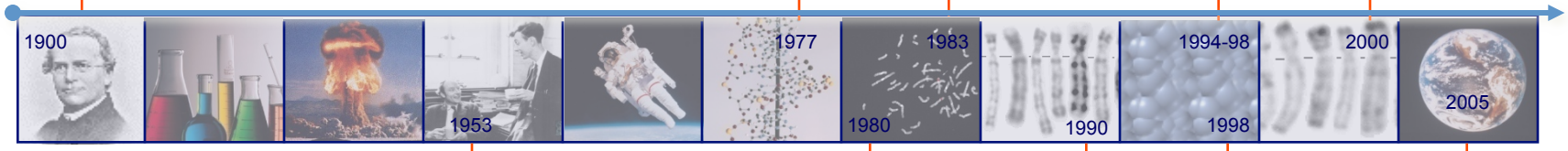
Genomics marked the beginning of a new age in biology and medicine



Rediscovery of Mendel's laws helps establish the science of genetics

Watson and Crick identify DNA (the double helix) as the Chemical basis of heredity

Sanger and Gilbert derive methods of sequencing DNA

DNA markers used to map human disease genes to chromosomal regions

Huntington disease gene mapped to chromosome 4

Human Genome Projects (HPG) begins-an international effort to map and sequence all the genes in the human genome

Genetic and physical mapping

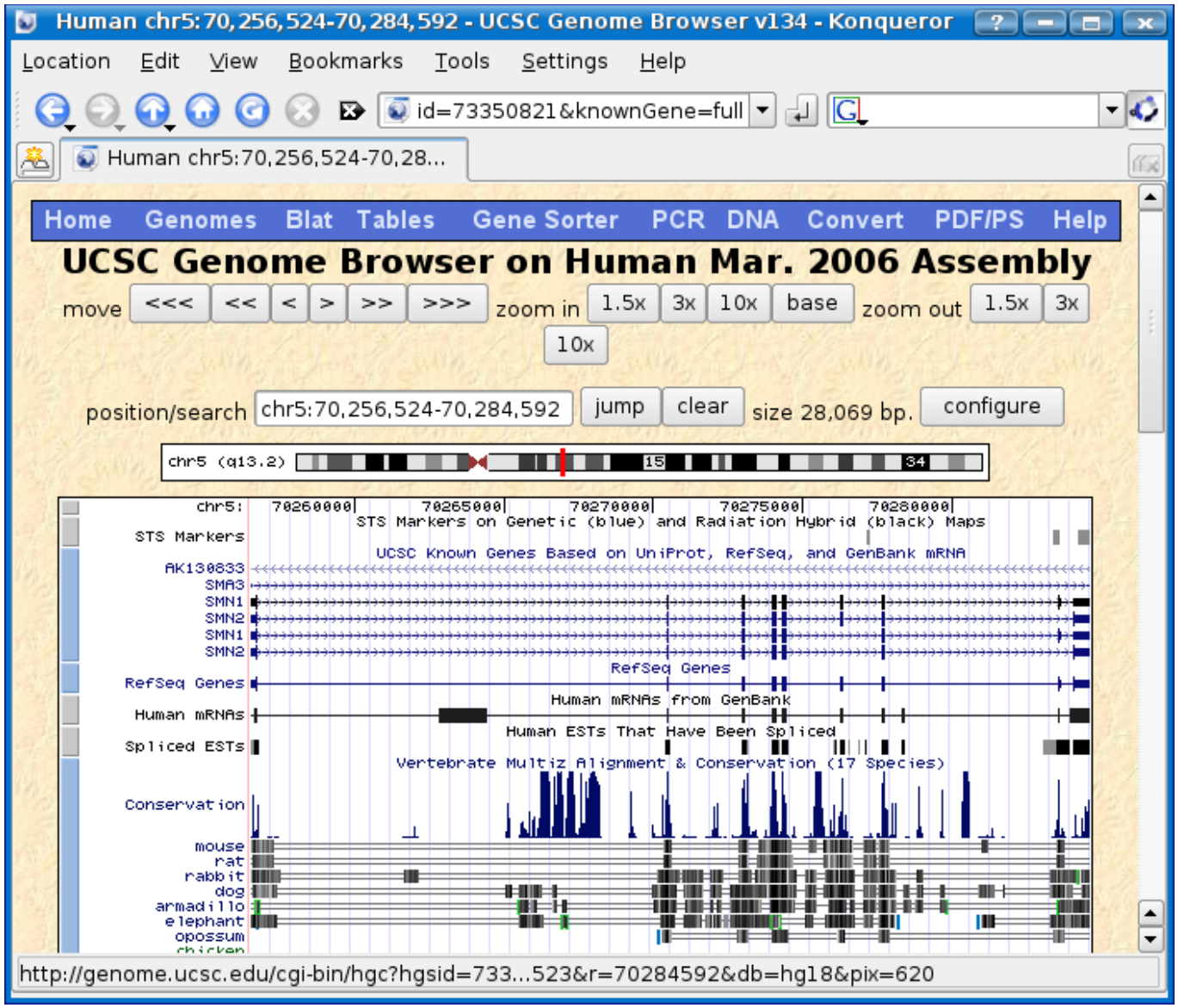
DNA markers used to map human disease genes to chromosomal regions

Working Draft of the human genome sequencing complete

Release of Human Genome Project

Source: Health Policy Research Bulletin, volume 1 issue2, September 2001

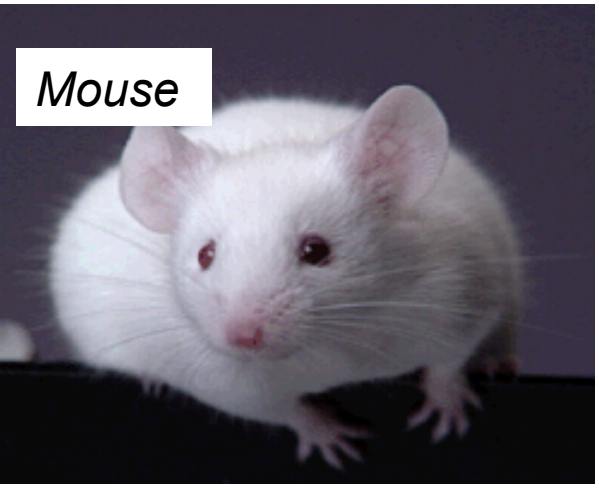
The genome browser



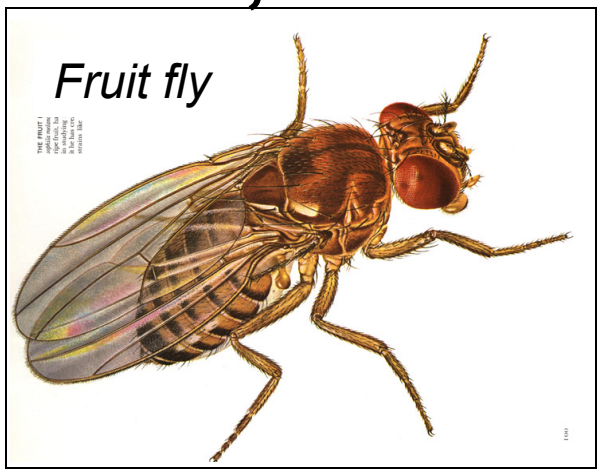
Link

Number of protein coding genes

20,210

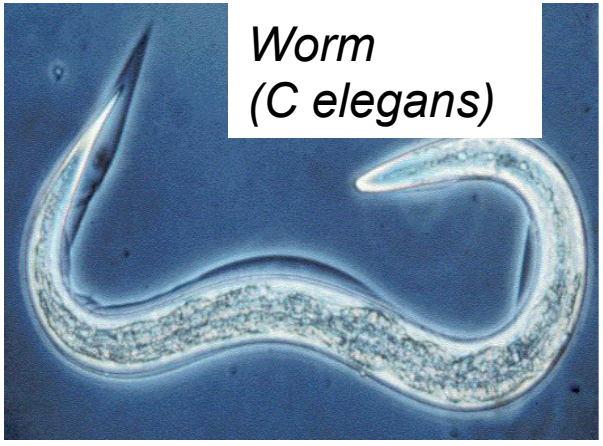


13,601

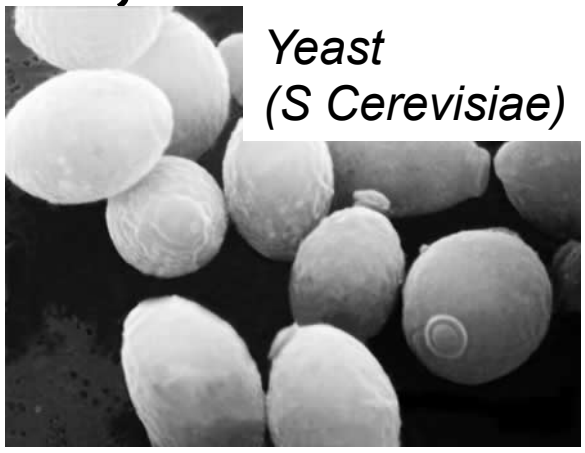


20,568

19,735



5,616

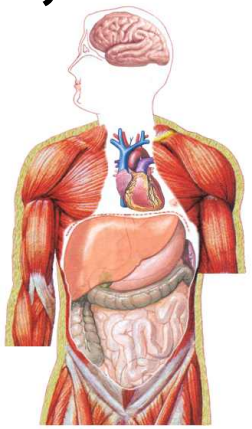


482

How comes we have so few genes give that we are so complex???

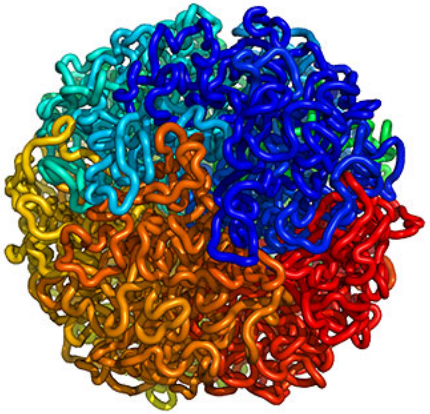
- We have many non-protein coding genes
- Our genes are longer and more complex
- Regulation of human genes activity is more complex
- Repeats (formerly known as “junk DNA” (yet not garbage) contribute to complexity
- Combinatorial interactions among genes and products

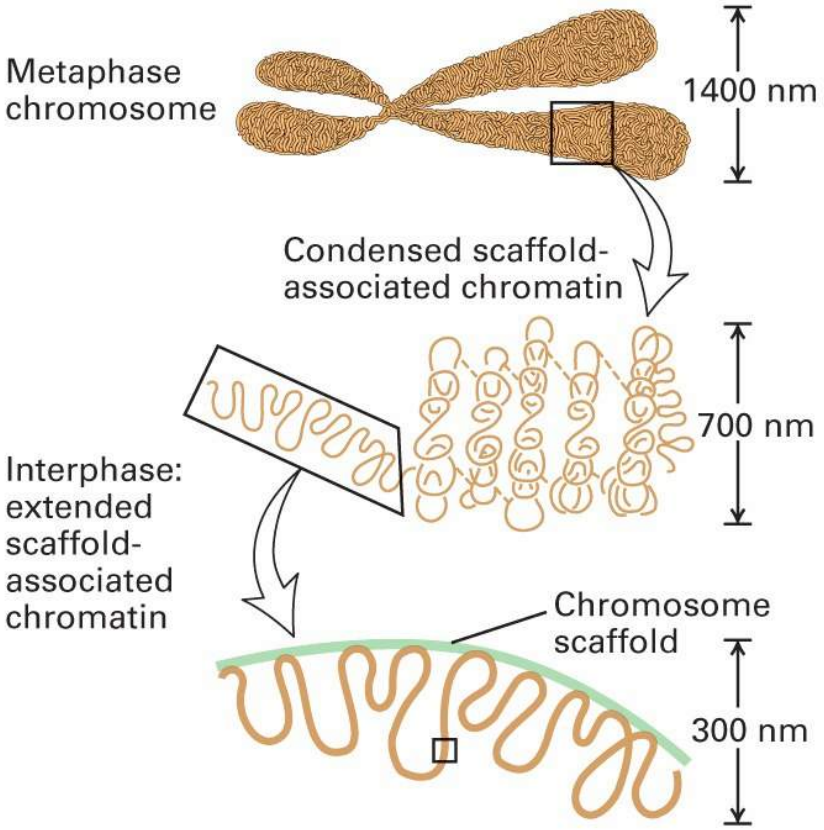
21,710

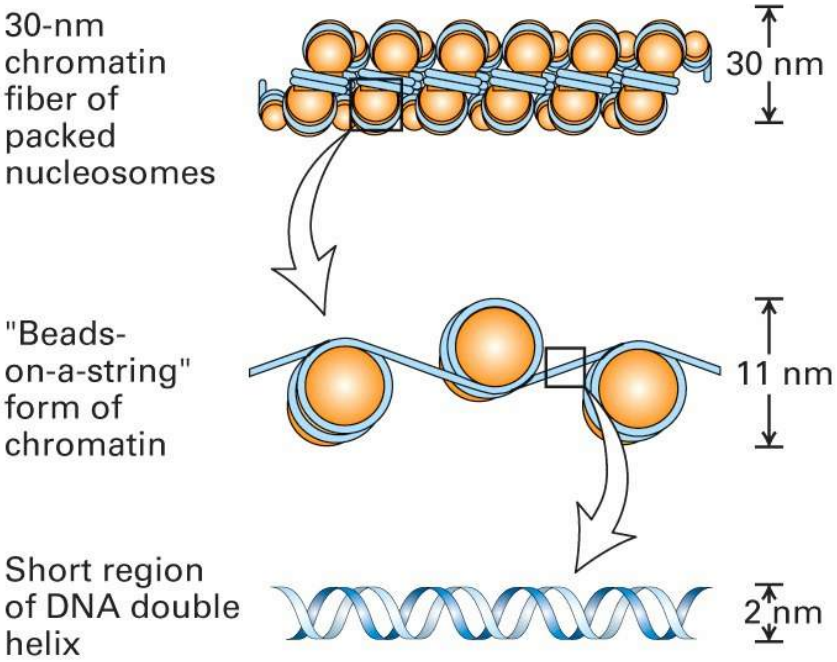


19,735

The hierarchical structure of the genome







Lodish et al. *Molecular Biology of the Cell* (5th ed.). W.H. Freeman & Co., 2003.

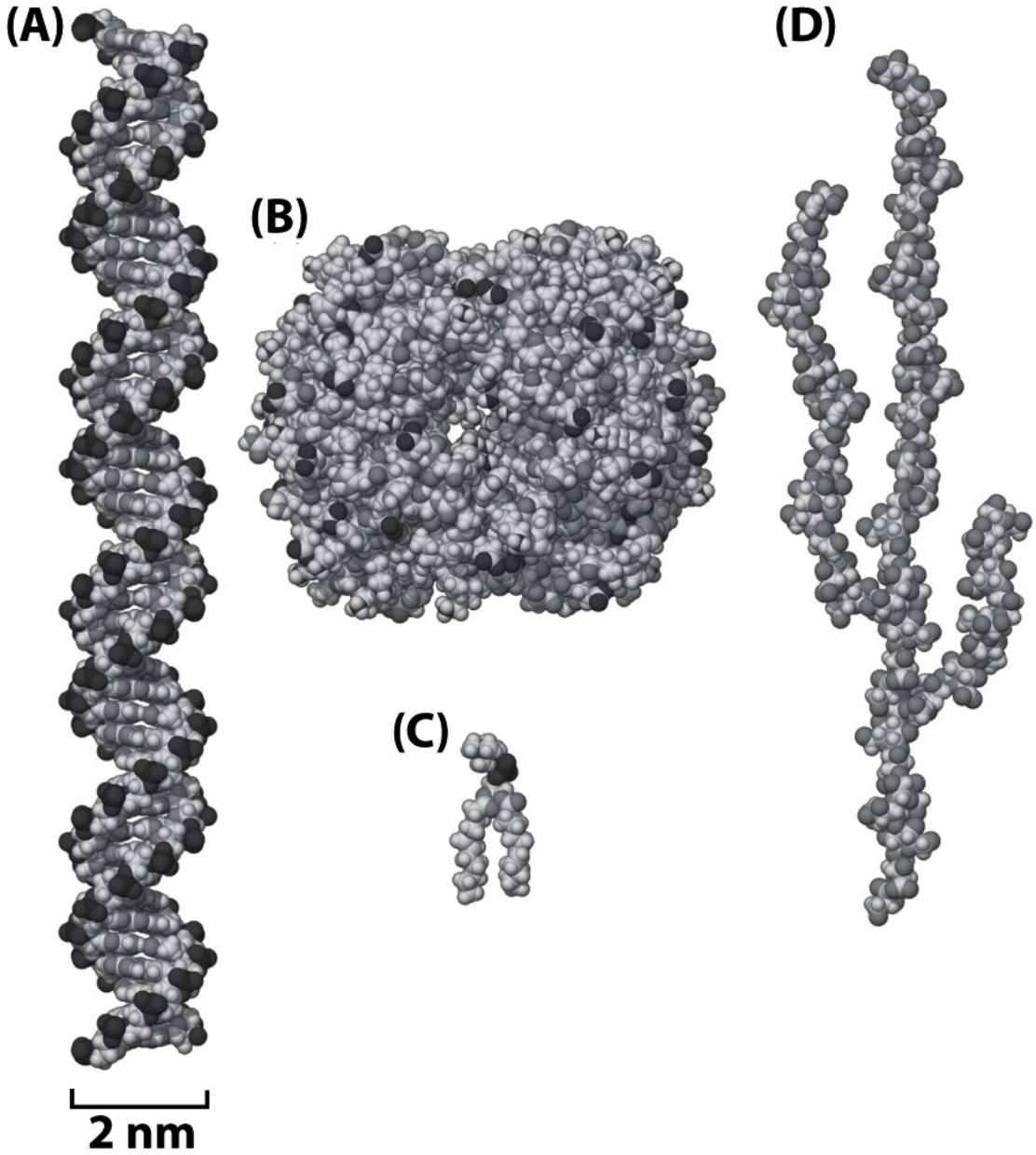


Figure 1.1 Physical Biology of the Cell (© Garland Science 2009)

Macromolecules (quasi-random hetero-polymers)

- Synthesis
  - Molecular "strings" made in cell by linking *monomers* (symbols) from a specified set (alphabet)

- Examples
  - Polysaccharides (sugar chains)
  - Proteins (amino acid chains)
  - DNA & RNA (Nucleic acids; nucleotide chains)

The central dogma (sic) of molecular biology

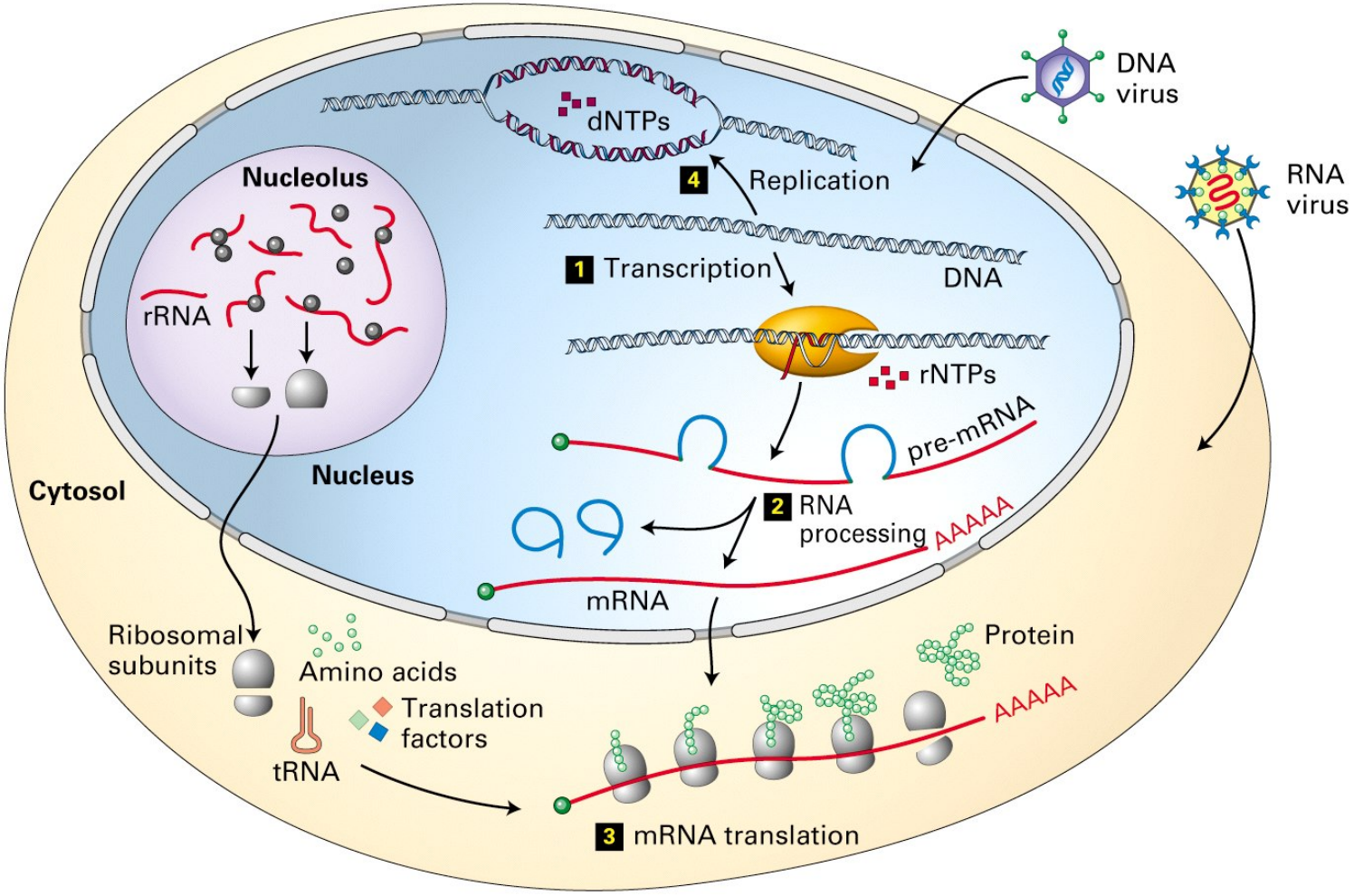
DNA ⇒ mRNA ⇒ Protein

RNA polymerase (an enzyme, a protein) *transcribes* a segment of DNA to a complementary messenger RNA

Primary messenger RNA is processed to mature mRNA

Mature mRNA is *translated* into protein by a *ribosome* (a complex of proteins and rRNA)

The Central Dogma: a cellular context



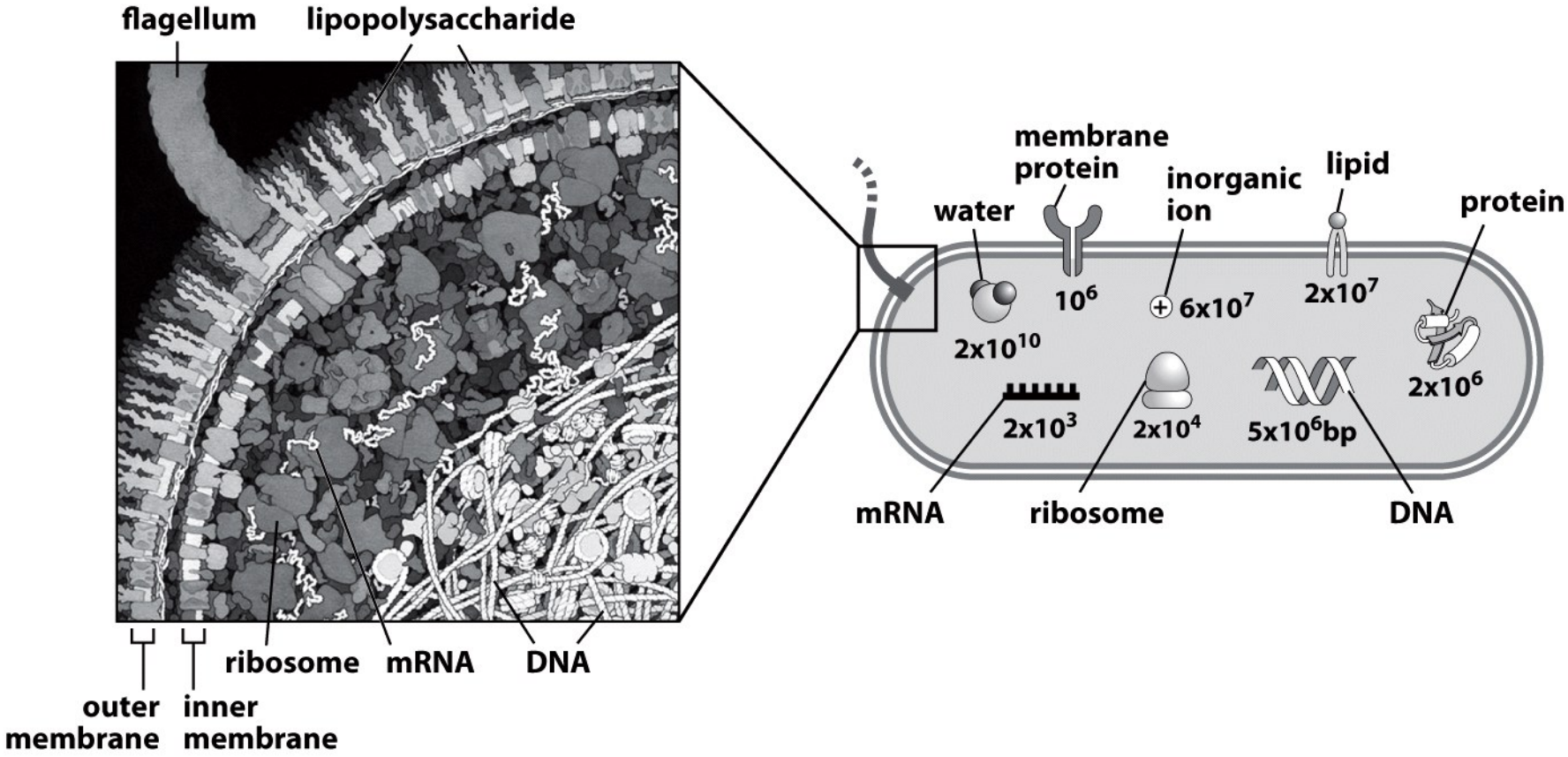
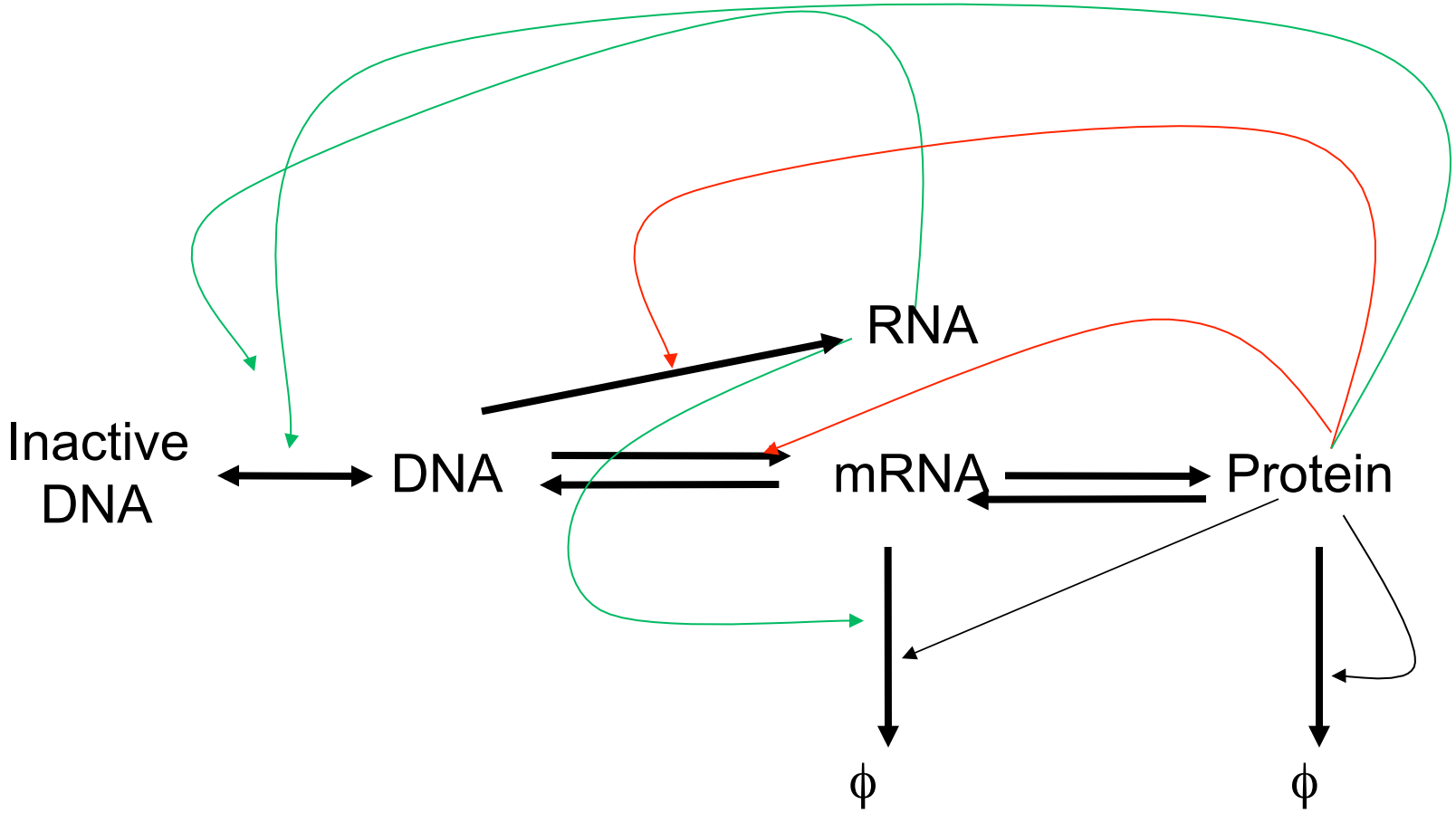
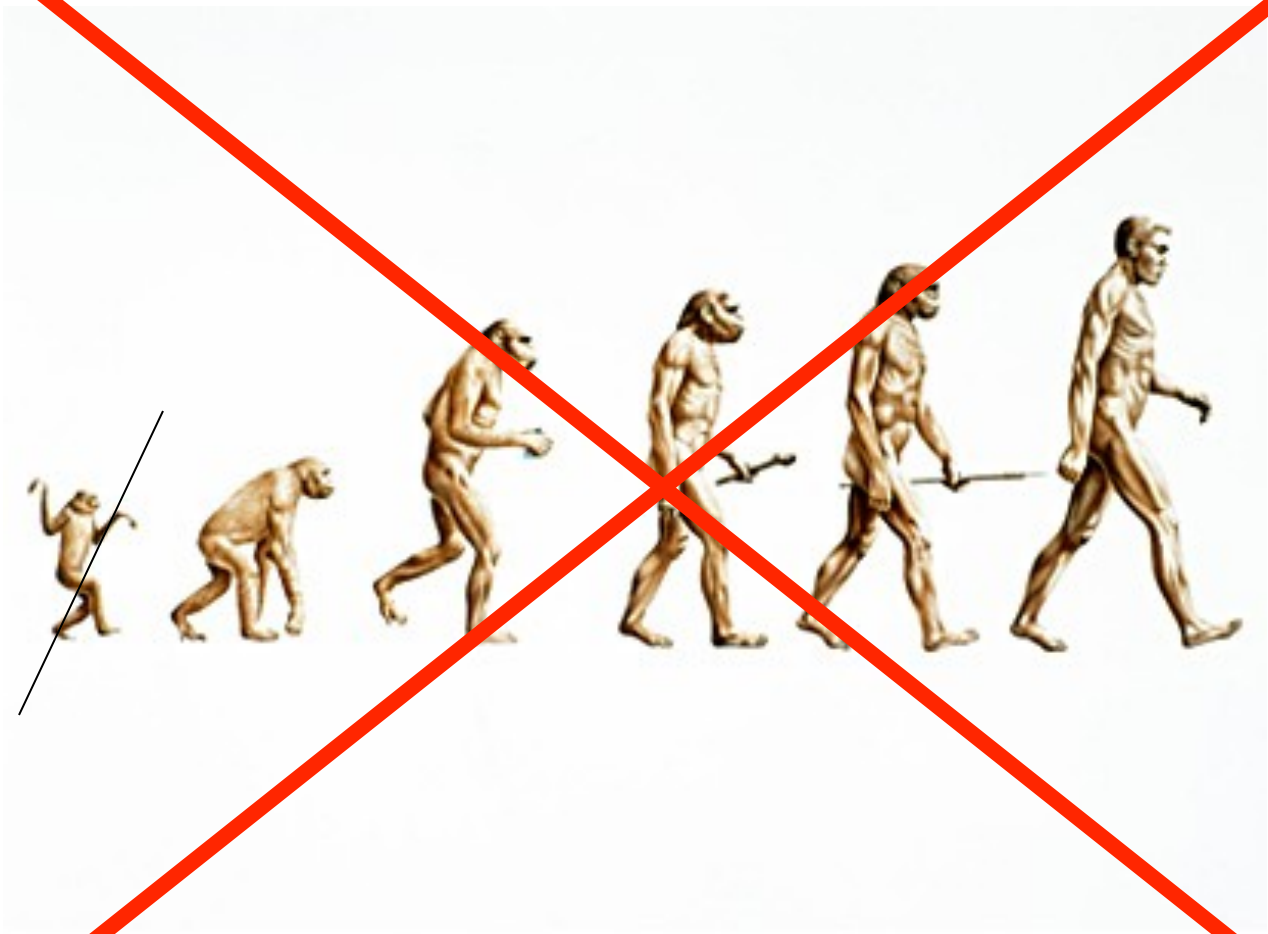


Figure 2.2 Physical Biology of the Cell (© Garland Science 2009)

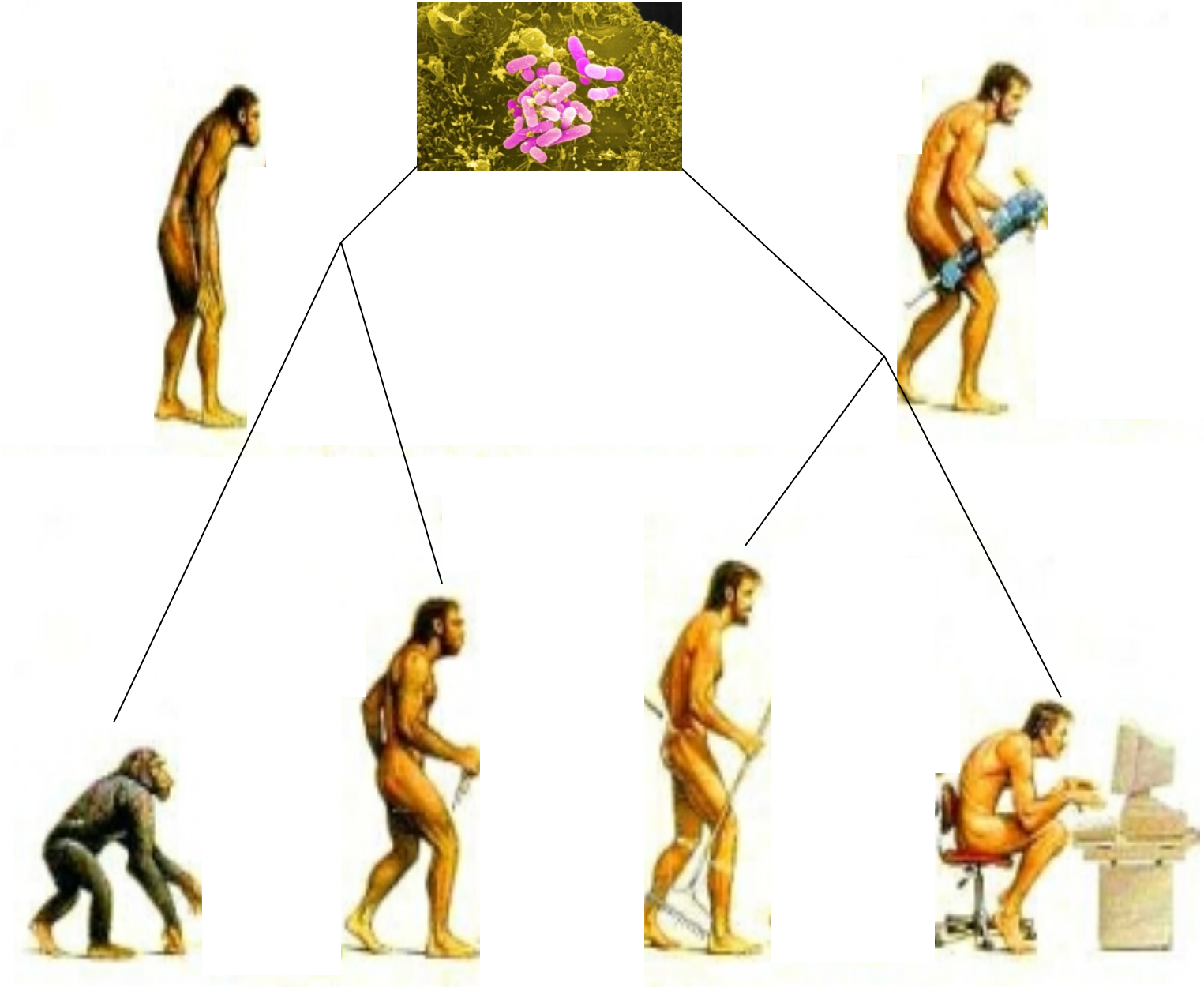
The Central Dogma of Molecular Biology
Expressing the genome



Evolution



Corrected view of evolution



The tree of life

Phylogenetic trees: evolutionary vs newtonian time

(A)

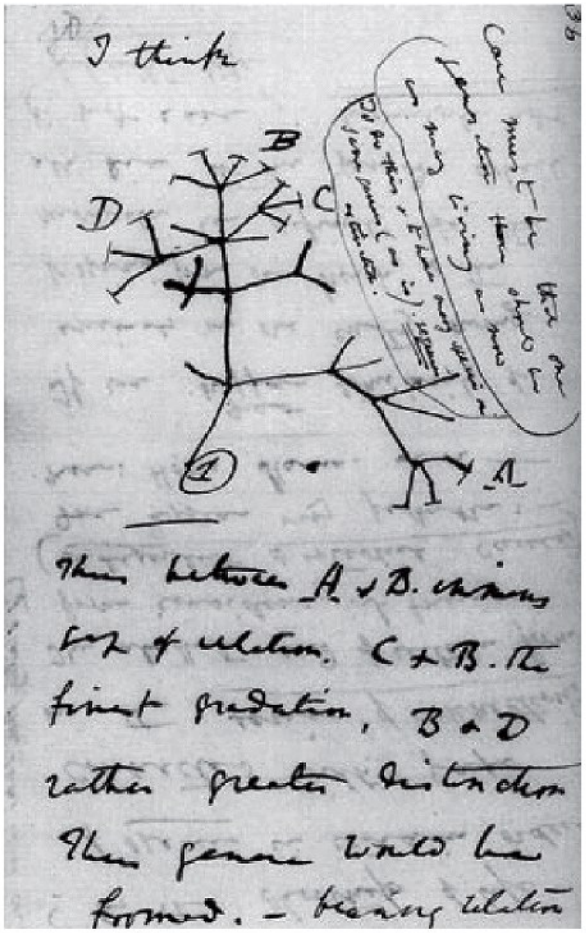


Figure 3.4 Physical Biology of the Cell (© Garland Science 2009)

(B)

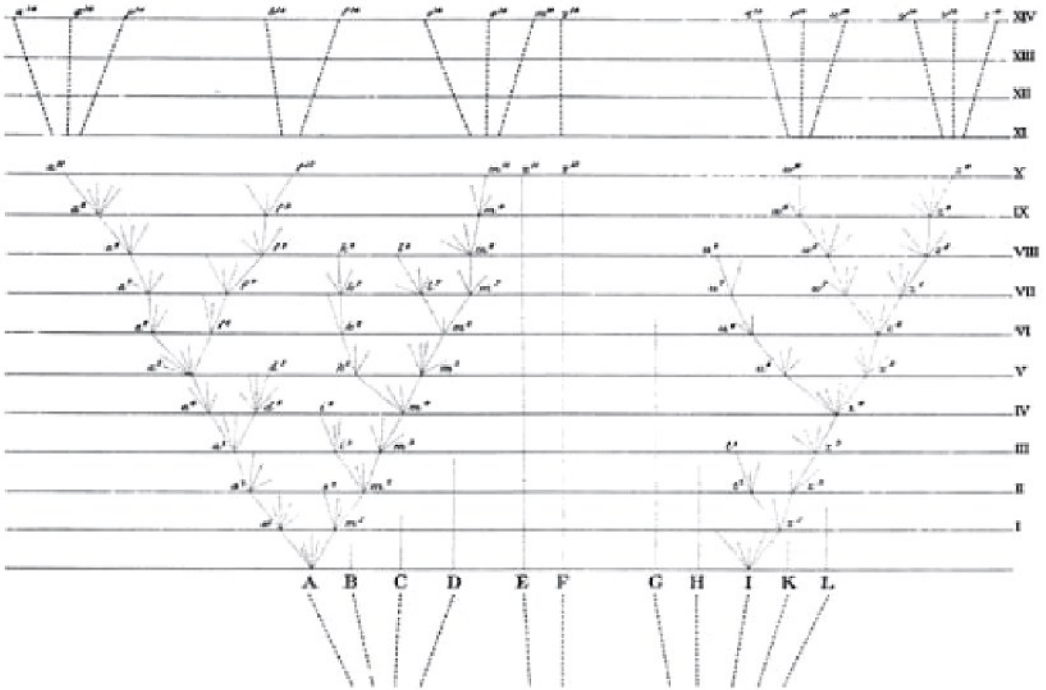
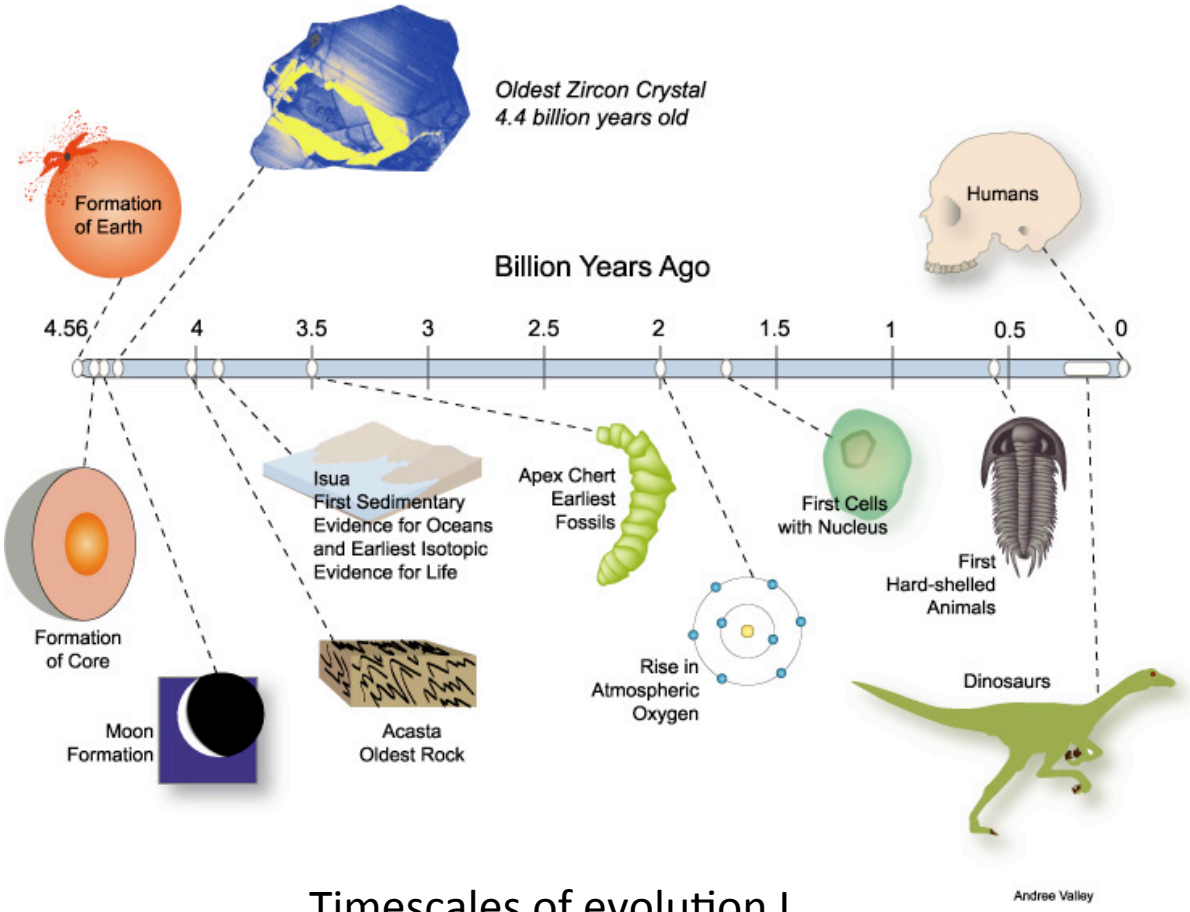


Figure 3.4 Two versions of Darwin's phylogenetic tree. (A) In his notebooks, Darwin drew the first version of what we now recognize as a common schematic demonstrating the relatedness of organisms. He introduced this speculative sketch with the words "I think" as his theory was beginning to take form. (B) In the final published version of *On the Origin of Species*, the tree had assumed more detail showing the passage of time and explicitly indicating that most species have gone extinct. (Adapted from C. Darwin, *On the Origin of Species*, London, John Murray, 1859. Courtesy of The American Museum of Natural History.)



Timescales of evolution I

Timescales of evolution II

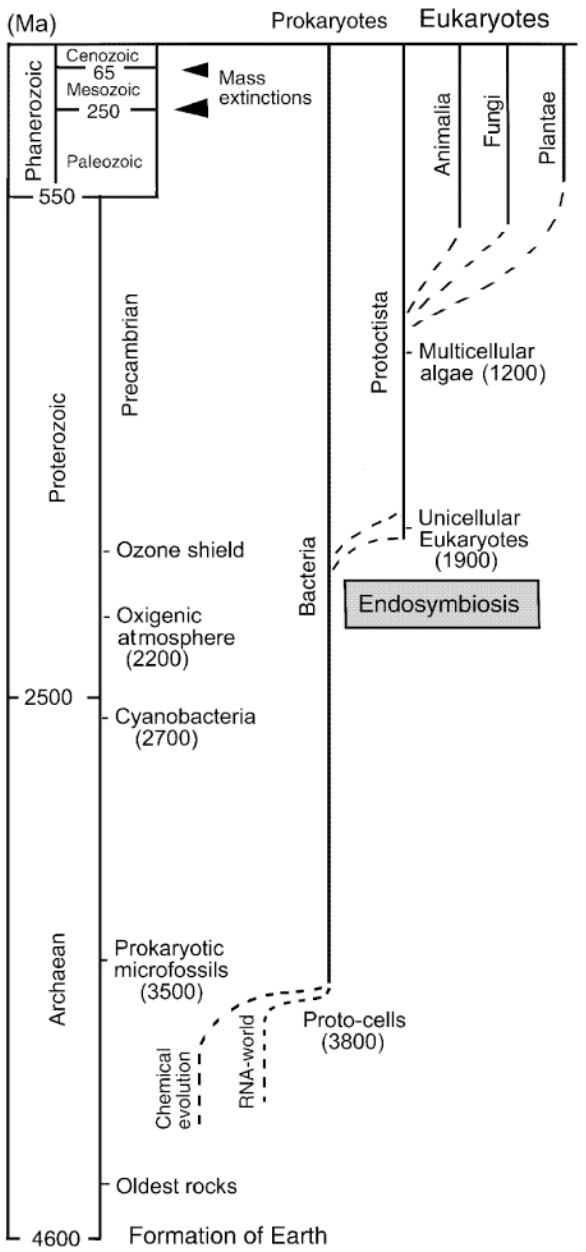


Fig. 2 Geological time scale with key events in the history of life, from the formation of the Earth to the present. All five kingdoms of organisms are included (Bacteria, Protoctista, Animalia, Fungi, Plantae). *Ma* millions of years

La Sapienza Università di Roma, Dipartimento di Fisica
a.a. 2011-2012, Corso di Biofisica
Andrea Giansanti

The case of Carl Woese

The tree of life (tolweb.org)

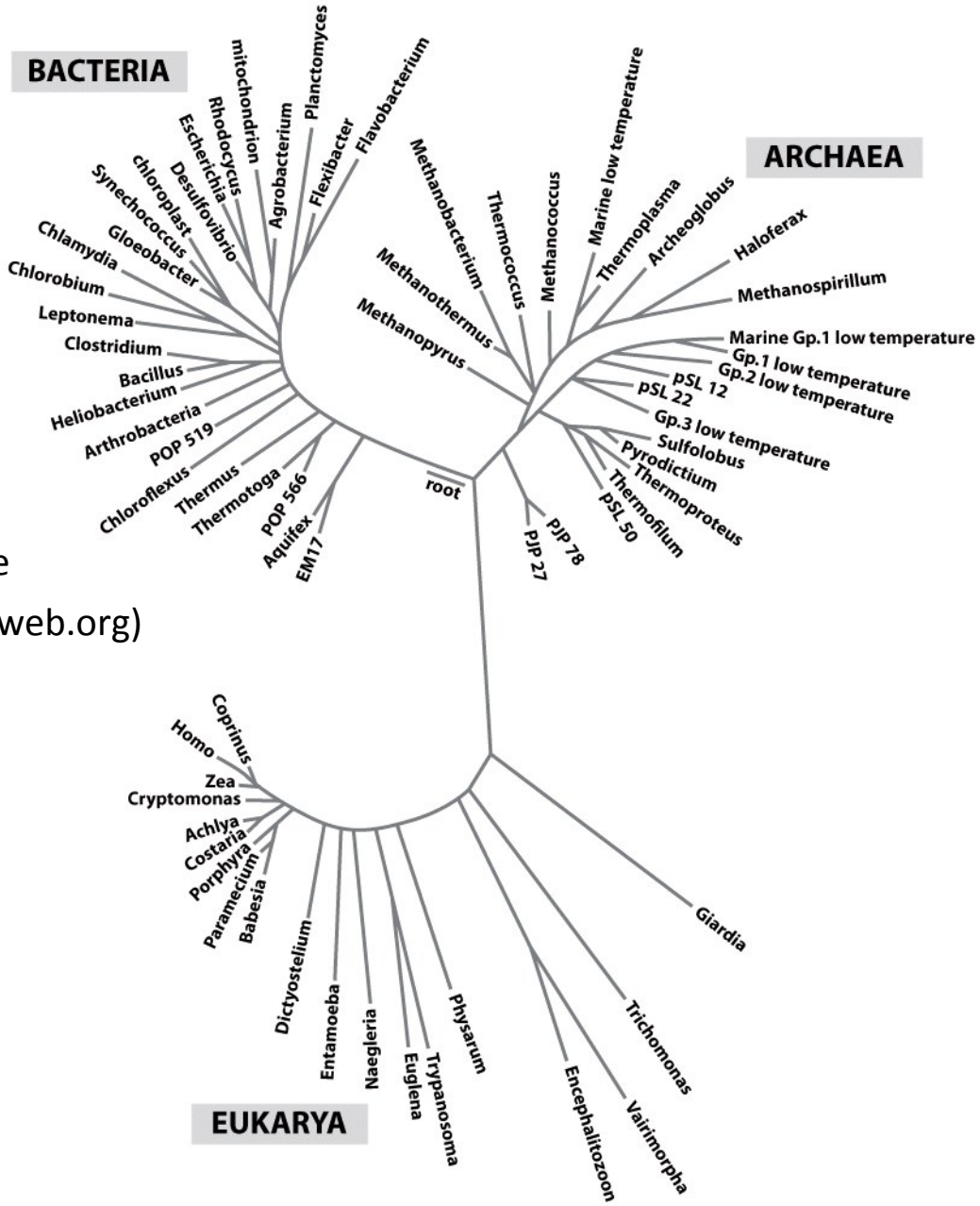


Figure 3.6 Physical Biology of the Cell (© Garland Science 2009)

How genomes evolve?

Consider two distinct possibilities:

- Genomes evolve by lots of *de-novo* "inventions"
- Genomes evolve predominantly by mixing and matching existings parts

History

- 1866 Mendel discovered genetics
- 1869 DNA discovered
- 1944 Avery & McCarty demonstrated DNA as carrier of genetic info
- 1953 Watson & Crick deduced 3D struct of DNA
- 1960 Elucidation of genetic code, mapping DNA to protein
- 1970 Development of DNA sequencing techniques: sequence segmentation and electrophoresis
- 1980 Development of PCR: exploiting natural replication, amplify DNA samples so that they are enough for doing expt
- 1990 Human Genome Project
- 2002 Human genome published
- Now Understanding the detail mechanism of the cell

Body

- Our body consists of a number of organs
- Each organ composes of a number of tissues
- Each tissue composes of cells of the same type

Cell

- Performs two types of function
  - Chemical reactions necessary to maintain our life
  - Pass info for maintaining life to next generation
- In particular
  - Protein performs chemical reactions
  - DNA stores & passes info
  - RNA is intermediate between DNA & proteins

All living organisms are made of cells

The specific smallness of cells selects a peculiar regime where noise and deterministic forces match

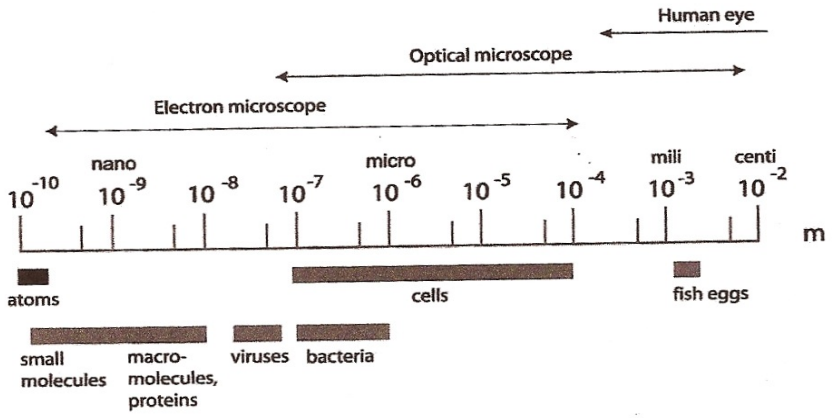


Fig. 11.1. Log scale of comparative dimensions.

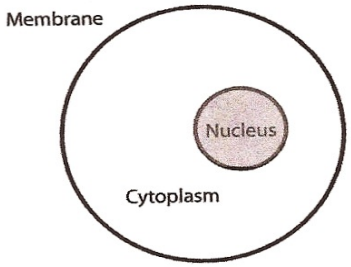
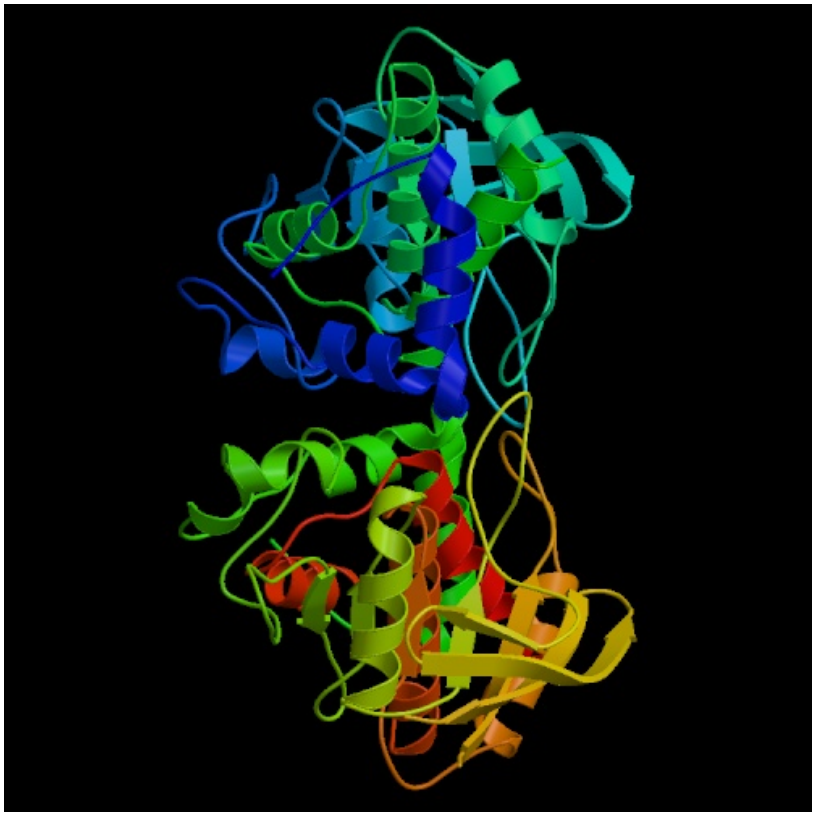


Fig. 11.2. Schematic structure of a cell.

From: K. Huang, Lectures on Statistical Physoics and protein folding

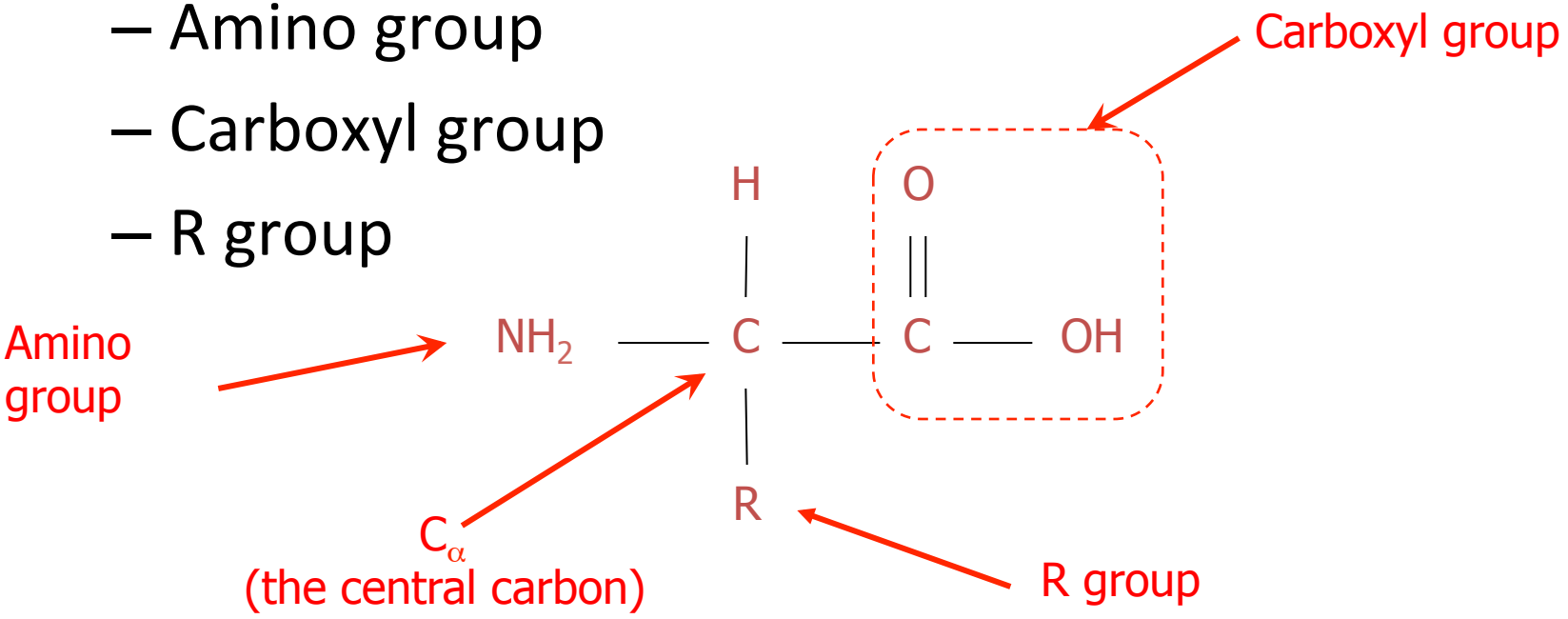
Protein

- A sequence composed from an alphabet of 20 amino acids
  - Length is usually 20 to 5000 amino acids
  - Average around 350 amino acids
- Folds into 3D shape, forming the building blocks & performing most of the chemical reactions within a cell



Amino Acid

- Each amino acid consist of
  - Amino group
  - Carboxyl group
  - R group



Classification of Amino Acids

- Amino acids can be classified into 4 types.

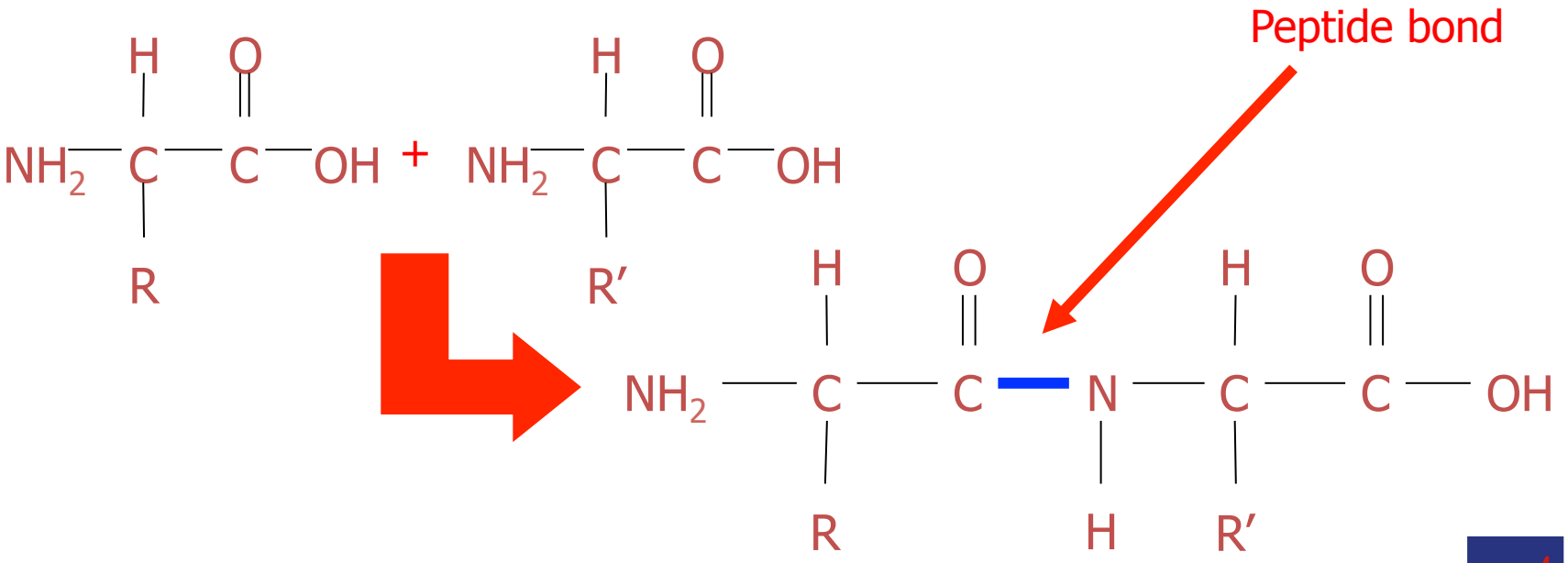
- Positively charged (basic)
  - Arginine (Arg, R)
  - Histidine (His, H)
  - Lysine (Lys, K)
- Negatively charged (acidic)
  - Aspartic acid (Asp, D)
  - Glutamic acid (Glu, E)

Classification of Amino Acids

- Polar (overall uncharged, but uneven charge distribution. can form hydrogen bonds with water. they are called hydrophilic)
  - Asparagine (Asn, N)
  - Cysteine (Cys, C)
  - Glutamine (Gln, Q)
  - Glycine (Gly, G)
  - Serine (Ser, S)
  - Threonine (Thr, T)
  - Tyrosine (Tyr, Y)
- Nonpolar (overall uncharged and uniform charge distribution. cant form hydrogen bonds with water. they are called hydrophobic)
  - Alanine (Ala, A)
  - Isoleucine (Ile, I)
  - Leucine (Leu, L)
  - Methionine (Met, M)
  - Phenylalanine (Phe, F)
  - Proline (Pro, P)
  - Tryptophan (Trp, W)
  - Valine (Val, V)

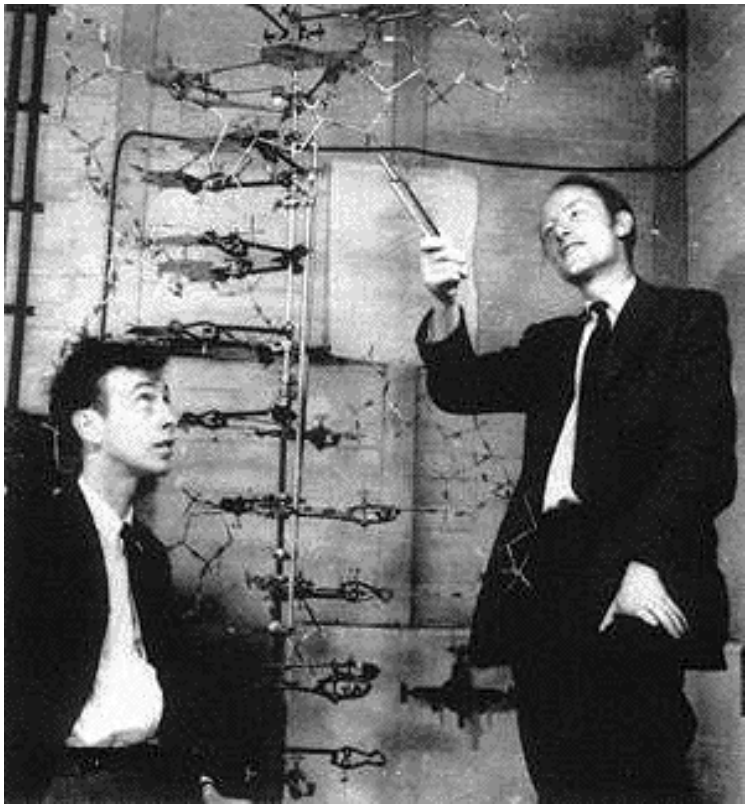
Protein & Polypeptide Chain

- Formed by joining amino acids via peptide bond
- One end the amino group, called N-terminus
- The other end is the carboxyl group, called C-terminus



DNA

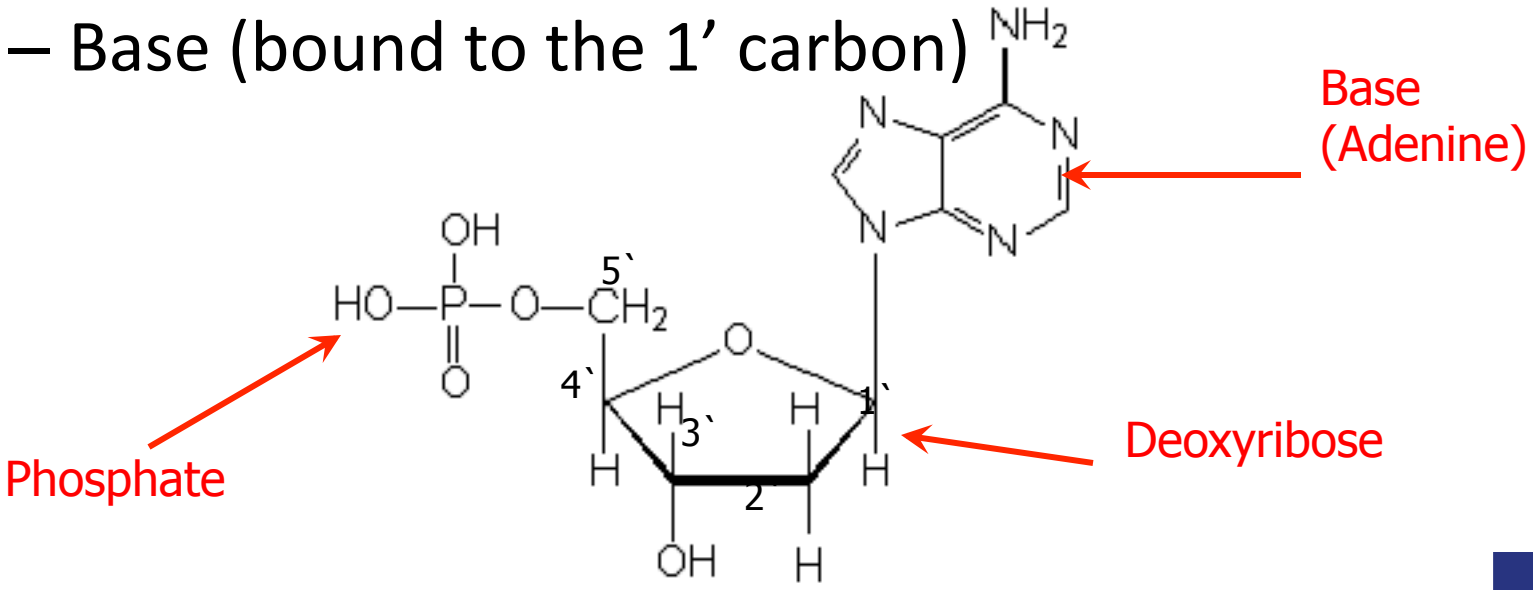
- Stores instruction needed by the cell to perform daily life function
- Consists of two strands interwoven together and form a double helix
- Each strand is a chain of some small molecules called nucleotides



Francis Crick shows James Watson the model of DNA in their room number 103 of the Austin Wing at the Cavendish Laboratories, Cambridge

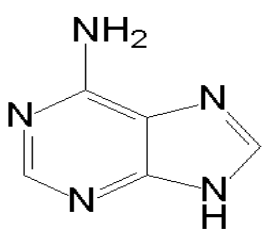
Nucleotide

- Consists of three parts:
  - Deoxyribose
  - Phosphate (bound to the 5’ carbon)
  - Base (bound to the 1’ carbon)

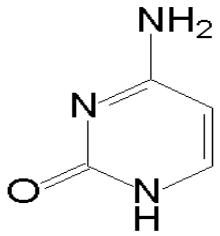


Classification of Nucleotides

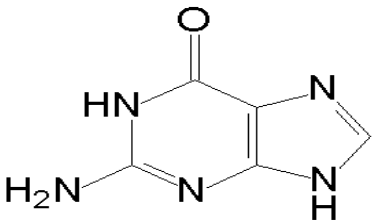
- 5 diff nucleotides: adenine(A), cytosine(C), guanine(G), thymine(T), & uracil(U)
- A, G are purines. They have a 2-ring structure
- C, T, U are pyrimidines. They have a 1-ring structure
- DNA only uses A, C, G, & T



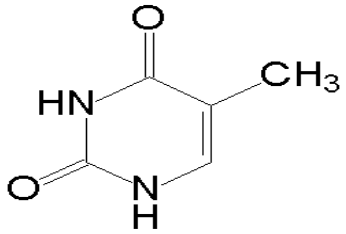
A



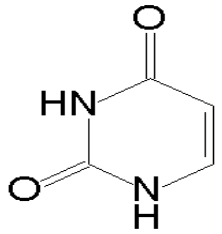
C



G



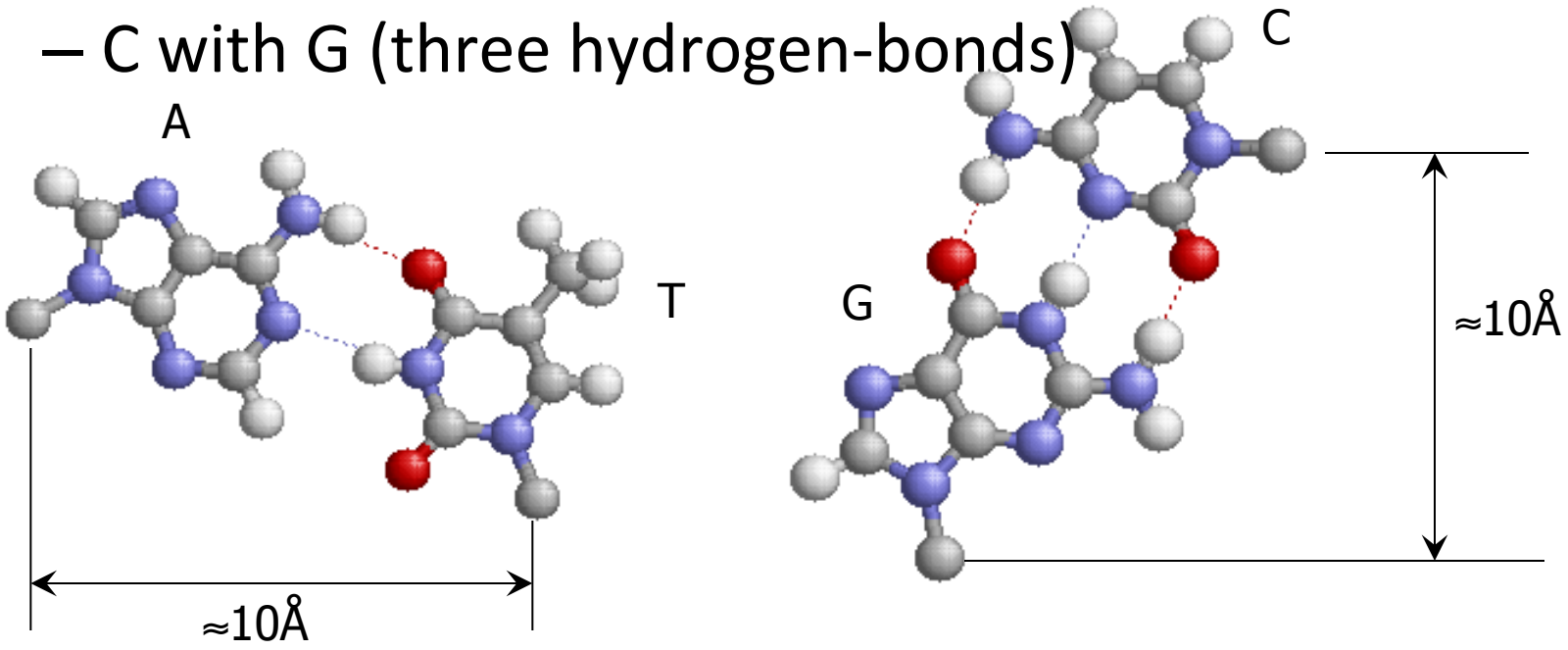
T



U

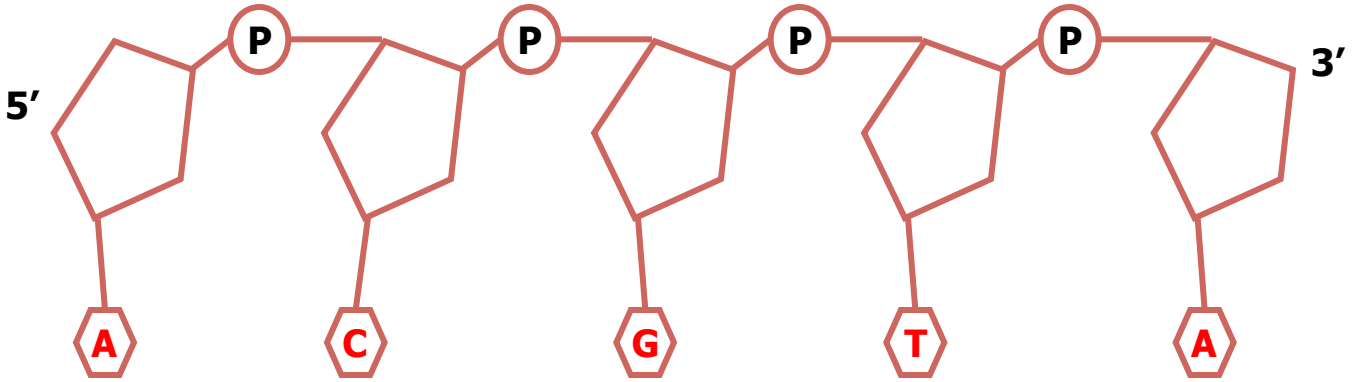
Watson-Crick rules

- Complementary bases:
  - A with T (two hydrogen-bonds)
  - C with G (three hydrogen-bonds)

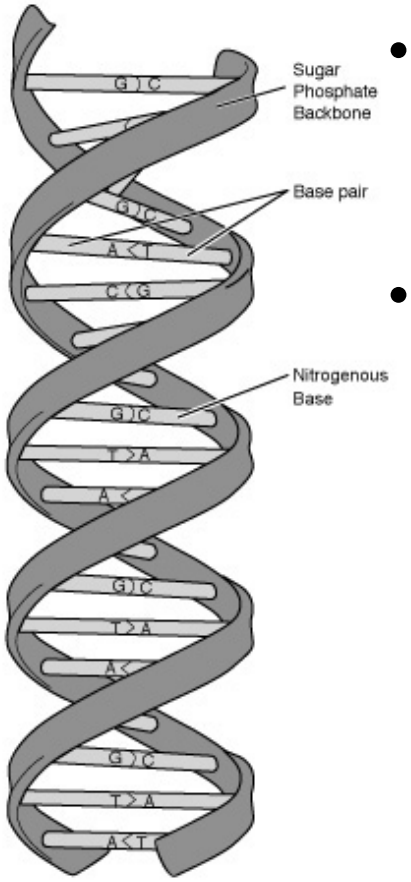


Orientation of a DNA

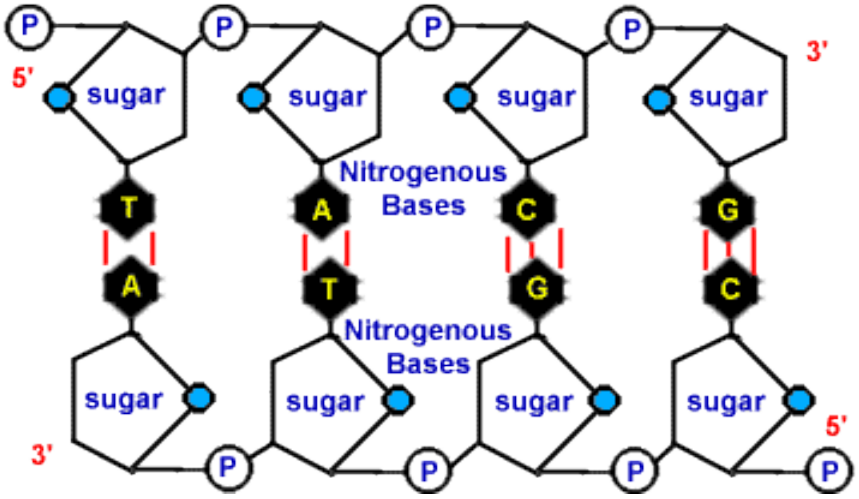
- One strand of DNA is generated by chaining together nucleotides, forming a phosphate-sugar backbone
- It has direction: from 5’ to 3’, because DNA always extends from 3’ end:
  - Upstream, from 5’ to 3’
  - Downstream, from 3’ to 5’



Double Stranded DNA



- DNA is double stranded in a cell. The two strands are anti-parallel. One strand is reverse complement of the other
- The double strands are interwoven to form a double helix



Locations of DNAs in a Cell?

- Two types of organisms
  - Prokaryotes (single-celled organisms with no nuclei. e.g., bacteria)
  - Eukaryotes (organisms with single or multiple cells. their cells have nuclei. e.g., plant & animal)
- In Prokaryotes, DNA swims within the cell
- In Eukaryotes, DNA locates within the nucleus

Chromosome

- DNA is usually tightly wound around histone proteins and forms a chromosome
- The total info stored in all chromosomes constitutes a genome
- In most multi-cell organisms, every cell contains the same complete set of chromosomes
  - May have some small different due to mutation
- Human genome has 3G base pairs, organized in 23 pairs of chromosomes

Gene

- A gene is a sequence of DNA that encodes a protein or an RNA molecule
- About 30,000 – 35,000 (protein-coding) genes in human genome
- For gene that encodes protein
  - In Prokaryotic genome, one gene corresponds to one protein
  - In Eukaryotic genome, one gene can corresponds to more than one protein because of the process “alternative splicing”

Locations of DNAs in a Cell?

- Two types of organisms
  - Prokaryotes (single-celled organisms with no nuclei. e.g., bacteria)
  - Eukaryotes (organisms with single or multiple cells. their cells have nuclei. e.g., plant & animal)
- In Prokaryotes, DNA swims within the cell
- In Eukaryotes, DNA locates within the nucleus

Chromosome

- DNA is usually tightly wound around histone proteins and forms a chromosome
- The total info stored in all chromosomes constitutes a genome
- In most multi-cell organisms, every cell contains the same complete set of chromosomes
  - May have some small different due to mutation
- Human genome has 3G base pairs, organized in 23 pairs of chromosomes

Gene

- A gene is a sequence of DNA that encodes a protein or an RNA molecule
- About 30,000 – 35,000 (protein-coding) genes in human genome
- For gene that encodes protein
  - In Prokaryotic genome, one gene corresponds to one protein
  - In Eukaryotic genome, one gene can corresponds to more than one protein because of the process “alternative splicing”

Complexity of Organism vs. Genome Size

- Human Genome: 3G base pairs
- *Amoeba dubia* (a single cell organism): 600G base pairs

$\Rightarrow$ Genome size has no relationship with the complexity of the organism

Number of Genes vs. Genome Size

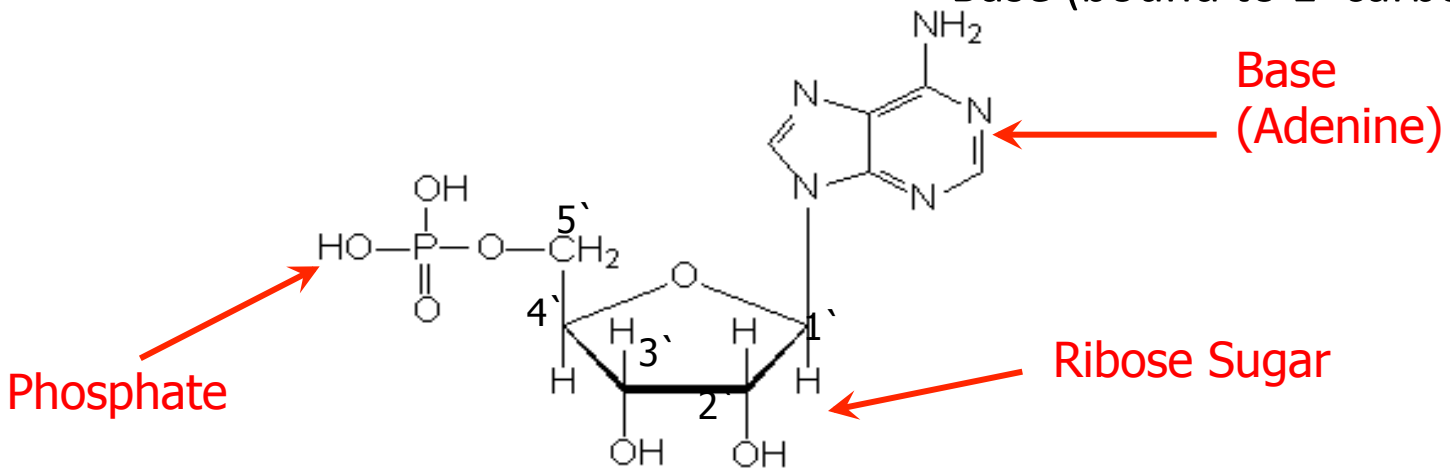
- Prokaryotic genome (e.g., *E. coli*)
  - Number of base pairs: 5M
  - Number of genes: 4k
  - Average length of a gene: 1000 bp
- Eukaryotic genome (e.g., human)
  - Number of base pairs: 3G
  - Estimated number of genes: 30k – 35k
  - Estimated average length of a gene: 1000-2000 bp

- ~ 90% of *E. coli* genome are of coding regions.
- < 3% of human genome is believed to be coding regions

⇒ Genome size has no relationship with the number of genes!

RNA

- RNA has both the properties of DNA & protein
  - Similar to DNA, it can store & transfer info
  - Similar to protein, it can form complex 3D structure & perform some functions
- Nucleotide for RNA has of three parts:
  - Ribose Sugar (has an extra OH group at 2')
  - Phosphate (bound to 5' carbon)
  - Base (bound to 1' carbon)

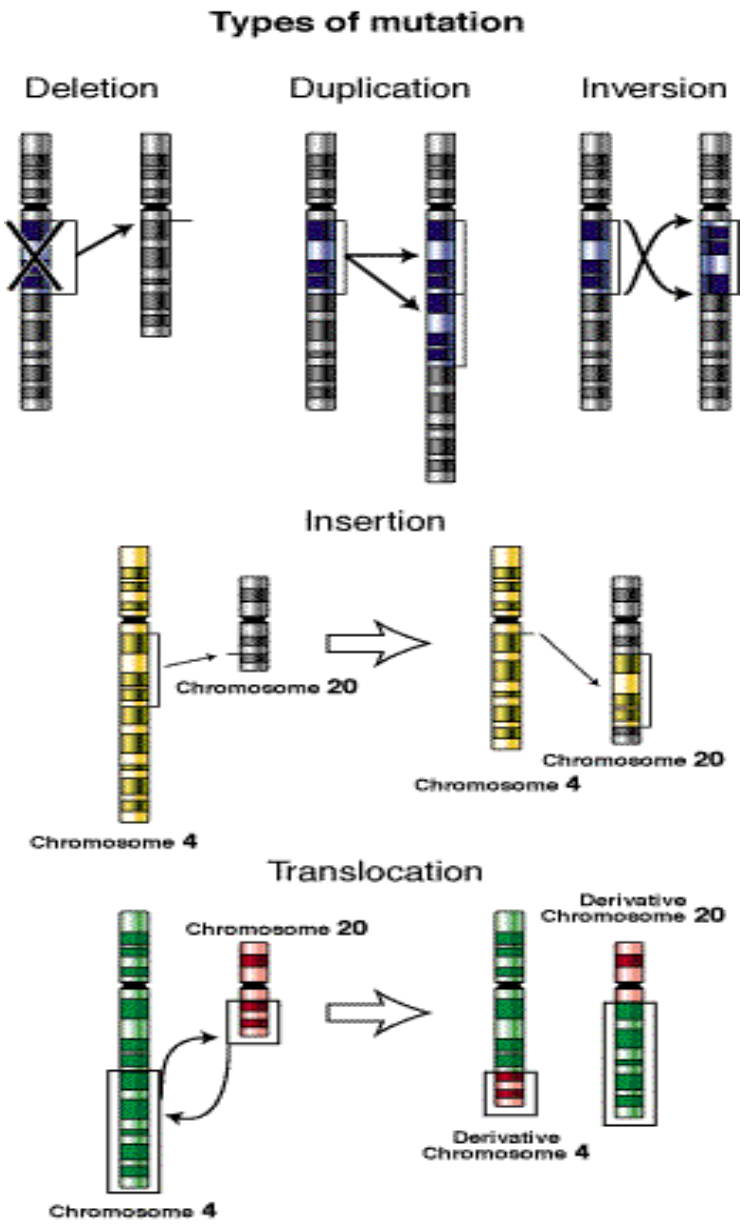


RNA vs DNA

- RNA is single stranded
- Nucleotides of RNA are similar to that of DNA, except that have an extra OH at position 2'
  - Due to this extra OH, it can form more hydrogen bonds than DNA
  - So RNA can form complex 3D structure
- RNA use the base U instead of T
  - U is chemically similar to T
  - In particular, U is also complementary to A

Mutation

- Sudden change of genome
- Basis of evolution
- Cause of cancer
- Can occur in DNA, RNA, & Protein



Introns and exons

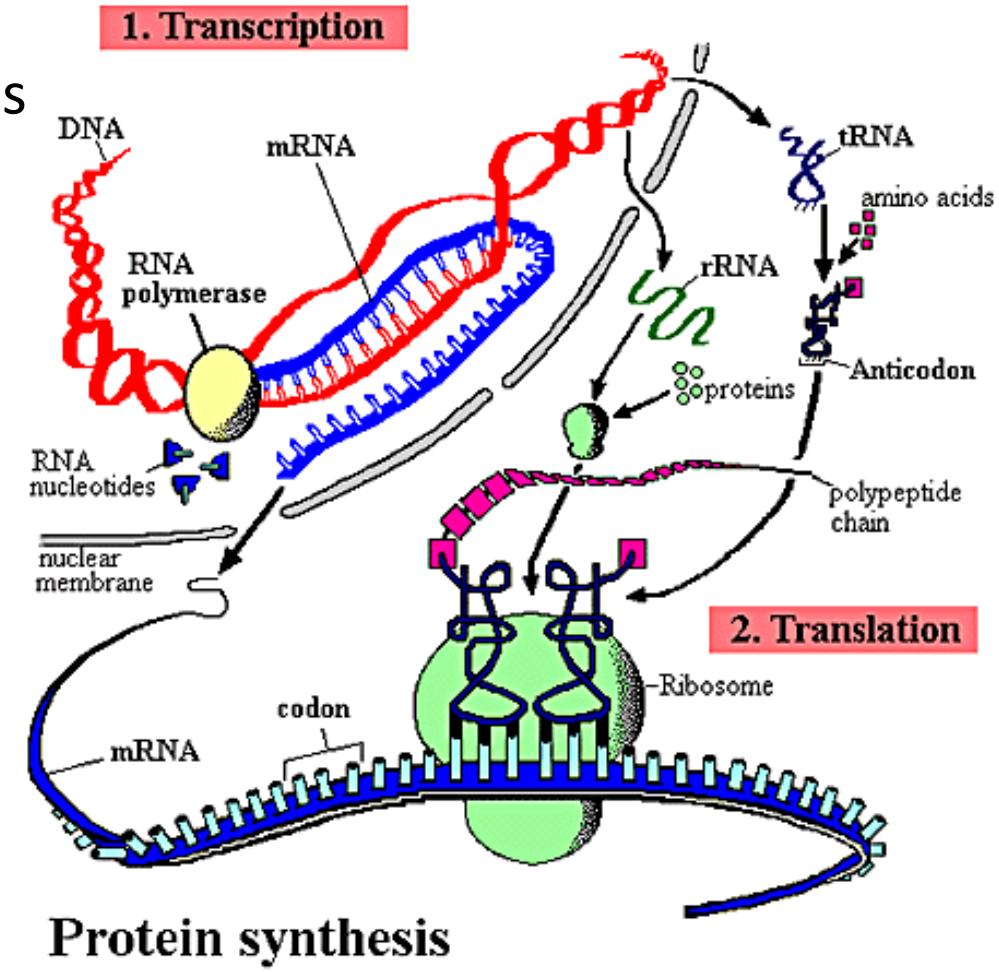
- Eukaryotic genes contain introns & exons
  - Introns are seq that are ultimately spliced out of mRNA
  - Introns normally satisfy GT-AG rule, viz. begin w/ GT & end w/ AG
  - Each gene can have many introns & each intron can have thousands bases
- Introns can be very long
- An extreme example is a gene associated with cystic fibrosis in human:
  - Length of 24 introns ~1Mb
  - Length of exons ~1kb

Central Dogma

- Gene expression consists of two steps
  - Transcription

    DNA → mRNA
  - Translation

    mRNA → Protein



The genetic code: codons code for amino acids

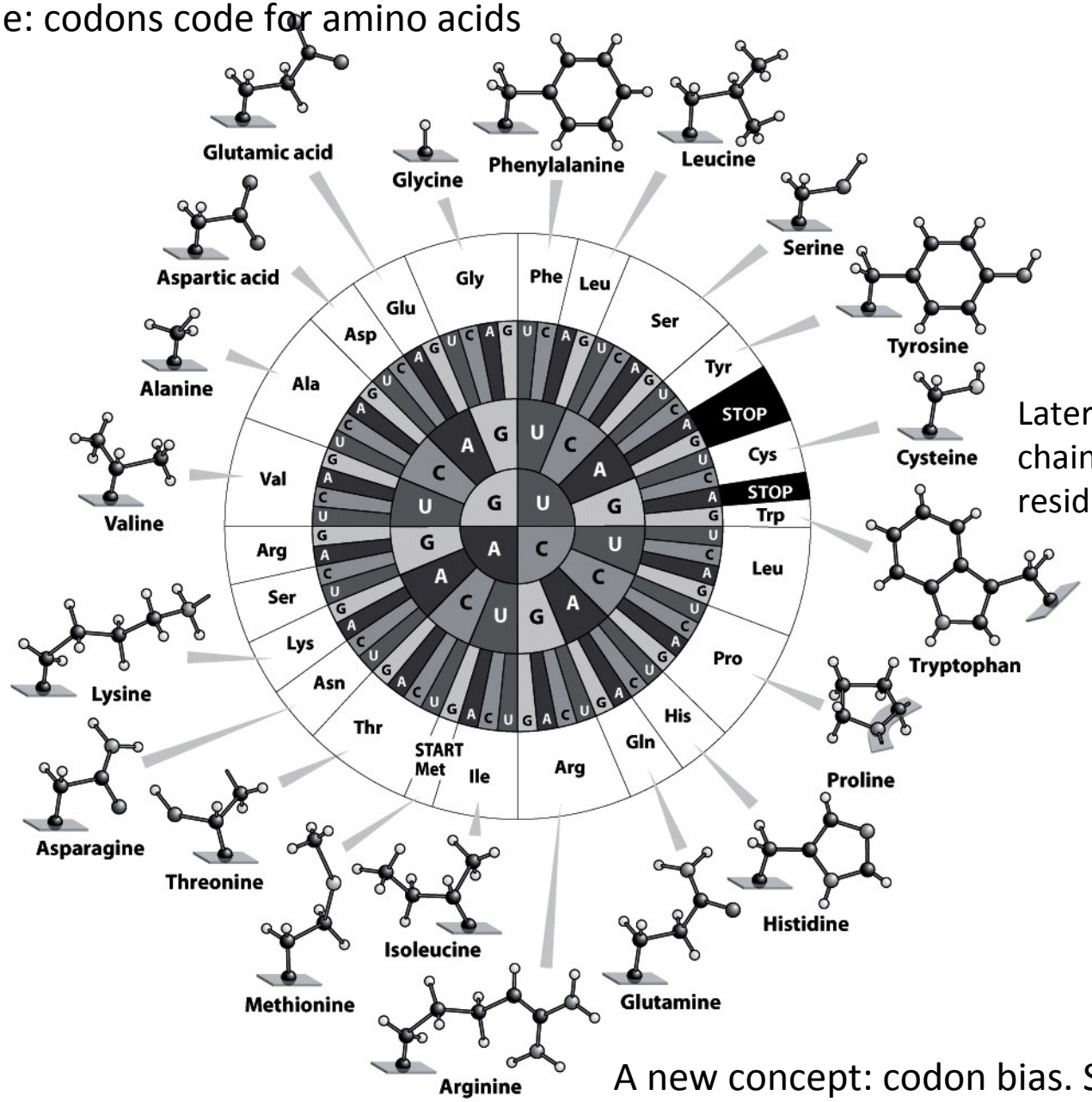


Figure 1.4 Physical Biology of the Cell (© Garland Science 2009)

Lateral chains: or residues

A new concept: codon bias. Slow/fast codon

THE CODON BIAS PROBLEM

SCALES AND INFORMATION CONTENT

organisms	genome	genes
virus	10-100,000	10-100s
bacteria	5 Mb	4,000
single euk. cell (yeast)	15Mb	6,000
simple animal	100Mb	15,000
Homo sapiens	3,000Mb	30-40,000

Examples of Computational Problems

- Physical and Genetic Maps
- Genome assembly
- Pairwise and Multiple Alignments
- Motif Detection/Discrimination/Classification
- Data Base Searches and Mining
- Phylogenetic Tree Reconstruction
- Gene Finding and Gene Parsing
- Protein Secondary Structure Prediction
- Protein Tertiary Structure Prediction
- Protein Function Prediction
- Comparative genomics and evolution
- DNA microarray analysis
- Molecular docking/Drug design
- Gene regulation/regulatory networks
- Systems biology

Machine Learning

- Extract information from the data automatically (inference) via a process of model fitting (learning from examples).
- Model Selection: Neural Networks, Hidden Markov Models, Stochastic Grammars, Bayesian Networks
- Model Fitting: Gradient Methods, Monte Carlo Methods,…
- Machine learning approaches are most useful in areas where there is a lot of data but little theory.

Intuitive, Bayesian Approach

- Look at ALL available data, background information, and hypothesis
- Use probabilities to express PRIOR knowledge
- Use probabilities for inference, model selection, model comparison, etc. by computing POSTERIOR distributions and deriving UNIQUE answers

Deduction and Inference

- DEDUCTION:

  If A => B and A is true,

  then B is true.

- INDUCTION:

  If A => B and B is true,

  then A is more plausible

Exercise, what is **abduction**?

Probability as Degree of Belief

- **Sum Rule**:

$$P(\text{non } A) = 1 - P(A)$$

- **Product Rule**:

$$P(A \text{ and } B) = P(A)\, P(B|A)$$

- **BayesTheorem**:

$$P(B|A) = P(A|B)P(B)/P(A)$$

- **Induction Form**:

$$P(M|D) = P(D|M)P(M)/P(D)$$

- **Equivalently**:

$$\log[P(M|D)] = \log[P(D|M)] + \log[P(M)] - \log[P(D)]$$

- **Recursive Form**:

$$P(M|D_1, D_2, \ldots, D_{n+1}) = P(D_{n+1}|M)\, P(I|D_1, \ldots, D_n) \,/\, P(D_{n+1}|D_1, \ldots, D_n)$$

The relevance of Bayes' theorem: see DILL & BROMBERG: EXAMPLE1.11 …BIOINFORMATIC CONTEXT

EXAMPLE 1.11 Applying Bayes' rule: Predicting protein properties. *Bayes' rule*, a combination of Equations (1.11) and (1.15), can help you compute hard-to-get probabilities from ones that are easier to get. Here's a toy example. Let's figure out a protein's structure from its amino acid sequence. From modern genomics, it is easy to learn protein sequences. It's harder to learn protein structures. Suppose you discover a new type of protein structure, call it a *helicoil h*. It's rare; you've searched 5000 proteins and found only 20 helicoils, so $p(h) = 0.004$. If you could discover some special amino acid *sequence feature*, call it sf, that predicts the h structure, you could search other genomes to find other helicoil proteins in nature. It's easier to turn this around. Rather than looking through 5000 sequences for patterns, you want to look at the 20 helicoil proteins for patterns. How do you compute $p(\text{sf} \mid h)$? You take the 20 given helicoils and find the fraction of them that have your sequence feature. If your sequence feature (say alternating glycine and lysine amino acids) appears in 19 out of the 20 helicoils, you have $p(\text{sf} \mid h) = 0.95$. You also need $p(\text{sf} \mid \bar{h})$, the fraction of non-helicoil proteins (let's call those $\bar{h}$) that have your sequence feature. Suppose you find $p(\text{sf} \mid \bar{h}) = 0.001$. Combining Equations (1.11) and (1.15) gives Bayes' rule for the probability you want:

$$p(h \mid \text{sf}) = \frac{p(\text{sf} \mid h)p(h)}{p(\text{sf})} = \frac{p(\text{sf} \mid h)p(h)}{p(\text{sf} \mid h)p(h) + p(\text{sf} \mid \bar{h})p(\bar{h})}$$

$$= \frac{(0.95)(0.004)}{(0.95)(0.004) + (0.001)(0.996)} = 0.79. \qquad (1.16)$$

In short, if a protein has the sf sequence, it will have the h structure about 80% of the time.

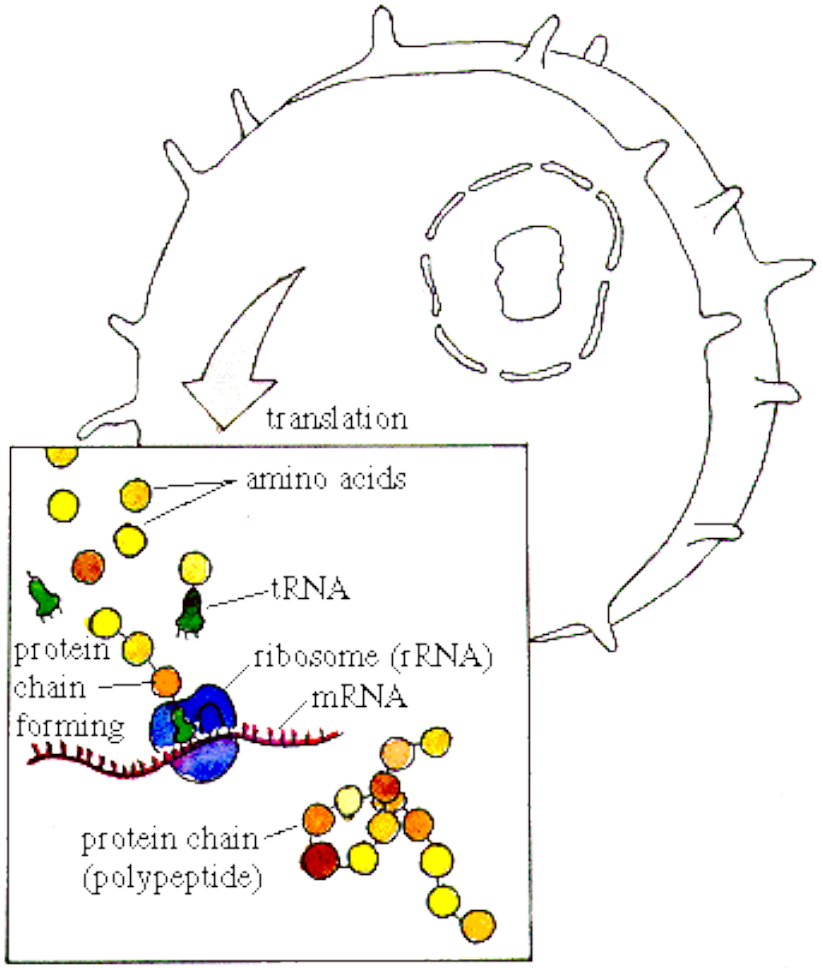
Learning

- MODEL FITTING AND MODEL COMPARISON
- MAXIMUM LIKELIHOOD AND MAXIMUM A POSTERIORI

$$\log p(M \mid D) = \log p(D \mid M) + \log p(M) - \log p(D)$$

RNA functions and structure

http://physwww.mcmaster.ca/~higgsp/RNA.htm

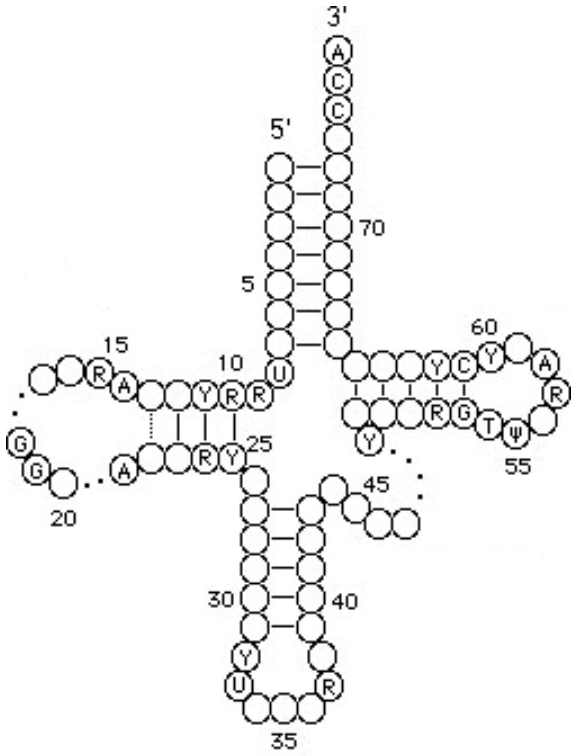


Roles of RNA

- mRNA (messenger)
- rRNA (ribosomal)
- tRNA (transfer)
- other ribonucleoproteins (e.g. spliceosome, signal recognition particle, ribonuclease P)
- viral genomes
- artificial ribozymes

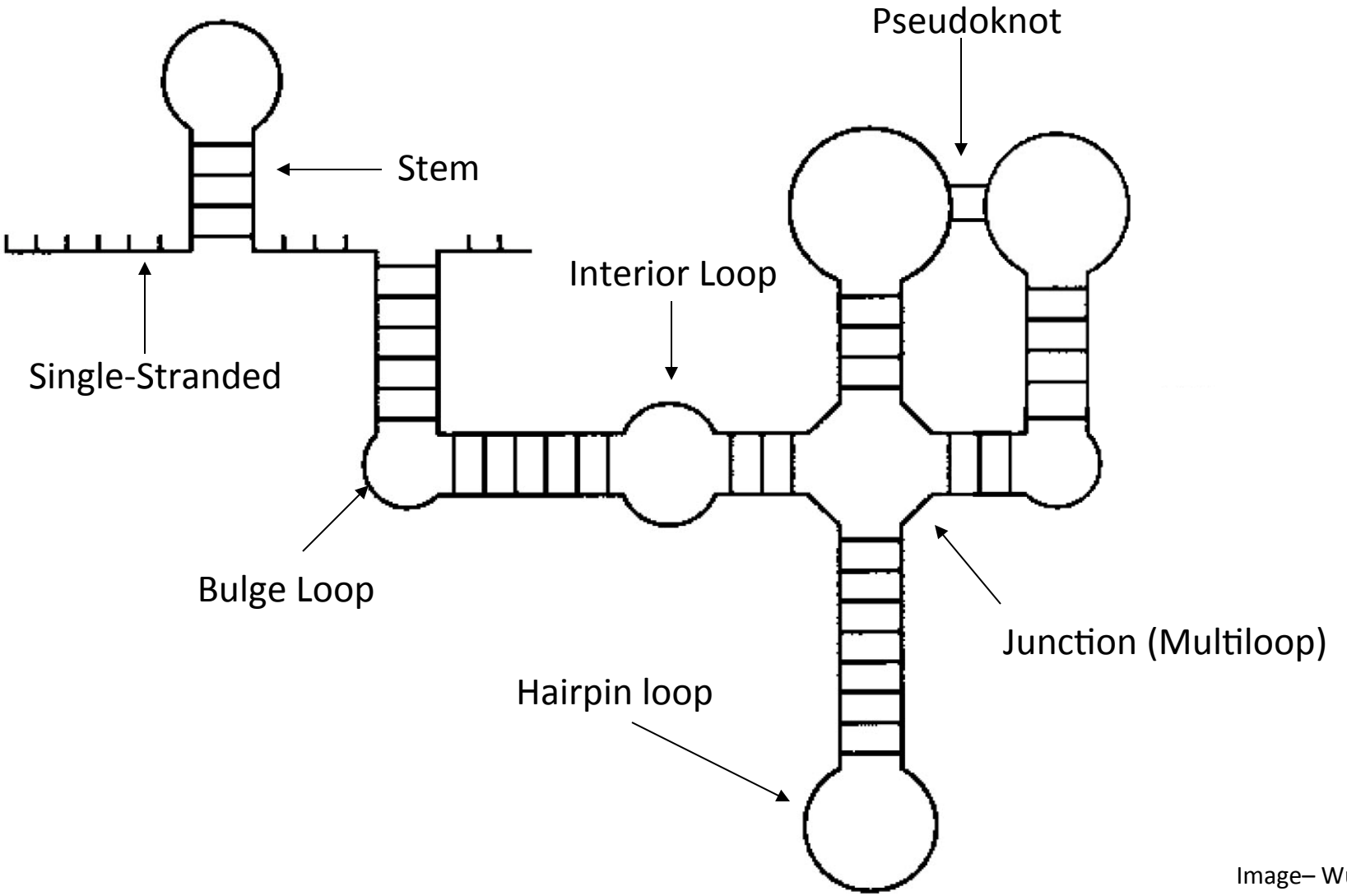
RNA Basics

- transfer RNA (tRNA)
- messenger RNA (mRNA)
- ribosomal RNA (rRNA)
- small interfering RNA (siRNA)
- micro RNA (miRNA)
- small nucleolar RNA (snoRNA)



http://www.genetics.wustl.edu/eddy/tRNAscan-SE/

RNA Secondary Structure



Image– Wuchty

RNA Structure and Evolution

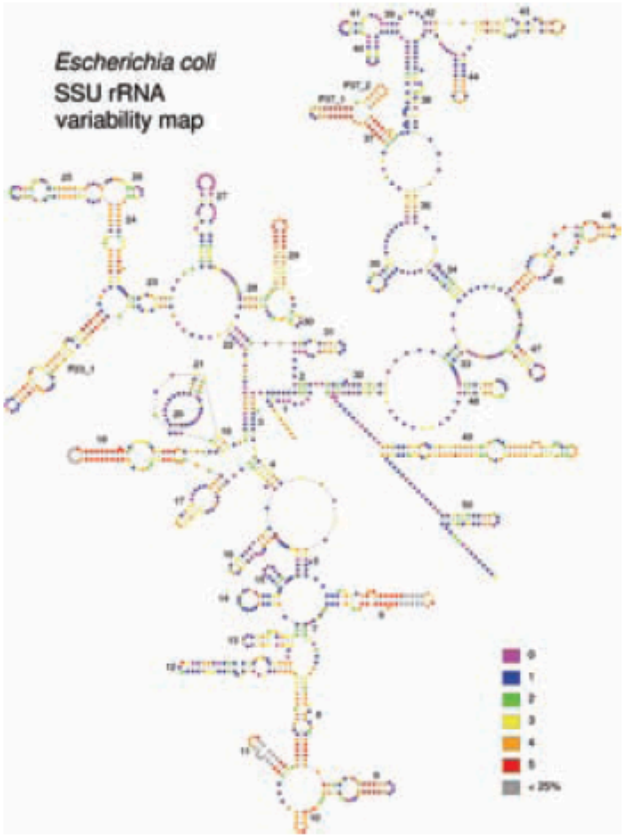


Plate 11.1 The secondary structure of SSU rRNA in *E. coli*. The color scheme shows the degree of variability of the sequence across the bacterial domain. Category 0 (purple) sites are completely conserved. Categories 1 to 5 range from very slowly evolving (blue) to rapidly evolving (red). The gray sites are present in less that 25% of the species considered, hence no measure of evolutionary rate was made. Reproduced with permission from the European Ribosomal RNA database http://oberon.fvms.ugent.be:8080/rRNA/index.html

EVOLUTION SEQUENCES STRUCTURES (see PBC chap. 18)

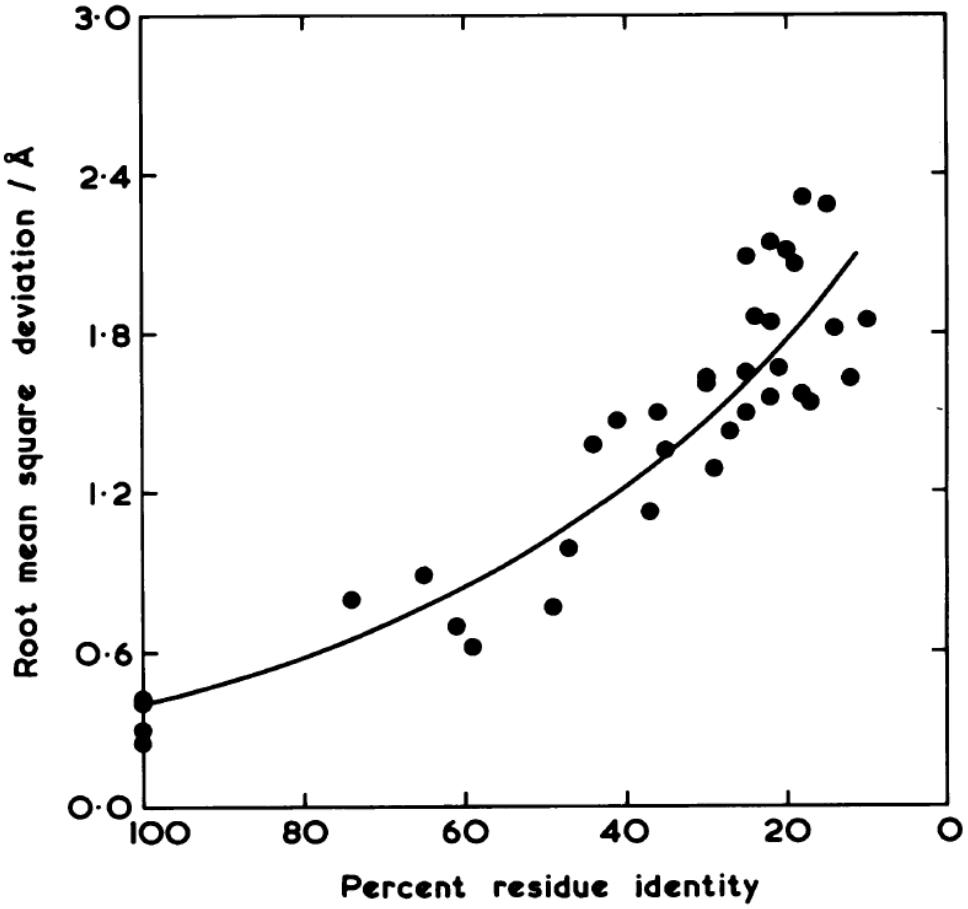


Fig. 2. The relation of residue identity and the r.m.s. deviation of the backbone atoms of the common cores of 32 pairs of homologous proteins (see Table II).

From: Chothia and Lesk

The EMBO Journal vol.5 no.4 pp.823–826, 1986

RNA in comparison to Proteins

Both have well defined 3d structures

RNA folding problem is easier because secondary structure separates from tertiary structure more easily - But it is still a complex problem.

RNA model has real parameters therefore you can say something about real molecules. RNA folding algorithm is simple enough to be able to do statistical physics. (cf. 27-mer lattice protein models).

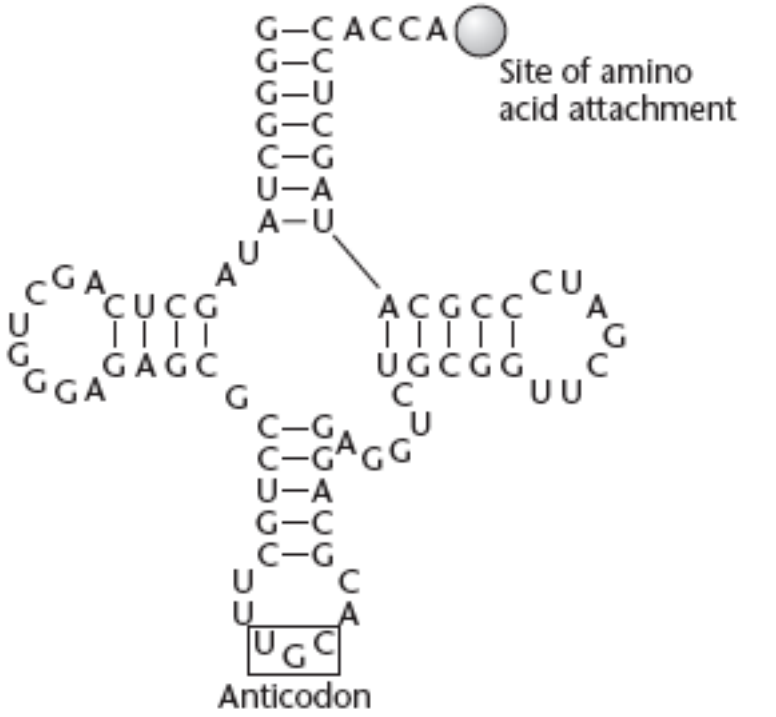
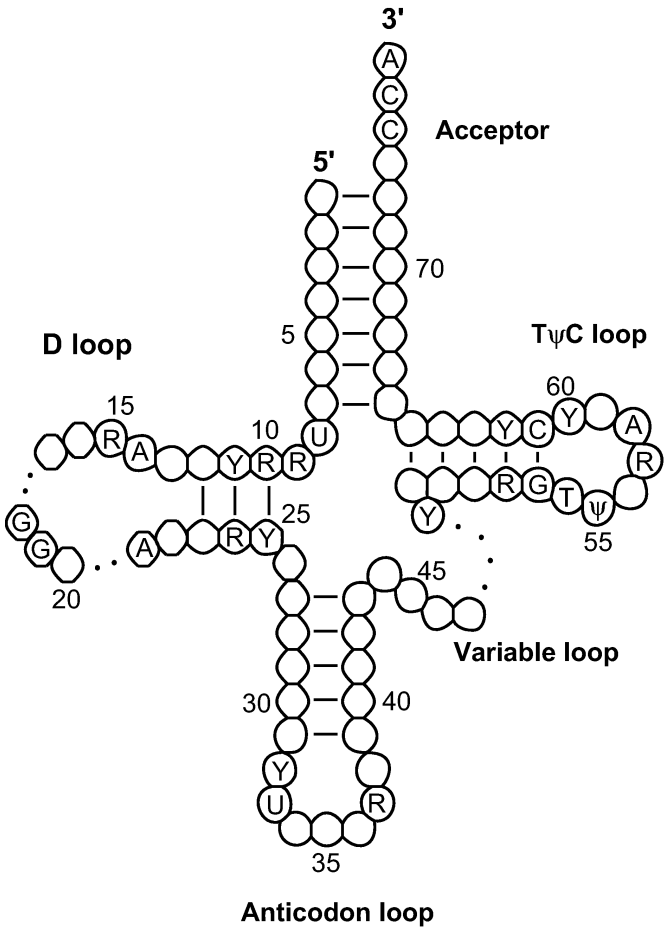
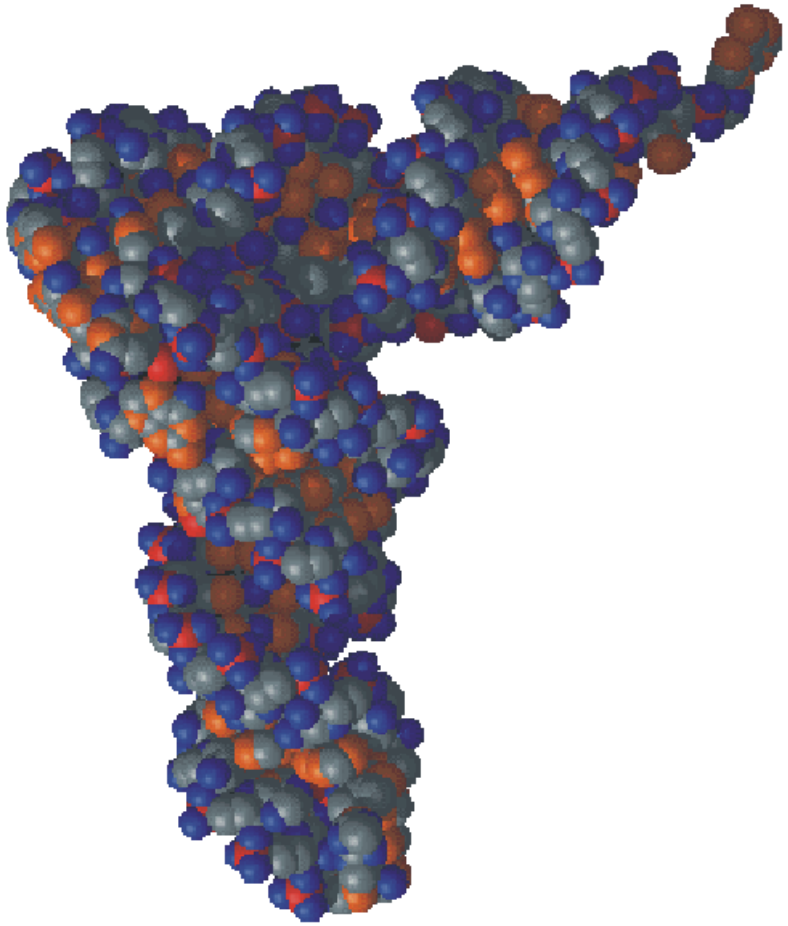


Fig. 2.4 Secondary structure of tRNA-Ala from *Escherichia coli* showing the anticodon position and the site of amino acid attachment.

Typical transfer RNA structure





RNA world

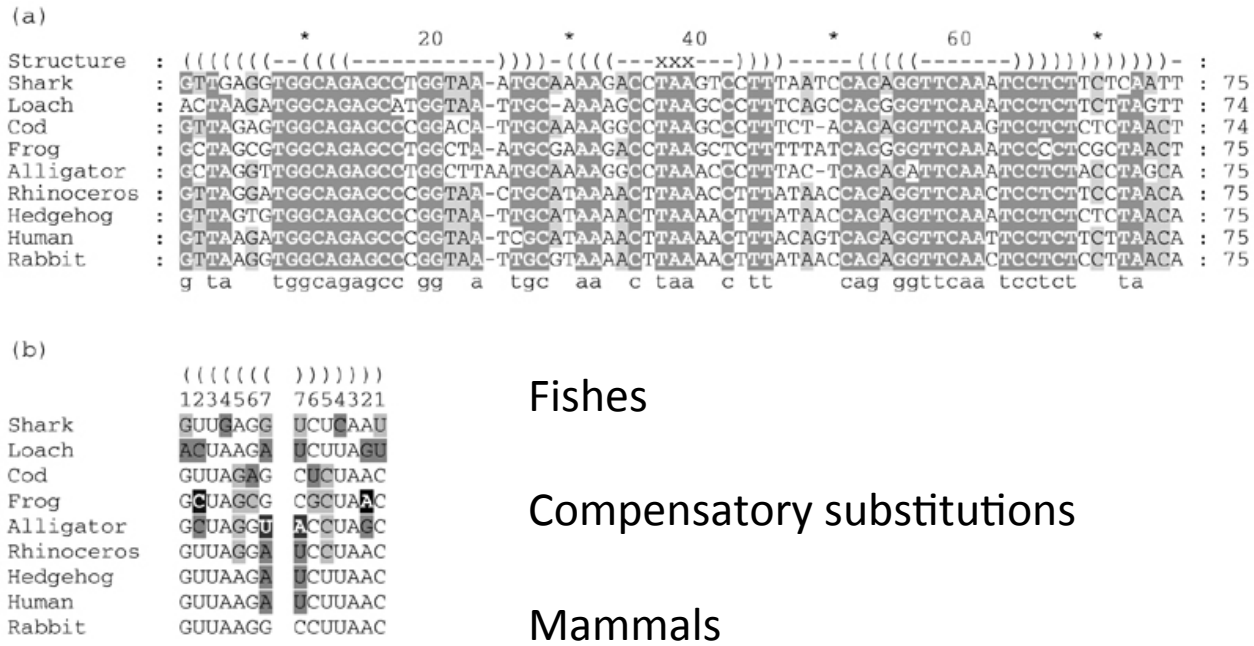
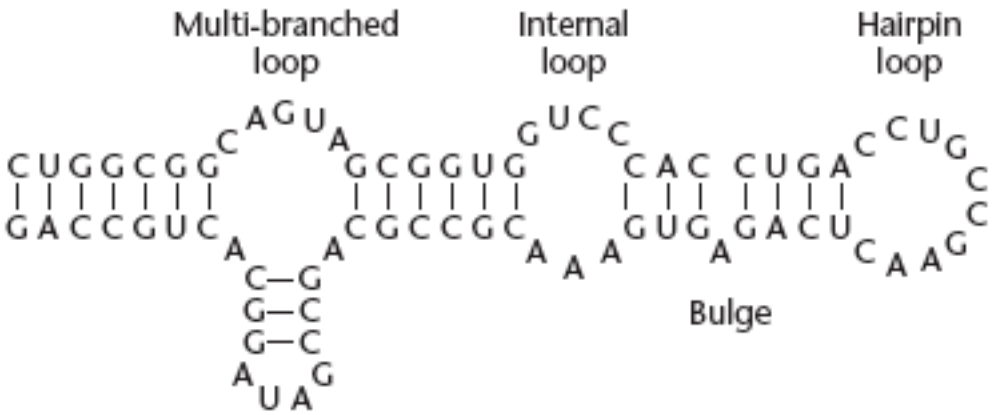


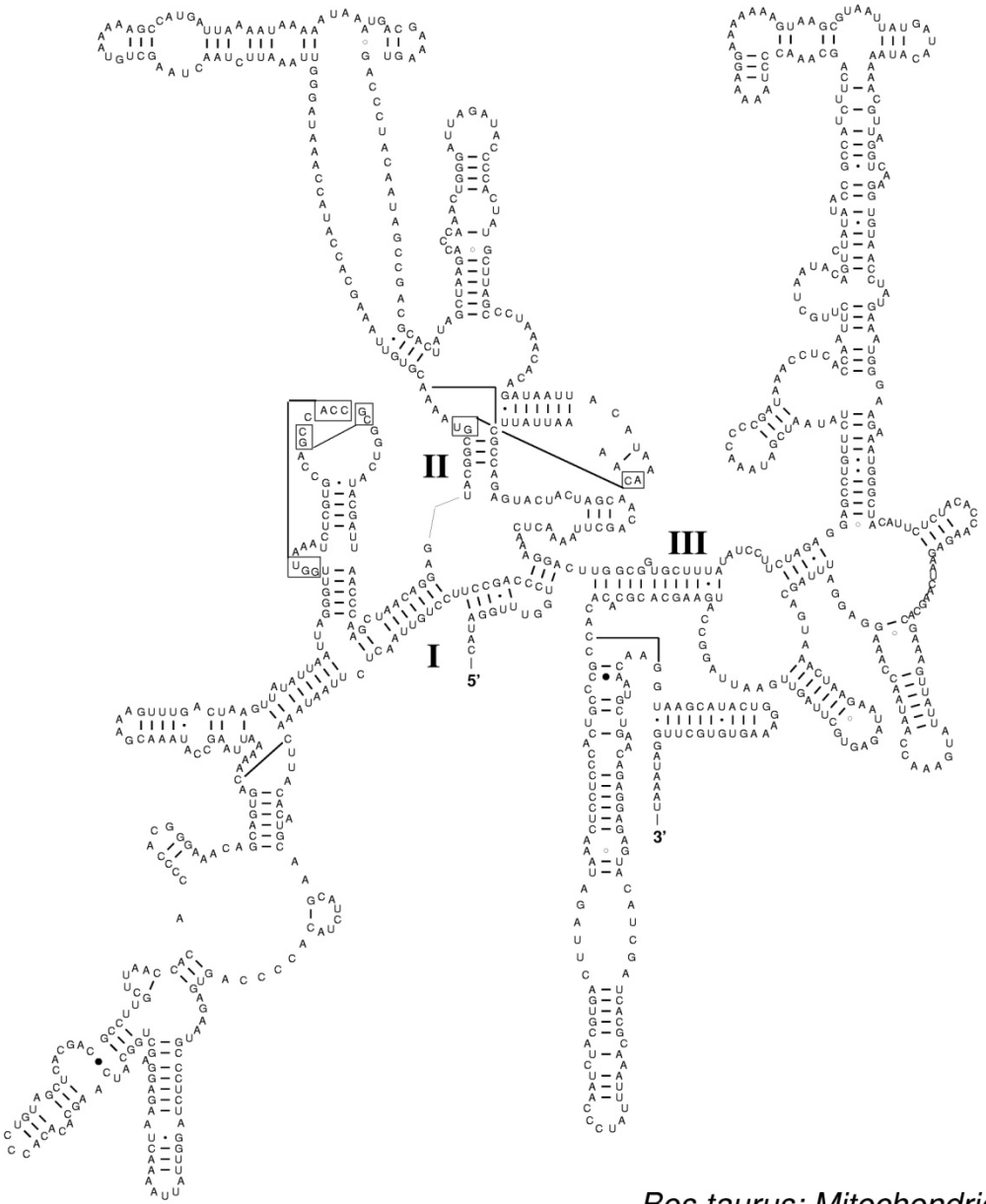
Fig. 11.1 (a) Alignment of tRNA-Leu genes from mitochondrial genomes, with conserved secondary structure illustrated using bracket notation. Gray-scale shading illustrates sequence conservation. (b) Alignment of the two halves of the aminoacyl acceptor stem of the tRNA, with shading added to illustrate compensatory substitutions.

http://www.rna.icmb.utexas.edu/
http://oberon.fvms.ugent.be:8080/rRNA/index.html (non attivo dal 2007)

Pairing: the comparative method (co-variation), Woese & Pace 1993



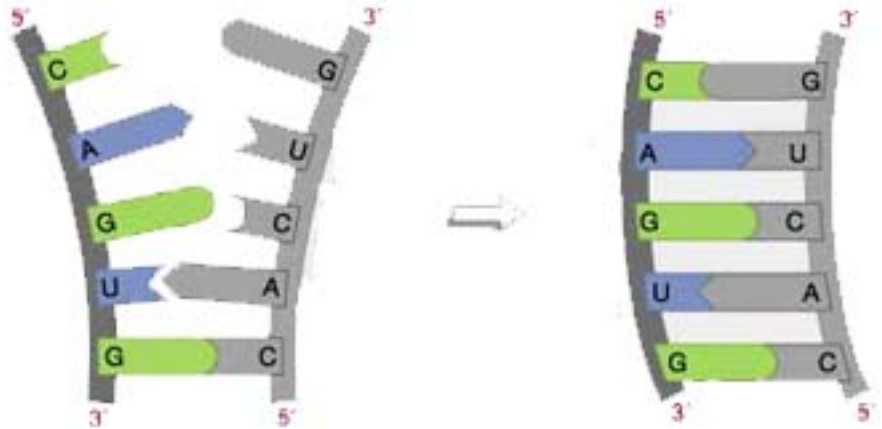
Secondary Structure: small subunit ribosomal RNA



Bos taurus: Mitochondrion

Sequence Alignment as a method to determine structure

- Bases pair in order to form backbones and determine the secondary structure
- Aligning bases based on their ability to pair with each other gives an algorithmic approach to determining the optimal structure



Dot Matrix Analysis

Sequence

Complementary sequence

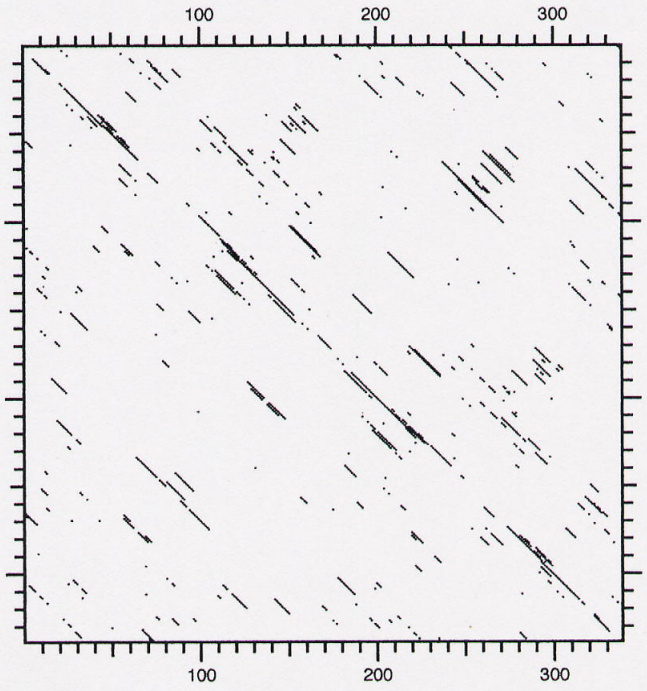


Figure 5.5. Dot matrix analysis of the potato tuber spindle viroid for RNA secondary structure using the MATRIX function of DNA Strider v. 1.2 on a Macintosh computer.

The repeated regions can potentially self-hybridize to form double strands

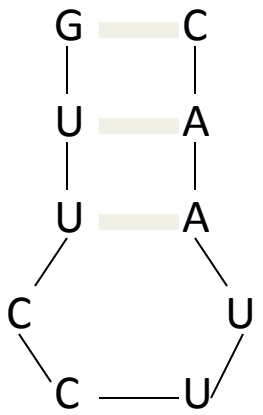
(usually ignore U—G pair in this simple analysis)

Introduction to RNA folding problem

Thermodynamics parameters are measured on real molecules.

Helix formation = hydrogen bonds + stacking

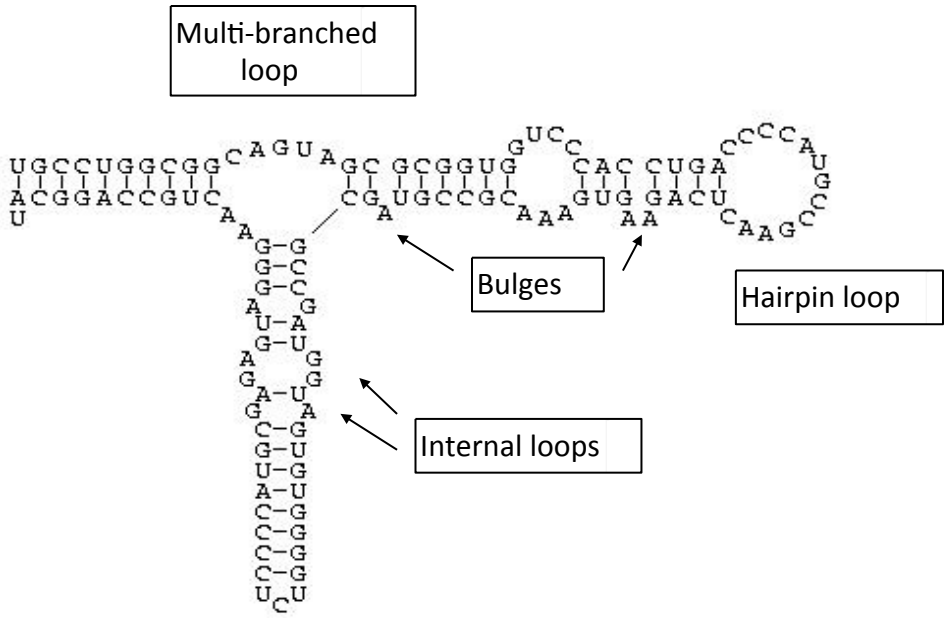
Entropic penalty for loop formation.



} ΔG = -2.1 kcal/mol

} ΔG = -1.2 kcal/mol

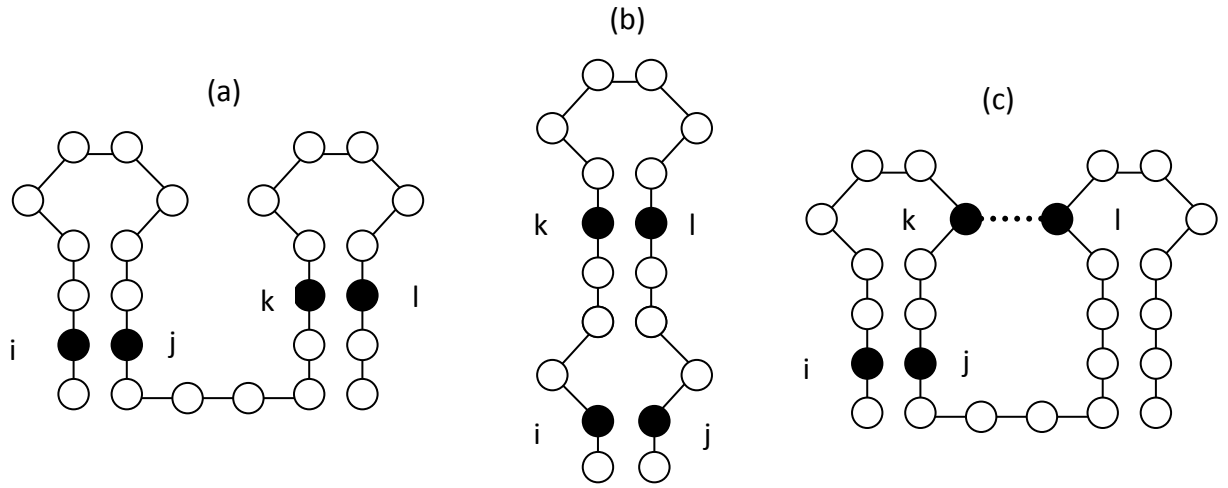
loop ΔG = + 4.5 kcal/mol



Sum up contributions of helices and loops over the whole structure.

Pairs i-j and k-l are compatible if (a) i < j < k < l , or (b) i < k < l < j .

(c) is called a pseudoknot: i < k < j < l . Usually not counted as secondary structure.



Bracket notation is used to represent structure:

```
a:    ((((....)))).((((....))))
b:    ((.((((....)))).))
```

Basic problem: Want an algorithm that considers every allowed secondary structure for a given sequence and finds the lowest energy state.

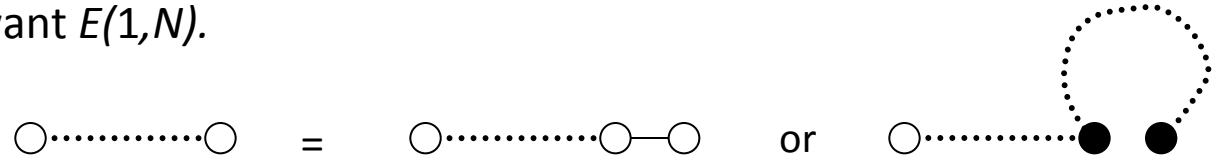
Simplest case: find structure which maximizes number of base pairs.

Let $\varepsilon_{\iota\varphi} = -1$ if bases can pair and +∞ if not.

Ignore loop contributions.

E(i,j) = energy of min energy structure for chain segment from i to j.

We want *E(1,N).*



Algorithms that work by recursion relations like this are called dynamic programming.

The algorithm is $O(N^3)$ although the number of structures increases exponentially with N.

Also need to do backtracking to work out the minimum energy structure:

Set *B(i,j) = k* if *j* is paired with *k*, or 0 if unpaired.