# THE BIOLOGICAL SETTING

Andrea Giansanti

Dipartimento di Fisica, Sapienza Università di Roma

Andrea.Giansanti@roma1.infn.it

CB_23_24 Lecture  N.4
Rome 3rd OCT 2023

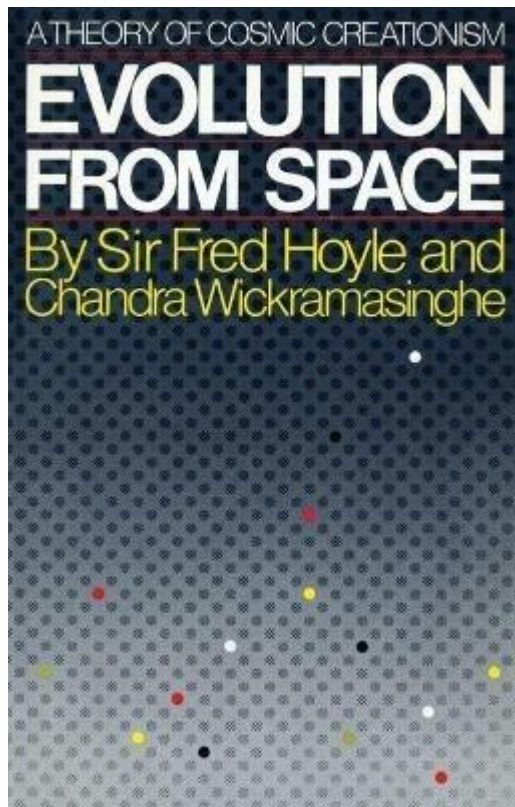DIPARTIMENTO DI FISICA

SAPIENZA
UNIVERSITÀ DI ROMA

# outline

- chemical cosmic evolution

- biomolecules

- evolutionary thinking

- central dogma of molecular biology

- genomic revolution

- time scales of evolution: phylogeny
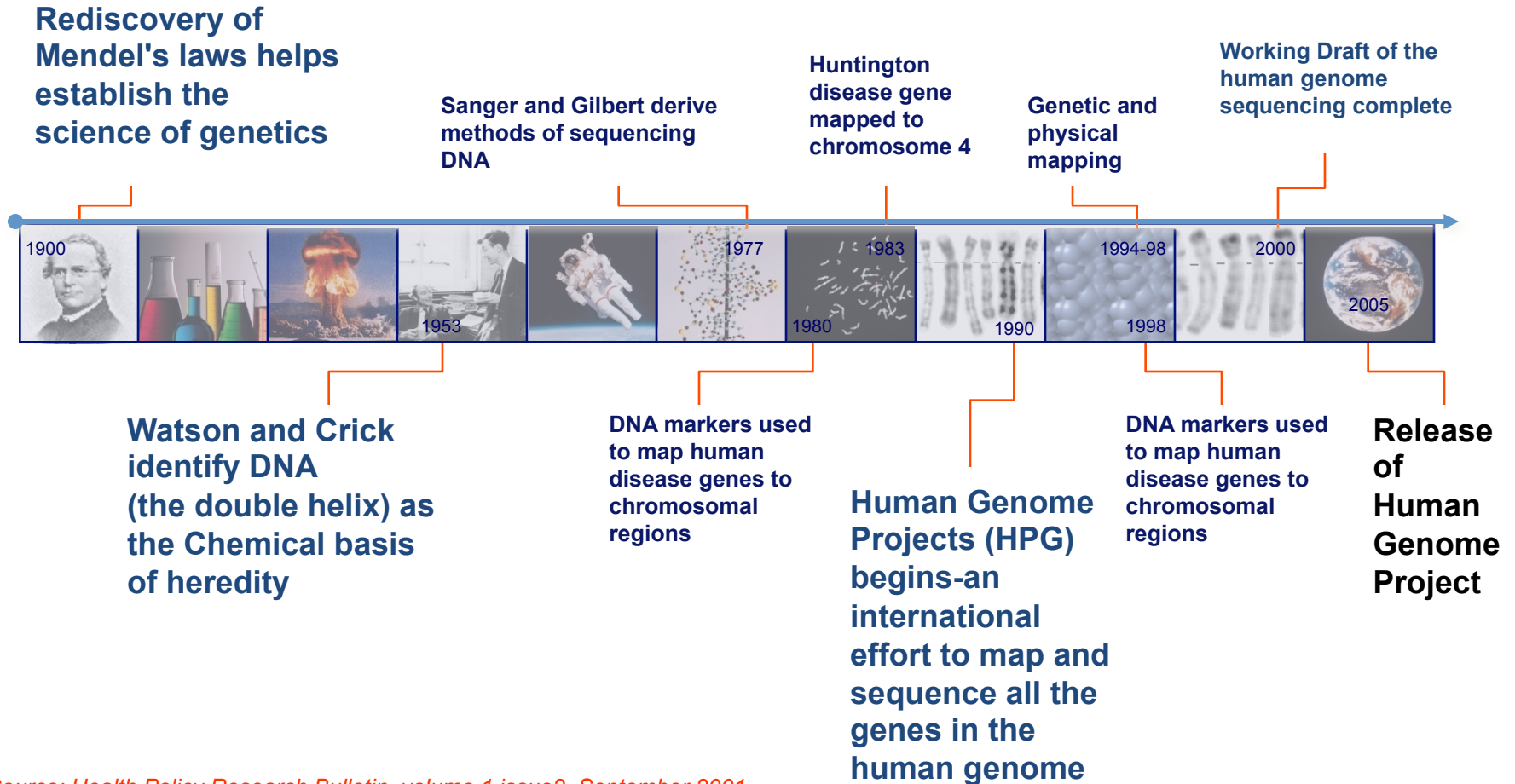
**we have touched in lecture n. 4...**

- **Basic biology and computational biophysics, at the molecular level**:
- semantics: biological physics/physical biology; biologically inspired physics/biophysics
- evolutionary distances (slides CB_21_22_L1-4.pdf, integrate with further self-study)
- what is life (see HA 3.1 and Weinberg_chap1.pdf).
- genomes and the genetic code: Central dogma of molecular biology, informational biomolecules: nucleic acids & proteins: peptide bonds formation and planarity, protein synthesis on the ribosome, genetic code and its degeneration (codon bias) (HA 2.2, 2.3)
- what is darwinian evolution (HA 3.1,3.2, 3.3 (mutations), 3.4 (coalescence) 3.6 (neutral evolution and adaptation, codon bias))
- the space of biological sequences as the archive of evolution (molecules as documents of evolutionary history, see: ZuckerkandlPauling1965
- molecular evolution of nucleotide sequences
- essential genes, homologous/paralogous genes
- evolutionary distances vs. sequence identity

- SEQUENCE->STRUCTURE/UNSTRUCTURE->DYNAMICS-FUNCTIONS

# Cosmochemistry

- pioneers: Fred Hoyle & Vikramasinghe
- Cecilia Ceccarelli Grenoble (see nice seminar organized by R. Schneider: https://www.youtube.com/watch?v=ylJNUoyImVA )
- Giovanna Tinetti UCL (see Fermi Lectures in our dept)

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

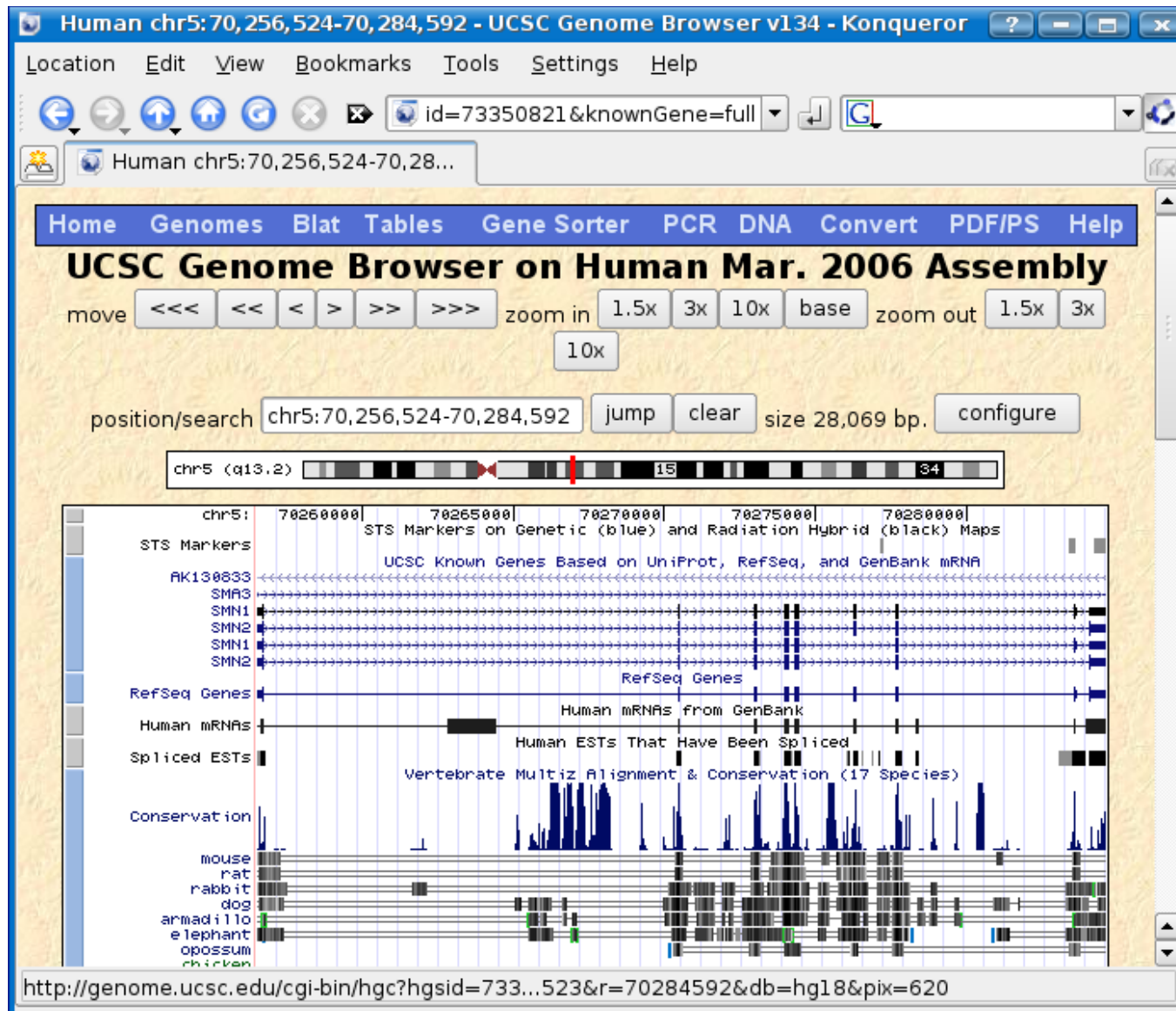**Rediscovery of Mendel's laws helps establish the science of genetics**

**Sanger and Gilbert derive methods of sequencing DNA**

**Huntington disease gene mapped to chromosome 4**

**Genetic and physical mapping**

**Working Draft of the human genome sequencing complete**

1900 · 1953 · 1977 · 1980 · 1983 · 1990 · 1994-98 · 1998 · 2000 · 2005

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

**DNA markers used to map human disease genes to chromosomal regions**

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

**DNA markers used to map human disease genes to chromosomal regions**

**Release of Human Genome Project**

# 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

# The genome browser
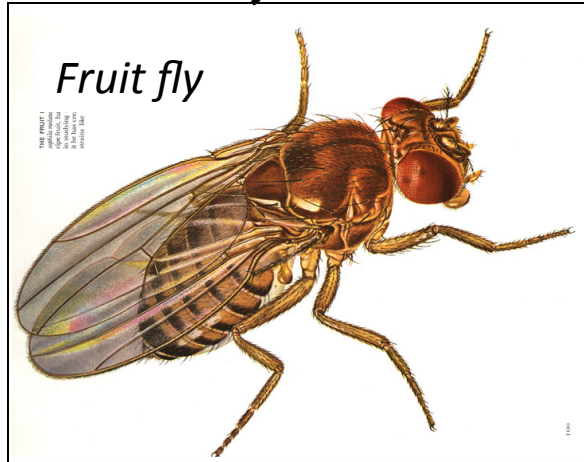
Link

# Number of protein coding genes

**20,210**

*Mouse*
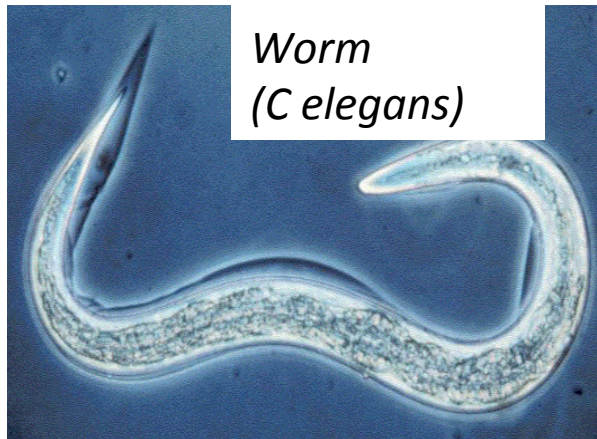
**13,601**

*Fruit fly*

**20,568**

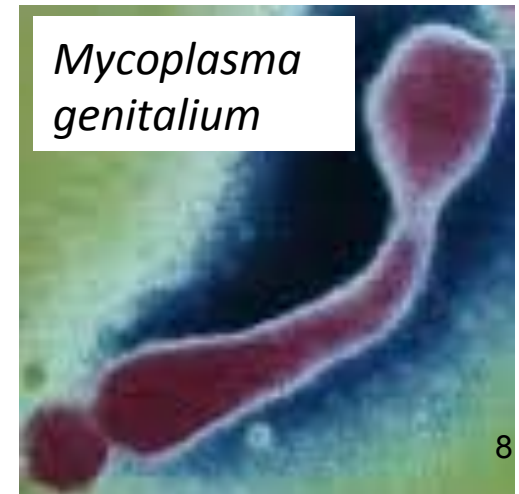*Mustered (Arabidopsis)*

**19,735**

*Worm (C elegans)*

**5,616**

*Yeast (S Cerevisiae)*
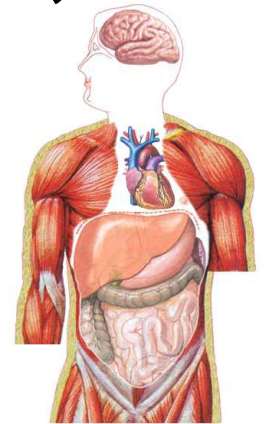
**482**

*Mycoplasma genitalium*

# 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

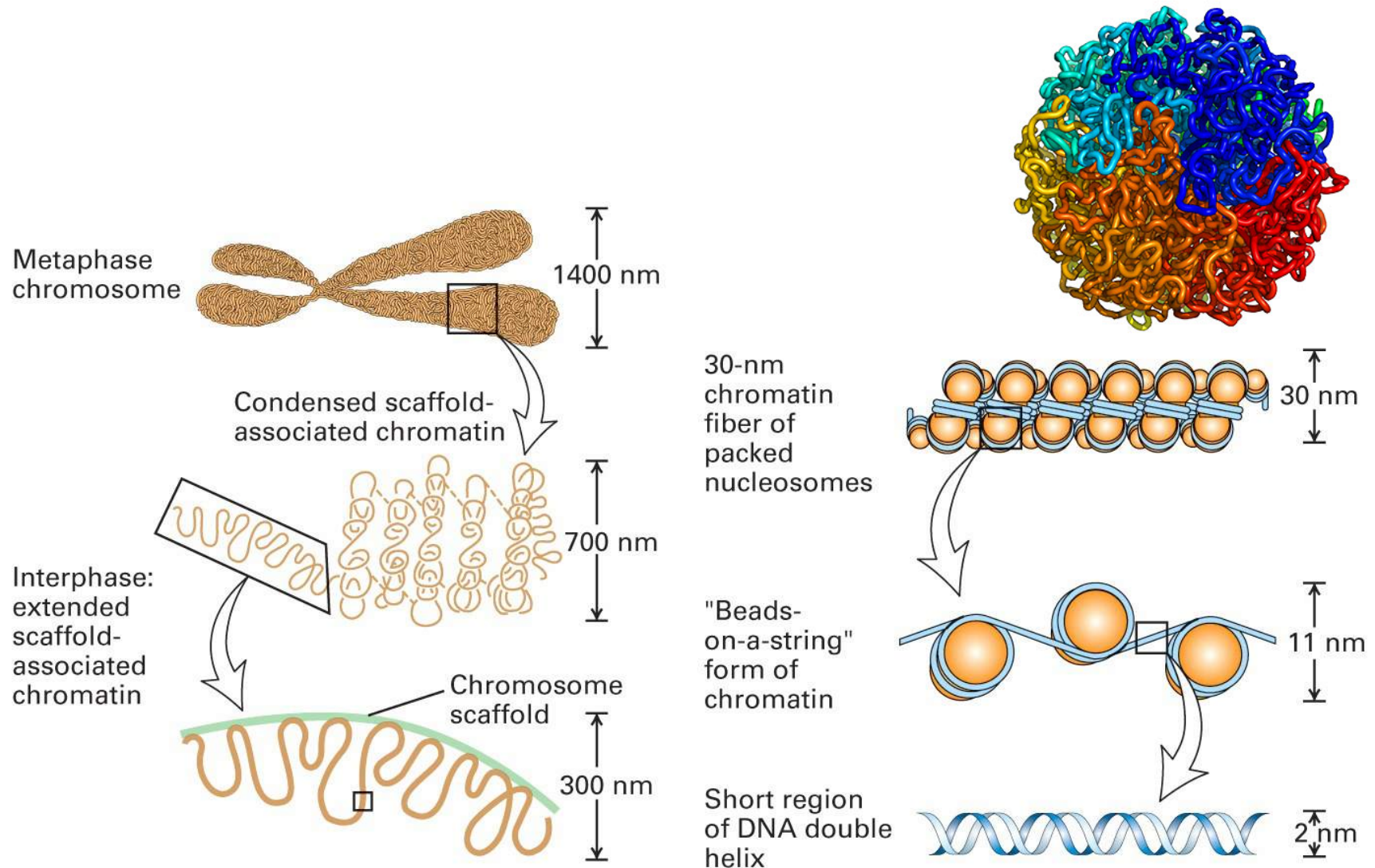- Combinatorial interactions among genes and products

21,710


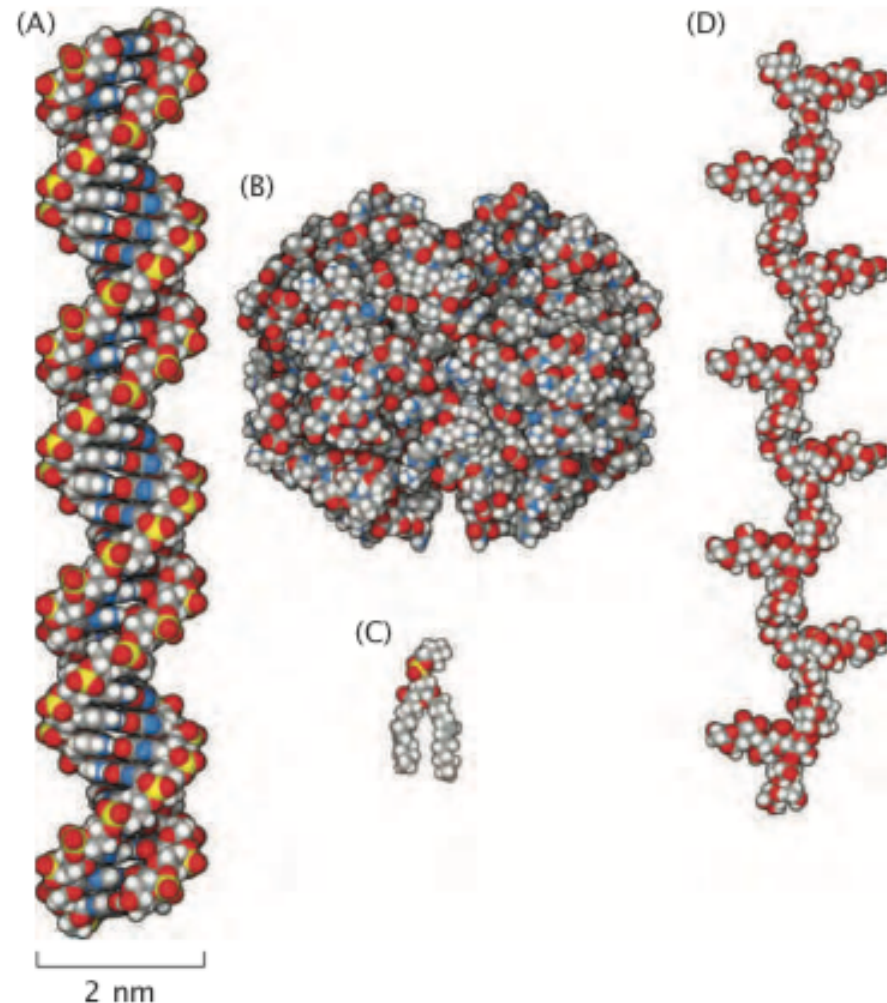
19,735

# The hierarchical structure of the genome

Metaphase chromosome — 1400 nm

Condensed scaffold-associated chromatin — 700 nm

Interphase: extended scaffold-associated chromatin

Chromosome scaffold — 300 nm

30-nm chromatin fiber of packed nucleosomes — 30 nm

"Beads-on-a-string" form of chromatin — 11 nm

Short region of DNA double helix — 2 nm

# 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)

# Biomolecular stuff: nucleic acids, proteins, lipids, sugars from R.Phillips et. al Physical biology of the Cell. 2nd ed.



**Figure 1.1:** Atomic-level structural representation of members of each of the major classes of macromolecules, all drawn at the same scale. Nitrogen is colored in blue, oxygen in red, phosphorus in yellow, carbon in gray, and hydrogen in white. (A) Atomic structure of a small fragment of the nucleic acid DNA in the B form, (B) atomic structure of the oxygen-carrying protein hemoglobin (PDB 1hho), (C) phosphatidylcholine lipid molecule from a cell membrane, and (D) branched complex carbohydrate (M41 capsular polysaccharide) from the surface of the bacterium *Escherichia coli* (PDB 1cap). (Illustrations courtesy of D. Goodsell.)

(A)

(B)

(C)

(D)

2 nm

# The central dogma (sic) of molecular biology
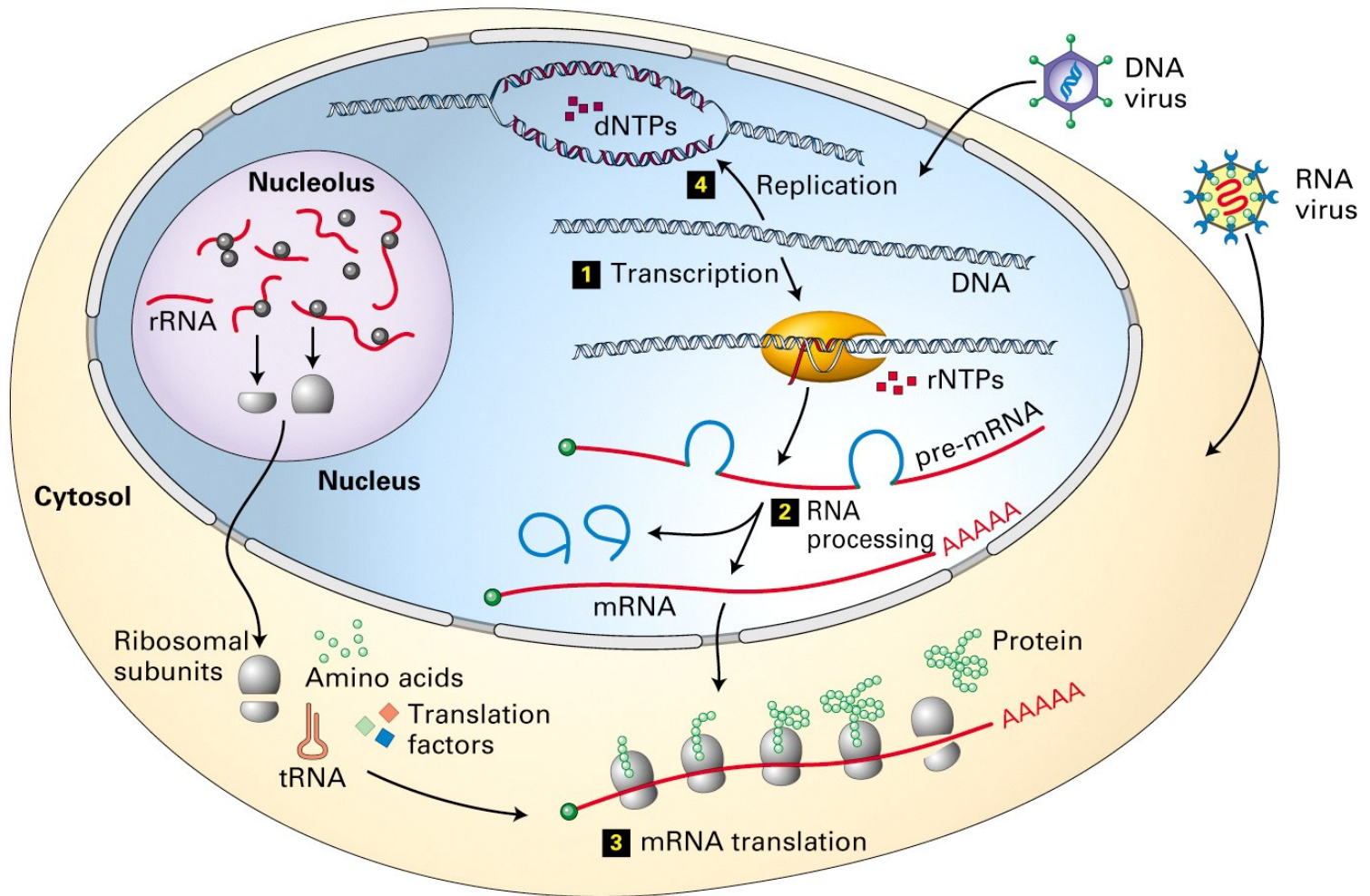
$$DNA \Rightarrow mRNA \Rightarrow Protein$$

RNA polymerase ( an enzyme, a protein) *transcribes* a segment of DNA to a complementary messenger RNA
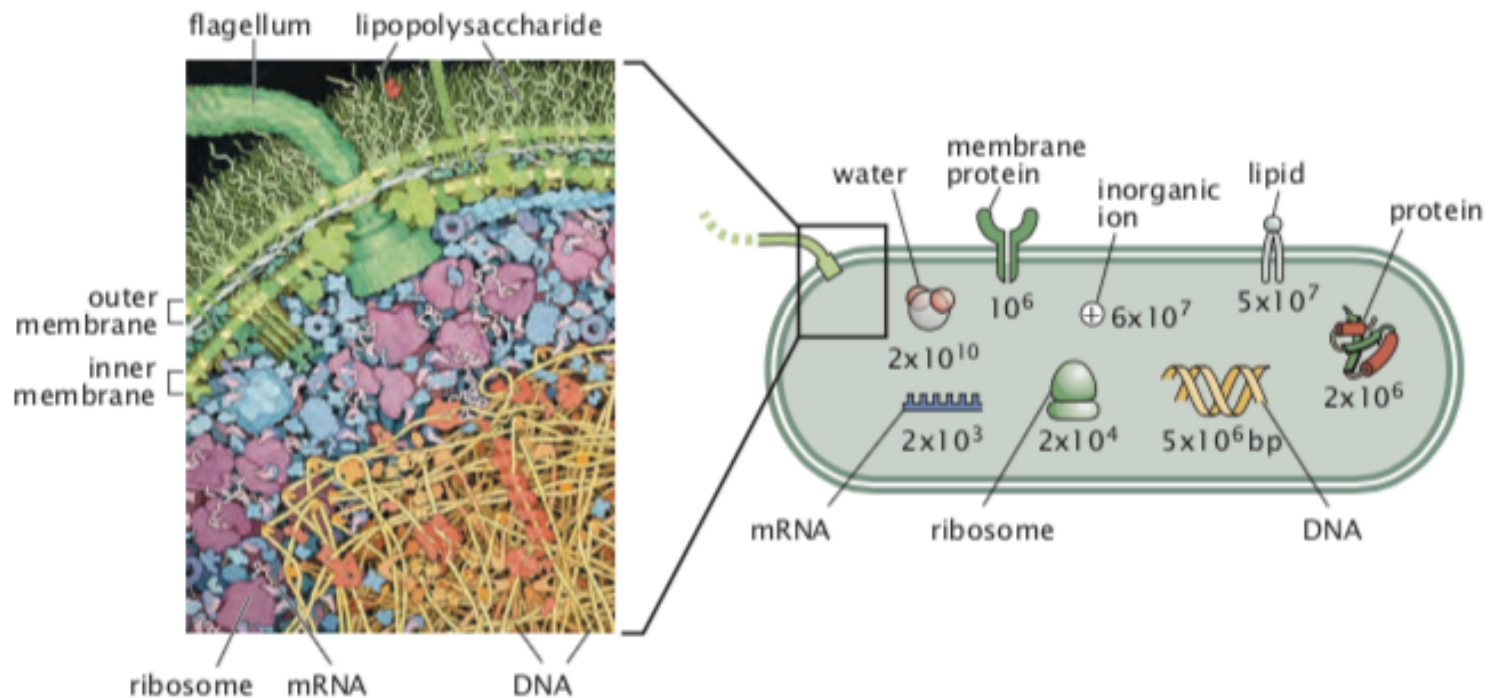
Primary messenger RNA is processed to mature mRNA

Mature mRNA is *translated* into protein by a *ribosome*

(a complex of proteins and rRNA)

# The Central Dogma: a cellular context

**Figure 2.4:** Molecular contents of the bacterium *E. coli*. The illustration on the left shows the crowded cytoplasm of the bacterial cell. The cartoon on the right shows an order-of-magnitude molecular census of the *E. coli* bacterium with the approximate number of different molecules in *E. coli*. (Illustration of the cellular interior courtesy of D. Goodsell.)
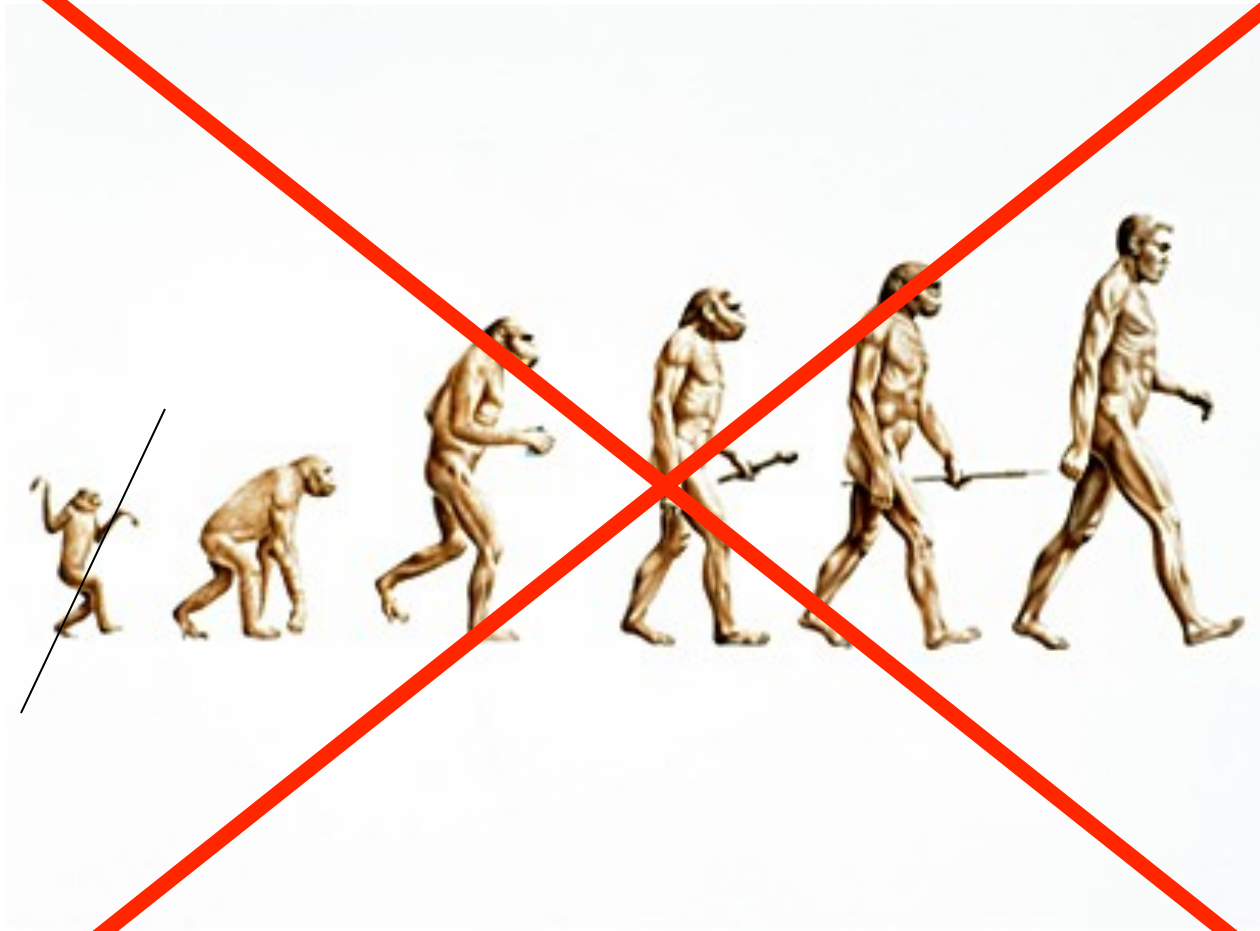
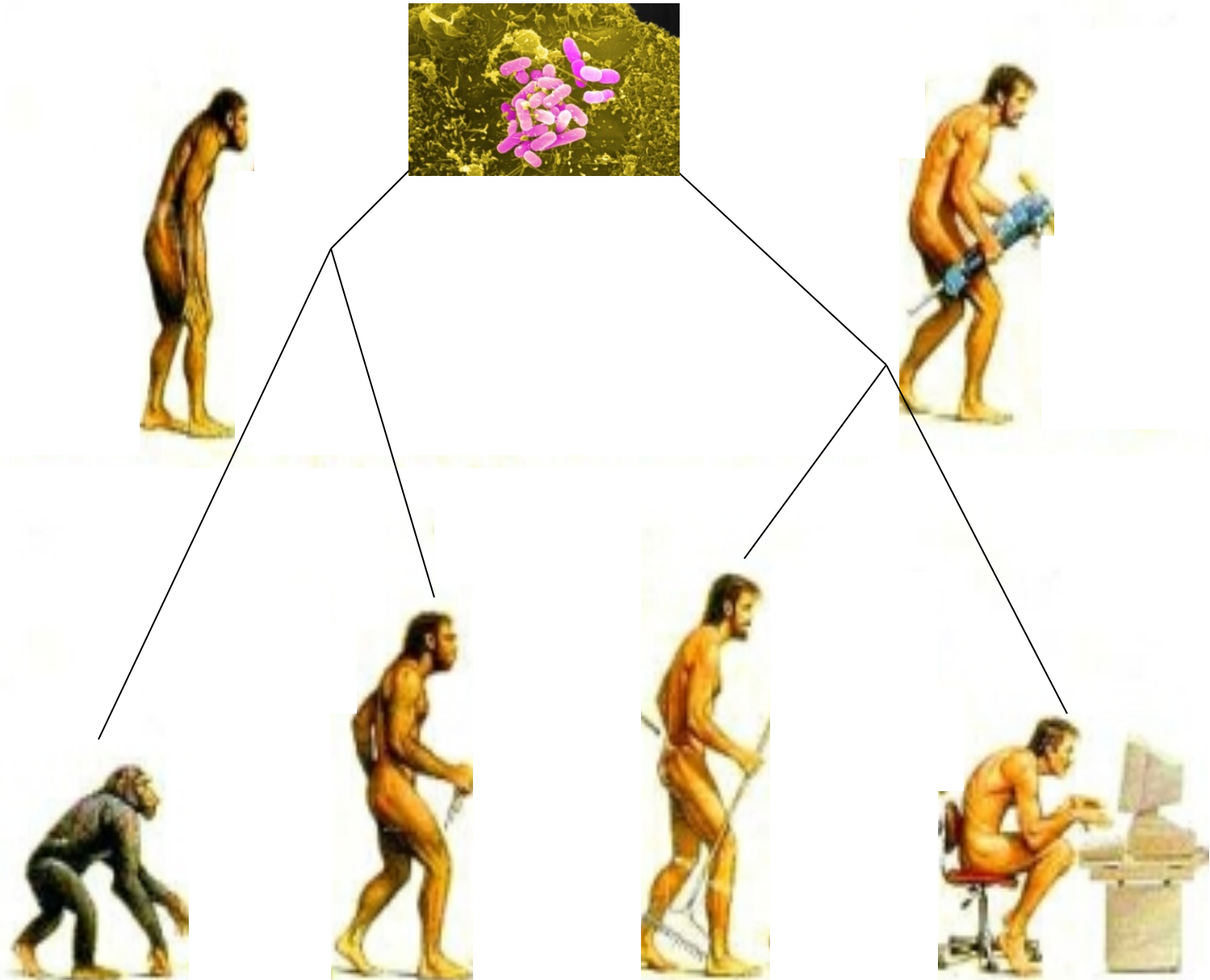# doing biology by the numbers
## Rob Phillips, Ron Milo

**Table 2.1:** Observed macromolecular census of an *E. coli* cell. (Data from F. C. Neidhardt et al., Physiology of the Bacterial Cell, Sinauer Associates, 1990 and M. Schaechter et al., Microbe, ASM Press, 2006.)

| Substance | % of total dry weight | Number of molecules |
|---|---|---|
| **Macromolecules** | | |
| Protein | 55.0 | $2.4 \times 10^6$ |
| RNA | 20.4 | |
|   23S RNA | 10.6 | 19,000 |
|   16S RNA | 5.5 | 19,000 |
|   5S RNA | 0.4 | 19,000 |
|   Transfer RNA (4S) | 2.9 | 200,000 |
|   Messenger RNA | 0.8 | 1,400 |
| Phospholipid | 9.1 | $22 \times 10^6$ |
| Lipopolysaccharide (outer membrane) | 3.4 | $1.2 \times 10^6$ |
| DNA | 3.1 | 2 |
| Murein (cell wall) | 2.5 | 1 |
| Glycogen (sugar storage) | 2.5 | 4,360 |
| **Total macromolecules** | **96.1** | |
| **Small molecules** | | |
| Metabolites, building blocks, etc. | 2.9 | |
| Inorganic ions | 1.0 | |
| **Total small molecules** | **3.9** | |

# Naïve Linear Evolution ?

# Complex view of evolution tree
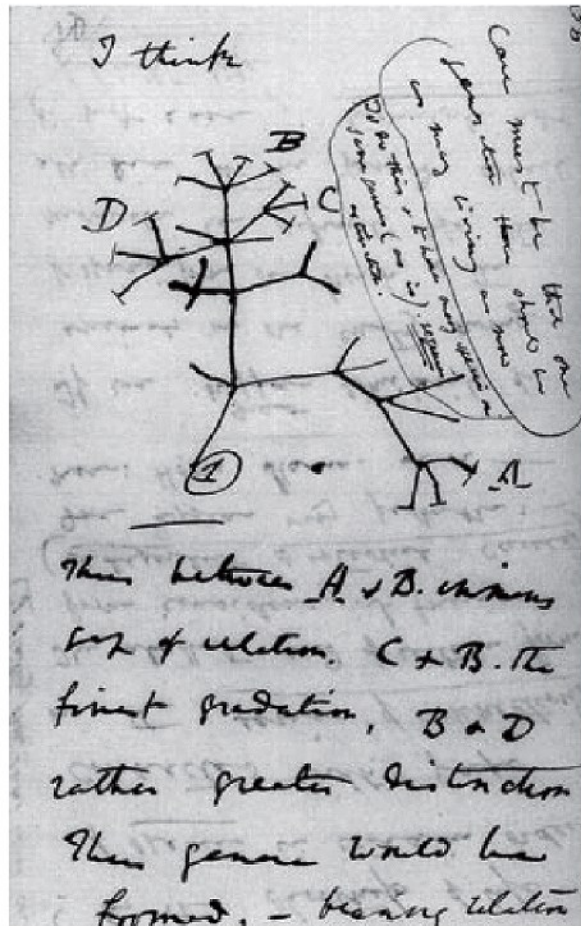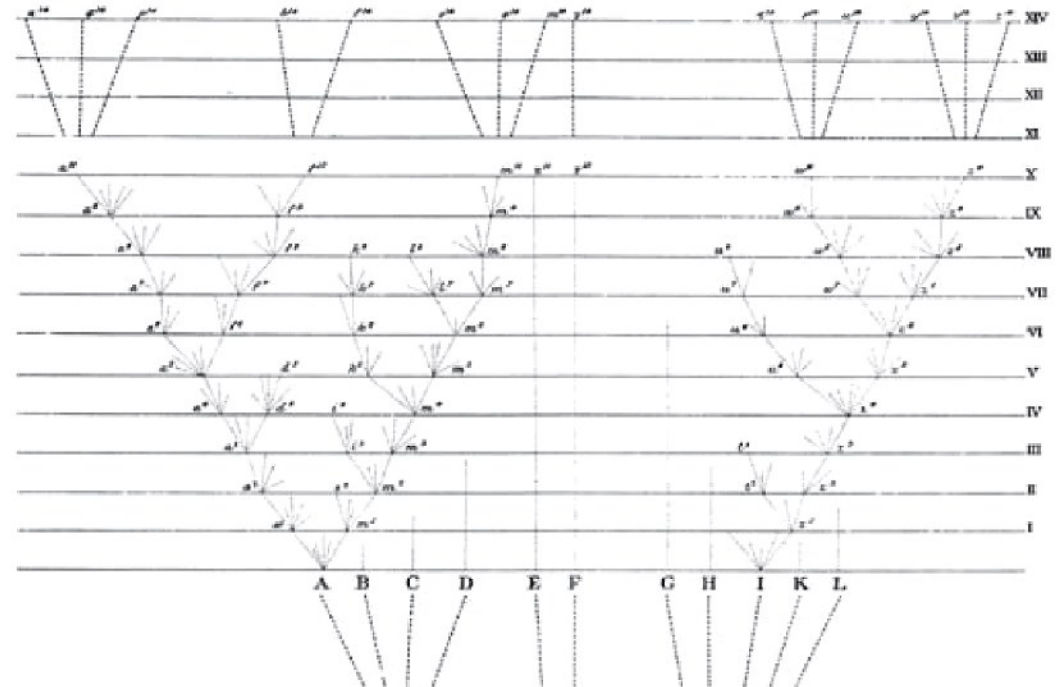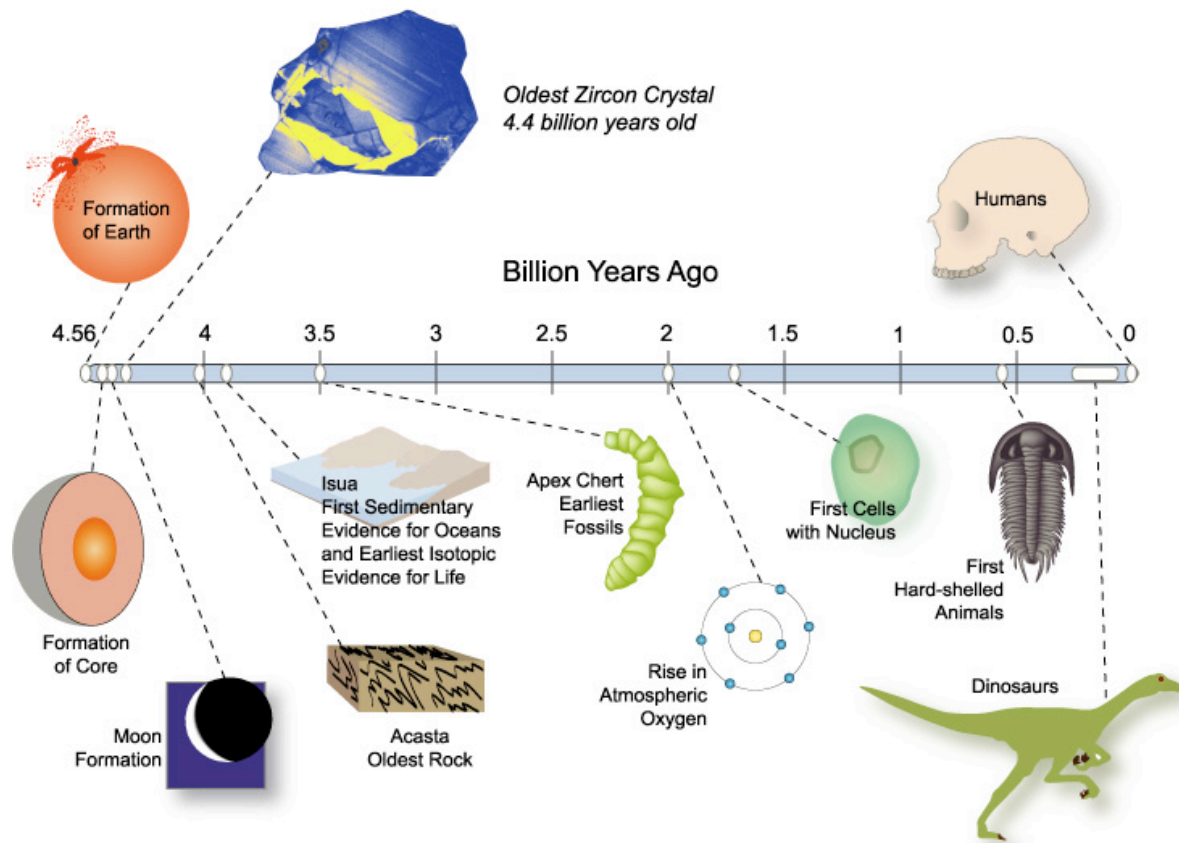
# The tree of life

**(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
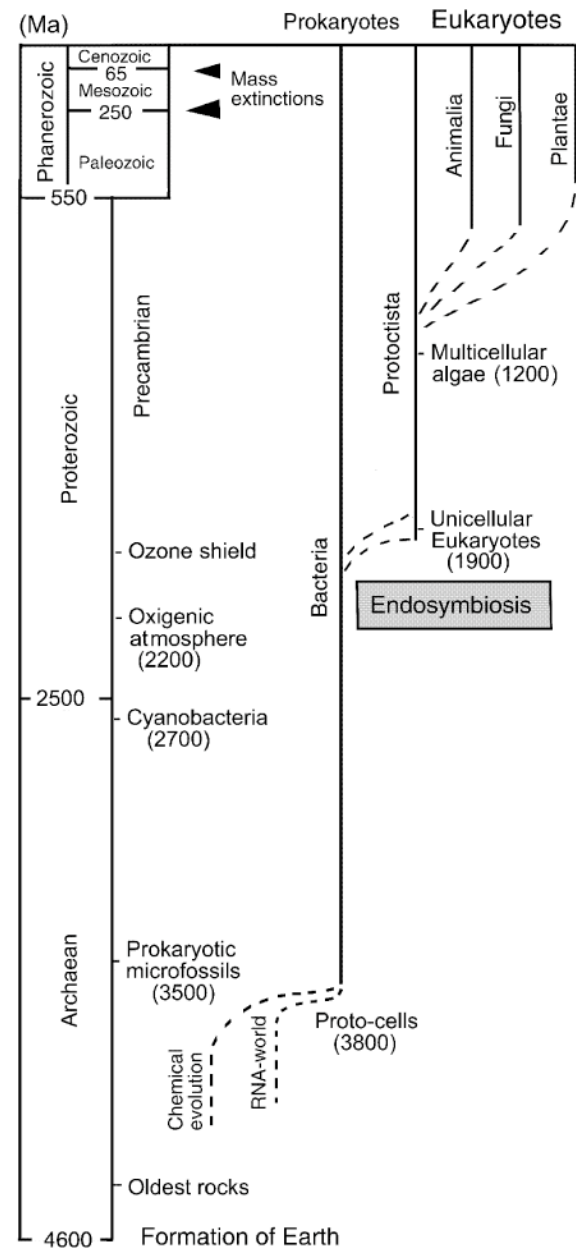
Timescales of evolution II



**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
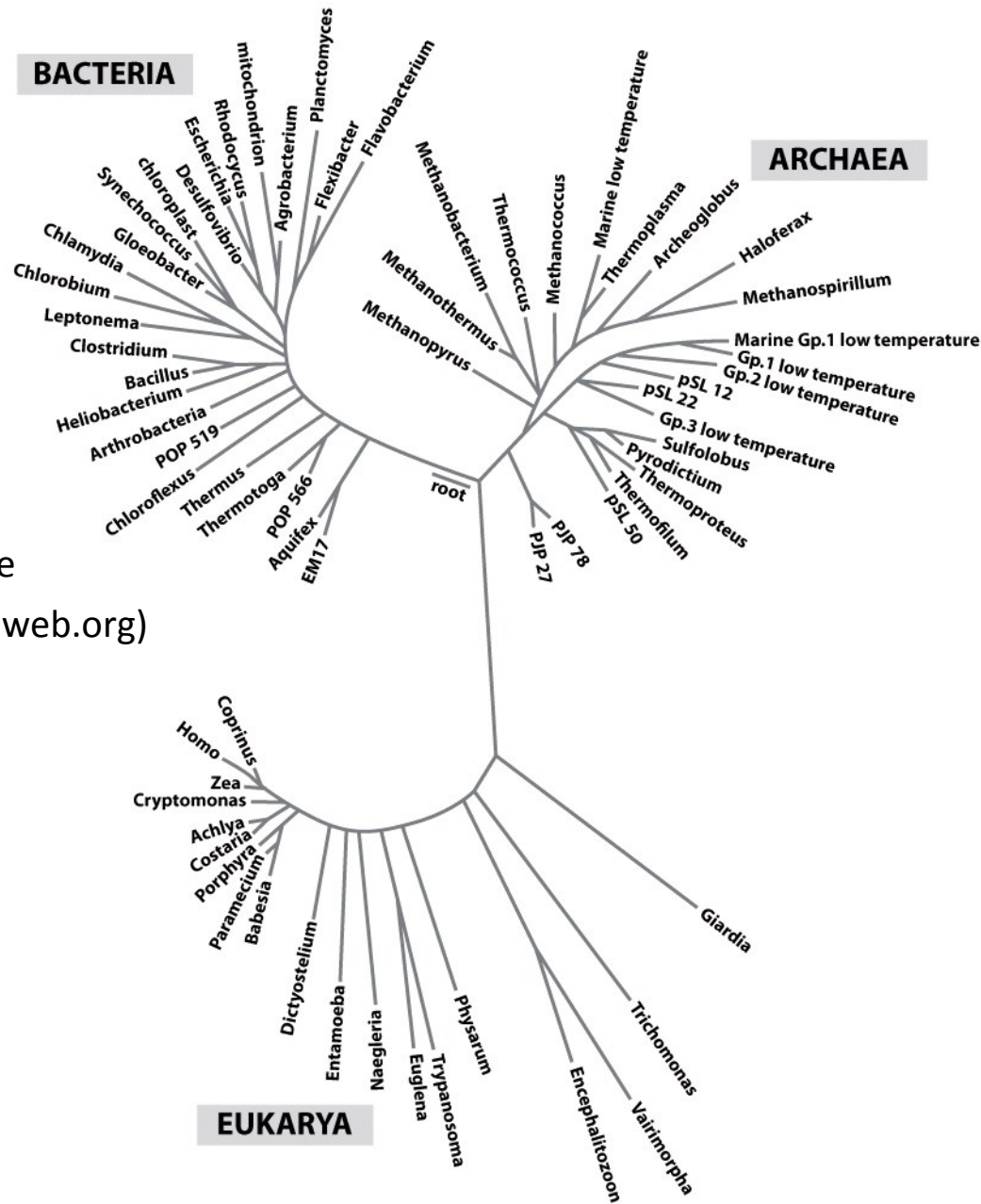
The tree of life (tolweb.org)

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

# Molecular evolution and biological diversity
## The **tree** (graph, network) **of life**

# the big picture on buoinformaticsreferences

- Comparing nucleotide sequences (genetic material) of two or more organisms often reveal that changes have been accumulated, at the DNA level, even if all the sequences come from functionally equivalent regions (suggesting a **neutral, non selective drift**)

- Actually, it is not uncommon that, during the evolution, **homologous sequences** have become so different as to make it very difficult to obtain reliable alignments (the problem of **remote homology**)

- The analysis of both the number and the type of substitutions, that have been occurred during the evolution, are of central importance for the study of **molecular evolution ----> BIOINFORMATICS & MOLECULAR EVOLUTION (higgs&attwood) or e.g. Evolutionary Bioinformatics by D. Forsdyke , structural bioinformatics (next Monday)**
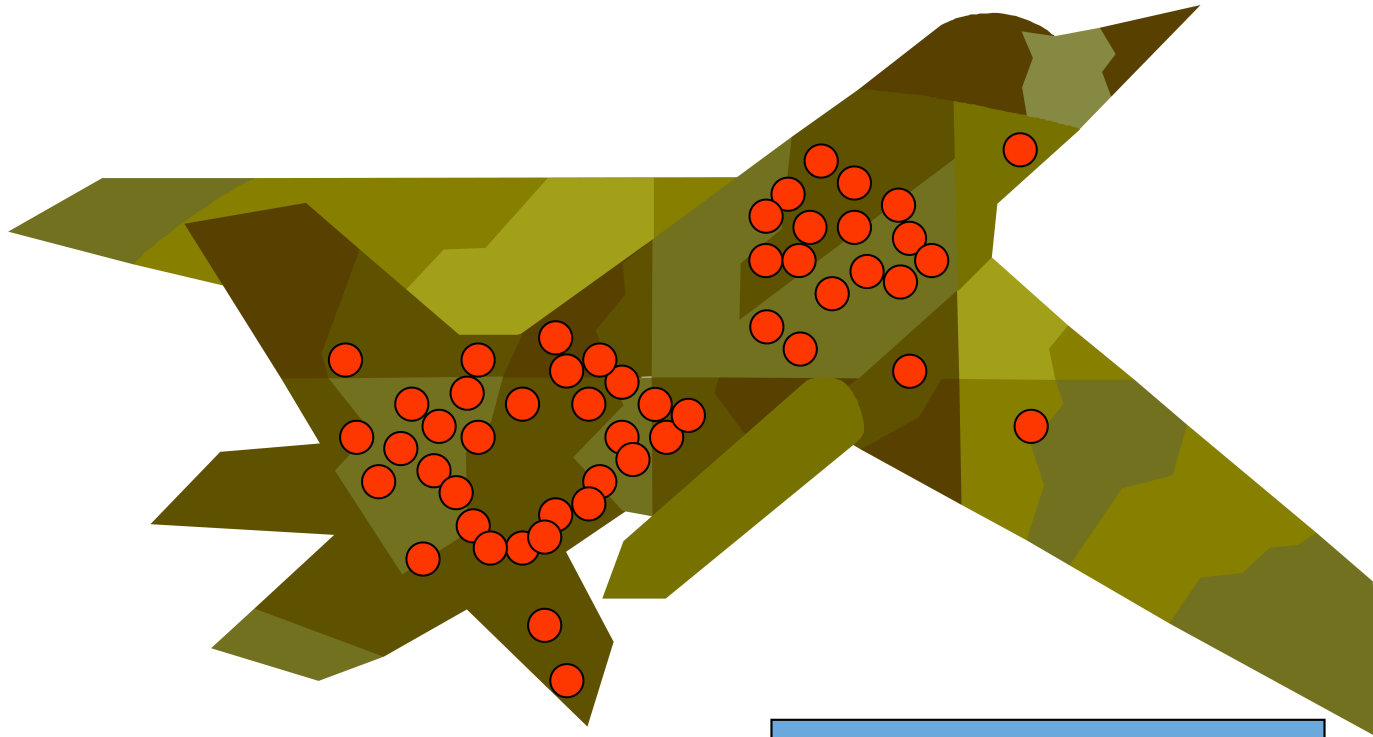
# ABOUT MOLECULAR EVOLUTION

◆ *DNA molecules are not only the key to heredity, but they are "document of evolutionary history"* (E. Zuckerkandl)---> **DNAs, genomes are tghe ARCHIVES of Evolution (just fancy idea?)**

◆ Molecular evolution integrates evolutionary biology, molecular biology, and population genetics

- It describes the process of evolution (changes in time, Being vs Becoming) of **DNAs, mRNAs, tRNAs, ncRNAs and proteins**
- It includes the study of **rates** of sequence change, relative importance of **adaptive** and **neutral** changes, and changes in **genome structure (e.g. chromatin structure, Hi-C maps)**
- It deals with **patterns** (diagrams, models) and studies the evolution of...
  - ✖ ...molecular entities, like genes, genomes, proteins, introns, chromosomal arrangements
  - ✖ ...organisms and biological systems, i.e. species, systems that co-evolve, ecological niches, migration patterns

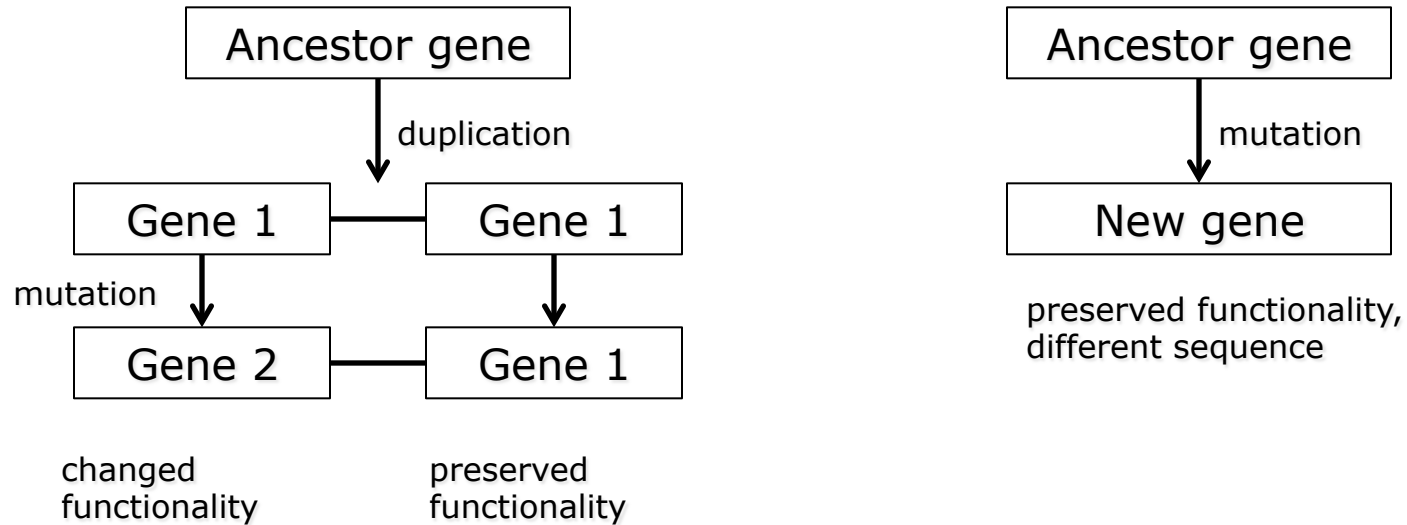  using **molecular** data (the pioneer has been **Carl Woese**).

# Natural Selection

➤ The **process** of <span style="color:red">natural selection</span> is truly effective in removing harmful (**fitness** reducing) changes, molecular evolution also serves to recognize and characterize the genome portions that are **more important (conserved, invariant, subject to purifying selection)** from the functional point of view (<span style="color:green">Noether's theorem...</span>)

➤ ...the rates (see then the Jukes –Cantor model) of nucleotide substitutions are different in different areas of the same gene, for different genes, and across species, and may be used as a measure of the functional significance of a particular sequence (and, therefore, it accounts for the need of its <span style="color:green">"conservation"</span>)

# Evolution in a nutshell (remembering Anna Tramontano, pioneer of bioinformatics)



Manguel M, Samaniego F.J.,
***Abraham Wald's Work on Aircraft Suvivability,***
J. American Statistical Association. 79, 259-270, (1984)

# Genes and proteins - 1



e.g. active kitinases vs chitinase-like proteins

**Paralogy / orthology/synteny (topology)**

# Genes and proteins - 2

- Orthologous genes: similar genes, found in organisms related to each other
  - The speciation phenomenon leads to the divergence of genes and, therefore, of the proteins that they encode
  - Example: Human and mouse β-globins started to diverge about 80 million years ago, when the evolutionary event, that gave rise to primates and rodents, took place
- Paralogous genes: genes originated from the duplication of a single gene in the same organism
  - Example: Human α-globins and β-globins began to diverge due to the duplication of an ancestral globin gene
- In both cases, there is homology

# Genes and proteins – 3

COWS

**Human ribonuclease**
(digestive enzyme)



Orthologous
genes

speciation

**Bovine ribonuclease**
(digestive enzyme)

duplication

Paralogous
genes

**Angiogenin**
(It stimulates the growth of blood vessels)

# How proteins change

✦ A protein present in a particular organism can change as a result of some mutations in its coding sequence

✦ Mutations can be point-like or frame-shift

- **Point mutations**: substitution of a single nucleotide

- **Insertion** : one or more nucleotides are inserted

- **Deletion** :one or more nucleotides are removed

- **Inversion** : a DNA stretch is reversed

- What about **transposons?**

This theme is covered in our textbook: Higgs & Attwood, Bioinformatics and Molecular Evolution

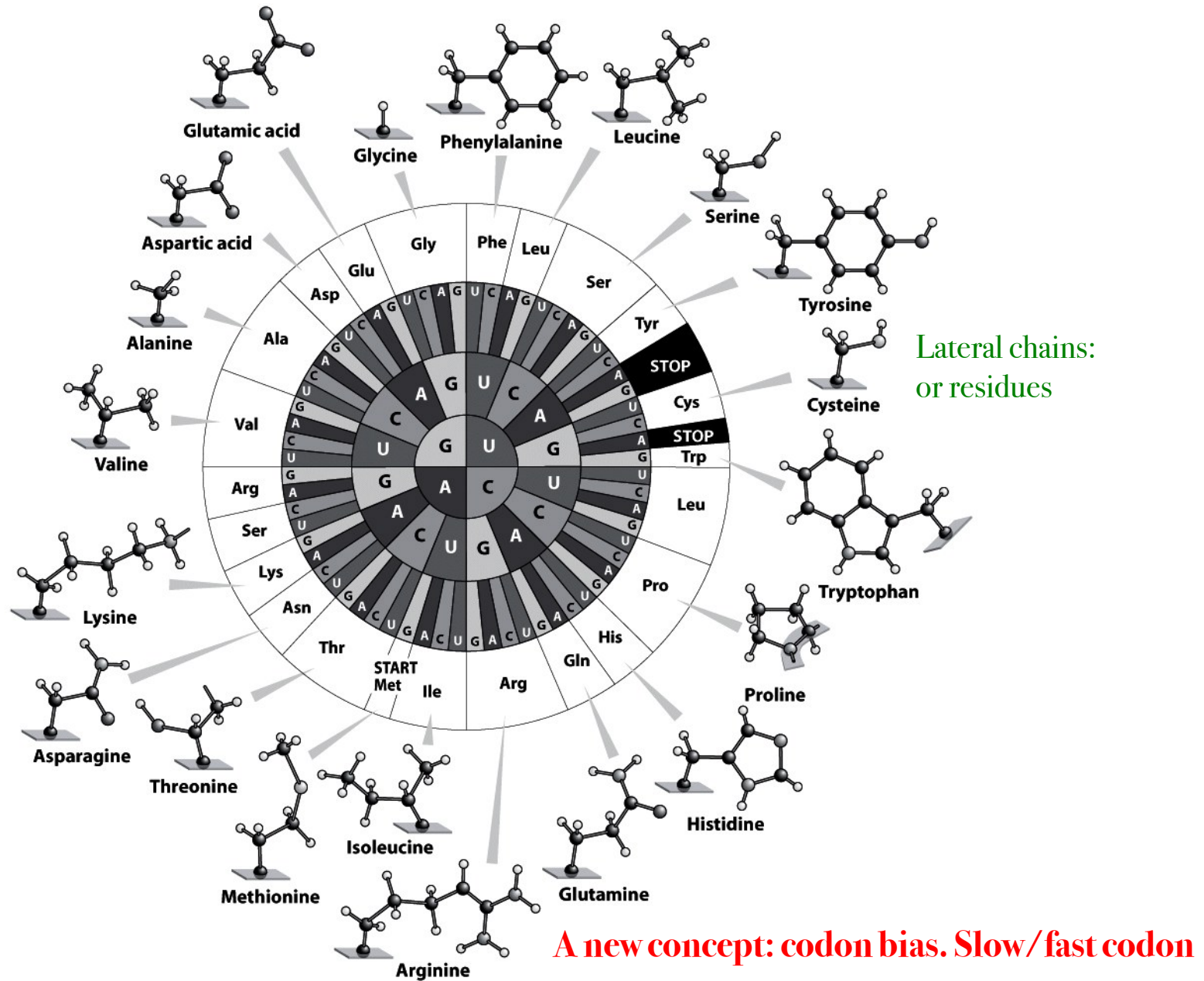# The genetic code: codons code for amino acids



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

Lateral chains: or residues

A new concept: codon bias. Slow/fast codon

# Substitution patterns

*"Organic life, we are told, has developed gradually from the protozoon to the philosopher, and this development, we are assured, is indubitably an advance. Unfortunately it is the philosopher, not the protozoon, who gives us this assurance."*

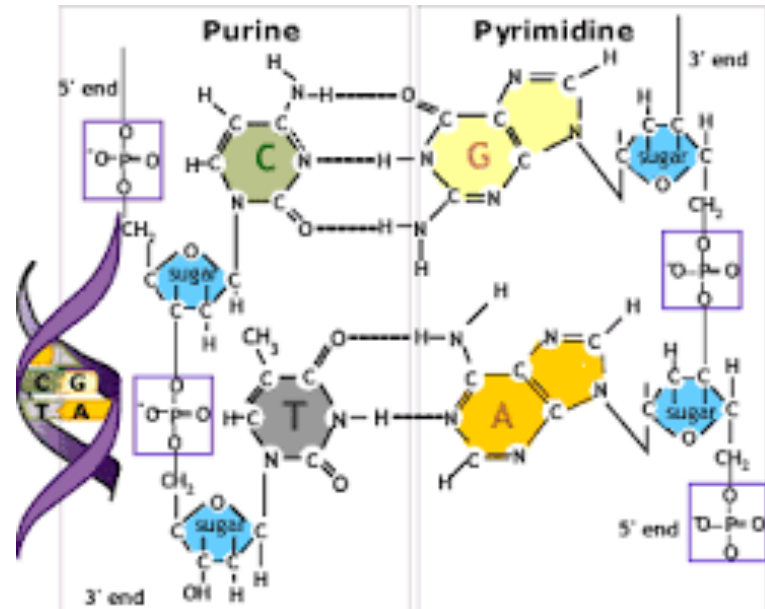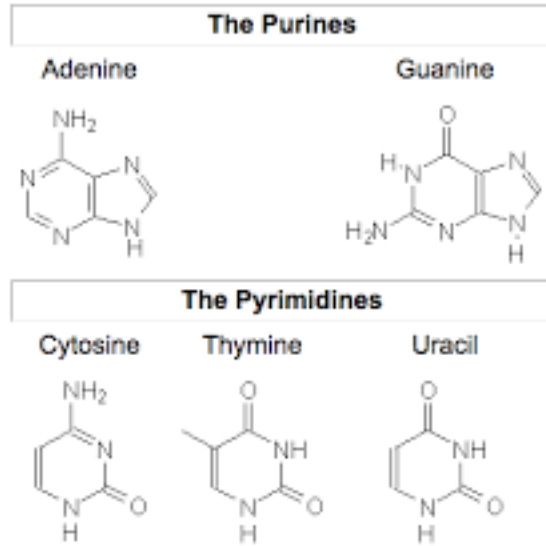**(B. Russel,** *Mysticism and Logic,* **1918)** **…viruses as angels…aha!**

# How proteins can change - 2

- The genetic code is redundant and, therefore, a substitution does not always lead to a change of an amino acid

  - A silent mutation occurs if the protein remains functionally unchanged

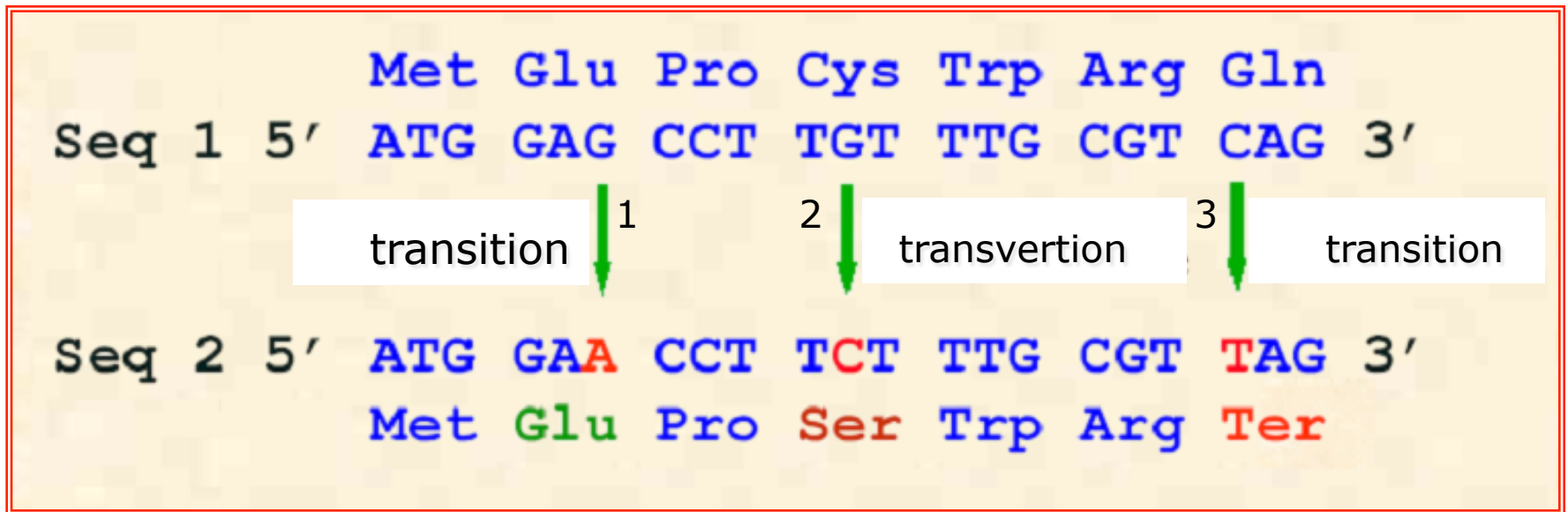- In other cases, from the mutation point onwards, the amino acids change, and the protein can become "unrecognizable" and definitely lose its functionality
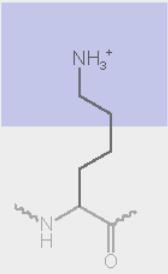
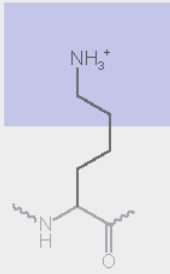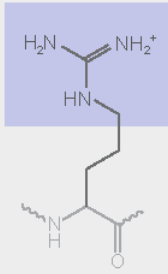# Purines vs pyrimidines

transition

transversion



The Purines

Adenine     Guanine

The Pyrimidines

Cytosine     Thymine     Uracil



Purine     Pyrimidine

# How proteins can change - 3

Look at point mutations (purines vs pirimidines)

Met Glu Pro Cys Trp Arg Gln
Seq 1 5′ ATG GAG CCT TGT TTG CGT CAG 3′

transition    1    2   transvertion    3    transition

Seq 2 5′ ATG GAA CCT TCT TTG CGT TAG 3′
      Met Glu Pro Ser Trp Arg Ter

(1) Glutamic acid → Glutamic acid
(2) Cysteine → Serine (amino acids with a polar, chiral molecule)
(3) Glutamine → Stop codon

# How proteins can change - 4



| No mutation | Point mutations | | | |
| | Silent | Nonsense | Missense | |
| | | | conservative | non-conservative |
| **DNA level** | TTC | TTT | ATC | TCC | TGC |
| **mRNA level** | AAG | AAA | UAG | AGG | ACG |
| **protein level** | Lys | Lys | STOP | Arg | Thr |

basic ▇
polar ▇

Arginine and lysine are both basic amino acids (positively charged), while threonine is a polar amino acid (hydrophilic)

◆ If two sequences have a significant **degree of similarity (how to measure? Hamming distance? Distance in amino acid composition)** for all their length, it is very likely that this is due to a sort of "memory" of their evolutionary relationship (do evolutionarily related proteins have a memory in amino acidic composition? **E.G. chitinases**)

◆ Two sequences that do not show a strong similarity, however, can still be homologous (sharing a very **remote** common ancestor, or having subdue to a **very rapid** evolutionary dynamics)
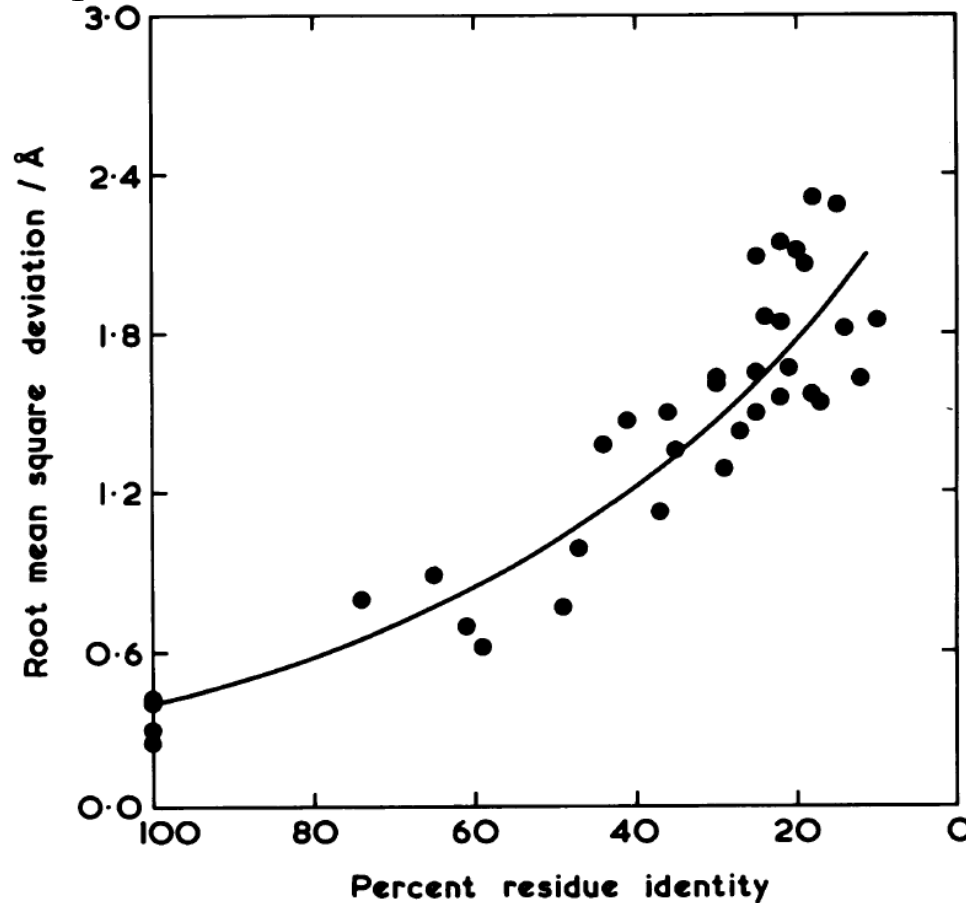
◆ Note that... Similarity ≠ Homology

**Sequence Similarity** is a quantitative information, based on the chosen metric, and it is independent from assumptions about the cause of the similarity itself

**Sequence Homology** is a qualitative information related to the ontology, that stands for the common phylogenetic origin of two sequences. The **evolutionary distance** is related to homology. There is a twilight zone: elaborate on that.

# SEQUENCE SIMILARITY VS HOMOLOGY

# 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