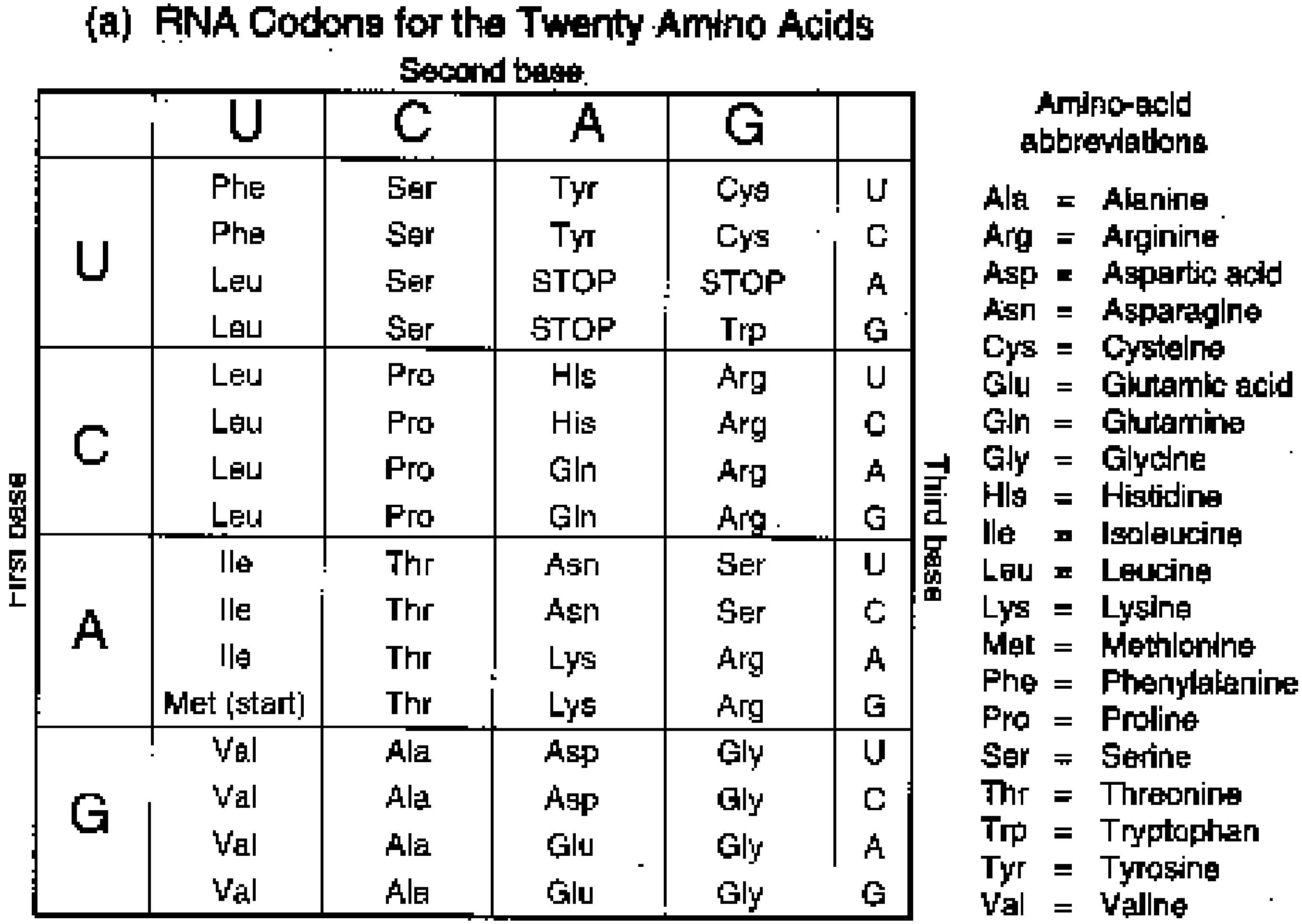
Proteins

- Chain of amino acids, of 20 kinds
- Proteins are the major functional elements in cells
  - Structural
  - Enzymes (catalyze chemical reactions)
  - Receptors
  - Transcription factors
  - …

The “Central Dogma”

- Genes encode proteins
- DNA transcribed into messenger RNA
- RNA translated into proteins
- Triple code (codons)

The Genetic Code



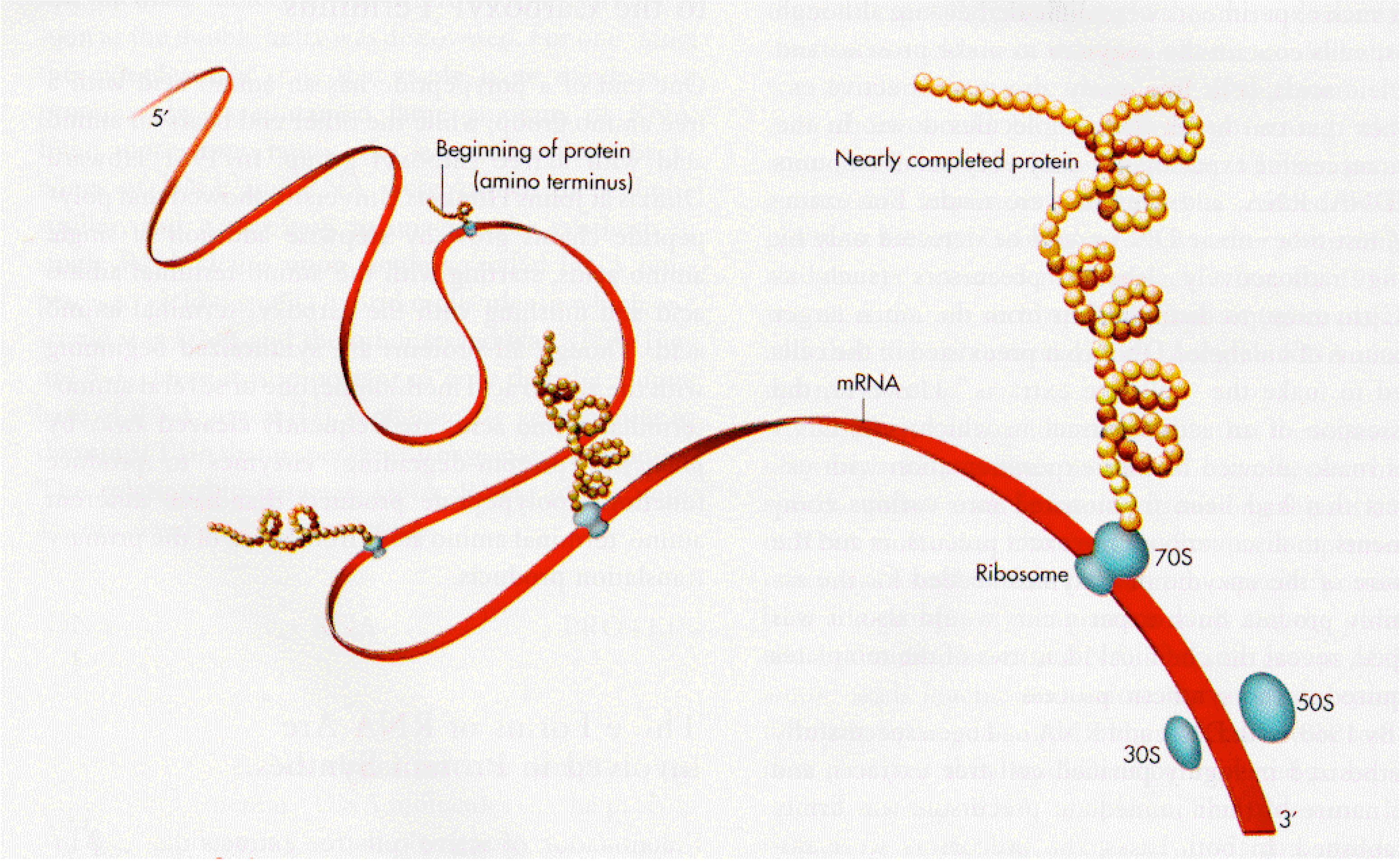
(a) RNA Codons for the Twenty Amino Acids

First base	Second base: U	C	A	G	Third base
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met (start)	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Amino-acid abbreviations

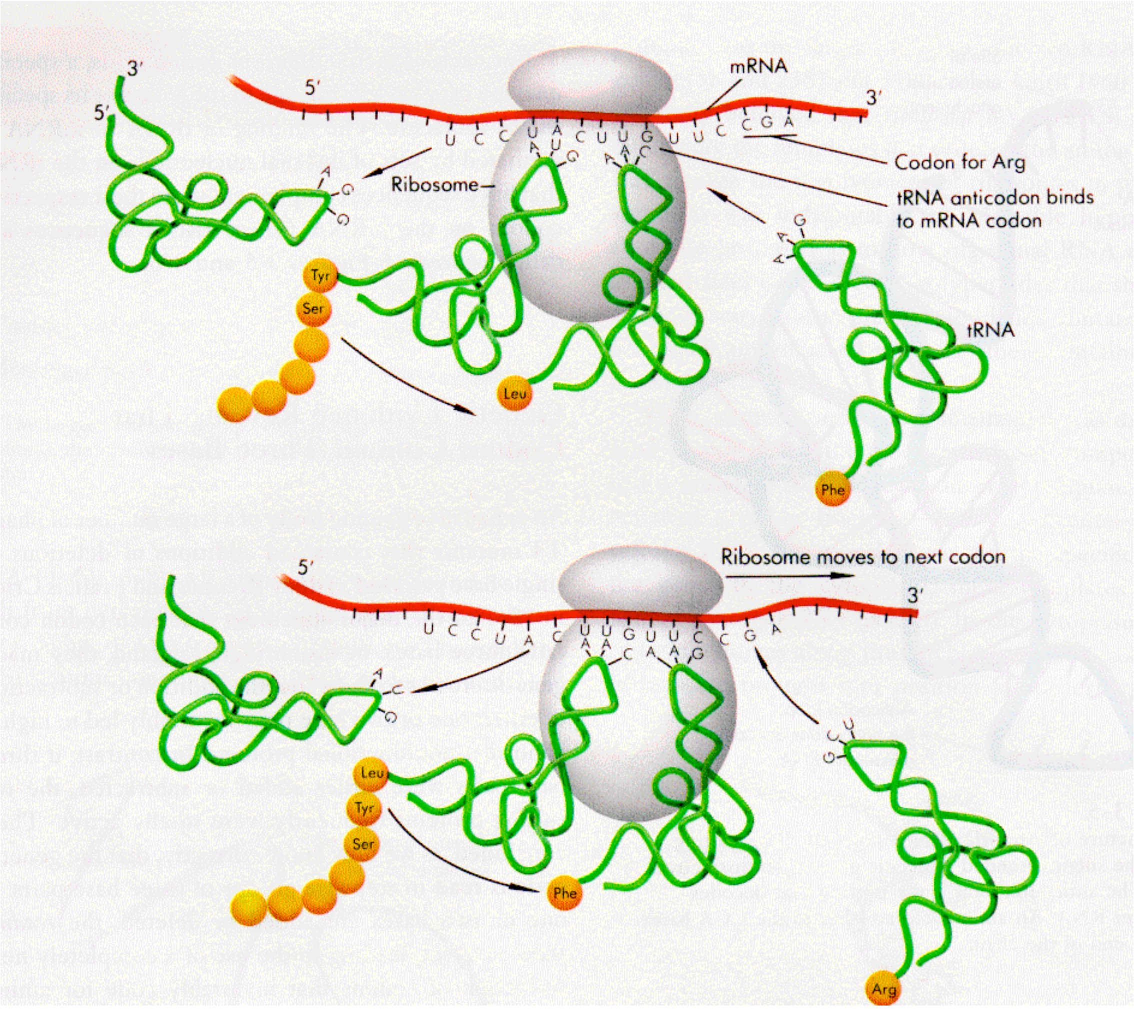
Ala	=	Alanine
Arg	=	Arginine
Asp	=	Aspartic acid
Asn	=	Asparagine
Cys	=	Cysteine
Glu	=	Glutamic acid
Gln	=	Glutamine
Gly	=	Glycine
His	=	Histidine
Ile	=	Isoleucine
Leu	=	Leucine
Lys	=	Lysine
Met	=	Methionine
Phe	=	Phenylalanine
Pro	=	Proline
Ser	=	Serine
Thr	=	Threonine
Trp	=	Tryptophan
Tyr	=	Tyrosine
Val	=	Valine

Translation: mRNA □ Protein



Watson, Gilman, Witkowski, & Zoller, 1992

Ribosomes



Watson, Gilman, Witkowski, & Zoller, 1992

Gene Structure

- Transcribed 5’ to 3’
- Promoter region and transcription factor binding sites precede 5’
- Transcribed region includes 5’ and 3’ untranslated regions
- In eukaryotes, most genes also include introns, spliced out before export from nucleus, hence before translation

Genome Sizes

	Base Pairs	Genes
Mycoplasma genitalium	580,073	483
E. coli	4,639,221	4,290
Saccharomyces cerevisiae	12,495,682	5,726
Caenorhabditis elegans	95.5 x 10^6	19,820
Arabidopsis thaliana	115,409,949	25,498
Drosophila melanogaster	122,653,977	13,472
Humans	3.3 x 10^9	~25,000

… and much more …

- Read one of the many intro surveys or books for much more info.