Molecular Biology II

Department of Molecular Pharmacy

Mgr. Marie Brázdová, Ph.D. (FAF MUNI, BFU AVČR)
Mgr. Marek Petr (FAF MUNI, BFU AVČR)
Mgr. Daniel Renčiuk, Ph.D., (BFU AVČR)
PharmDr. Jakub Treml, Ph.D. (FAF)

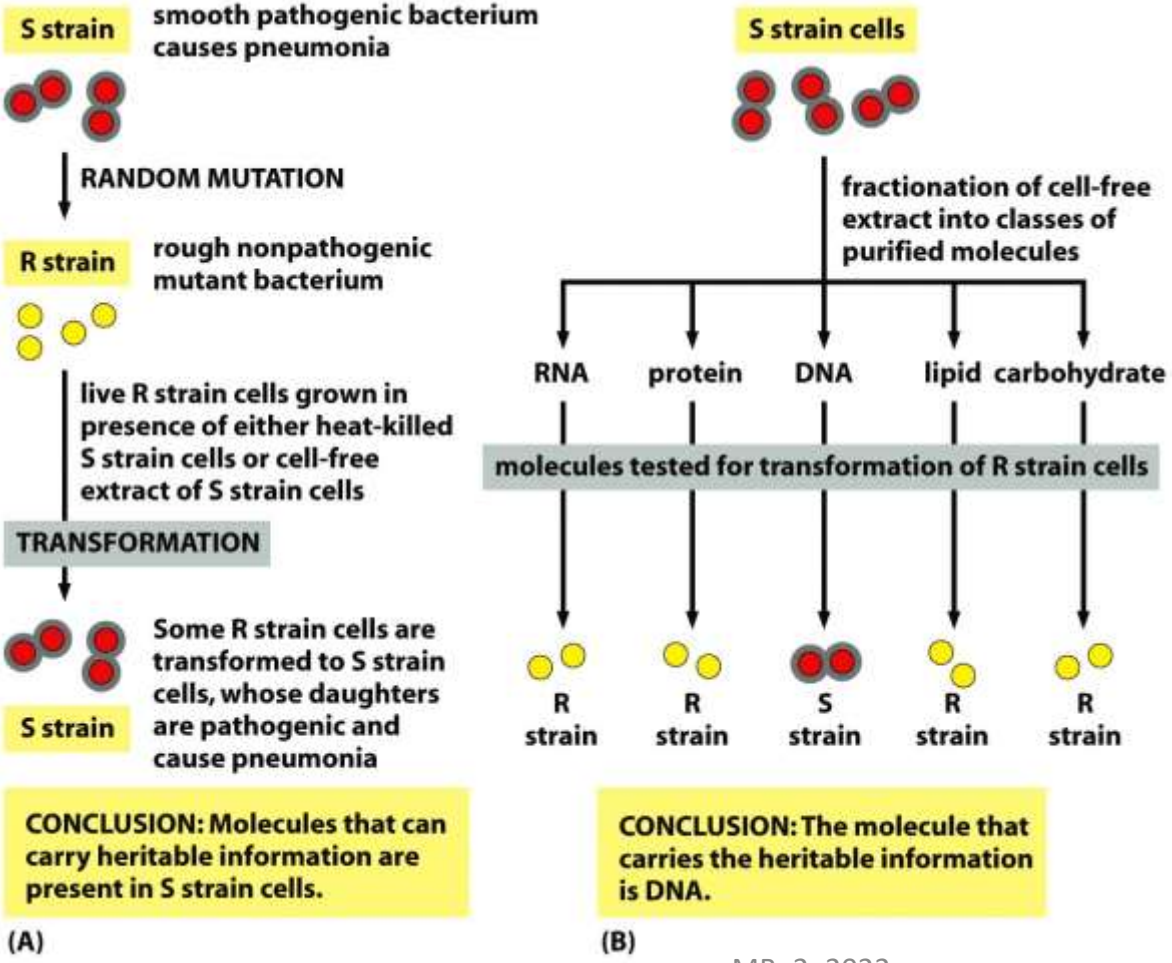
brazdovam@pharm.muni.cz,
petrmpharm.muni.cz,

The structure and function of biopolymers during the transitions of genetic information

Daniel Renčiuk

Genetic information is coded by DNA

The experiment combining two strains of Streptococcus pneumoniae bacteria.



O. Avery
C. MacLeod
M.McCarty

STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES

INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III

BY OSWALD T. AVERY, M.D., COLIN M. MACLEOD, M.D., AND MACLYN MCCARTY,* M.D.

(From the Hospital of The Rockefeller Institute for Medical Research)

PLATE 1

(Received for publication, November 1, 1943)

Biologists have long attempted by chemical means to induce in higher organisms predictable and specific changes which thereafter could be transmitted in series as hereditary characters. Among microörganisms the most striking example of inheritable and specific alterations in cell structure and function that can be experimentally induced and are reproducible under well defined and adequately controlled conditions is the transformation of specific types of Pneumococcus. This phenomenon was first described by Griffith (1) who succeeded in transforming an attenuated and non-encapsulated (R) variant derived from one specific type into fully encapsulated and virulent (S) cells of a heterologous specific type. A typical instance will suffice to illustrate the techniques originally used and serve to indicate the wide variety of transformations that are possible within the limits of this bacterial species.

Nuclein

- Nuclein – acidic substance rich in nitrogen and phosphorus
- J. F. Miescher - 1864
- Isolated from blood of wounded patients and cleaved by pepsin (proteolytic enzyme)

Roles of genetic material

Genotype – storage of genetic information and its transition to the offspring

Phenotype – expression of genetic information to particular properties of an individual

Evolutionary role – adaptation of an organism/species to the environment through the changes in genetic information

Terminology

Gene – several “definitions” depending on the point of view:

- classic genetics (Mendel) – elementary unit of hereditary genetic information
- molecular genetics – part of DNA coding for RNA (and as a consequence coding for some property of the individual)
  - structural genes coding for mRNA/protein (+ regulatory regions)
  - genes coding for functional RNA (miRNA, …)
- strict – structural gene – part of DNA that codes for protein sequence

Allele – particular variant of the respective gene

Genome – complete DNA of organism (molecular) x complete genetic information of organism x sum of genes (classic)

Genotype – the combination of particular alleles of all genes in individual

Phenotype – the sum of actual individual properties (as a result of expression of particular genotype in the respective environment)

Genophore – the carrier of genetic information, usually a molecule of DNA (often used for bacteria)

Information biopolymers

Deoxyribonucleic acid (DNA)

• linear heteropolymer composed from 2-deoxyribonucleotides connected by phosphodiester bonds
• usually as a stable and resistant double helix
• serves as a storage of genetic information, as a template for its reproduction (replication) and as a template for the expression of genetic information to the phenotype (transcription)

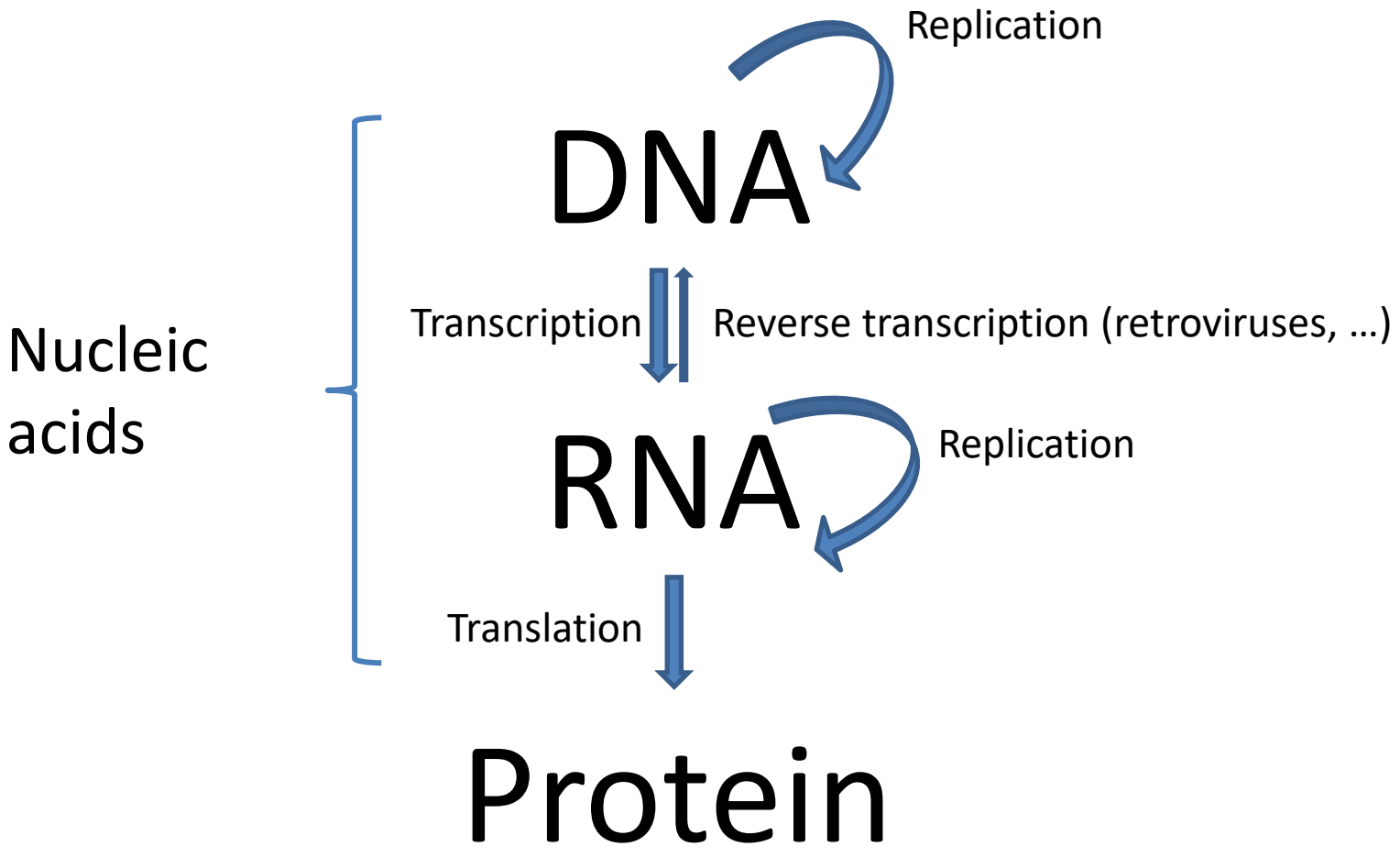
Ribonucleic acid (RNA)

• linear heteropolymer composed from ribonucleotides connected by phosphodiester bonds
• usually as a single-stranded structure of variable length, structure and reactivity
• many functions depending on type of RNA

Protein

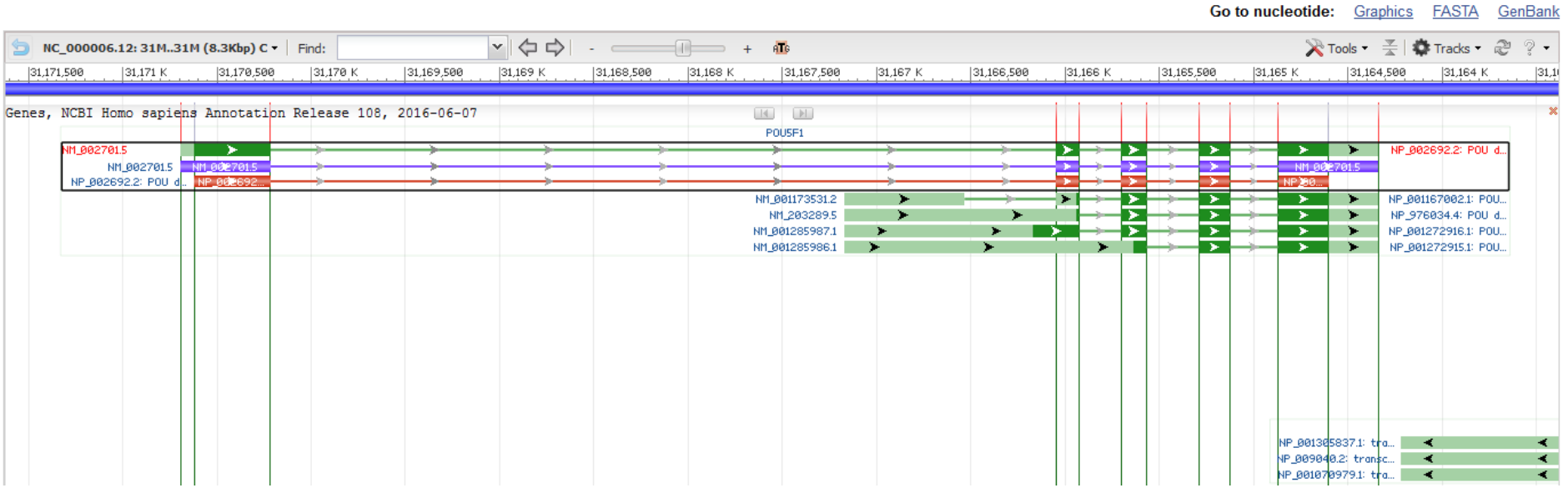
• linear heteropolymer composed from 20 (21) amino acids connected by peptide bonds
• highly variable structures, properties and functions

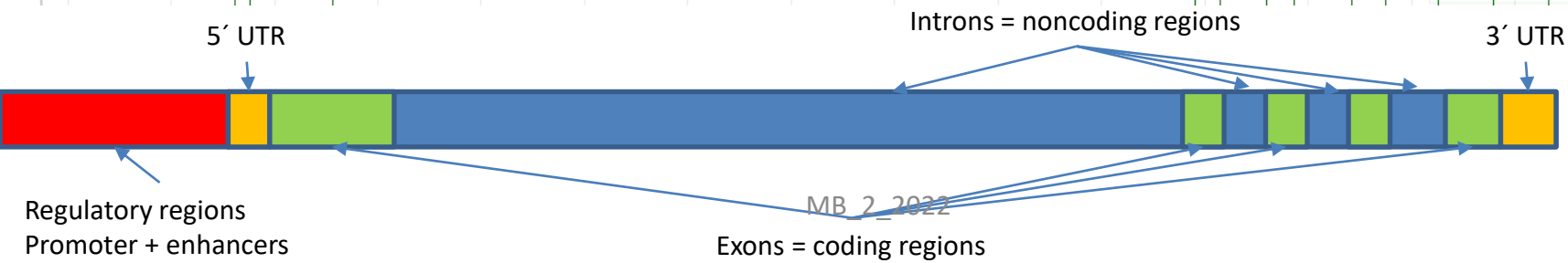
Central dogma of molecular biology



Gene

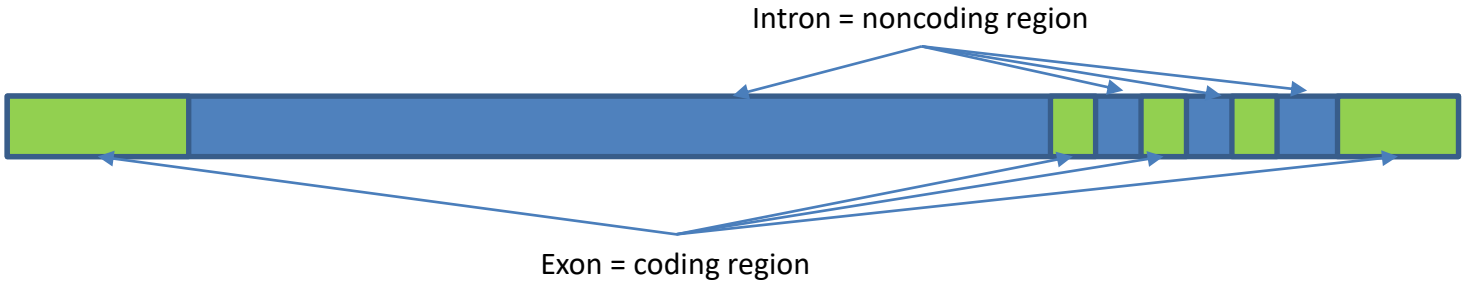
https://www.ncbi.nlm.nih.gov/gene/5460





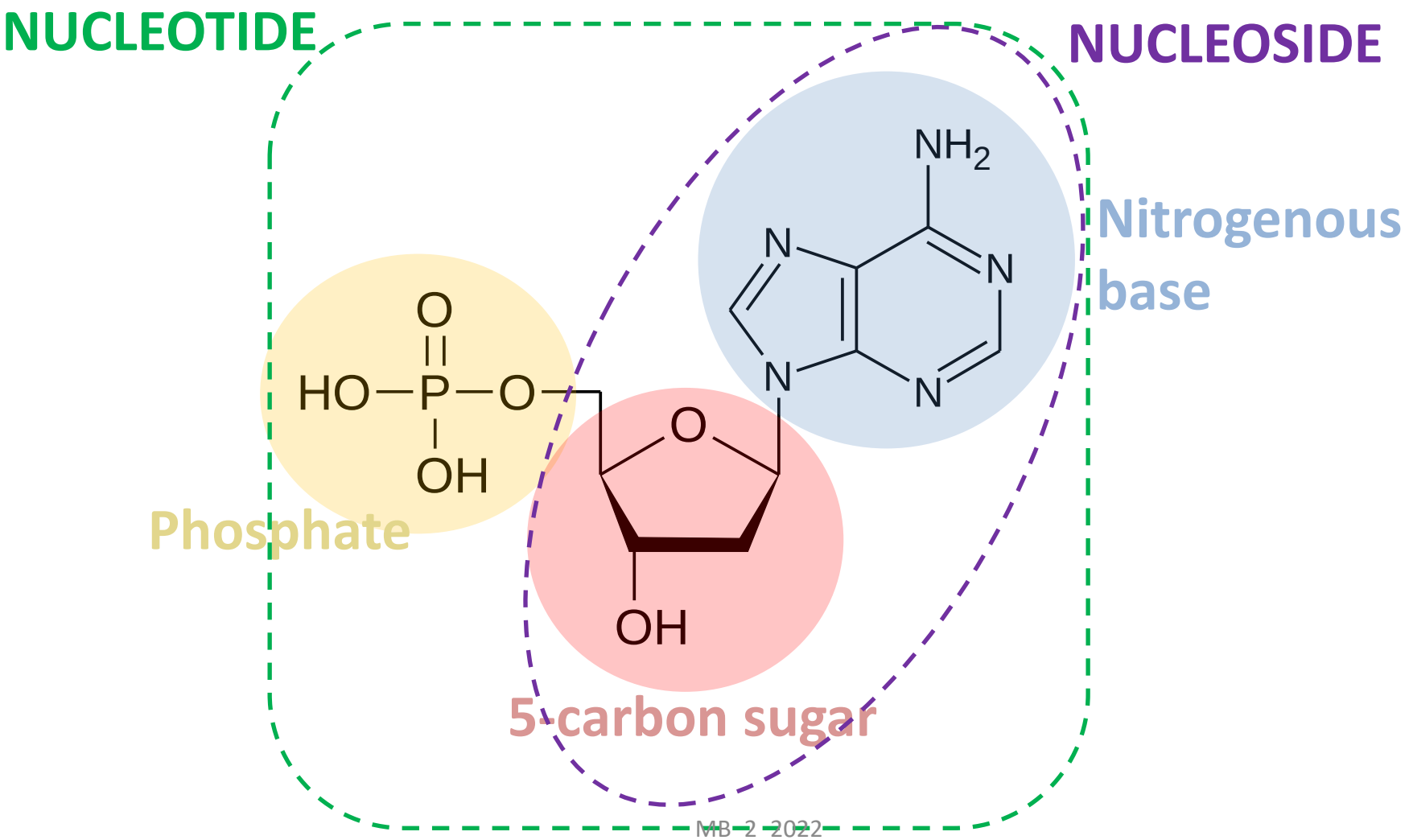
Genes

Most **eukaryotic** genes contain **introns**, that are transcribed into primary RNA transcript and introns are consecutively removed by **splicing** process on **spliceosome** to form final **mRNA**.



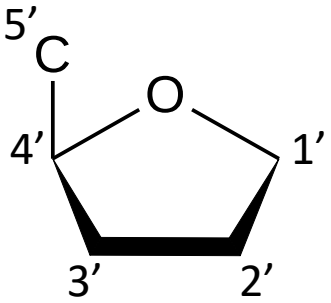
Prokaryotic genes do not contain **introns** and they are directly transcribed into mRNA that serves as a template for translation into protein.

Nucleic acids

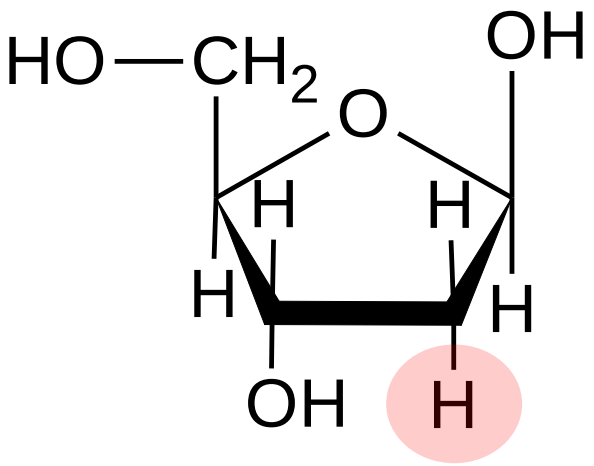


5-carbon sugar - pentose

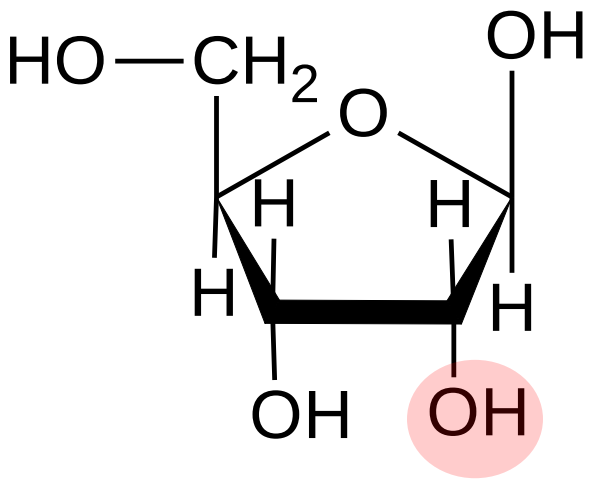
DNA



RNA

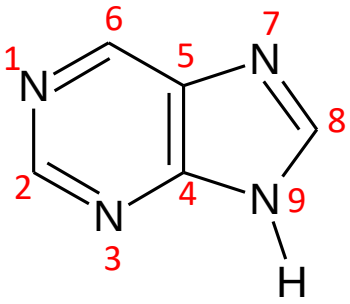


β-D-2-deoxyribose

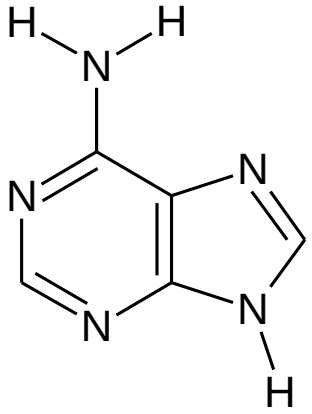


β-D-ribose

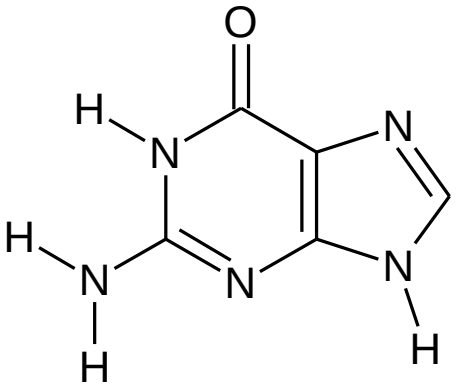
Base



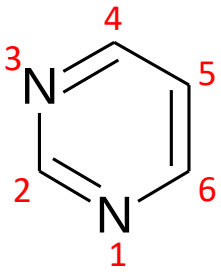
Purine



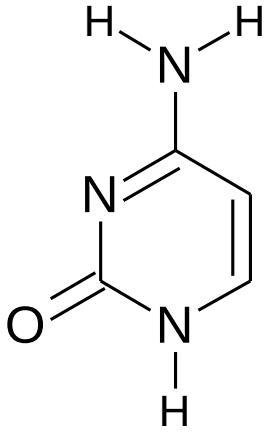
Adenine



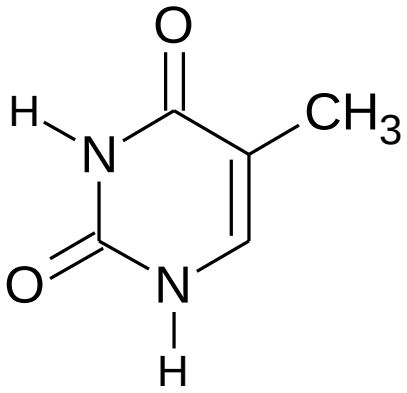
Guanine



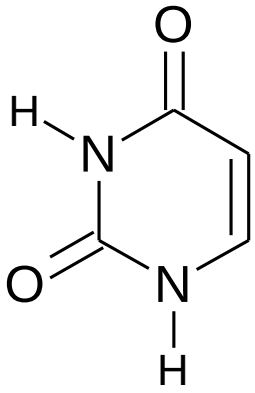
Pyrimidine



Cytosine



Thymine



Uracil

Nucleic acid nomenclature

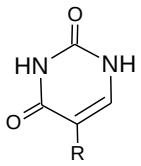
	Base	Nucleoside	Nucleotide NTP
G	Guanine	Guanosine	Guanosine triphosphate
A	Adenine	Adenosine	Adenosine triphosphate
T	Thymine	Thymidine	Tymidine triphosphate
C	Cytosine	Cytidine	Cytidine triphosphate
U	Uracil	Uridine	Uridine triphosphate

DNA – prefix deoxy-

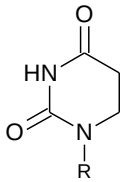
NMP = monophosphate
NDP = diphosphate

Modified bases

tRNA

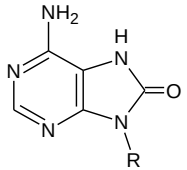


pseudouracil

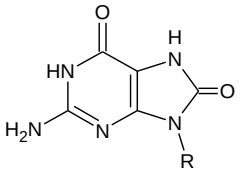


dihydrouracil

Oxidative damage

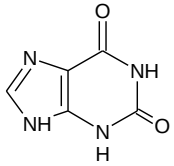


8-oxo adenine

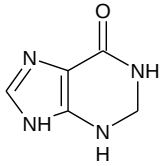


8-oxo guanine

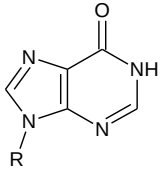
Metabolism



xanthine

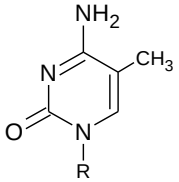


hypoxanthine

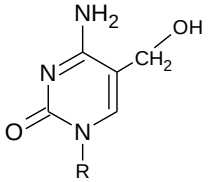


inosine

Epigenetics

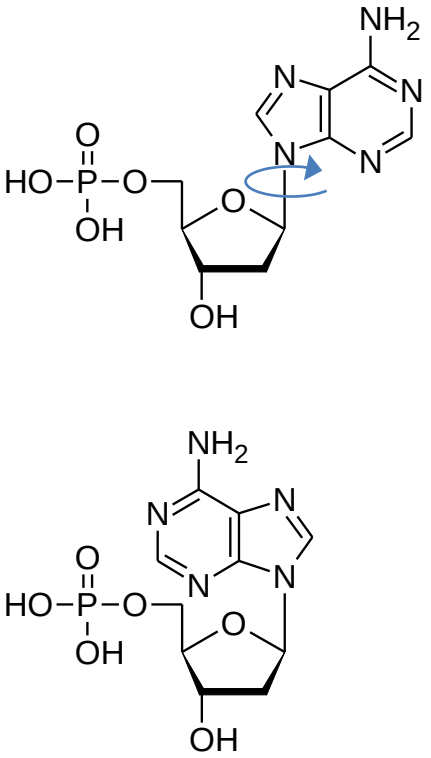


5-methyl cytosine



5-hydroxymethyl cytosine

Conformation of N-glycosidic bond



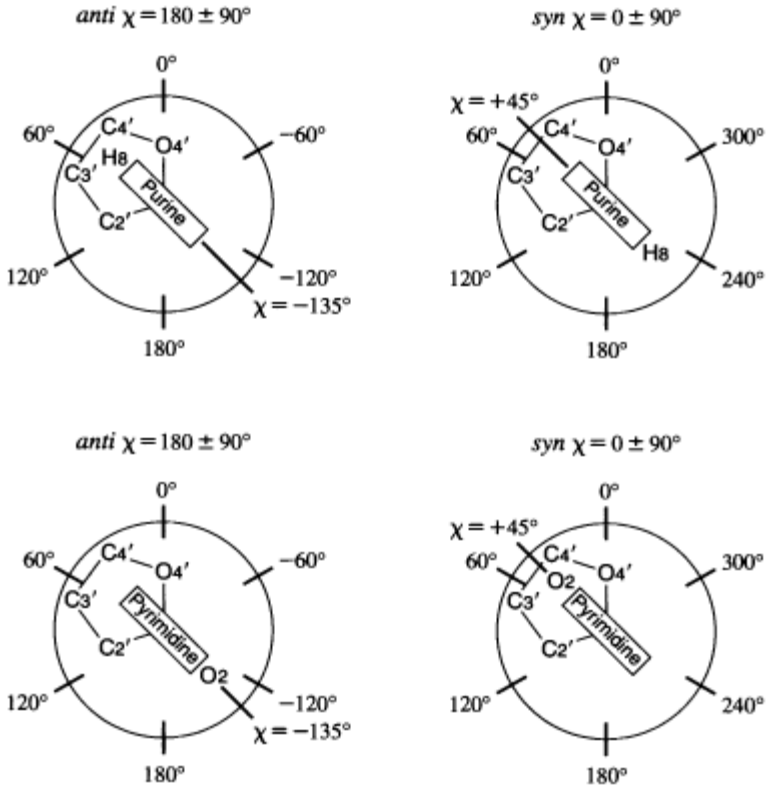
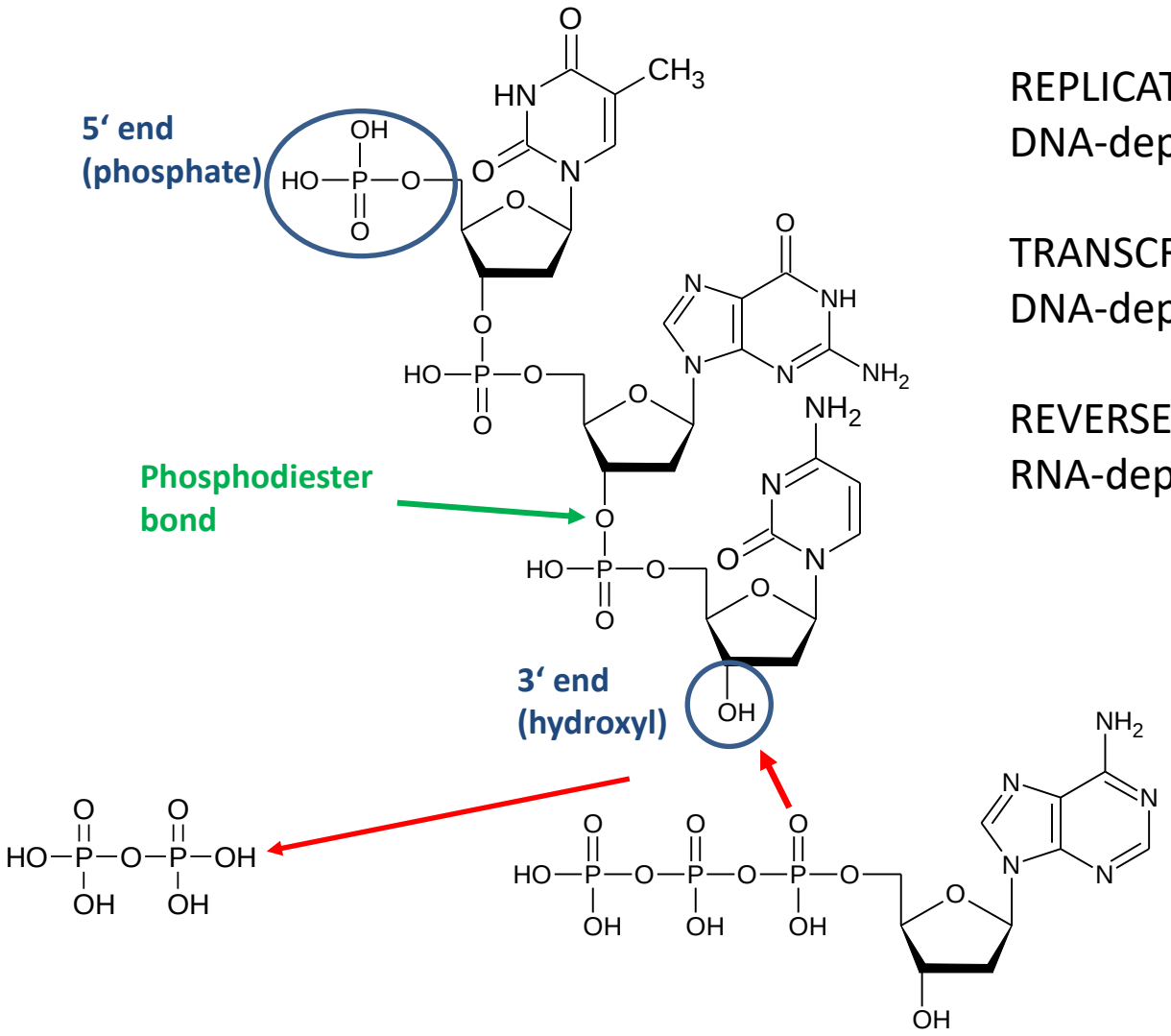


Figure 2-8
The glycosidic torsion angle χ is defined by O4′– C1′–N9–C4 for purines and O4′–C1′–N1–C2 for pyrimidines. When $\chi = 0°$ the O4′–C1′ is eclipsed by the N9–C4 bond in purines and by the N1–C2 bond in pyrimidines. The syn conformations correspond to $0° \pm 90°$; anti conformations correspond to $180° \pm 90°$. In nucleotides steric hindrance limits the conformations actually found to a much narrower range of angles that depend on sugar pucker and base. The syn conformations are usually found with $\chi = 45° \pm 45°$; anti conformations are usually found with $\chi = -135° \pm 45°$.

Formation of sugar-phosphate backbone



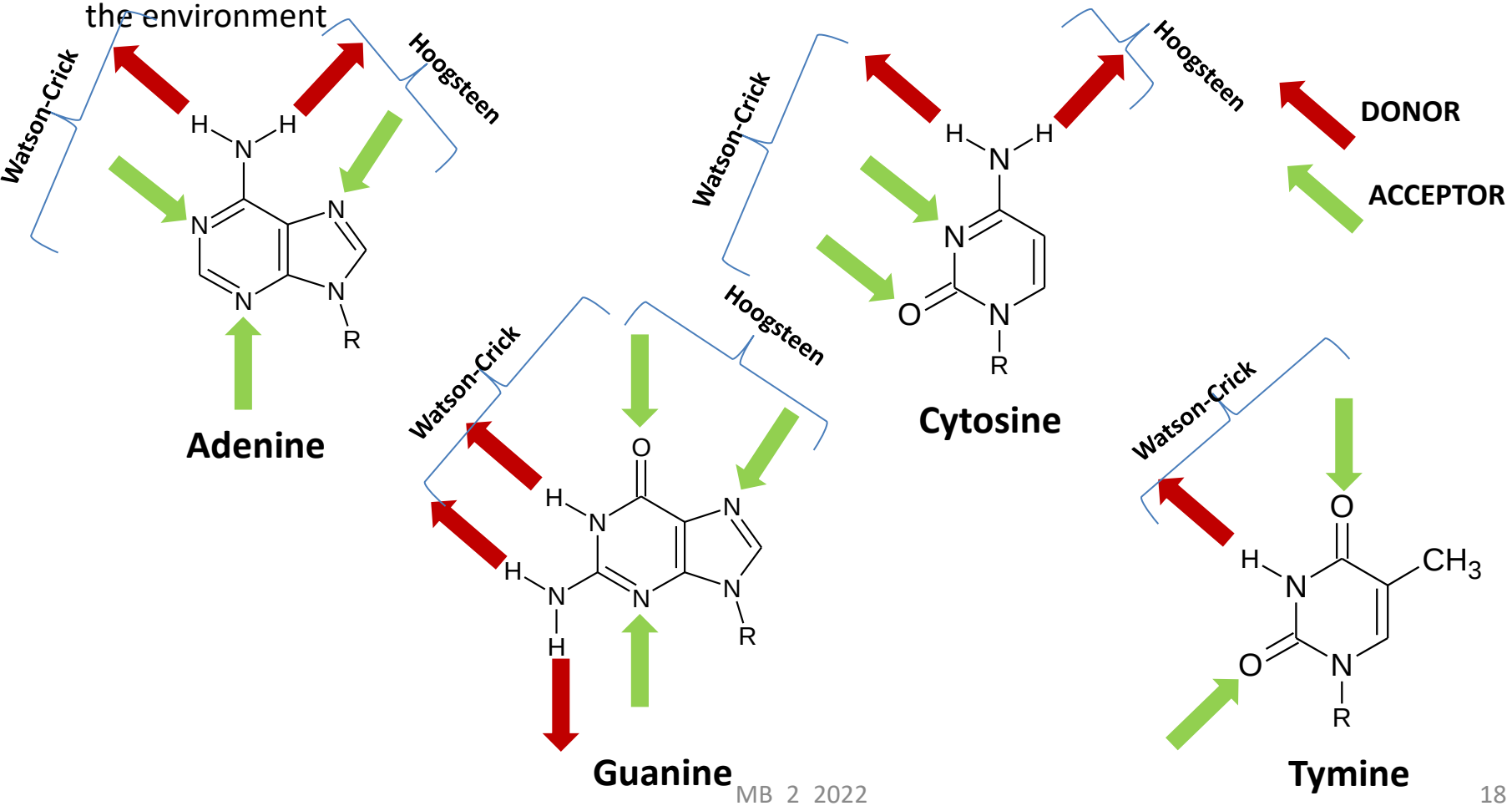
REPLICATION
DNA-dependent DNA polymerase

TRANSCRIPTION
DNA-dependent RNA polymerase

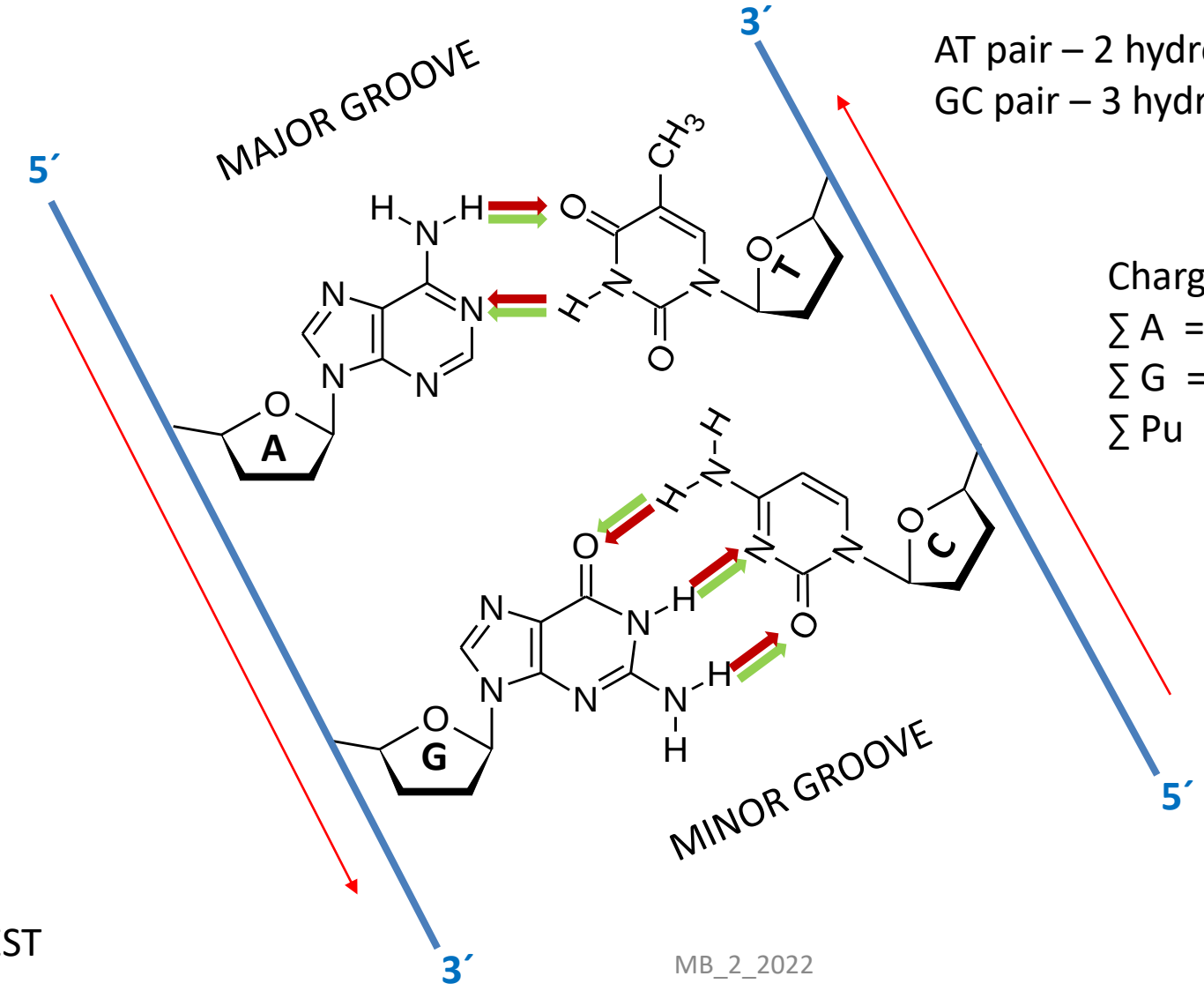
REVERSE TRANSCRIPTION
RNA-dependent DNA polymerase

Base reactivity – hydrogen bonds

Hydrogen bond – weak electrostatic interaction of two polar groups – one covalently binds hydrogen (DONOR – usually -N-H a -O-H); the second (ACCEPTOR) is usually N or O
Length: ~ 2.8 A (2 - 3,4 A) Energy: < 1 kcal/mol both depend on particular atoms and on the environment

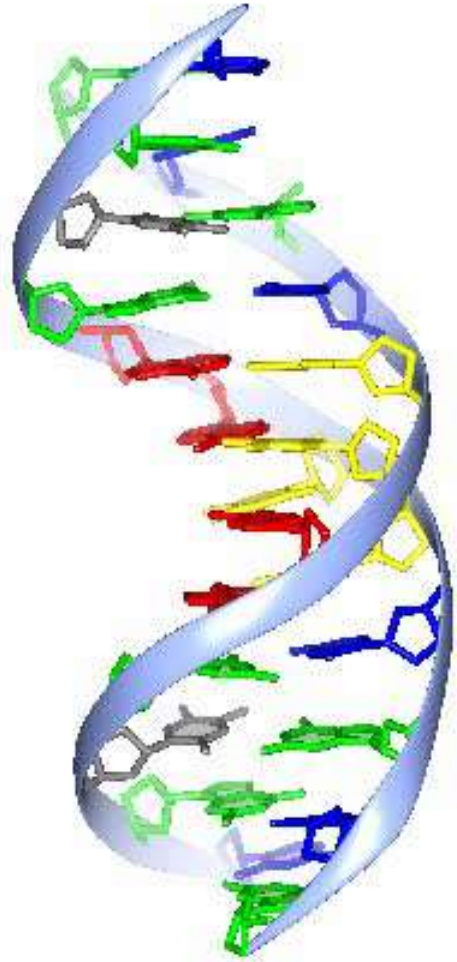


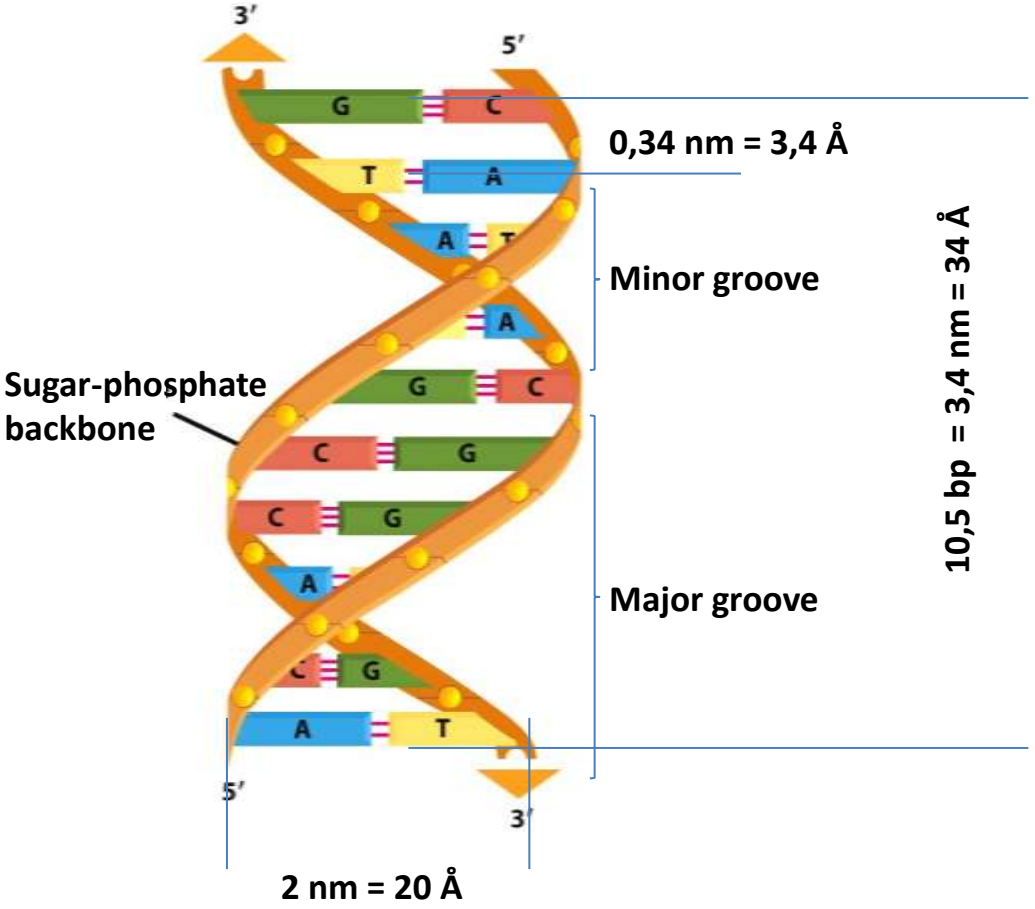
Watson-Crick base pairing



TEST

DNA double helix





DNA double helix

- **two** molecules (**strands**) of DNA
- the helix is **right-handed**
- the strands are **antiparallel** – their 5´-3´ direction is opposite in the context of double helix direction
- similar content of purines and pyrimidines; content of A = T, G = C (Chargaff rules)
- result – the strands are **complementary** – i.e. according to the Watson-Crick base pairing rules we can predict the sequence of one strand according to the sequence of the other
- on average the double helix contains **10,5 base pairs** per turn of the helix, which is about **3,4 nm** in length

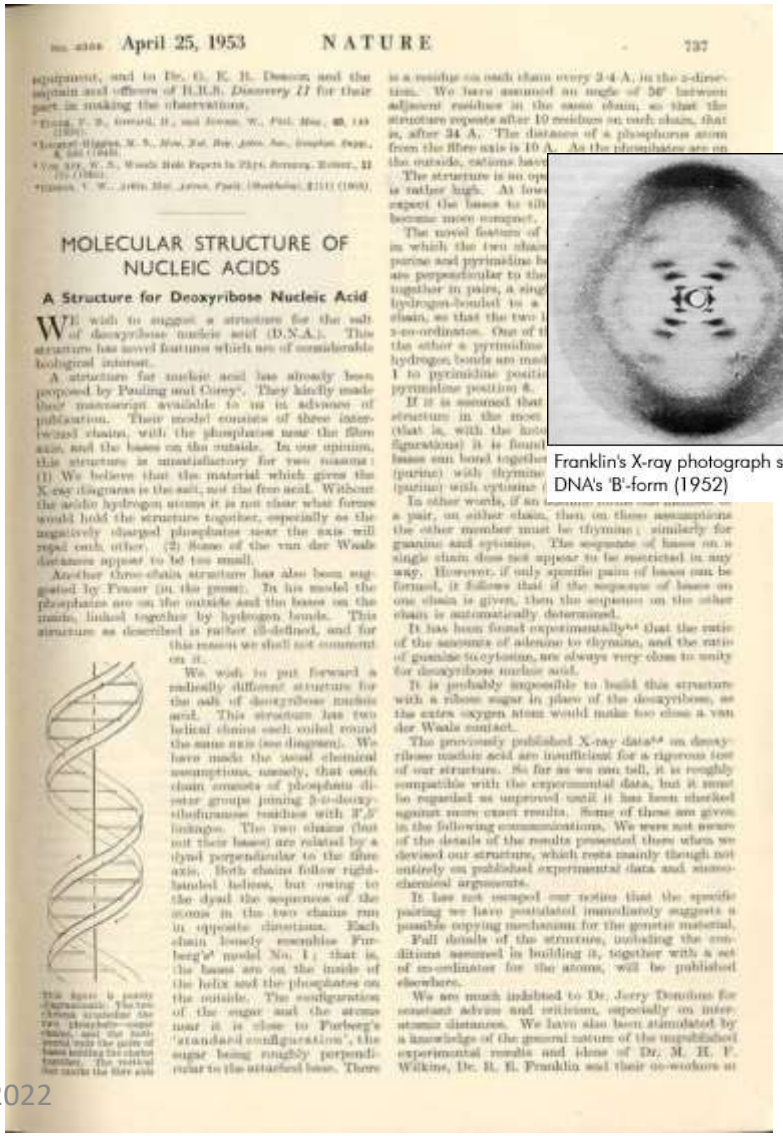
DNA structure –Watson and Crick model

F. Crick J.D. Watson

M. Wilkins

R. Franklin

E. Chargaff



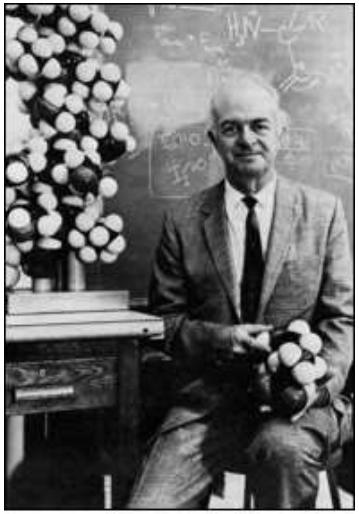
April 25, 1953 NATURE 737

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

Franklin's X-ray photograph shows DNA's 'B'-form (1952)

DNA structure – Pauling model



Linus Pauling

A PROPOSED STRUCTURE FOR THE NUCLEIC ACIDS

By Linus Pauling and Robert B. Corey

Gates and Crellin Laboratories of Chemistry,* California Institute of Technology

Communicated December 31, 1952

The nucleic acids, as constituents of living organisms, are comparable in importance to the proteins. There is evidence that they are involved in the processes of cell division and growth, that they participate in the transmission of hereditary characters, and that they are important constituents of viruses. An understanding of the molecular structure of the nucleic acids should be of value in the effort to understand the fundamental phenomena of life

which are involved in ester linkages. This distortion of the phosphate group from the regular tetrahedral configuration is not supported by direct experimental evidence; unfortunately no precise structure determinations have been made of any phosphate di-esters. The distortion, which corresponds to a larger amount of double bond character for the inner oxygen atoms than for the oxygen atoms involved in the ester linkages, is a reason-

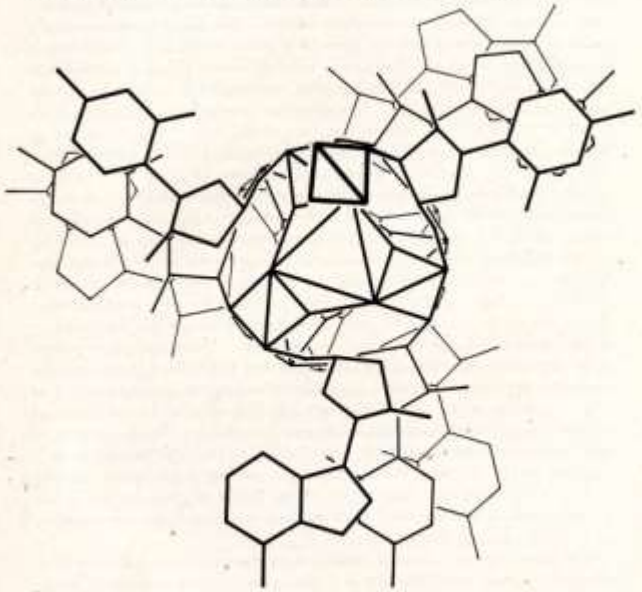


FIGURE 6

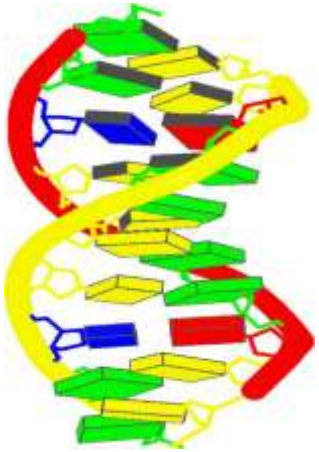
Plan of the nucleic acid structure, showing several nucleotide residues.

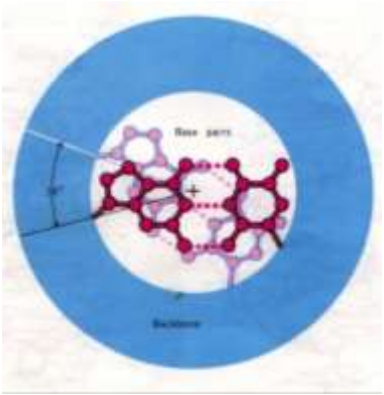
able one, and the assumed distances are those indicated by the observed values for somewhat similar substances, especially the ring compound S_3O_9, in which each sulfur atom is surrounded by a tetrahedron of four oxygen atoms, two of which are shared with adjacent tetrahedra, and two unshared. The O—O distances within the phosphate tetrahedron are 2.32 Å (between the two inner oxygen atoms), 2.46 Å, 2.55 Å, and 2.60 Å. The

Various types of double helix

B-DNA

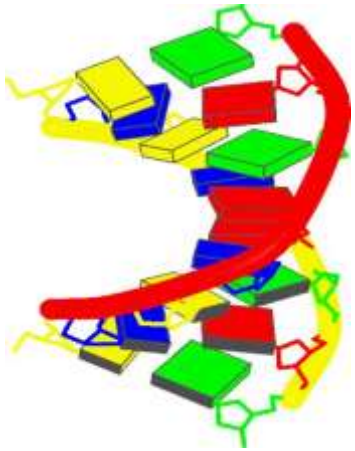
- DNA in water/salt soulutions

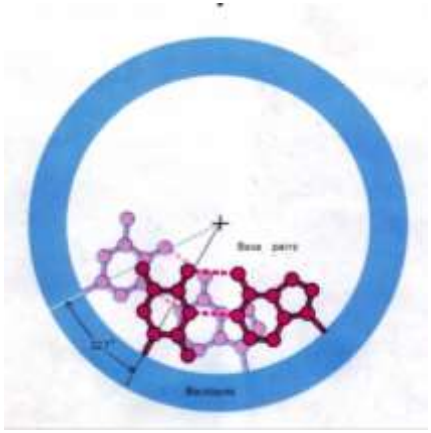




A-DNA

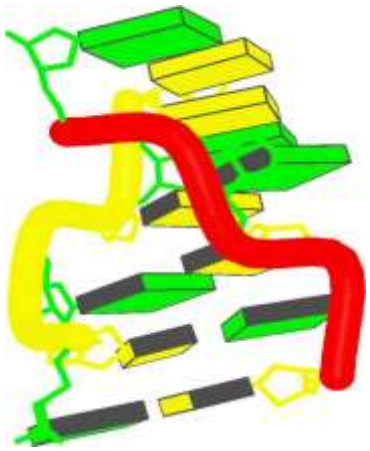
- DNA in crowding solutions
- RNA





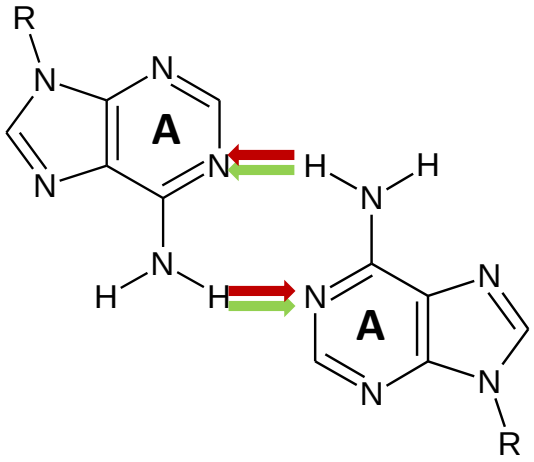
Z-DNA

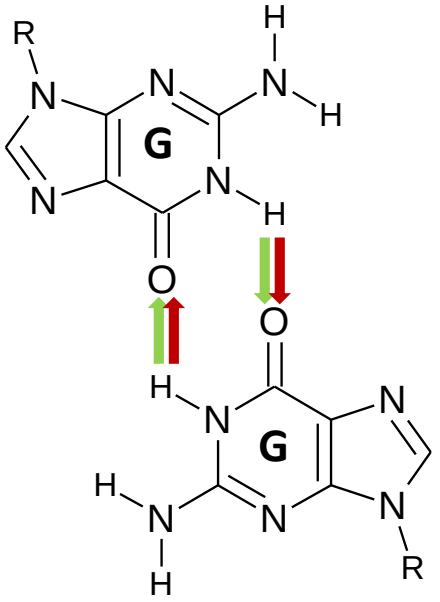
- CpG sequences in crowding conditions

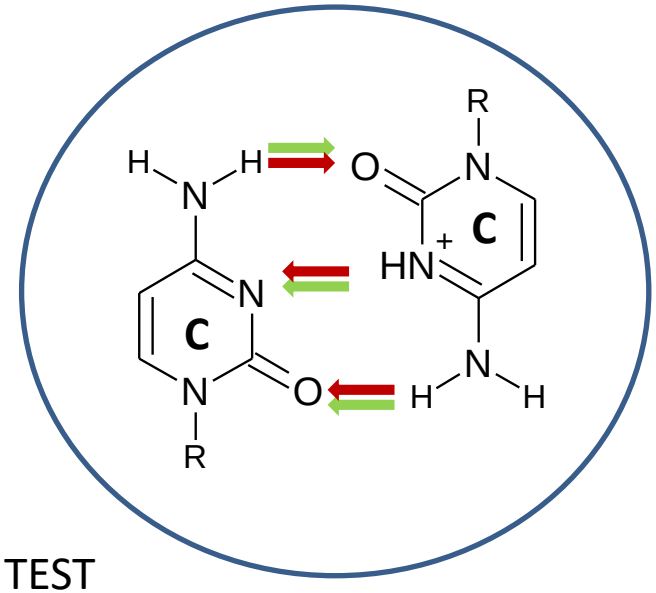


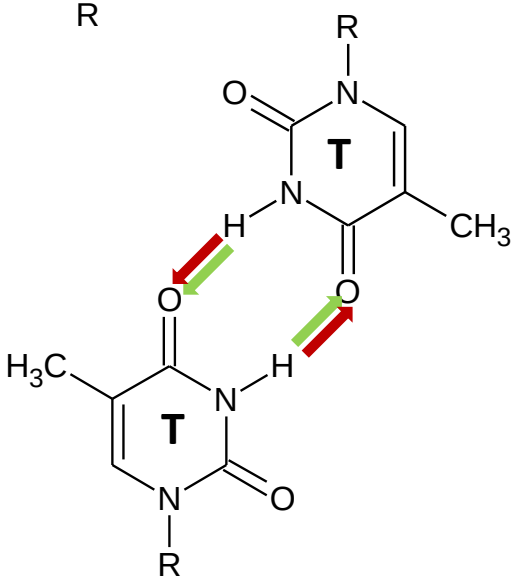
Left handed
Zig-zag step

Reversed Watson-Crick pairing

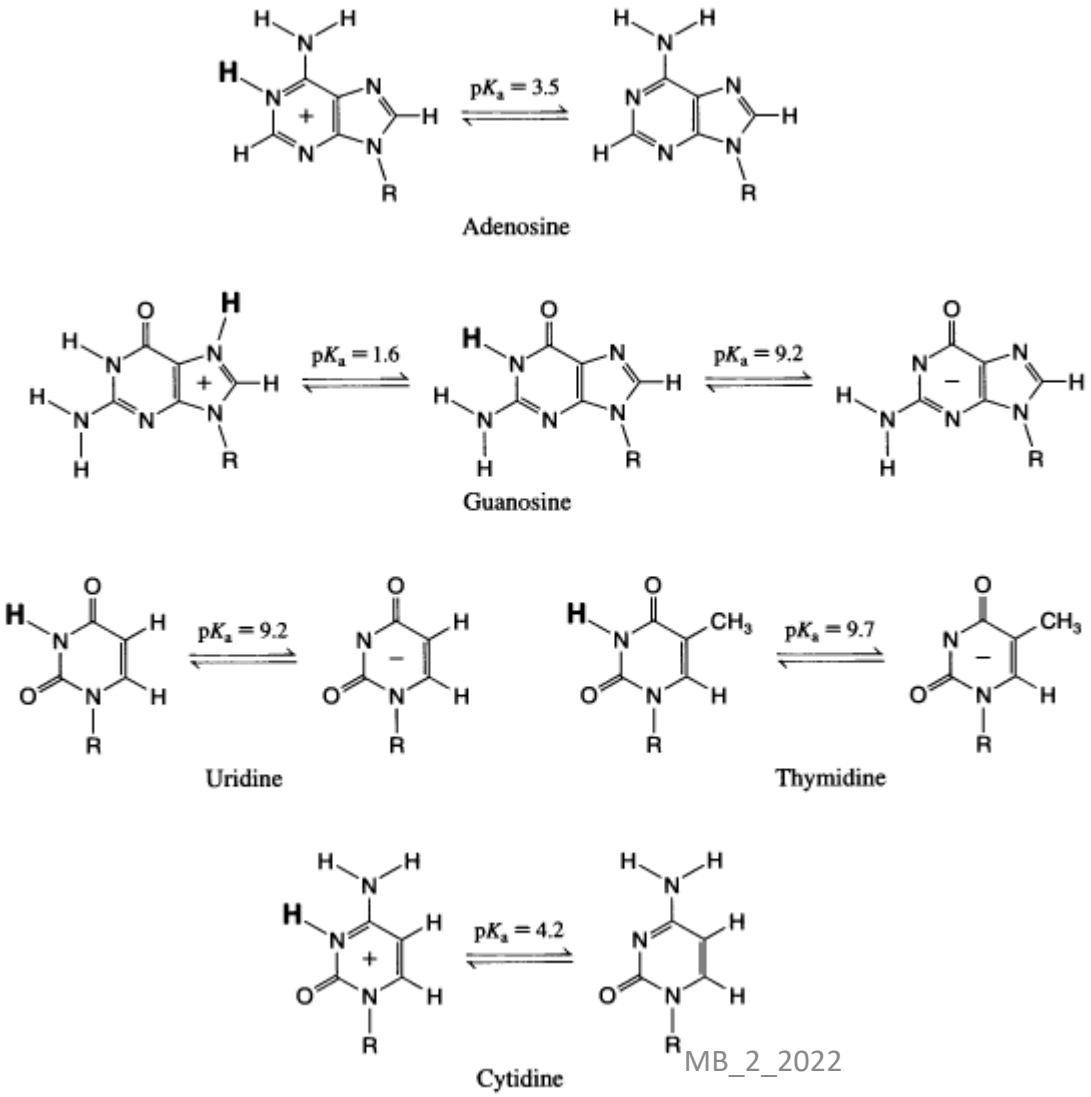








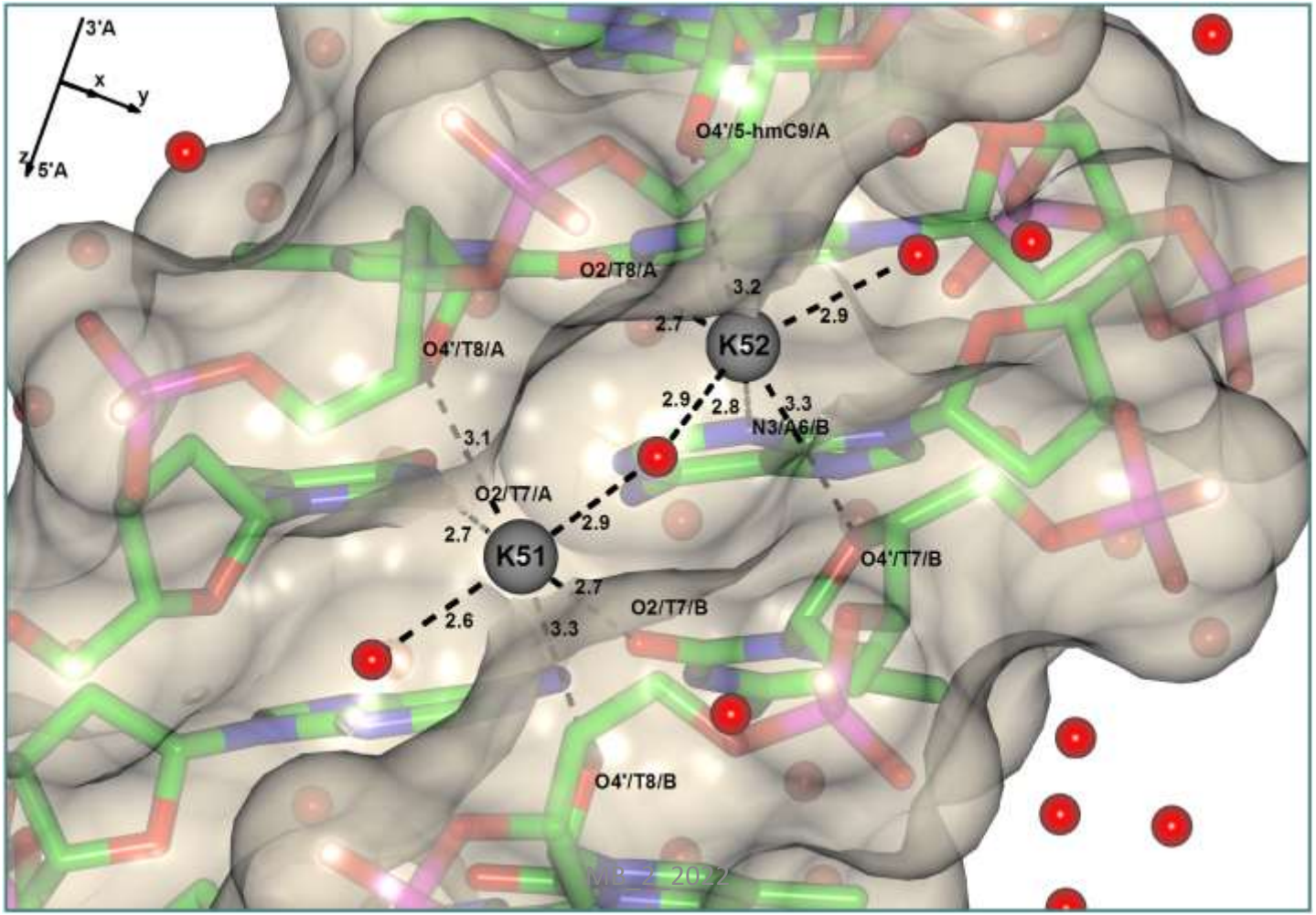
Base protonation



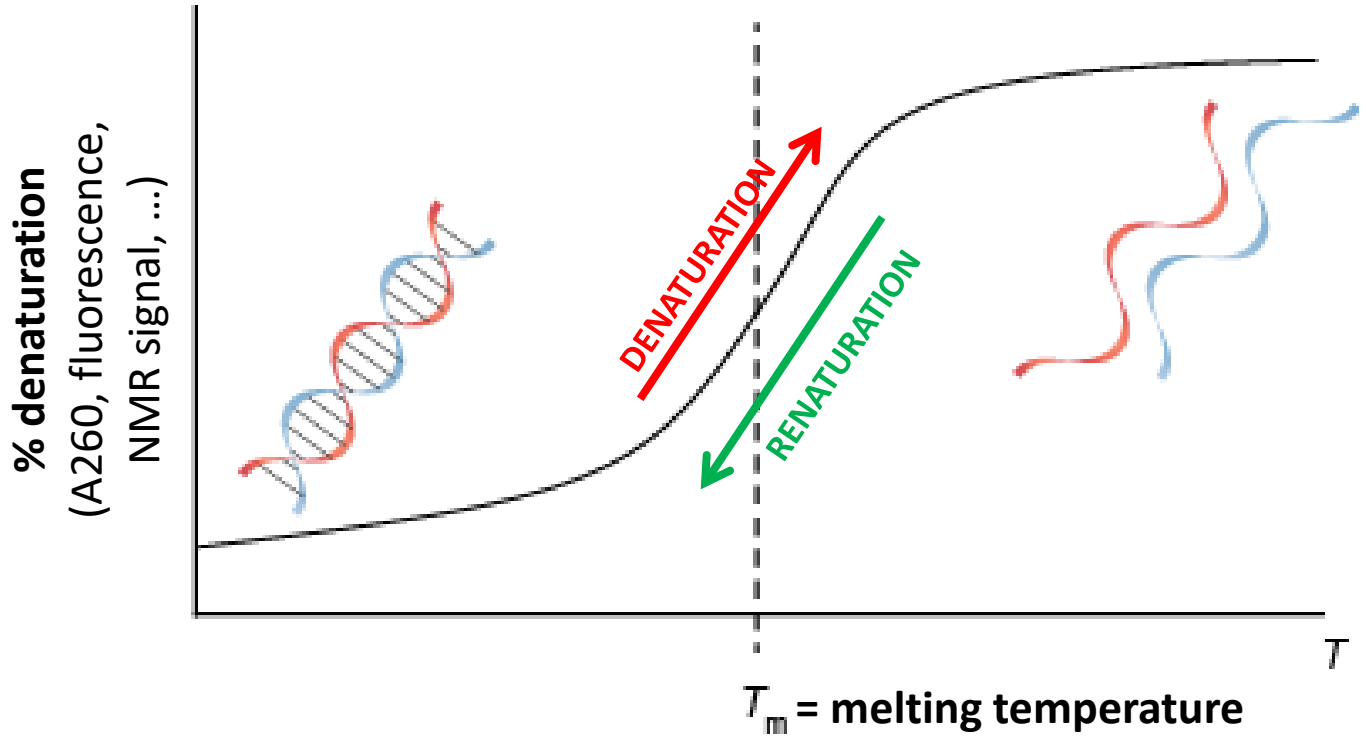
- base protonation might alter the base reactivity
- free bases have pK far from physiological
- pK of bases in DNA might be closer to pH 7.4
- cytosine in C_n sequences has pK~7 – cytosine i-motif

DNA double helix x ions / water

- phosphates in DNA backbone are negatively charged – repulsion
- this is compensated by interaction with ions (Na^+, K^+, Mg^{2+}, …) or water (H-H bonds)

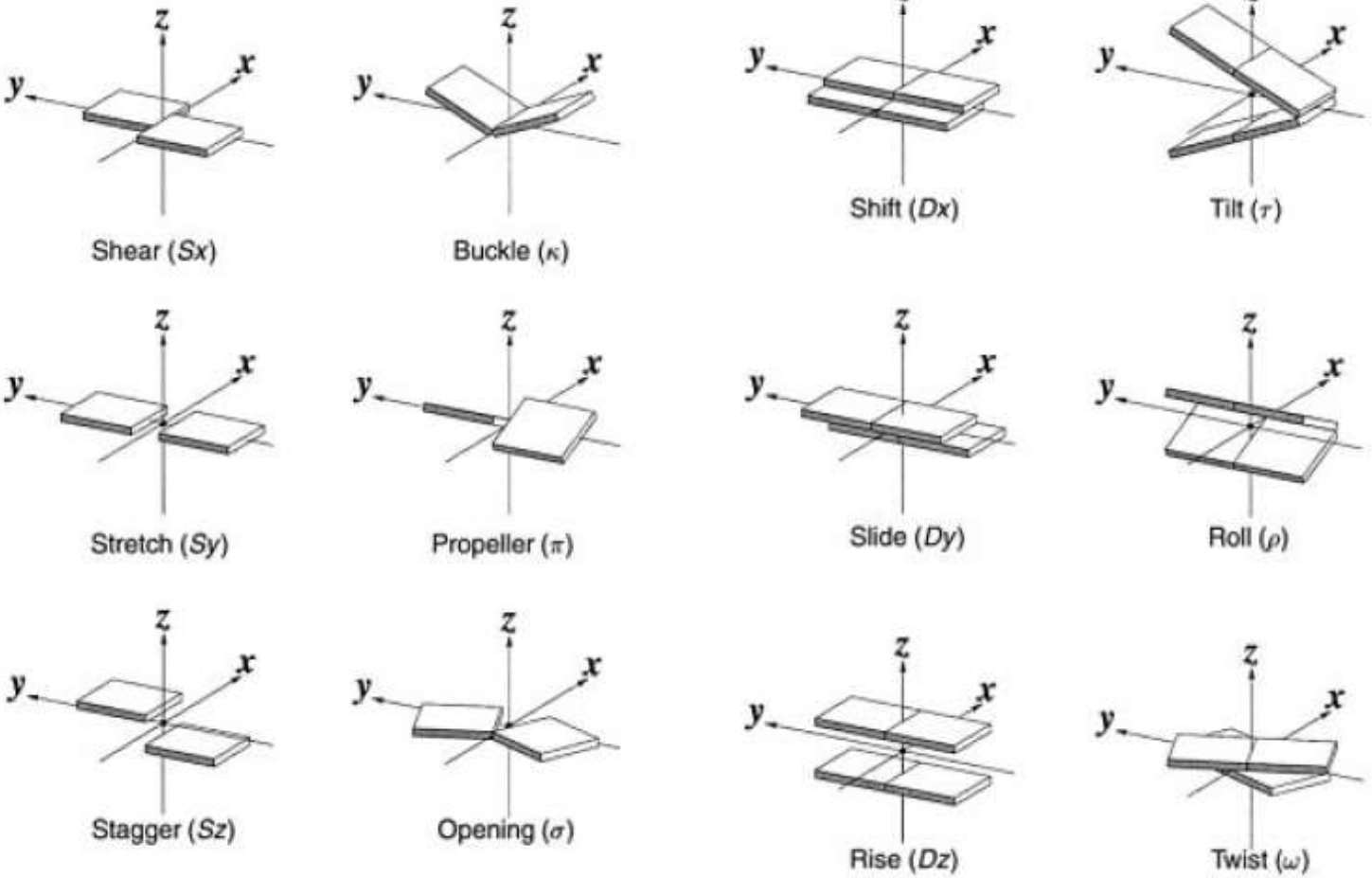


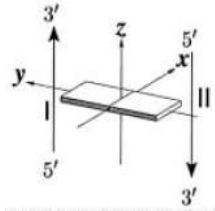
Stability of DNA double helix



- hydrogen bonds — AT = 2 x GC = 3 — Tm increases with GC and length
- base stacking — various — Tm increases with length and ions
- repulsion of backbone phosphates — Mg^{2+}>Na^{+} — Tm increases with ions

Base-pair parameters in double helix



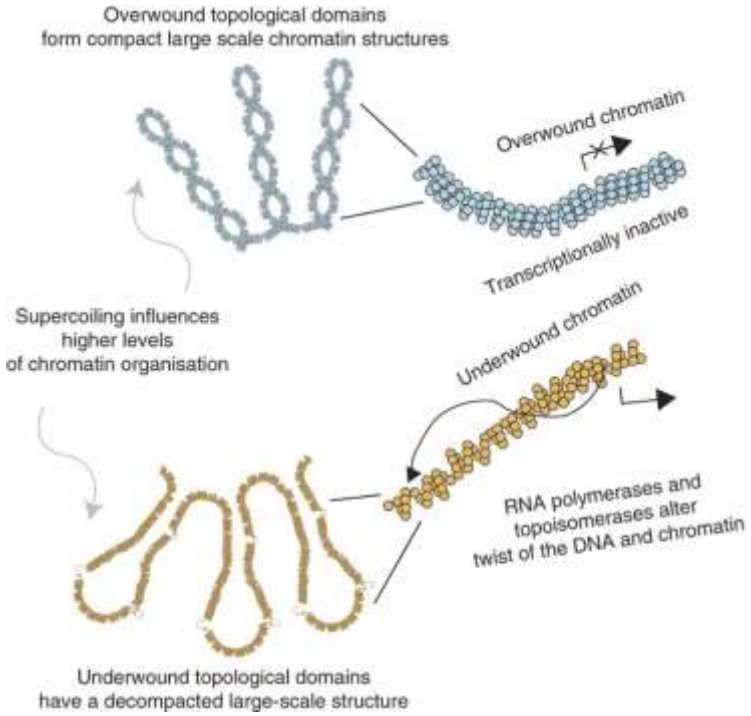


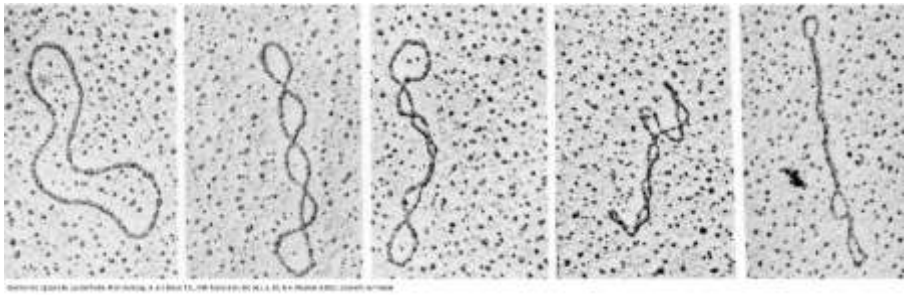
Lu et al., 2003, NAR

Types of nucleic acids

- **linear** (human chromosome) **x circular** (bacterial genome)
- **single-stranded** (most RNAs) **x double-stranded** (human DNA)

Superhelicity





Superhelicity happen mostly as a result of transition of polymerase complex and unwinding of DNA (helicase, …) during replication and transcription.

Topoisomerases

- Enzymes that relax the superhelicity
- Topo I – works on 1 DNA strand
- Topo II – works on 2-strand DNA

Reactivity of bases with amino acids

Double-stranded NA:
Interaction of Hoogsteen side with amino acid in major groove.

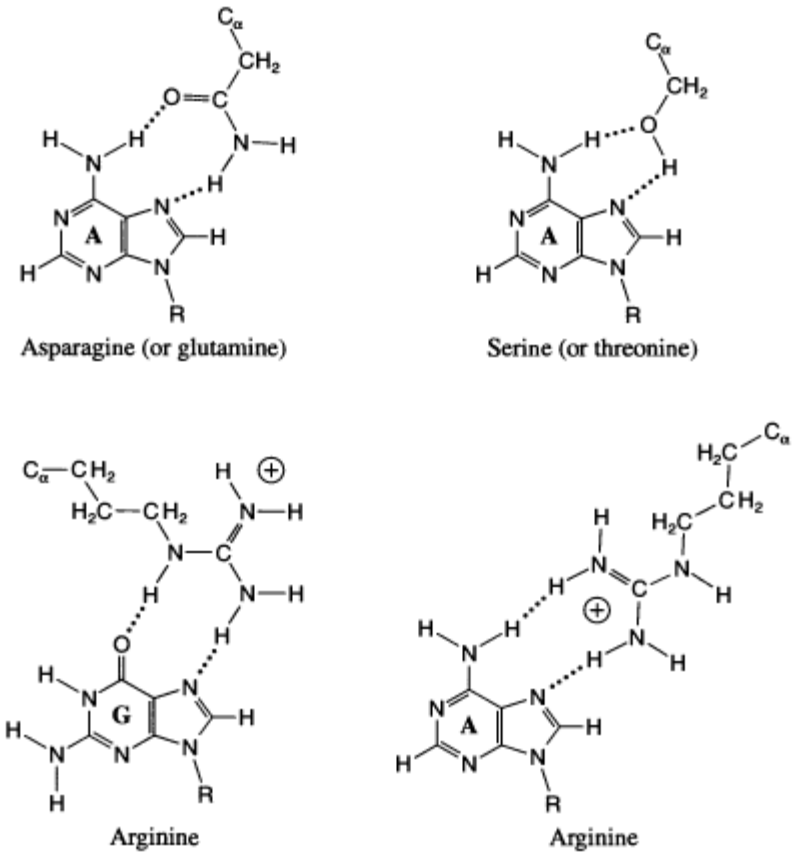


Figure 2-16
Interactions involving two hydrogen bonds between amino acids and bases that can occur through the major groove of a double helix.

Reactivity of bases with amino acids

Single-stranded NA:
Interaction of Watson-Crick side with amino acid.

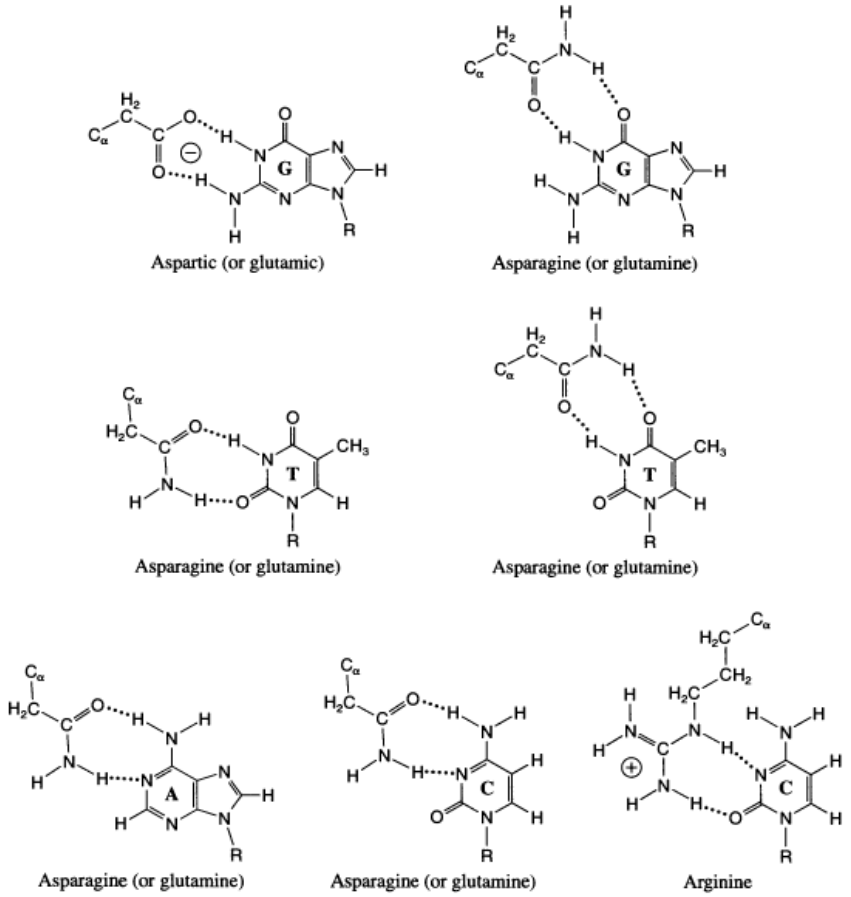
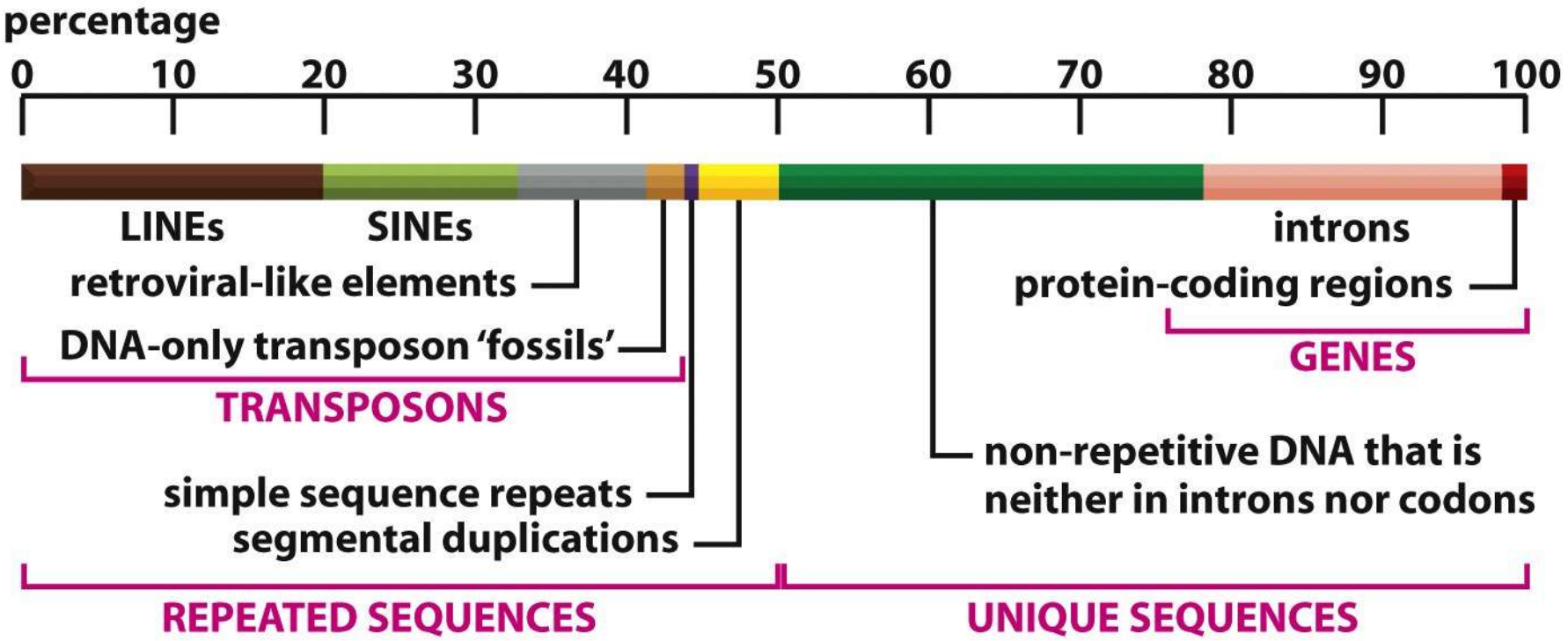


Figure 2-17
Interactions involving two hydrogen bonds between amino acids and bases that take the place of Watson–Crick base pairing.

Genome composition



Repetitive sequences - repeats

Some sequences in genome are **unique**, usually the genomic sequences (both coding and non-coding). In contrast, other sequences exist in many copies – **repetitive sequences (repeats)**. The length of repeat (microsatelites 2-6 bp x LINE 6-7000 bp), as well as the number of copies (several – 1.5M SINE in human) is highly variable.

Structure:

- **direct** repeats

```
5´...AGTC...AGTC...3´
3´...TCAG...TCAG...5´
```

```
5´...AGTC...CTGA...3´
3´...TCAG...GACT...5´
```

- **inverted** repeats + **palindromes**

```
5´...AGTC...GACT...3´
3´...TCAG...CTGA...5´
```

Position:

- **Tandem** repeats

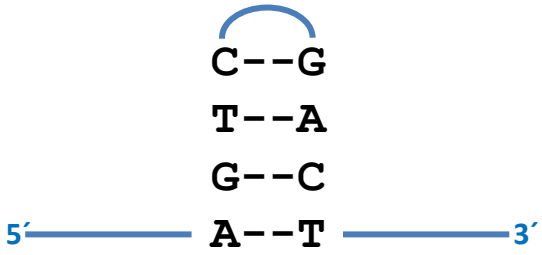
- **Interspersed** repeats

Inverted repeats

Palindrome

5´ AGTCGACT 3´

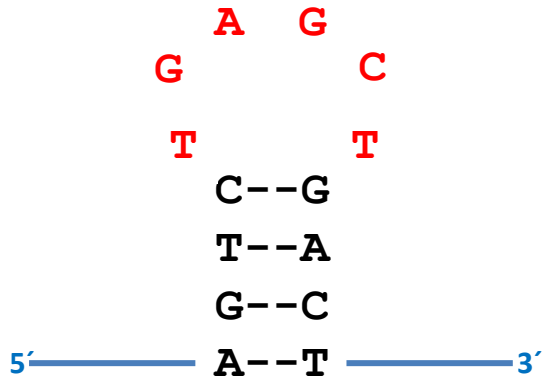
Hairpin



Inverted repeat

5´ AGTCTGAGCTGACT 3´

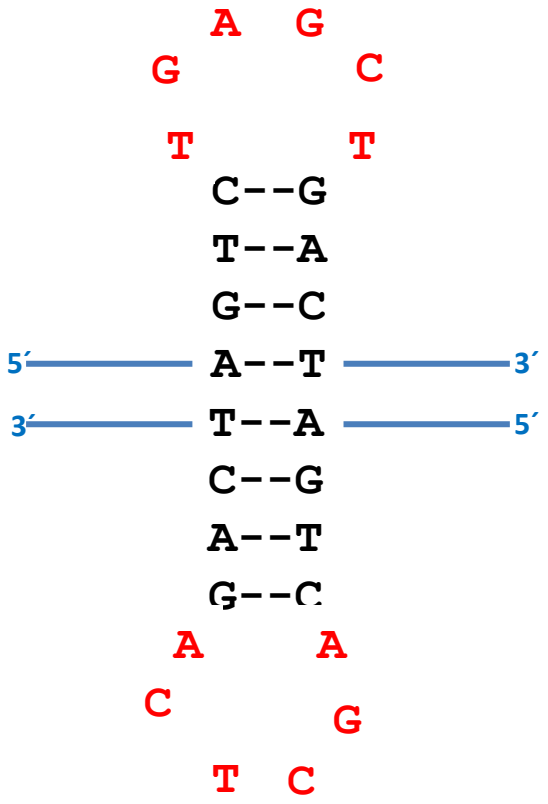
Hairpin with loop



Inverted repeats

Inverted repeat

Cruciform



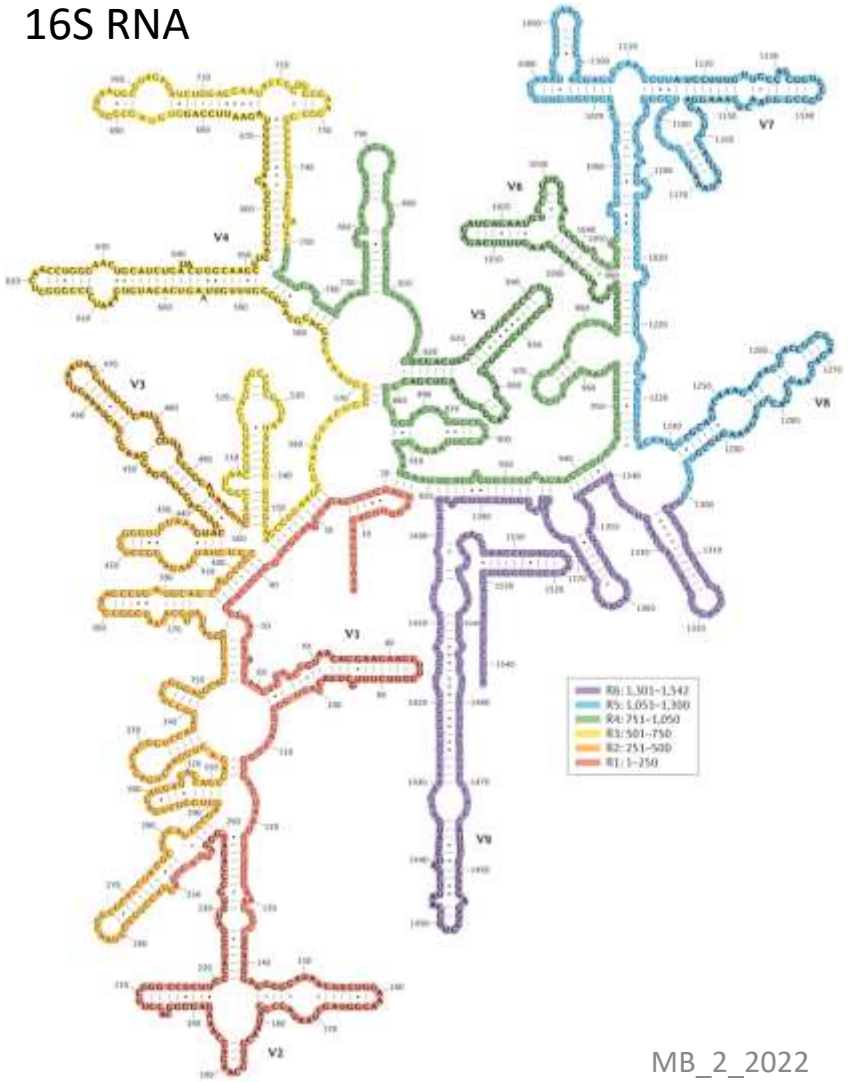
Special types of repetitions - transposons

Interspersed repetitions with various lengths and number of copies.

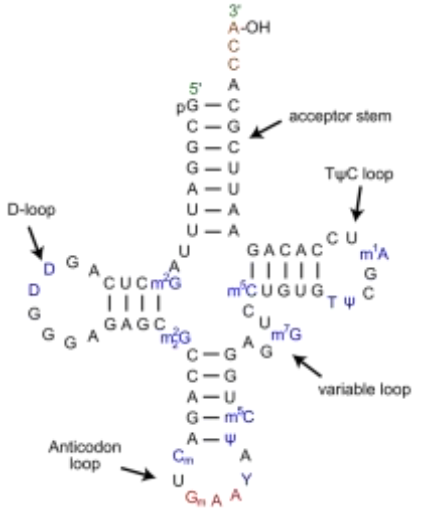
LTR – long terminal repeat - 100 bp – 5 kbp – variant of retrotransposons
LINE – long interspersed nuclear elements – up to 6 kbp – human> 500k copies
- 3 types (L1, L2, L3) – only some L1 are able to transpose
SINE – (Alu, …) short interspersed nuclear elements – up to 500 bp – human ~ 1,5M copies

Loops and hairpins in RNA

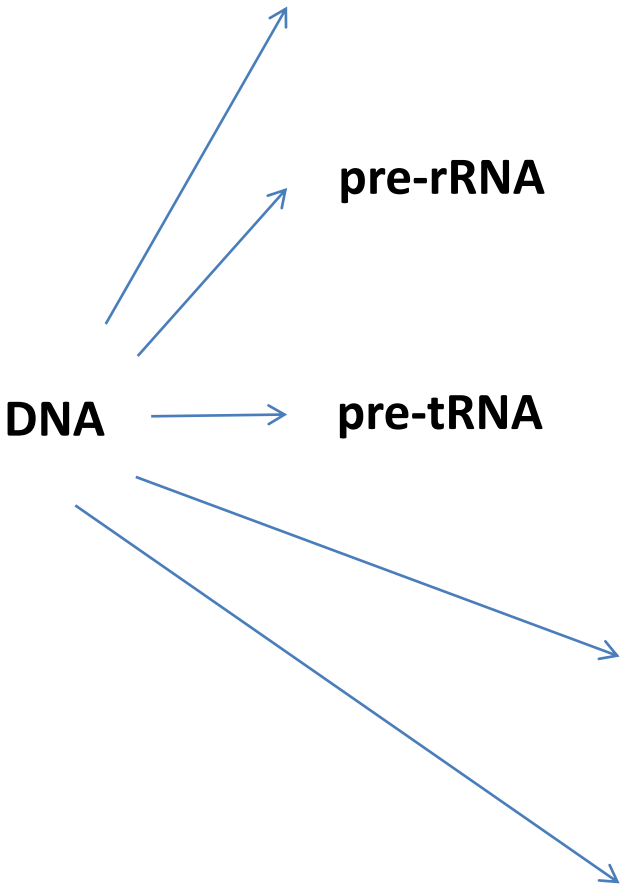
16S RNA



tRNA (Lys)



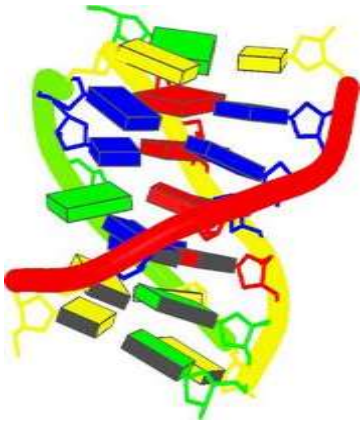
Functional types of RNA



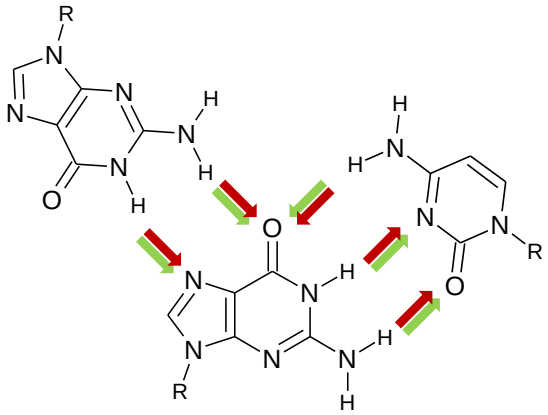
hnRNA	**mRNA**	Translation; codes for protein sequence
pre-rRNA	**rRNA**	Translation; part of ribosome
pre-tRNA	**tRNA**	Translation; amino acid carrier
	snRNA, snoRNA, …	Splicing/modification of RNA
	siRNA, miRNA, …	Gene expression regulation

Hoogsteen pairing - triplexes

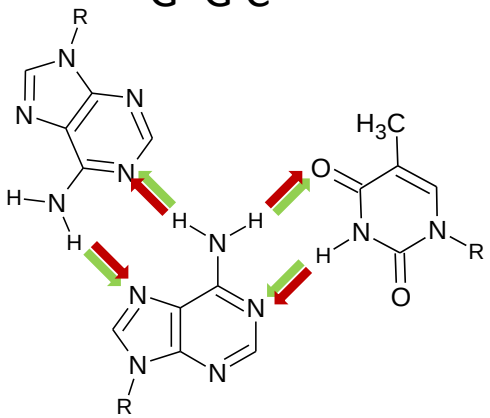
- gene expression regulation



Pu*Pu Py

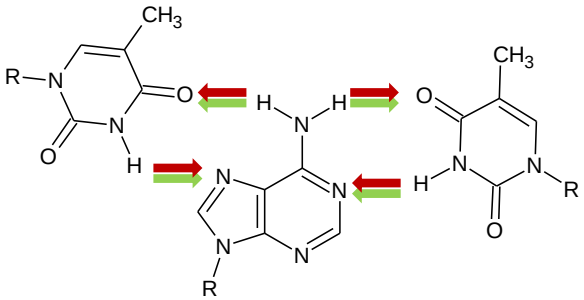


G*G C

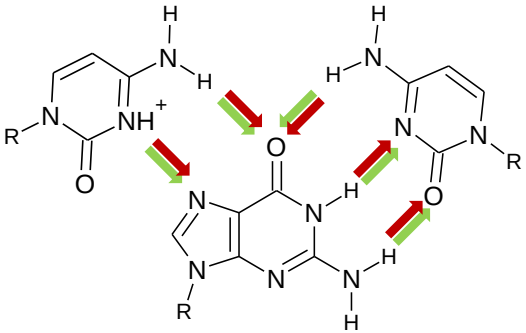


A*A T

Py*Pu Py



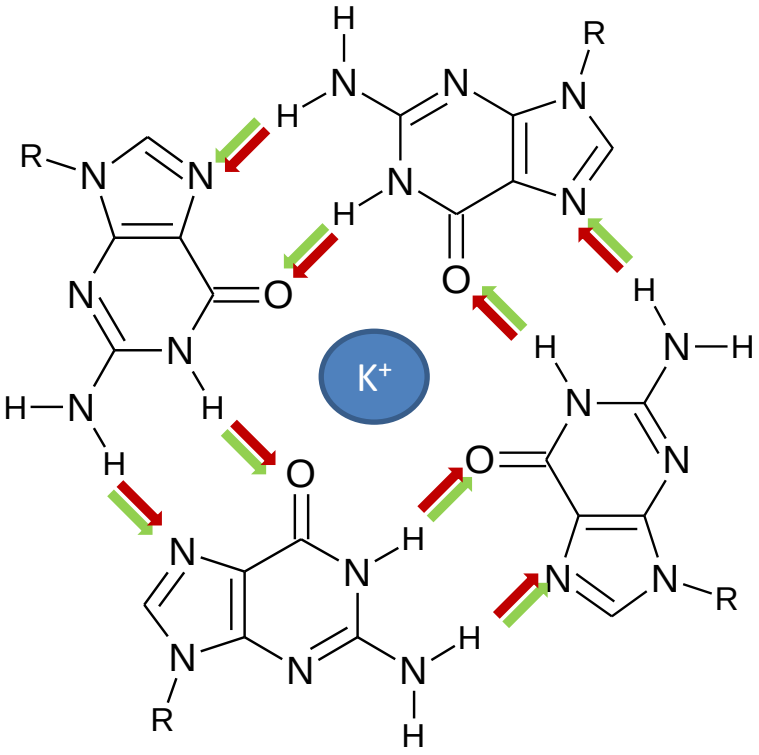
T*A T



C⁺*G C

TEST

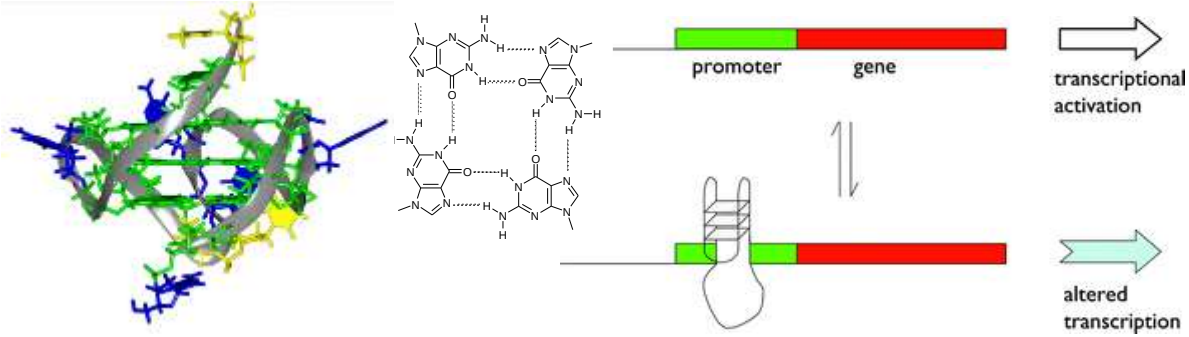
Hoogsteen pairing – G-quadruplexes



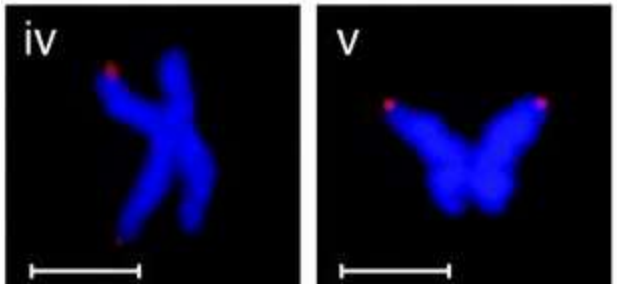
Guanine quadruplexes

$GGGN_nGGGN_nGGGN_nGGG$

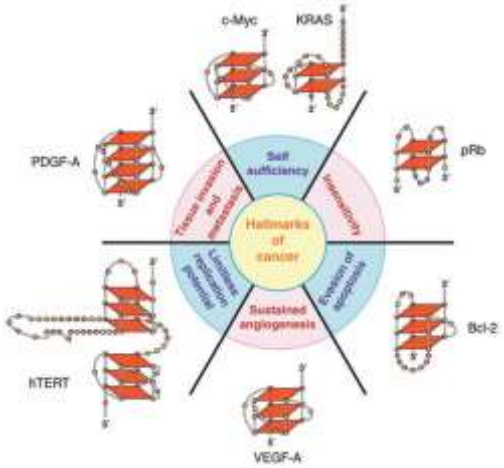
- gene expression regulation
- telomere structure



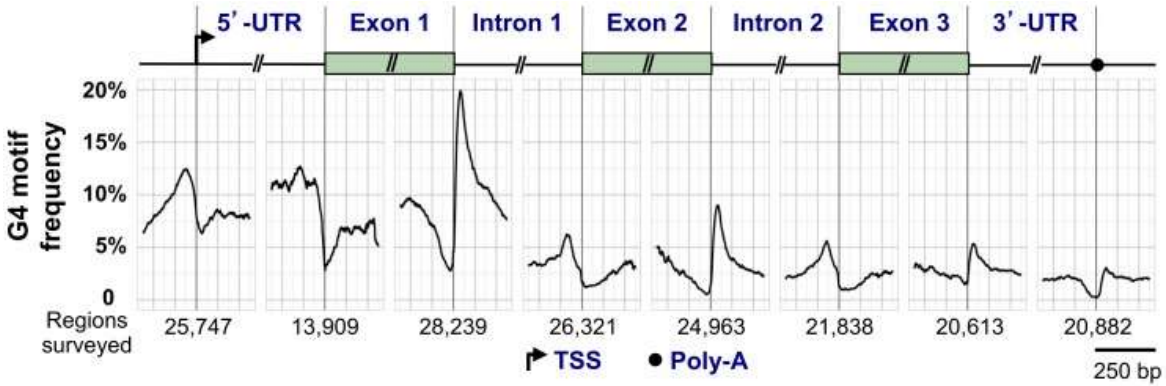
(Huppert J.L., Chem Soc Rev, 2008)



(Biffi G., et al., Nat Chem, 2013)



(Brooks T. A., et al., FEBS J, 2010)

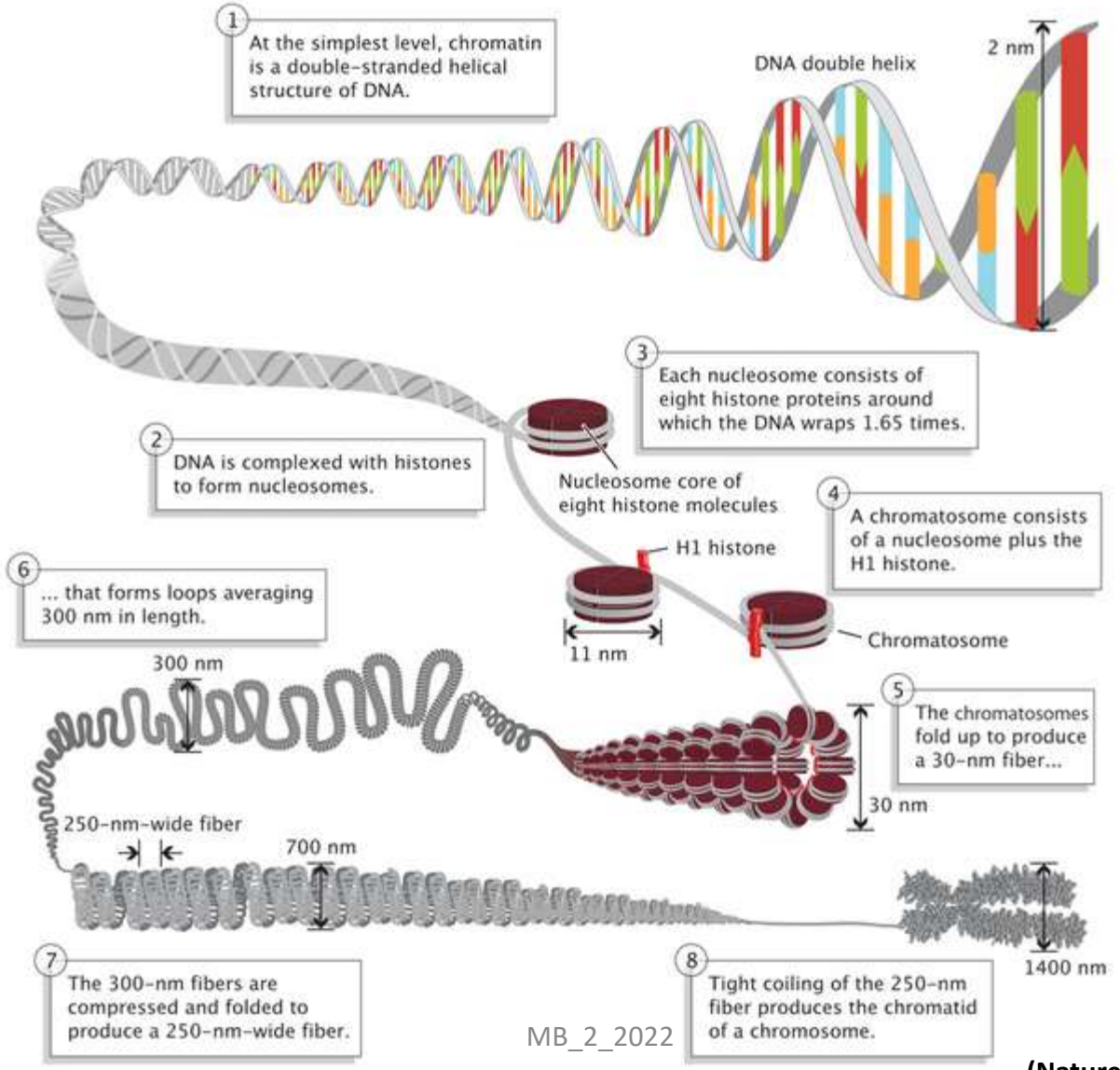


(Maizels N. and Gray L.T., PloS Genet., 2013)

Base reactivity

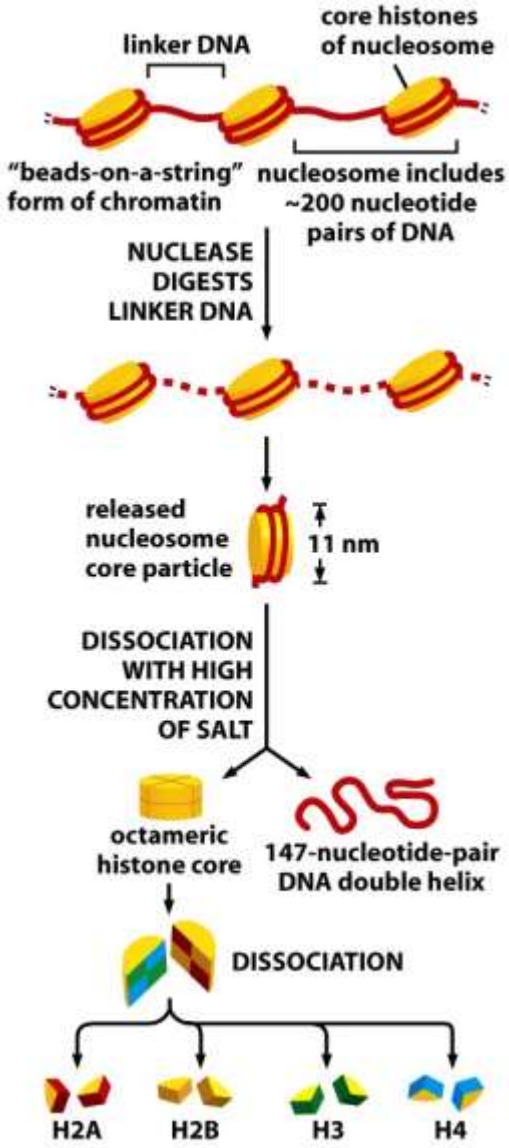
Hydrofobic bases with high ability to form hydrogen bonds are reluctant to be freely expressed into water environment around – if there is any chance to avoid this and lower the base exposition to the environment by any type of base pairing or base stacking, the bases tend to form a structure. Even the “single-stranded” RNA or DNA forms, in fact, compact structure with number of base pairs.

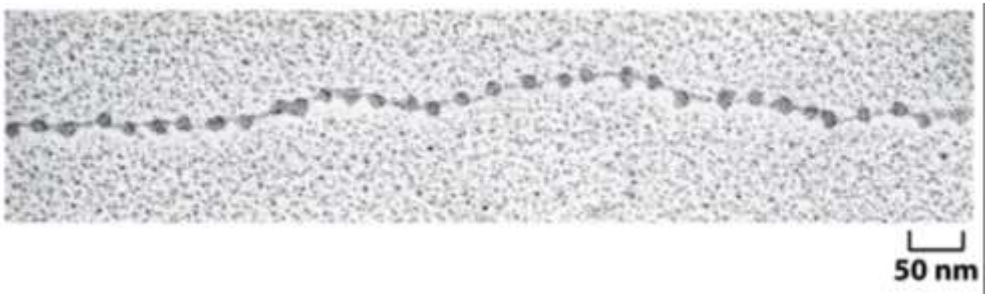
Packing of DNA into chromosome

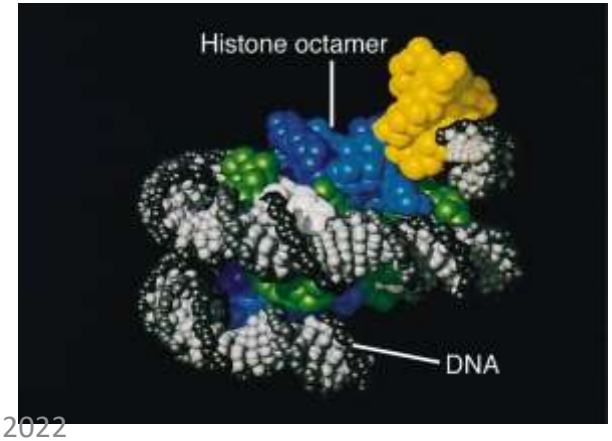


(Nature Education, 2013)

Binding of DNA to a histone octamer

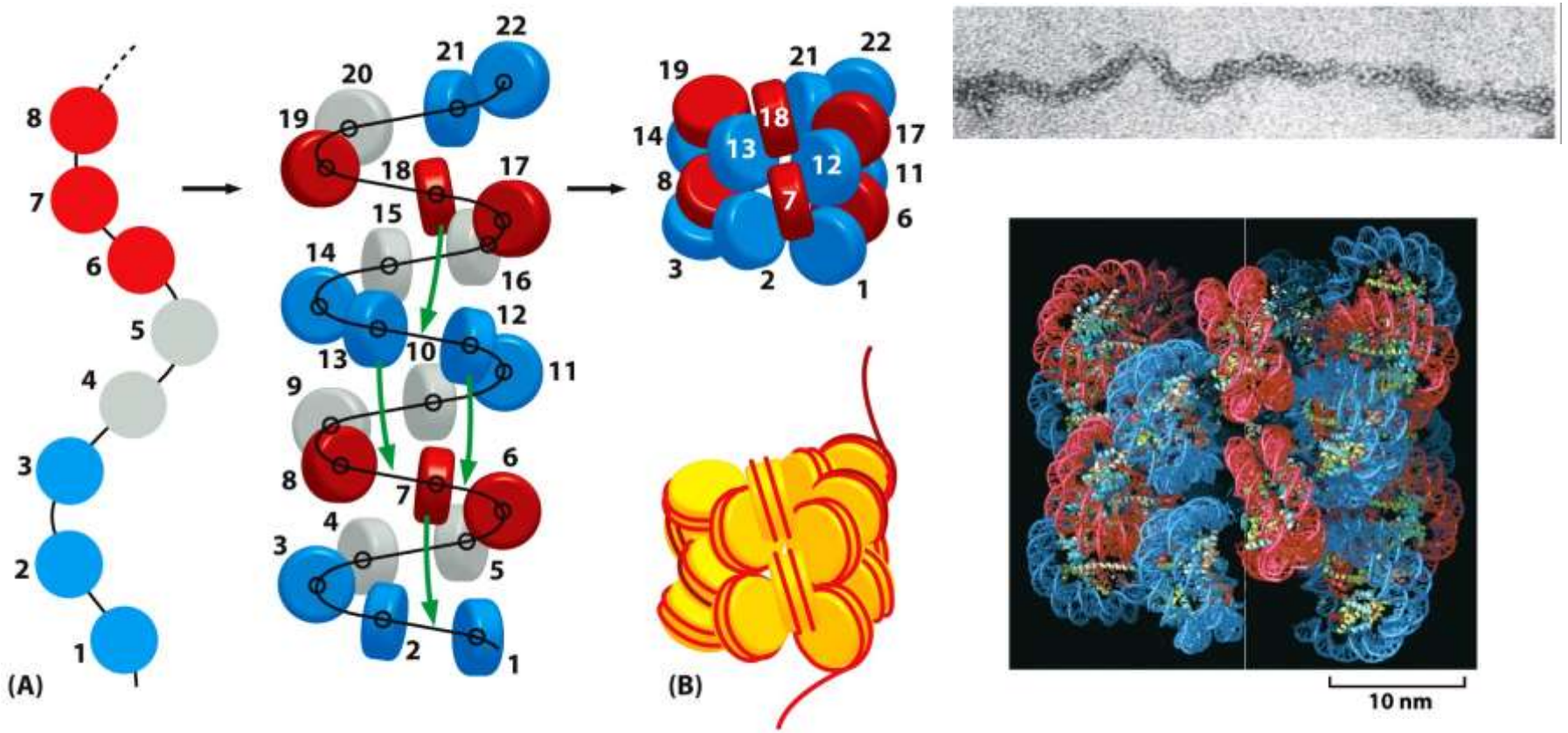




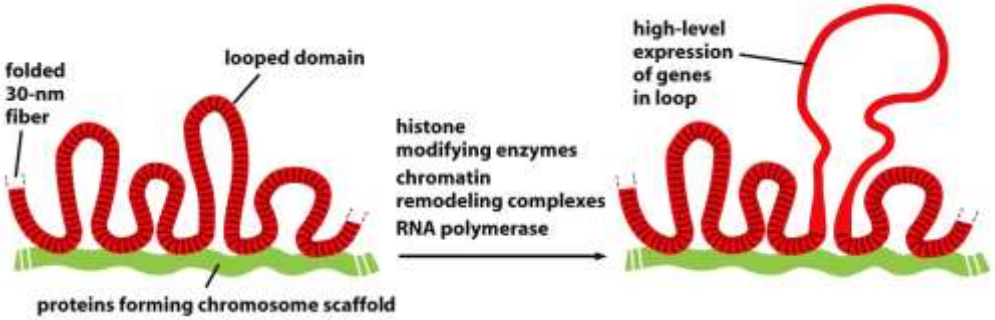


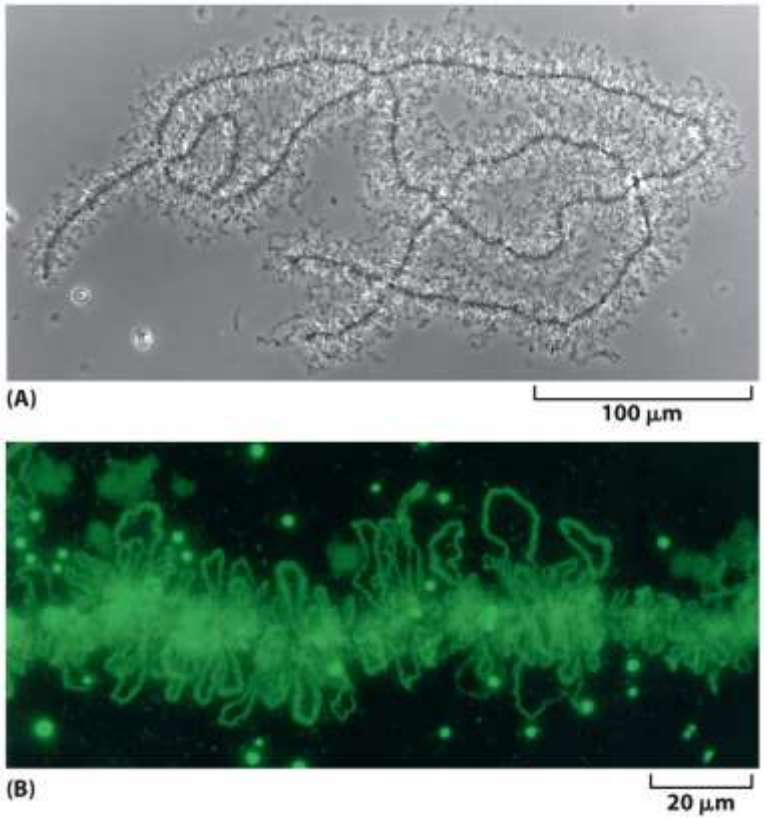
Molecular Biology of the Cell (© Garland Science 2008)

Folding of nucleosomes into 30 nm fiber

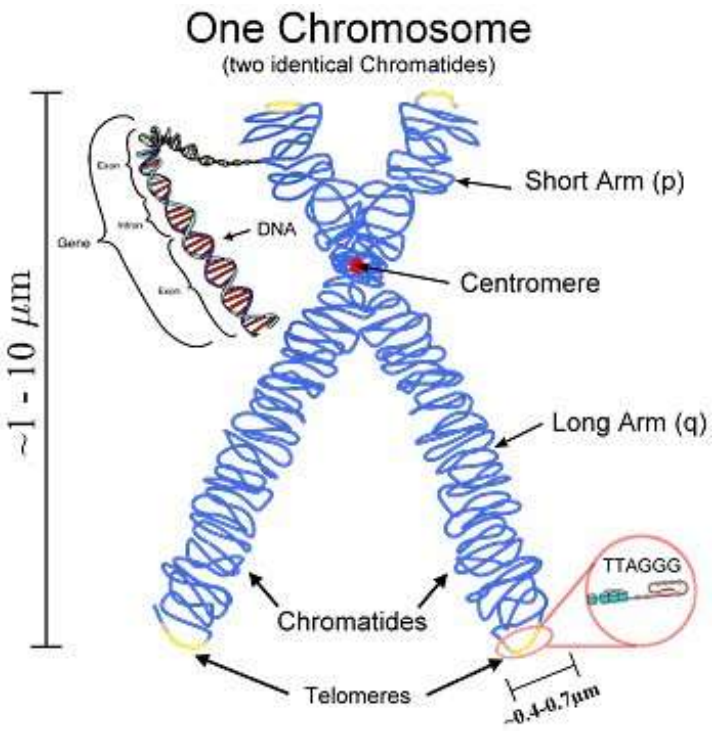


30 nm fiber binds to protein scaffold



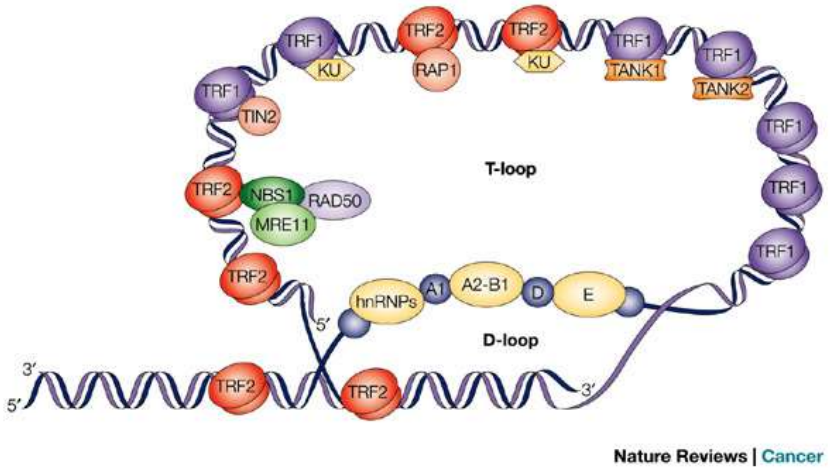


Chromosome



Centromere – here are the chromosomes connected to the system of cellular microtubules – important for chromosome segregation during cell division

Telomere – terminal part of chromatides that protect the end from being recognised as a double-strand break by a DNA repair machinery



Molecular Biology of the Cell (© Garland Science 2008)

Chromosome

Fully condensed chromosomes are present only during the cell division, otherwise they are more or less decondensed to a lower levels of structure, especially in transcriptionally active sites (**euchromatin**). Transcriptionally inactive parts of DNA, as well as repetitive regions or telomeres are much more condensed (**heterochromatin)**. Various types of chromatin differ in **epigenetic** markers of both DNA (5-methyl cytosine) and histones (methylation a acetylation).

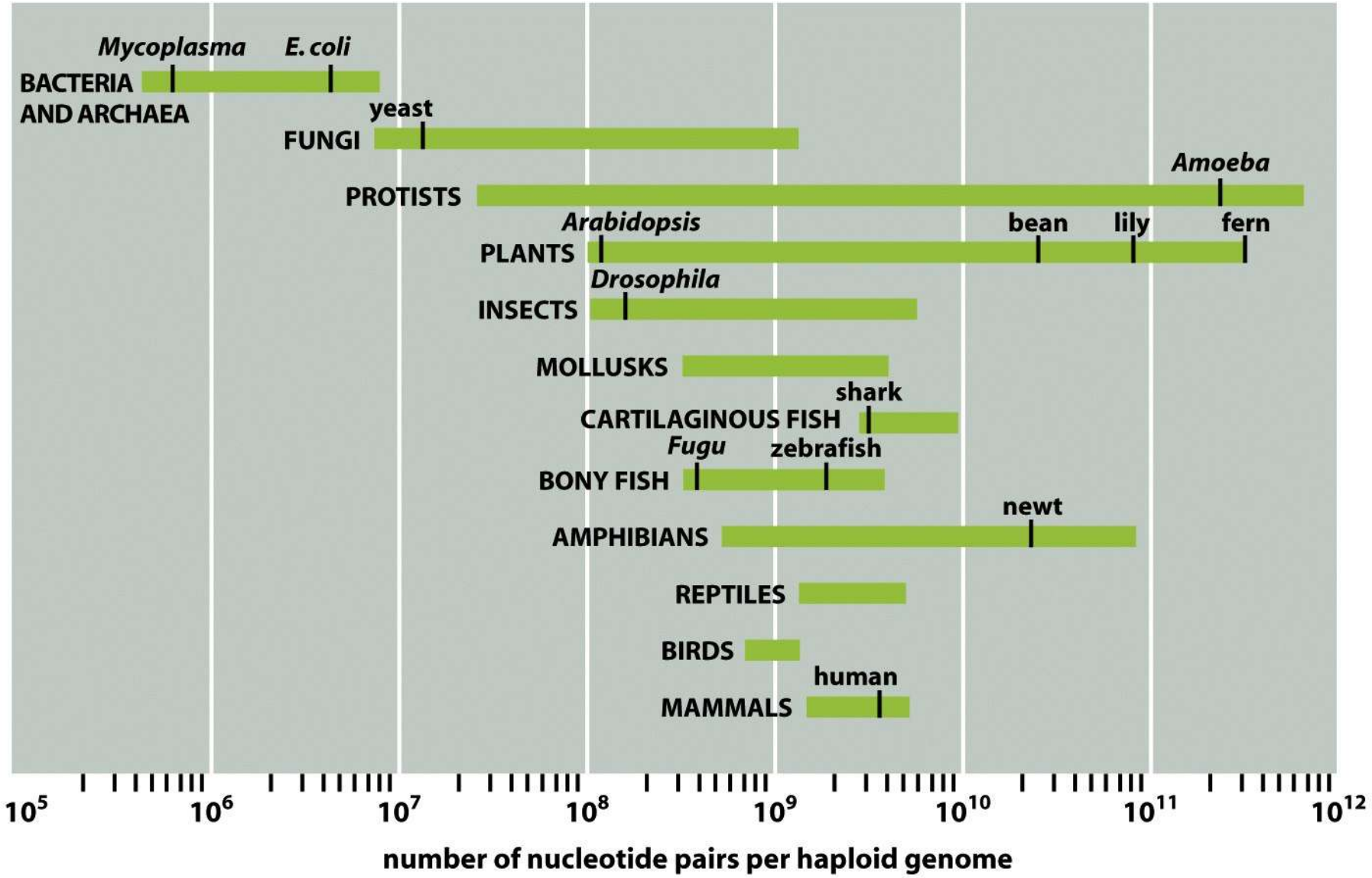


Figure 1-37 *Molecular Biology of the Cell*, Fifth Edition (© Garland Science 2008)

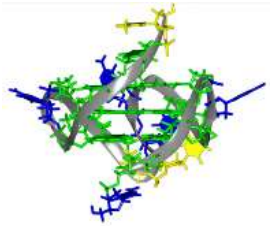
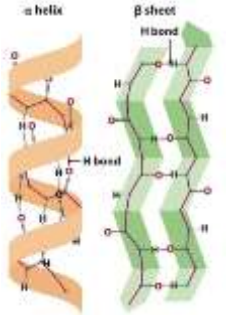
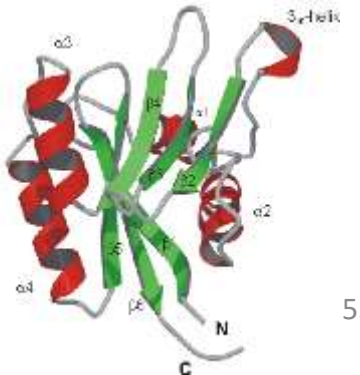
Table 1–1 Some Genomes That Have Been Completely Sequenced

SPECIES	SPECIAL FEATURES	HABITAT	GENOME SIZE (1000s OF NUCLEOTIDE PAIRS PER HAPLOID GENOME)	ESTIMATED NUMBER OF GENES CODING FOR PROTEINS
ARCHAEA				
Methanococcus jannaschii	lithotrophic, anaerobic, methane-producing	hydrothermal vents	1664	1750
Archaeoglobus fulgidus	lithotrophic or organotrophic, anaerobic, sulfate-reducing	hydrothermal vents	2178	2493
Nanoarchaeum equitans	smallest known archaean; anaerobic; parasitic on another, larger archaean	hydrothermal and volcanic hot vents	491	552
EUCARYOTES				
Saccharomyces cerevisiae (budding yeast)	minimal model eucaryote	grape skins, beer	12,069	~6300
Arabidopsis thaliana (Thale cress)	model organism for flowering plants	soil and air	~142,000	~26,000
Caenorhabditis elegans (nematode worm)	simple animal with perfectly predictable development	soil	~97,000	~20,000
Drosophila melanogaster (fruit fly)	key to the genetics of animal development	rotting fruit	~137,000	~14,000
Homo sapiens (human)	most intensively studied mammal	houses	~3,200,000	~24,000

Genome size and gene number vary between strains of a single species, especially for bacteria and archaea. The table shows data for particular strains that have been sequenced. For eucaryotes, many genes can give rise to several alternative variant proteins, so that the total number of proteins specified by the genome is substantially greater than the number of genes.

Table 1-1 (part 2 of 2) *Molecular Biology of the Cell*, Fifth Edition (© Garland Science 2008)

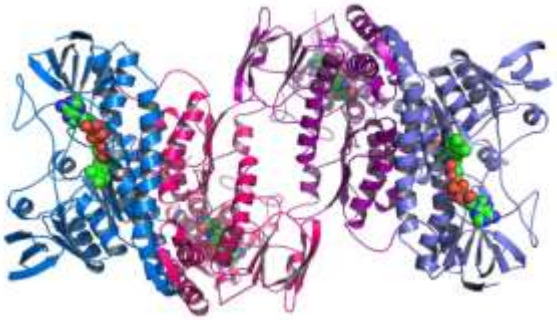
Levels of structure of biopolymers

	DNA	RNA	Protein
Primary	AGGCTGCAAGTCGAT	UCCGACGUUCAGCUA	Ser-Asp-Val-Gln-Leu
Secondary		 	
Tertiary			

Genetic code

Set of rules that assign a sequence of aminoacids in the protein to the sequence of nucleotides in DNA or RNA.

Transcription
+
Translation



M.W.Nirenberg

RNA CODEWORDS AND PROTEIN SYNTHESIS, III.
ON THE NUCLEOTIDE SEQUENCE OF A CYSTEINE AND A LEUCINE RNA CODEWORD

By Philip Leder and Marshall W. Nirenberg

National Heart Institute, National Institutes of Health

Communicated by Richard B. Roberts, October 1, 1964

Previous studies utilizing randomly ordered synthetic polynucleotides to direct amino acid incorporation into protein in *E. coli* extracts indicated that RNA codewords corresponding to valine, leucine, and cysteine contain the bases (UUG).[1–4] The activity of chemically defined trinucleotides in stimulating the binding of a specific C^{14}-aminoacyl-sRNA to ribosomes, prior to peptide bond formation,[3] provided a means of investigating base sequence of RNA codewords and showed that the sequence of a valine RNA codeword is GpUpU.[5]

Properties of genetic code

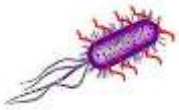
• genetic code is based on **triplets** – one aminoacid in protein is coded by a sequence of three nucleotides in DNA (RNA)

mRNA CGUGGUACGAUUGGAUGU

Protein **Arg Gly Thr Ile Gly Cys**

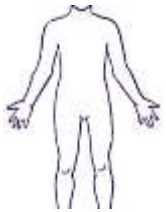
Triplet = Codon x anticodon = complementary sequence on particular tRNA that carries the respective aminoacid

• genetic code is **universal** – individual triplets code for the same aminoacid in almost all organisms (x mitochondria)



CGU = **Arginine**

CGU = **Arginine**



CGU = **Arginine**

• genetic code is **degenerated** – one aminoacid might be coded by several different triplets (but the opposite is not true)

Genetic code

	Second nt				
First nt	U	C	A	G	Third nt
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP/Sel	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met/START	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Reading frames

Genetic code is based on triplets – three possible ways of reading (**reading frames**), but only one is correct.

mRNA CGUGGUACGAUUGGAUGU

Protein1 **Arg Gly Thr Ile Gly Cys**

mRNA CGUGGUACGAUUGGAUGU

Protein2 **Val Val Arg Leu Asp**

mRNA CGUGGUACGAUUGGAUGU

Protein3 **Trp Tyr Asp Trp Met**

Genetic code

Although the genetic code is universal, the usage of particular codons, as well as the amount of particular tRNAs and aminoacyl transferases differ

Modification of synthetic genes for recombinant protein production according to the expression system used (Bacteria, human, …) might be highly beneficial.