Cells: building blocks of living organisms

Two kinds of cells (with and without nucleus)

Prokaryote (procaryote, prokaryotic cell, procaryotic organism): Cell or organism lacking a membrane-bound, structurally discrete nucleus and other sub-cellular compartments. Bacteria are prokaryotes

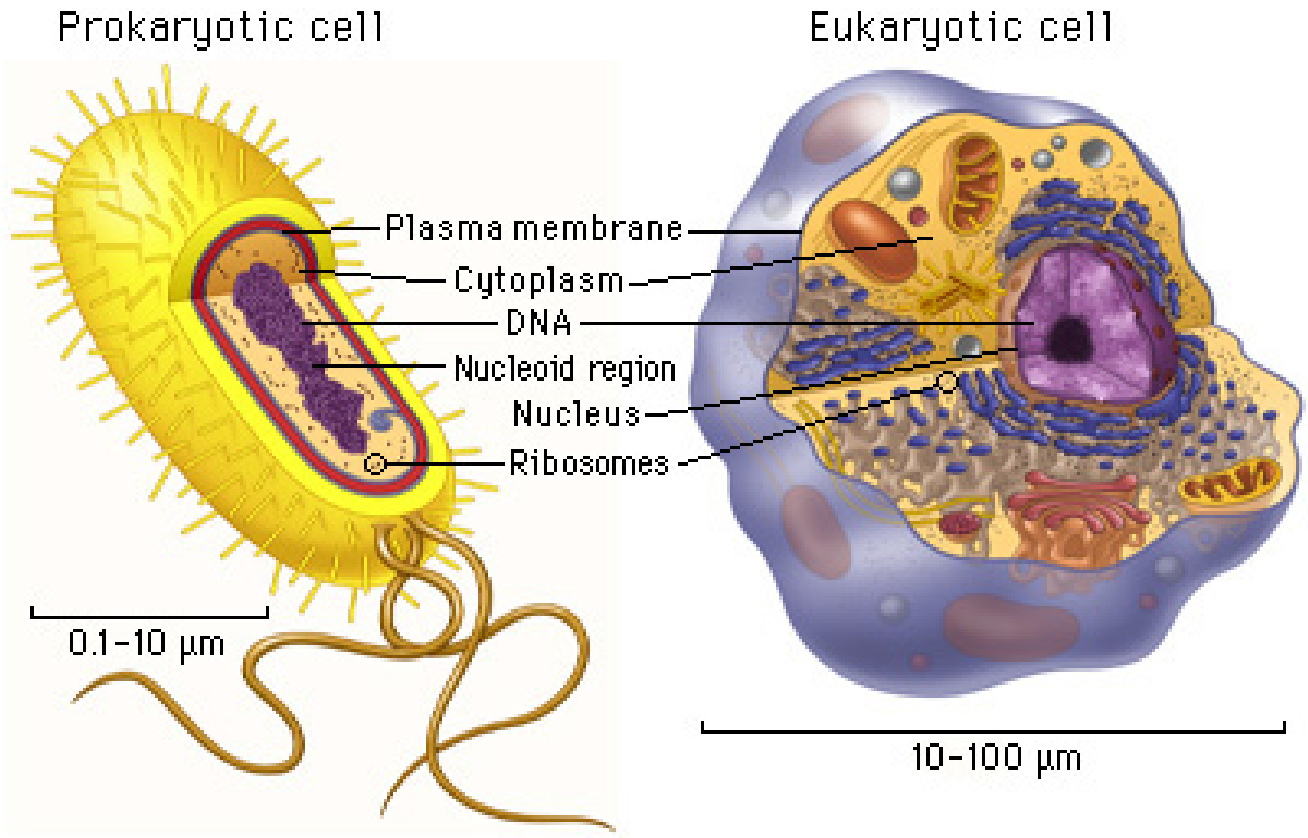
Eukaryote (eucaryote, eukaryotic cell, eucaryotic cell): Cell or organism with a membrane-bound, structurally discrete nucleus and other well-developed sub-cellular compartments. Eukaryotes include all organisms except viruses, bacteria, and cyanobacteria (blue-green algae)

Cells: building blocks of living organisms

Eukaryotic cells are generally larger than prokaryotic cells.
The packaging of the genetic information (DNA) is much more structured and compact in Eukaryotes compared to prokaryotes.

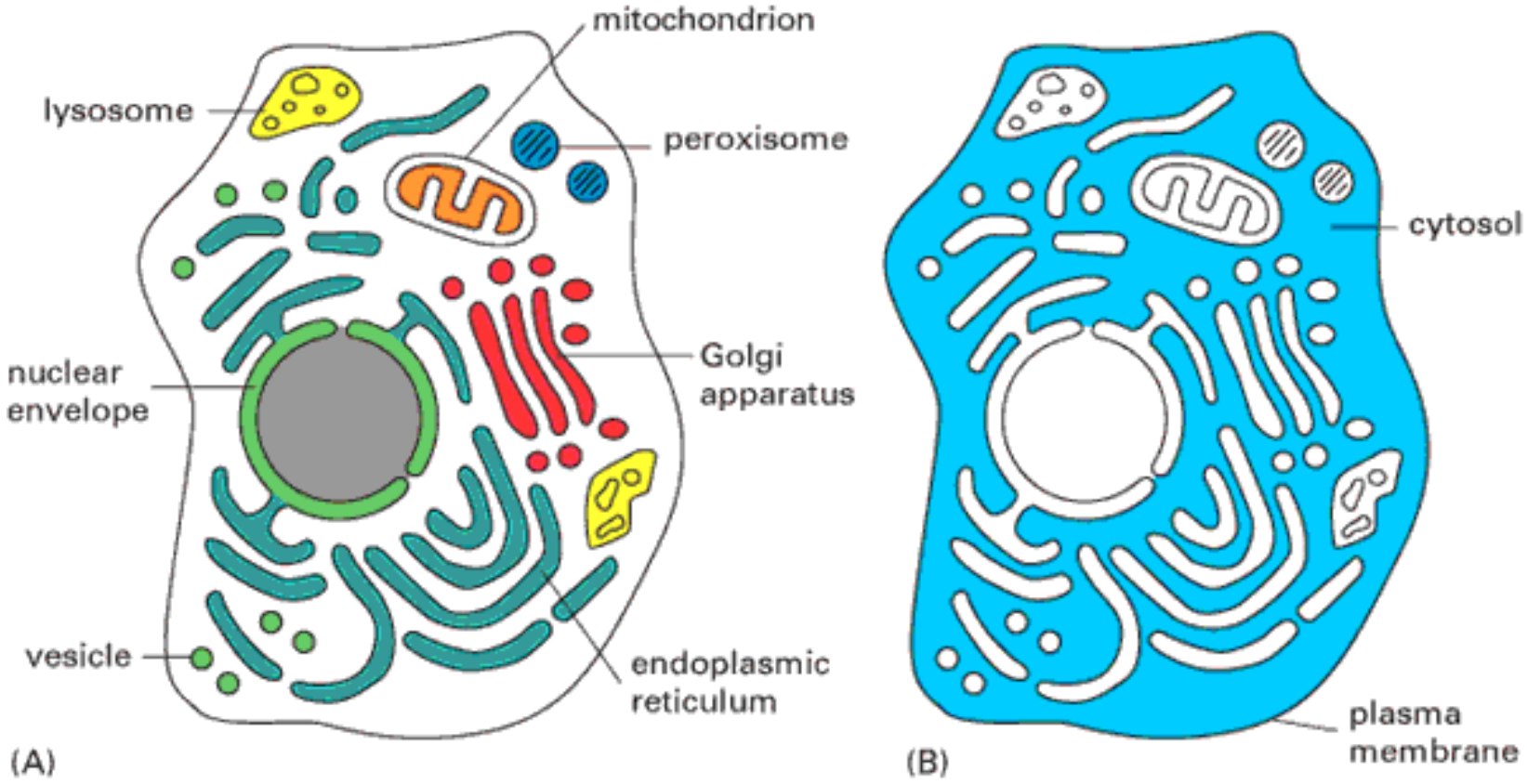
$\Rightarrow$ Cell theory: 1939 by Matthias Schleiden and Theodor Schwann.

Prokaryotic vs eukaryotic cell



www.phschool.com/science/biology_place/biocoach/cells/common.html

Organisation of an eukaryotic cell



Organelle genomes

- Organelles are discrete structures having specialized functions
- Mitochondria are energy-generating organelles (cellular power plants)
- Mitochondria contain DNA and a small number of genes, which are sometimes called extrachromosomal genes or mitochondrial genes
- Several organelles are believed to be engulfed prokaryotes (endosymbiotic theory made popular by Lynn Margulis)
- Mitochondria make it clear why certain genes are inherited from the mother only

Bioinformaticist's point of view

- The organization of genes (genome structure) is quite different between the two kinds of cell
- Consequently the gene-finding algorithms must be adapted
- Eukaryotic cells being more complex provide a richer set of problems:
  e.g. protein sub-cellular localisation problem

(3) kingdoms of life

Prokarya: the cells of those organisms, prokaryotes, do not have a nucleus. Representative organisms are *cyanobacteria* (blue-green algae) and *Escherichia coli* (a common bacteria)

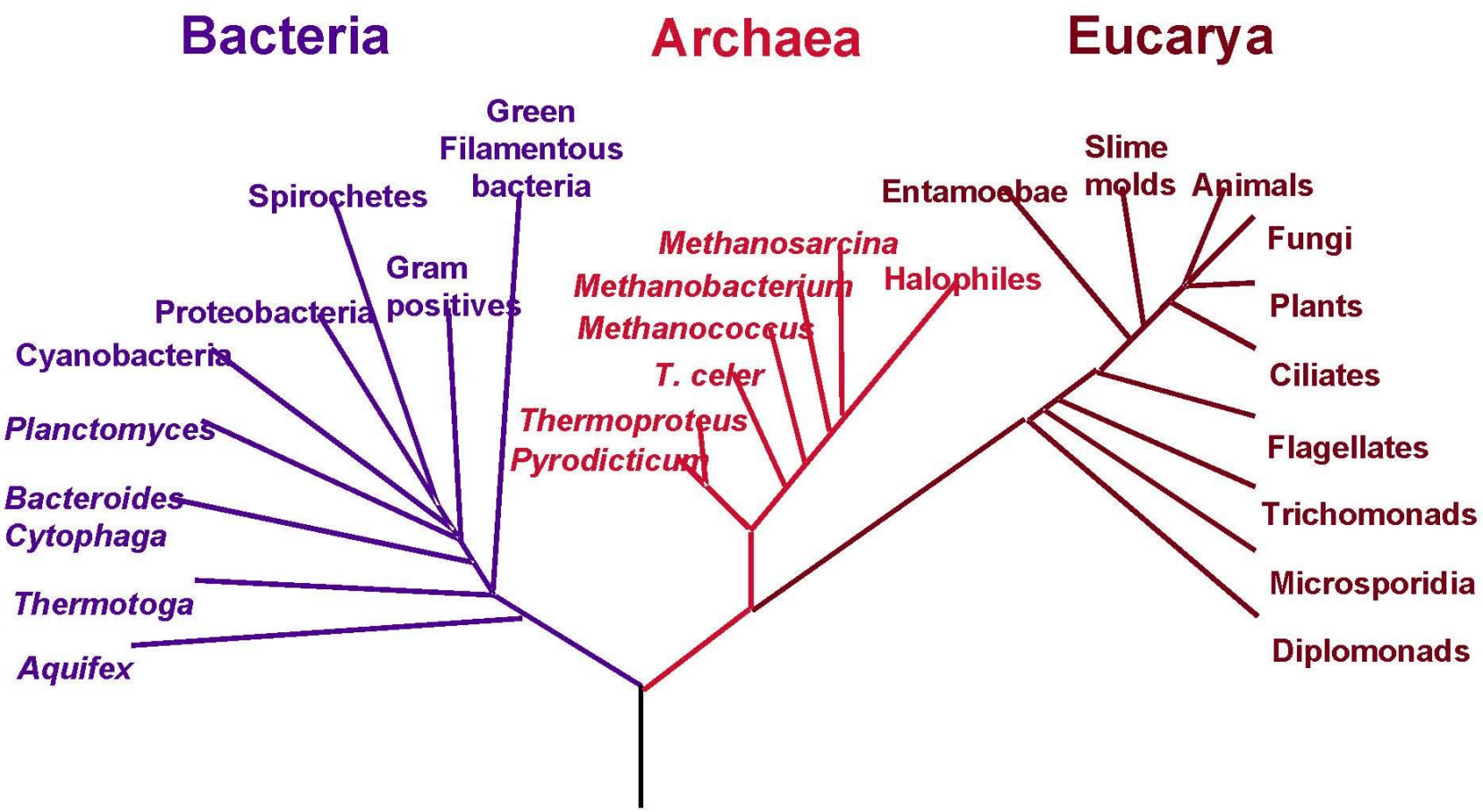
Eukarya: the cells of those organisms, eukaryotes, all have a nucleus. Representative organisms are *Trypanosoma brucei* (unicelluar organism which can cause sleeping sickness) and *Homo sapiens* (multicellular organism)

Archaea: (archaebacteria) like the prokaryotes they lack the nuclear membrane but have transcription and translation mechanisms close to those of the eukaryotes

(3) kingdoms of life: Archaea

Methanococcus jannaschii is an methane producing archaebacterium which had its complete genome sequenced in 1996. This organism was discovered in 1982 in white smoker of a hot spot at the bottom of the Pacific ocean: depth 2600 meters, temperature 48-94° C (thermophilic), optimum at 85° C, 1.66 Mega bases, 1738 genes. 56% of its genes are unlike any known eukaryote or prokaryote, one kind of DNA polymerase (other genomes have several).

Phylogenetic Tree of Life



Phylogenetic tree

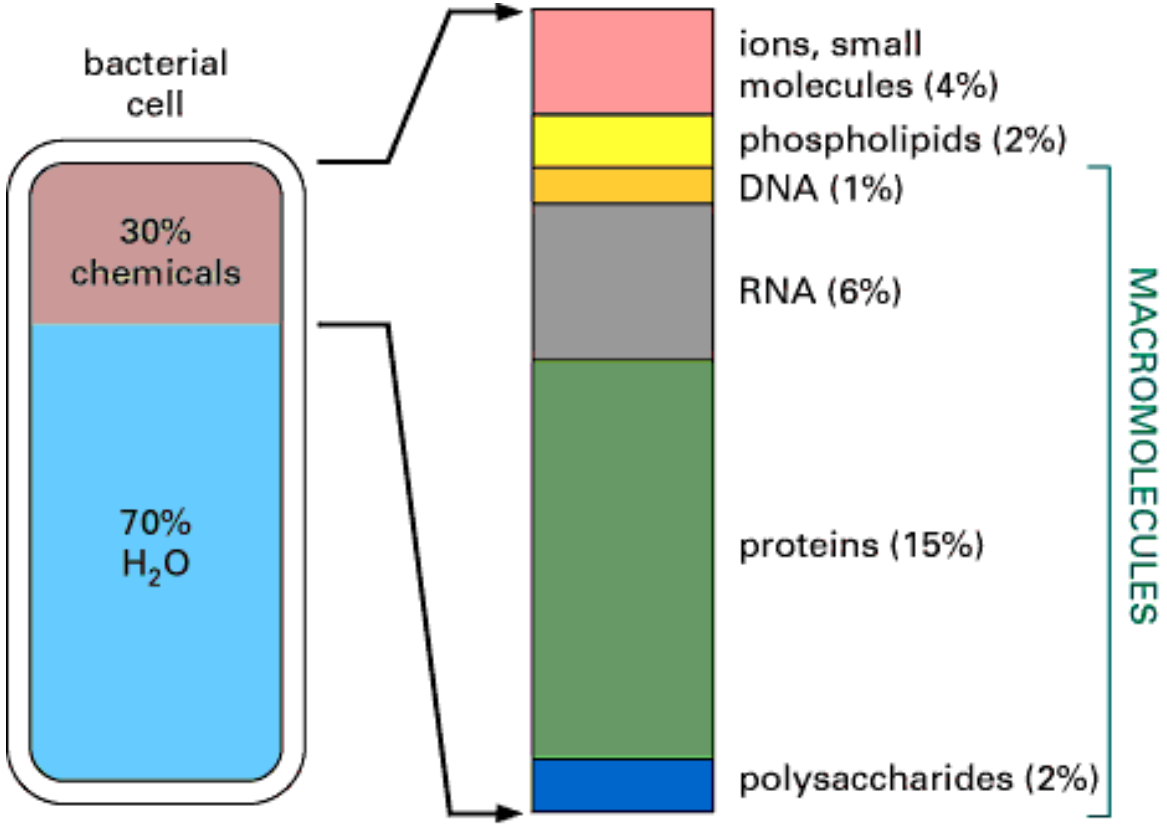
- "The objectives of phylogenetic studies are (1) to reconstruct the correct genealogical ties between organisms and (2) to estimate the time of divergence between organisms since they last shared a common ancestor."
- "A phylogenetic tree is a graph composed of nodes and branches, in which only one branch connects any two adjacent nodes."
- "The nodes represents the taxonomic units, and the branches define the relationships among the units in terms of descent and ancestry."
- "The branch length usually represents the number of changes that have occurred in that branch." (or some amount of time)

$\Rightarrow$ Li, W.-H. and Graur, D. (1991) Fundamentals of Molecular Evolution. Sinauer.

Bioinformaticist's point of view

- **Large phylogeny problem**: Reconstructing phylogenetic trees from molecular sequence data
- **Small phylogeny problem**: Reconstructing ancestral molecular sequences

Composition of the Cell



$\Rightarrow$ DNA, RNA and proteins will be the main focus of the course.

Macromolecules: DNA (deoxyribonucleic acid), RNA (ribonucleic acid) and Protein

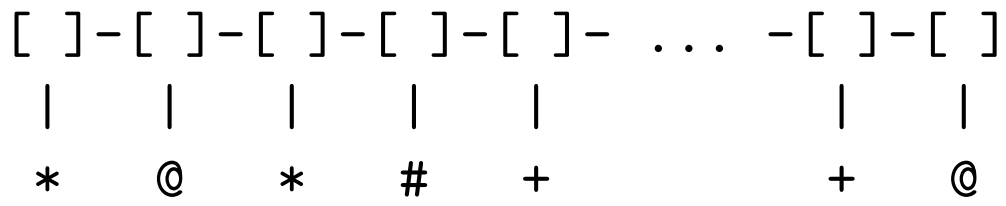
Bioinformatics is mainly concerned with three classes of molecules: DNA, RNA and proteins — collectively called **macromolecules** or **biomolecules**.

Macromolecules: DNA (deoxyribonucleic acid), RNA (ribonucleic acid) and Protein

All three classes of macromolecules are **polymers**, that is they are composed of smaller units (molecules), called monomers, that are linked sequentially one to another forming unbranched linear structures.

Macromolecules: DNA (deoxyribonucleic acid), RNA (ribonucleic acid) and Protein

Generally speaking, the units (monomers) consits of two distinct parts, one that is **common** to all the monomers and defines the **backbone** of the molecule, and another part that confers the **identity** of the unit, and therefore its **properties**.



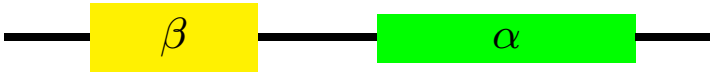
Structure

It's useful to distinguish between four levels of abstraction or structure: **primary**, **secondary**, **tertiary** and **quaternary** structure.

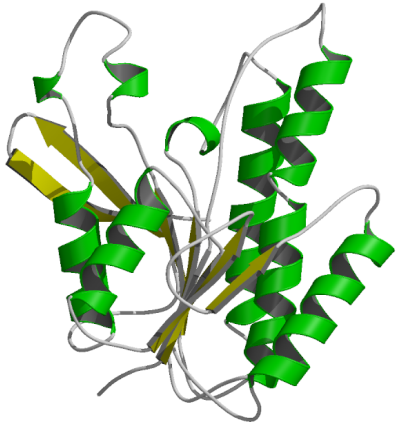
1, 2, 3, . . .

EARRVLVYGGRGALGSRCVQNW . . . (236) . . .

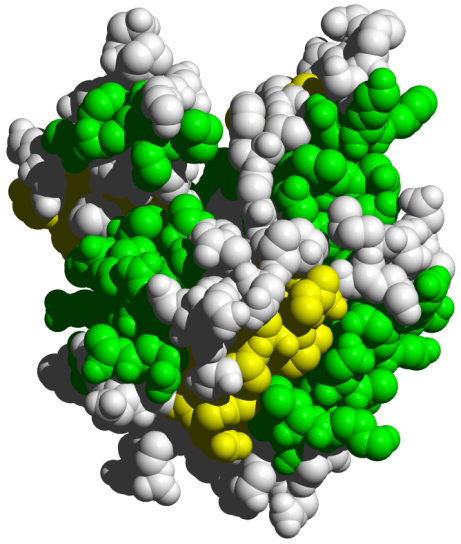
(a) primary structure



(b) secondary structure



(c) tertiary structure - ribbon



(d) tertiary structure - all atoms

Word Analogy

- Words are made of letters and the particular order of the letters defines the identity the word (each word has a semantic)
- The grammatical structure of the words in a sentence is highly related to the meaning of the sentence
- Polymers are made of monomers and the order of the monomers defines their structure and function
- In biology, but also in general, structure and function are inter-related

Macromolecules: DNA (deoxyribonucleic acid), RNA (ribonucleic acid) and Protein

The primary structure or **sequence** is an ordered list of characters, from a given alphabet, written contiguously from left to right.
DNA : 4 letters alphabet,
$\Sigma = \{A, C, G, T\}$
RNA : 4 letters alphabet,
$\Sigma = \{A, C, G, U\}$
Proteins : 20 letters alphabet,
$\Sigma = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$

In the case of nucleic acids (DNA and RNA), the building blocks are called nucleotides, whilst in the case of proteins they are called amino acids.
Examples of DNA, RNA and protein sequences.

```
> Chimpanzee Chromosome 1; A DNA sequence (size = 245,522,847 nt)
TAACCCTAACCCTAACCCTAACCCTAACC ... TCTCATGACAGTGAGTGAGTTCTCATGATC
```

```
> A01592; An RNA sequence (coding Beta Globin gene) (size = 438 nt)
GUGCACCUGACUCCUGAGGAGAAGUCUGC ... GCAAGGUGAACGUGGAUGAAGUUGGUGGUG
```

```
> Beta Globin; A protein sequence (size = 147 aa)
MVHLTPEEKSAVTALWGKVNVDEVGGEAL ... FFESFGDLSTPDAVMGNPKVKAHGKKVLGA
```

Bioinformaticist's point of view

- Exact string (sequence) comparison, approximate matching ($k-$mismatches), comparison under the edit-distance, significance of match, multi-way sequence comparison
- Finding repeats, approximate repeats, finding interesting patterns
- Secondary, tertiary and quaternary structure inference

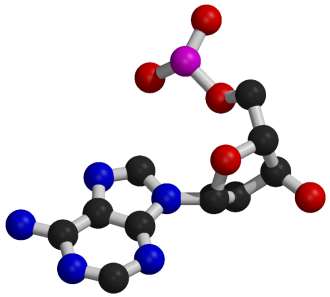
Challenges

- What is the longest word from the Oxford English Dictionary that also appears in the Human genome sequence?
- What is the word from the Oxford English Dictionary that appears the most frequently in the Human genome?
- What is the time/space complexity of a naive algorithm for answering the above two questions?
- Is it feasible to answer these two questions with today's computers?
- Come up with a clever strategy for solving these two problems. Can you do it in optimal linear time?

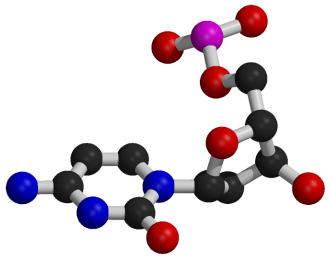
Deoxyribonucleic acids (DNA)

- DNA was discovered by Johann Friedrich Miescher in 1869. Who discarded the possibility that DNA might be related to heredity!
- The double-helical structure of DNA was proposed in 1953 by James Watson and Francis Crick (who died on July 28, 2004).
- This discovery is often referred to as the most important breakthrough in biology of the 20th century.
- The proposed model finally explained Chargaff's rule (same amount of adenine and thymine, same amount of guanine and cytosine).
- More importantly, the model finally explains how DNA and heredity are linked! (replication)

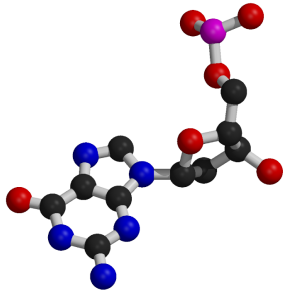
DNA's building blocks: ACGT



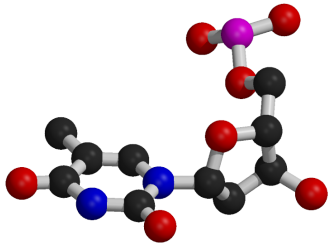
Adenine (A)



Cytosine (C)



Guanine (G)



Thymine (T)

$\Rightarrow$ Identify the common and unique parts of each monomer.

DNA/RNA's building blocks

- The common part of the nucleotides is formed of a deoxy-ribose (pentose, sugar) and a phosphate group.
- The part that is unique is called the (nitrogenous) base.
- If you look carefully you'll see big (two rings) and small (one ring) bases, respectively called **purines** (A,G) and **pyrimidines** (C,T).
- In the case of DNA, the bases are Adenine (A), Cytosine (C), Guanine (G) and Thymine (T).
- In the case of RNA, the bases are Adenine (A), Cytosine (C), Guanine (G) and Uracil (U).

DNA/RNA's building blocks

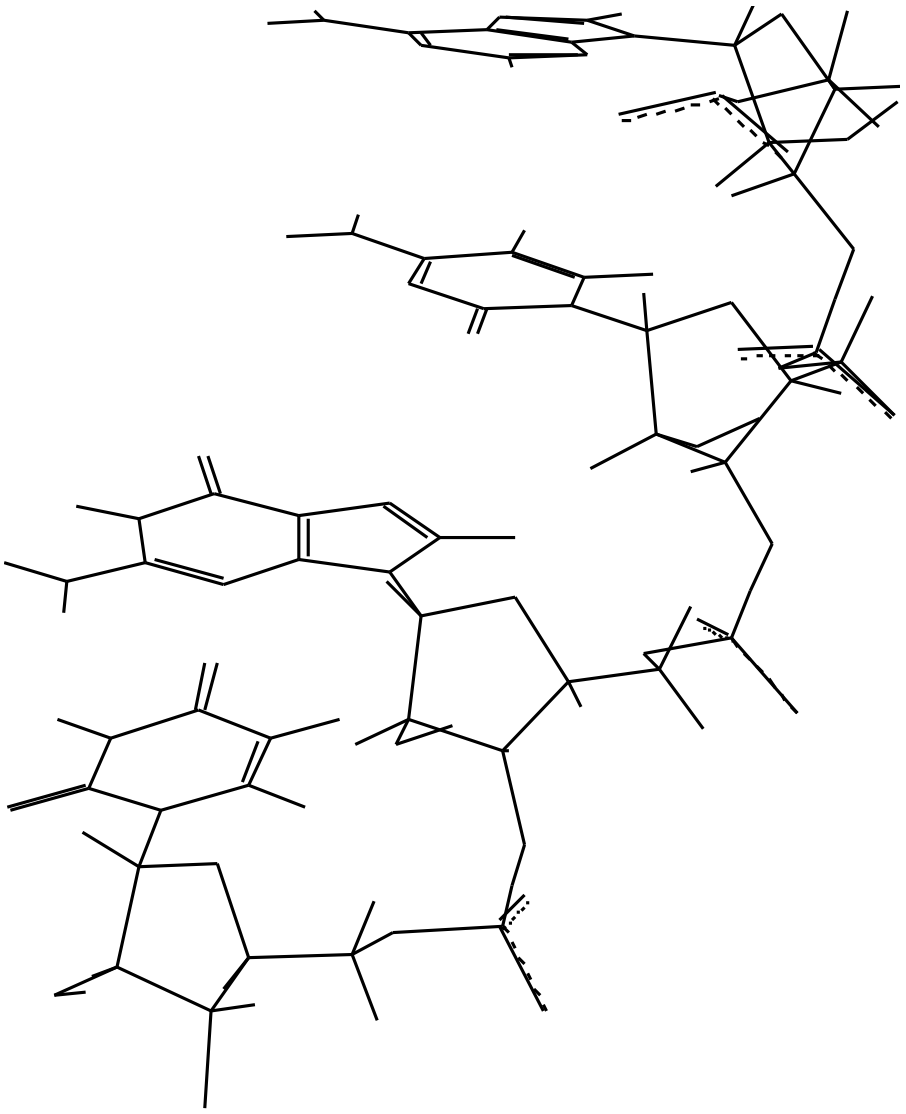
The length of a DNA/RNA molecule is often expressed in **bases**, e.g. a 10 **mega base** long region. Or, since nucleic acids molecules **hybridize** (bind together) to form a duplex (double helical) structure, the length of a molecule is often expression is base pairs to avoid confusion, e.g. a 10 **mega base** pairs region.

DNA/RNA building blocks

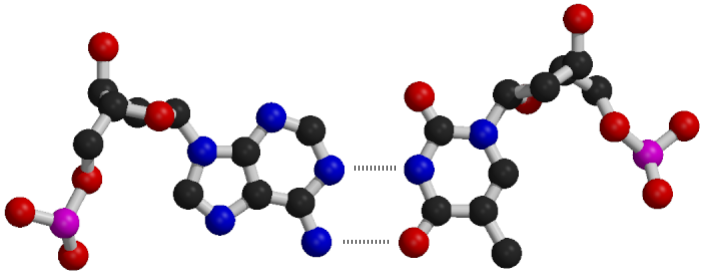
- DNA stands for deoxyribonucleic acid, and deoxy comes from the fact that the C2' carbon of the sugar has no oxygen; while RNA has one. RNA's O2' oxygen is key to its functional versatility!
- The other difference is the use of T (thymine) in the case of DNA vs U (uracil) in the case of RNA.
- Nucleotides are always attached one to another in the same way (well, almost always): the C3' atom of the nucleotide i is covalently linked to the phosphate group of the nucleotide $i + 1$.

DNA/RNA building blocks

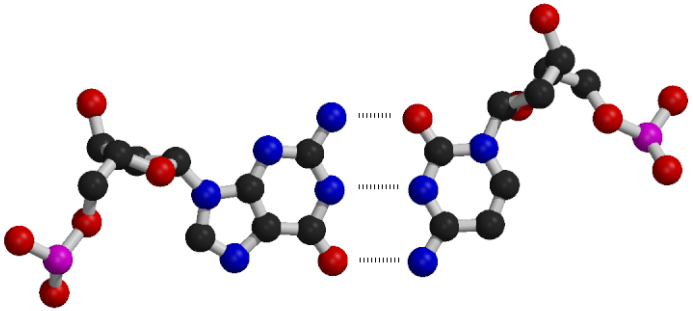
- The orientation of a DNA molecule is important; just like the orientation of words are important in natural languages.
- The convention is to enumerate the string from its 5' end; this correspond to the order into which information is process for certain key steps, to be described later. The features that are occurring before the 5' are said to be upstream while those occurring after the 3' end are downstream, upstream and downstream signals.



Watson-Crick (Canonical) base pairs



(Adenosine) A : T (Thymine)

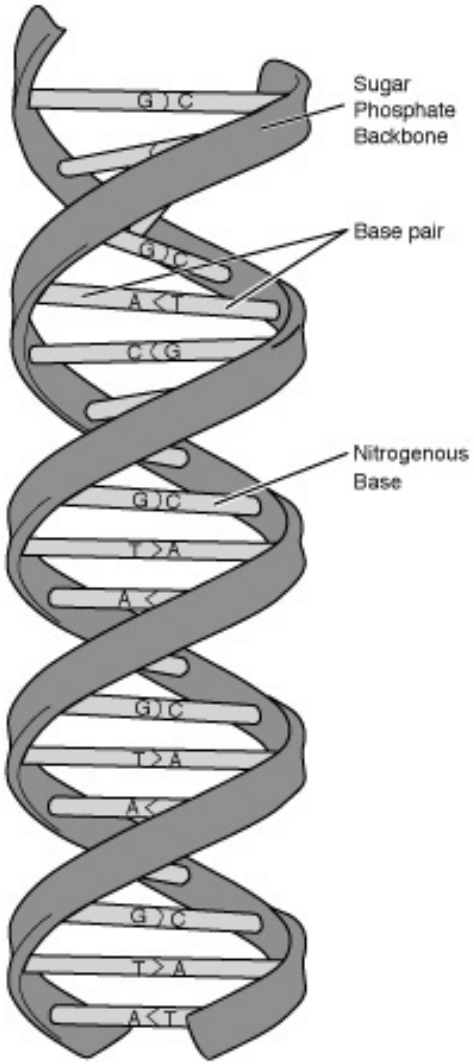


(Guanine) G : C (Cytosine)

$\Rightarrow$ One of the two base pairs is stronger that the other, which one?

Watson-Crick (Canonical) base pairs

In the case of DNA, bases interact, i.e. form hydrogen bonds, primarily through the following set of rules:
A interacts with T (and vice versa)
G interacts with C (and vice versa)
Those rules are the consequence of the fact that A:T and G:C pairs position the backbone atoms roughly at the same three-dimensional location and therefore both produces the same double helical structure; isosteric base pairs
DNA molecules generally form right-hand side helices in B form, while RNA are A form, also right-hand side. A left-hand side helix exists that is called Z DNA.
DNA molecules cannot exist as a single strand, they are degraded, i.e. cut into pieces.
A DNA molecule is made of two complementary strands running in opposite directions.


Sugar
Phosphate
Backbone
Base pair
Nitrogenous
Base

This explains how information can be copied from one generation to the next, or simply from one parent cell to its daughter cells during replication.
Before replication

```
5'  - GATACA -> 3' A
      ||||||
3' <- CTATGT -  5' B
```

A serves as a template to produce B'

```
5'  - GATACA -> 3' A
```

```
5'  - GATACA -> 3' A
      ||||||
3' <- CTATGT -  5' B'
```

Whilst B serves as a template to produce A'

```
5' -  TGTATC -> 3' B
```

```
5' -  TGTATC -> 3' B
      ||||||
3' <- ACATAG -> 5' A'
```

Parent cell (AB)

```
5’  - GATACA -> 3’ A
      ||||||
3’ <- CTATGT -  5’ B
```

Daughter cell (AB’)

```
5’  - GATACA -> 3’ A
      ||||||
3’ <- CTATGT -  5’ B’
```

Daughter cell (A’B)

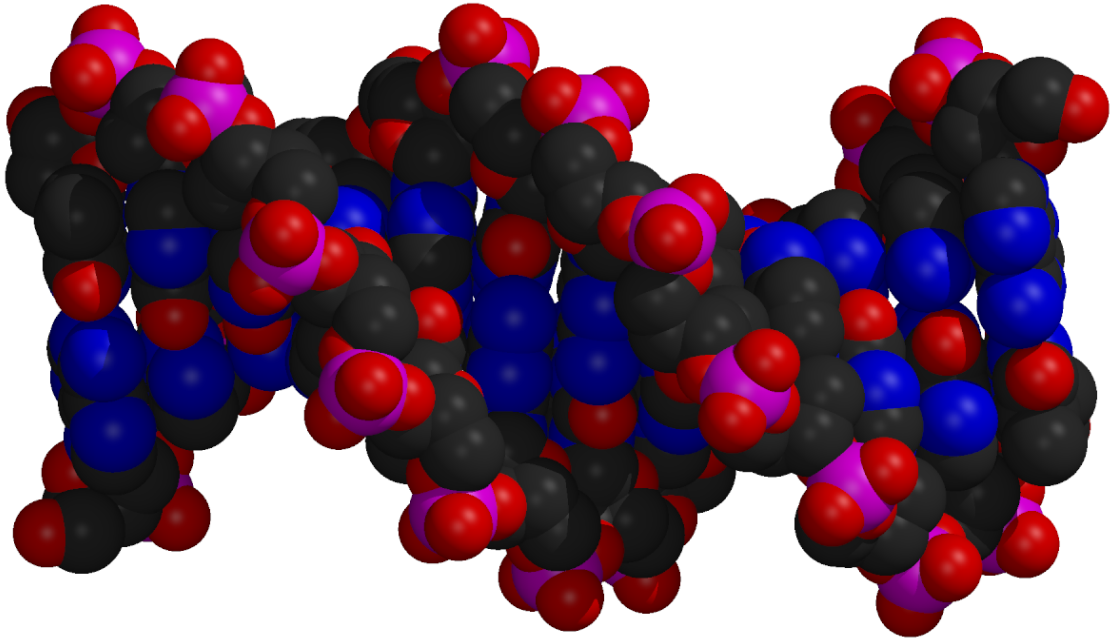
```
5’ -  TGTATC -> 3’ B
      ||||||
3’ <- ACATAG -> 5’ A’
```

Two daughter cells, identical to their parent.

Remarks

- Complex organisms are growing from a single cell to billions of cells. Each cell contains an exact copy of the DNA of its parent cell.
- The information is redundant, the information on the second strand can be inferred from the information on the first strand. This is the basis of DNA repair mechanisms. A base that is deleted can be replaced. A mismatch can be detected.

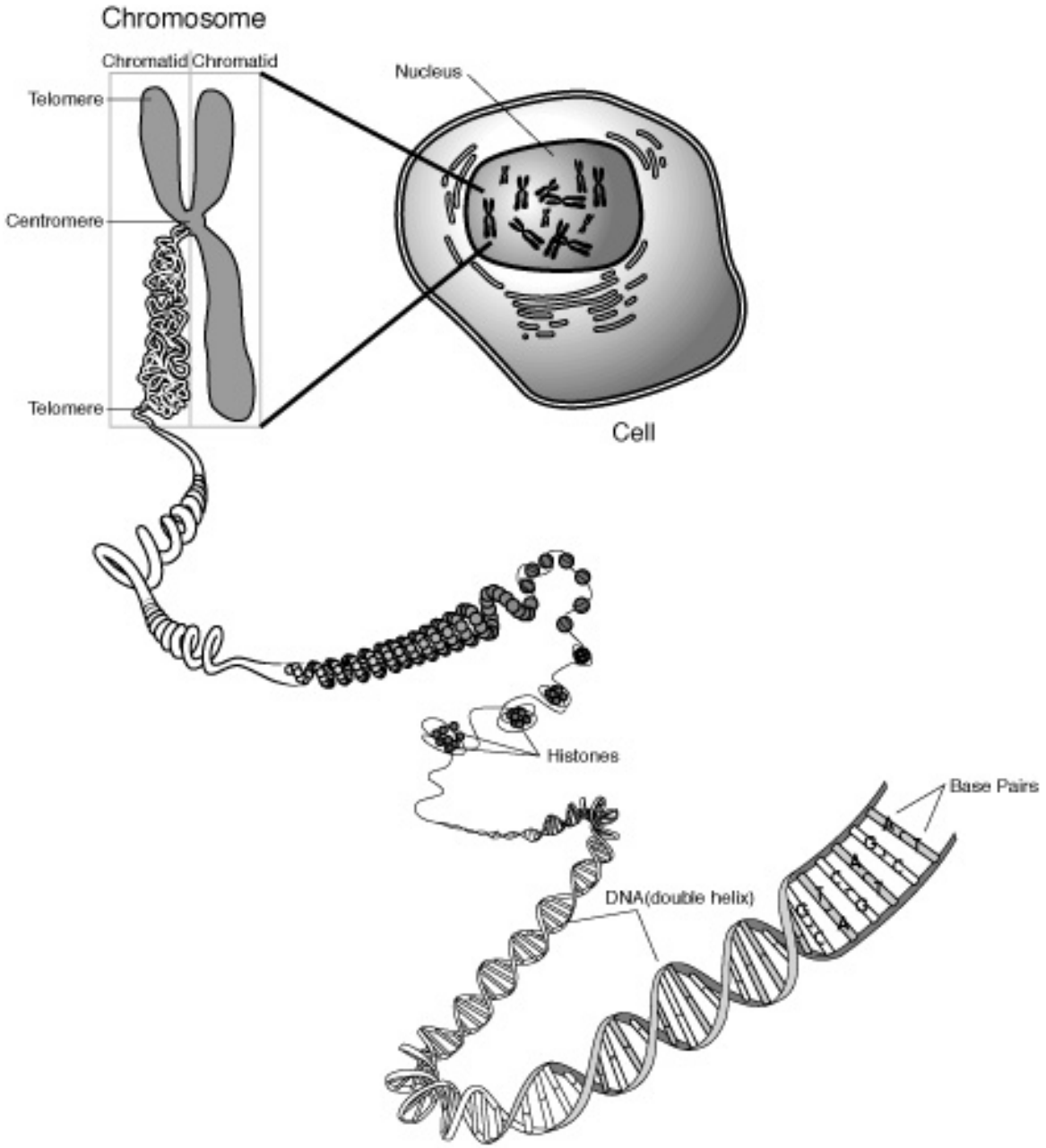
CPK representation of a fragment of a DNA helix (B form)

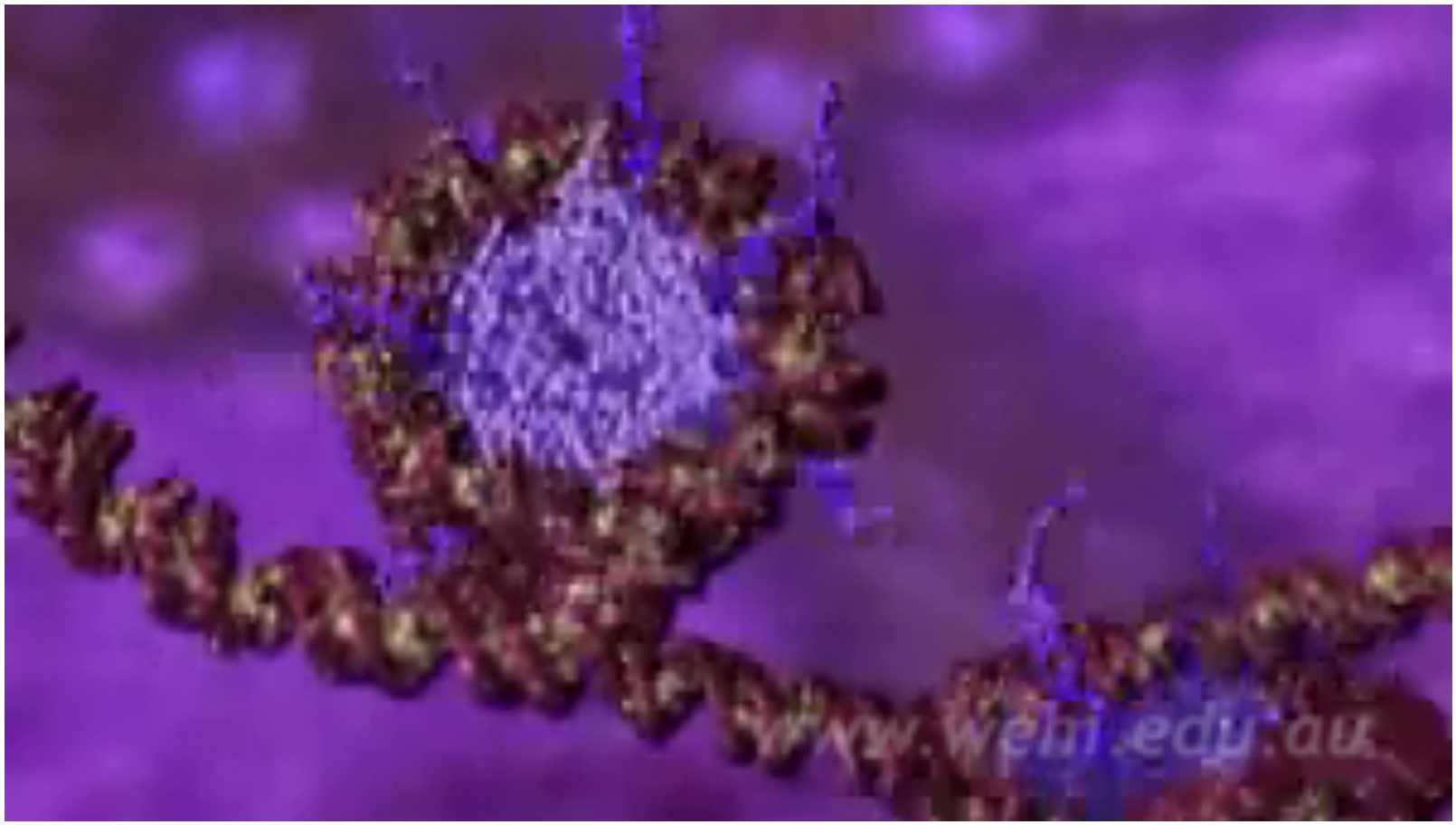


```
TAAGTTATTA                                AAAAAAATAC
|||||||||| ... (580,074 bp) ... ||||||||||
ATTCAATAAT                                TTTTTTTATG
```

About CPK

CPK stands for Corey-Pauling-Koltun representation. Every atom is represented as a sphere, with radius proportional to its van der Walls radius. The usual color scheme is to represent carbon atoms in black, nitrogen in blue, oxygen in red and phosphorus atoms in pink.


Chromosome
Chromatid
Chromatid
Telomere
Centromere
Telomere
Nucleus
Cell
Histones
Base Pairs
DNA(double helix)



www.wehi.edu.au/education/wehi-tv/dna/movies/Chromosome_Coil.mov

About the animation

- Histone proteins attach to the DNA.
- Histones interact one with another to form a complex called nucleosome, but also forcing the DNA to wrap around it.
- The histone, nucleosome and DNA models were derived from their PDB (http://www.rcsb.org/pdb/) structures and other published data.
- Macromolecular structures cannot be directly oberserved. A molecular bond is between 1 and 2 Å (angstrom – 10^{-10} m) long, wave length in the visible spectrum are 400 to 700 nm (10^{-9} m).

Bioinformaticist's point of view

- Given DNA sequence information alone, predict the locations where the histones will be binding
- Knowing the location of the histones might help predicting the location of genes

Genome sizes

Species	**Size**
Potato spindle tuber viroid (PSTVd)	360
Human immunodeficiency virus (HIV)	9,700
Bacteriophage lambda (λ)	48,500
Mycoplasma genitalium (bacterium)	580,000
Escherichia coli (bacterium)	4,600,000
Drosophila melanogaster (fruit fly)	120,000,000
Homo sapiens (human)	3,000 000,000
Lilium longiflorum (easter lily)	90,000,000,000
Amoeba dubia (amoeba)	670,000,000,000

Genome sizes

- *Haemophilus influenzae* (bacterium), dna = 1.8 Mbp
- *Escherichia coli* (baterium), dna = 4.6 Mbp
- *Saccharomyces cerevisiae* (yeast), dna = 12 Mbp
- *Caenorhabditis elegans* (worm), dna = 97 Mbp
- *Arabidopsis thaliana* (flowering plant), dna = 115 Mbp
- *Drosophila melanogaster* (fruit fly), dna = 137 Mbp
- Smallest Human chromosome (Y), dna = 50 Mbp
- Largest Human chromosome (1), dna = 250 Mbp
- Whole Human genome, dna = 3 Gbp
- *Mus musculus* (mouse), dna = 3 Gbp.

$\Rightarrow$ Mbp = million base pairs

DNA is organized into chromosomes

The self-replicating genetic structures of cells containing the cellular DNA that bears in its nucleotide sequence the linear array of genes. In prokaryotes, chromosomal DNA is circular, and the entire genome is carried on one chromosome. Eukaryotic genomes consist of a number of chromosomes whose DNA is associated with different kinds of proteins.

$\Rightarrow$ Work by Thomas Morgan in the 1920s established the connection between traits (genes) and chromosomes (DNA).

Genome of multicellular animals (including human)

The human genome has two parts:

Nuclear genome: Consists of 23 pairs of chromosomes; for a total of 24 distinct linear molecules (22 autosomes and 2 sex chromosomes X and Y). The shortest chromosome consists of approximately 50 million nucleotides. The longest chromosome is more than 205 million nucleotides long. The sum of all the nucleotides is 3,2 billion nucleotides long. The nuclear genome encodes approximately 35,000 protein genes.

Mitochondrial genome: Consists of one circular molecule 16,569 nucleotides long, multiple copies of which are found in the organelles called mitochondria. The mitochondrial genome consists of 37 protein genes.

Each cell has its own "identical" copy of the genome

- The adult human body consists of approximately 10^{13} cell.
- Each cell has its own copy of the genome.

Human

- Most human cells are **diploid**, which means they have two copies of the 22 autosomes and two sex chromosomes (XX for females or XY for males).
- Diploid cells are also called **somatic** cells
- Sex cells (or **gametes**) are **haploid** and therefore have a single copy of the 22 autosomes as well as one sex chromosome.

Bioinformaticist's point of view

The distinction between somatic and sex cells will be important for the discussion on evolutionary events, which is important for the comparison of molecular sequences, more later.

Genes

What are the genes?

> *The fundamental physical and functional unit of heredity. A gene is an ordered sequence of nucleotides located in a particular position on a particular chromosome that encodes a specific functional product (i.e., a protein or RNA molecule).*

biotech.icmb.utexas.edu/search/dict-search.html
Can be several thousands nt (nucleotides) long.
Occurs on either stand, not often but sometimes overlapping.

Genome

What is a genome?

> *All the genetic material in the chromosomes of a particular organism needed create and maintain the organism alive.*

Can be several millions or even billion letters long.
Most genomes consists of DNA (deoxyribonucleic acids) molecules.
However, some pathogens (some viruses, viroids and sub-viral agents) are made up of ribonucleic acids (RNA).

Genome organisation

Without going into to much details, in higher organisms, the genes are broken into subsegments that are called **exons**. The segments are separated by intervening sequences that are called **introns**.
Genomes are not packed with genes.
Human genome organisation.

- Up to 60 % repetitive sequences
  - $\frac{1}{3}$ satellite DNA: low complexity, short and highly repeated
  - $\frac{2}{3}$ complex repeats: transposons, etc.
- Unique sequences;
  - 1.2 % protein-coding
  - 20 % introns

Genome organisation

- "About one-half of the platypus genome consists of interspersed repeats derived from transposable elements."
- Genome analysis of the platypus reveals unique signatures of evolution. Nature (2008) vol. 453 (7192) pp. 175-183

Bioinformaticist's point of view

Repetitive sequences are an obstacle for the algorithms involved in sequence assembly.
Repetitive sequences are often linked to diseases, therefore, the detection of repetitive sequences is in itself an important study.

Summary

- Cells are building blocks of life (two kinds of cells: prokaryotic, eukaryotic);
- Three kinds of macromolecules (DNA, RNA, proteins);
- Nucleotides are the building blocks of DNA;

Reverse engineering analogy by Ken Fasman

A computer device of unknown origin has been found, here is what we know about the device.

- Its hard disk contains 3 Gb of information divided into 23 partitions
- The operating system is unknown
- The file format is unknown;
- Encoding is known (words of 3 bases)
- Files are fragmented
- The hardware design and BIOS specification are stored on the disk
- The disk contains pieces of deleted files but we don't know which are active and which are not

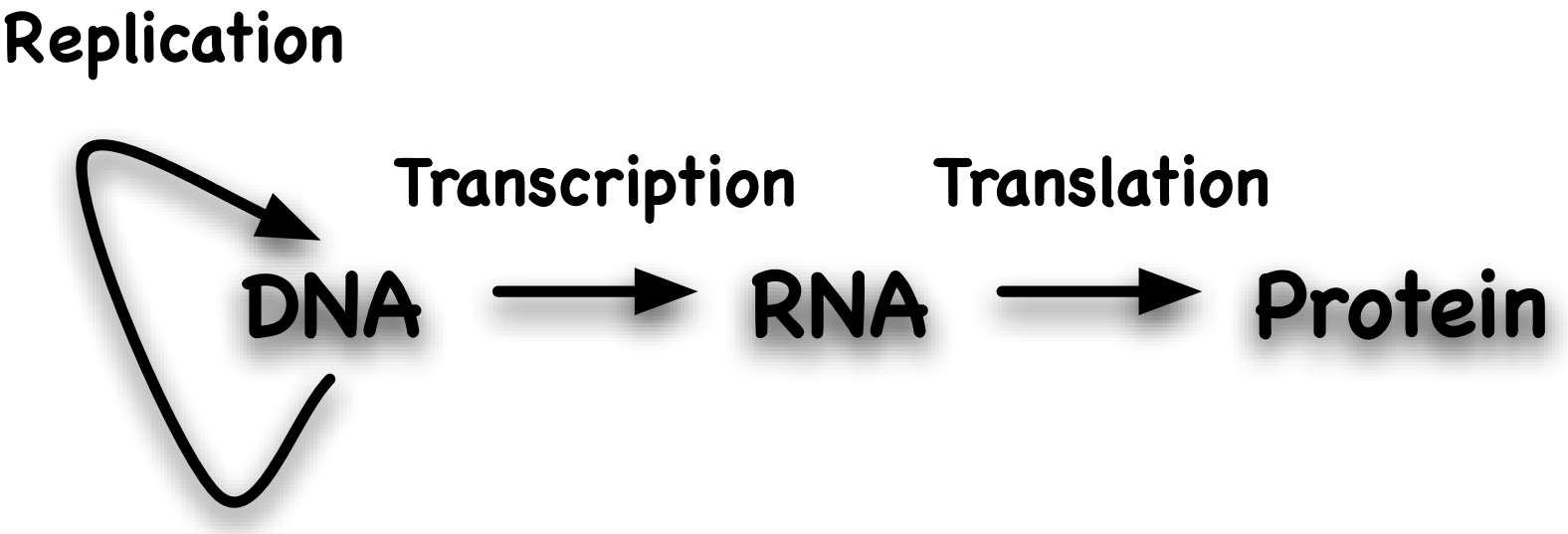
Reverse engineering analogy by Ken Fasman (contd)

It has been estimated that the operating system is made of $30,000$ to $100,000$ programs, we need to understand its functioning! Reading the content of a hard disk is a simple task, reading DNA sequences is not!

Bioinformaticist's point of view

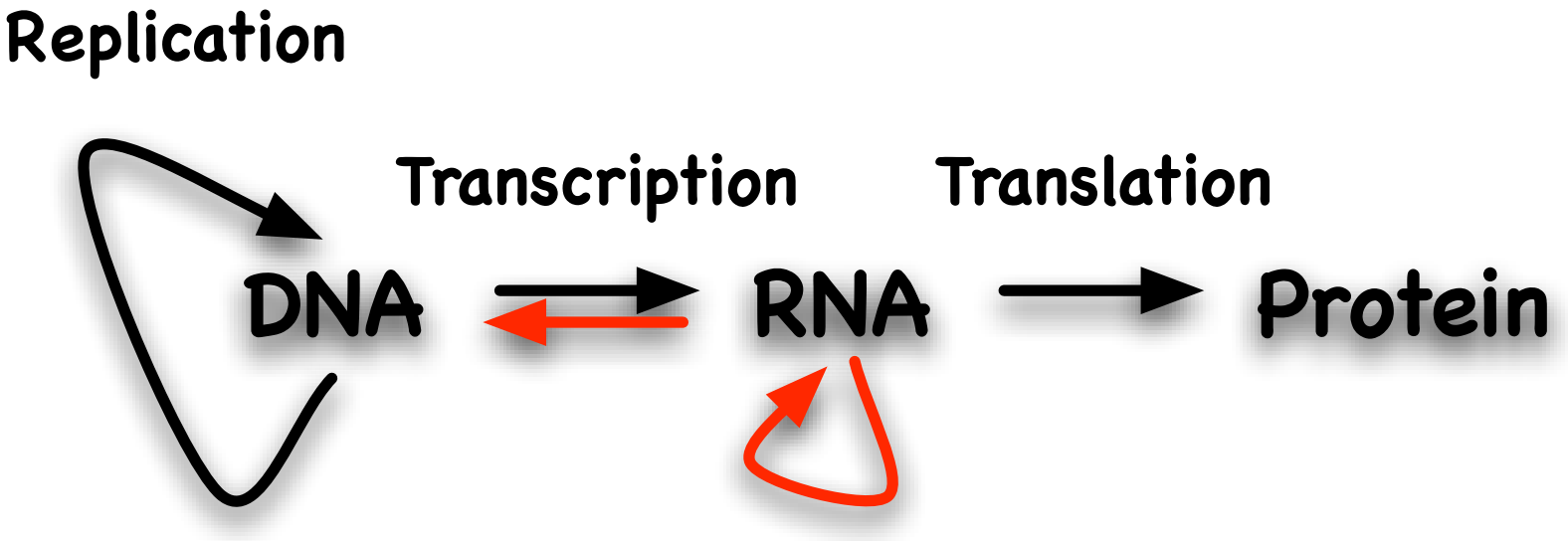
- DNA Sequencing (traditional or high-throughput)
- Gene finding (stochastic grammatical models)
- Identifying signals (pattern discovery)

Central Dogma (1958)



⇒ Francis Crick (1958) *Symposium of the Society of Experimental Biology* **12**:138-167.

Essential Cell Biology: Central Dogma (2009)



Central Dogma (contd)

DNA: stores genetic information (library of programs);

RNA: stores a copy a gene during protein synthesis (mRNA), adapter molecule involved proteins synthesis (tRNA), part of the ribosome (a ribo-protein complex), regulation/development (micro-RNAs, regulatory motifs, riboswitches, etc.);

Proteins: catalyse reactions (modulator), communication (signalling), transport, structure, etc.

Replication: DNA $\longrightarrow$ DNA

- Replication is catalyzed by an enzyme (protein) called DNA polymerase.
- The complementarity of the base pairs is fundamental to DNA replication mechanisms.
- Each strand of a DNA molecule serves as a template for producing a complementary copy.
- The result is two double helices identical to their parent; each daughter molecule has one strand of its parent (this is called a **semiconservative** system).
- It is a complex process (timing, topology, distribution to daughter cells). Some of its important steps were understood in the 1980s whilst the details are still an active research topic.
- Remember higher levels of organization of DNA?!

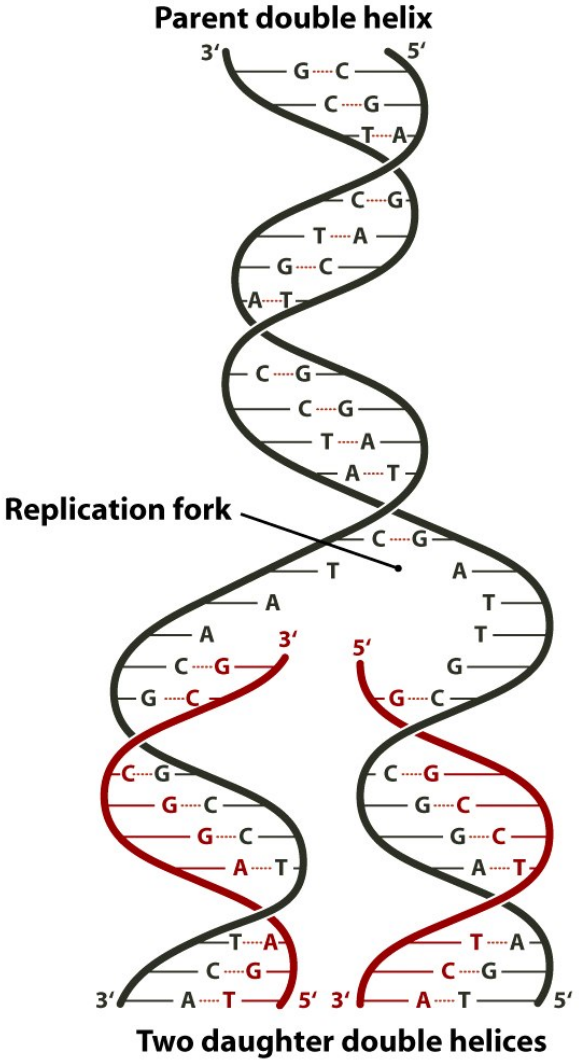


Figure 15-1 Genomes 3 (© Garland Science 2007)

Initiation of genome replication

- Replication is not initiated at random sites but rather at specific sites called **origin of replication**.
- The origin of replication is a particular sequence motif, which in the case of *E. coli* spans about 245 nucleotides.
- Type I DNA topoisomerase breaks one of the strands, the replication fork moves, DNA synthesis continues, repeat!
- Two replication forks move in opposite direction.
- The human genome comprises approximately 20,000 origins of replication.
- Replication origins in higher eukayryotes is still an active research subject.

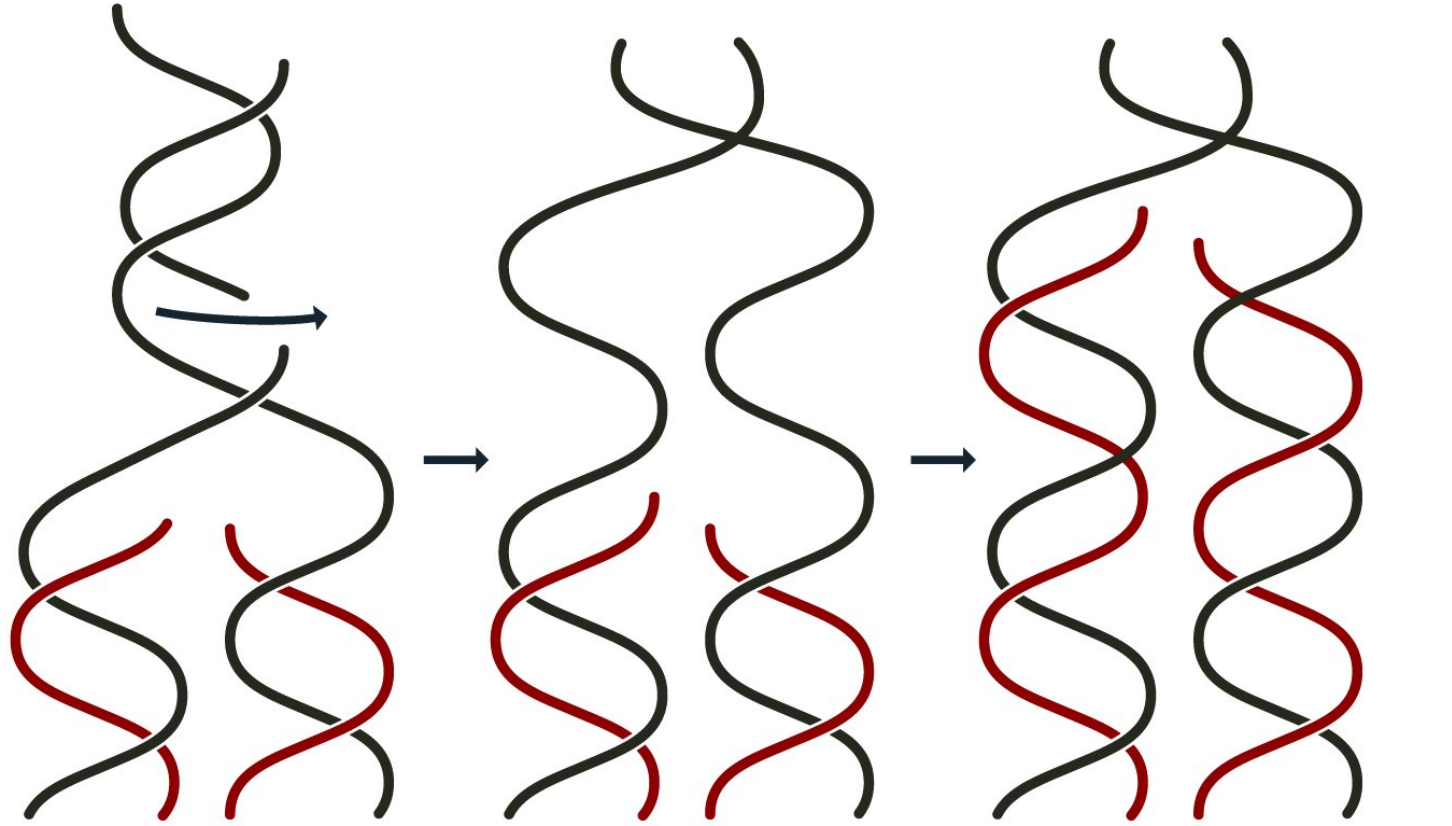


Figure 15-5 Genomes 3 (© Garland Science 2007)

Elongation phase of replication

Problems:

- DNA polymerase synthesize DNA in the $5' \rightarrow 3'$ direction;
- DNA polymerase requires a primer (short double stranded region) for the initiation of the synthesis.

The first point implies that only one strand, called the **leading strand**, is copied in a continuous manner. The other stand, called the **lagging strand**, is copied as discontinuous segments that are later ligated (connected).
The enzyme primase (a special RNA polymerase) synthesize the primers in bacteria. In eukaryotes, it is primase and DNA polymerase α that synthesize the initial primer.

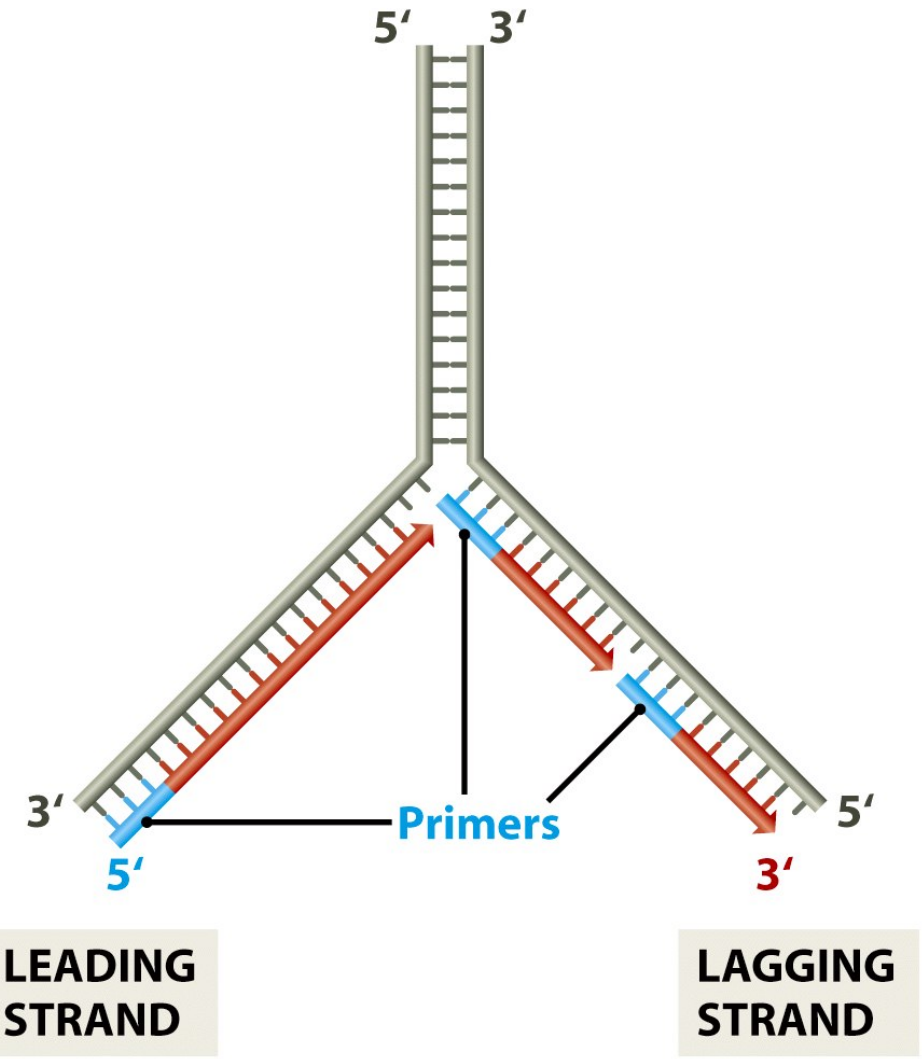


Figure 15-12 Genomes 3 (© Garland Science 2007)

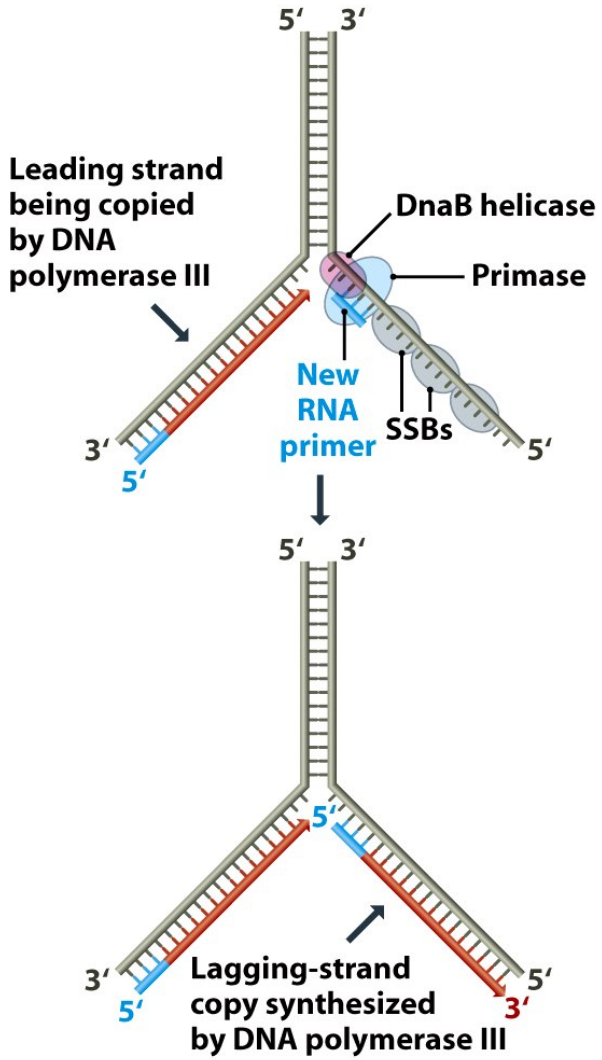


Figure 15-16 Genomes 3 (© Garland Science 2007)

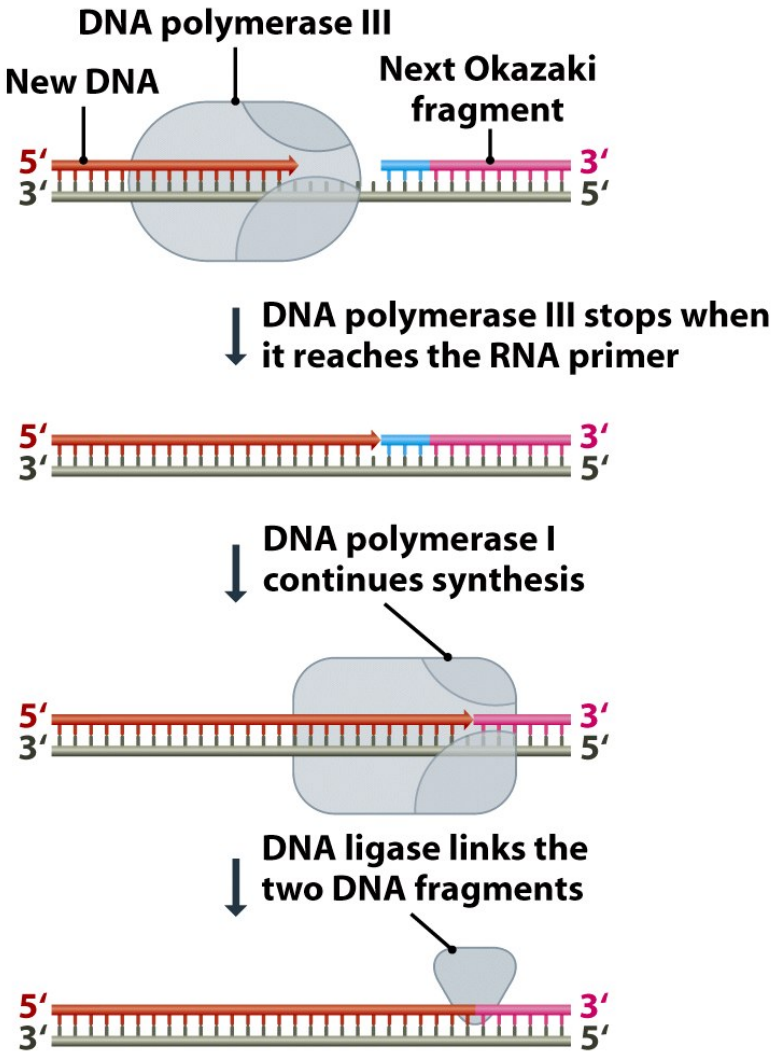
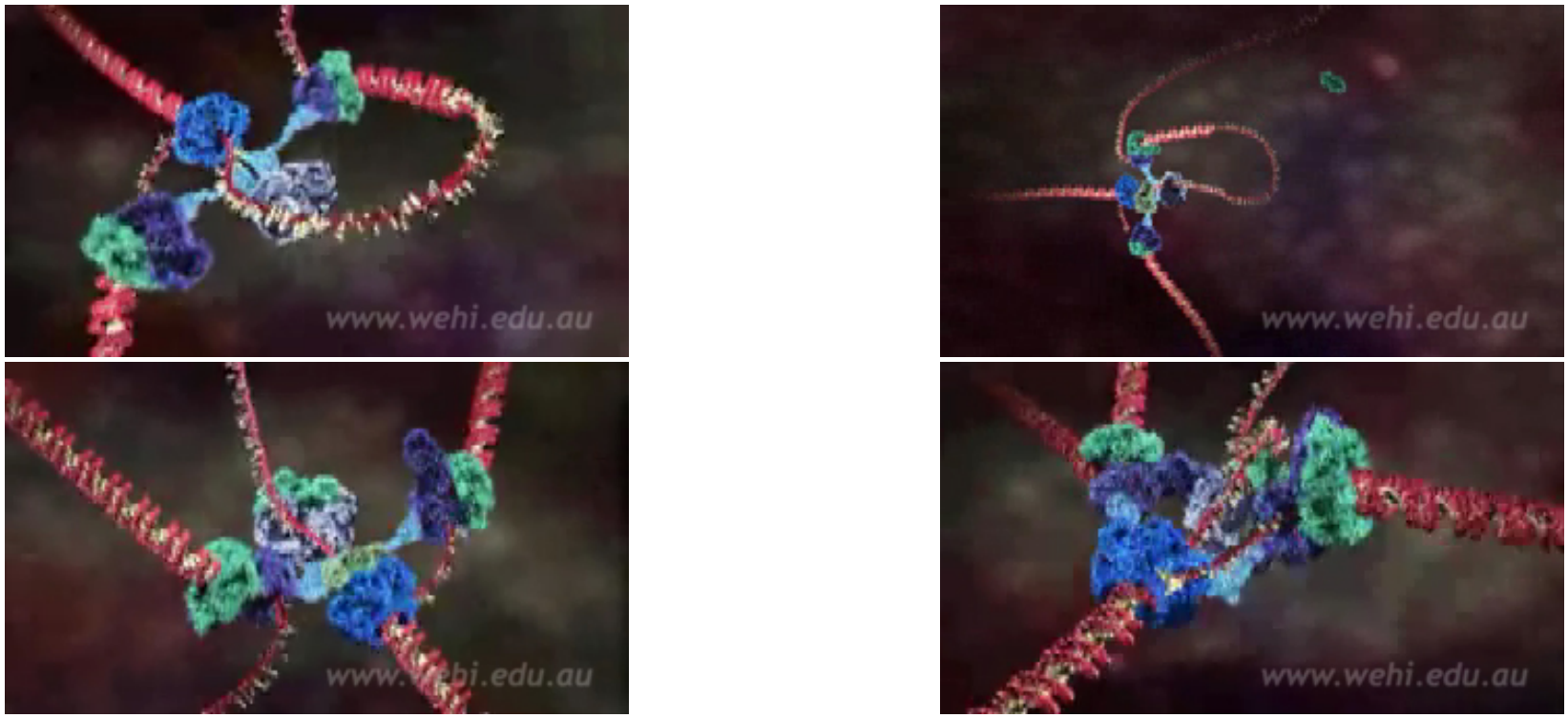


Figure 15-18 Genomes 3 (© Garland Science 2007)

Replication

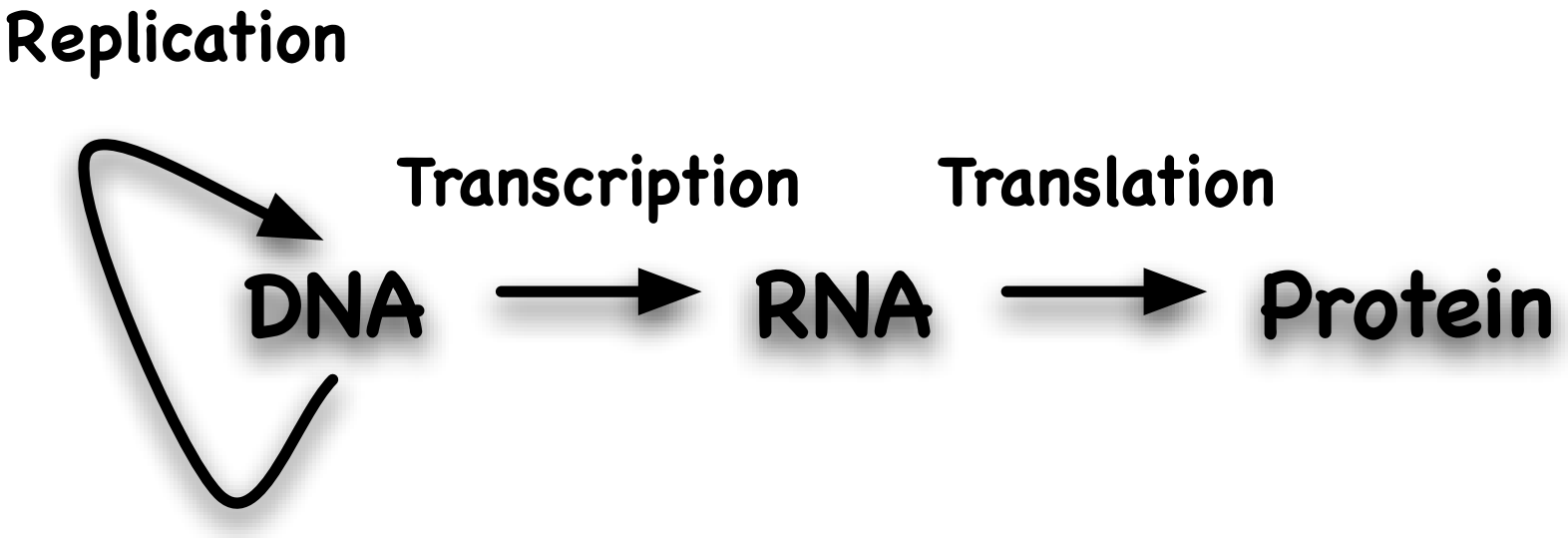
- DNA polymerase makes about 1 error in every 10^6 or 10^7 nucleotides. This is the primary enzyme involved in DNA replication in *E. coli*.
- But this enzyme also has a "proof reading" activity which corrects these errors.
- Further DNA repair mechanisms exist so that the overall error rate is about 1 in 10^{10} or 10^{11} nucleotides.
- On average, one replication error occurs every $1,000$ copies!

Animation: Replication



$\Rightarrow$ www.wehi.edu.au/education/wehi-tv/dna/replication.html

Central Dogma: transcription



Genes

"(...) a gene is a sequence of genomic DNA (...) that is essential for a specific function." Li & Graur 1991.
There are three (3) kinds of genes:

1. Protein-coding genes;
2. RNA-coding genes;
3. Regulatory genes.

1 & 2 are called structural gene (only 1 for some authors).
The genome is the sum of all the genes.

Transcription (contd)

Transcription of prokaryotic genes is under the control of one type of RNA polymerase, while 3 are involved in this process for the eukaryotic genes (rRNA by RNA polymerase I, **protein-coding genes by RNA polymerase II**, while small cytoplasmic RNA genes, such as tRNA-specifying genes are under the control of RNA polymerase III, small nuclear RNA genes are transcribed by RNA polymerase II and/or III (U6 transcribed by II or III)).

Transcription: DNA $\longrightarrow$ RNA

The need for an intermediate molecule. In Eukaryotes, it had been observed that proteins are synthesised in the cytoplasm (inside the cell but outside of the nucleus), whereas DNA is found in the nucleus.

- Carried out by a (DNA-dependent) RNA polymerase;
- Requires the presence of specific sequences (called signals) upstream of the start of transcription (in the case of protein-coding genes). This region is called the promoter;
- In Eukaryotes, the messenger RNA contains non-coding regions, called introns, that are removed through various processes, called intron splicing. Before splicing the transcript is called a pre-mRNA.

The collection of the transcripts is called the **transcriptome**.

DNA-RNA relationship

```
DNA: ... TAACCTACCGCGCCTATTACTGCCAGGAAGGAACTTGATC ...

DNA: ... TAACCTACCGCGCCTATTACTGCCAGGAAGGAACTTGATC ...
                 |||||
RNA:             AUGGC

DNA: ... TAACCTACCGCGCCTATTACTGCCAGGAAGGAACTTGATC ...
                 ||||||
RNA:             AUGGCG ...

...

DNA: ... TAACCTACCGCGCCTATTACTGCCAGGAAGGAACTTGATC ...
                 ||||||||||||||||||||||||||||||
RNA:             AUGGCGCCGAUAAUGUCGGUCCUUCCUUGA
```

Transcription (contd)

Conceptually simple, one to one relationship between each nucleotide of the source and the destination.

- G pairs with C;
- A pairs with U (not T);
- Uses ribonucleotides; instead of deoxyribonucleotides;

The result (product) is called a (pre-)messenger RNA or transcript.

Transcription (contd)

I don't understand, is it the whole of the genome that is transcribed? No, translation is is not initiated randomly but at specific sites, called promoters.
Here is the consensus sequence for the core promoter in *E. coli* (*Escherichia coli*):

`TTGACA(N){16,18}TATAAT`

What is the likelihood of this motif to occur?

Transcription (contd)

- Here size do matter, and it depends on your assumptions. How do you want to model the promoter sequence motif?
- The simplest model is *i.i.d.*, which stands for **independent and identically distributed**.
- What does it mean?
- First, since the positions are considered to be independent one from another, the probability of the motif is the product of the probabilities of occurrence of the nucleotides at each position.
- Second, we also assume that the probability distribution for the nucleotides is the same for all the positions.
- In general, the maximum likelihood estimators are used to estimated the probability distributions, which simply means that a large number of examples are collected and that the frequencies of occurrence are used as estimators.

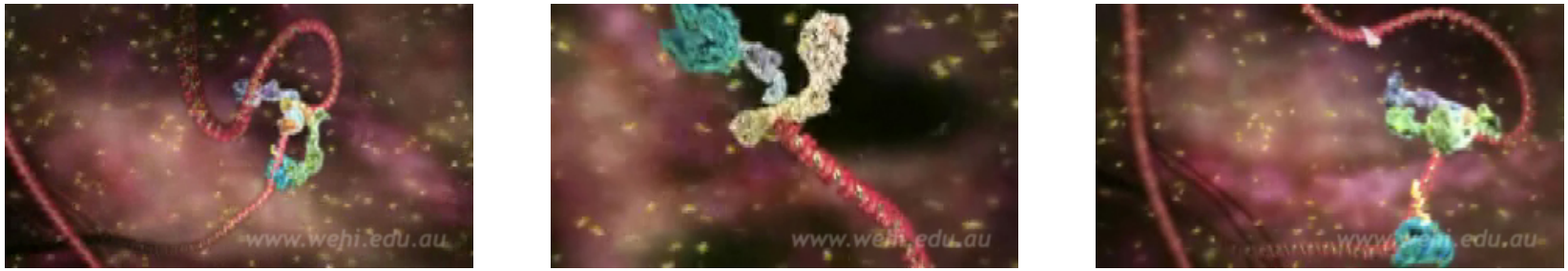
`TTGACA(N){16,18}TATAAT`

- To make the argument simple, we can assume the events to be equally likely, $p_A = p_C = p_G = p_T = \frac{1}{4}$, so that the probability of the motif is $\frac{1}{4^{12}} = 6 \times 10^{-8}$.
- How many promoters would you expect to find in the *E. Coli* genome?
- $6 \times 10^{-8} \times 4.6\ \mathrm{Mb} = 0.276 < 1$.
- Eukaryotic genomes are larger, often billions of bp, and accordingly their promoter sequence is more complex!
- Finally, other regulatory sequences exist, which are the binding site for regulatory proteins, which can enhance the transcription, positive regulation, or inhibit transcription, negative regulation.

Bioinformaticist's point of view

The discovery of (new) regulatory motifs (promotors, signals, etc.) is an active area of research.

Animation: Transcription initiation



$\Rightarrow$ http://www.wehi.edu.au/education/wehi-tv/dna/dogma.html

Animation: Transcription (contd)

Transcription factors assemble at a DNA promoter region found at the start of a gene. Promoter regions are characterised by the DNA's base sequence, which contains the repetition TATATA and for this reason is known as the "TATA box".
The TATA box is gripped by the transcription factor TFIID (yellow-brown) that marks the attachment point for RNA polymerase and associated transcription factors. In the middle of TFIID is the TATA Binding Protein subunit, which recognises and fastens onto the TATA box. It's tight grip makes the DNA kink 90 degrees, which is thought to serve as a physical landmark for the start of a gene.

Animation: Transcription (contd)

A mediator (purple) protein complex arrives carrying the enzyme RNA polymerase II (blue-green). It manoeuvres the RNA polymerase into place. Other transcription factors arrive (TFIIA and TFIIB - small blue molecules) and lock into place. Then TFIIH (green) arrives. One of its jobs is to pry apart the two strands of DNA (via helicase action) to allow the RNA polymerase to get access to the DNA bases.
Finally, the initiation complex requires contact with activator proteins, which bind to specific sequences of DNA known as enhancer regions. These regions can be thousands of base pairs away from the initiation complex. The consequent bending of the activator protein/enhancer region into contact with the initiation-complex resembles a scorpion's tail in this animation.

Animation: Transcription (contd)

The activator protein triggers the release of the RNA polymerase, which runs along the DNA transcribing the gene into mRNA (yellow ribbon).

$\Rightarrow$ http://www.wehi.edu.au/education/wehi-tv/dna/dogma.html

Animation: Transcription

The RNA polymerase unzips a small portion of the DNA helix exposing the bases on each strand. One of the strands acts as a template for the synthesis of an RNA molecule. The base-sequence code is transcribed by matching these DNA bases with RNA subunits, forming a long RNA polymer chain.

$\Rightarrow$ http://www.wehi.edu.au/education/wehi-tv/dna/dogma2.html

Transcriptome and gene regulation

- Messenger RNA are degraded minutes (prokaryotes) or hours (eukaryotes) after synthesis.
- Furthermore, information stored in the untranslated regions of the transcript is involved in regulation and transport.

Transcriptome

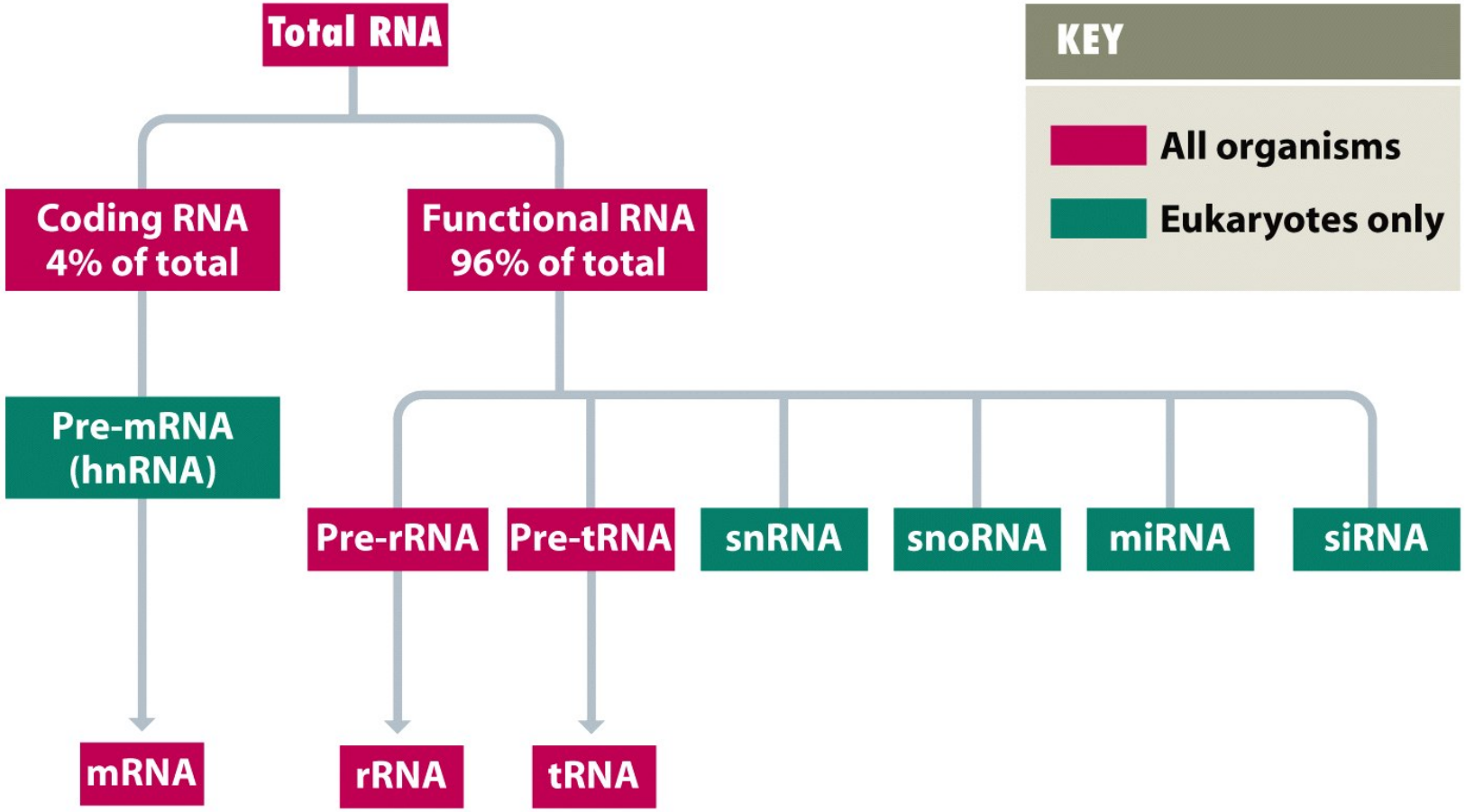
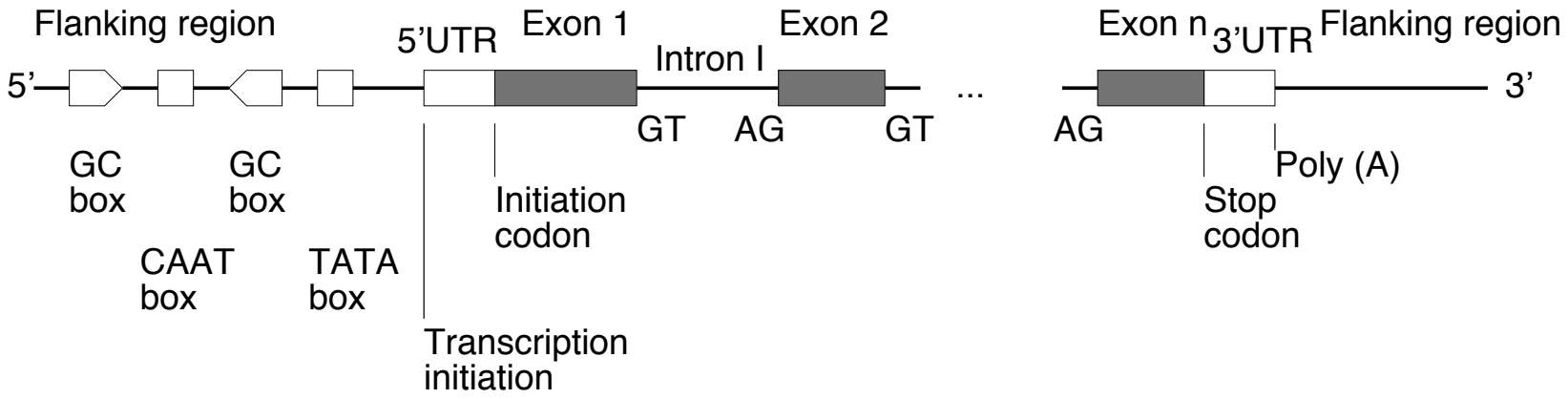


Figure 1-12 Genomes 3 (© Garland Science 2007)

Protein-coding gene structure



- 5' and 3' flanking region (non-transcribed).
- Contains several signals regulating the gene expression; where, when, what, rate.

Protein-coding gene structure

- A signal, in this context, is a specific sequence (sub-string). Here the signals promote the transcription of the gene and accordingly are called promoters.
- The 5' flanking region is called the promoter region.
- TATA box 19-17 bp upstream of the transcription starting point, the CAAT box, one or more GC box (GGGCGGG). CAAT + GC boxes act as a landing zone for the RNA polymerase (control the initial binding), while the TATA box determines the choice of starting point.
- There are no mandatory signals, some don't have a TATA box, CAAT, ...

Protein-coding gene structure

- The beginning and ending positions are somewhat fuzzy; i.e. we don't know yet (active bioinformatics/biology research topic).
- The result of the transcription is called a transcript or messenger RNA.

Protein-coding gene structure

In Eukaryotes, genes contain protein coding and non-coding regions, called exons and introns respectively. The initial transcript is called a pre-mRNA.
In Prokaryotes, the genes are continuous sequences.
The process by which introns are excised is called splicing, which occurs in the nucleus. The result is a mature RNA.
There exist several mechanisms to remove, splice, introns.

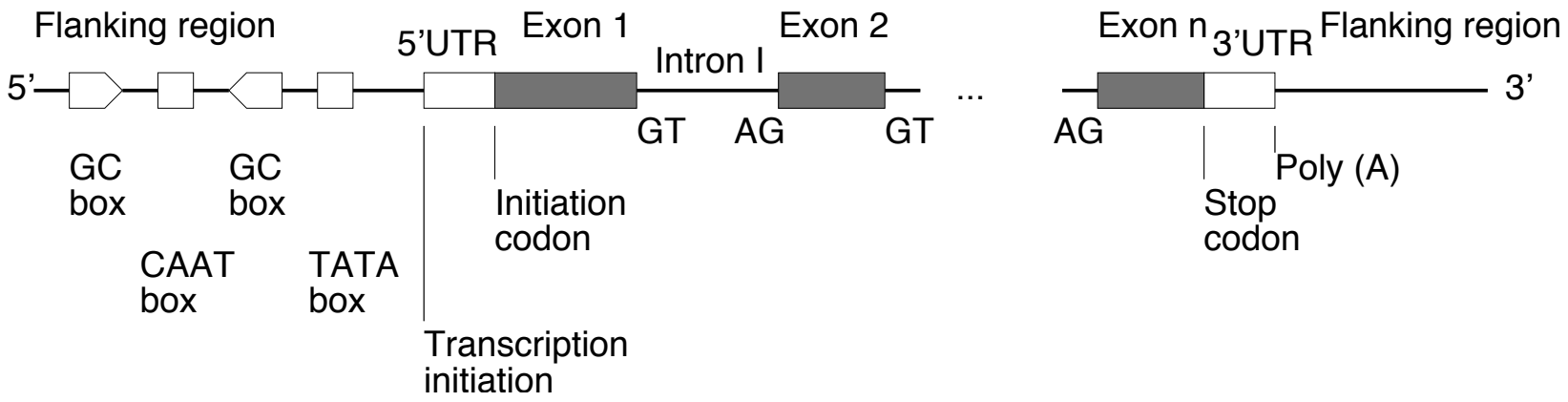
Protein-coding gene structure

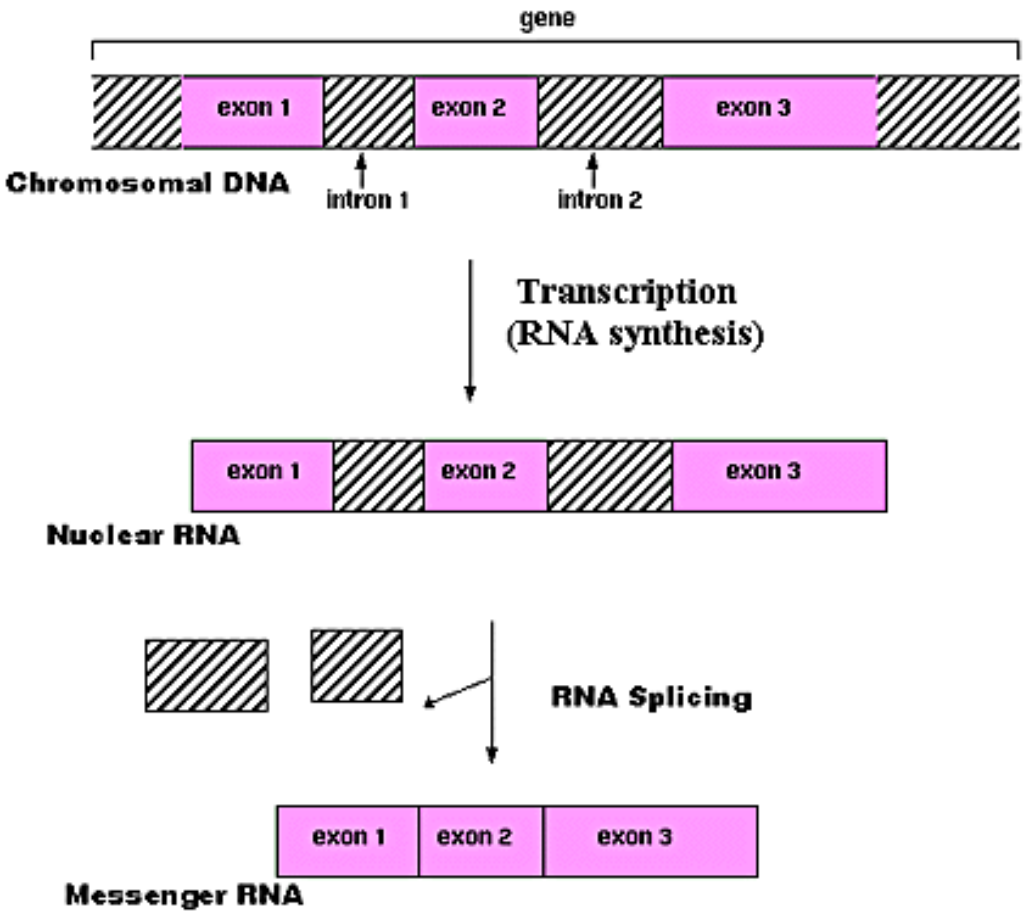
The junction between exons and introns is called the splice site. Generally an intron start with GT and ends with AG (donor and acceptor sites, GT-AG rule). Two nucleotides are not enough to characterize a slice site, i.e. you would expect to find GT once every 16 nucleotides, so the surrounding nucleotides are also important.

Why so much variation? **It's all a matter of regulation, the genes that need to be produced in large quantities will tend to have a strong promoter, others will have a weaker one.**

Gene regulation is a hot research topic!

Structure of an Eukaryotic protein-coding gene





RNA synthesis and processing

Translation: RNA $\longrightarrow$ Protein

- This step is called translation and it is under the control of a riboprotein complex called the **ribosome**, adapter RNA molecules, called tRNAs, and several other proteins to control the regulation, charging tRNA molecules with the appropriate amino acids.
- It is clear that what ever coding principle exists, there cannot be a one-to-one mapping! $4^1 < 20, 4^2 < 20, 4^3 > 20$!
- For each consecutive three nucleotide, this is called a codon (coding unit), correspond a unique amino acid.

$$4 \times 4 \times 4 = 64$$

- Contiguous, non-overlapping triplets.
- Since there are 64 possible codons, the code is said to be **degenerated**, i.e. several triples map onto the same amino acid.

Universal Genetic Code

	U		C		A		G		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
U	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	C
U	UUA	Leu	UCA	Ser	UAA	*Stop*	UGA	*Stop*	A
U	UUG	Leu	UCG	Ser	UAG	*Stop*	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
C	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
C	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
C	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
A	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
A	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
A	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg	G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
G	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
G	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
G	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

DNA-RNA-Protein relationships

```
    DNA: TAC CGC GCC TAT TAC TGC CAG GAA GGA ACT
    RNA: AUG GCG CCG AUA AUG ACG GUC CUU CCU UGA
Protein:   M   A   P   I   M   T   V   L   P   *
```

```
    DNA: TAC CGC GCC TAT TAC TGC CAG GAA GGA ACT
    RNA: AUG GCG CCG AUA AUG ACG GUC CUU CCU UGA
Protein: Met Ala Pro Ile Met Thr Val Leu Pro Stop
```

$\Rightarrow$ Example from Jones & Pevzner, p. 65.

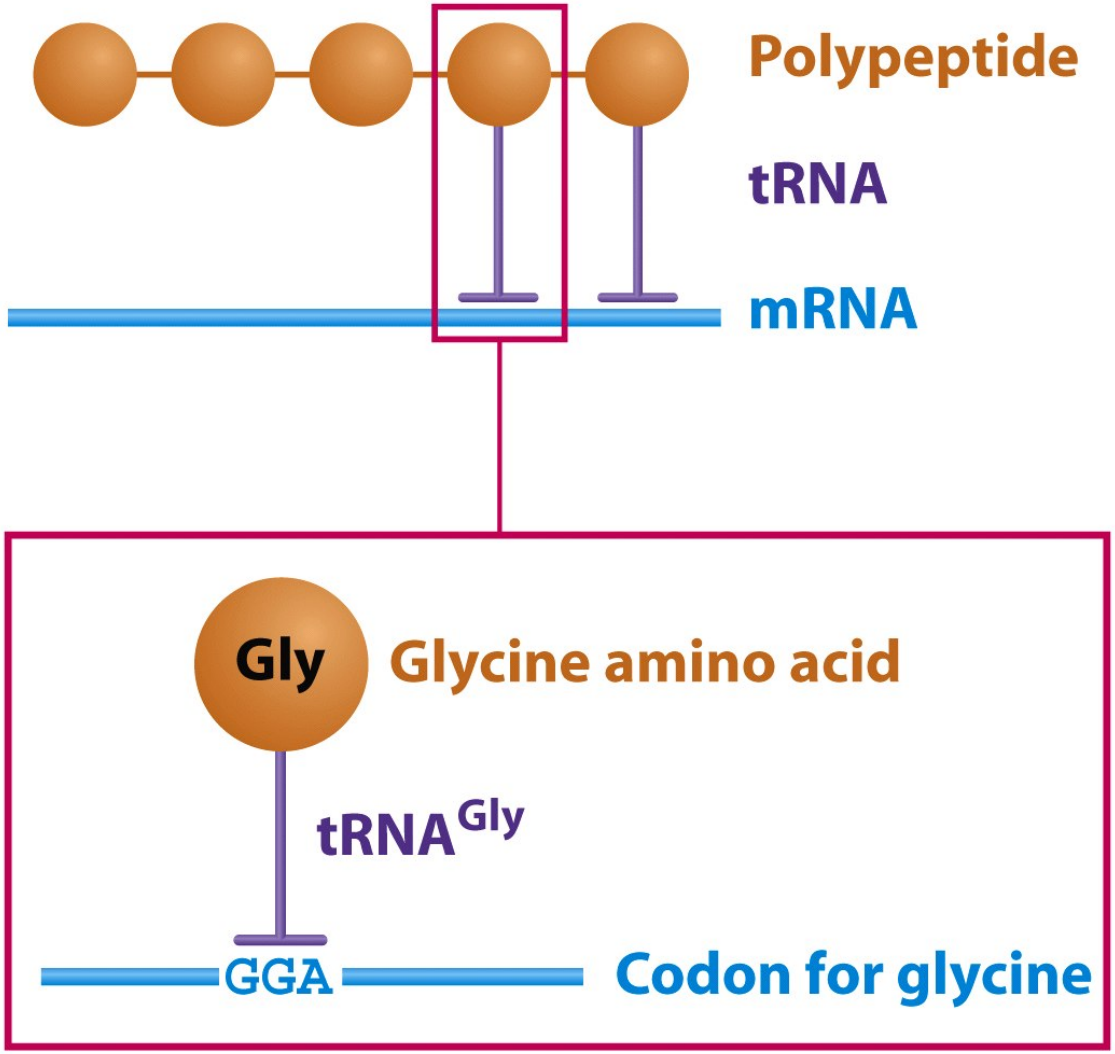
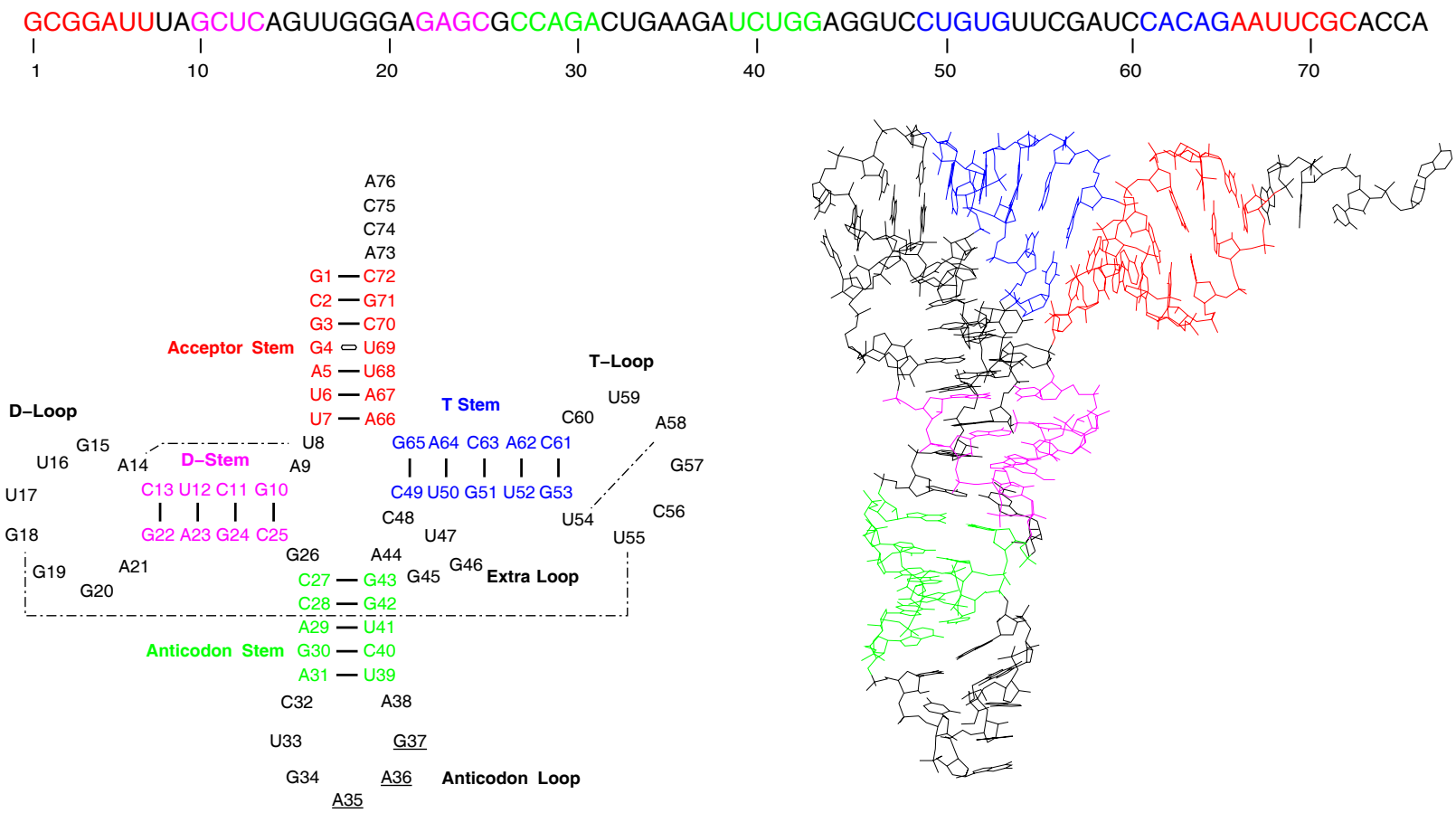


Figure 13-1 Genomes 3 (© Garland Science 2007)

tRNA: 1, 2, 3


GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA
1
10
20
30
40
50
60
70
A76
C75
C74
A73
G1 — C72
C2 — G71
G3 — C70
Acceptor Stem
G4 — U69
A5 — U68
U6 — A67
U7 — A66
T Stem
T–Loop
U59
C60
A58
D–Loop
G15
U8
G65 A64 C63 A62 C61
G57
U16
A14
D–Stem
A9
C13 U12 C11 G10
C49 U50 G51 U52 G53
U17
C48
C56
U54
G18
G22 A23 G24 C25
U47
U55
G26
A44
G46
G19
A21
C27 — G43
G45
Extra Loop
G20
C28 — G42
A29 — U41
Anticodon Stem
G30 — C40
A31 — U39
C32
A38
U33
G37
G34
A36
Anticodon Loop
A35

The transfer RNAs (tRNAs) are adaptor molecules. Bacteria have 30 to 45 different adaptors whilst some eukaryotes have up to 50 (48 in the case of humans). **Each tRNA is loaded (charged) with a specific amino acid at one end, and has a specific (triplet) sequence, called the anti-codon, at the other end.**
Notation: tRNA^{Phe} is a tRNA molecule specific for phenylalanine (one of the 20 amino acids).
The tRNA molecules are 70 to 90 nt long and virtually all of them fold into the same cloverleaf structure presented on the previous slide.

As will be seen next, it is quite important that **all the tRNAs have a similar structure** so that one molecular machine (the ribosome) can be used for the protein synthesis.
The enzymes responsible for "charging" the proper amino acid onto each tRNA are called aminoacyl-tRNA synthetases. Most organisms have 20 aminoacyl-tRNA synthetases, meaning that a given aminoacyl-tRNA synthetase is responsible for the attachment of a specific amino acid on all the isoacepting tRNAs (different tRNAs charged with the same amino acid type).
Each tRNA also has unique features so that it gets loaded with the right amino acid.

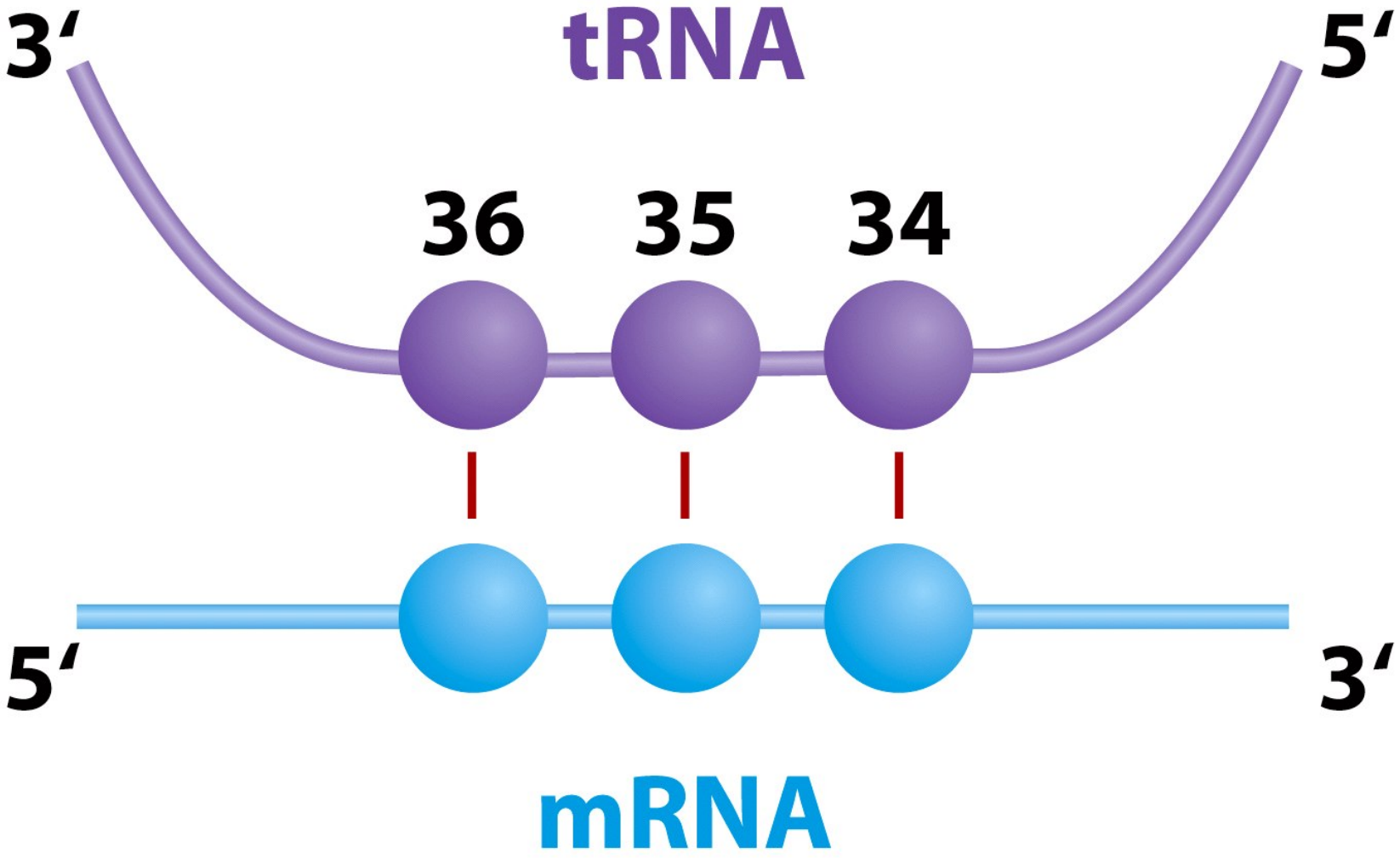


Figure 13-6 Genomes 3 (© Garland Science 2007)

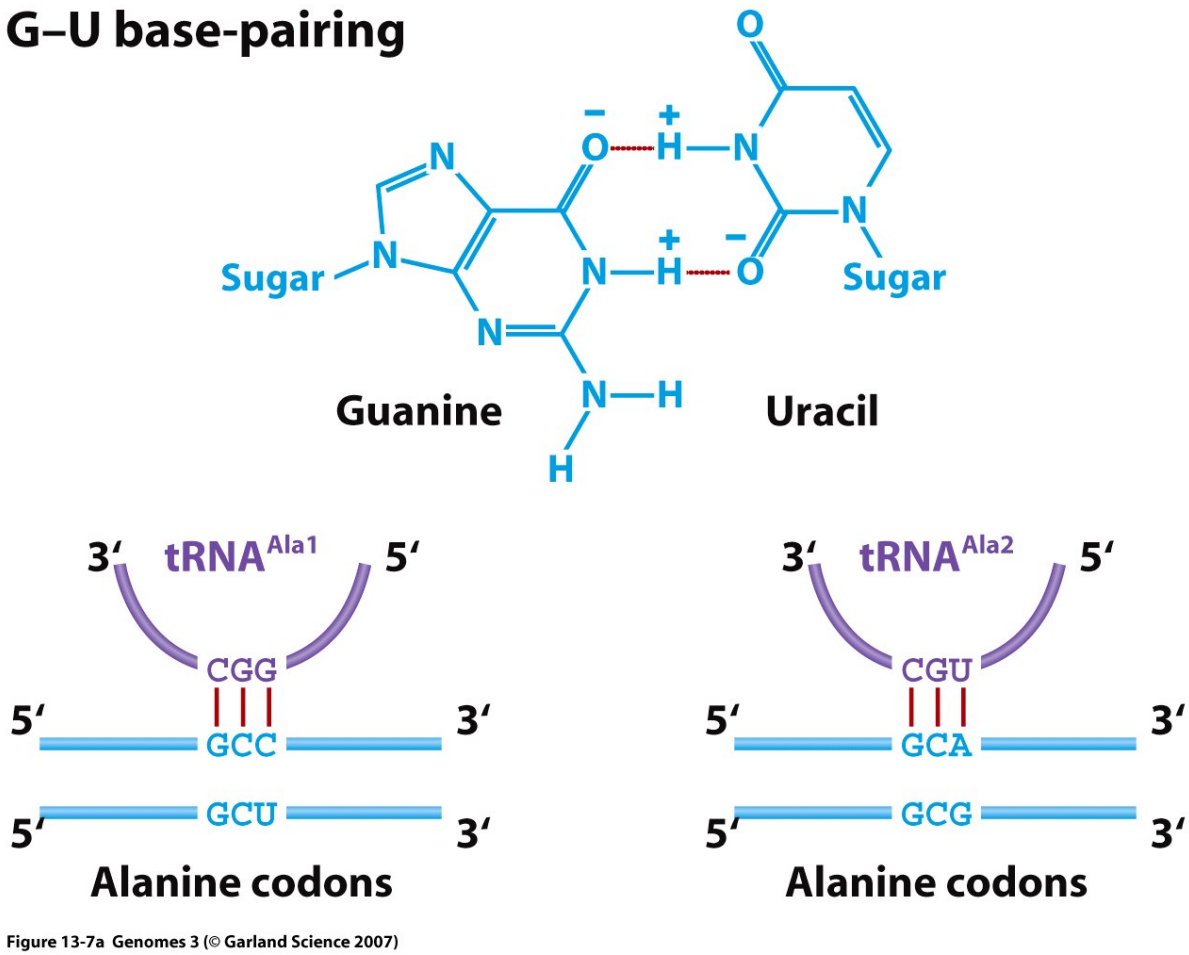


Figure 13-7a Genomes 3 (© Garland Science 2007)

Wobble base pairs are possible and reduce the number of tRNAs needed since the same tRNA binds 2 or possibly 3 codons.

Ribosomes play an essential role in translation

Large RNAs + proteins complex (the result of the association of 3 to 4 RNAs + 55 to 83 proteins!). In bacteria, there are approximately 20,000 ribosomes at any given time (more in eukaryotes).

- Coordinate protein synthesis by orchestrating the placement of the messenger RNAs (mRNAs), the transfer RNAs (tRNAs) and necessary protein factors;
- Catalyze (at least partially) some of the chemical reactions involved in protein synthesis.

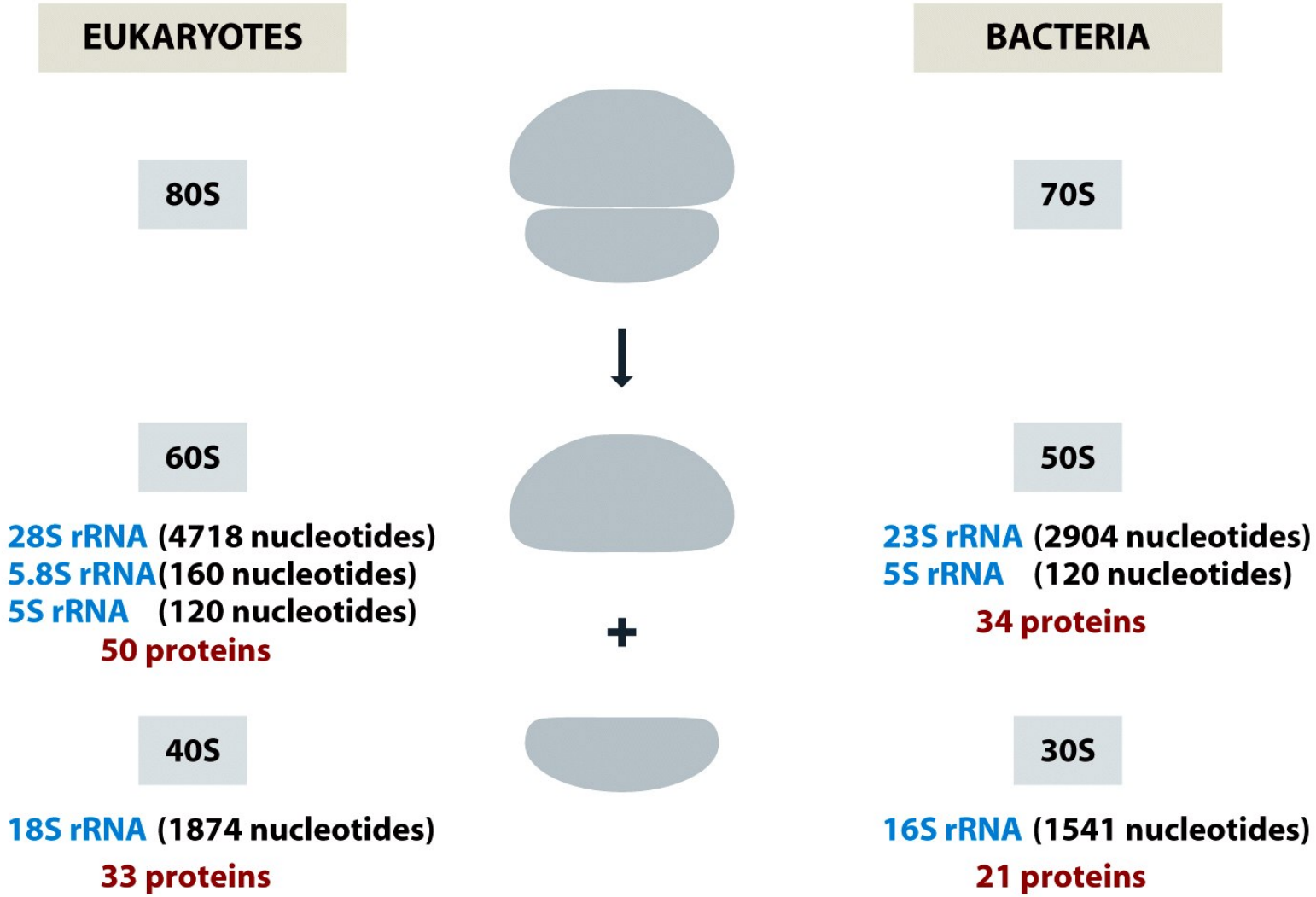


Figure 13-10 Genomes 3 (© Garland Science 2007)

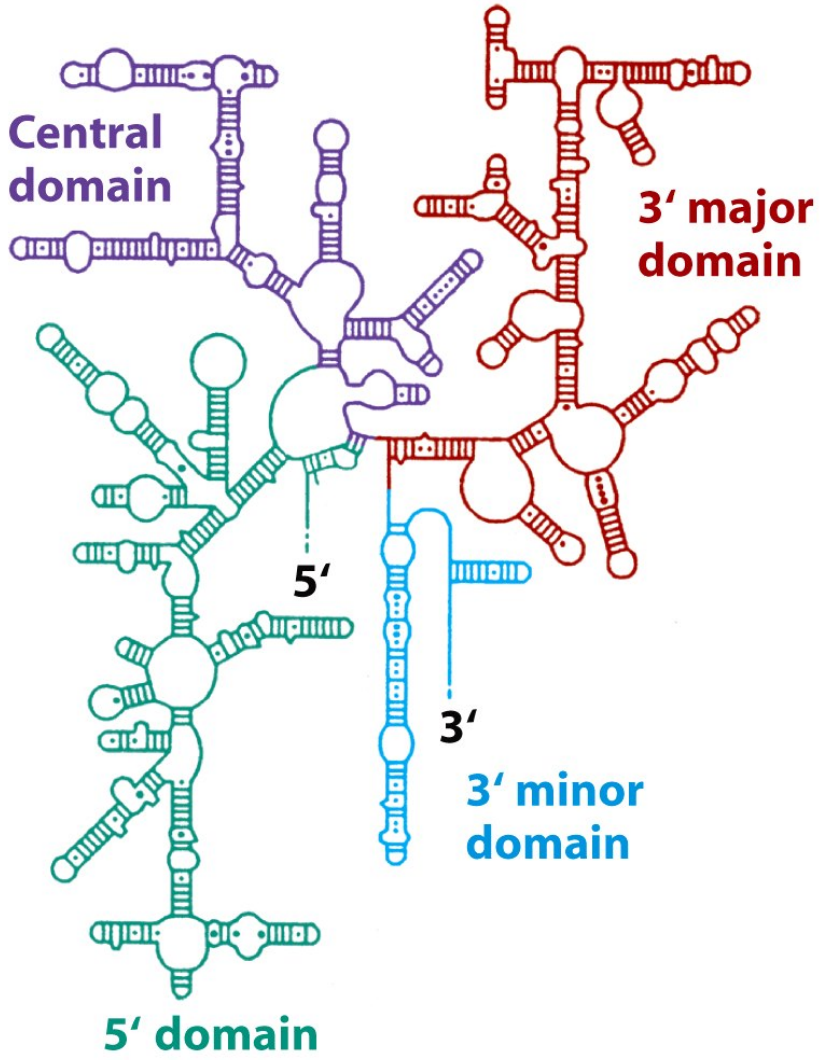


Figure 13-11 Genomes 3 (© Garland Science 2007)

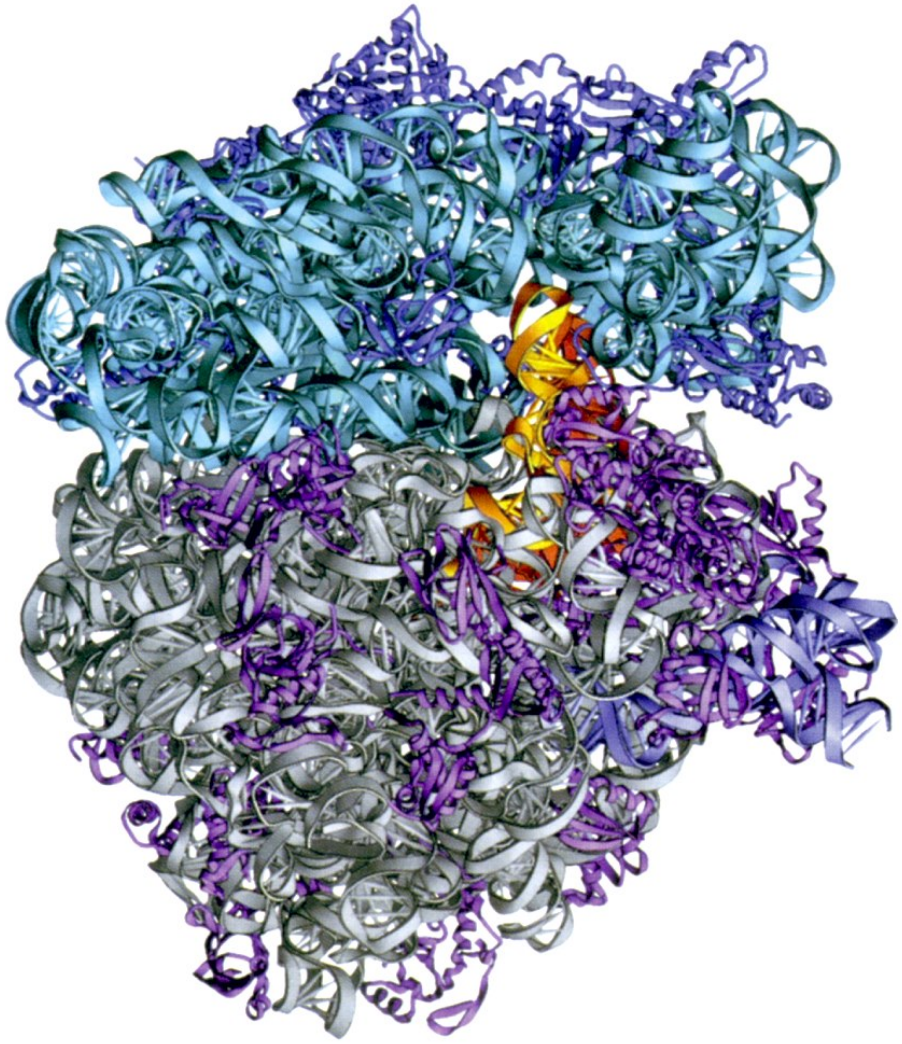


Figure 13-13 Genomes 3 (© Garland Science 2007)

Animation: Translation

http://www.wehi.edu.au/education/wehi-tv/dna/movies/Translation.mov

Animation: Translation (contd)

The message in mRNA (yellow) is decoded inside the ribosome (purple and light blue) and translated into a chain of amino acids (red).
The ribosome is composed of one large (purple) and one small subunit (light blue), each with a specific task to perform. The small subunit's task is to match the triple letter code, known as a codon, to the anticodon at the base of each tRNA (green). The large subunit's task is to link the amino acids together into a chain. The amino acid chain exits the ribosome through a tunnel in the large subunit, then folds up into a three-dimensional protein molecule.

Animation: Translation (contd)

As the mRNA is ratcheted through the ribosome, the mRNA sequence is translated into an amino acid sequence. The sequence of mRNA condons determines the specific amino acids that are added to the growing polypeptide chain. Selection of the correct amino acid is determined by complimentary base pairing between the mRNA's codon and the tRNA's anticodon. The codons are shown in this animation during the close up of the mRNA entering the ribosome. The codons are indicated as triplet groups of yellow-brown bases.
tRNA (green) is a courier molecule carrying a single amino acid (red tip) as its parcel.

Animation: Translation (contd)

During the amino acid chain synthesis, the tRNA steps through three locations inside the ribosome, referred to as the A-site, P-site and E-site. tRNA enters the ribosome and lodges in the A-site, where it is tested for a correct codon-anticodon match. If the tRNA's anticondon correctly matches the mRNA condon, it is stepped through to the P-site by a conformational change in the ribosome. In the P-site the amino acid carried by the tRNA is attached to the growing end of the amino acid chain.

Animation: Translation (contd)

The addition of amino acids is a three step cycle

1. The tRNA enters the ribosome at the A-site and is tested for a codon-anticodon match with the mRNA;
2. If it is a correct match, the tRNA is shifted to the P-site and the amino acid it carries is added to the end of the peptide chain. The mRNA is also ratcheted three nucleotides (1 codon);
3. The spent tRNA is moved to the E-site and then ejected from the ribosome.

Animation: Translation (contd)

A typical eukaryotic cell contains millions of ribosomes in its cytoplasm.
Many details, such as elongation factors (eg EFTu), have been omitted from this animation. This animation represents an idealised system with no incorrect tRNAs entering the ribosome, and consequently no error correction at the A-site.

http://www.wehi.edu.au/education/wehi-tv/dna/dogma3.html

DNA-RNA-Protein relationships

```
    DNA: TAC CGC GCC TAT TAC TGC CAG GAA GGA ACT
    RNA: AUG GCG CCG AUA AUG ACG GUC CUU CCU UGA
Protein:   M   A   P   I   M   T   V   L   P   *
```

```
    DNA: TAC CGC GCC TAT TAC TGC CAG GAA GGA ACT
    RNA: AUG GCG CCG AUA AUG ACG GUC CUU CCU UGA
Protein: Met Ala Pro Ile Met Thr Val Leu Pro Stop
```

$\Rightarrow$ Example from Jones & Pevzner, p. 65.

The translation starts at the **start codon**, ATG (AUG), and stops at a **stop codon**. The ATG codon determines the **reading frame** (phase).
[Most proteins start with a methionine. However, for certain mRNAs GUG or UUG are used as a start codon, or further processing removes the N-terminal part of the peptide (protein).]
3 stop codons (non sense)
61 codons correspond to 20 aa (called sense codons) one of which is the start codon (codes for Met)
The code is said to be **degenerated** because there are more than one code for each amino acid. Therefore, there is a unique translation, the same amino acid sequence can be encoded by more than one DNA sequence!
The collection of all the proteins is called the **proteome**; and **proteomics** studies the interactions of all the proteins.

The proteome is the sum of all the proteins at a given time. Just like the transcritome, the proteome is dynamic.

Proteins are the main players in the cell, constituting the structure of the cell, but more importantly by catalyzing most reactions.

"(...) understanding how a genome specifies the biochemical capability of a living cell is one of the major research challenge of modern biology." [**?**]

From hypothesis-driven reductionist approach to holistic, data-driven, systems-based approach.

Systems biology: "(...) the study of an organism, viewed as an integrated and interacting network of genes, proteins and biochemical reactions which give rise to life."
www.systemsbiology.org.

Summary

- The code consists of triplets, called codons;
- The start codon is Met, which is the codon for amino acid Methionine;
- There are 3 stop codons; signifying the end of the chain, no amino acid is added;
- There are approximately 30 to 50 adapter molecules, called transfer RNAs or tRNAs for short. Each tRNA is charged (loaded) with a specific amino acid, which correspond to its anti-codon. The tRNA molecules are nucleic acids and the recognition of the codon/anti-codon follows the normal base-pairing rules;
- An Open Reading Frame (ORF) is a contiguous sequence of codons starting with Met (Start) and ending with a Stop codon;

Summary

- Since the code is made of triplets, there are three possible translation frames in one strand, following that the start codon occurs at position $i \bmod 3 = 0, 1$ or 2;
- Since DNA is made of two complementary strands running anti-parallel, this makes a total of six translation frames.

A mutation occurring in a coding region will affect the gene product, the encoded protein.

Properties of the genetic code

- The number of codons per each amino acid vary: Arg has 6, Met has 1, Phe has 2, . . . ;
  The codons that are specifying the same amino acid are called synonymous codons;
- Look at the third position, Ser, Pro, Arg, Thr, Val, Leu, Gly, a mutation does not change the identity of the encoded aa;
- In general, it takes more than one mutational event to change the encoded aa into an aa of a different functional group;
- For a given amino acid, not all the codons are used with the same frequency.
- The code itself is mostly universal.

Properties of the genetic code

You can observe that a deletion or insertion would most likely change the identity of the encoded aa from the point of deletion (insertion) until a stop codon is found in the new reading frame.
The region that starts at an initiation codon and runs until a stop codon is called an "open reading frame", ORF for short.
In a given sequence, when looking for ORFs, one has to consider 6 possible reading frames.

Universal Genetic Code

	U		C		A		G		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
U	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	C
U	UUA	Leu	UCA	Ser	UAA	*Stop*	UGA	*Stop*	A
U	UUG	Leu	UCG	Ser	UAG	*Stop*	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
C	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
C	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
C	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
A	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
A	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
A	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg	G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
G	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
G	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
G	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

Gene expression: implications

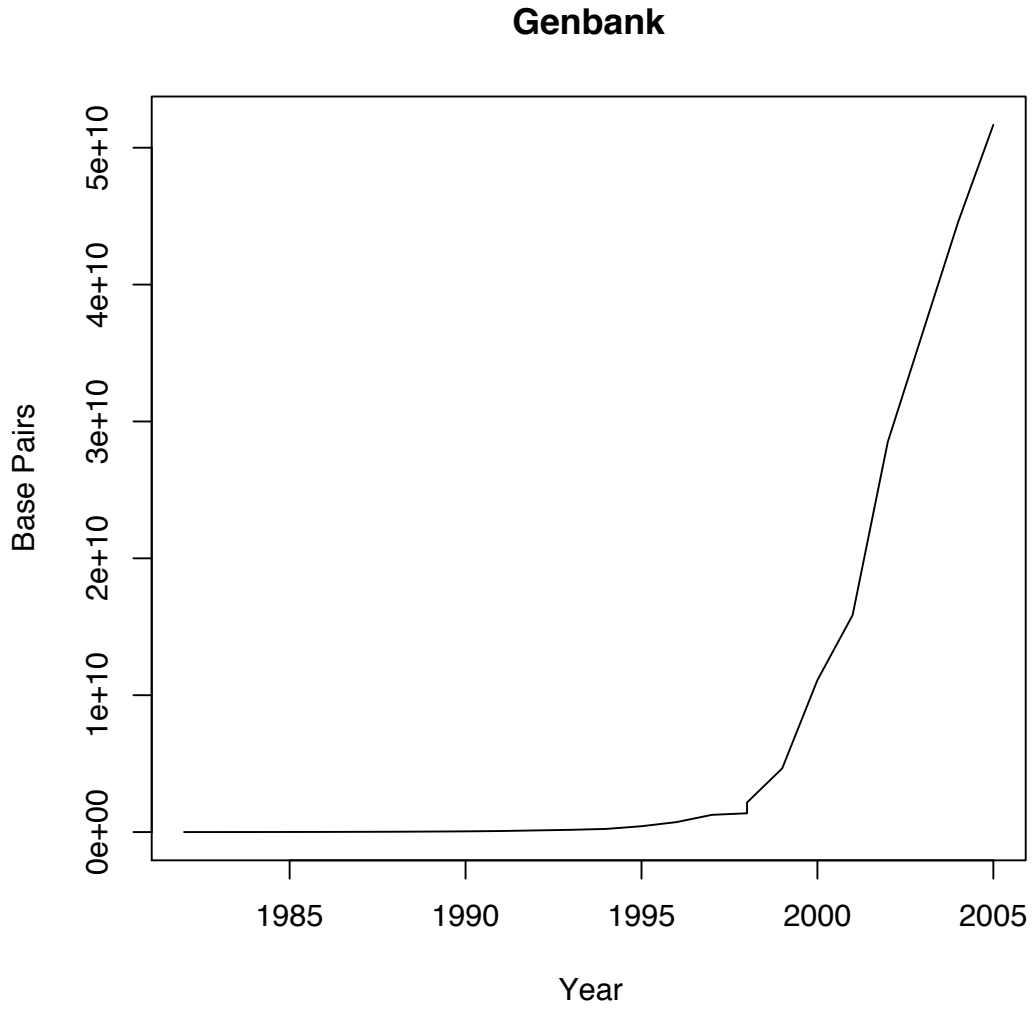
Each one of your cells has an exact copy of your DNA (well almost exact, we'll talk about this later) and the content of your DNA does not change with time (except for the rare mutations), unlike the concentration of proteins that constantly vary with time.

Statistics as of June 2002

The National Center for Biotechnology Information (US) is the host of GenBank, the NIH genetic sequence database, started in 1982, it now contains "20,649,000,000 bases in 17,471,000 sequence records as of June 2002":
www.ncbi.nlm.nih.gov/Genbank/
The European Molecular Biology Laboratory established the EMBL Nucleotide Sequence Database in 1980, it now contains 17,226,421 sequence entries comprising nucleotides 20,017,246,707 bases in 17,226,421 records (as of June 2002). "**In recent years the EMBL database has doubled in size nearly every year** . . . During the first 8 months of 1999 $>$ 1.6 million new entries (1.3 Gigabases) were made public, an average of 6400 entries (5.4 Mbases) per day.":
www.ebi.ac.uk/embl/
DDBJ (DNA Data Bank of Japan) began DNA data bank activities in earnest in 1986 at the National Institute of Genetics (NIG):
www.ddbj.nig.ac.jp/

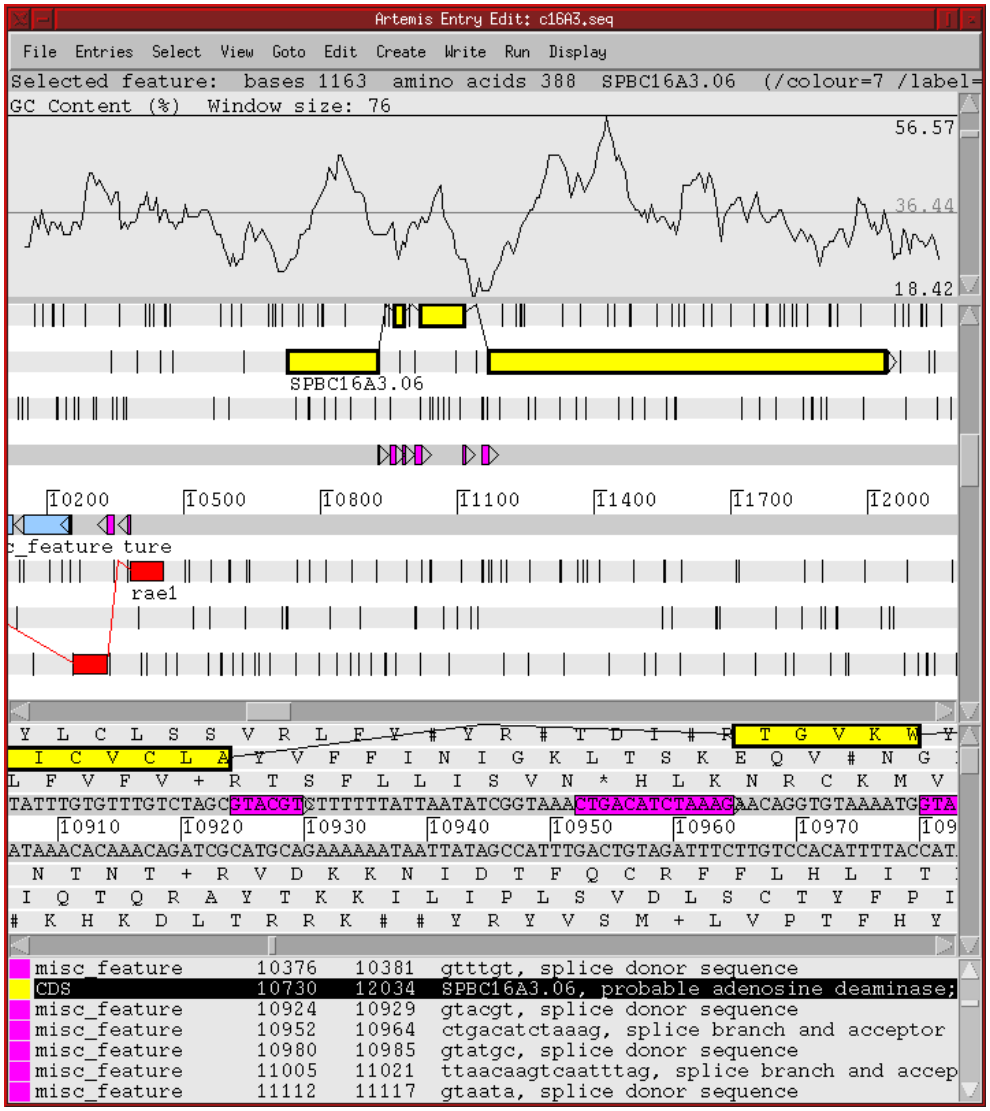
Statistics as of September 2005

- Genbank (August 15, 2005) — DNA sequences
  - 179 Gigabytes (sequences only)
  - 51,674,486,881 bases
  - 46,947,388 sequences
- UniProtKB/Swiss-Prot (September 13, 2005) — Protein sequences
  - 70,391,852 amino acids
  - 194317 sequences
- PDB (September 13, 2005) — Protein three-dimensional structures
  - 29,757 entries


Genbank
Base Pairs
0e+00
1e+10
2e+10
3e+10
4e+10
5e+10
1985
1990
1995
2000
2005
Year

Summary

- DNA replication allows to pass information unchanged to descendants;
- DNA and hereditary information are linked;
- Transciption occurs at promotor regions;
- Transciption: simple process, one-to-one relationship;
- Translation: involves triple codes, called codons;
- Genome: the sum of all the genetic information;
- Transcriptome: all the transcripts;
- Proteome: all the proteins.



www.sanger.ac.uk/Software/Artemis (File, dbfetch, U00096).

References

Resources: databases

Each January, Nucleic Acids Research Journal publishes a special issue about the main bioinformatics databases. Similarly, in July, it publishes an issue on Web servers.

- Database Issue 2009
- Web Server Issue 2009

Resources: biology

- Genomes 2 by T.A. Brown

  (http://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowTOC&rid=genomes.TOC&depth=2)

  - Several other books are available from NIH
    http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books
- The Biology Project of the University of Arizona
  www.biology.arizona.edu/default.html
  - Molecular Biology
    www.biology.arizona.edu/molecular_bio/molecular_bio.html
  - Cell Biology
    http://www.biology.arizona.edu/cell_bio/cell_bio.html
- DNA from the beginning
  www.dnaftb.org/
- The Beginner's guide to molecular biology
  www.rothamsted.ac.uk/notebook/courses/guide

Resources: biology (contd)

- Introduction: The Nature of Science and Biology
  www.emc.maricopa.edu/faculty/farabee/BIOBK/BioBookintro.html
- CS 262. Computational Genomics (Stanford Winter 2004)
  - Basics of Molecular Biology Part I
    www.stanford.edu/class/cs262/Notes/hong.pdf
  - Basics of Molecular Biology Part II
    www.stanford.edu/class/cs262/Notes/jdlh.pdf

References

Please don't print these lecture notes unless you really need to!