Geometry applied to DNA

Ángel Ferrández Izquierdo
aferr@um.es
webs.um.es/aferr
Departamento de Matemáticas
Universidad de Murcia

Why Geometry (or better, Geometry & Topology) applied?

Three (main and personal) reasons:

1. Geometry & Topology is the area where I was always working

2. I think that Geometry & Topology is the best area in Maths being useful to any other branch of Science

3. As time goes on, one would really wish to find some application of his own research

Why applied to DNA?

Out of curiosity, simply

However, at the end of 2002, Prof. A. Romero invited to me to publish in La Gaceta de la RSME.

Then I remembered that the 50th anniversary of double helix Watson & Crick model of DNA will be celebrated all over the world on april 2003.

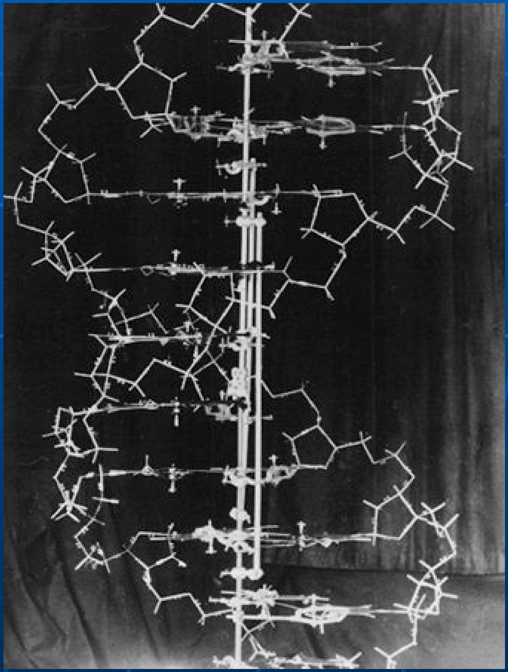
Then we took this nice opportunity to dust down a project that had been shelved years ago and write a paper on some mathematical topics involved in the DNA Watson & Crick model (*La Gaceta Vol. 6.3, 2003, 557-570*).

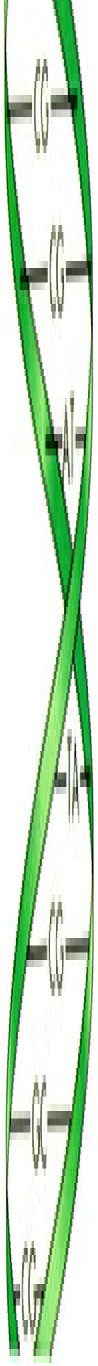
DNA Watson & Crick model

In 1953, James Watson and Francis Crick

analysing the X-ray photographs of DNA fibres by Rosalind Franklin

deduced the 3-dimensional structure of DNA.





The important features of the DNA W-C model are (1953):

(a) Two helical polynucleotide chains coiled around a common axis.

(b) The purine and pyrimidine bases are on the inside of the helix whereas the phosphate and deoxyribose units are situated on the outside.

(c) The diameter of the helix is 2 nm. Adjacent bases are separated by 0.34 nm along the helix axis. The helix repeats itself every 10 base pairs at intervals of 3.4 nm.

(d) The two chains held together by hydrogen bonds formed between pairs of bases. Pairing is highly specific. Adenine pairs with thymine, guanine always pairs with cytosine. A = T, G = C.

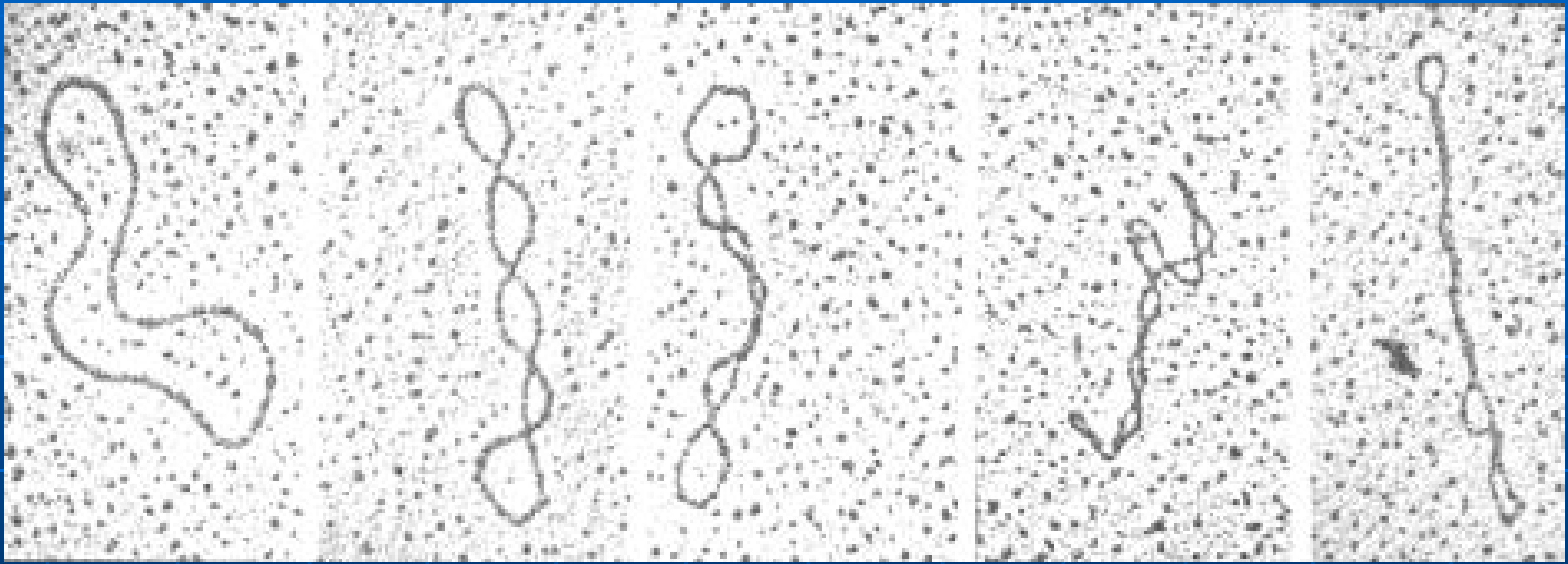
(e) The sequence of bases along the polynucleotide chain is not restricted. The precise sequence of bases carries the genetic information.

Some important features of the DNA today

- DNA is a flexible structure.
- Because their ends are free, linear DNA molecules can freely rotate to accommodate changes in the number of times the two chains of the double helix twist about each other.
- If the two ends are covalently linked to form a circular DNA molecule, then the absolute number of times the chains can twist about each other cannot change. Such a closed, circular DNA is said to be topologically constrained.
- DNA participates in numerous dynamic processes in the cell. For example, the two strands of the double helix, which are twisted around each other, must rapidly separate in order for DNA to be duplicated and to be transcribed into RNA.

Thus, understanding the Geometry and Topology of DNA and how the cell both accommodates and exploits topological constraints during DNA replication, transcription -and other chromosomal transactions- is of fundamental importance in molecular biology.

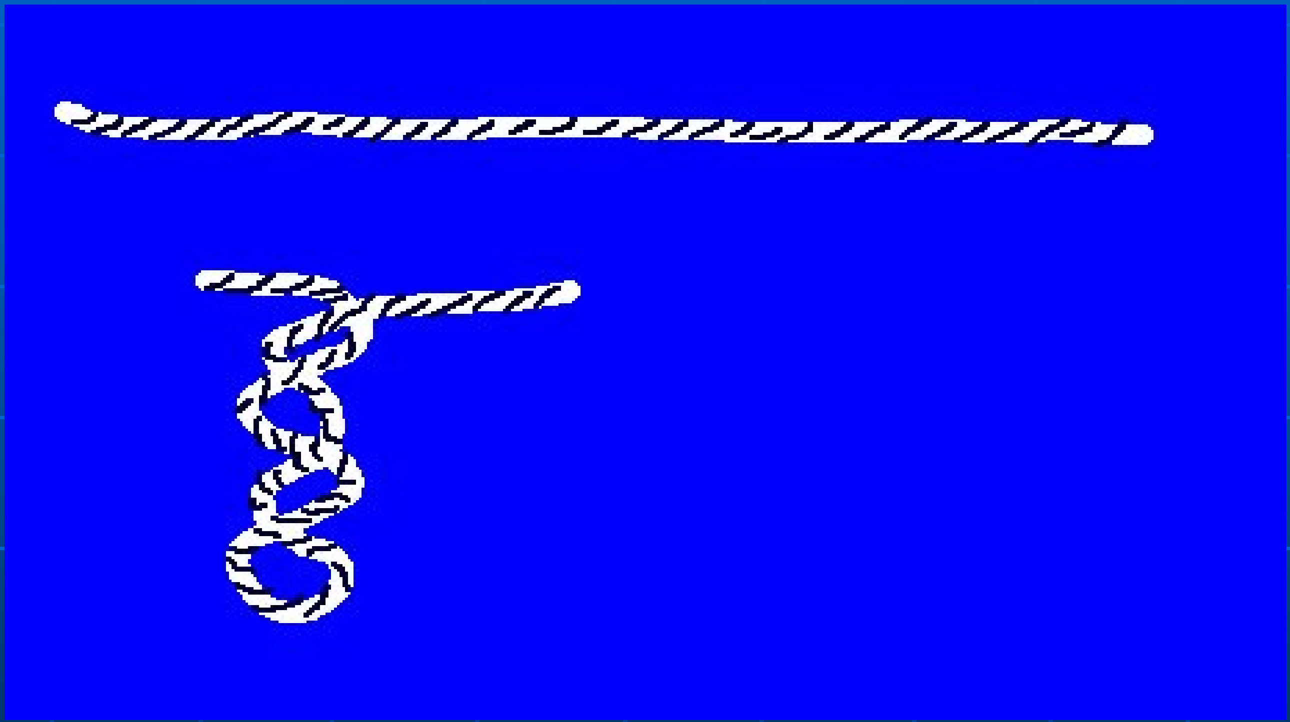
How does circular DNA molecule appear ?



Supercoiling of DNA

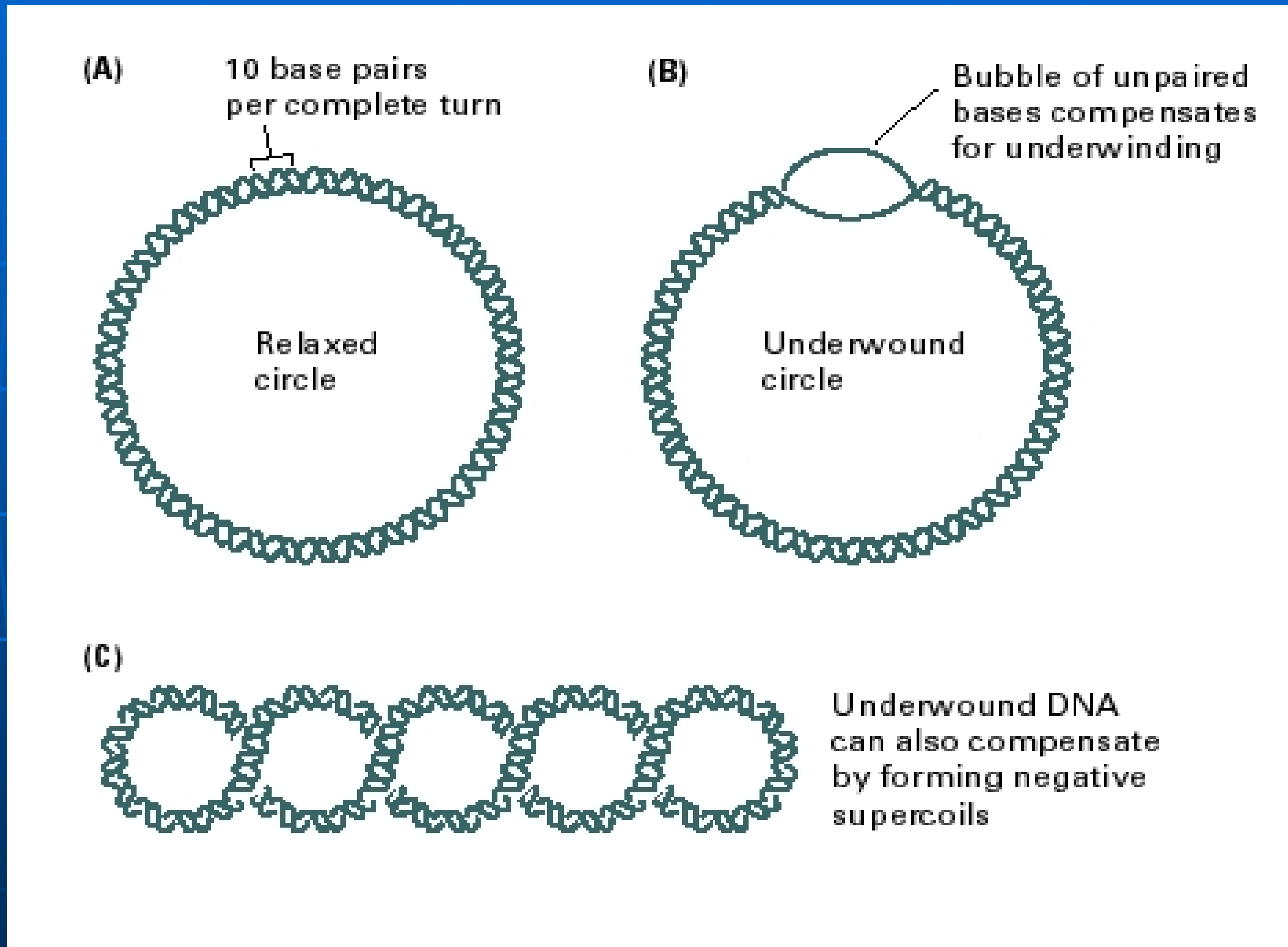
- The tension induced in a circular DNA molecule causes it to become supercoiled
- Supercoiling is the usual state for bacterial chromosomes, which consist of a number of independently supercoiled loops
- The process is controlled by topoisomerase enzymes that can cut and re-join one strand of the DNA
- Topoisomerases can also untangle DNA

Supercoiled DNA I



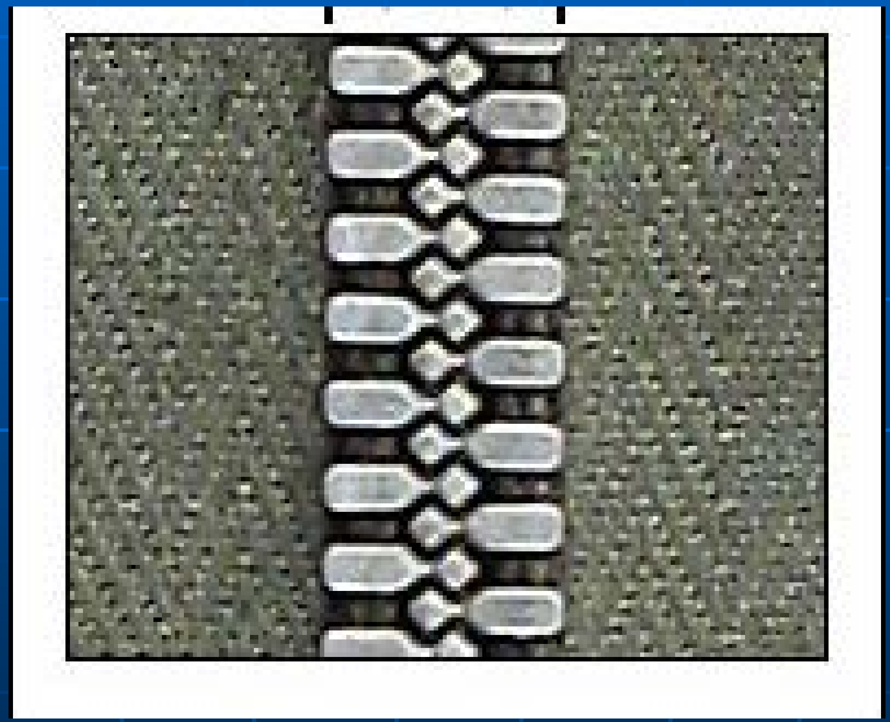
DNA, like a telephone cable

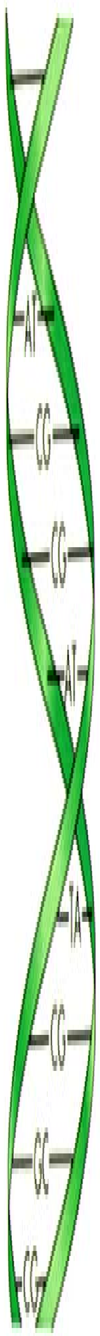
Supercoiled DNA II



How to imagine a 2-helix DNA

A zip-fastener or zipper





Some personal notes I

- The double strand helical configuration adopted by DNA was not produced by chance.
- Actually, the Nature is guided by its own rules.
- Some time ago, Prof. M Barros and myself stated the following natural question:

What kind of helical configurations do really occur in Nature ?

- To answer this question we have proposed a model to describe the genuine helices which we will meet in the real world.

Some personal notes II

- Now, the model should be subject to three main principles:

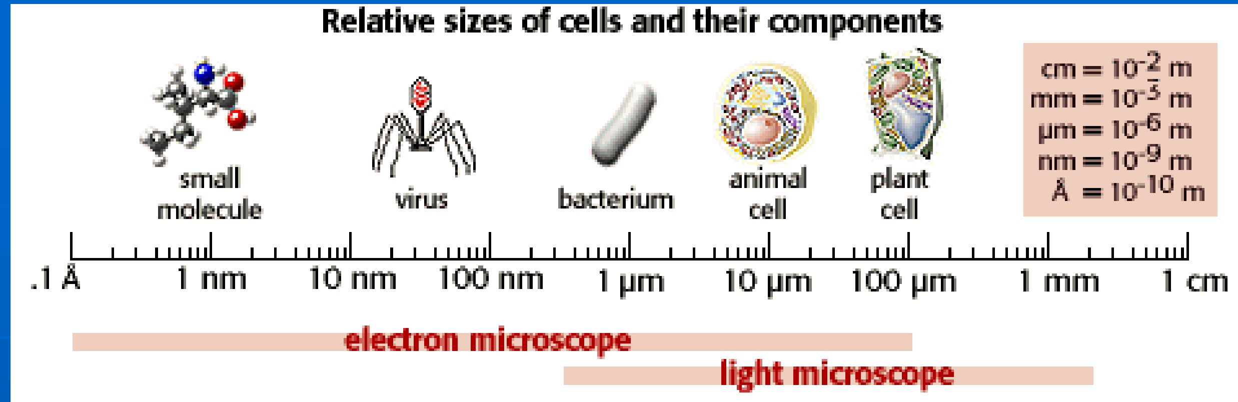
<u>*Variational:*</u> Admissible helical structures in Nature should be, as soon as possible, extremals of a reasonable elastic energy action.

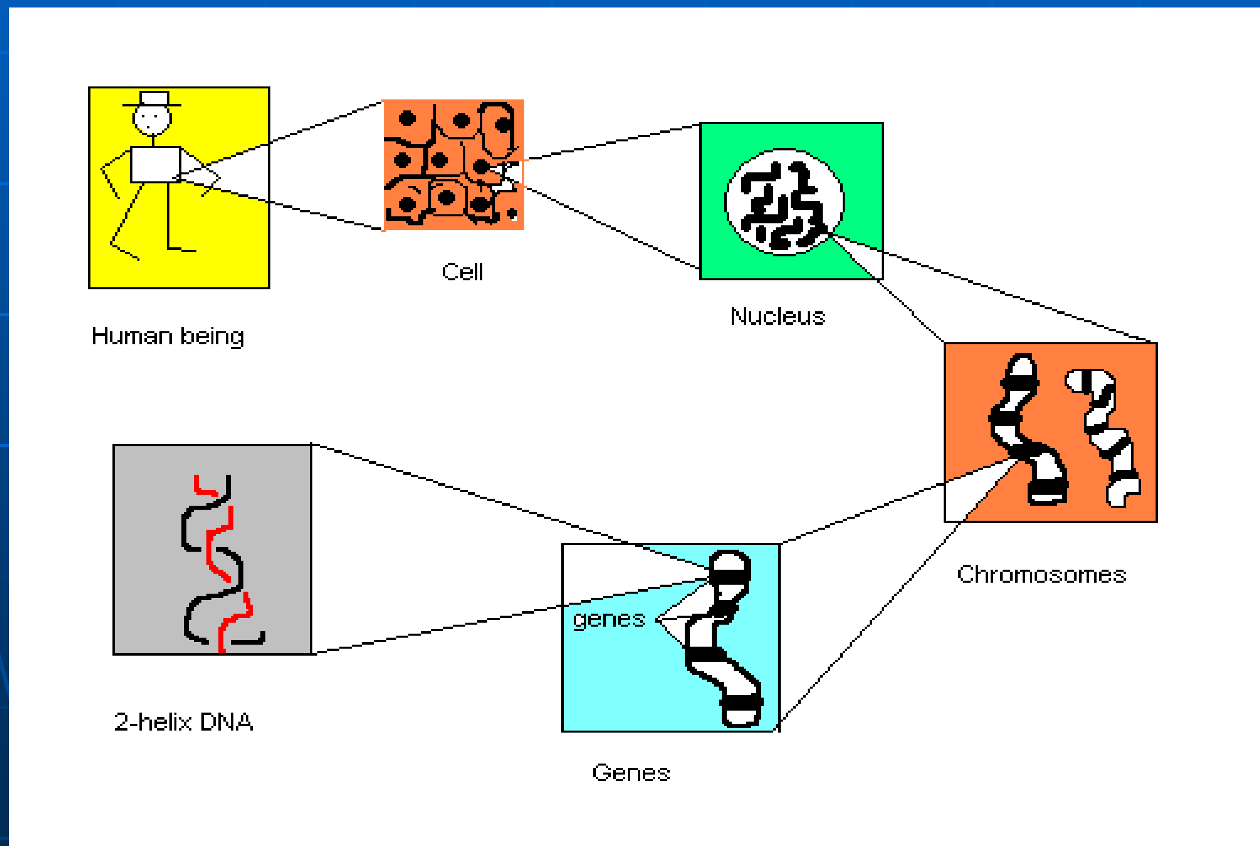
<u>*Topological:*</u> The model should be able to provide closed helical structures.

<u>*Quantization:*</u> Some important properties associated with the solutions of this model should be quantized.

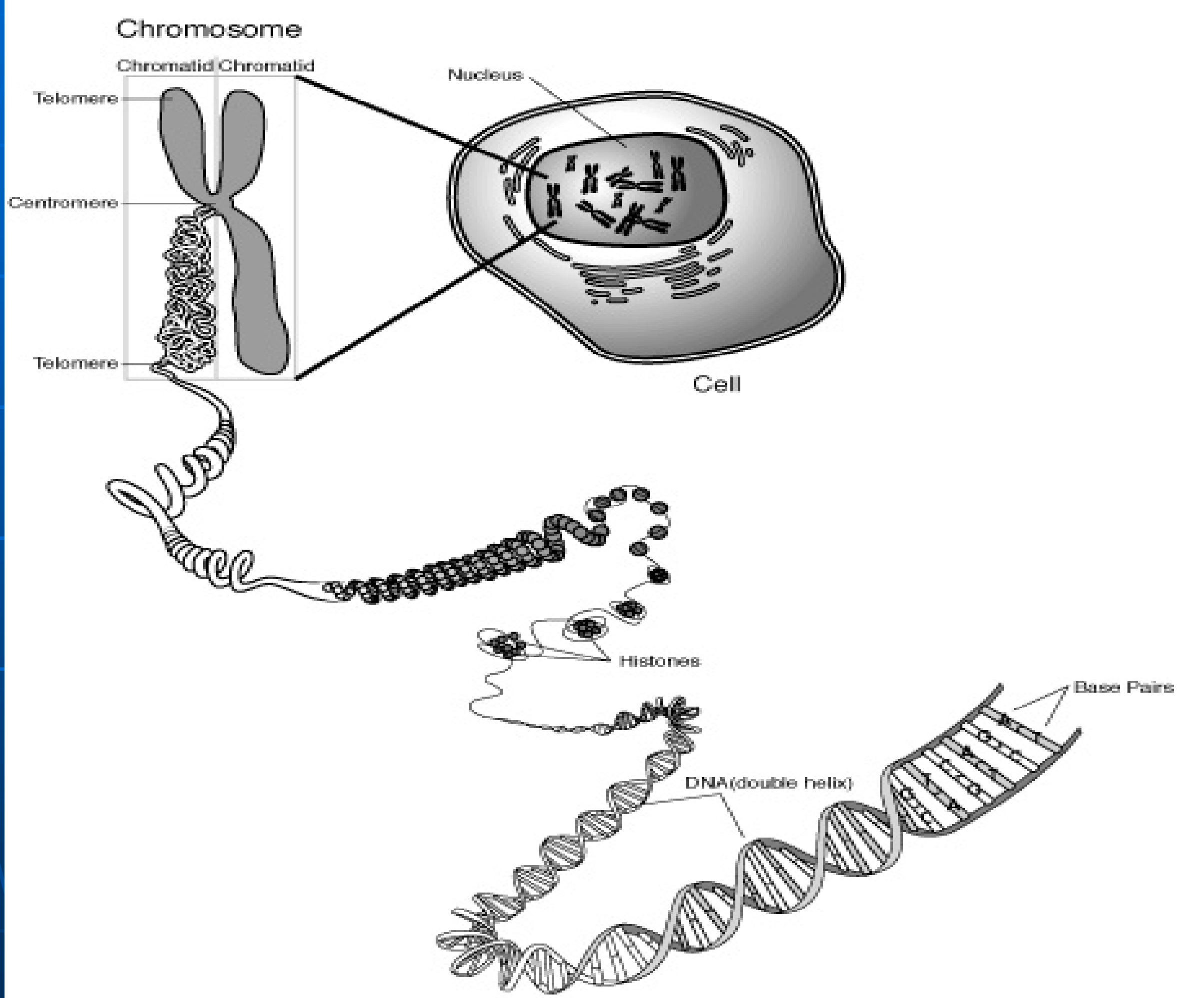
OUR ANSWER: All helices should be of Lancret-type,
i. e., loxodromes in some surface
(you can get a copy of our results)

Where are we looking at?



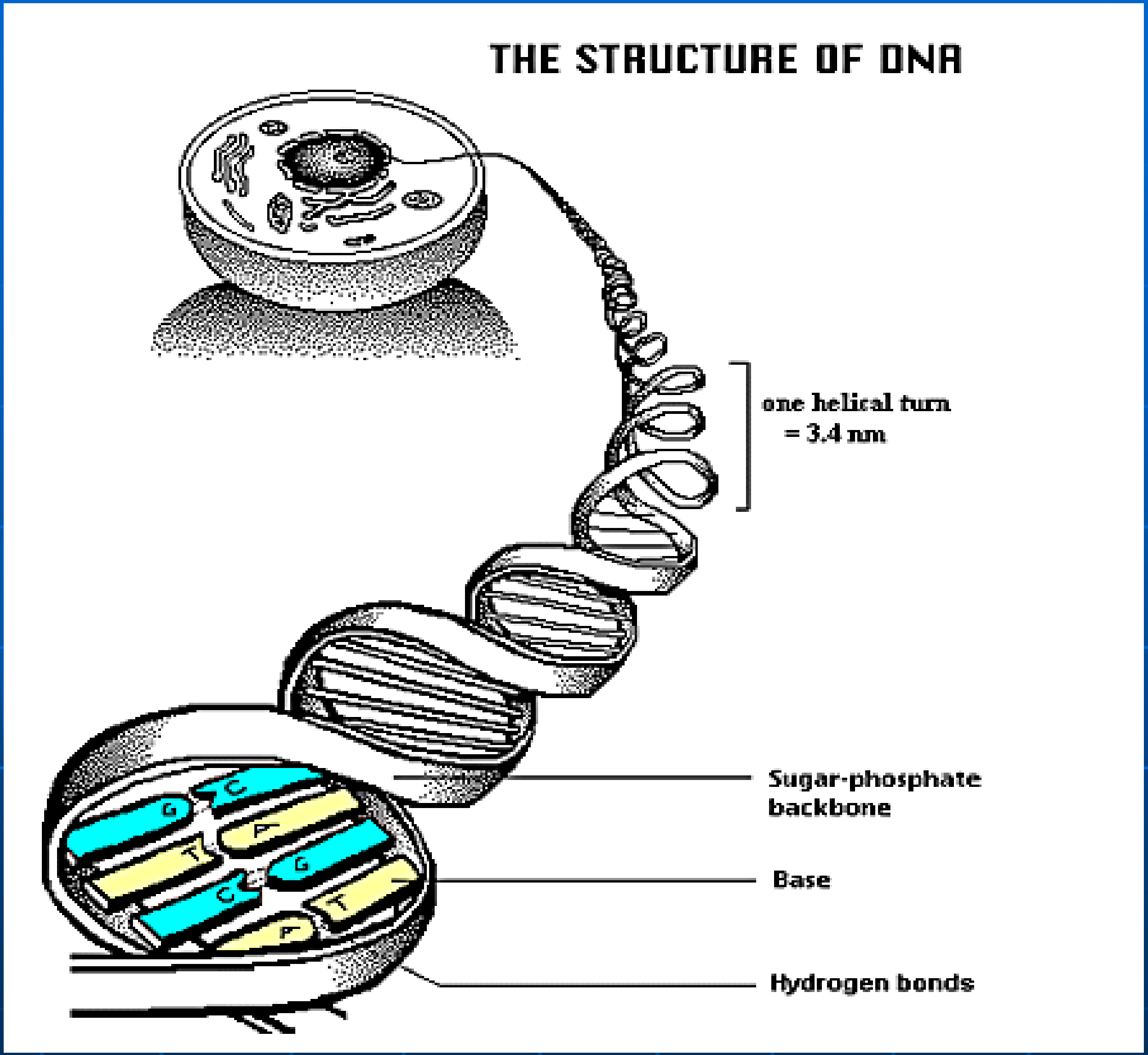


Getting down to business



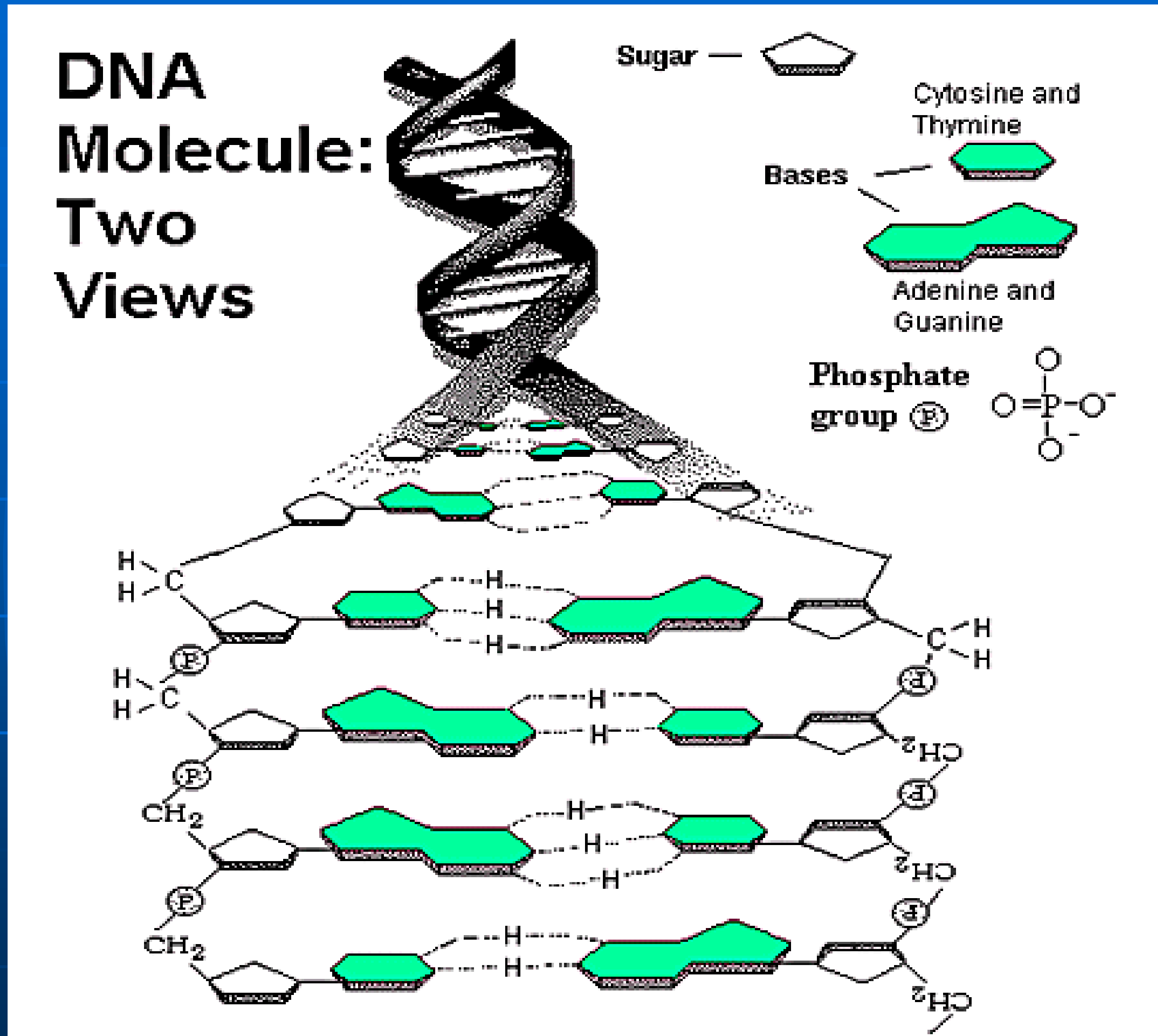
http://www.accessexcellence.org/AB/GG/chromosome.html

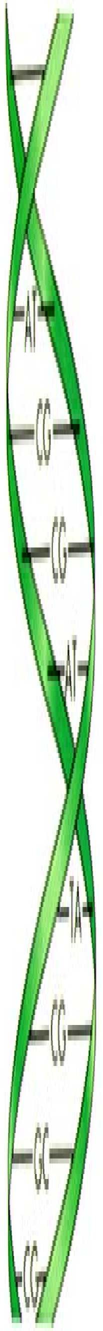
More details



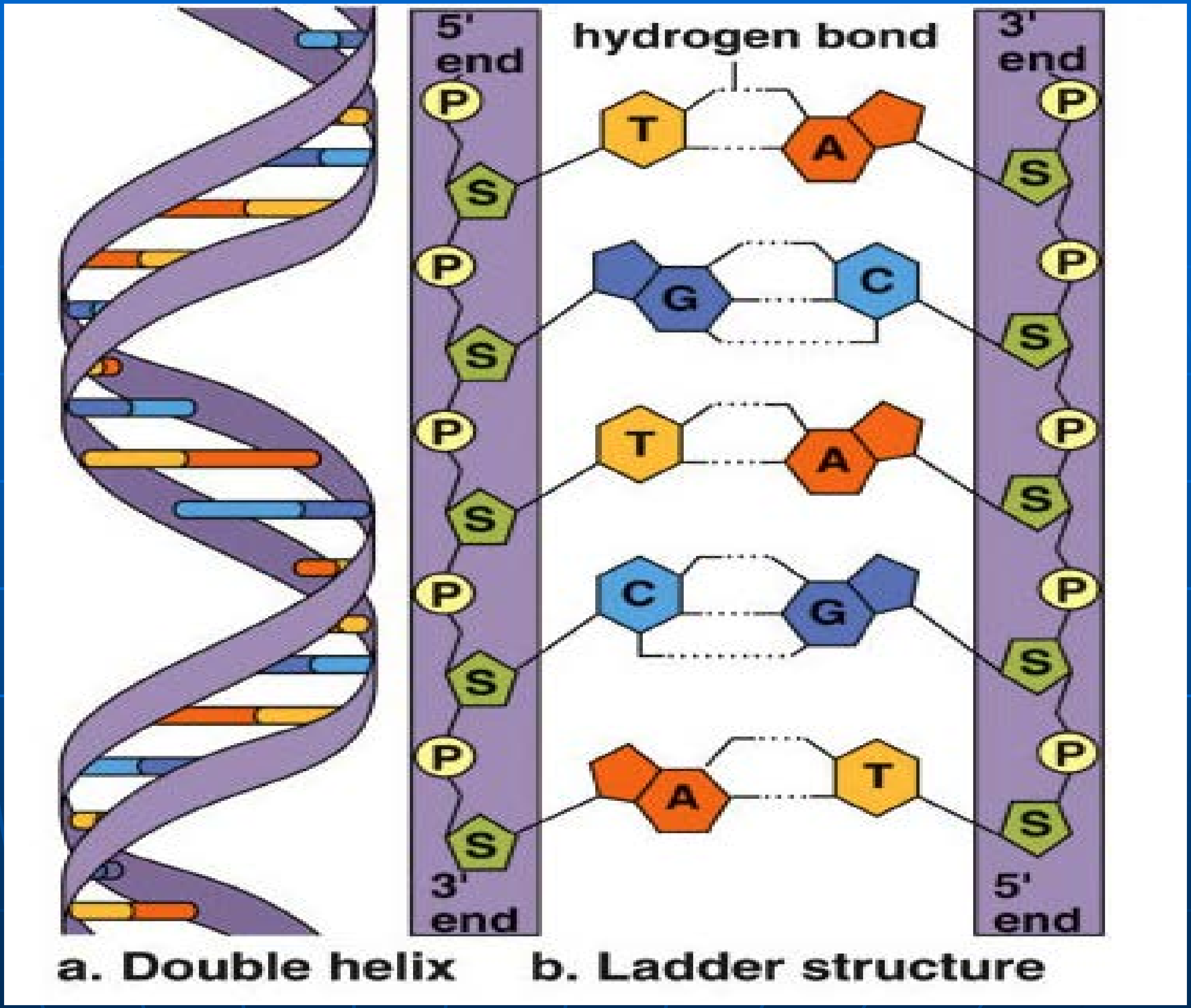
Relative sizes

By scaling the cell nucleus up to the size of a basketball, the DNA inside scales to the size of thin fishing line, so that 200 km of that fishing line are inside the nuclear basketball.

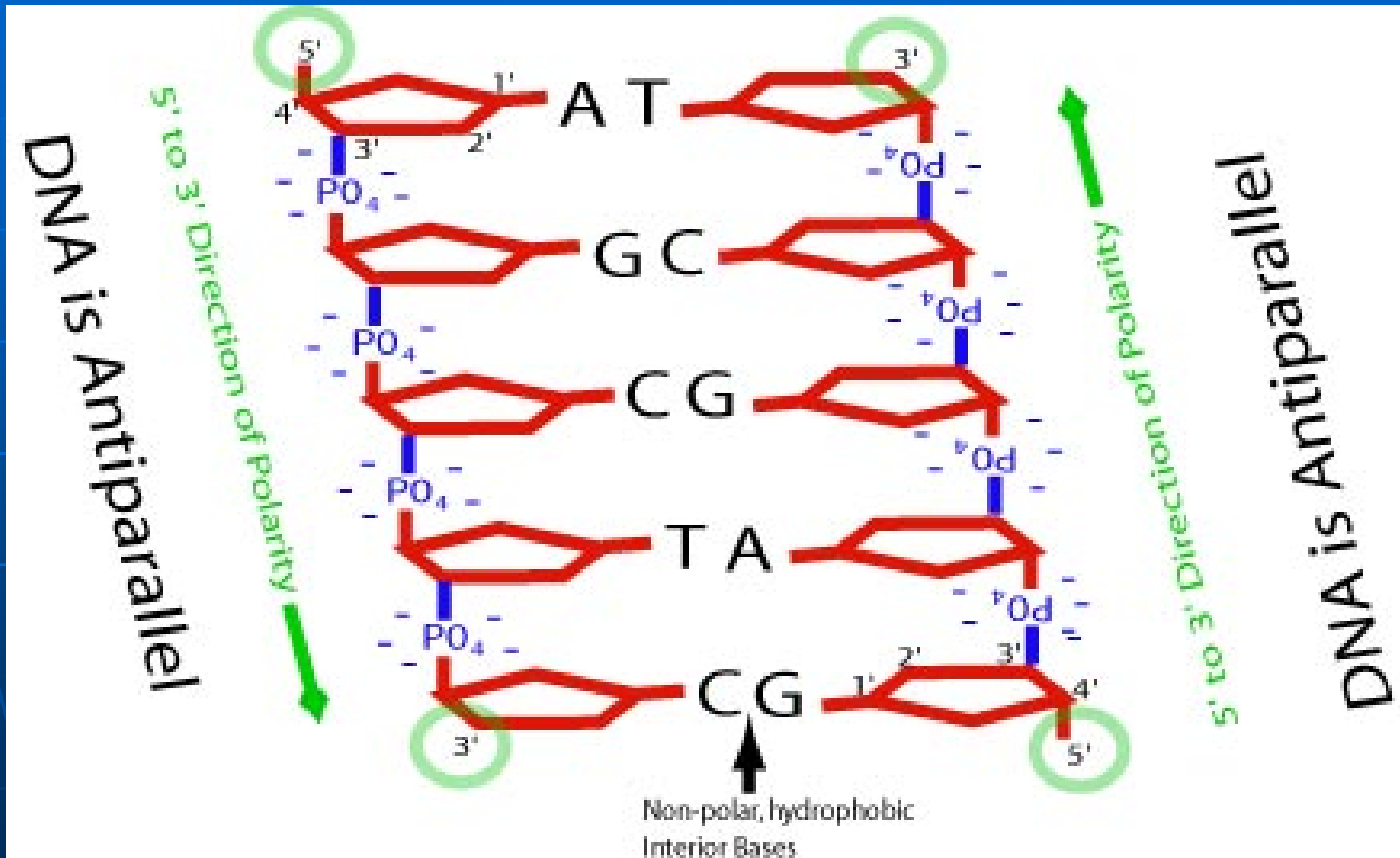

DNA
Molecule:
Two
Views
Sugar
Cytosine and
Thymine
Bases
Adenine and
Guanine
Phosphate
group (P)

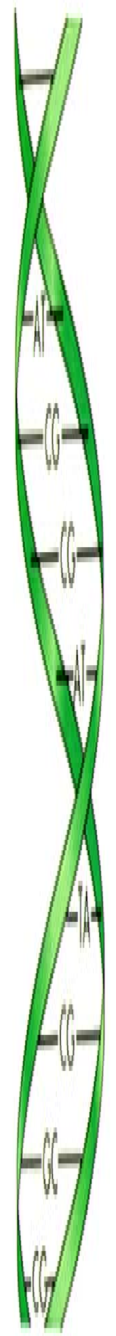


Planar view point

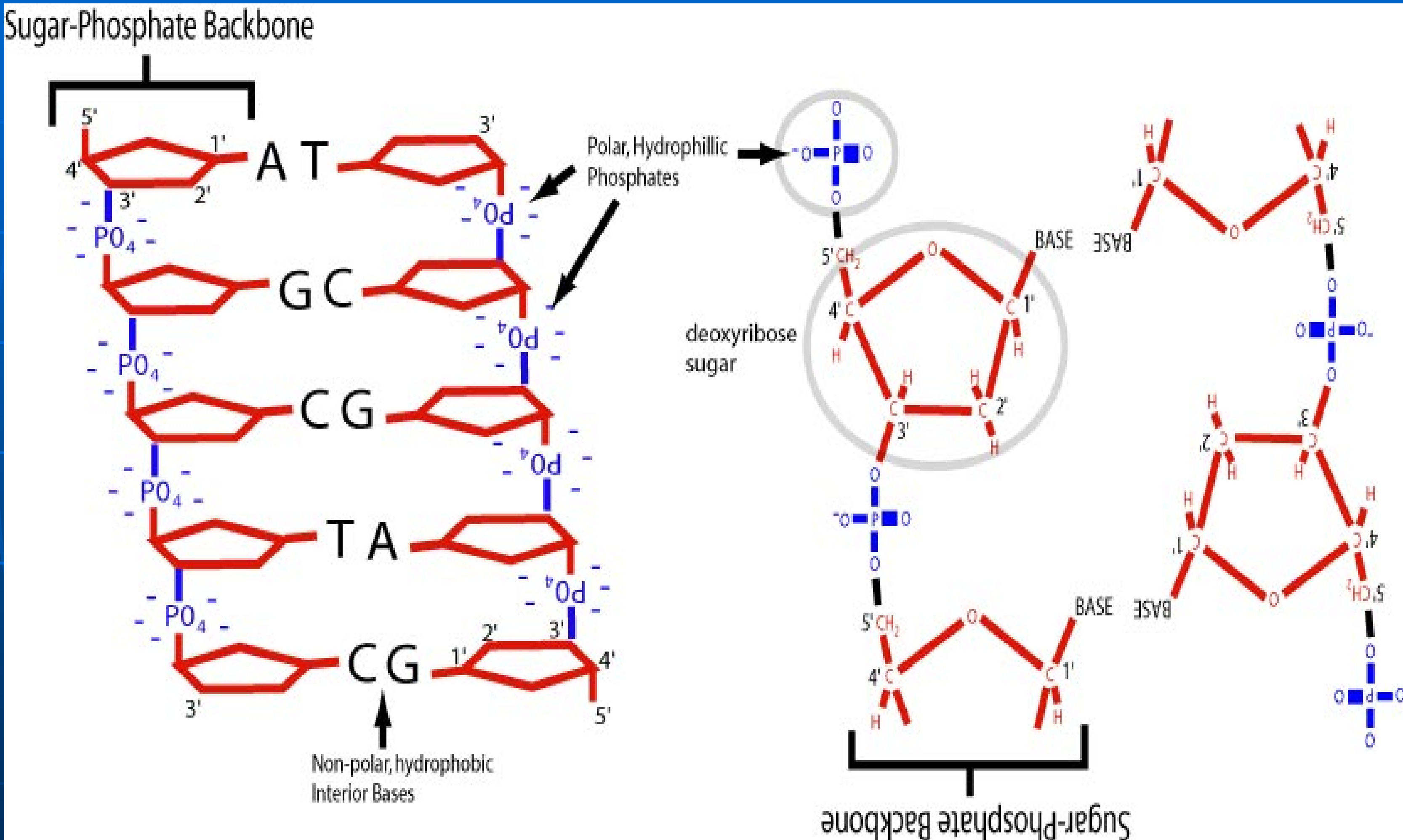


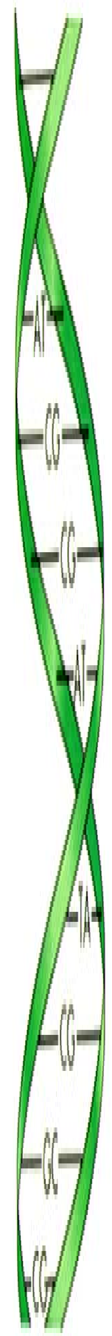
Detailed vision I



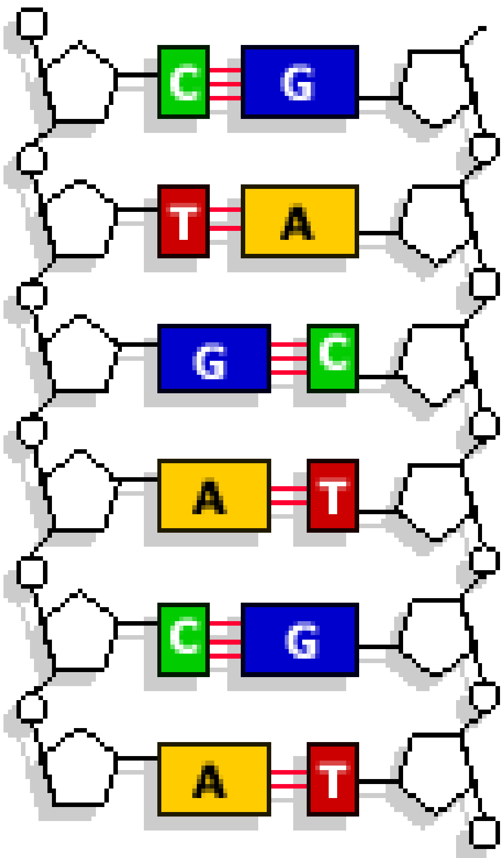


Detailed vision II





In short



- The double-stranded DNA molecule is held together by bases: Adenine (A) bonds with thymine (T); cytosine(C) bonds with guanine (G).
- These letters form the "code of life"; there are about 2,9x10^9 base pairs in the human genome wound into 24 distinct bundles, or chromosomes
- Written in the DNA are 30.000-40.000 genes which human cells use as templates to make proteins, which build and maintain our bodies

The two chief points

1. The central dogma of Genetics

2. Geometry & Topology to understand the actions of enzymes

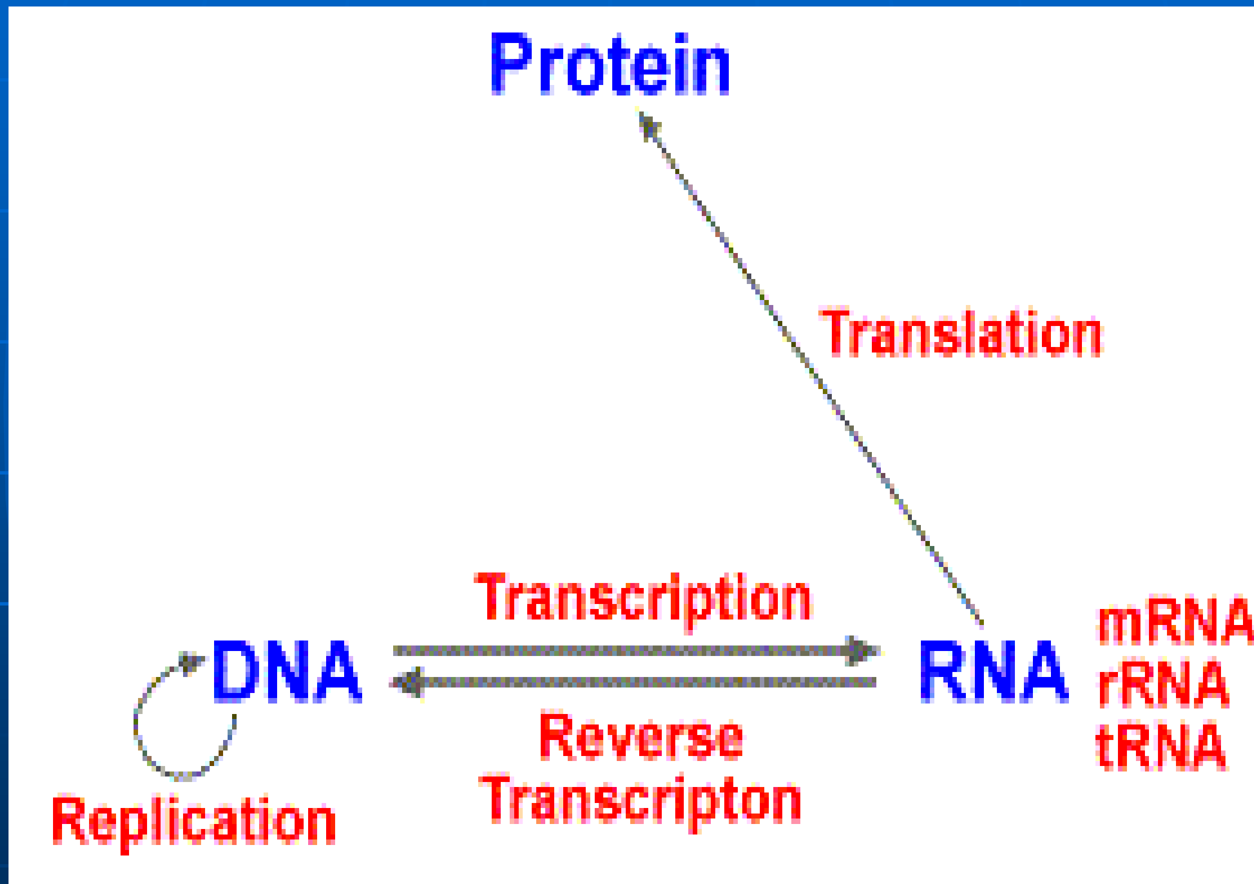
The Central Dogma

- ✓ The *central dogma of genetics* concerns the DNA-to-protein sequence involving ***transcription*** and ***translation***.

- ✓ DNA has a sequence of bases that is transcribed into a sequence of bases in mRNA (messenger RNA).

- ✓ Every three bases is a "codon" that stands for a particular amino acid. A protein is a chain of amino acids

The Central Dogma: a synopsis

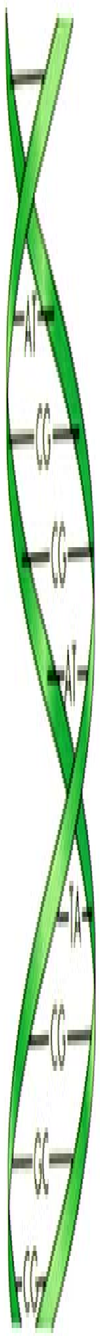


The Main Ideas in the < DNA-to-protein sequence >

1. ***DNA replication*** : DNA duplicates itself.

2. ***DNA transcription*** : DNA produces RNA (mRNA).

3. ***DNA translation*** : RNA produces proteins, the base of life.



DNA replication I

- *DNA replication* occurs during chromosome duplication.

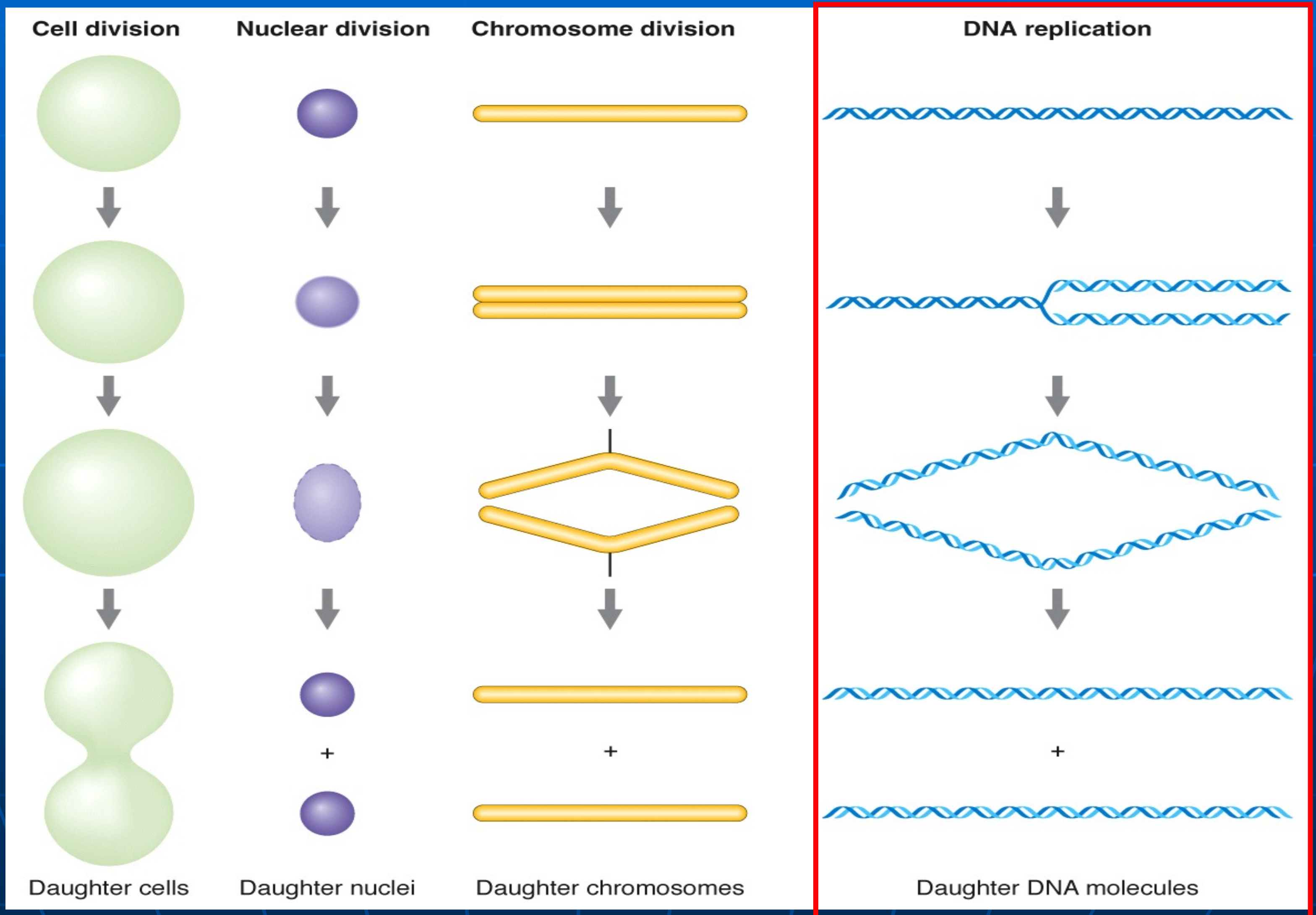
- An exact copy of the DNA is produced with the aid of *DNA polymerase.*

- Hydrogen bonds between bases break and enzymes “unzip” the molecule.

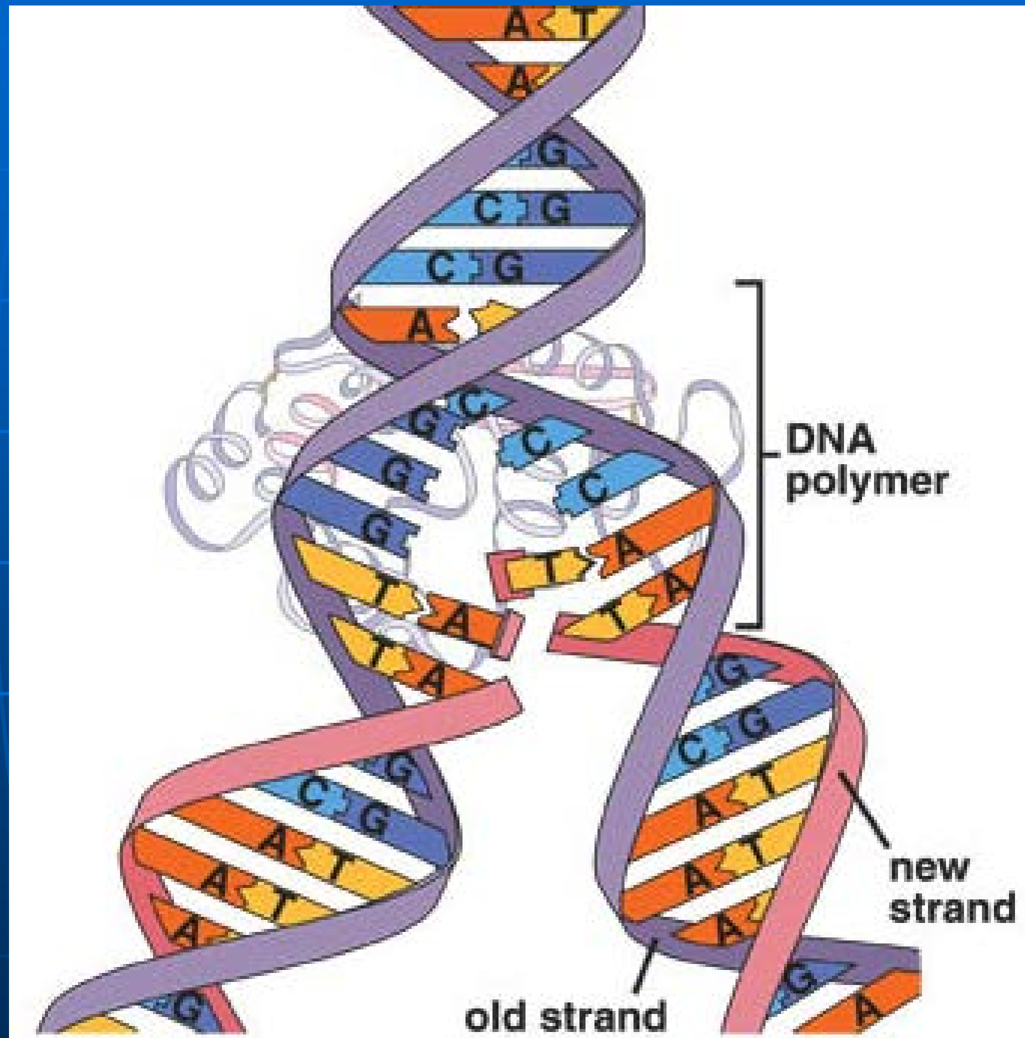
- Each old strand of nucleotides serves as a *template* for each new strand.

- New nucleotides move into complementary positions are joined by DNA polymerase.

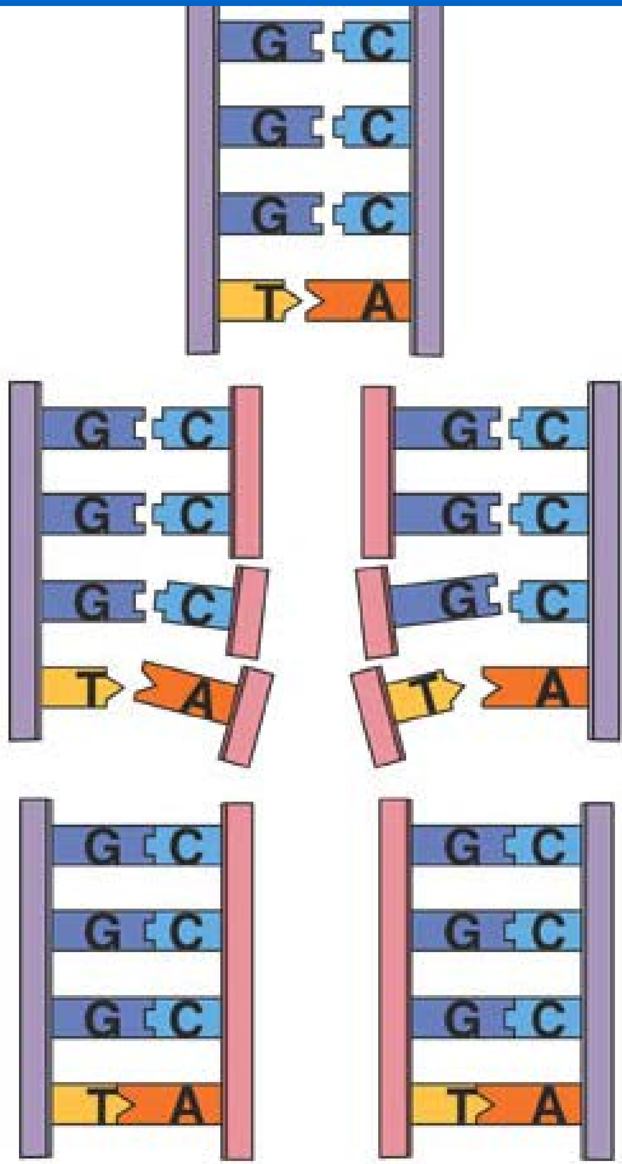
Schematic overview I



Schematic overview II



Flat schematic overview III



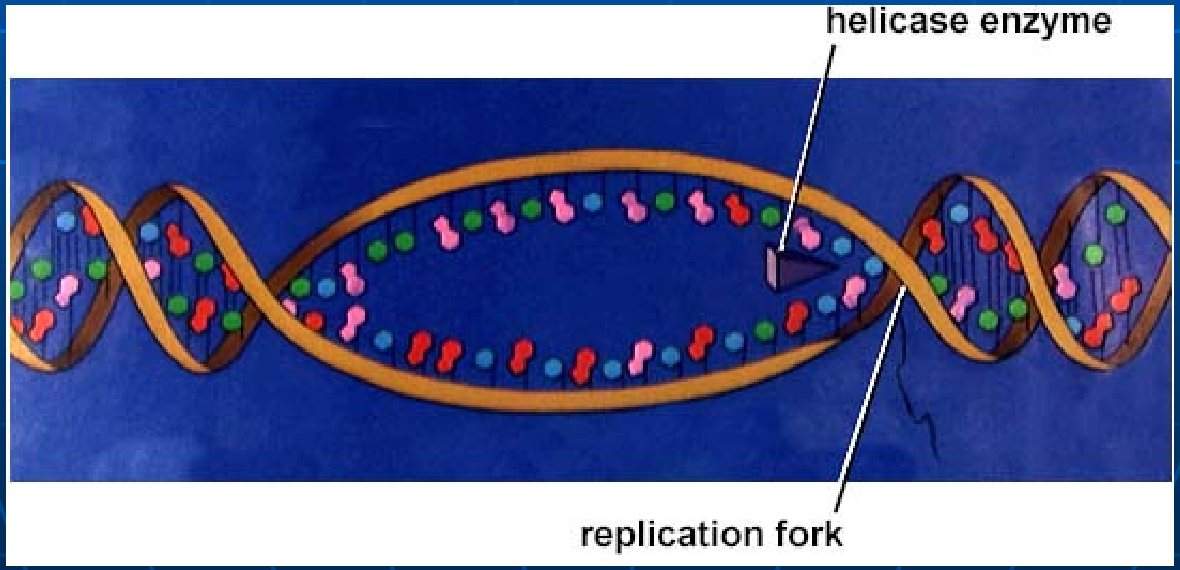
Parental DNA molecule contains so-called old strands hydrogen-bonded by complementary base pairing.

Region of replication. Parental DNA is unwound and unzipped. New nucleotides are pairing with those in old strands.

Replication is complete. Each double helix is composed of an old (parental) strand and a new (daughter) strand.

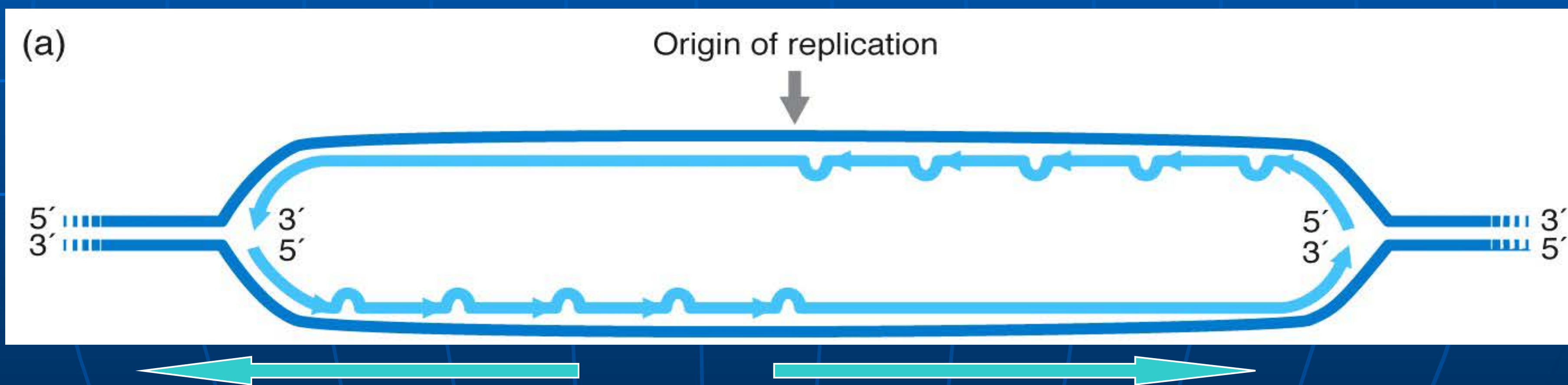
DNA Replication: a synopsis

- Replication: each DNA strand serving as template for synthesis of the complementary strand
- DNA synthesis occurs in an origin of replication and two diverging replication forks (bidirectional)
- Each replication fork contains a complex of enzymes, including DNA polymerase. Each cell contains a family of more than thirty enzymes to insure the accurate replication of DNA.
  - Other enzymes include primase, helicase, topoisomerase



Bidirectional DNA Replication

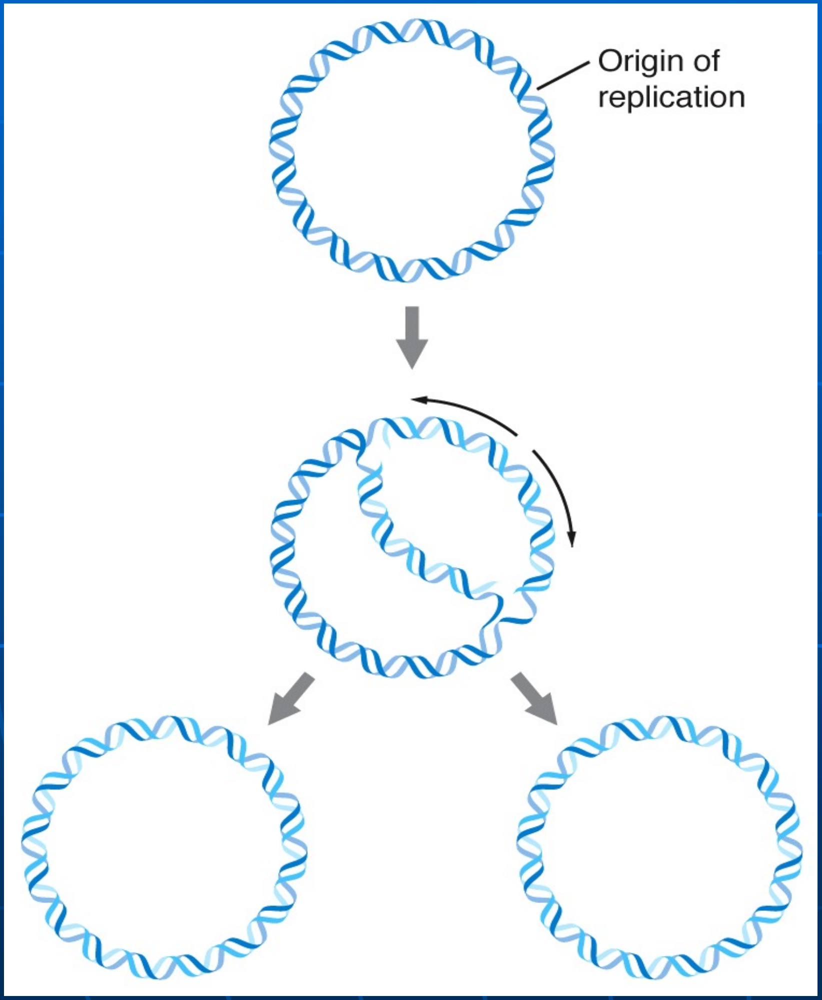
An origin of replication and two replication forks (bidirectional)



Fork movement

Fork movement

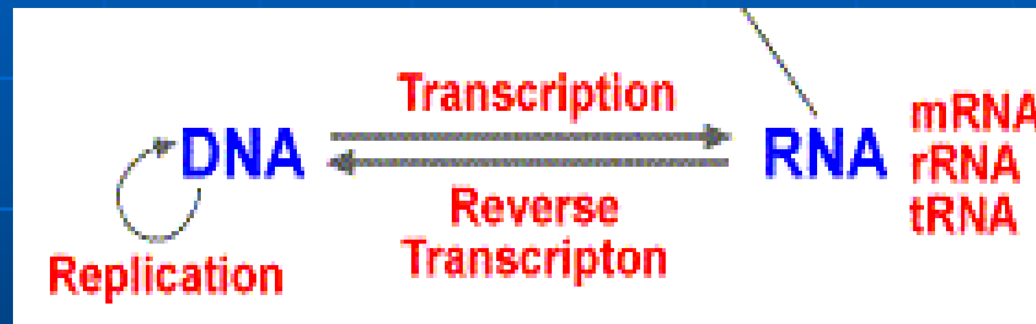
Replication of Circular Chromosome



- Replication begins from a single origin
- Replication is bi-directional
  - -intermediate is called a theta structure

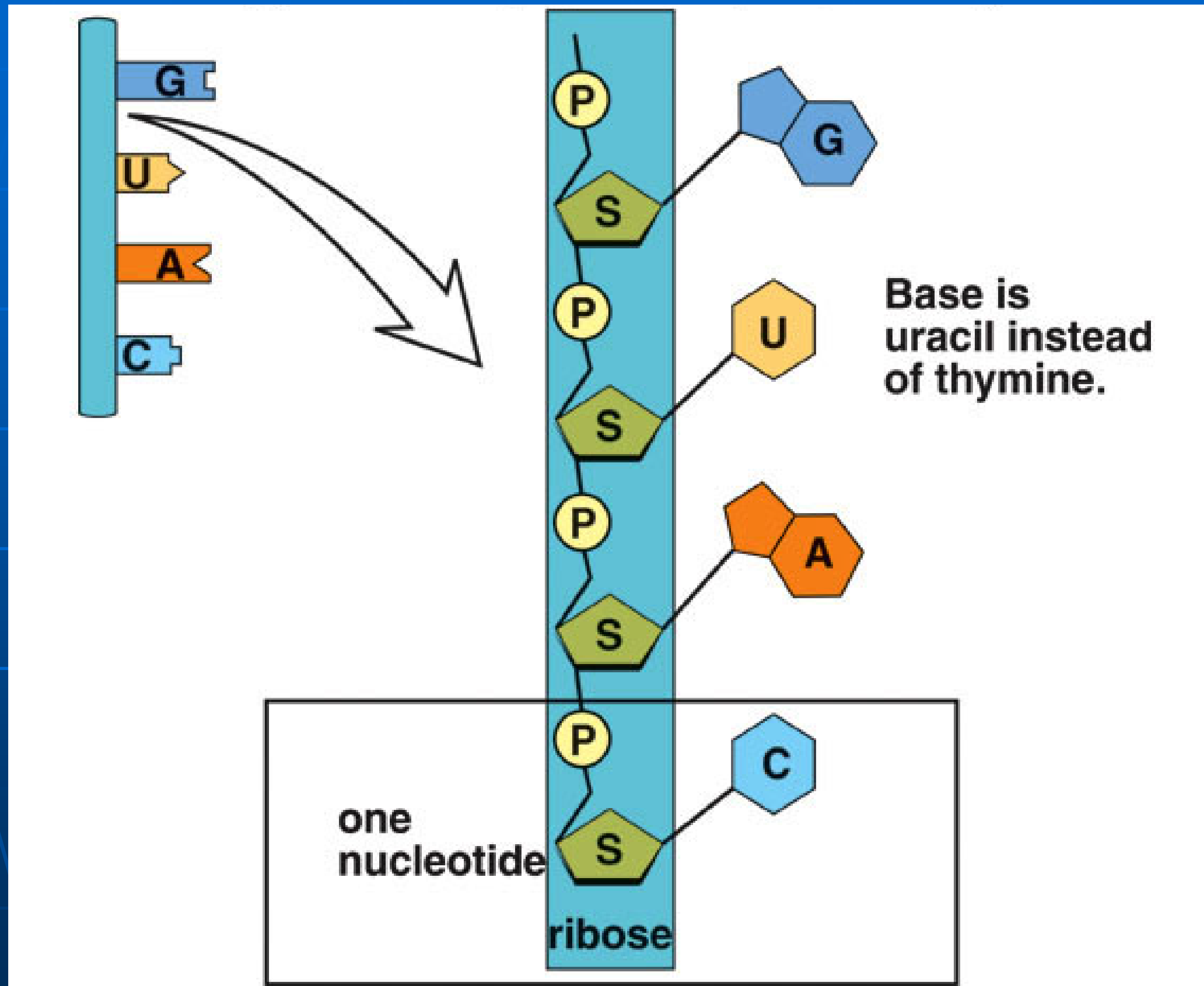
DNA Transcription

Transcription makes an RNA molecule complementary to a portion of DNA (mRNA formation).

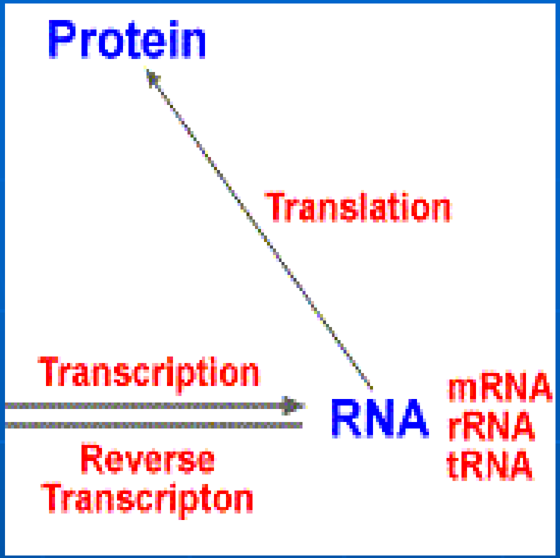


RNA polymerase joins the RNA nucleotides so that the codons in mRNA are complementary to the triplet code in DNA.

Structure of RNA



Translation



Translation is the second step by which gene expression leads to protein synthesis.

During translation, the sequence of codons in mRNA specifies the order of amino acids in a protein.

Translation requires several enzymes and two other types of RNA: *transfer RNA* and *ribosomal RNA*.

Synthesis of proteins

Two processes are involved in the synthesis of proteins in the cell:

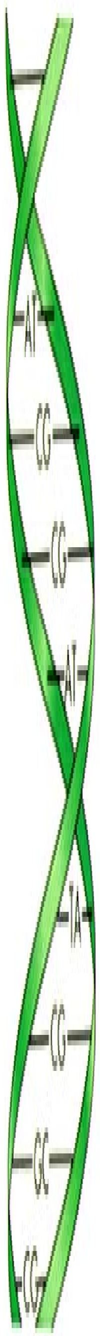
Transcription makes an RNA molecule complementary to a portion of DNA.

Translation occurs when the sequence of bases of mRNA directs the sequence of amino acids in a protein.

The genetic code

- It is the system that maps the language of DNA to that of proteins; it is indispensable for the manufacture of proteins by the cell. The genetic code is a reference system that allows a cell to make proteins by reading its DNA . It “translates” “words” formed by DNA, and constructed of 3 *nucleotides* , into amino acids . For example, the DNA word, AAT, maps to the amino acid “leucine.” The genetic code is universal: the same triplet of nucleotides corresponds to the same amino acid in all living beings.
-
- Most amino acids have more than one codon; there are 20 amino acids with a possible 4^3=64 different triplets.

- The code is nearly universal among living organisms.



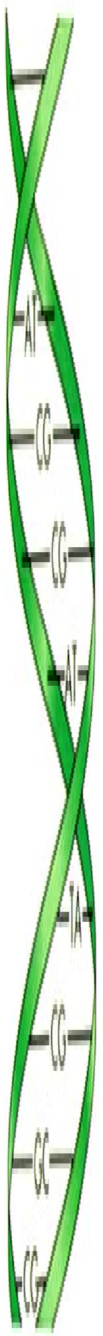
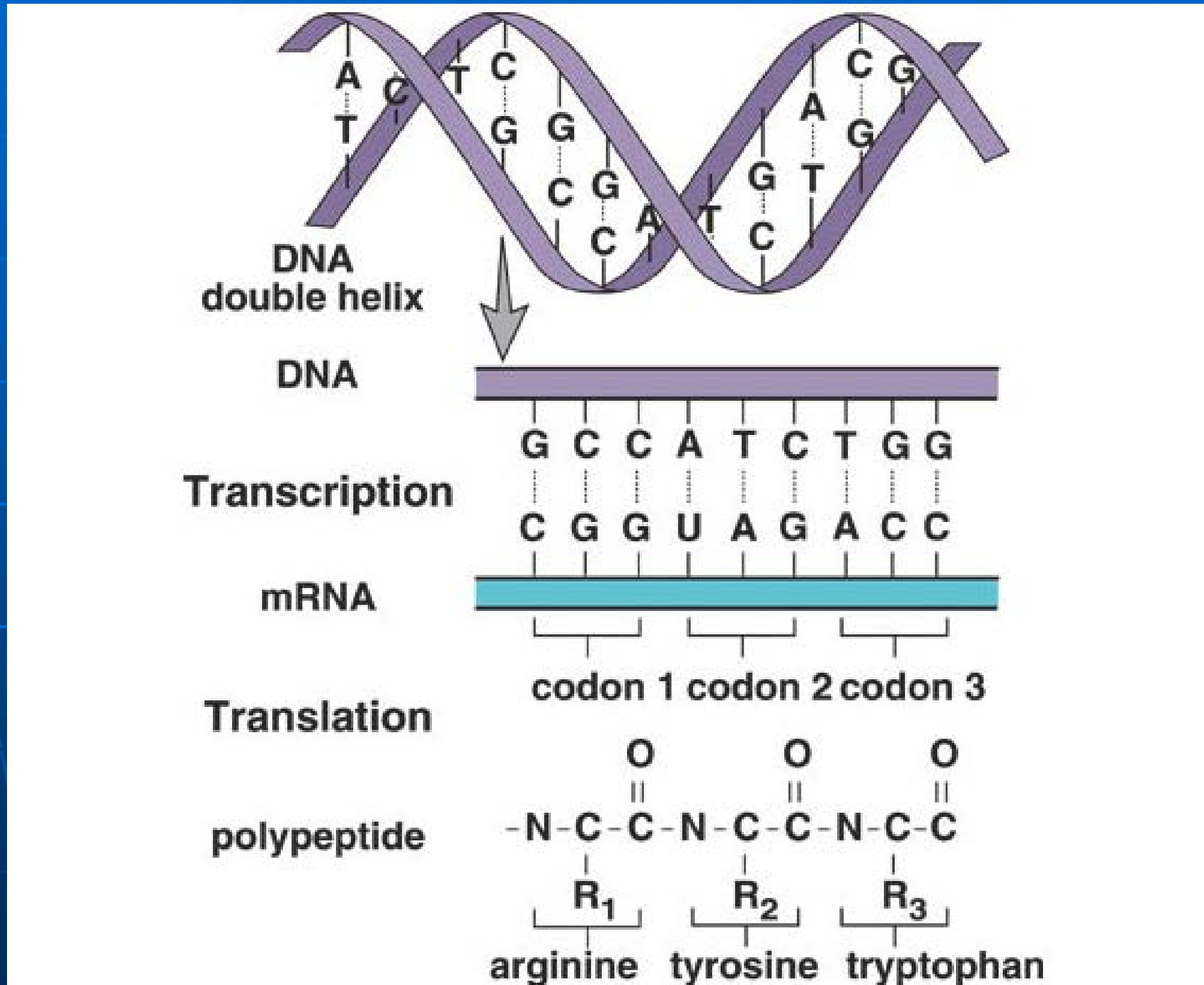


Table of codons and amino acids

First Base	Second Base: U	Second Base: C	Second Base: A	Second Base: G	Third Base
U	UUU phenylalanine	UCU serine	UAU tyrosine	UGU cysteine	U
U	UUC phenylalanine	UCC serine	UAC tyrosine	UGC cysteine	C
U	UUA leucine	UCA serine	UAA stop	UGA *stop*	A
U	UUG leucine	UCG serine	UAG stop	UGG tryptophan	G
C	CUU leucine	CCU proline	CAU histidine	CGU arginine	U
C	CUC leucine	CCC proline	CAC histidine	CGC arginine	C
C	CUA leucine	CCA proline	CAA glutamine	CGA arginine	A
C	CUG leucine	CCG proline	CAG glutamine	CGG arginine	G
A	AUU isoleucine	ACU threonine	AAU asparagine	AGU serine	U
A	AUC isoleucine	ACC threonine	AAC asparagine	AGC serine	C
A	AUA isoleucine	ACA threonine	AAA lysine	AGA arginine	A
A	AUG (*start*) methionine	ACG threonine	AAG lysine	AGG arginine	G
G	GUU valine	GCU alanine	GAU aspartate	GGU glycine	U
G	GUC valine	GCC alanine	GAC aspartate	GGC glycine	C
G	GUA valine	GCA alanine	GAA glutamate	GGA glycine	A
G	GUG valine	GCG alanine	GAG glutamate	GGG glycine	G

Overview of gene expression



What kind of curiosity we had ?

The Calugareanu-White-Fuller formula

$$L = T + W$$

which gives the amount of coiling and supercoiling of circular DNA.

We wish to know its implications in the dynamics involving proteins, i. e., in any biological function.

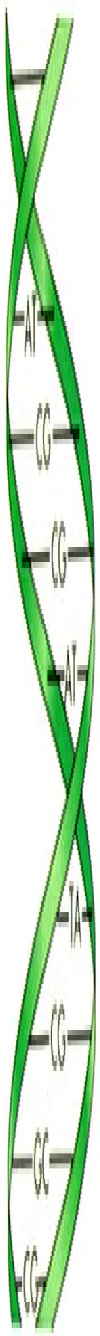
Biological function ?

Main laws of biochemistry

- Every process is catalyzed by an enzyme
- Every enzyme is encoded by a (set of) genes

Therefore:

Genes → enzymes → processes → life



In short

Life = f(processes(enzymes(genes)))

Life depends on processes,
which depend on enzymes,
which depend on genes

Therefore, we want to understand how processes depend on enzymes and why

Geometry and Topology becomes a very useful tool to understand current life and Nature !

$L = T + W$

- Linking number
- How two links are connected in a chain
- Always an integer number
- Topology

L = T + W

- Twisting number
- The number of times that the two strands are twisted about each other
- Usually not integer
- Geometry

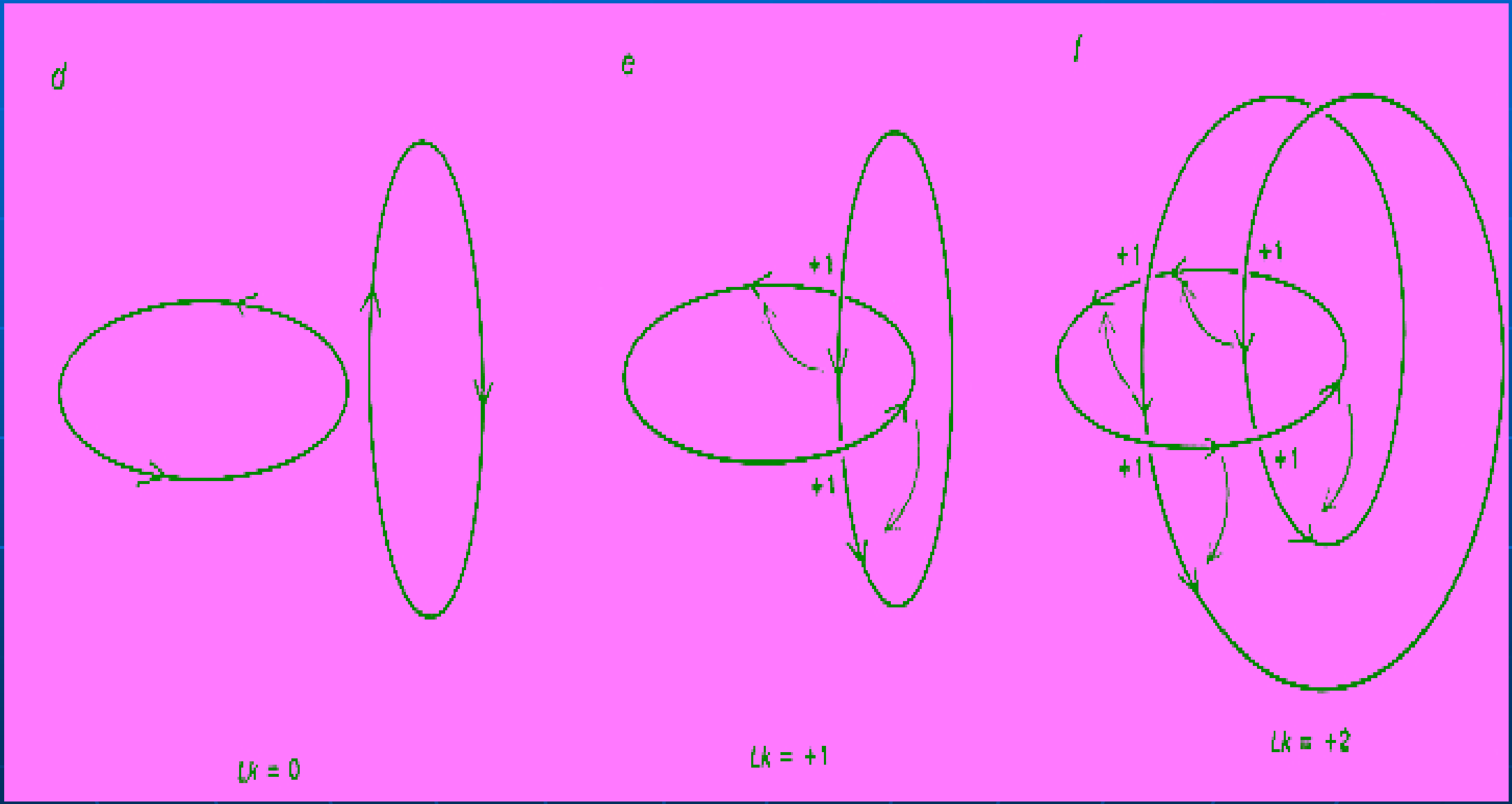
$L = T + W$

- Writing number
- The number of times that the the DNA helix is coiled about itself in 3-space
- How flat (W=0) is the strand
- Usually not integer
- Geometry

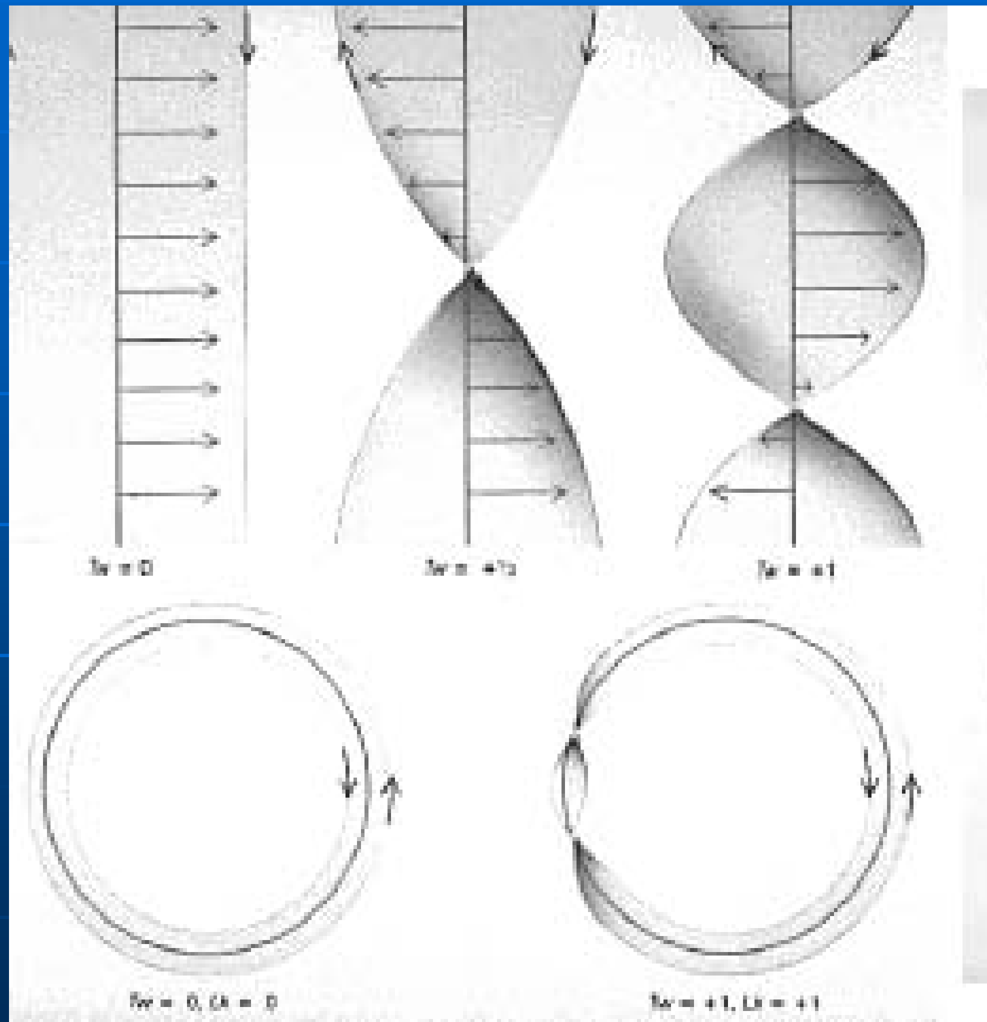
In short

The formula-theorem states that the topological invariant L is the sum of two other terms whose proper definitions will be given later and which individually depend on Geometry rather than Topology: the twist T, which is a measure of how much the ribbon is twisted about its own axis, and the writhe W, which is a measure of non-planarity (and non-sphericity) of the axis curve.

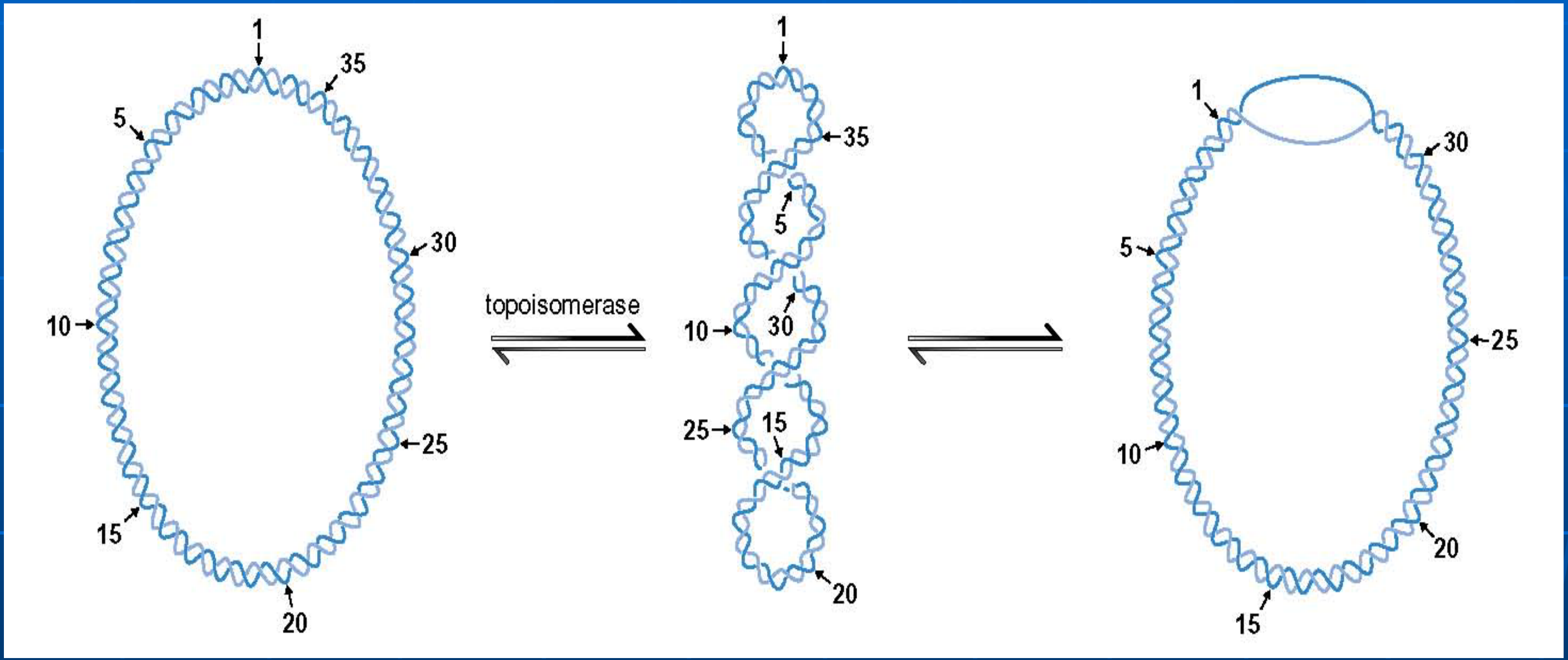
Some examples: L



Some examples: T

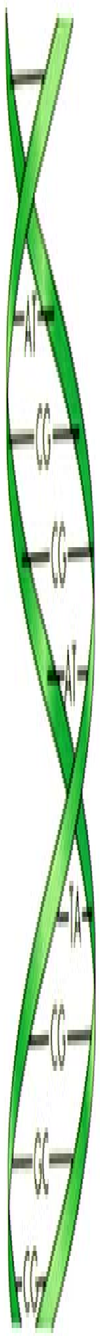


Some examples: W



Importance of DNA Geom & Topo

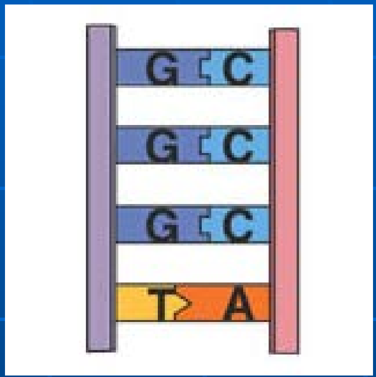
- The Geometry & Topology (3-dimensional arran-gement) of DNA becomes important every time DNA has to:
  - Replicate during cell division
  - Be transcribed
  - Be packaged into cell
  - Be repaired if mutated
- These topics will be discussed tomorrow by Prof Pastor



Summarizing I

- DNA can be viewed as two very long curves that are intertwined millions of times, linked to other curves, and subjected to four or five successive orders of coiling to convert it into a compact form for information storage.

- A ladder whose sides are the backbones and whose rungs are hydrogen bonds is formed by hydrogen bonding between base pairs, A only with T, and C only with G.



- In the classic Crick-Watson double helix model for DNA, the ladder is twisted in a right hand helical fashion, with an average and nearly constant pitch of approximately 10.5 base pairs per full helical twist. The local helical pitch of duplex DNA is a function of both the local base pair sequence and the cellular environment in which the DNA lives.

Summarizing II

The packing, twisting, and topological constraints all taken together mean that topological entanglement poses serious functional problems for DNA. This entanglement would interfere with processes of replication and transcription.

During replication, the daughter DNA molecules become entangled and must be disentangled to complete the replication. After the process is finished, the original DNA conformation must be restored. Some enzymes maintain proper geometry and topology by passing one strand of DNA through another. Other enzymes break the DNA apart and recombine the ends.

The description of the 3-dimensional structure of DNA and the changes in DNA structure due to the action of these enzymes have required the serious use of geometry and topology. This use of mathematics as an analytic tool is especially important because there is no experimental way to observe the dynamics of enzymatic action directly.

Summarizing III

In the experimental study of DNA structure and enzyme mechanism, biologists developed the topological approach to enzymology. The DNA knots and links of circular DNA are observed by gel electrophoresis and electron microscopy. The enzyme mechanism can be described by observing the changes in geometry (supercoiling) and topology (knotting and linking) in DNA caused by an enzyme.

The topological approach to enzymology poses an interesting challenge for mathematics: from the observed changes in DNA geometry and topology, how can one deduce enzyme mechanisms? This requires the construction of mathematical models for enzyme action and the use of these models to analyze the results of topological enzymology experiments. The entangled form of the product DNA knots and links contains information about the enzymes that made them.

That's all folks !

Thanks for your attention

Appendix I

Gene = region of DNA that determines the structure of a protein
= unit of heredity in living organisms

Section of DNA that represents the plan for the construction of a protein. A gene is a sequence of DNA , contained within the nucleus of our cells , that carries the 'instructions' for the manufacture of a protein . The information they hold describes the myriad characteristics of an individual (e.g. eye colour). Humans beings have approximately 30 000 genes, which amounts to only around 5% of their total DNA. Some animal and plant species have more genes than humans.

Protein = (from greek protos, of primary importance)
= chain of amino acids

Macromolecules present in all living beings, essential for the life of the cell, as well as the entire organism. Their functions are greatly varied. A protein is a macromolecule present in all living beings. Indispensable for the life of the cell and the entire organism, proteins are constructed by our cells through DNA and the genetic code . From those that make up our hair (keratin) to those that defend us against microbes (antibodies), proteins have a wide variety of functions and responsibilities.

Appendix II

• **Genetic code** = System that maps the language of DNA to that of proteins; it is indispensable for the manufacture of proteins by the cell. The *genetic code* is a reference system that allows a cell to make proteins by reading its DNA . It "translates" "words" formed by DNA, and constructed of 3 *nucleotides* , into amino acids . For example, the DNA word, AAT, maps to the amino acid "leucine." The *genetic code* is universal: the same triplet of nucleotides corresponds to the same amino acid in all living beings.

• **Nucleotide** = Basic unit of nucleic acids, made of a sugar (deoxyribose (DNA) or ribose (RNA)), a phosphate group, and a base: A, G, C or T (DNA); A, G, C or U (RNA)