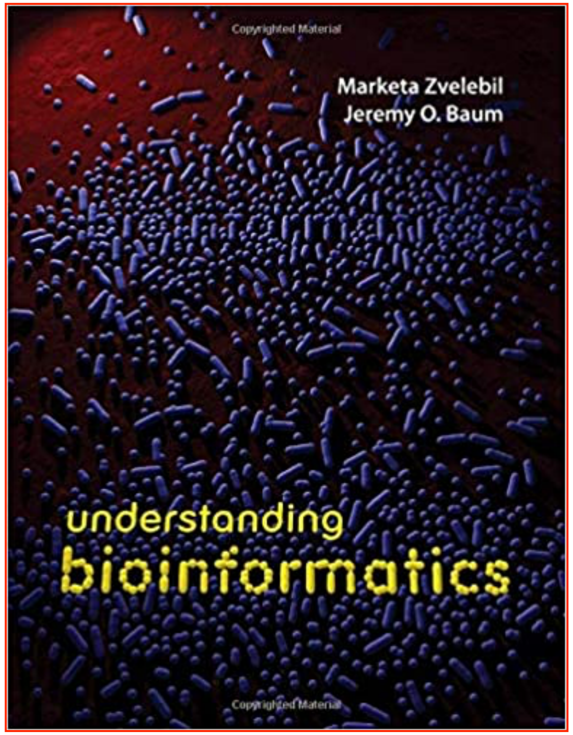
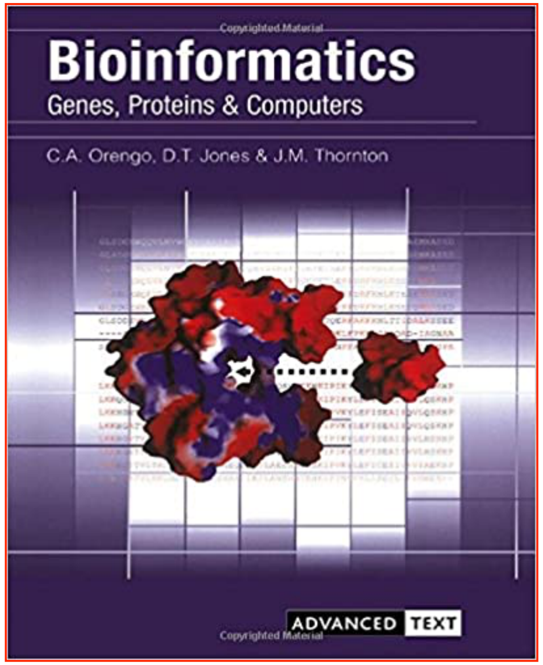
Bioinformatics

Prof Romina Oliva

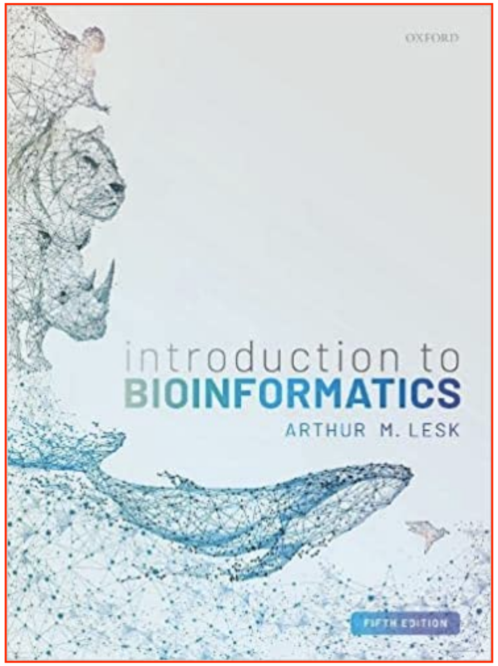
romina.oliva@uniparthenope.it



M.J. Zvelebil & J.O. Baum
Understanding Bioinformatics
Garland Science,
Taylor & Francis Group



C.A. Orengo, D.T. Jones,
J.M. Thornton
Bioinformatics
CRC Press,
Taylor & Francis Group



A.M. Lesk
Introduction to Bioinformatics
Oxford University Press

Lezioni *on-line*

https://elearning.uniparthenope.it

Lessons 1&2. Content

1. Introduction to bioinformatics

2. DNA: sequence, structure, replication and translation

3. Genomes: evolution and information

Bioinformatics: collection, storage, organization and interpretation of large-scale biological data

Janet M. Thornton

Bioinformatics is a way of understanding biology using the power of computers and data

The informatics component is about organising and handling datasets, which requires a lot of technology

Once you capture the data, you need to understand what it means

Janet M. Thornton

Bioinformatics is a multidisciplinary subject, whose main components are:

biology B

informatics I

statistics S

Bioinformatics is what bioinformatics does

Eric C. Snowdeal III

At the end of the course, you should be able to say:

What issues are of competence of bioinformatics

What issues bioinformatics can address nowadays, with what reliability rate

The main challenges for the near future

Ideas/suggestions for applications in your fields of interest?...

B

Life processes are due to the concerted action of biological (macro)molecules, mainly proteins

Instructions for the protein synthesis are contained in the genomes (DNA).

The main role of DNA, from unicellular bacteria to multicellular plants and animals, is information storage

All the information required to make and maintain an organism is stored in its DNA

Information about nucleic acids and proteins is the raw material of bioinformatics

BIOinformatics = genes + proteins + informatics

(part of computational biology, biocomputing)

GENE: DNA segment which codes for a specific protein and determines an hereditary feature

PROTEIN: expression product of a gene and EFFECTOR of the biochemical function whose information is stored in the gene

RNA: fundamental role in the gene expression regulation

A little history…

1951	Pauling: alfa e beta
1953	DNA double helix
1955	Insulin sequence
1959	Myoglobin 3D structure
1960	Anfinsen
1967	Dayoff collection
1968	PAM
1970	Nedleman and Wunsch
1977	PDB
1977	Chou and Fasman
1977	DNA sequencing (Sanger)
1980	Wutrich
1981	Greer
1985	FASTP
1986	Chothia and Lesk
1990	Blast
1991	Fold recognition
1993	PHD
1994 2001	CASP/CAPRI

A little history…

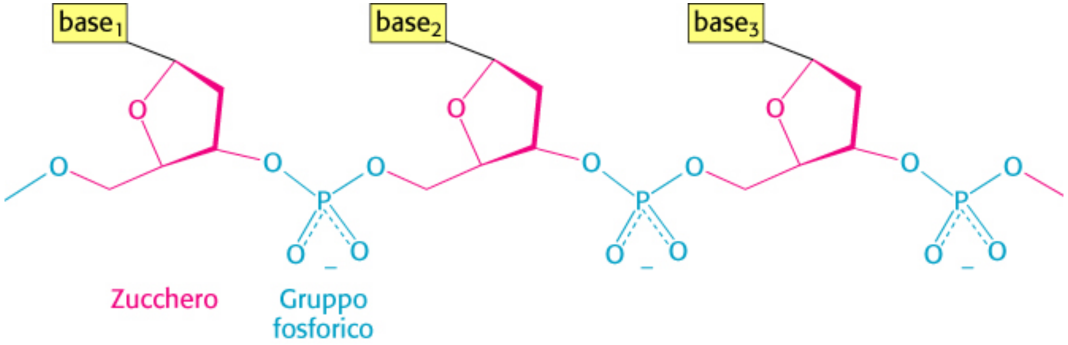
1951	Pauling: alfa e beta
1953	DNA double helix
1955	Insulin sequence
1959	Myoglobin 3D structure
1960	Anfinsen
1967	Dayoff collection
1968	PAM
1970	Nedleman and Wunsch
1977	PDB
1977	Chou and Fasman
1977	DNA sequencing (Sanger)
1980	Wutrich
1981	Greer
1985	FASTP
1986	Chothia and Lesk
1990	Blast
1991	Fold recognition
1993	PHD
1994 2001	CASP/CAPRI
2007	Next Generation Sequencing
2020	AlphaFold

A little history…

1951	Pauling: alfa e beta
1953	DNA double helix
1955	Insulin sequence
1959	Myoglobin 3D structure
1960	Anfinsen
1967	Dayoff collection
1968	PAM
1970	Nedleman and Wunsch
1977	PDB
1977	Chou and Fasman
1977	DNA sequencing (Sanger)
1980	Wutrich
1981	Greer
1985	FASTP
1986	Chothia and Lesk
1990	Blast
1991	Fold recognition
1993	PHD
1994 2001	CASP/CAPRI
2007	Next Generation Sequencing
2020	AlphaFold

DNA, Deoxy-riboNucleic Acid

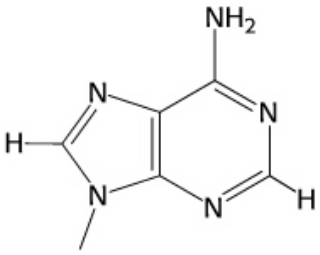
B



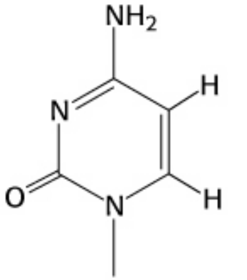
The DNA **backbone** is made of alternate phosphate groups and sugars (deoxyriboses)

purine

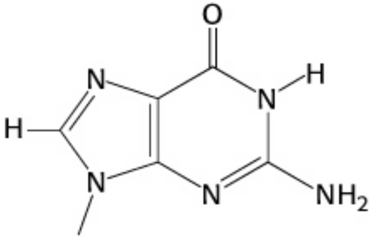
pyrimidine



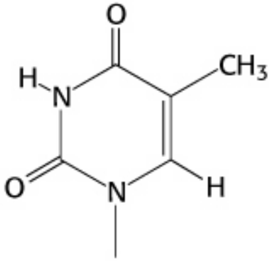
Adenina (A)



Citosina (C)



Guanina (G)

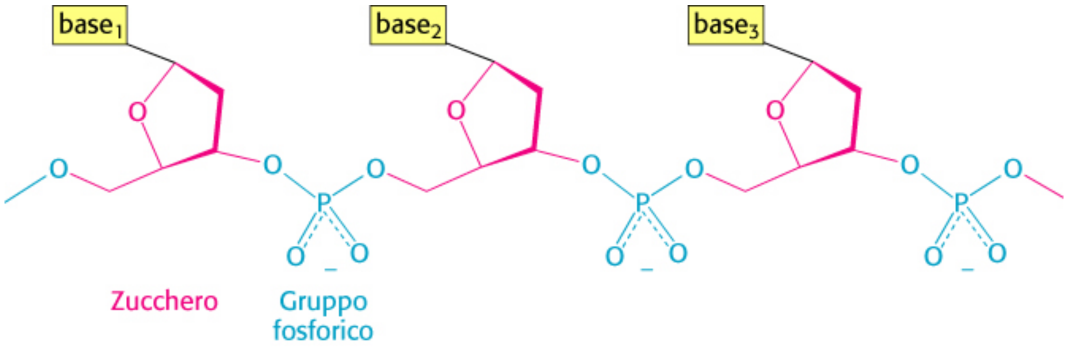


Timina (T)

One of **4 nitrogen-containing bases**: Adenine (**A**), Cytosine (**C**), Guanine (**G**) e Thymine (**T**) is linked to each sugar at the C1′ position

DNA, Deoxy-riboNucleic Acid

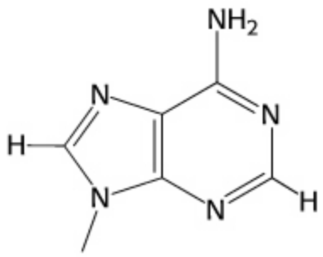
B



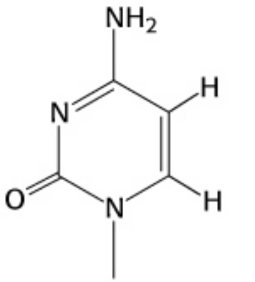
The DNA **backbone** is made of alternate phosphate groups and sugars (deoxyriboses)

One of **4 nitrogen-containing bases**: Adenine (**A**), Cytosine (**C**), Guanine (**G**) e Thymine (**T**) is linked to each sugar at the C1′ position

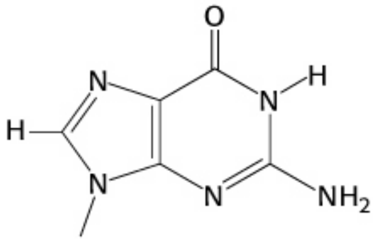
purine pyrimidine



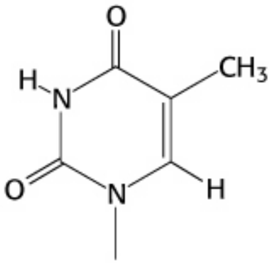
Adenina (A)



Citosina (C)



Guanina (G)



Timina (T)

DNA sequences ONLY differ in the combination of the same 4 nucleobases: **A, C, G, T**

Example of DNA sequence

```
CCAGCTGGGTGAGCCTGGGAGGGAGGAGGTGAGTTGGGCTGGACTCAGGGACCGACTCTT

CCCGTCTCATGACTGTGTTTACTGGGCTGGATTTTGGGAAGGGGCCAGATTGCATCAGAC

AGGGCCTGATGGGCTGGAGCCAGACTGTGGTCTGAGGAGGAGACACAGCCTTATAAGCTG

AGGGAGTGGAGAGGCCCGGGGCCAGGAAAGCAGAGACAGACAAAGCGTTAGGAGAAGAAG

AGAGGCAGGGAAGACAAGCCAGGCACGATGGCCACCTTCCCACCAGCAACCAGCGCCCCC

CAGCAGCCCCCAGGCCCGGAGGACGAGGACTCCAGCCTGGATGAATCTGACCTCTATAGC

TTCATCCTTGTTAAA
```

DNA sequences ONLY differ in the combination of the same 4 nucleobases: **A, C, G, T**

Example of DNA sequence

```
CCAGCTGGGTGAGCCTGGGAGGGAGGAGGTGAGTTGGGCTGGACTCAGGGACCGACTCTT

CCCGTCTCATGACTGTGTTTACTGGGCTGGATTTTGGGAAGGGGCCAGATTGCATCAGAC

AGGGCCTGATGGGCTGGAGCCAGACTGTGGTCTGAGGAGGAGACACAGCCTTATAAGCTG

AGGGAGTGGAGAGGCCCGGGGCCAGGAAAGCAGAGACAGACAAAGCGTTAGGAGAAGAAG

AGAGGCAGGGAAGACAAGCCAGGCACGATGGCCACCTTCCCACCAGCAACCAGCGCCCCC

CAGCAGCCCCCAGGCCCGGAGGACGAGGACTCCAGCCTGGATGAATCTGACCTCTATAGC

TTCATCCTTGTTAAA
```

DNA sequences are written in
a 4-letter alphabet

DNA sequences ONLY differ in the
combination of the same 4
nucleobases: **A, C, G, T**

Example of DNA sequence

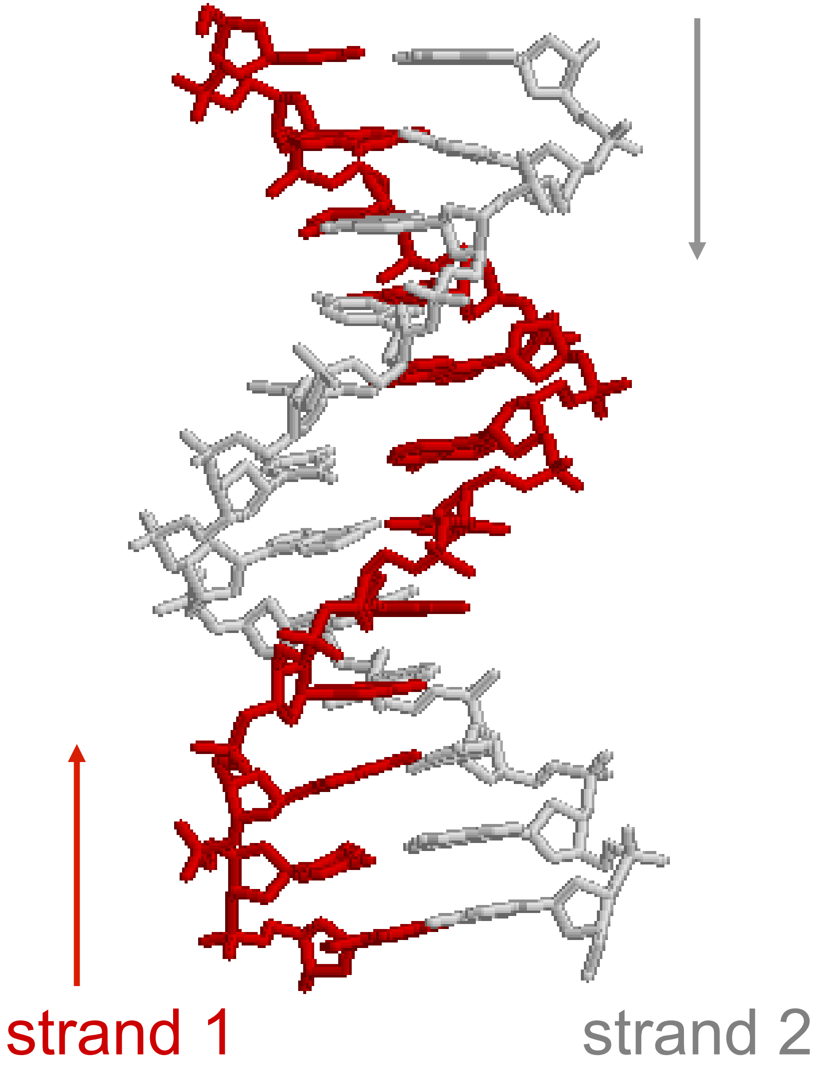
```
CCAGCTGGGTGAGCCTGGGAGGGAGGAGGTGAGTTGGGCTGGACTCAGGGACCGACTCTT
CCCGTCTCATGACTGTGTTTACTGGGCTGGATTTTGGGAAGGGGCCAGATTGCATCAGAC
AGGGCCTGATGGGCTGGAGCCAGACTGTGGTCTGAGGAGGAGACACAGCCTTATAAGCTG
AGGGAGTGGAGAGGCCCGGGGCCAGGAAAGCAGAGACAGACAAAGCGTTAGGAGAAGAAG
AGAGGCAGGGAAGACAAGCCAGGCACGATGGCCACCTTCCCACCAGCAACCAGCGCCCCC
CAGCAGCCCCCAGGCCCGGAGGACGAGGACTCCAGCCTGGATGAATCTGACCTCTATAGC
TTCATCCTTGTTAAA
```

Genetic information is stored in the nucleobases sequence of a DNA chain

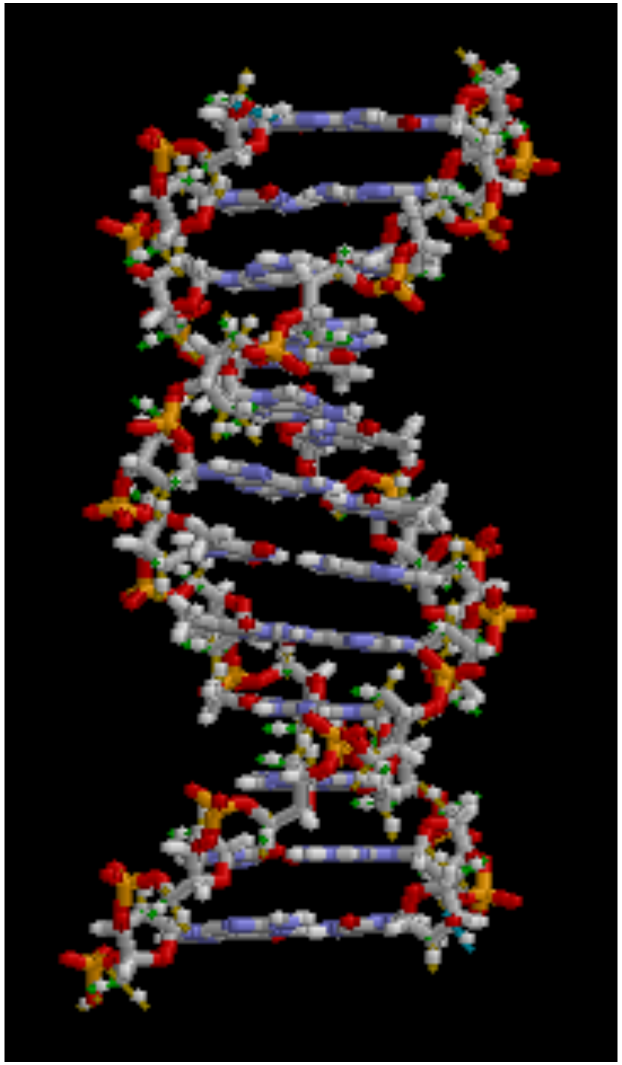
DNA sequences are written in a 4-letter alphabet

DNA sequences ONLY differ in the combination of the same 4 nucleobases: **A, C, G, T**

DNA: Double helix structure, Watson & Crick 1953

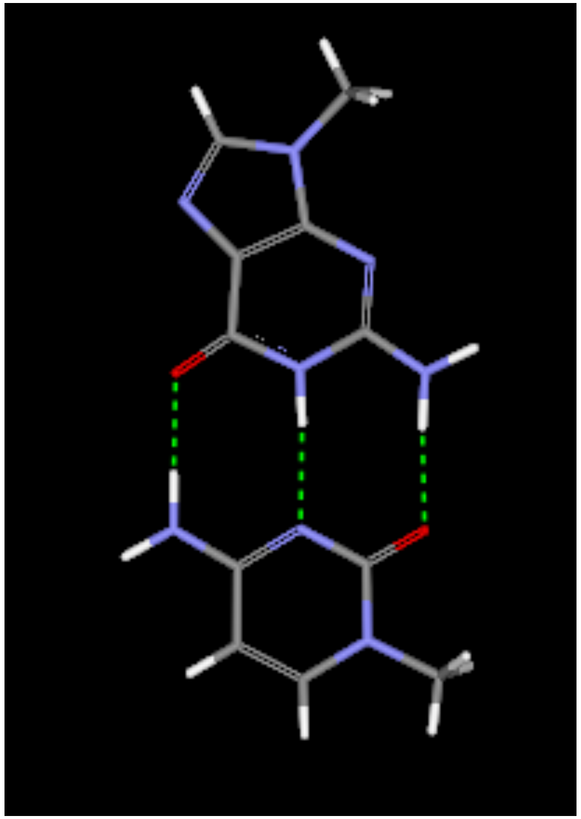


DNA: Double helix structure, Watson & Crick 1953

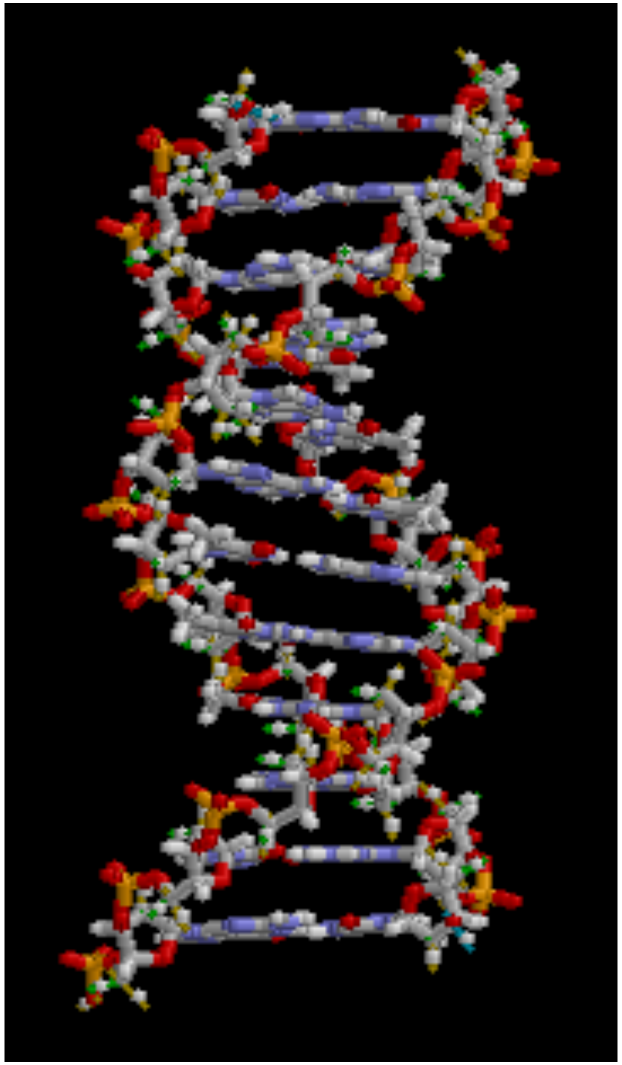


DNA: Double helix structure, Watson & Crick 1953

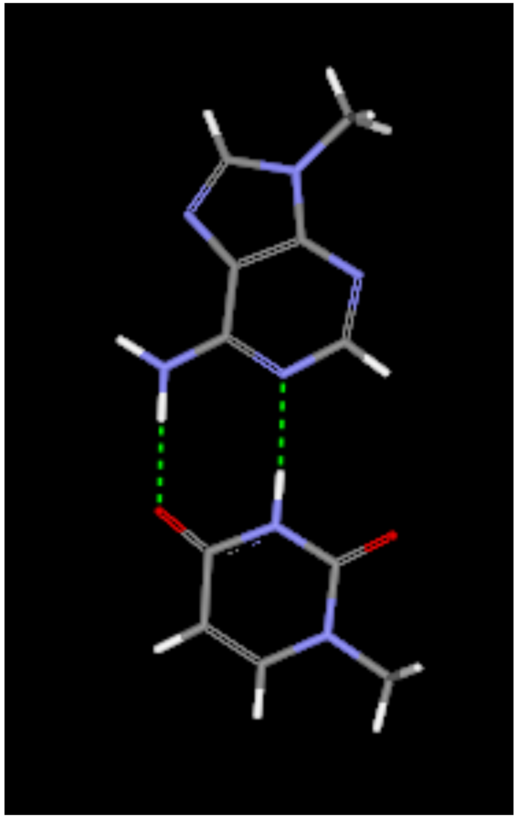
Guanine



Cytosine



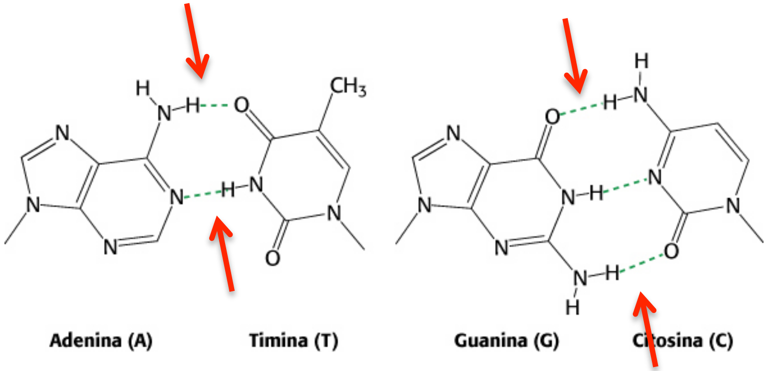
Adenine



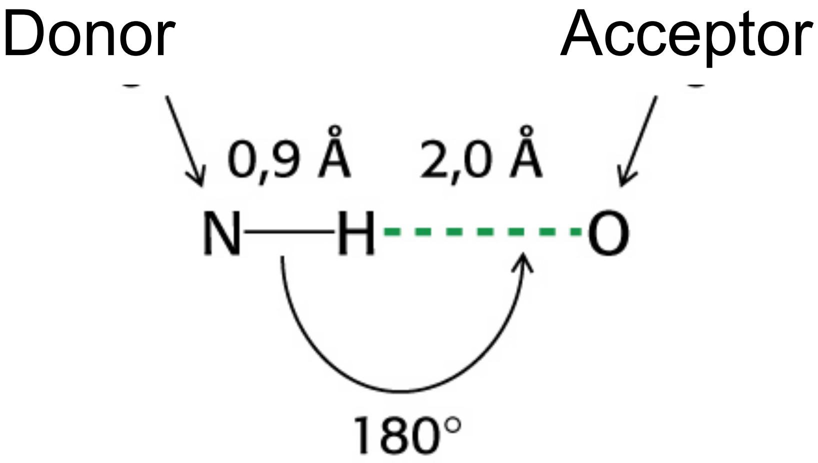
Thymine

G-C Watson & Crick base pairing A-T

What is a hydrogen(H)-bond ?



In a covalent bond two atoms (in the same molecule) share their valence electrons



Donor

Acceptor

0,9 Å

2,0 Å

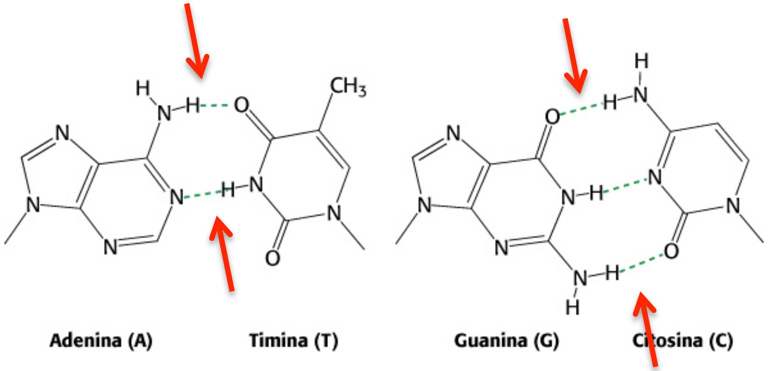
N—H- - - - - - - -O

180°

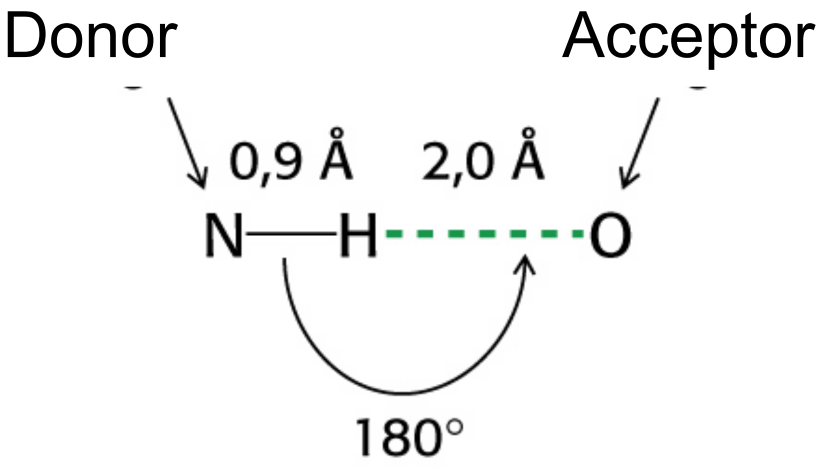
A H-bond is roughly 20-fold weaker than a covalent bond, energy of ≈2-5 kcal/mol

In a hydrogen(H)-bond, an hydrogen is covalently bonded to an electronegative atom (the donor) and interacts electrostatically with another electronegative atom (the acceptor)

What is a hydrogen(H)-bond ?



H-bonds are the strongest inter-molecular interactions



However, they can be broken when needed, e.g. during the DNA replication process

A H-bond is roughly 20-fold weaker than a covalent bond, energy of ≈2-5 kcal/mol

The key for copying the genetic material !

The key for copying the genetic material !

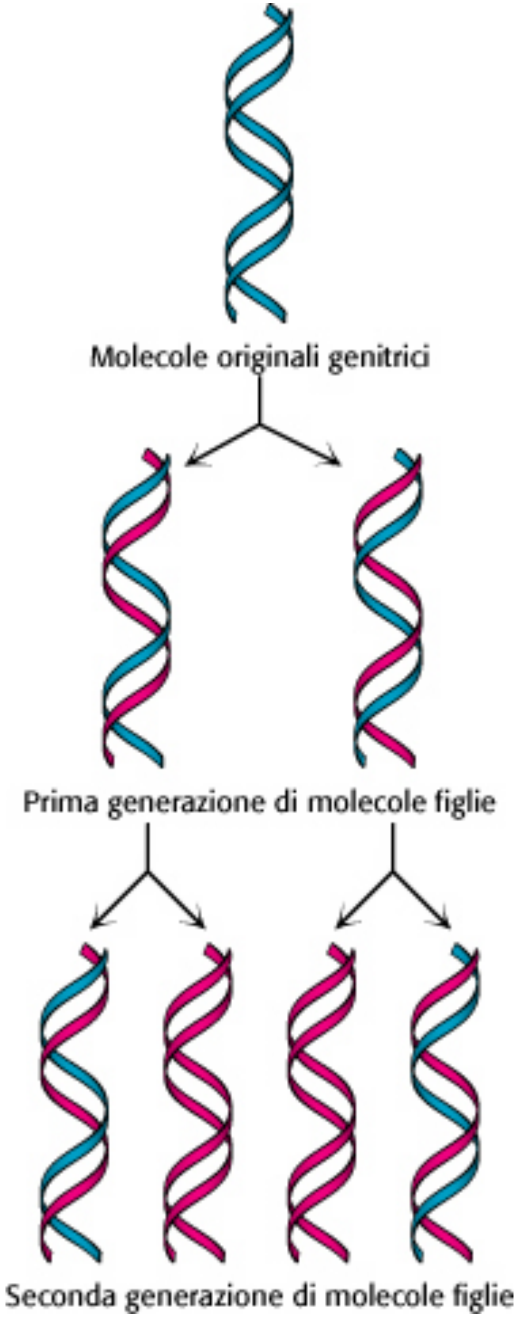
“It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.”

Watson J.D. & Crick F.H.C., Nature Vol. 171 (1953)

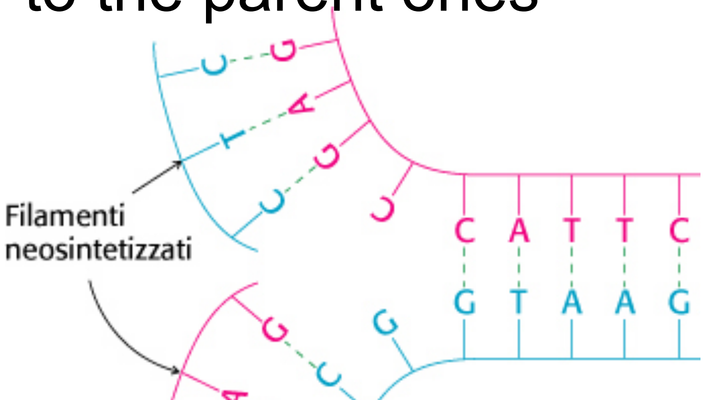
“*But what is all this ignorance compared to the perplexity in which phenomena such as the amazing memory that would be the hereditary transmission of acquired qualities put us? The impossibility, or only the suspicion of being able to conceive a mechanical explanation of such performances of the cellular substance is unbridgeable*”

Thomas Mann, “Enchanted mountain”, 1924

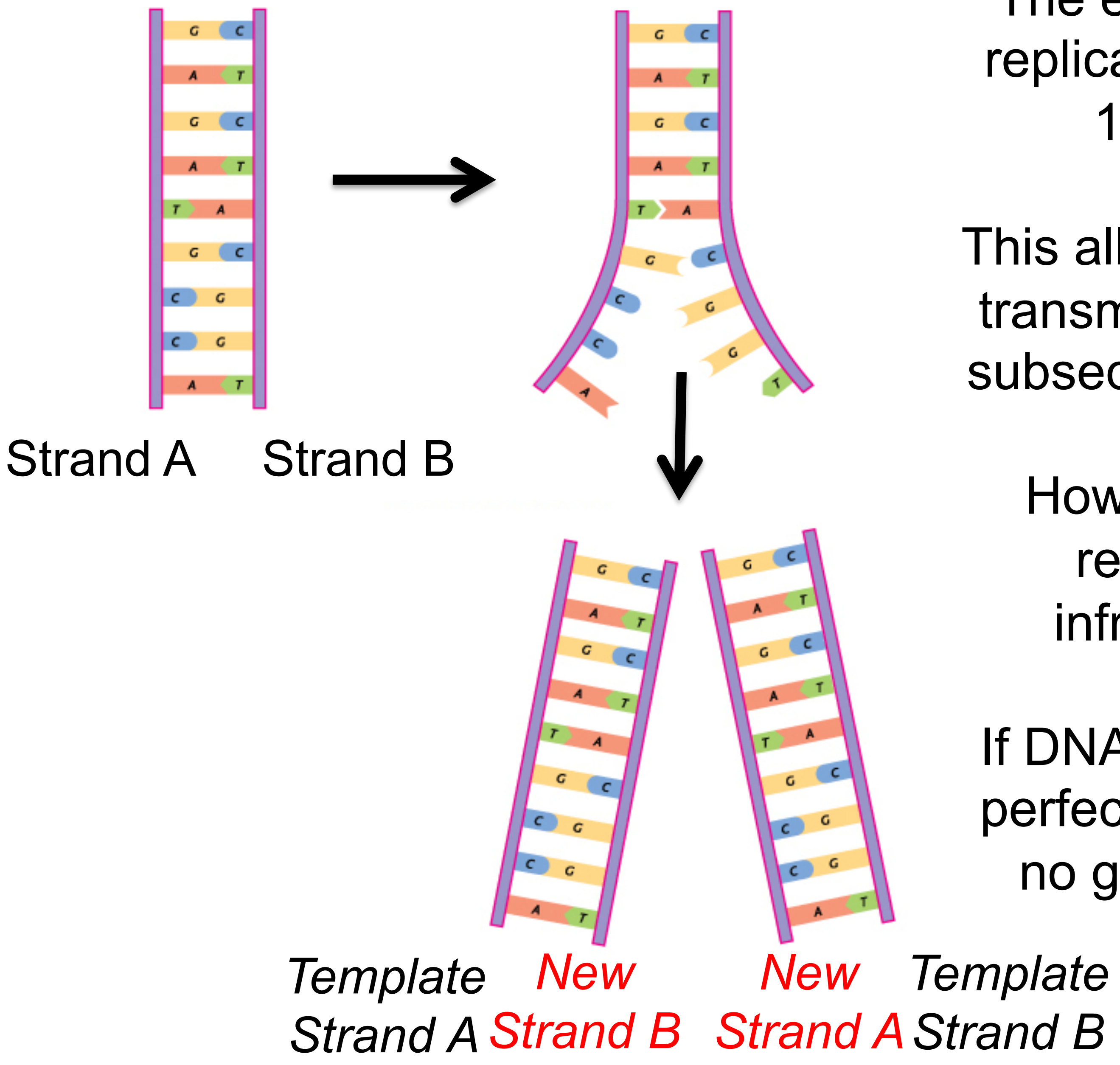
Replication of DNA (and of genetic information)


Molecole originali genitrici
Prima generazione di molecole figlie
Seconda generazione di molecole figlie

Thanks to the selective complementarity of the G-C & A-T base pairs (Watson-Crick base pairing), DNA can replicate itself generating novel chains identical to the parent ones



Replication of DNA (and of genetic information)



Strand A Strand B

Template Strand A *New Strand B* *New Strand A* *Template Strand B*

The error rate in DNA replication is as low as 1 base in 10^9

This allows to accurately transmit the genome to subsequent generations

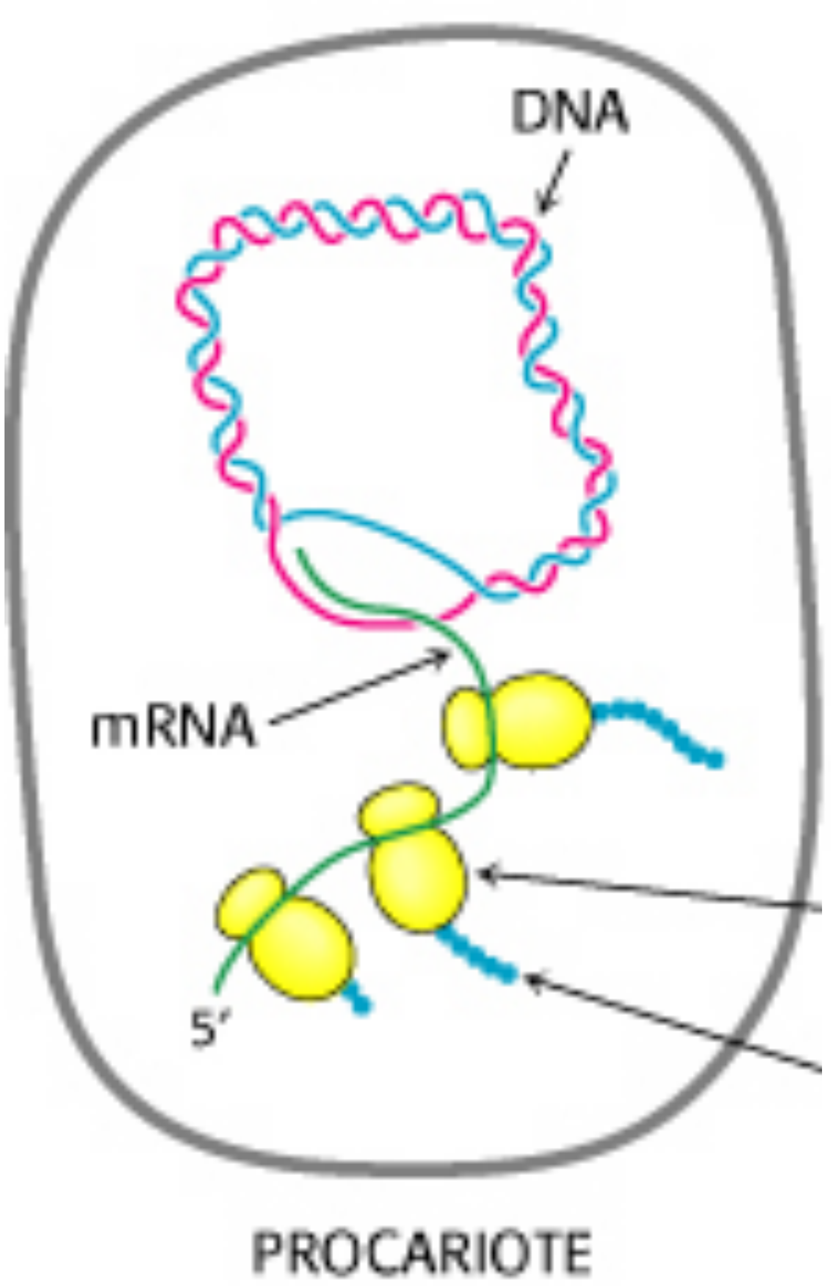
However, evolution relies on these infrequent errors

If DNA replication were perfect, there would be no genetic variation

DNA Transcription & Tranlastion: the central dogma

$$\text{DNA} \xrightarrow{\text{transcription}} \text{RNA} \xrightarrow{\text{translation}} \text{protein}$$

DNA Transcription & Tranlastion: the central dogma



Ribosome

Protein under synthesis

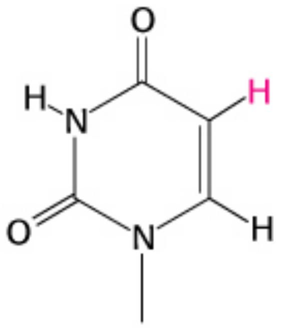
1. DNA is **transcripted** into a messanger RNA (mRNA) chain ▶

2. mRNA is then **translated** into proteins ▶

What’s the difference between RNA (RiboNucleic Acid) e DNA (Deoxy-riboNucleic Acid) ?

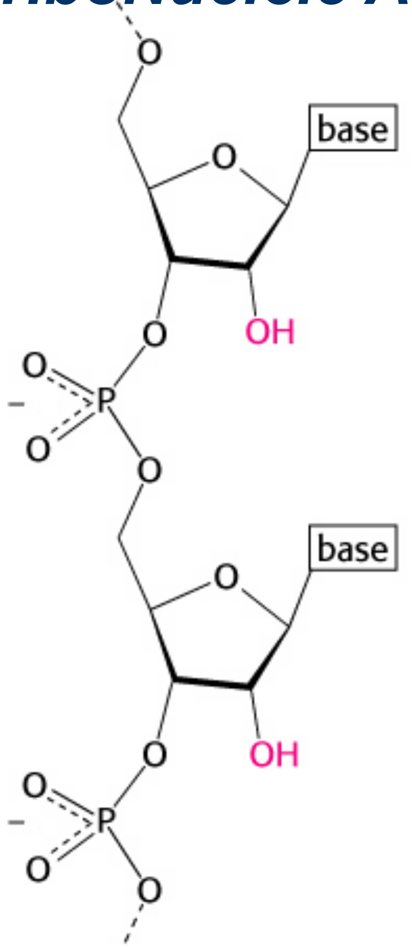
RNA sequence is made of

A, C, G, U

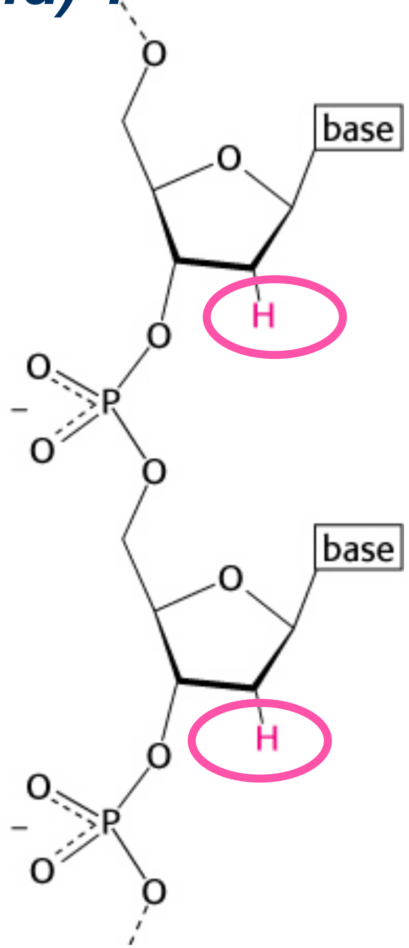


Uracile (U)

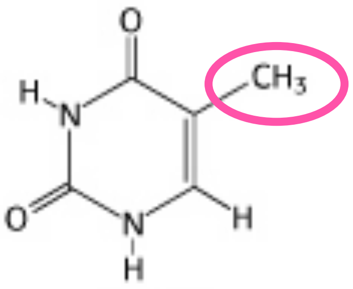
WC pairing: A-U, G-C



RNA



DNA

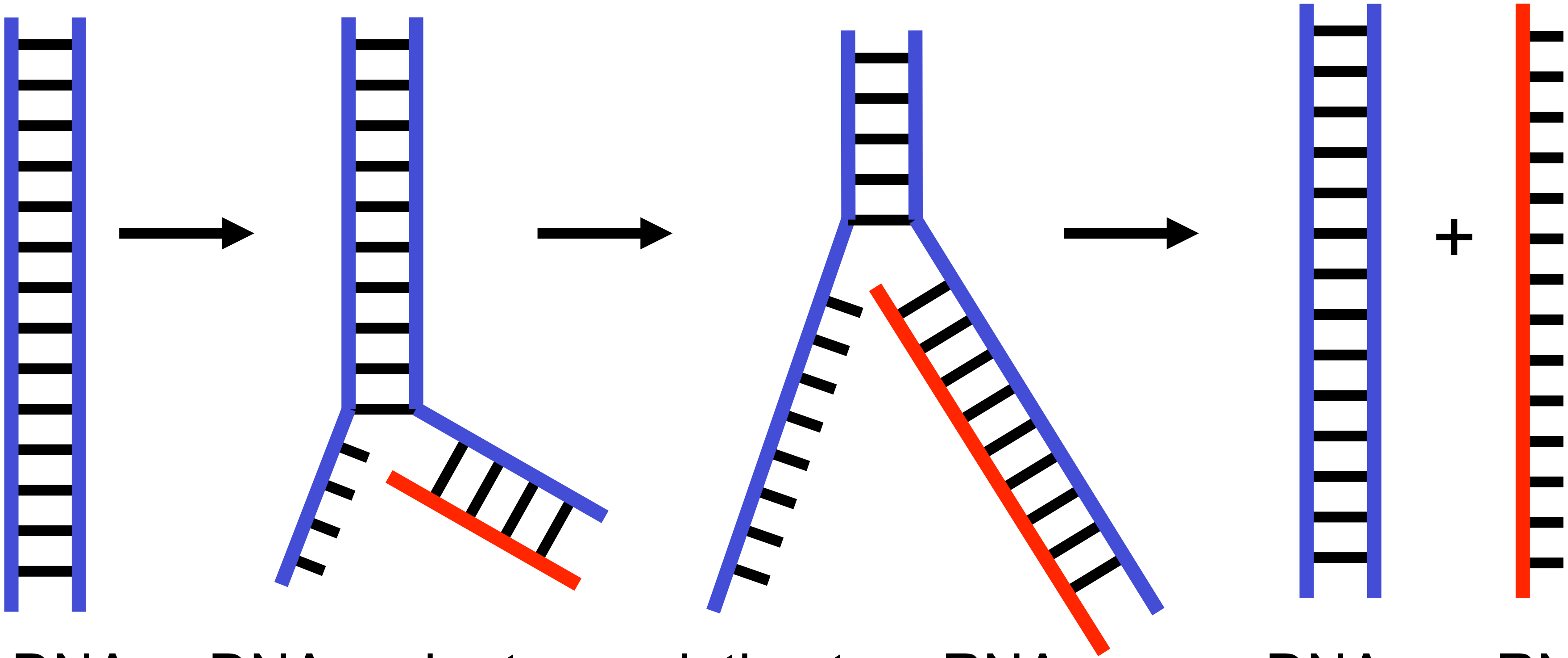


Timina

RNA is identical to DNA apart from the presence of a hydroxyl group (OH) on the C2’ atom of the sugar (ribose!) and the substitution of the Thymine (T) base with the Uracile (U) base

TRANSCRIPTION

1) Transcription of the gene information DNA to mRNA by a RNA-polymerase

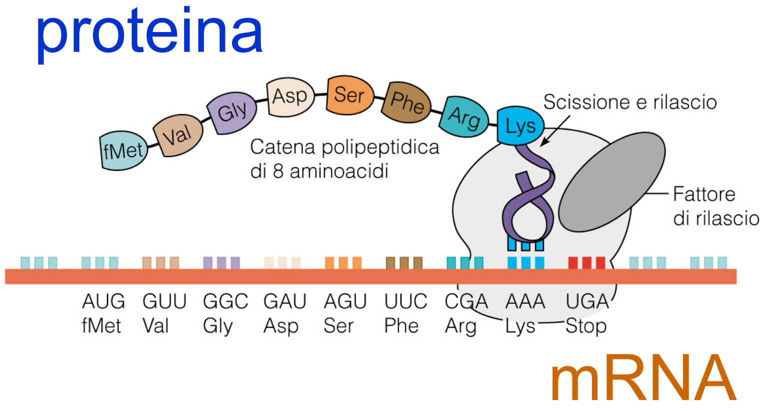


DNA DNA under transcription to mRNA DNA + mRNA

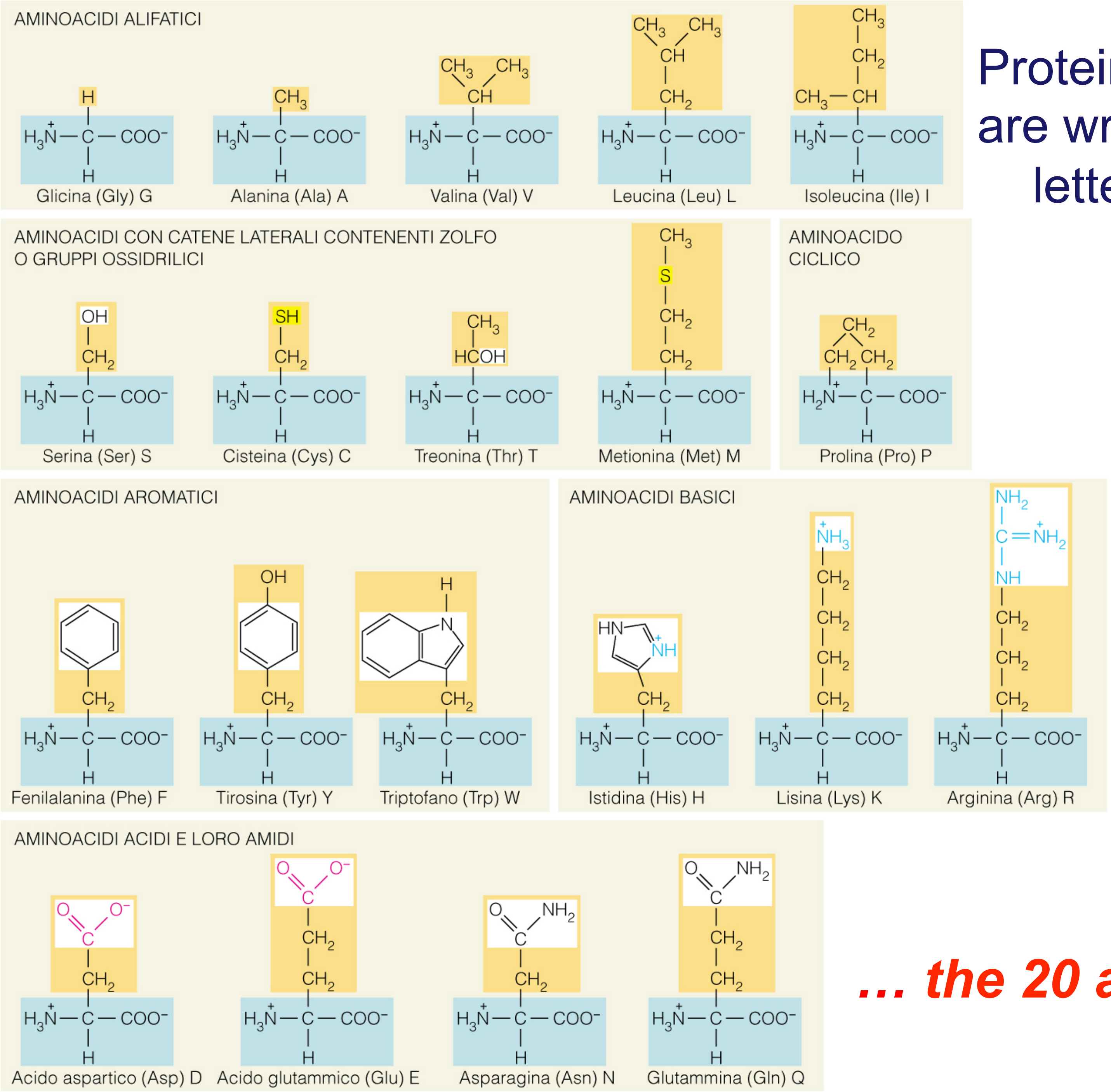
TRANSLATION

How is a nucleotide sequence (DNA/RNA, 4 nucleotides) translated into a protein sequence (20 amino acids)?

A **code** is needed !



Protein sequences are written in a 20-letter alphabet...



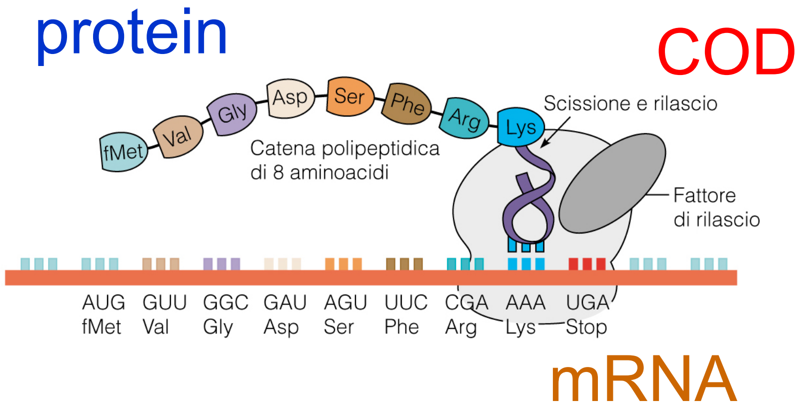
… the 20 amino acids

TRANSLATION

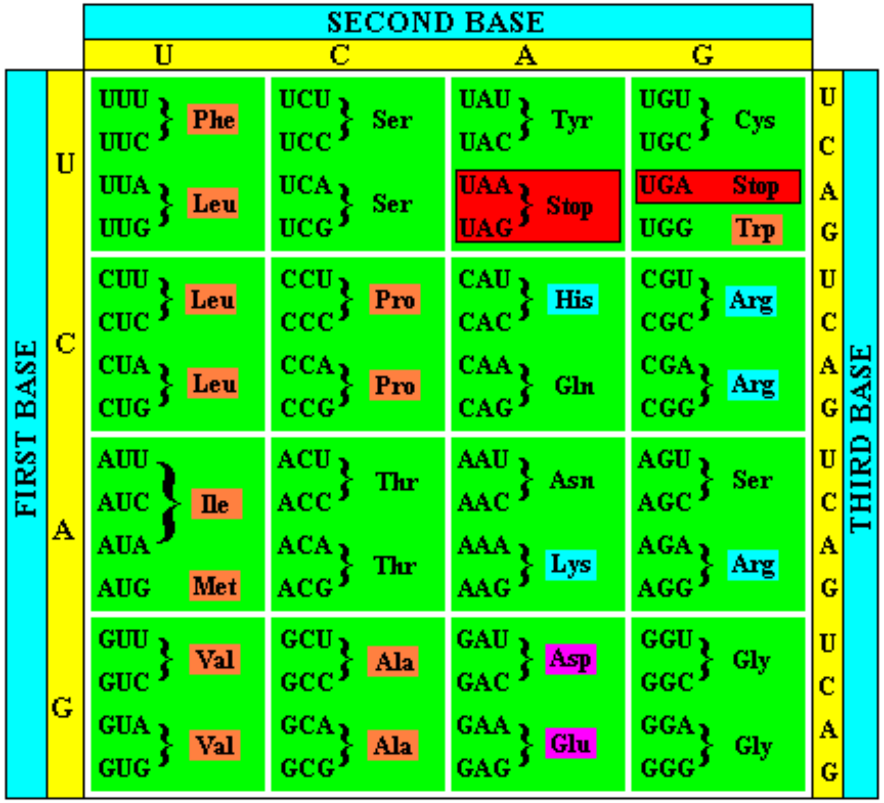
How is a nucleotide sequence (DNA/RNA, 4 nucleotides) translated into a protein sequence (20 amino acids)?

A **code** is needed !

UNIVERSAL GENETIC CODE



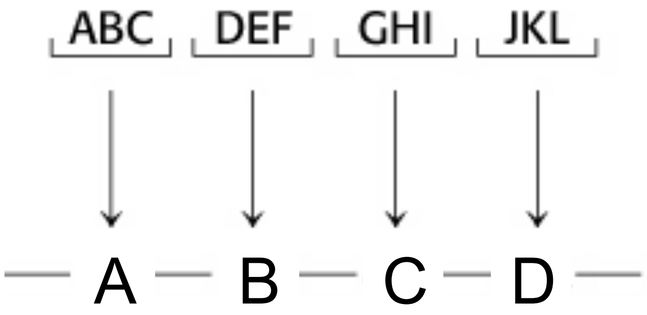
**Alternative genetic codes appear in organelles – chloroplasts and mitochondria – and in some species*



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

TRANSLATION

How is a nucleotide sequence (DNA/RNA, 4 nucleotides) translated into a protein sequence (20 amino acids)?



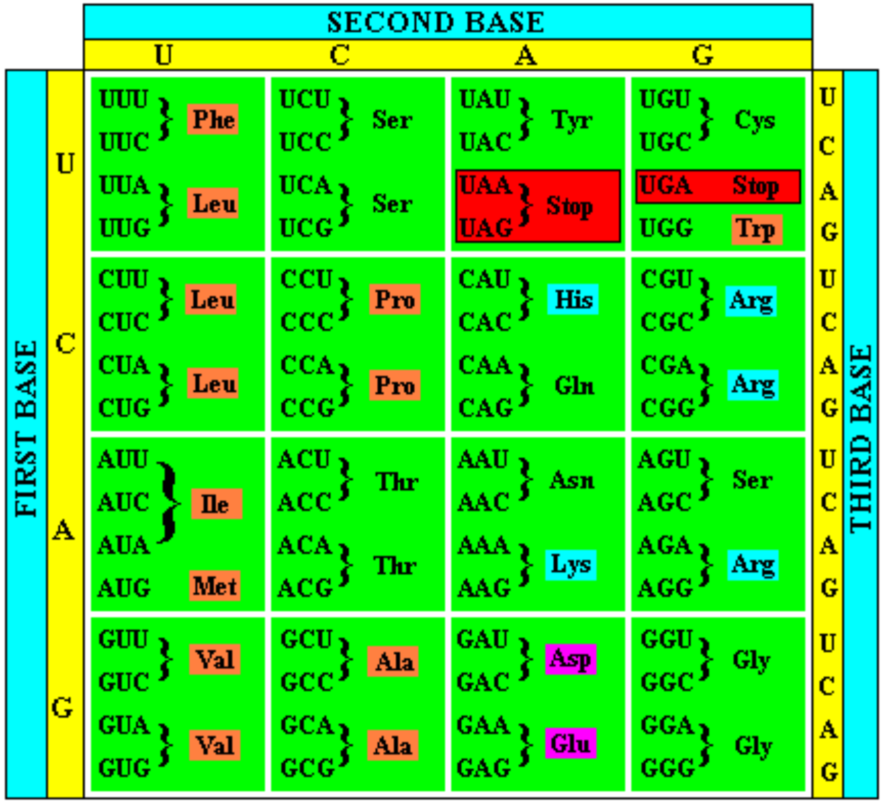
Sequenza di basi

mRNA

Sequenza di amminoacidi

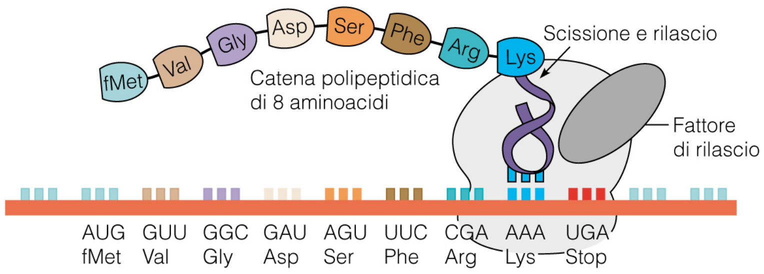
protein

UNIVERSAL GENETIC CODE



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

protein

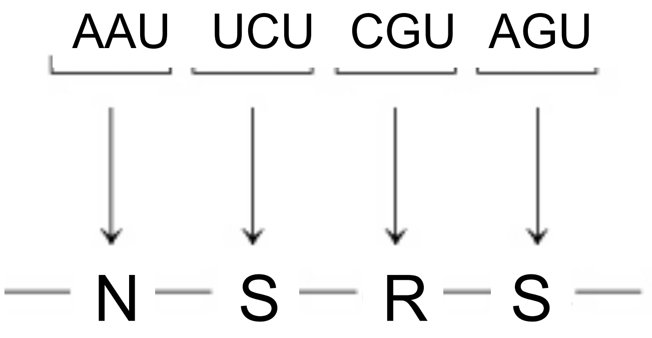

fMet
Val
Gly
Asp
Ser
Phe
Arg
Lys
Scissione e rilascio
Catena polipeptidica di 8 aminoacidi
Fattore di rilascio
AUG GUU GGC GAU AGU UUC CGA AAA UGA
fMet Val Gly Asp Ser Phe Arg Lys Stop

mRNA

TRANSLATION

How is a nucleotide sequence (DNA/RNA, 4 nucleotides) translated into a protein sequence (20 amino acids)?

EXAMPLE:

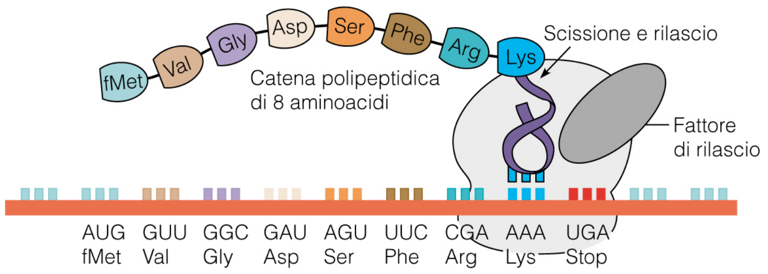


Sequenza di basi — *mRNA*

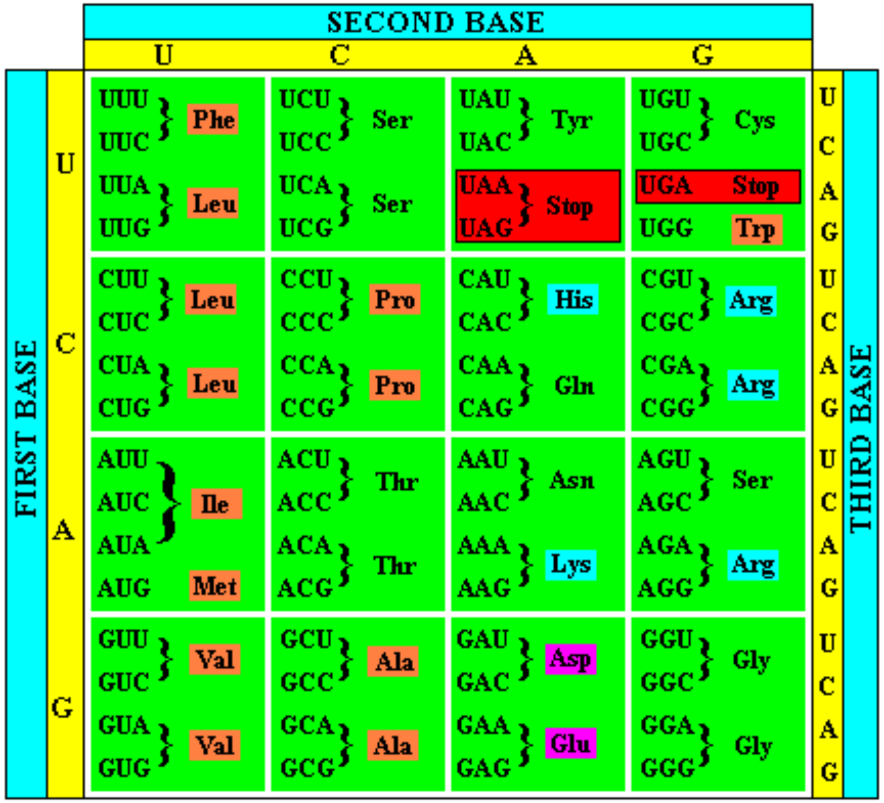
Sequenza di amminoacidi — *protein*

UNIVERSAL GENETIC CODE

protein



mRNA

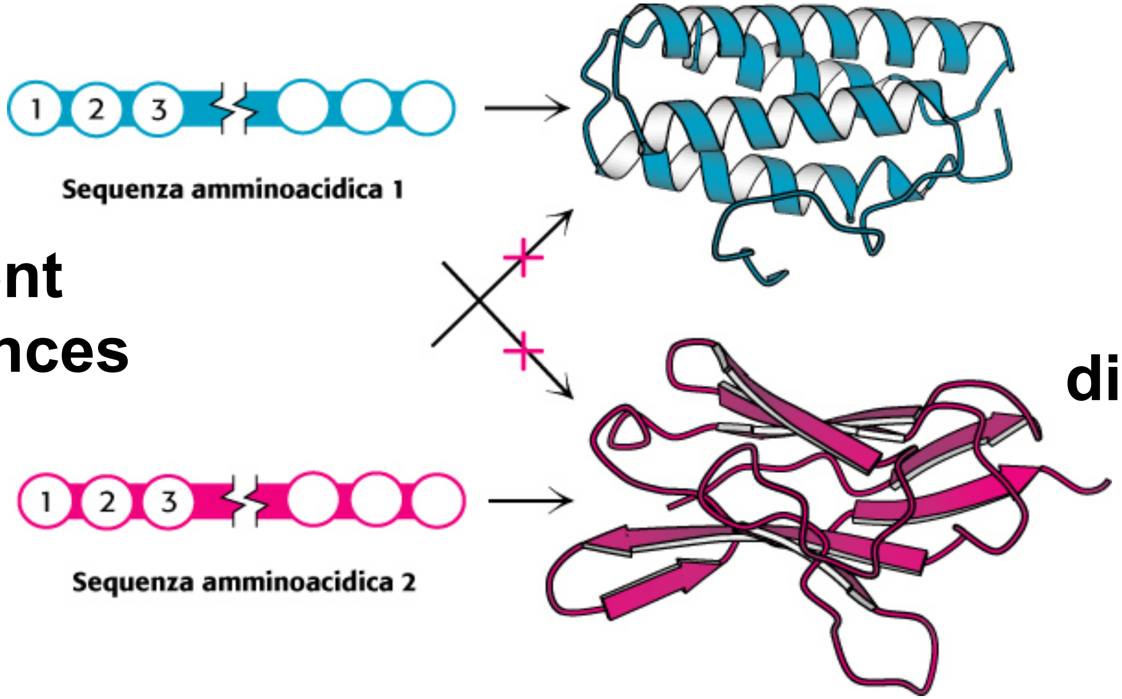


FIRST BASE	SECOND BASE U	SECOND BASE C	SECOND BASE A	SECOND BASE G	THIRD BASE
U	UUU UUC Phe	UCU UCC Ser	UAU UAC Tyr	UGU UGC Cys	U C
	UUA UUG Leu	UCA UCG Ser	UAA UAG Stop	UGA Stop UGG Trp	A G
C	CUU CUC Leu	CCU CCC Pro	CAU CAC His	CGU CGC Arg	U C
	CUA CUG Leu	CCA CCG Pro	CAA CAG Gln	CGA CGG Arg	A G
A	AUU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

Proteins are made by 20 amino acids, different in size and physico-chemical nature:

ACDEFGHIKLMNPQRSTVWY

Different sequences



Fold into very different 3D structures responsible for different biological functions!

RNA: current view

The RNA molecule involved in the central dogma is the messenger RNA, mRNA

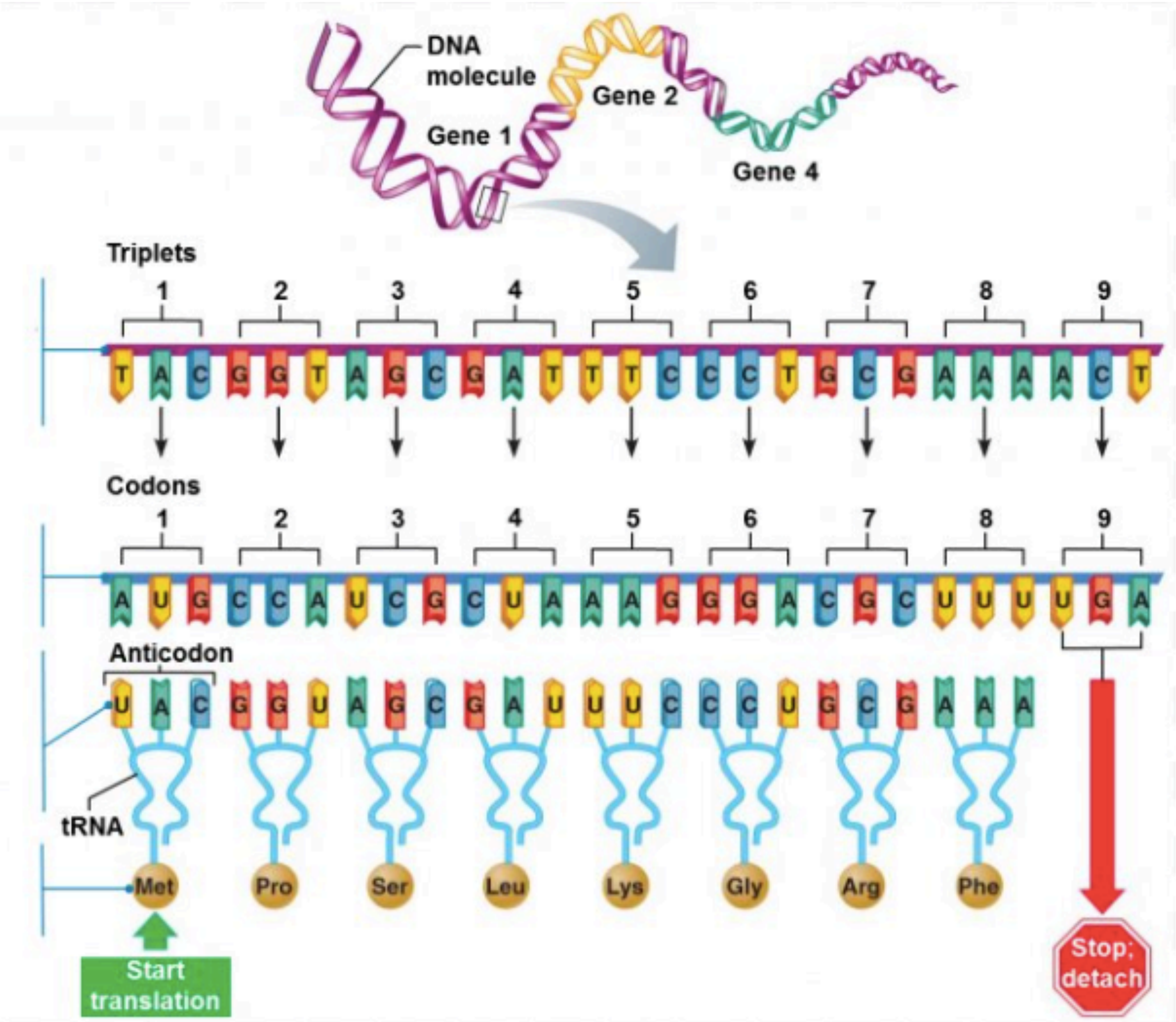
Traditionally, besides mRNA, the functional role of ribosomal RNA, rRNA, and transfer RNA, tRNA was recognized, both involved in the translation (to proteins) process

Nowadays it is instead recognized that RNA molecules have a variety of complex 3D structures and functions

Ribozymes are RNA molecules performing a catalytic activity,
While riboswitches and other RNAs such small interfering RNA (siRNA), microRNA (miRNA) and piwi-interacting RNA (piRNA) function to control translation

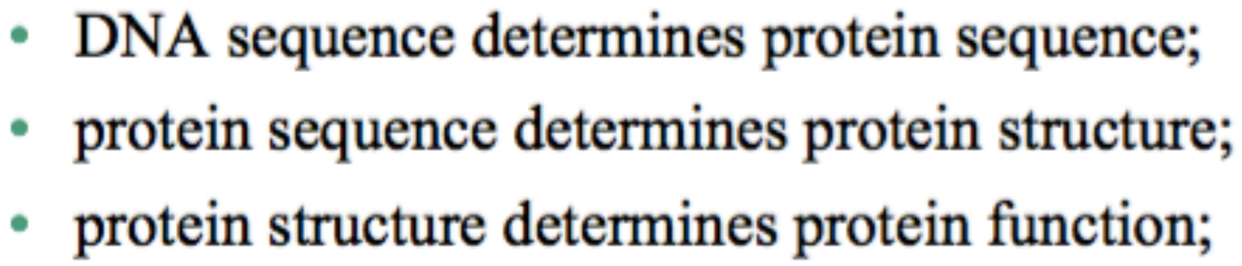
Example: **Tetrahymena ribozyme**

Information transfer from DNA to RNA to polypeptide (protein)



DNA molecule
Gene 1
Gene 2
Gene 4
Triplets
1 2 3 4 5 6 7 8 9
T A C G G T A G C G A T T T C C C T G C G A A A A C T
Codons
1 2 3 4 5 6 7 8 9
A U G C C A U C G C U A A A G G G A C G C U U U U G A
Anticodon
U A C G G U A G C G A U U U C C C U G C G A A A
tRNA
Met
Pro
Ser
Leu
Lys
Gly
Arg
Phe
Start translation
Stop; detach

It is all matter of information



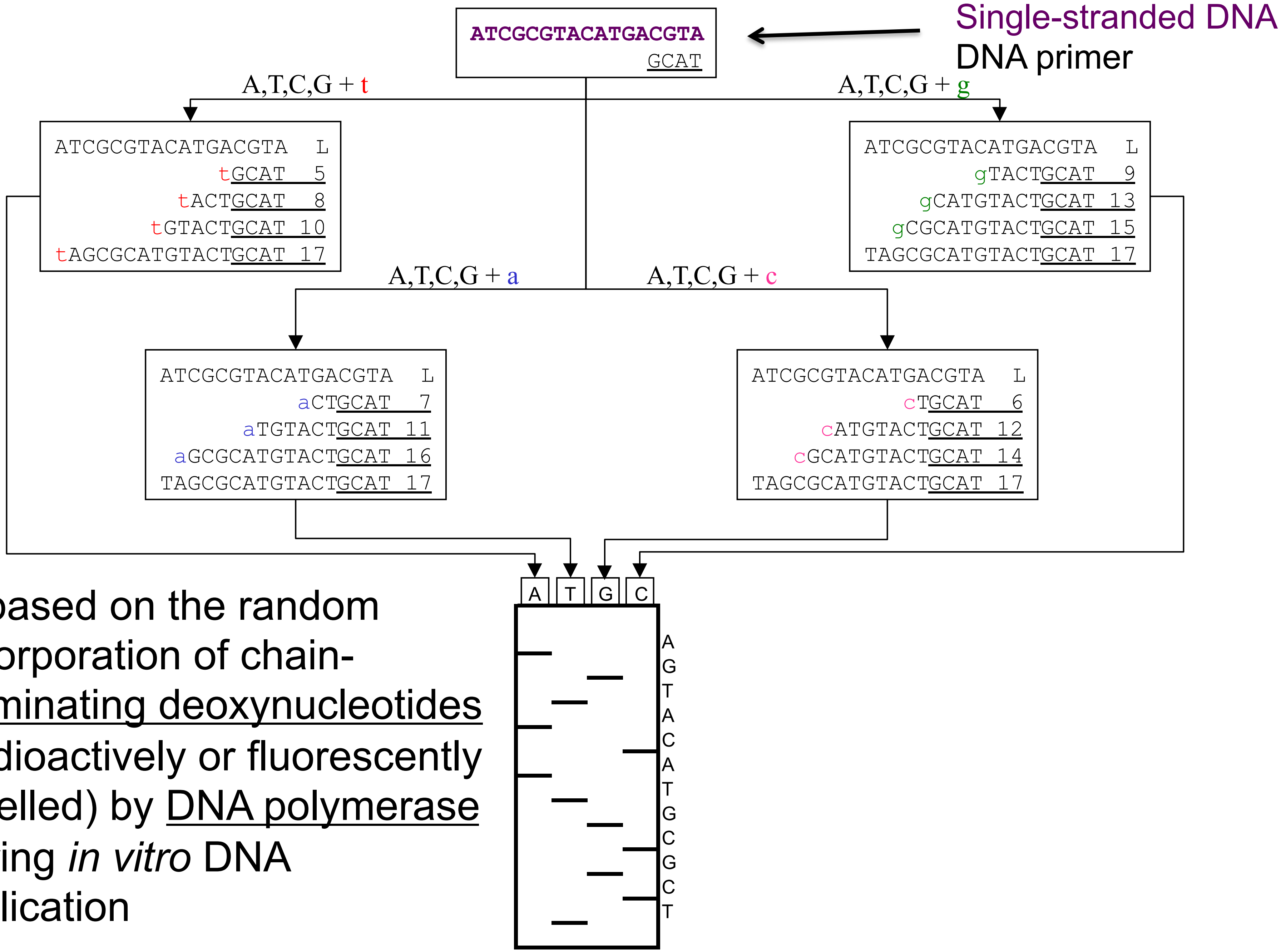
- DNA sequence determines protein sequence;
- protein sequence determines protein structure;
- protein structure determines protein function;

In addition, many regulatory mechanisms depend on the binding of proteins to other proteins, DNA, RNA or small molecules

Much of the Bioinformatics activities is focused on the analysis of the data related to the above processes

(We are overlooking by the moment the role of epigenetics, an upcoming field that studies the gene regulation - by modification of histone proteins, methylation of DNA, chromatin modeling, RNA-mediating silencing -, which changes the physiology of cells without altering the DNA sequence)

An efficient technique to sequence DNA, Sanger 1977



Single-stranded DNA

DNA primer

ATCGCGTACATGACGTA
GCAT

A,T,C,G + t

ATCGCGTACATGACGTA L
tGCAT 5
tACTGCAT 8
tGTACTGCAT 10
tAGCGCATGTACTGCAT 17

A,T,C,G + g

ATCGCGTACATGACGTA L
gTACTGCAT 9
gCATGTACTGCAT 13
gCGCATGTACTGCAT 15
TAGCGCATGTACTGCAT 17

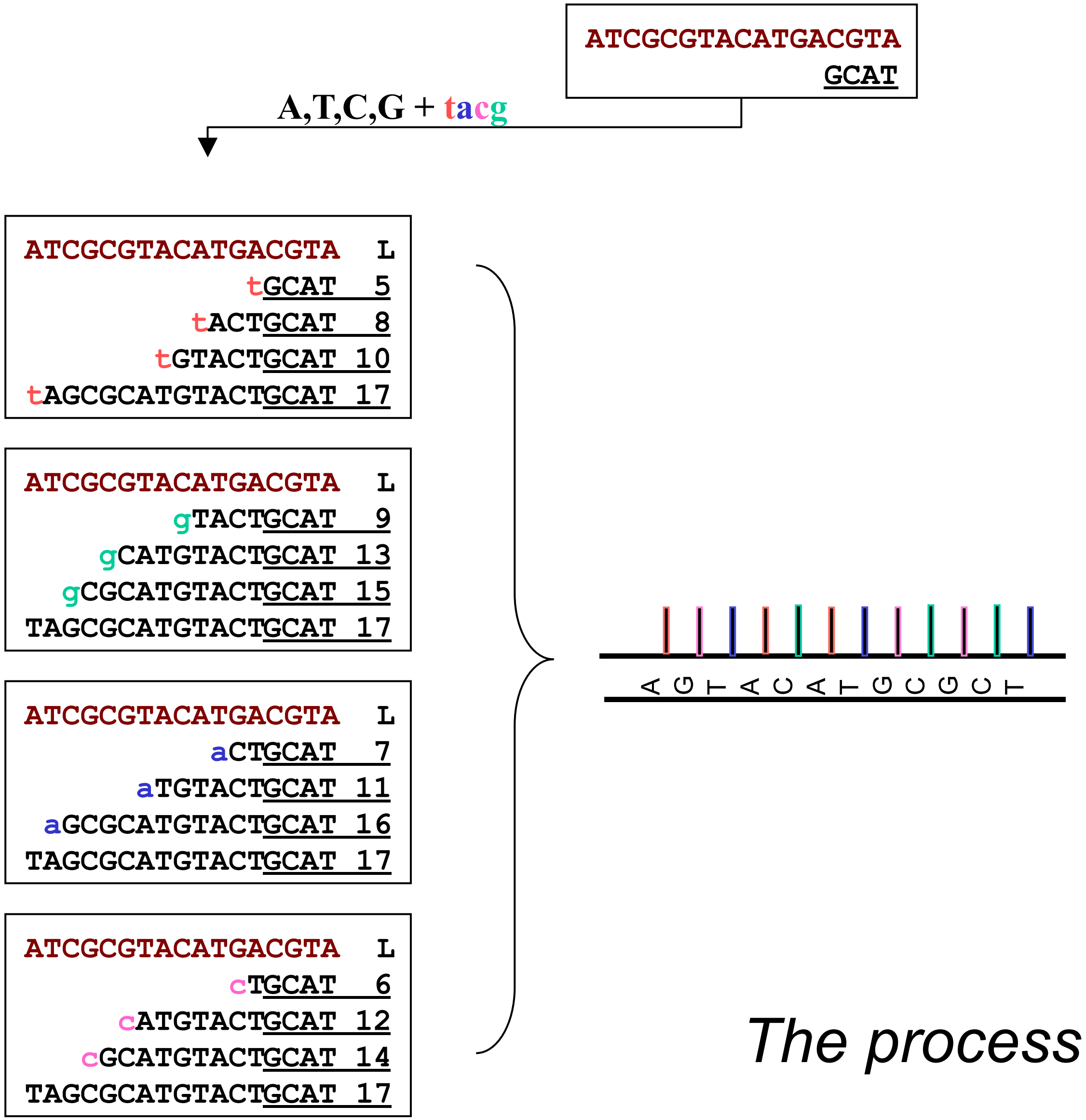
A,T,C,G + a

ATCGCGTACATGACGTA L
aCTGCAT 7
aTGTACTGCAT 11
aGCGCATGTACTGCAT 16
TAGCGCATGTACTGCAT 17

A,T,C,G + c

ATCGCGTACATGACGTA L
cTGCAT 6
cATGTACTGCAT 12
cGCATGTACTGCAT 14
TAGCGCATGTACTGCAT 17

Is based on the random incorporation of chain-<u>terminating deoxynucleotides</u> (radioactively or fluorescently labelled) by <u>DNA polymerase</u> during *in vitro* DNA replication



ATCGCGTACATGACGTA
GCAT

A,T,C,G + tacg

ATCGCGTACATGACGTA	L
tGCAT	5
tACTGCAT	8
tGTACTGCAT	10
tAGCGCATGTACTGCAT	17

ATCGCGTACATGACGTA	L
gTACTGCAT	9
gCATGTACTGCAT	13
gCGCATGTACTGCAT	15
TAGCGCATGTACTGCAT	17

ATCGCGTACATGACGTA	L
aCTGCAT	7
aTGTACTGCAT	11
aGCGCATGTACTGCAT	16
TAGCGCATGTACTGCAT	17

ATCGCGTACATGACGTA	L
cTGCAT	6
cATGTACTGCAT	12
cGCATGTACTGCAT	14
TAGCGCATGTACTGCAT	17

The process can be automated...

Although largely substituted by Next Generation Sequencing, it is still actively used in projects requiring high quality outputs, e.g. efforts for public health such as sequencing the spike protein from SARS-CoV-2

“*Bio*informatician” problems

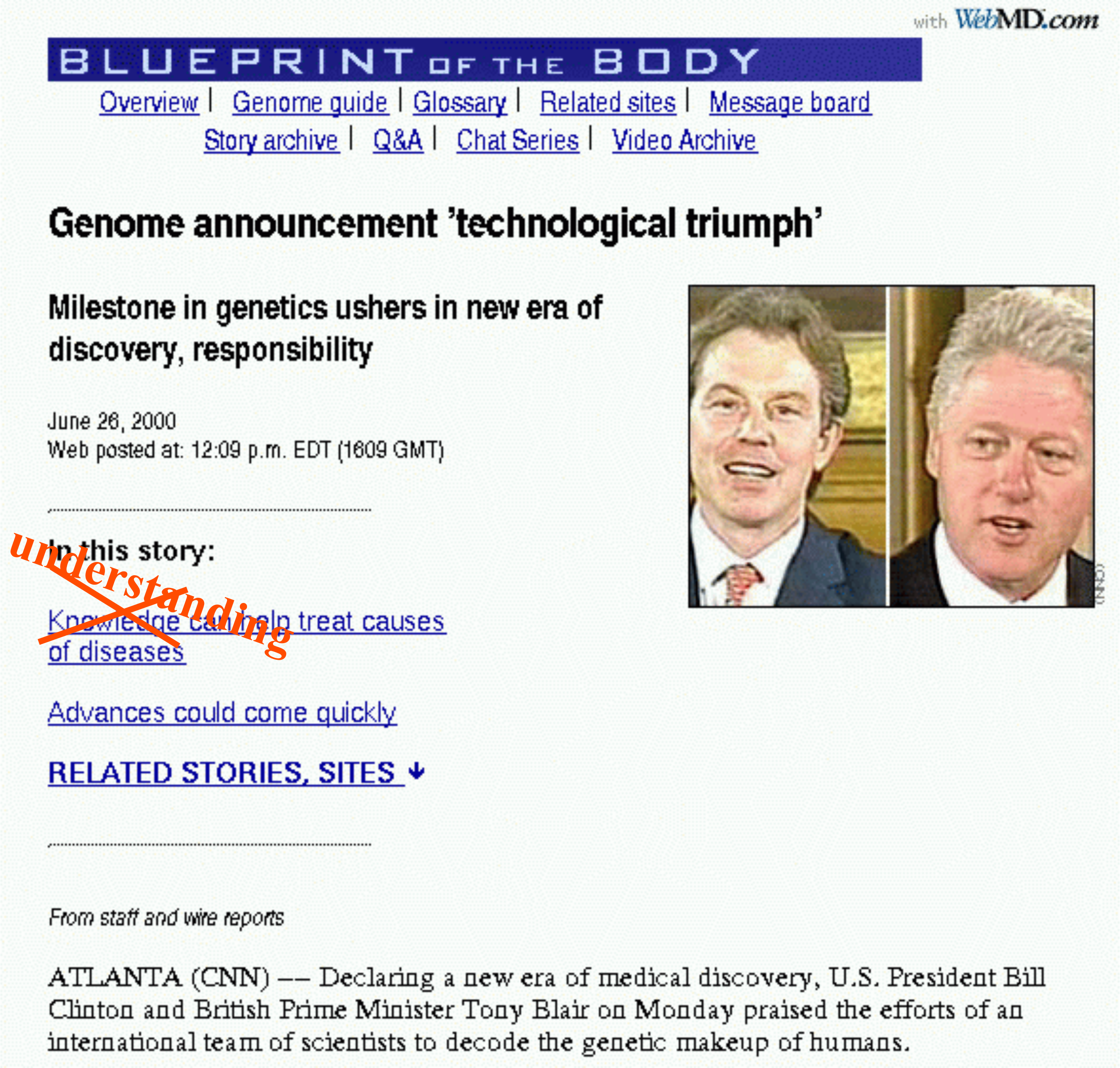
Storing DNA sequences

Concatenating DNA sequence fragments

Calculating the sequence complementary to a given DNA strand

Transcribing DNA sequences in RNA sequences

Translating DNA sequences in protein sequences (through the universal genetic code)



with WebMD.com

BLUEPRINT OF THE BODY

Overview | Genome guide | Glossary | Related sites | Message board

Story archive | Q&A | Chat Series | Video Archive

Genome announcement 'technological triumph'

Milestone in genetics ushers in new era of discovery, responsibility

June 26, 2000
Web posted at: 12:09 p.m. EDT (1609 GMT)

In this story:

~~Knowledge~~ understanding can help treat causes of diseases

Advances could come quickly

RELATED STORIES, SITES ↓

From staff and wire reports

ATLANTA (CNN) -- Declaring a new era of medical discovery, U.S. President Bill Clinton and British Prime Minister Tony Blair on Monday praised the efforts of an international team of scientists to decode the genetic makeup of humans.

≈ 98.5% of human DNA is non-coding

Where is the gene?

```
>cD0826Q1_425-22425 Main
ggcataagaatgatacaatggactttggggacctgagaggaaaggtgggagggggcaagg
gatactgctcaggtgataggtgcaccaaaatctcacaaatcatcactaaagaacttactc
atgtaaccaaatactacctgtaccactataacctacgggggaaaaaagcaacataaccat
gaaccaactaataaaaaacaaccttgccttcagtctgcatcctaccctagagacactctc
tctgtgtcctcacacttggagctaagcttctgacttttgtctccagtacacccctgagga
tcctctcatcacggccatcagaaacctctgtagaaggtcaaatccagtgggttcttgtca
gtgcctctgacttgagttactgataatatttgcaccataatccacttctttctaatgagc
tactctgtccttatttttctcctatttactgaatcctccttatcatcctttgaaatctcc
tcttaattattatgttctctcatcataccctgagatccctgcatttctgatttttggcac
tcttcctggaaaagctcatctaacctgcacctatgcttgatgactctcagttctctggct
taaactcctctactgagaccacccatcatacaaaaatgtttacatattatttttccttag
ataacttttagatattctaagtgcaatagccccacactgaactcagtctcttctctcagt
caggctgtcttctctcattacccttttttaatgaatggaatcaagatgtttgcattgggtt
ggggagatgttggtcaaaggatacatccatttcatttcatttaggatacatttcaaaaga
tacatttcatttagattggaggaataattttaagagttttattgtataacatggactata
gttgctaacaatgtattgttgaaaattgctaaaagggtggattttaagtgttctcaccac
aaaaaataagtatgtgaggtgagccataagttctttagcttgatgtagccggtccatgat
gtacatacatttcaaaacaacatattatacatgataaatataaataattttttgtcaatca
aaataatttagaaaagtgacacacacttacacacacacacacaaaagagatgattgcatt
ggccagtctaggaataagagttatctgggagttttctaagtcggatgccaccgacatcac
tcaccaataatccctttaatgtcaatcaaattaagtcctcttcttccatcattttactcc
tatgcccatttcctcactctttgttcaggcactattagtcttgcctcttgaaccaacttc
tttcactcatgctgcccactgttgccgtagtgatcttcctaaattgcaaatgcgccatca
ctctcctgcttaaaatccttcaatgattccttatgacttccaggacagagtagccactcc
tgagctttgcatgtaacatctgtcatgatccagcccctgcctgtctatttttcctttttt
cttgctgctgttccacatccaaagctggctccattcatactgaagcagctgaagttcttc
agatatgtcattgccacactgggcccacacttttgaacctgcttcctcctgtgtgagaag
tggcttctgccctgttttcggactgcctacattgaagccatctgttccccaggaagcctt
ccctgatgccttgacagcagcatcttgtgcctgccccatatctgcacttatccatctggg
cctgctgttgtcttgtcacttgtgttctcttctgtgaactgtaaacatcaggaggacaag
acctatgtcttacttttatttgaatatttagcatctaacaatgttcgacatatagtaggc
ttttgatactatttttttactatgacattgtagtatatgttaatatccagtaggacatag
gatatattctctctgttttcaatttttcattgtttacacacatttataattctatctata
aggatttacaattatttacatgaaatgaatgaaataaatagagaatgttagatattaaga
gacagtgtggaaagccaggctgggactagggatgcacttaccttaggtgcaaaatttagg
aggataccaaaagaactcagtaataaaagtcaatcatattttaatgaaatatcttaagaa
atctaaattaatggaaaatatataatgaacaaaatgtcaaaagagaactattcaaagaaa
atggagaagcagagaggcagaagaattagtagaatatactggcacataagccaaggaggt
aaagatttccaggaaggaggaagtagagtggagtcagaagttcaacagaagtcatttcag
aaatcttaccttggttttgaaatcctttcagagagcagttttacataatgtgagcaatta
tttctccttcatccccatcattccagaattgagcttcttctctggcttcagaaatgtggc
ccttccccttgtcaggatatgttggcgacatgatgcatgcggatgccctcaaagtcagct
ggggtttgggggtgaaattaattgactttagggaactccttgaatgctaagttctgttca
cctggaggaccagagagggcacagagatgaccacctagcttctgcctgggacctaaacag
ggcagagaaataggaggatcaggtataaagggagcagggaagatgggtctgggcttacag
```

Hay in a haystack
(A. Tramontano)

Where is the gene?

The DNA length in the human genome is approximately **3.2 billion nucleotides**

Such a nucleotide sequence (combination of A, C, G, T) is not random !!!

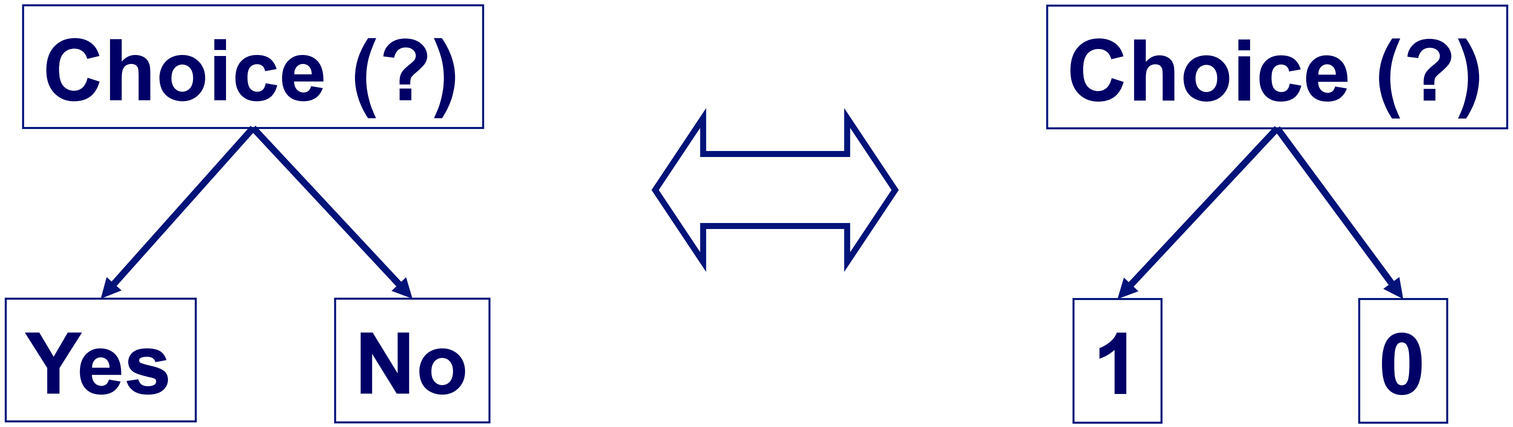
Information theory (Claude Shannon)

Information is a ***universal measure*** of order and can be applied to any structure or system.

Order refers to the structural disposition of a system.

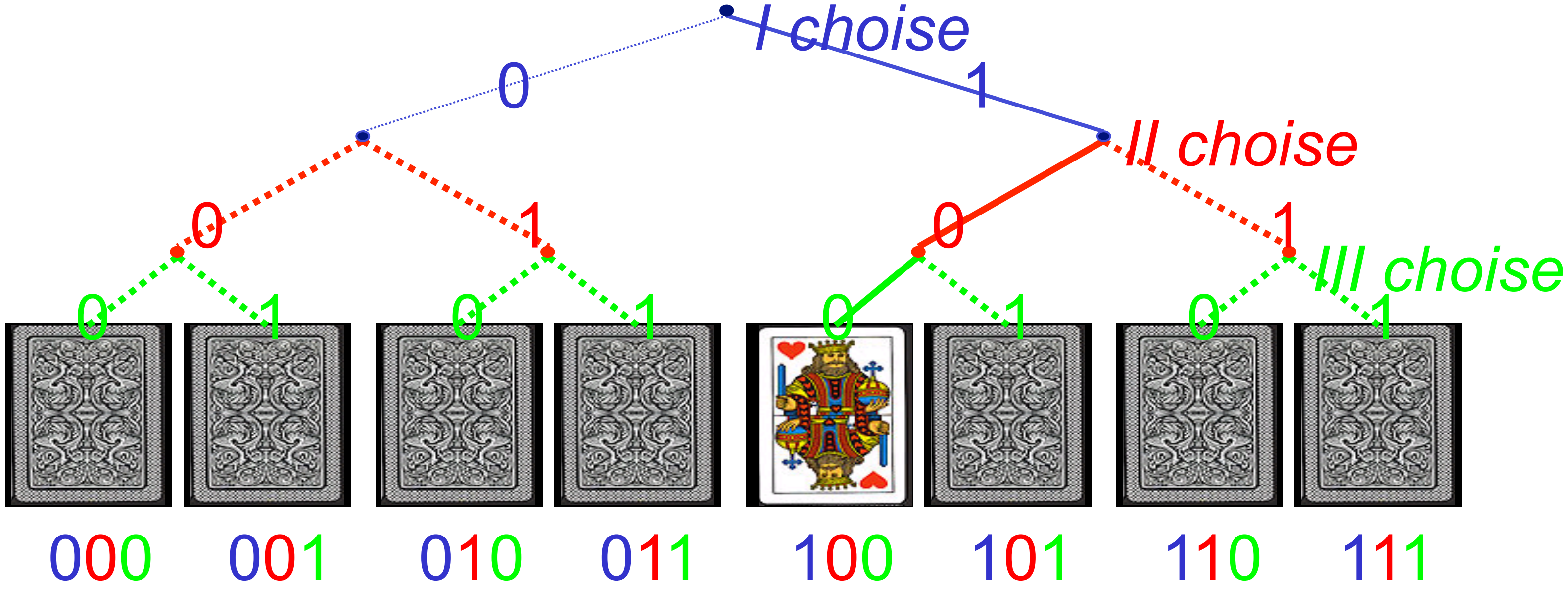
```
MSKGPAVGIDLGTTYSCVGVFQHG
KVEIIANDQGNRTTPSYVAFTDTE
RLIGDAAKNQVAMNPTNTVFDAKR
LIGRRFDDAVVQSDMKHWPFMVVN
DAGRPKVQVEYKGETKSFYPEEVS
SMVLTKMKEIAEAYLGKTVTNAVV
TVPAYFNDSQRQATKDAGTIAGLN
VLRIINEPTAAAIAYGLDKKVGAE
RNVLIFDLGGGTFDVSILTIEDGI
FEVKSTAGDTHLGGEDFDNRMVNH
FIAEFKRKHKKDISENKRAVRRLR
```

Information quantifies the instructions needed to produce a certain organization and can be (parsimoniously) achieved in terms of *binary choices* expressed in bits



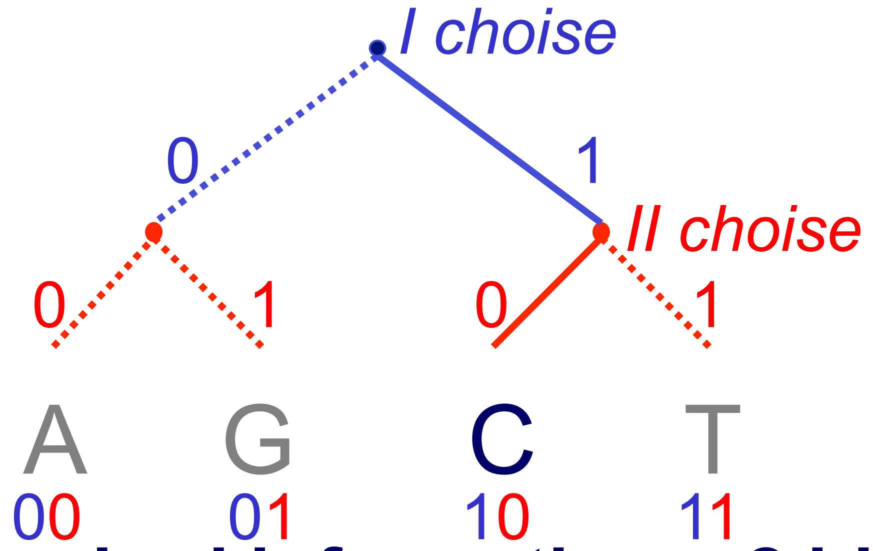
Shannon's informational entropy is the number of binary digits required to encode a message

Let’s choose a playing card from a 8-card pack



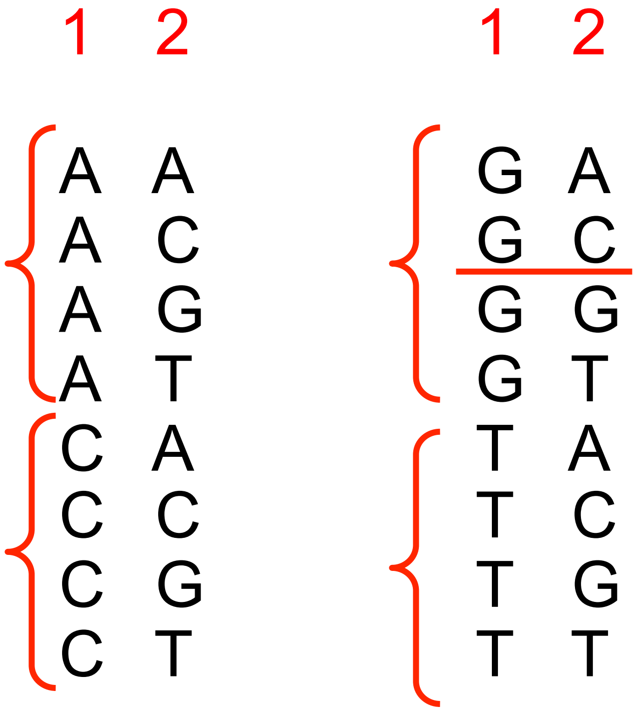
Required information = 3 bits

Let’s choose a nucleotide from the 4-letter alphabet



Required information = 2 bits

How much information is contained in a dinucleotide sequence, e.g. ‘GC’ ?



1	2	1	2
A	A	G	A
A	C	G	C
A	G	G	G
A	T	G	T
C	A	T	A
C	C	T	C
C	G	T	G
C	T	T	T

In information theory a GC sequence corresponds to **4 bits**

How “ordered” is the human genome

Overlooking spontaneous somatic mutations, we can say that APPROXIMATELY DNA molecules of a given individual feature the same sequence and compute its information content.

The approximate length of DNA in the human genome is
≈ $3.2*10^9$, corrisponding to **$6.4*10^9$ bits** !!

From a statistical point of view

S

How high is the probability that the ‘GC’ nucleotide sequence is spontaneously (randomly) generated?

S

The success probability of an event is given by:
The ratio between the number of favorable cases and the total number of cases

$$probability = \frac{favorable\ cases}{total\ cases}$$

S

The success probability of an event is given by:
The ratio between the number of favorable cases and the total number of cases

$$probability = \frac{favorable\ cases}{total\ cases}$$

Example:

If we throw 2 dices, how high is the probability of the **12** outcome?

Favorable cases:

Total cases:

The success probability of an event is given by:
The ratio between the number of favorable cases and the total number of cases

$$probability = \frac{favorable\ cases}{total\ cases}$$

Example:

If we throw 2 dices, how high is the probability of the **12** outcome?

Favorable cases:

Total cases:

$\boxed{N_{(tot)} = y^x = 6^2}$

N(tot) = nb of possible states
x = nb of available positions (dice 1, dice 2)
y = nb of possible choices for each position(6)

S

The success probability of an event is given by:
The ratio between the number of favorable cases and the total number of cases

$$\mathit{probability} = \frac{\mathit{favorable\ cases}}{\mathit{total\ cases}}$$

Example:

If we throw 2 dices, how high is the probability of the **12** outcome?

Favorable cases: **1**

Total cases: **$6^2 = 36$**

S

The success probability of an event is given by:
The ratio between the number of favorable cases and the total number of cases

$$probability = \frac{favorable\ cases}{total\ cases}$$

Example:

If we throw 2 dices, how high is the probability of the **12** outcome?

Favorable cases: **1**

Total cases: **$6^2 = 36$**

Probability = 1/36 = 0.027

S

The success probability of an event is given by:
The ratio between the number of favorable cases and the total number of cases

$$probability = \frac{favorable\ cases}{total\ cases}$$

Example:

If we throw 2 dices, how high is the probability of the **7** outcome?

Favorable cases:

Total cases:

The success probability of an event is given by:
The ratio between the number of favorable cases and the total number of cases

$$probability = \frac{favorable\ cases}{total\ cases}$$

Example:

If we throw **2** dices, how high is the probability of the **7** outcome?

Favorable cases: **6** (1+6,6+1,2+5,5+2,3+4,4+3)

Total cases: **6^2 = 36**

Probability = 6/36 = $0.\overline{16}$

S

The success probability of an event is given by:
The ratio between the number of favorable cases and the total number of cases

$$probability = \frac{favorable\ cases}{total\ cases}$$

Example:

If we throw **3** dices, how high is the probability of the **4** outcome?

Favorable cases:

Total cases:

The success probability of an event is given by:
The ratio between the number of favorable cases and the total number of cases

$$probability = \frac{favorable\ cases}{total\ cases}$$

Example:

If we throw **3** dices, how high is the probability of the **4** outcome?

Favorable cases: **3**

Total cases: $\mathbf{6^3 = 216}$

Probability = 3/216 = $0.013\overline{8}$

How high is the probability that the sequence of the human genome is spontanously (randomly) generated?

$$probability = \frac{favorable\ cases}{total\ cases}$$

How high is the probability that the sequence of the human genome is spontanously (randomly) generated?

$$probability = \frac{favorable\ cases}{total\ cases}$$

$$N_{(tot)} = 4^{3.200.000.000}$$

$x \cong 3.2*10^9$

$y = 4$ (A, G, T, C)

The number of possibilities N(tot) is larger than the estimated number of atoms in the universe!!!

But nature choses **ONLY ONE**…

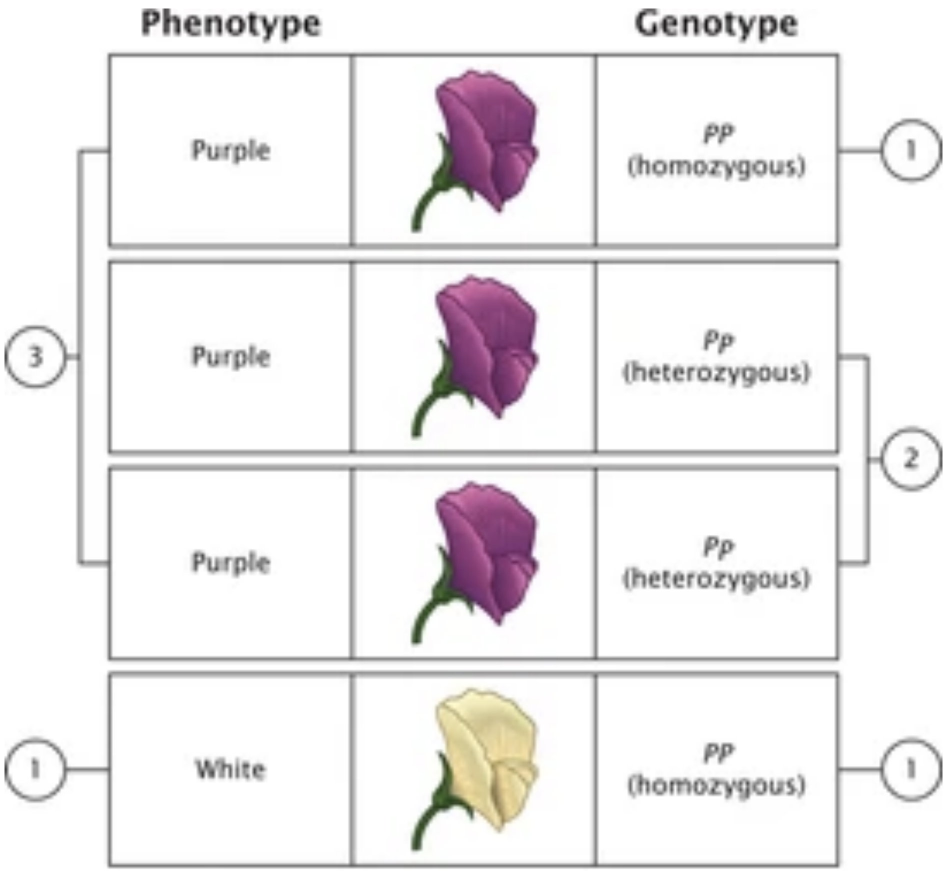
- The *information content* of the genomes of organisms belonging to the various species is *huge*

- Nucleotide sequences of genomes are not randomly generated

- Information relative to biological systems in nature gradually accumulates through processes of *casual variation of the genotype* and *natural selection* (Charles Darwin, *Origin of Species*)

The large molecules in living organisms offer the most striking example of information density in the universe

Genotype *vs* phenotype



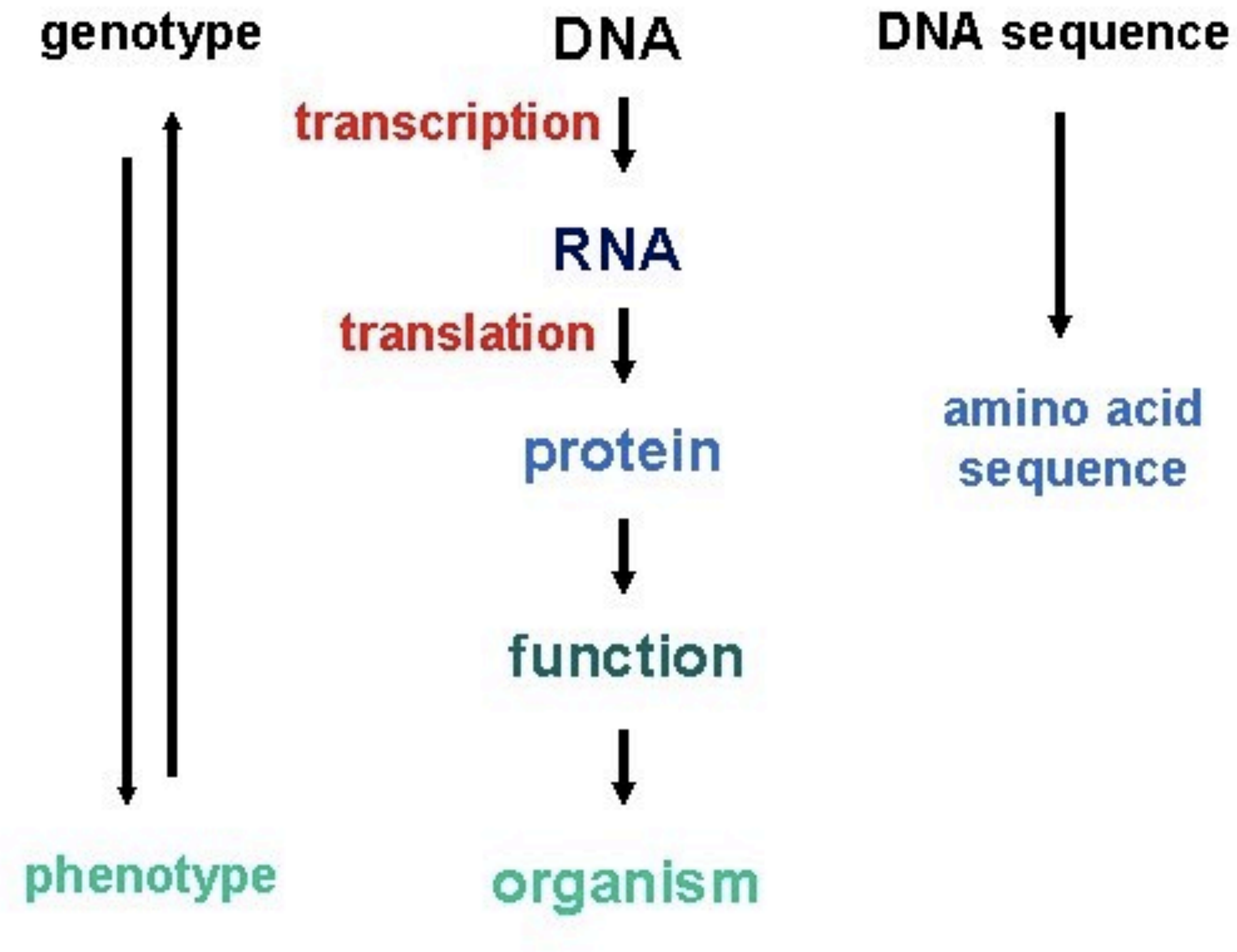
Phenotype		Genotype
Purple		*PP* (homozygous)
Purple		*Pp* (heterozygous)
Purple		*Pp* (heterozygous)
White		*pp* (homozygous)
Ratio 3:1		Ratio 1:2:1

shutterstock.com · 1462778021

The combination of alleles that an individual possesses for a specific gene is their **genotype**

Phenotype is determined by the genotype, but is also influenced by epigenetic modifications, environmental and lifestyle factors

Molecular basis for genotype vs phenotype



genotype

phenotype

DNA

transcription

RNA

translation

protein

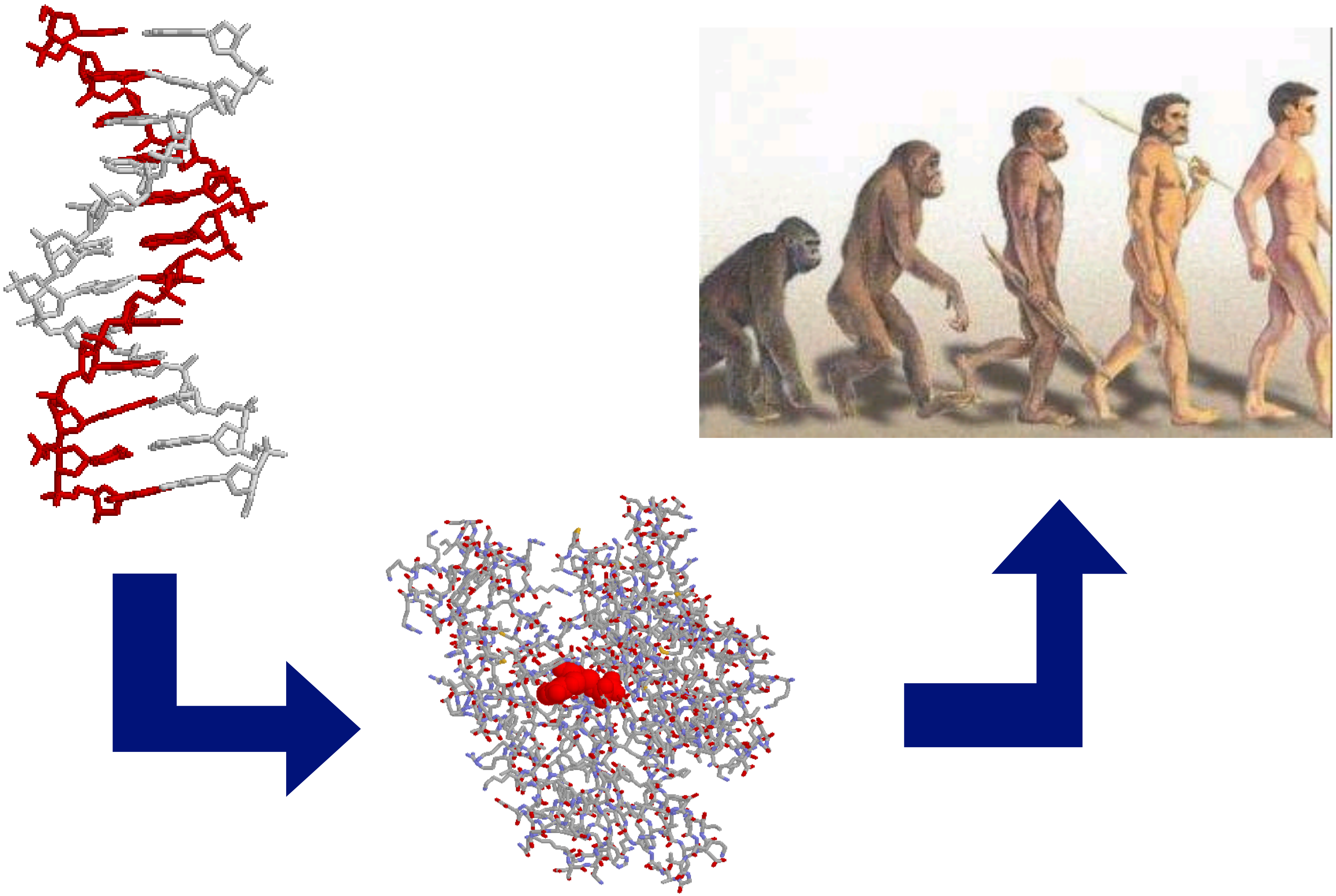
function

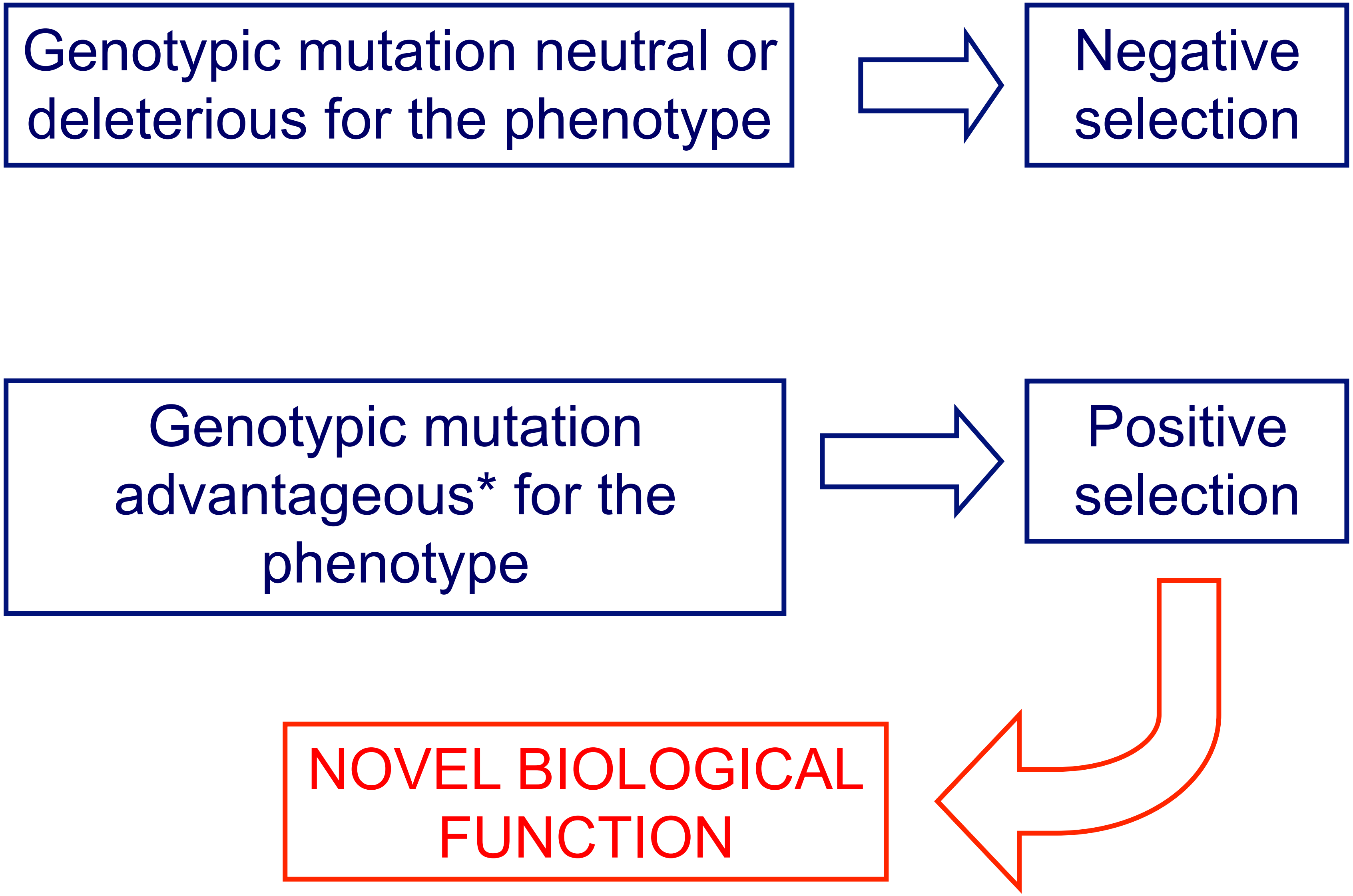
organism

DNA sequence

amino acid sequence

MOLECULAR EVOLUTION





Genotypic mutation neutral or deleterious for the phenotype

Negative selection

Genotypic mutation advantageous* for the phenotype

Positive selection

NOVEL BIOLOGICAL FUNCTION

* Advantageous mutations are rare as compared to the neutral and deleterious ones

Basic principles of evolution

B

All living species have evolved from other species

All living species are related to each other at different rates through common ancestors

All living species have a common descent, maybe existed 3.5 to 3.8 billion years ago (*L.U.C.A.: Last Universal Common Ancestor*)

The process through which a species evolves into another species involves **casual mutations**, of which those resulting in a **survival advantage** spread and persist more than the neutral or deleterious ones

Universal phylogenetic tree

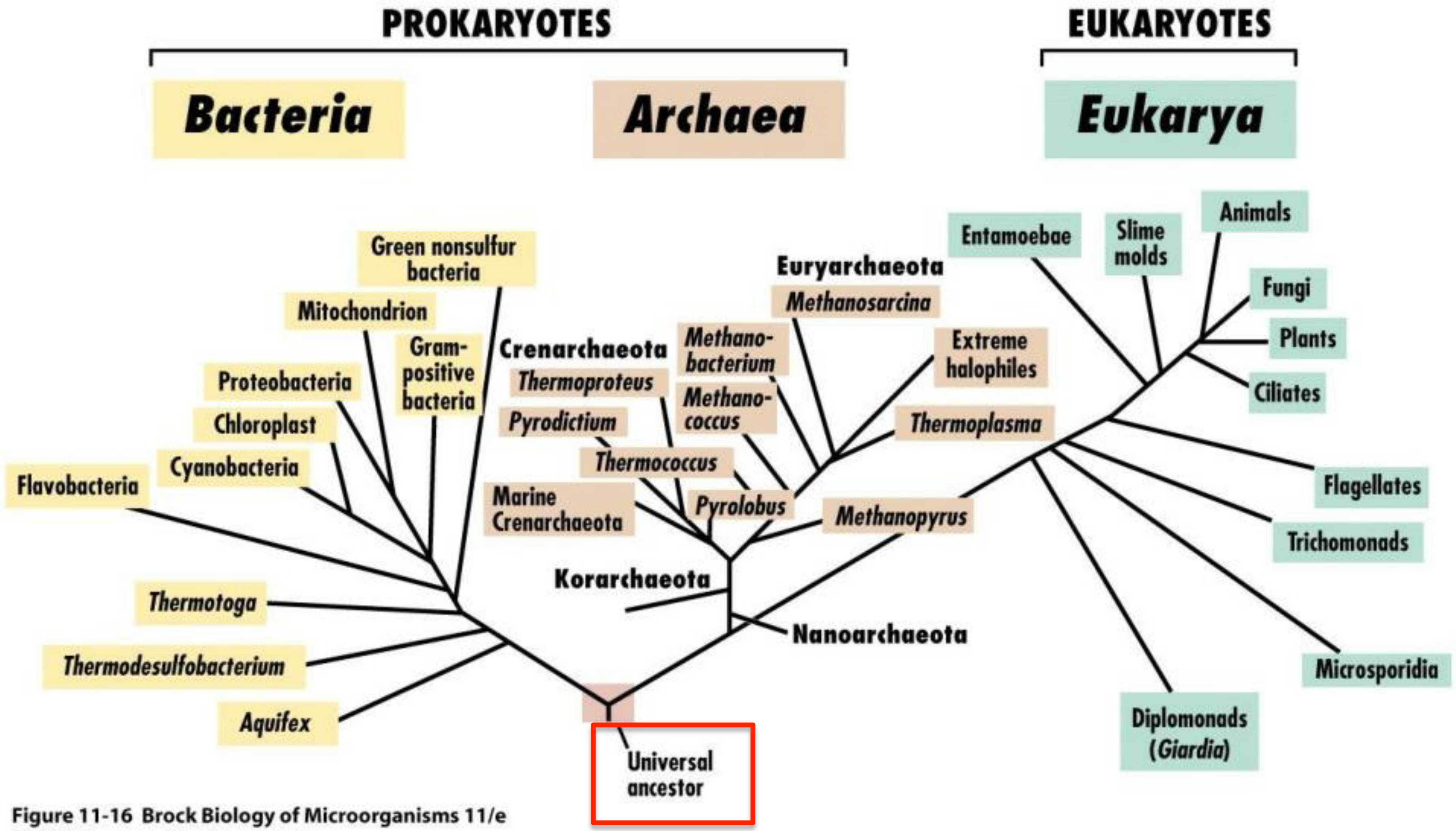
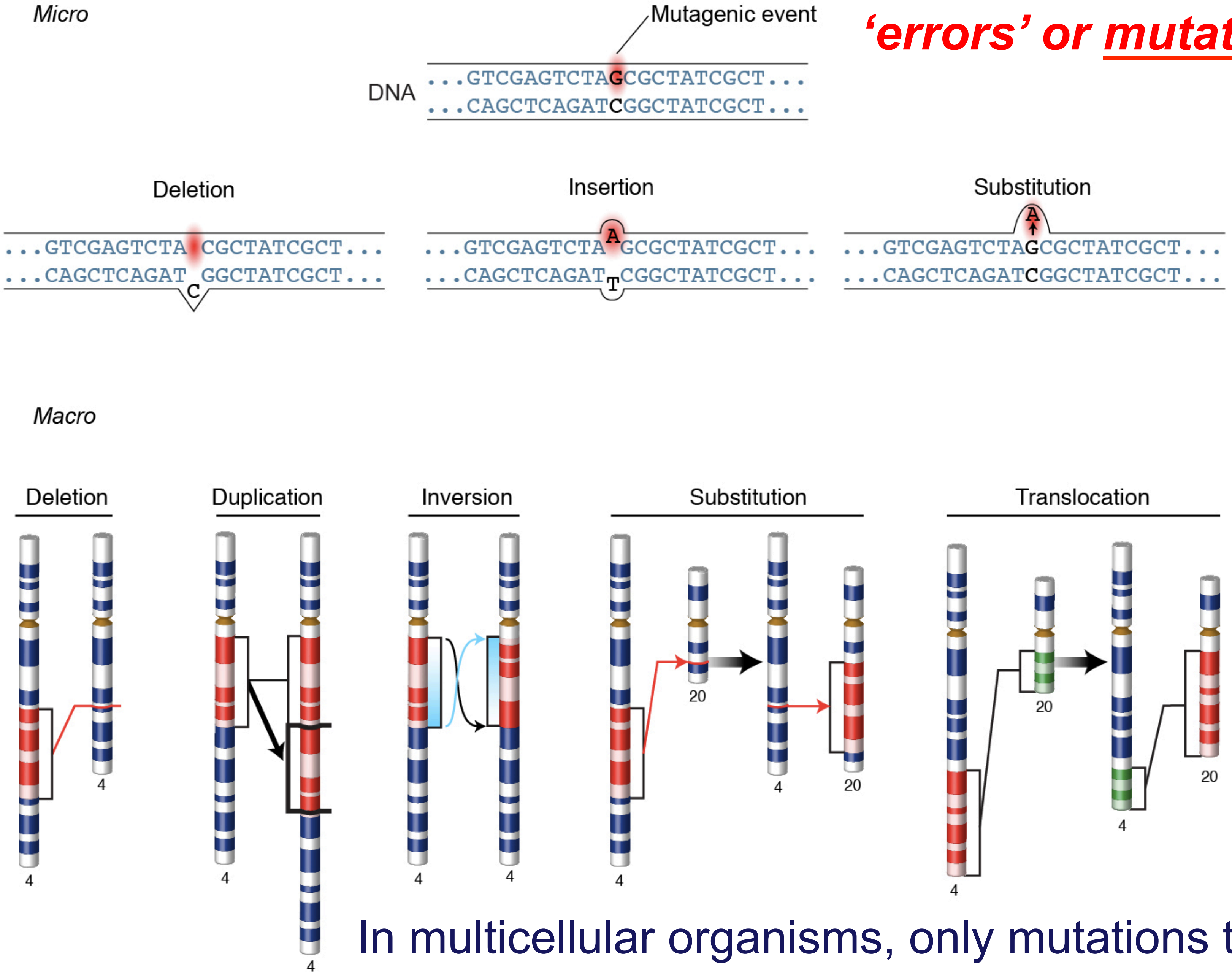


Figure 11-16 Brock Biology of Microorganisms 11/e
© 2006 Pearson Prentice Hall, Inc.

All living organisms belong to one of the three life kingdoms: **bacteria**, **archaea** and **eukarya**, depicting the "Tree of Life."

In the DNA replication ‘errors’ or mutations can occurr


Micro
Mutagenic event
DNA
...GTCGAGTCTAGCGCTATCGCT...
...CAGCTCAGATCGGCTATCGCT...
Deletion
...GTCGAGTCTA CGCTATCGCT...
...CAGCTCAGAT GGCTATCGCT...
C
Insertion
A
...GTCGAGTCTA GCGCTATCGCT...
...CAGCTCAGAT CGGCTATCGCT...
T
Substitution
A
...GTCGAGTCTAGCGCTATCGCT...
...CAGCTCAGATCGGCTATCGCT...
Macro
Deletion
Duplication
Inversion
Substitution
Translocation
4
20

In multicellular organisms, only mutations that occur in the germ cells are relevant to genome evolution

Not all mutations are equally important !

con chi vai nel bus

von chi vai nel bus

coc hiv ain elb us

Not all mutations are equally important !

the hat can fit her

phe hat can fit her

thh atc anf ith er

INsertions/DELetions (INDELs) are usually deleterious mutations, in which case they are removed by negative selection

Not all mutations are equally important !

the hat can fit her

phe hat can fit her

thh atc anf ith er

INsertions/DELetions (INDELs) are usually deleterious mutations, in which case they are removed by negative selection

Not all mutations are equally important !

she can fix the hat

phe can fix the hat

shc anf ixt heh at

INsertions/DELetions (INDELs) are usually deleterious mutations, in which case they are removed by negative selection

Not all mutations are equally important !

she can fix the hat

phe can fix the hat

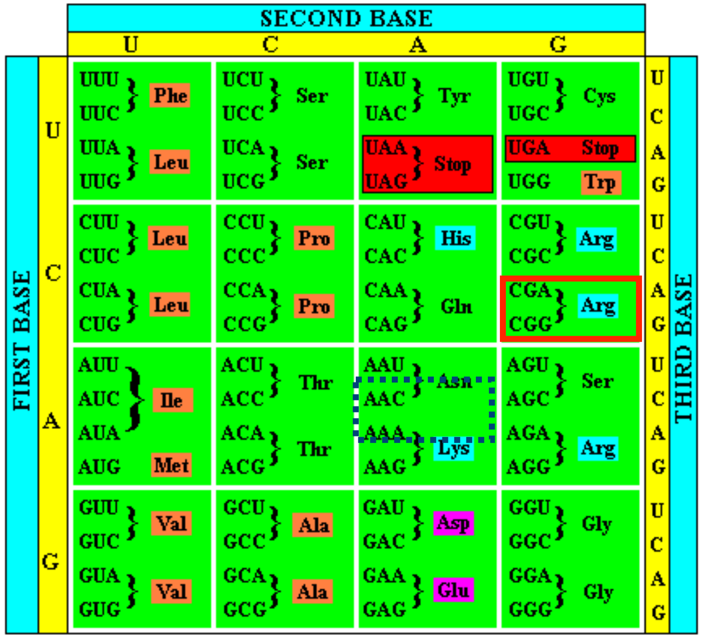
shc anf ixt heh at

Classification of nucleotide substitutions

Synonymous: does not cause a change in the coded amino acid (CG**A** ⇔ CG**G**, both code for an Arg (R))

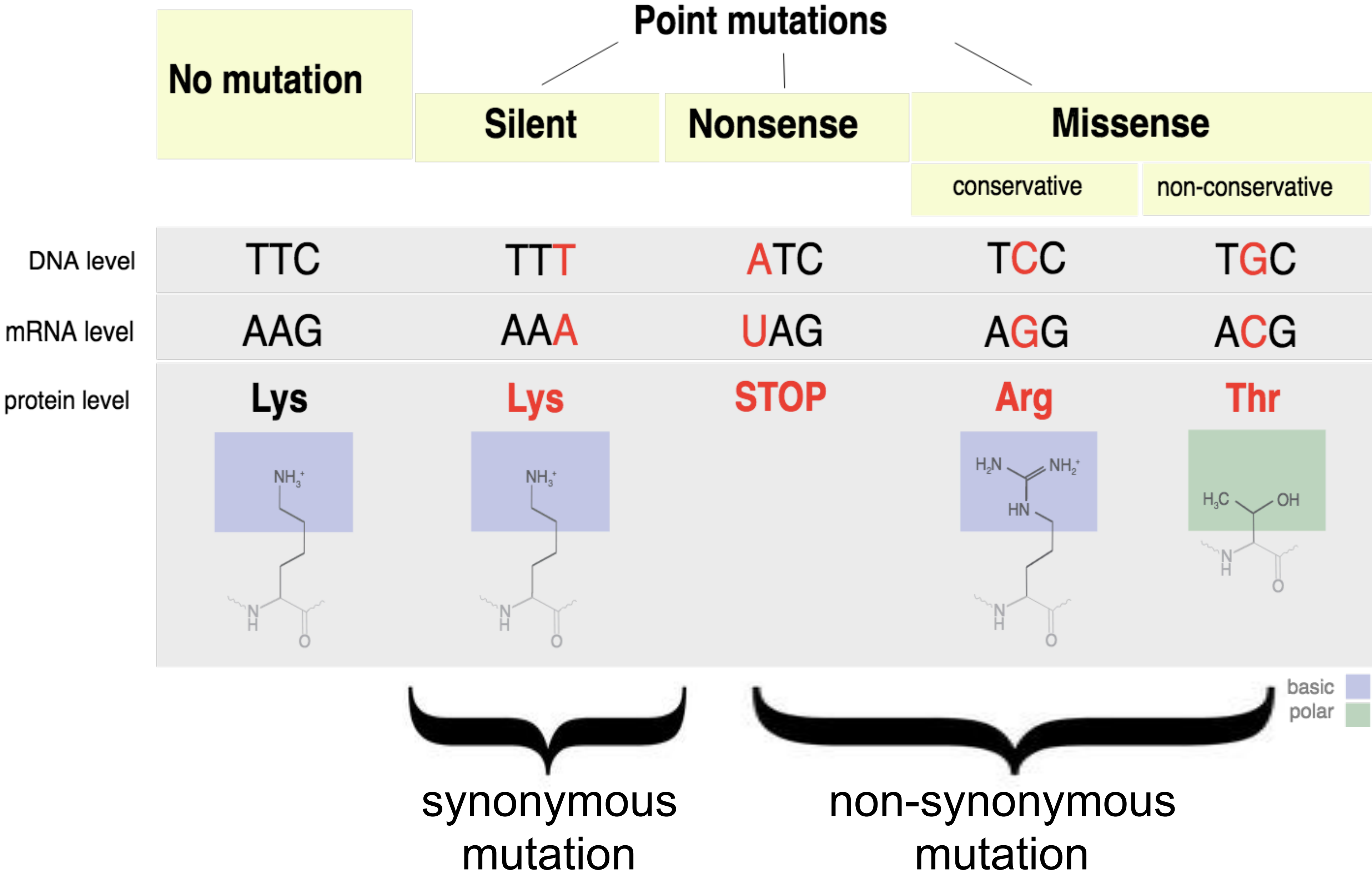
Non-synonymous: causes a change in the coded amino acid (AA**C** ⇔ AA**A**, Asn (N) ⇔ Lys (K))

A synonymous mutation has no effect on the sequence of the coded protein !



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

Classification of nucleotide substitutions



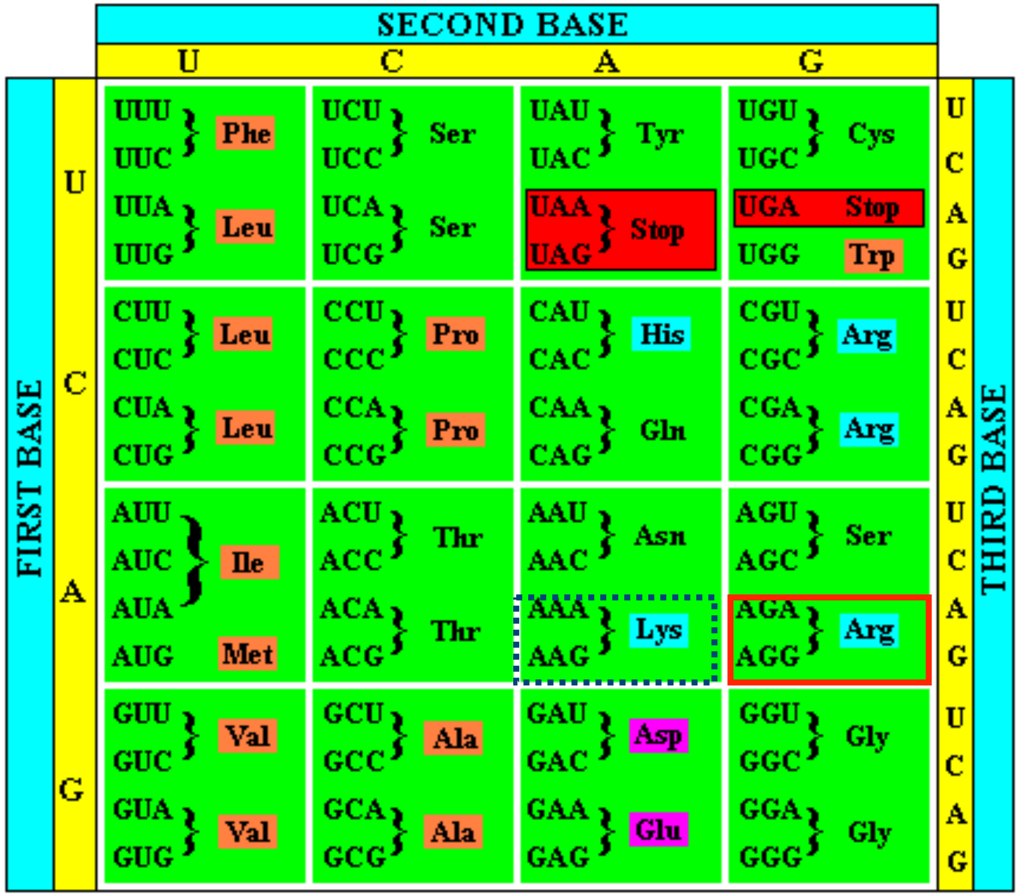
		Point mutations			
	No mutation	Silent	Nonsense	Missense	
				conservative	non-conservative
DNA level	TTC	TTT	ATC	TCC	TGC
mRNA level	AAG	AAA	UAG	AGG	ACG
protein level	Lys	Lys	STOP	Arg	Thr

basic 🟦
polar 🟩

synonymous mutation

non-synonymous mutation

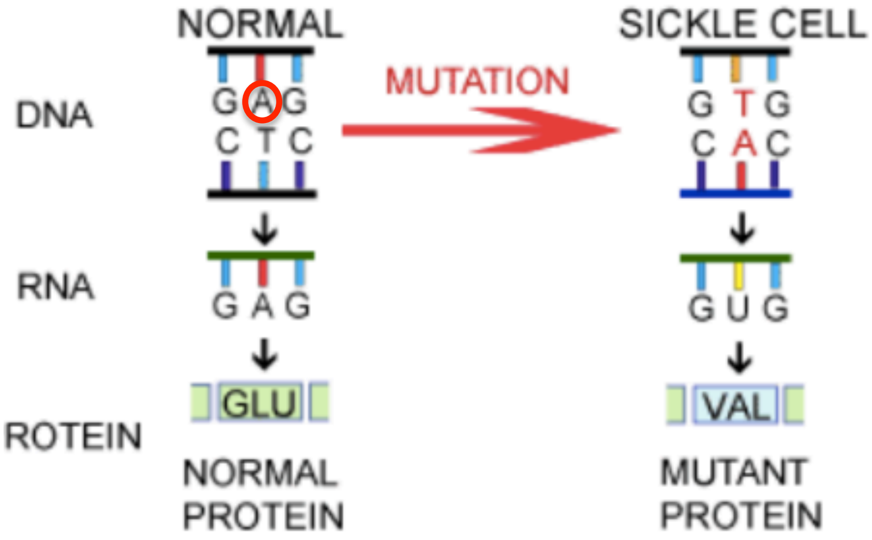
Classification of nucleotide substitutions

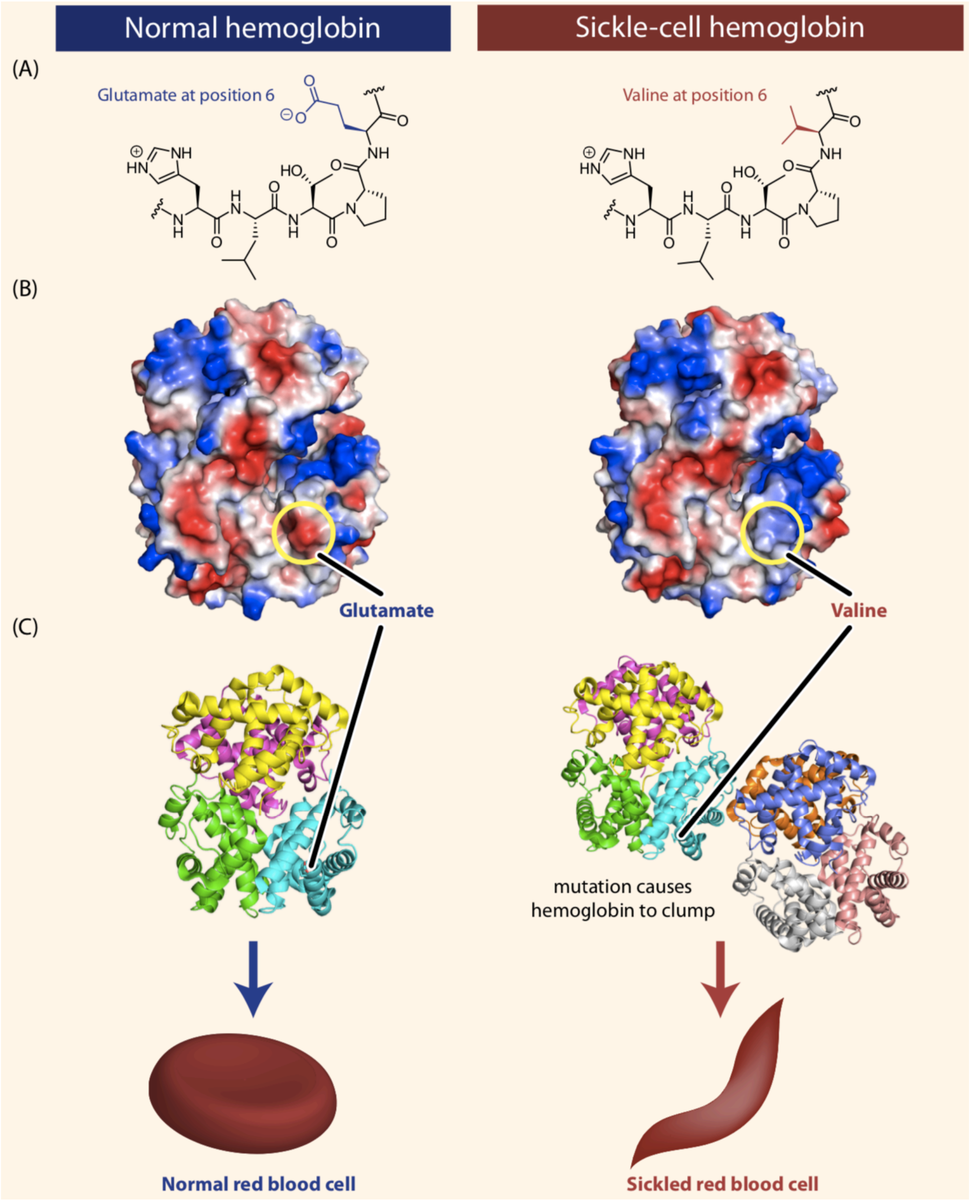


	SECOND BASE								
FIRST BASE	U		C		A		G		THIRD BASE
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG		UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile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

Due to the **degeneracy** of the genetic code – with codons differing by the 3rd position usually translating into the same amino acid –, nearly 70% of substitutions at the 3rd position are synonymous, while all substitutions at the 2nd position and most of the substitutions at the 1st positions are nonsynonymous

Example of a crucial point mutation





Errors in the copying of genetic material (DNA/RNA)

System	Estimated error rate (Mut/$N_{(Pos)}$)
Chemical reaction	0.05-0.1 (5-10/100)
RNA virus (flu, HIV)	10^{-2}-10^{-5}
Prokaryotes (*E.Coli*)	10^{-10}-10^{-11}
Eukaryotes(*H. Sapiens*)	$3*10^{-8}$

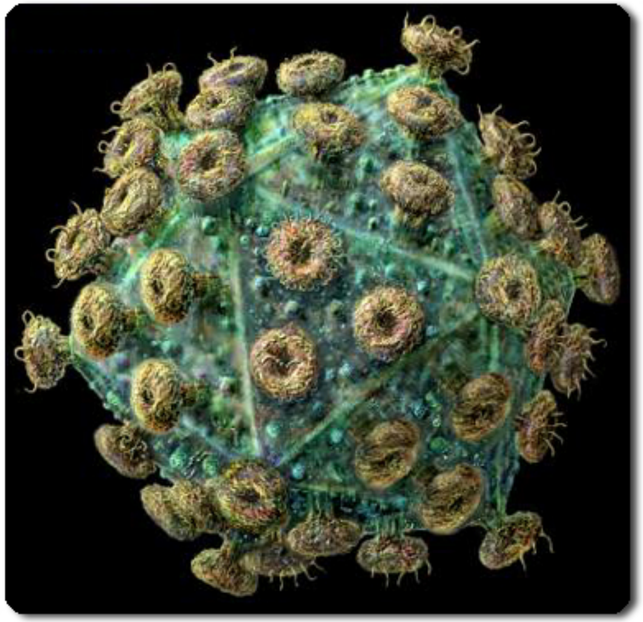
RNA-polymerase

DNA-polymerases
&
repair
mechanisms

Other mutations arise from exposure to excessive UV light, X-rays etc. and for reaction of the DNA with a mutagen chemical

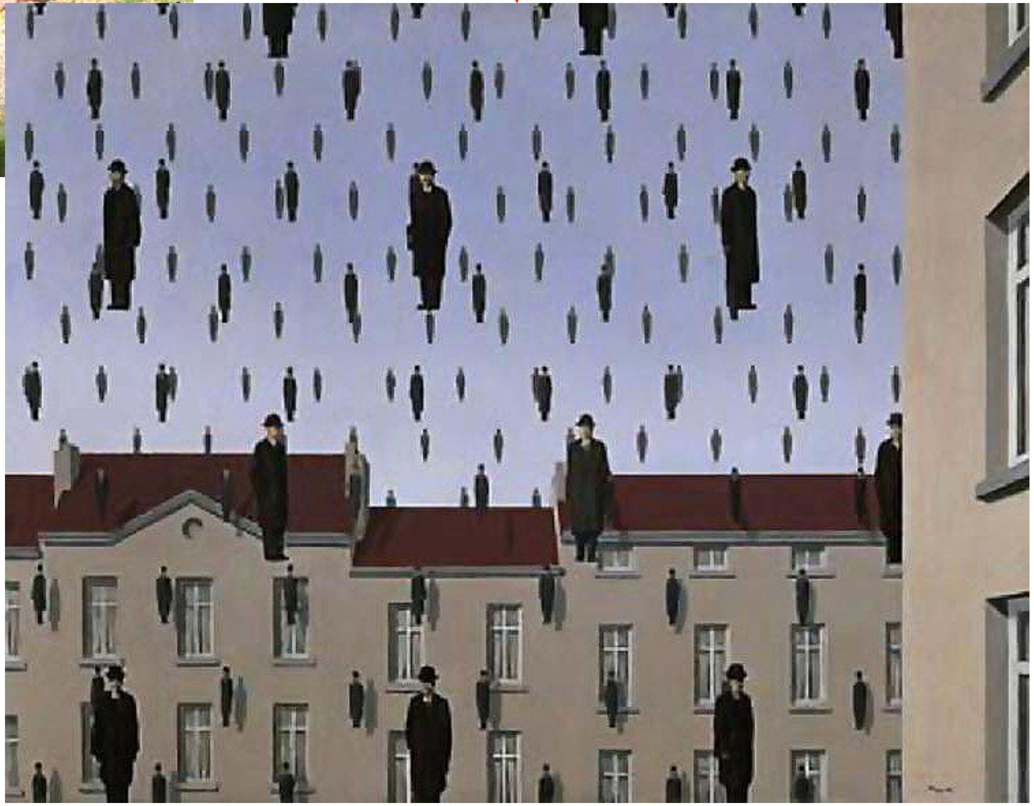
Intra-species variability is responsible for the *survival* of the species itself

The higher the intra-species variability, the higher the probability that positive mutations occurr and that the species can adapt to novel environmental conditions and survive longer



RNA-viruses (influenza, HIV, polio, SARS-CoV-2) are among the organisms most genetically variable and this is why they are so difficult to be treated pharmacologically

Human/chimpanzee inter-species variability = ~1-2 %

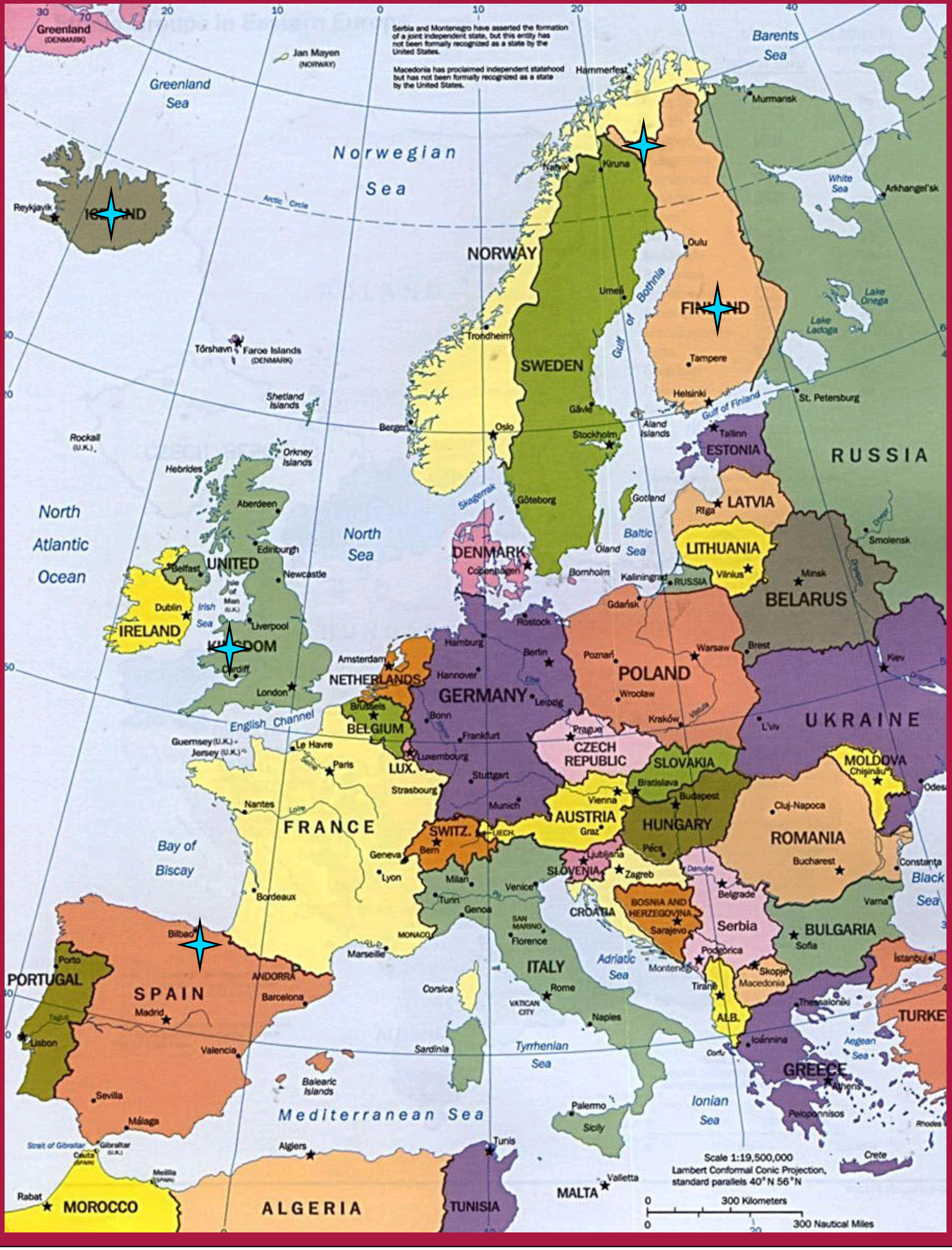


Chimpanzee intra-species variability = ~0,4 %

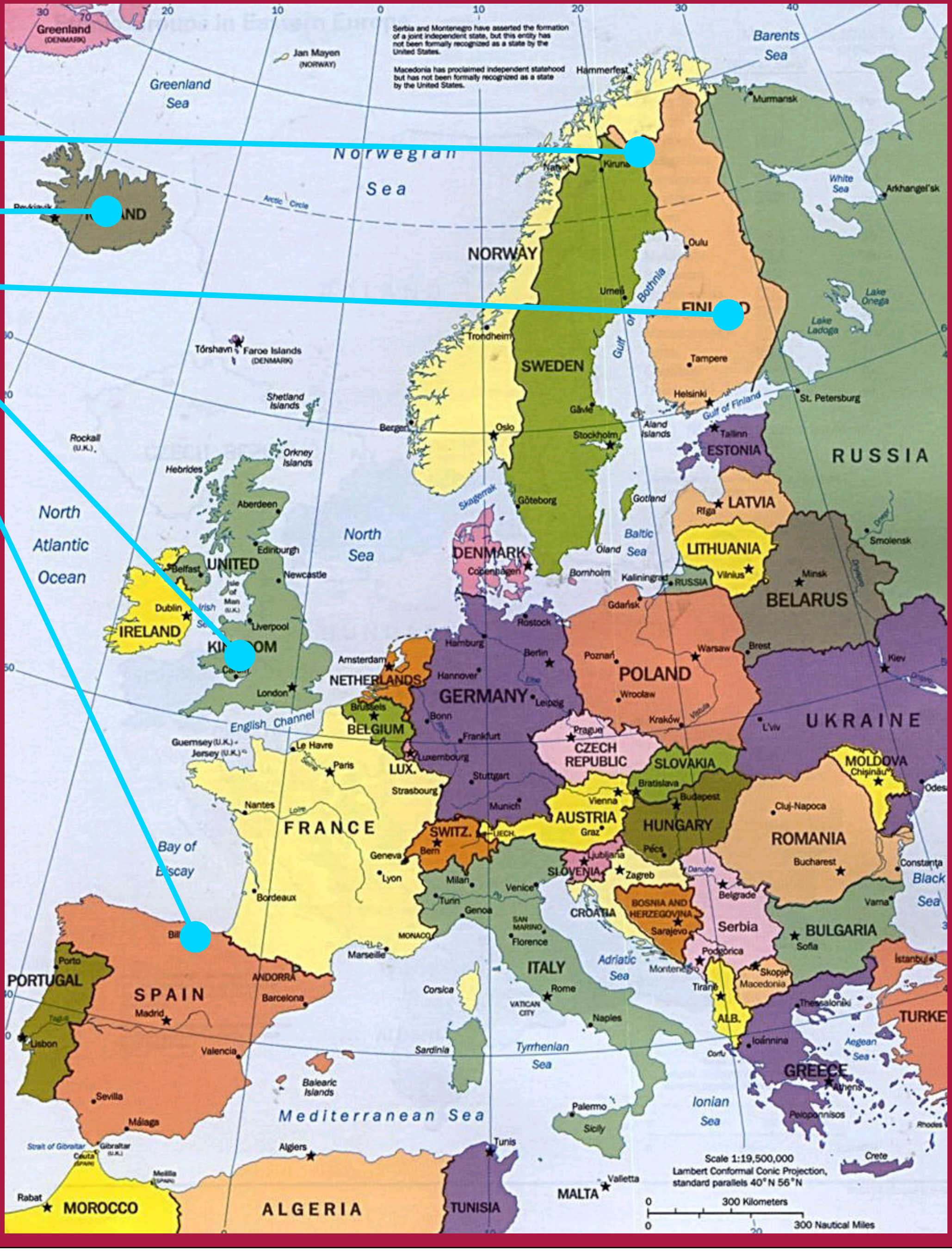
Human intra-species variability = ~0,1 %

Genetically isolated populations in Europe

Lapps
Icelanders
Finns
Welsh
Basque



Genetically isolated populations in Europe



Lapps

Icelanders

Finns

Welsh

Basque

For such populations the intra-species variability is particularly low

This is advantageous for highlighting the effect of specific mutations

Mechanisms of selection

Most life consists of discrete organisms. A *population* is a group of similar organisms that interact, interbreed and compete for resources

Evolution alters the composition and distribution of the gene pools and phenotypes in populations

Within a *population*, individuals with different phenotypes show different success at reproduction

Natural selection – i.e. enhanced reproduction by ‘fitter’ individuals – is the most important mechanism of evolution

Another mechanism of evolution is *genetic drift*, the random change in allelic frequencies, which is not in response to selection

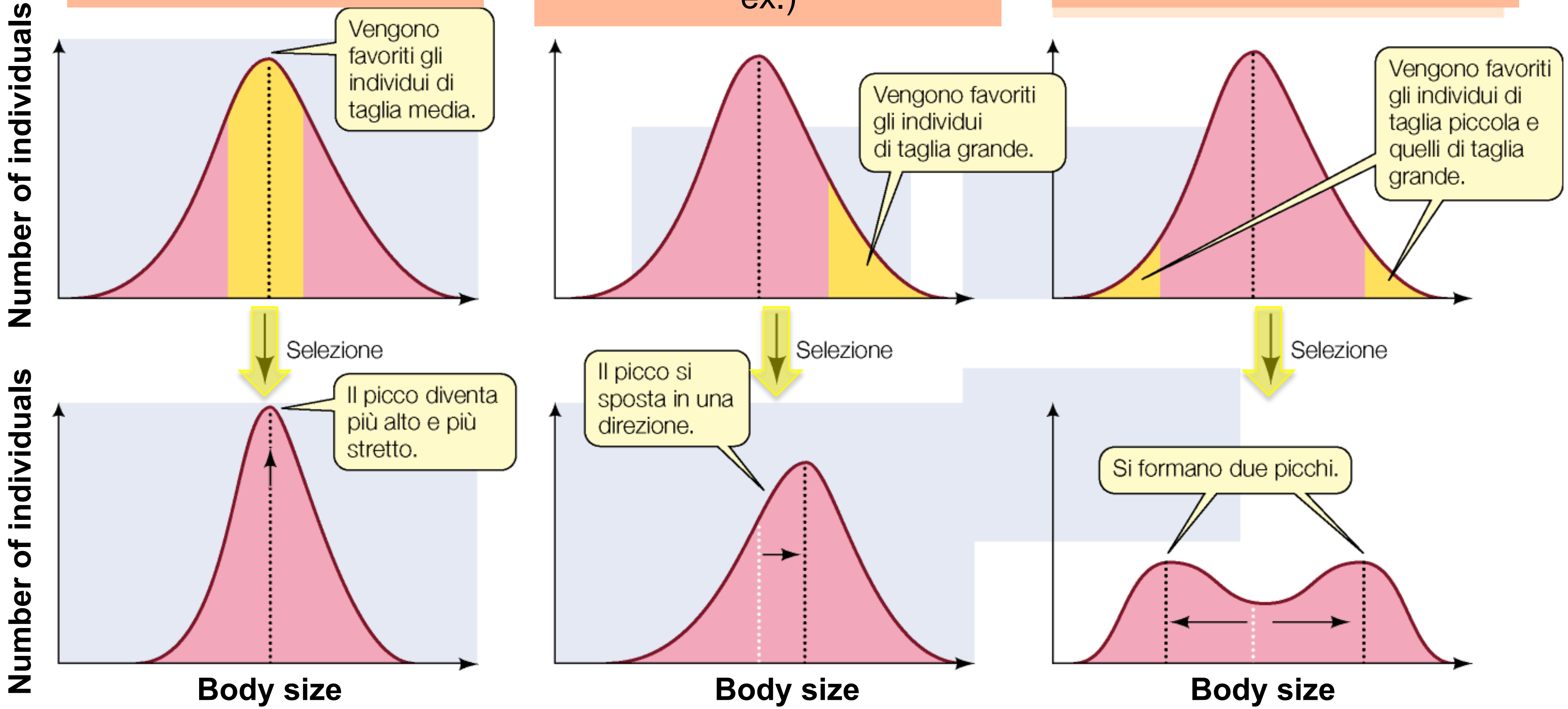
Processes of natural selection

B

Stabilizing selection reduces variability not affecting the average

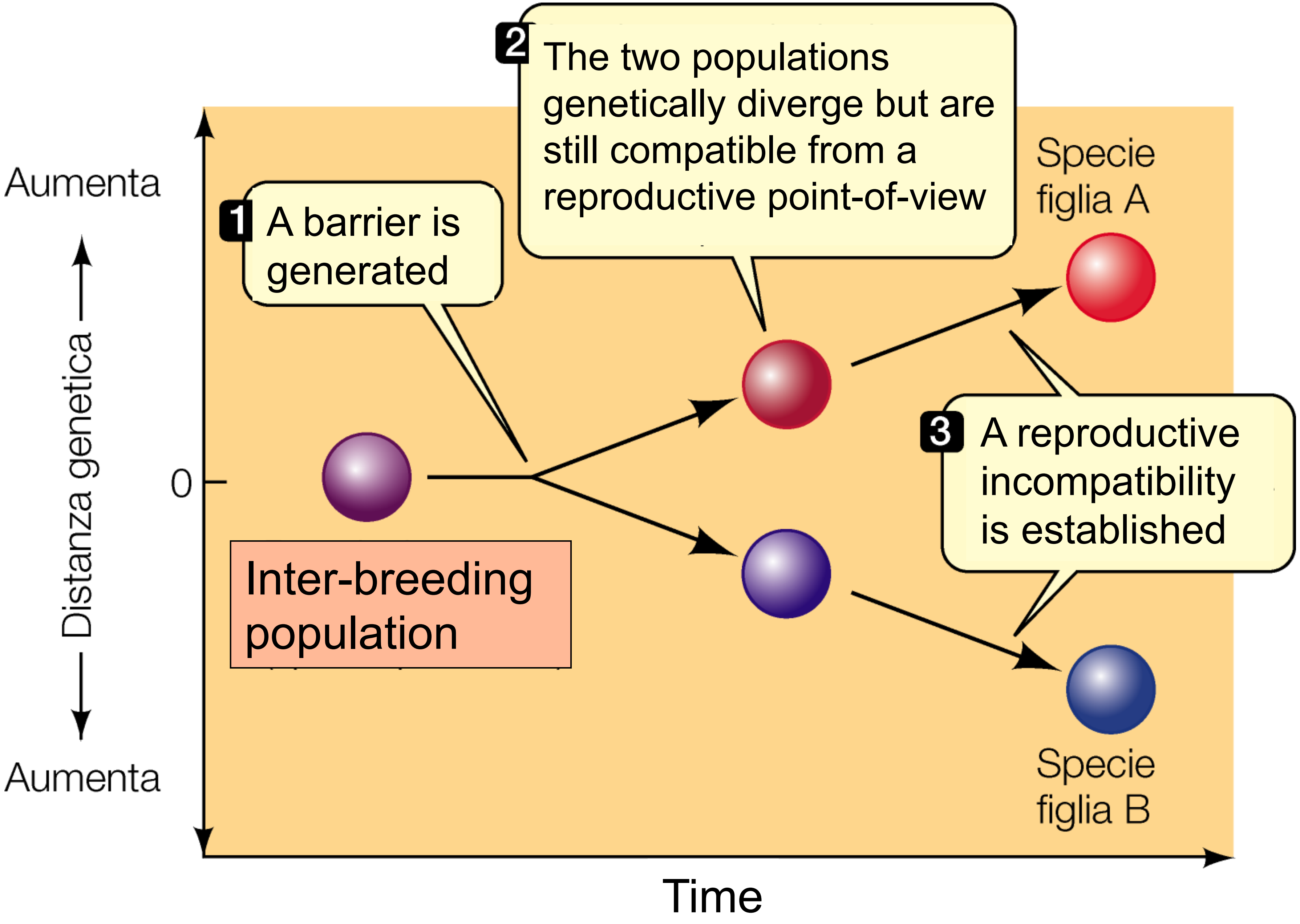
Directional selection changes the average (towards a higher size in this ex.)

Diversifying selection favors the extreme values producing two peaks

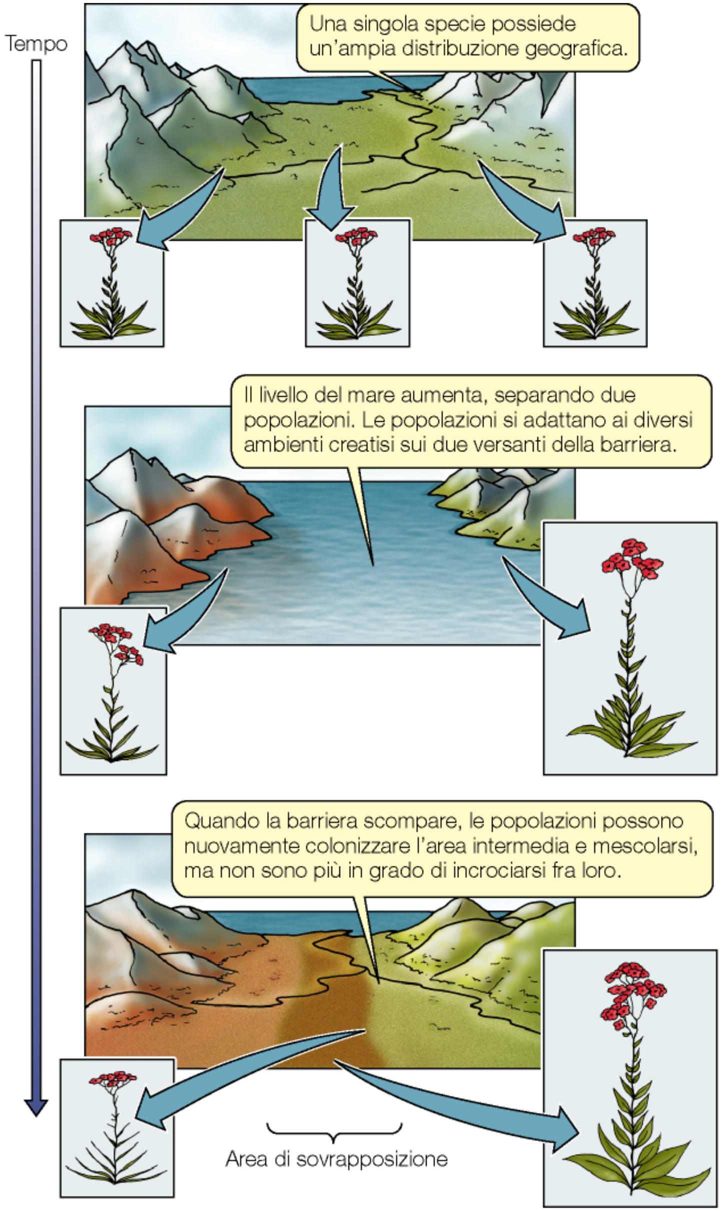


The phenotype here is the body size. Natural selection responds to the environmental conditions

Speciation events


Aumenta
Distanza genetica
0
Aumenta
1 A barrier is generated
2 The two populations genetically diverge but are still compatible from a reproductive point-of-view
Specie figlia A
3 A reproductive incompatibility is established
Inter-breeding population
Specie figlia B
Time

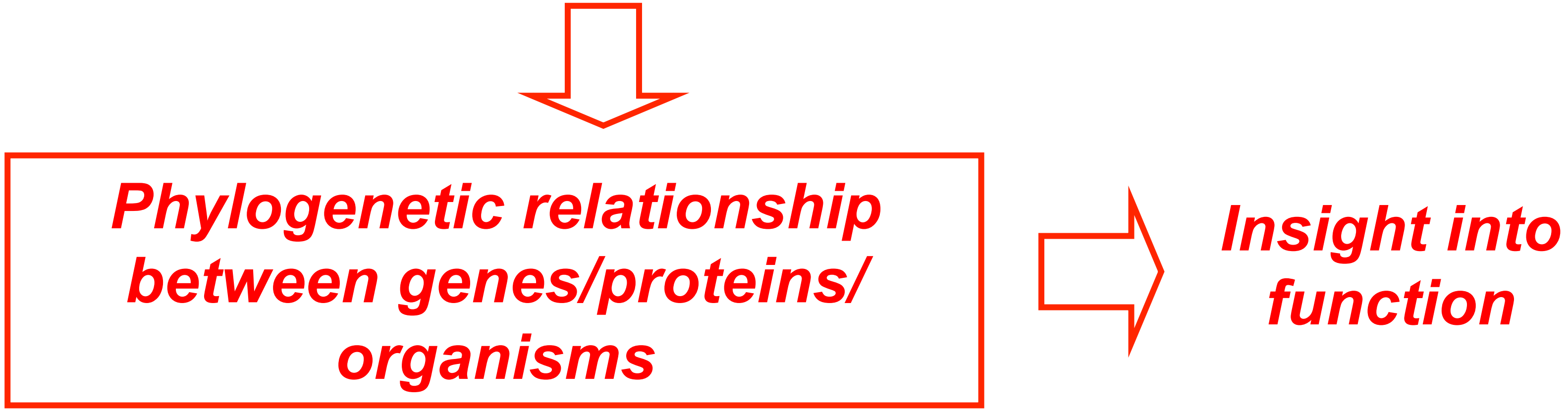
Speciation events


Tempo
Una singola specie possiede un'ampia distribuzione geografica.
Il livello del mare aumenta, separando due popolazioni. Le popolazioni si adattano ai diversi ambienti creatisi sui due versanti della barriera.
Quando la barriera scompare, le popolazioni possono nuovamente colonizzare l'area intermedia e mescolarsi, ma non sono più in grado di incrociarsi fra loro.
Area di sovrapposizione

Example of speciation due to geographical isolation

Bioinformatics and evolution

- Bioinformatics searches for and uses the molecular ***record*** of evolution, provided by results of the **genotypic changes**
- The closer are two species evolutionarily, the more similar are the corresponding **genomic sequences** and their **expression products** (i.e. proteins)
- Whereas sequences have undergone so large variations that they cannot be detected anymore, the corresponding **3D structures of proteins** may have preserved a significant similarity



Phylogenetic relationship between genes/proteins/ organisms

Insight into function

Lessons 1&2. Content

1. Introduction to bioinformatics. Multidisciplinary science, open to multiple applications.

2. DNA: sequence, structure, replication and translation. Contains complex information, only a small portion of it is translated to proteins. Its 3D structure is crucial for replication.

3. Genomes: evolution and information. Evolution has collected over time a huge amount of information. Results of evolution thus do not correspond to random probabilities!