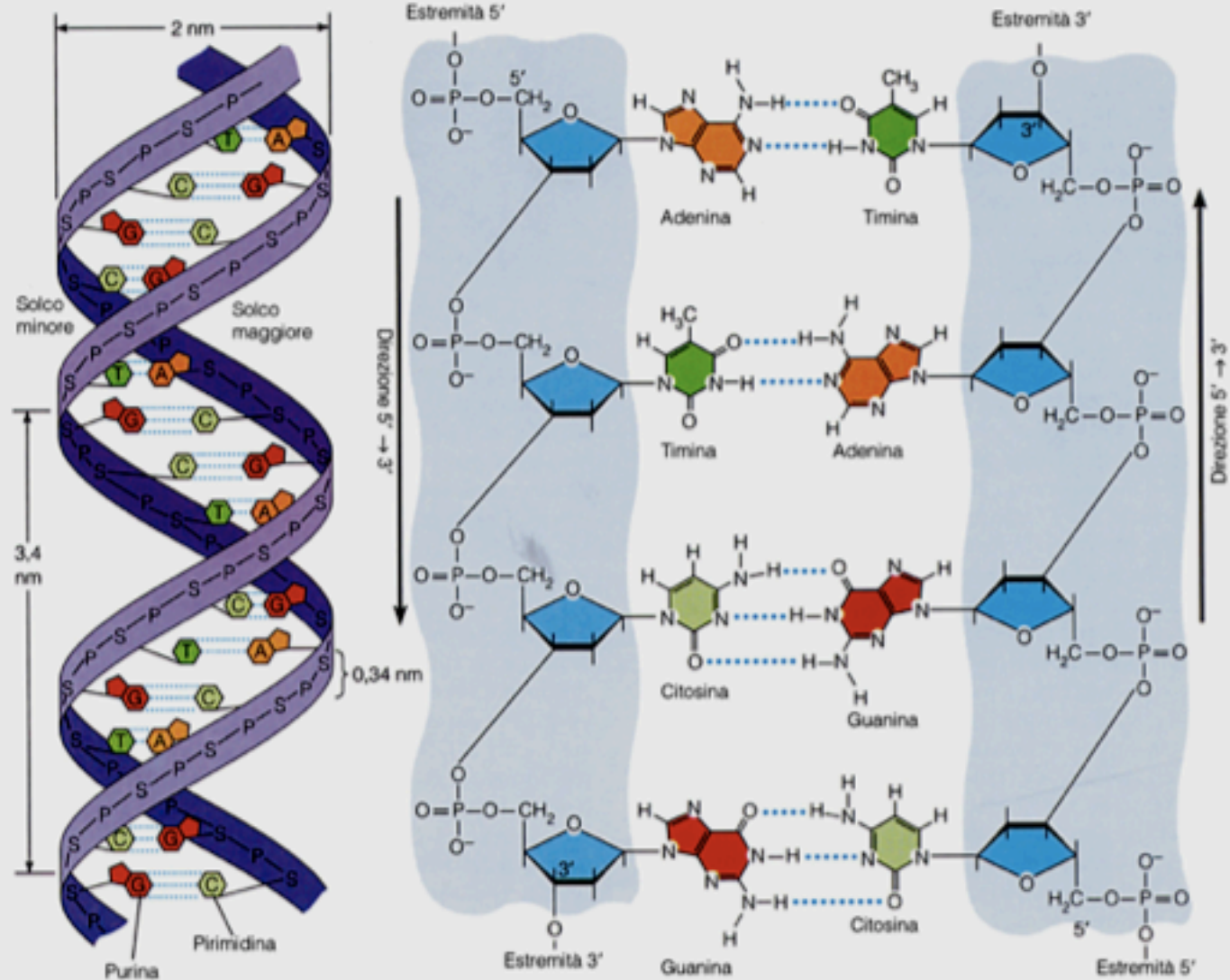


Genes and genomes



(a) Doppia elica

(b) Orientamento antiparallelo dei filamenti

Figura 15-4

Modern genomes

THE CENTRAL DOGMA

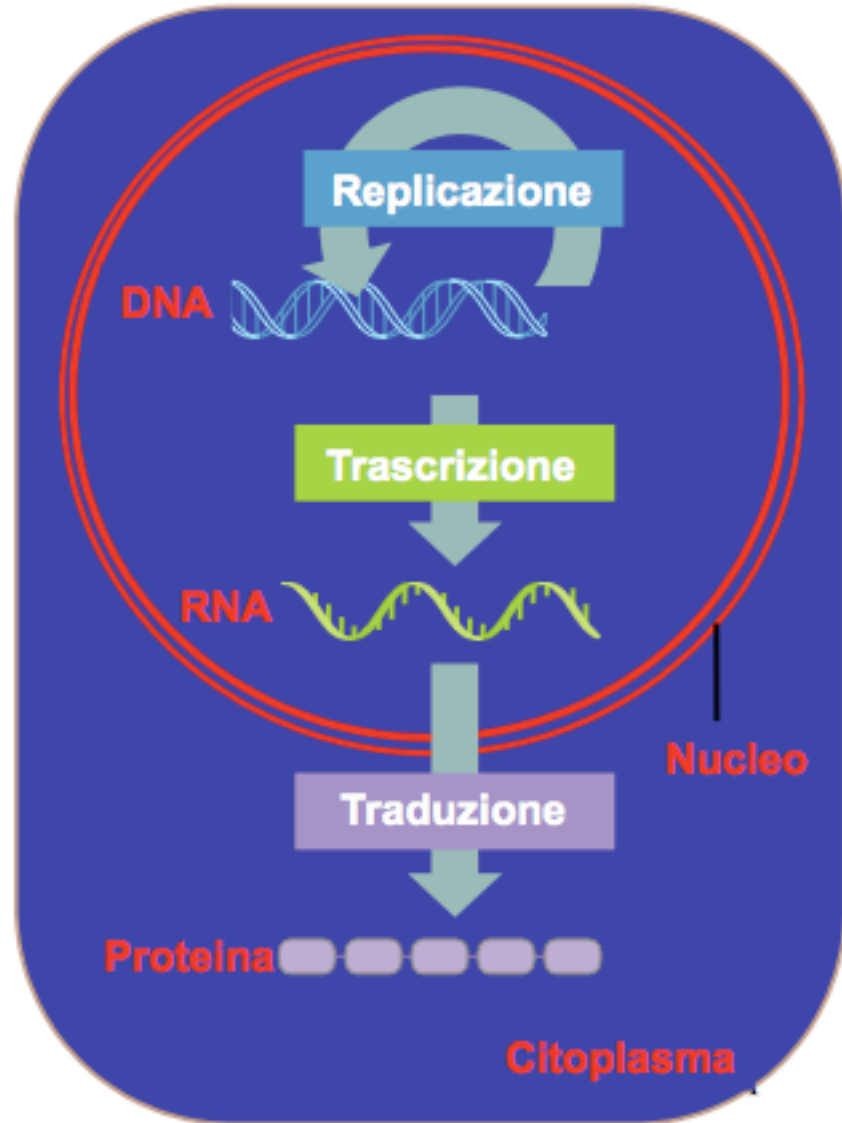
DNA carries the genetic information

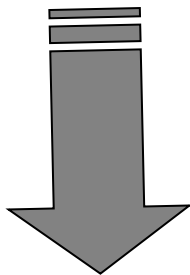
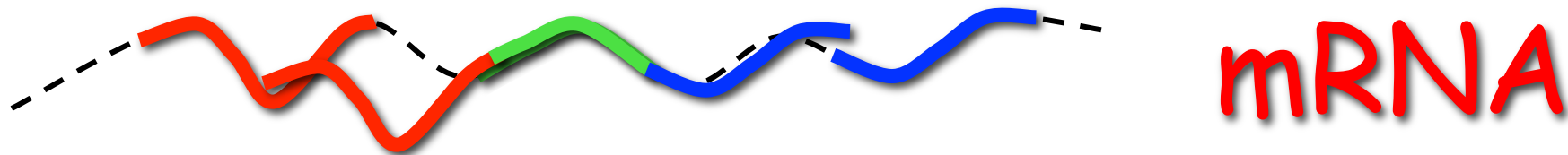
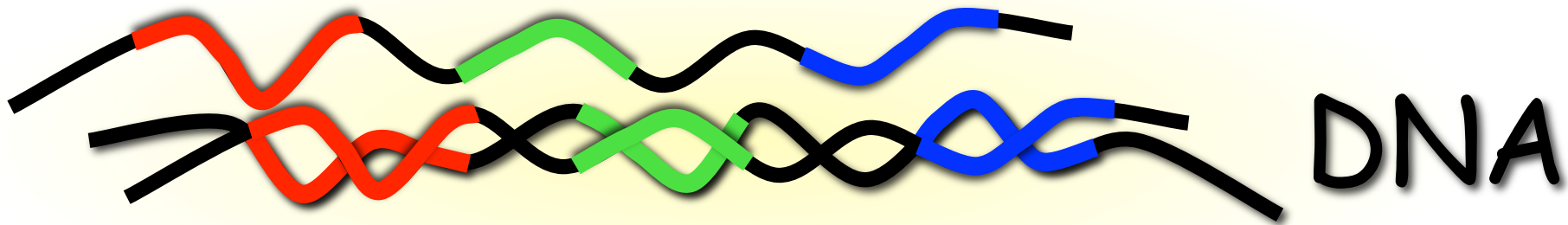
The replication process makes new DNA molecules

Transcription makes RNA by Copying DNA

RNA makes proteins

DNA → RNA → proteins

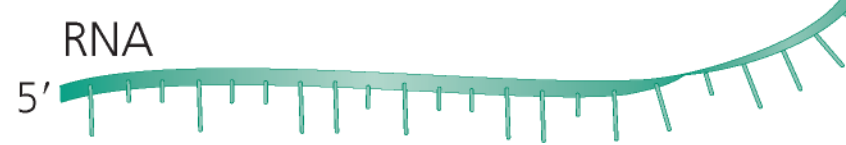
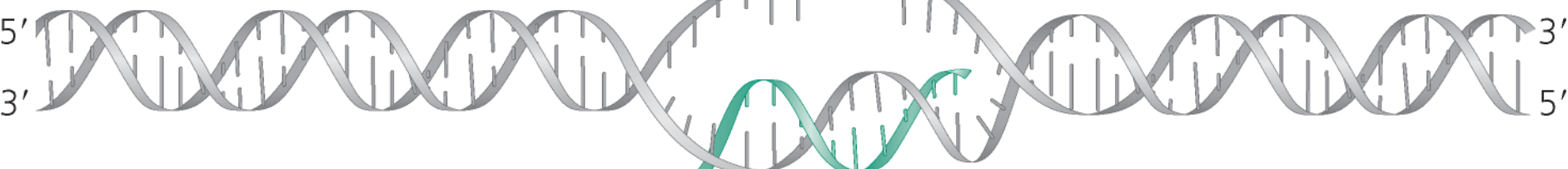




proteins

Transcription

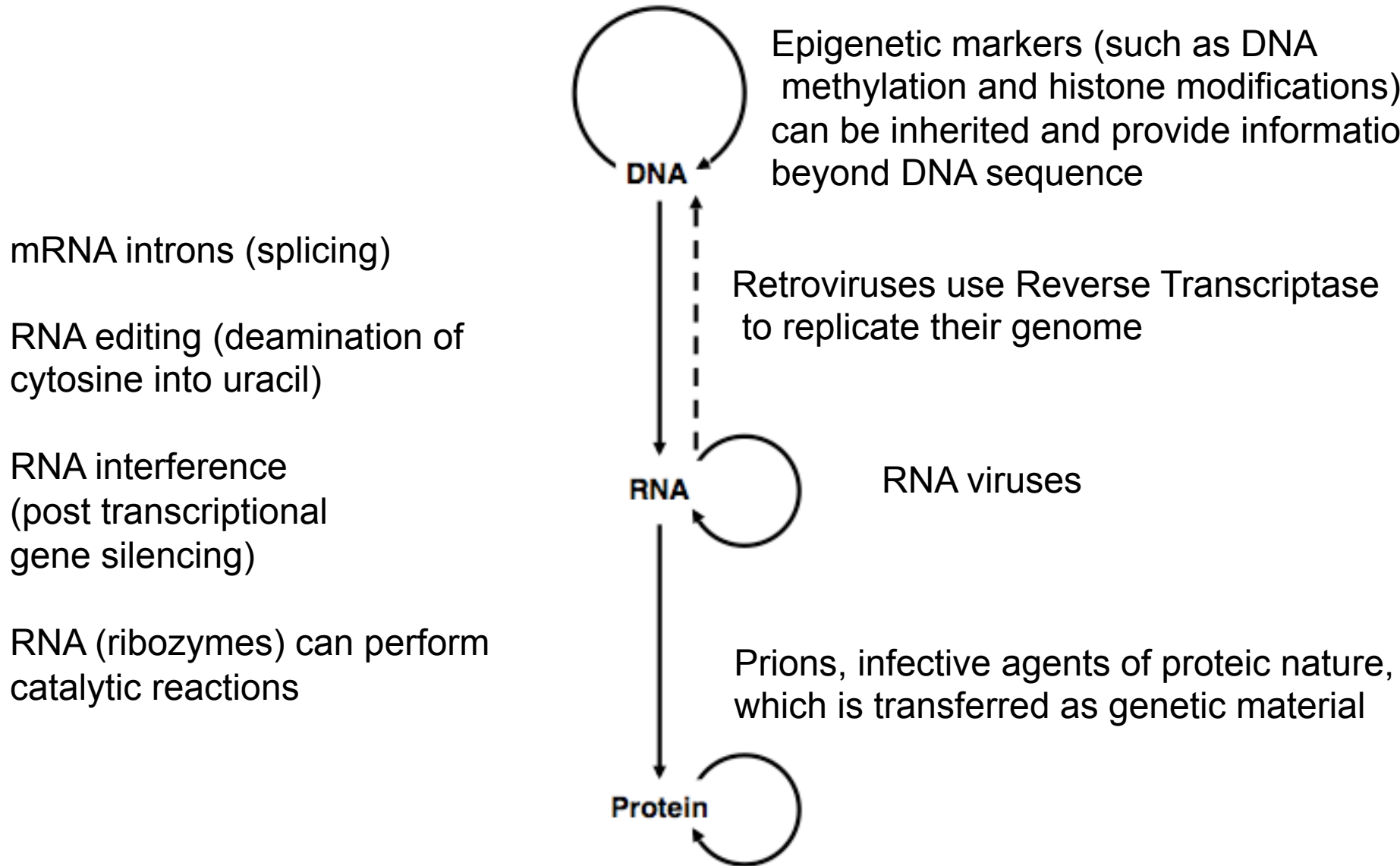
DNA duplex



filamento stampo



EXEPTIONS TO THE CENTRAL DOGMA

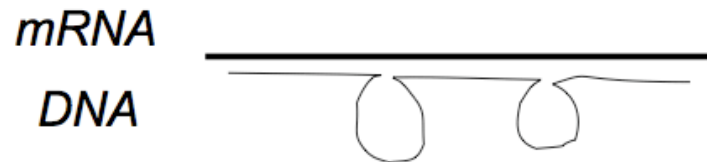


Discovery of the eukaryotic gene structure: exons and introns

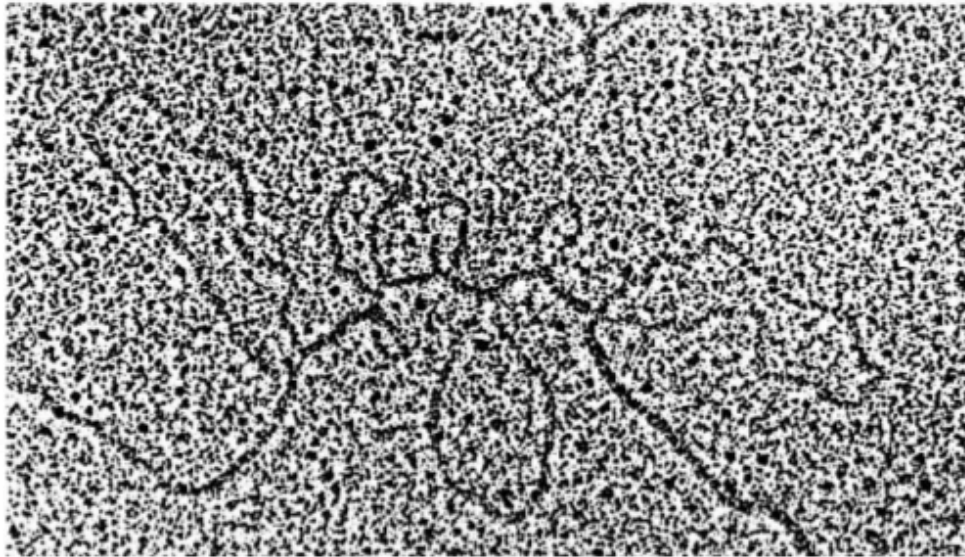
Eukaryotic genes contain sequences that are maintained in the mRNA but they also have sequences, called introns, that are not retained in the mRNA

P. Sharp / Roberts (anni 1977-1978)

R-loop experiments RNA/DNA hybridization and electron microscope analysis



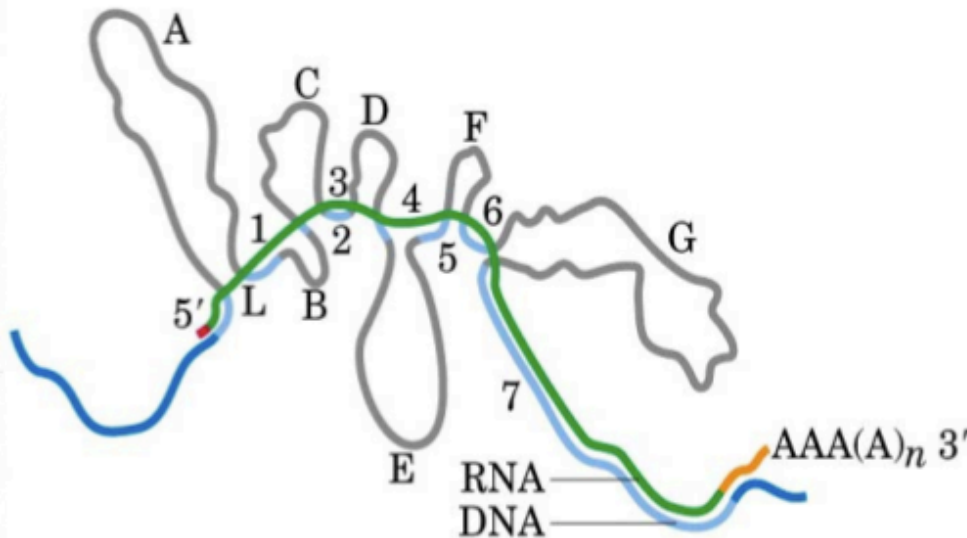
Gene dell' ovalbumina di pollo ibridato con il suo mRNA e visualizzato al microscopio elettronico



Canonical mammalian gene:
16kb with 7-8 esoni

Exons: short (100-200 bp)

Introns: very long (also 100 kb)



structure of the eukaryotic gene

promoter



exon

intron

exon

AAGCTGGCTAGCGCCCAATGGCTAGCTTACAGgtaacacgtggtcttttaaattctccagGTAATACTTTCTGAATTCagtg

pre-mRNA

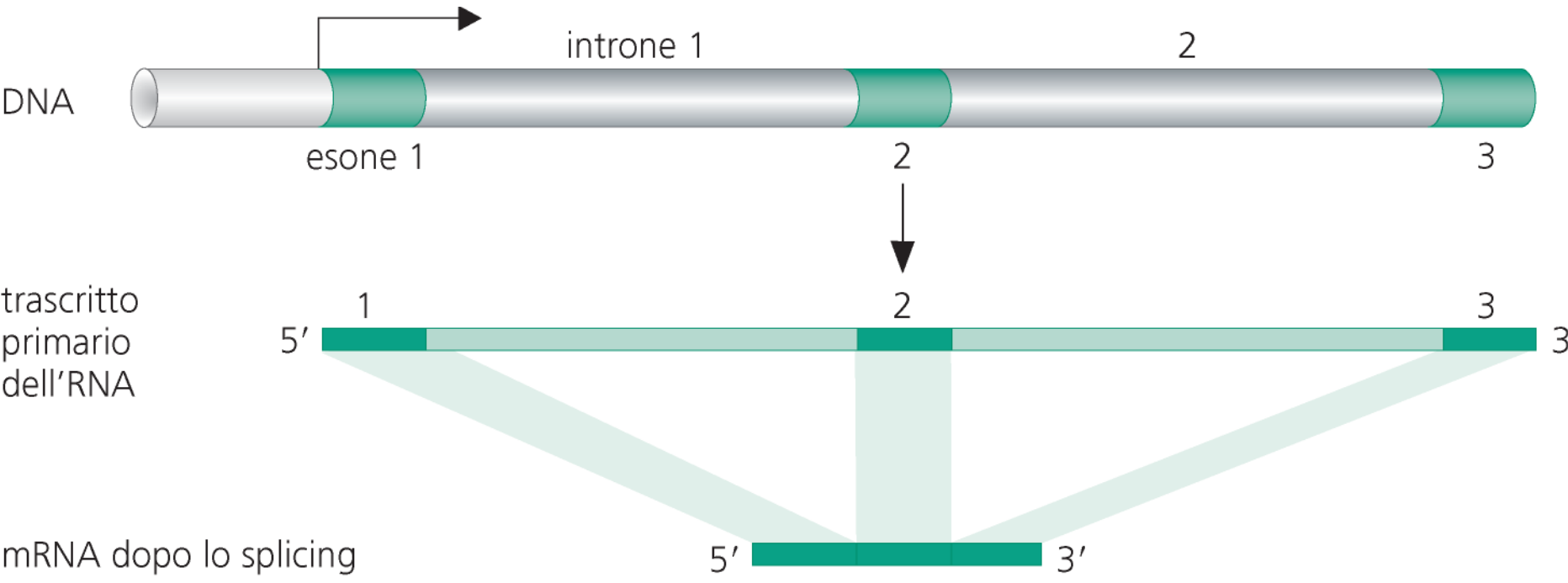
Met Ala Ser Leu Gln Val Asn Thr Phe ter
GCCCAAUGGCUAGCUUACAGguaacacguggucuaaauccagGUAAAUACUUUCUGAAU

mRNA

Met Ala Ser Leu Gln Val Asn Thr Phe ter
GCCCAAUGGCUAGCUUACAGGUAAAUACUUUCUGAAUU
C

proteina





The GENOMIC ERA

– at the beginning of the XXI century, one of the major question was:

how many genes in the human genome?

The huge popular interest in counting the number of genes present in the human genome led even to a public wager named Gene Sweepstake, with an extensive media coverage (nyt Wade 2003)



Procarioti

	<i>genoma (Mb)</i>	<i>N° geni</i>
E.coli	4.6	4.400
S.pneumoniae	2.2	2.300

Funghi

S.cerevisiae	12	5.800
--------------	----	-------

Protozoi

T.thermophila	125	27.000
---------------	-----	--------

Invertebrati

C.elegans	103	20.000
D.melanogaster	180	14.700

Vertebrati

M.musculus	2.600	22.000
H.Sapiens	3.200	20.000

Piante

A.thaliana	120	26.000
Z.mais	2.200	45.000

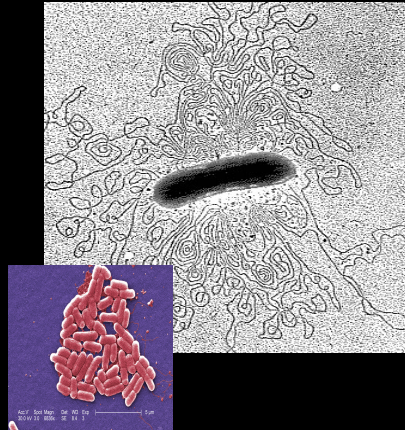
Eucarioti

Genome size and organism complexity

E. coli

C. elegans

H. sapiens

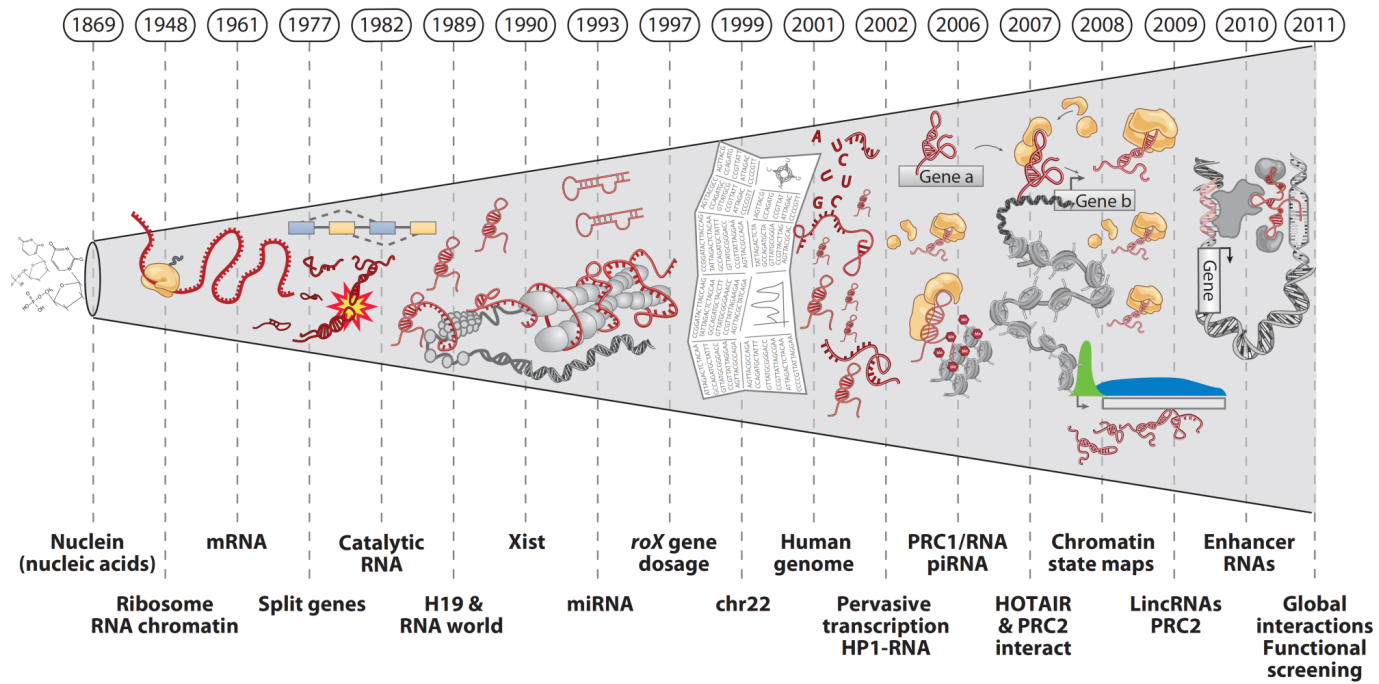


Genome	5×10^6 bp	1×10^8 bp	3×10^9 bp
Chromosomes	1	6	23
Coding genes	6692	20541	21995
ncDNA	5%	60%	98%
non-coding RNA genes	15	23136	ca. 40000
miRNAs	0	224	4274
pseudogenes	21	1522	10616

The GENOMIC ERA

how many genes in the human genome?

Lander *et al.*, 2001
Venter *et al.*, 2001

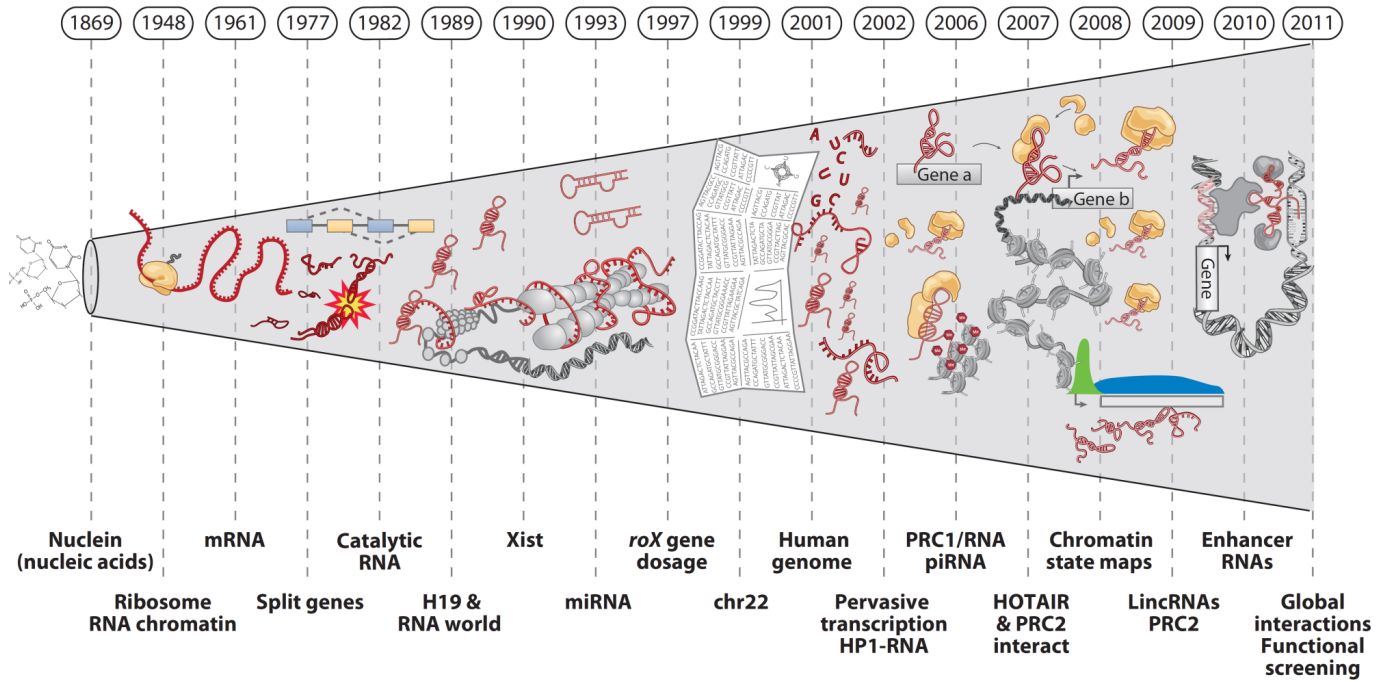


Rinn & Chang Annu Rev Biochem 2012

The Human Genome Project produced the first complete

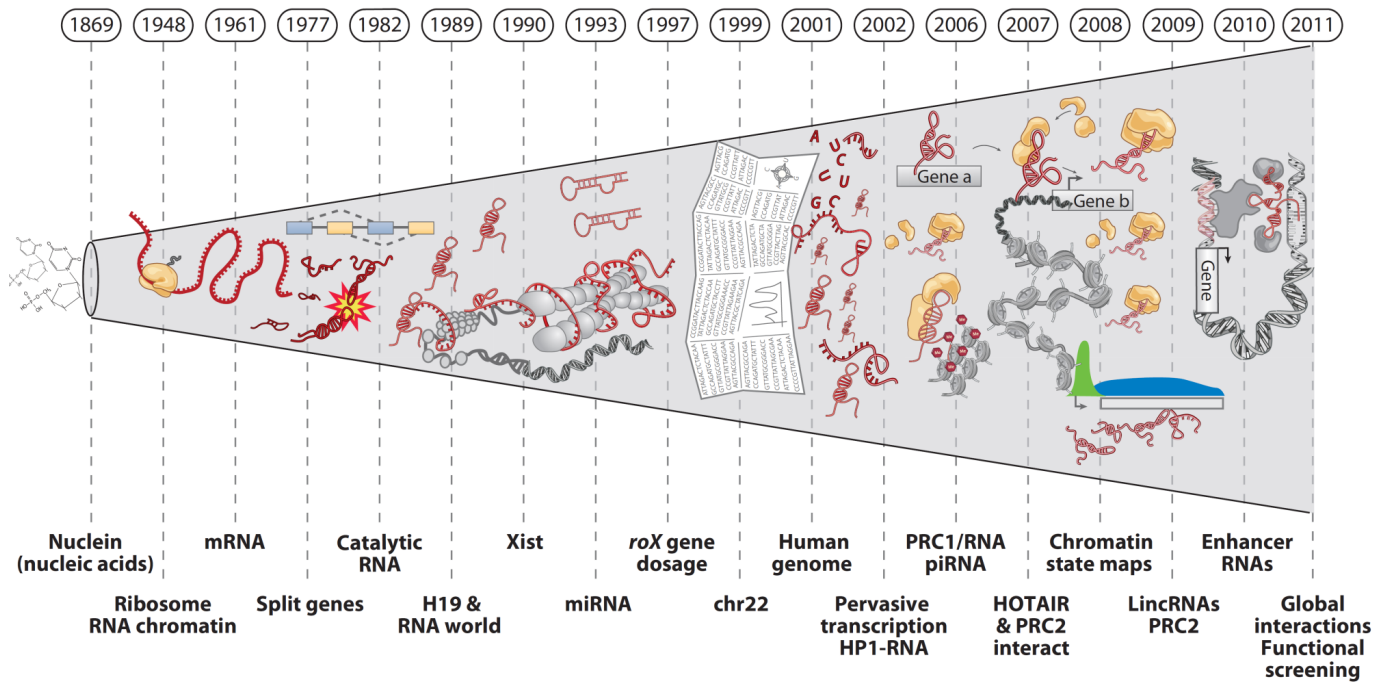
Transcriptome analysis

The FANTOM3 Consortium, 2005



Deep sequencing technologies – identification of low abundance transcripts

454 Pyrosequencing, Illumina,
SOLiD, Heliscope and RNA-Seq



Transcriptome analysis

ENCODE

ARTICLE

doi:10.1038/nature11247

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

The Encyclopedia of DNA Elements (ENCODE) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the **human** genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

22000 genes encoding for proteins

FANTOM 5

A promoter level mammalian expression atlas

Alistair R.R. Forrest *et al.*, *submitted*

CAGE analysis of the following libraries:

573 human primary cell samples

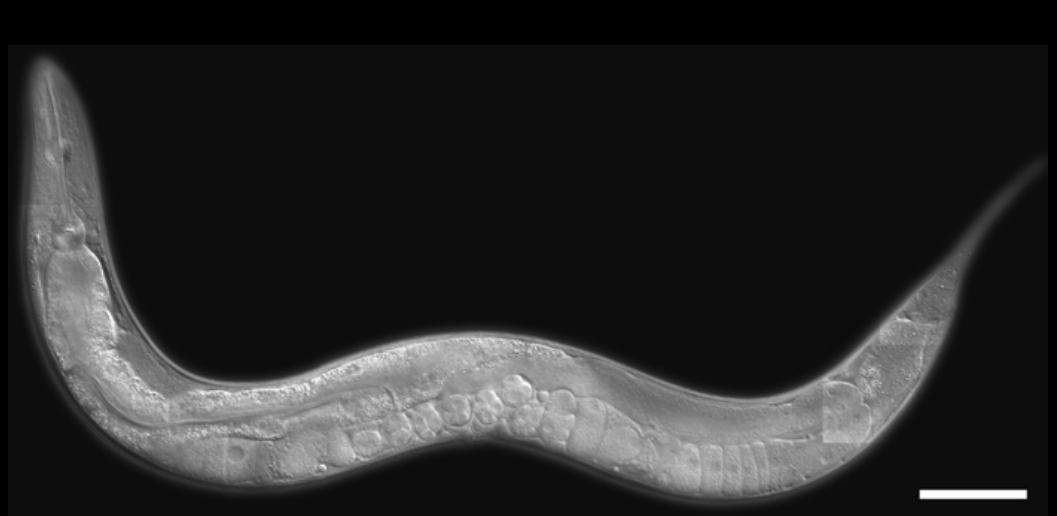
128 mouse primary cell samples

250 different cancer cell lines samples

152 human post-mortem tissues samples

271 mouse developmental tissue samples

The genetic basis of developmental complexity



C.elegans - **1000 cells**
H.sapiens - **10^{14} cells - and 10^{11} neurons!!!**

Both have approximately **20.000 proteins**

- Most of the proteins are orthologous and have similar functions from nematodes to humans, and many are common with yeast.
- **Where is the information that programs our complexity?**

Transcriptome analysis

ENCODE

ARTICLE

doi:10.1038/nature11247

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

The Encyclopedia of DNA Elements (ENCODE) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the **human** genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

22000 genes encoding for proteins

>40000 long non-coding RNAs and growing.....

>50% of the genome is functional

FANTOM5

A promoter level mammalian expression atlas

Alistair R.R. Forrest *et al.*, *submitted*

CAGE analysis of the following libraries:

573 human primary cell samples

128 mouse primary cell samples








250 different cancer cell lines samples

152 human post-mortem tissues samples

271 mouse developmental tissue samples

Although the central role of RNA in cellular functions and organismal evolution has been advocated periodically during the last 50 years, only recently has RNA received a remarkable level of attention from the scientific community.

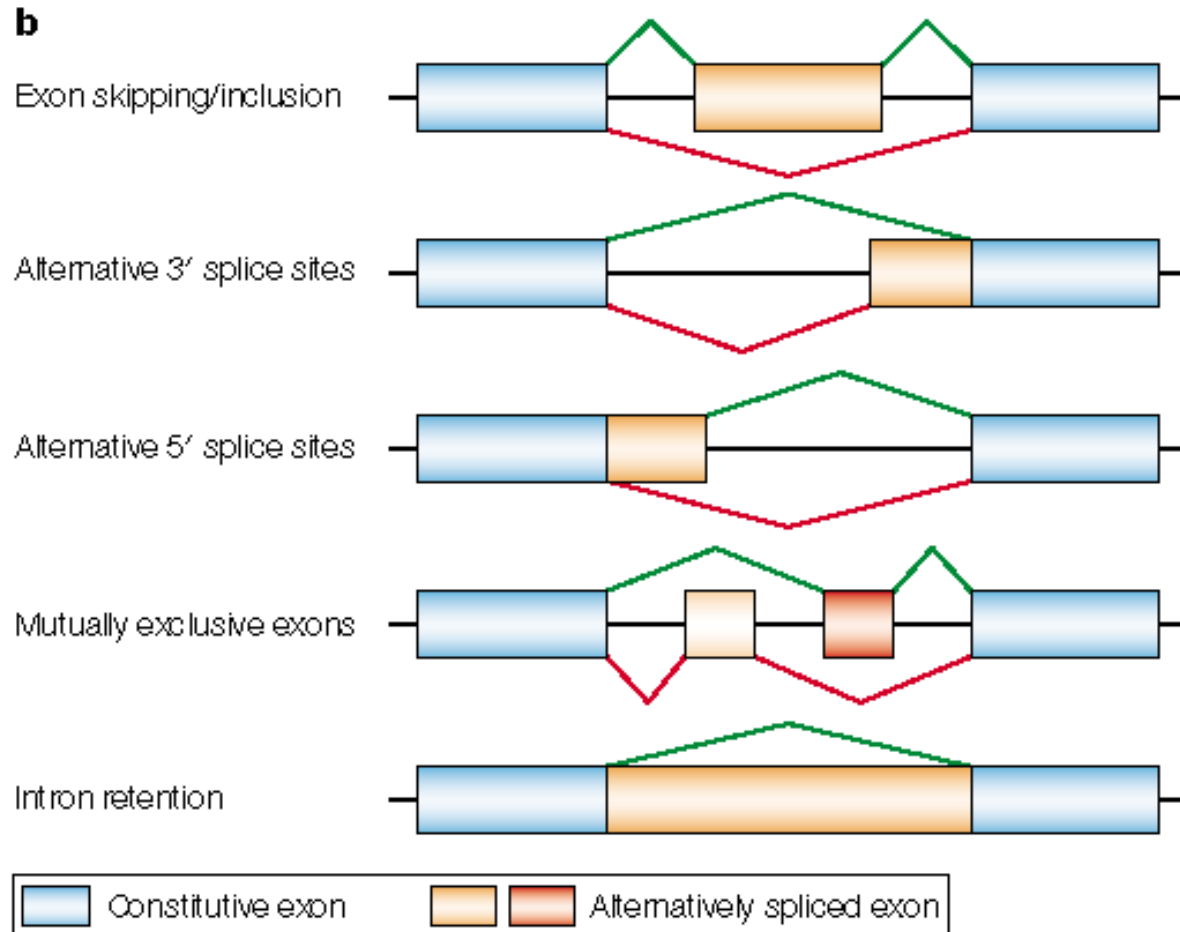
The number of protein coding genes cannot justify the evolutionary gap between eukaryotes

Gene number		~6,5k		~25k
Transcript number		~15k		~25k
		~20k		~31k
		~30k		

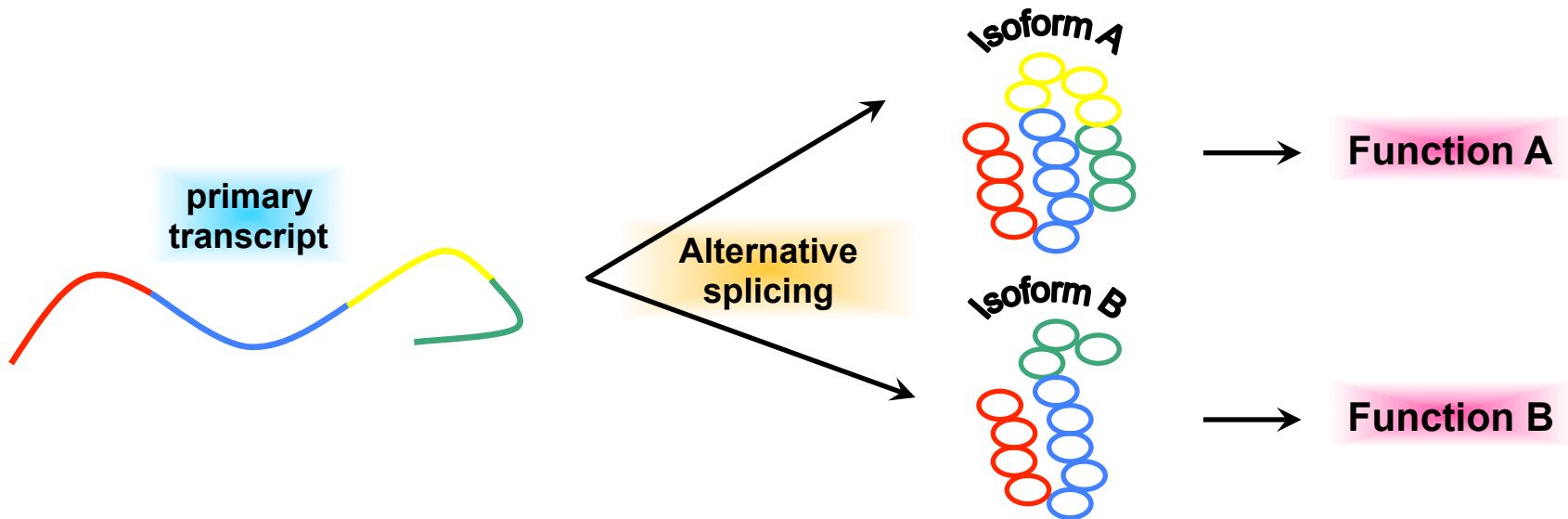
Splicing increases the coding potential of a gene

Several different alternative splicing isoforms from a single gene

Increase of the coding potential through alternative splicing



- Regulation of gene expression -
increase the complexity by increasing the combinatorial
use of exons by alternative splicing

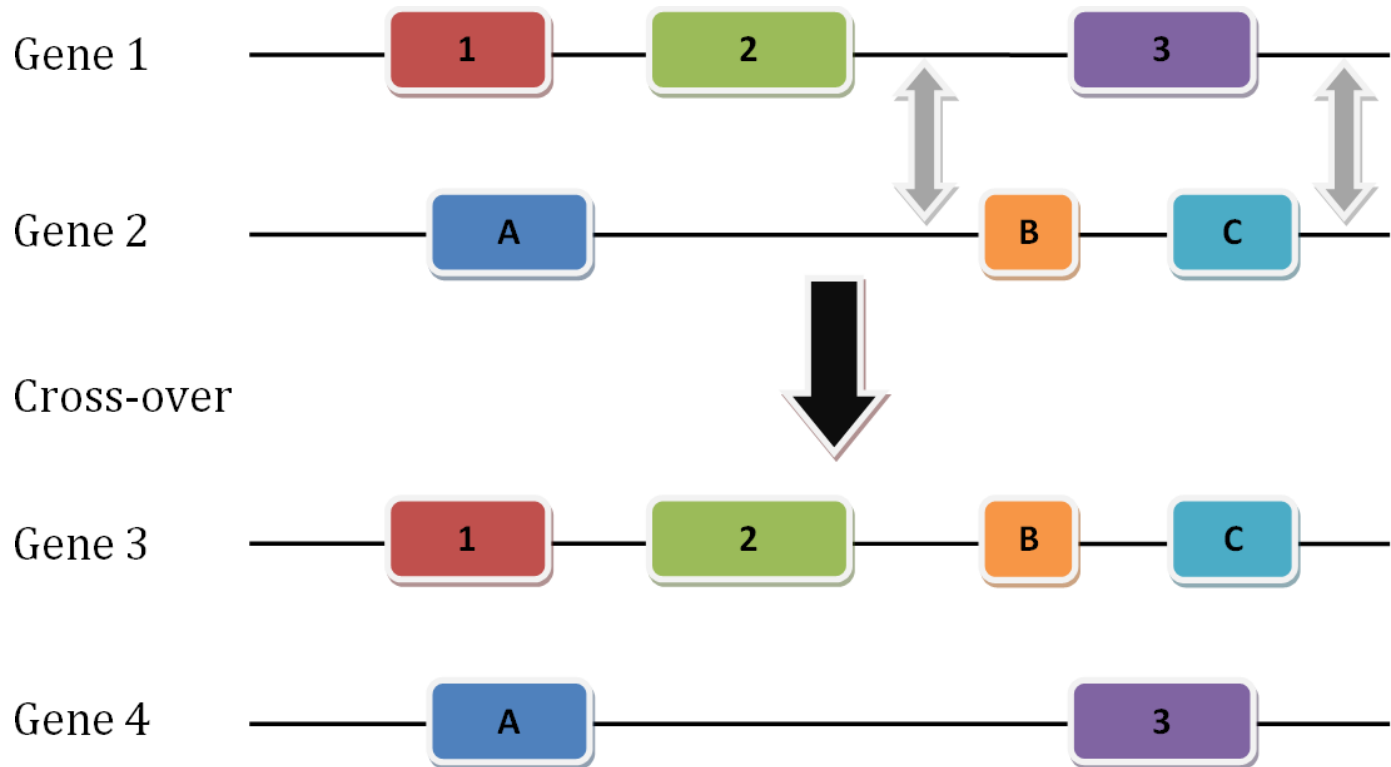


Utility of introns:

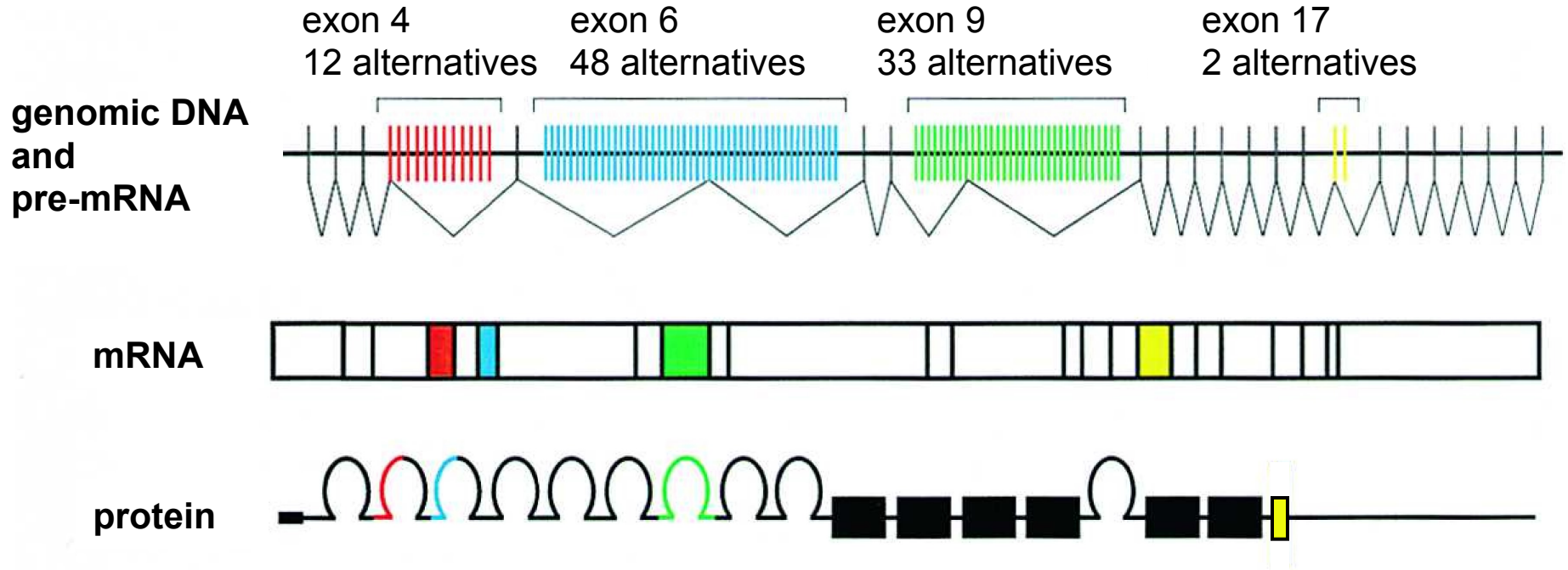
they increase the recombination frequency among coding regions (together with gene duplication they allow the spreading of functional modules to other genes - **exon shuffling**)

Exon shuffling

Exon shuffling is a theory, introduced by Walter Gilbert in 1977, in which different exons either within a gene or between two nonallelic genes are occasionally mixed. Gilbert suggested that exons might each encode a single protein domain, establishing a kind of modular property. In this fashion, it would be possible for exons to essentially be "mixed and matched" to produce a variety of different proteins, yielded from different combinations of such exons and their resulting domain combinations



Current world record holder for alternative mRNA splicing...
the *Drosophila* DSCAM gene, whose protein products
function as axon guidance receptors in the fly









The 61 kb DSCAM gene generates an 8 kb mRNA containing 24 exons. Exons 4, 6, 9, and 17 are encoded as tandem arrays of mutually exclusive alternative exons, so this one gene could in principle generate as many as $12 \times 48 \times 33 \times 2$ or 38,016 different mRNAs and proteins. See Graveley (2005) *Cell* 123, 65-73.

I genomi a rapida replicazione (virus, batteri...) avrebbero perso gli introni perché:

- organismi in nicchie ecologiche in cui non c'è necessità di variare, raggiunto il massimo dell'adattamento. Ricombinazione dannosa

Geni alla fine del loro "percorso evolutivo" avrebbero perso gli introni - GENI PER GLI ISTONI

Still the number of coding sequences cannot justify the evolutionary gap between eukaryotes

Gene number		~6,5k		~48k
		~20k		~48k
Transcript number		~23k		
		>30k		

Britten RJ, Davidson EH.

1969

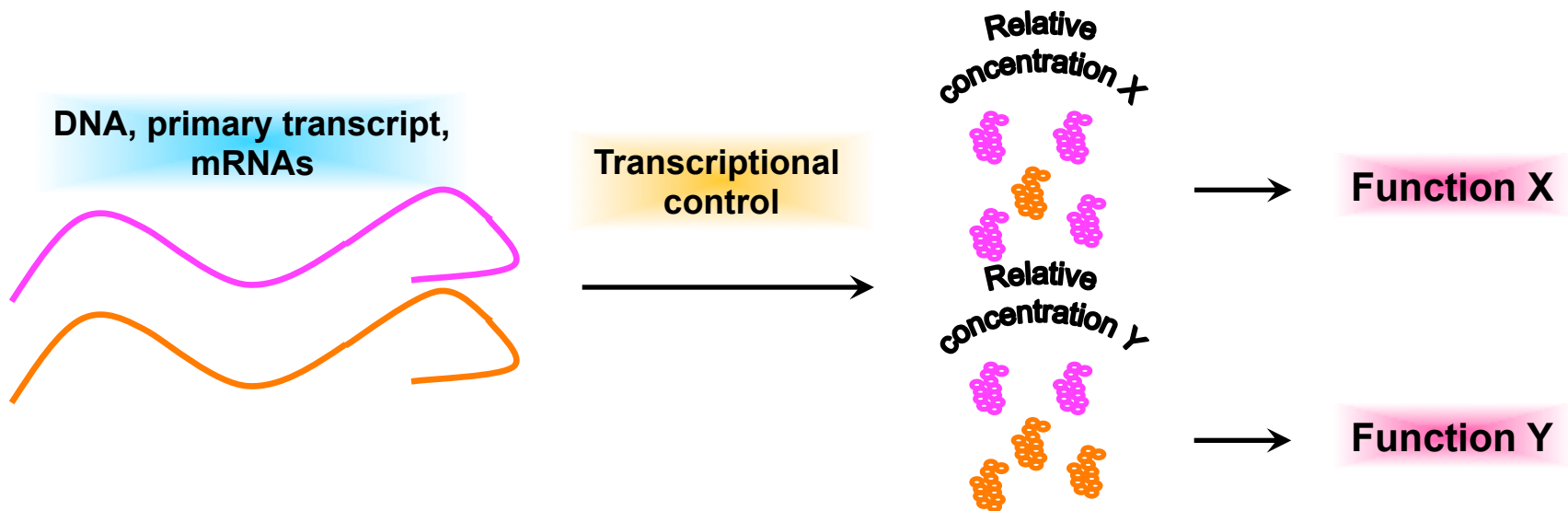
Gene regulation for higher cells: a theory.

Science 165:349-57

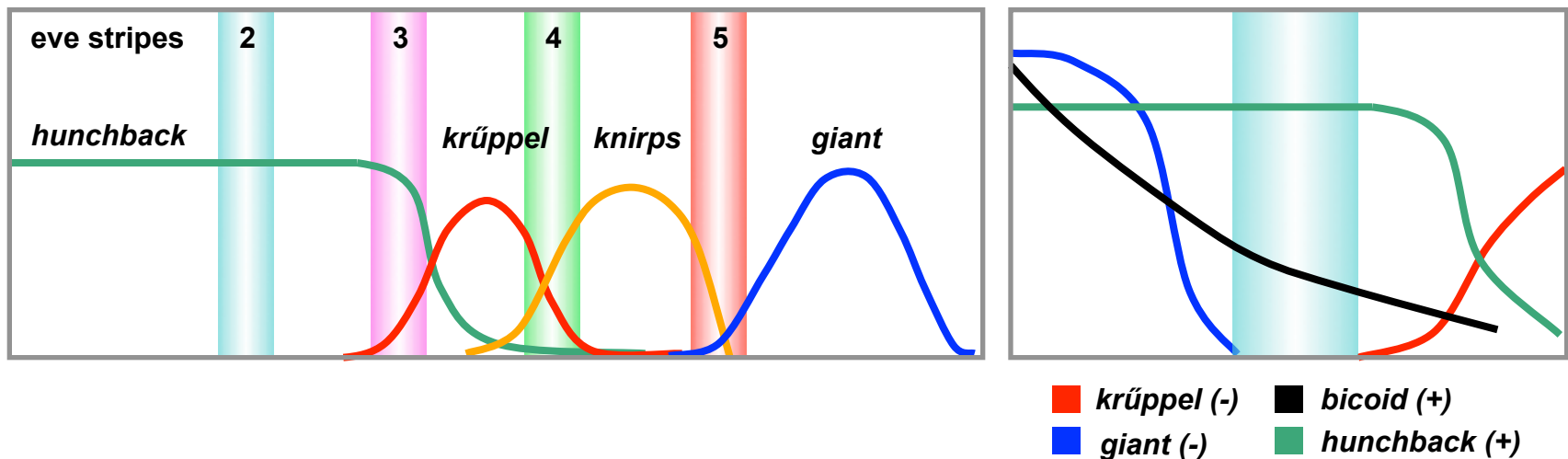
The principal difference between a poriferan and a mammal could lie in the degree of integrated cellular activities and thus in a vastly increased complexity of regulation rather than a vastly increase number of producer genes.

Role of interspersed repetitive sequences!

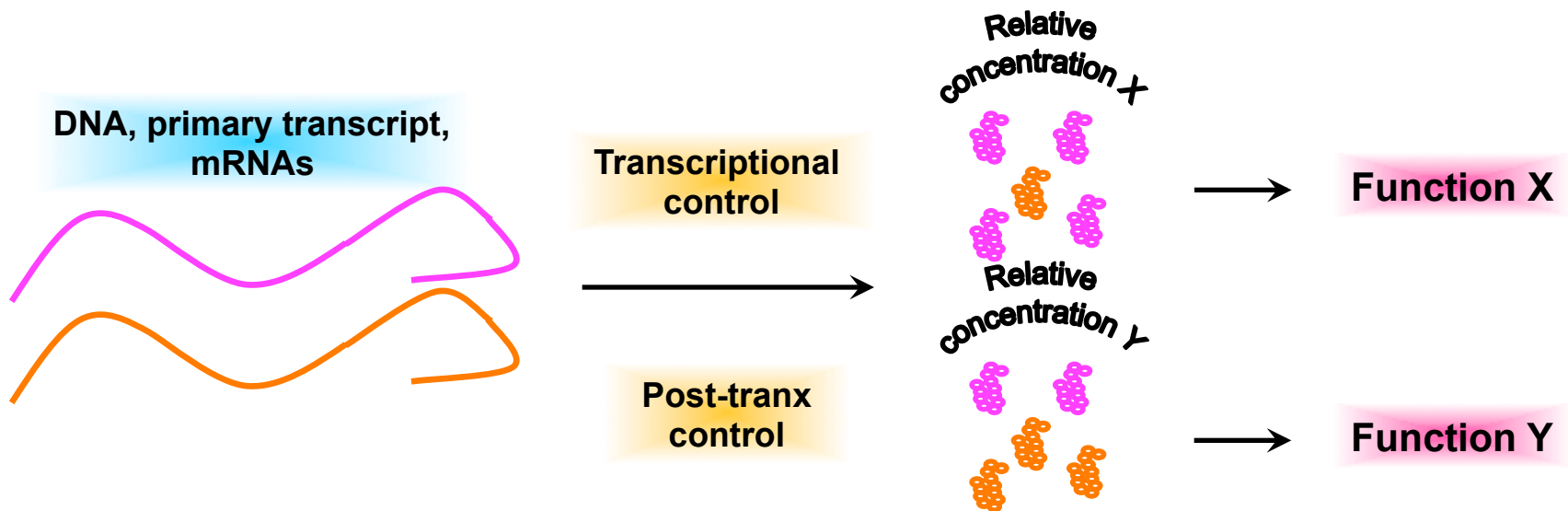
- Regulation of gene expression -
increase the complexity by different stoichiometric
combinations of a fixed number of proteins



Only specific combinations of factors activate specific genes in different body compartments - Drosophila early development



- Regulation of gene expression -
increase the complexity by different stoichiometric
combinations of a fixed number of proteins



Quale è la differenza tra l'RNA ed il DNA?

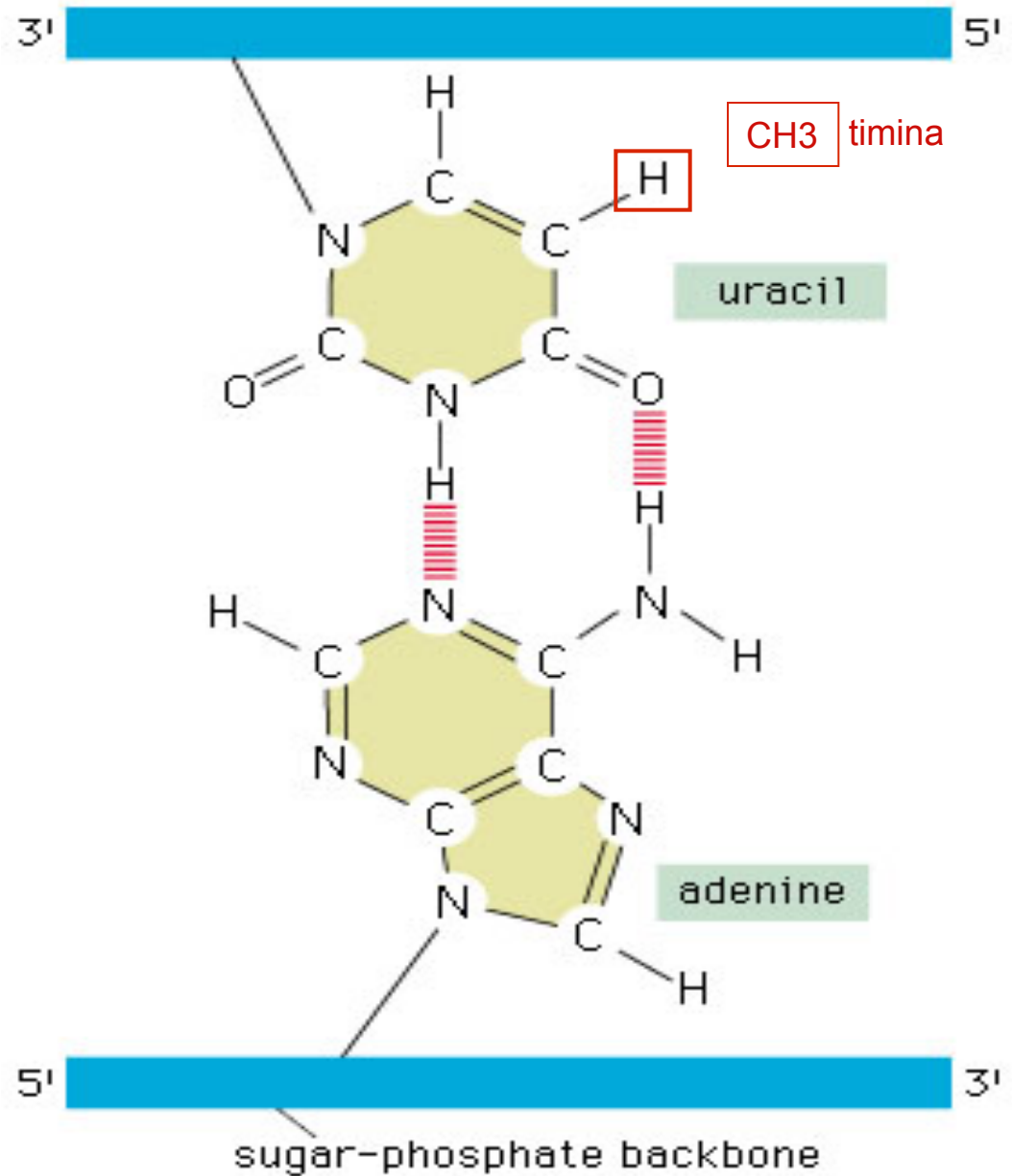
L' RNA ha l'**uracile** al posto della timina

L' RNA ha il **ribosio** al posto del desossiribosio

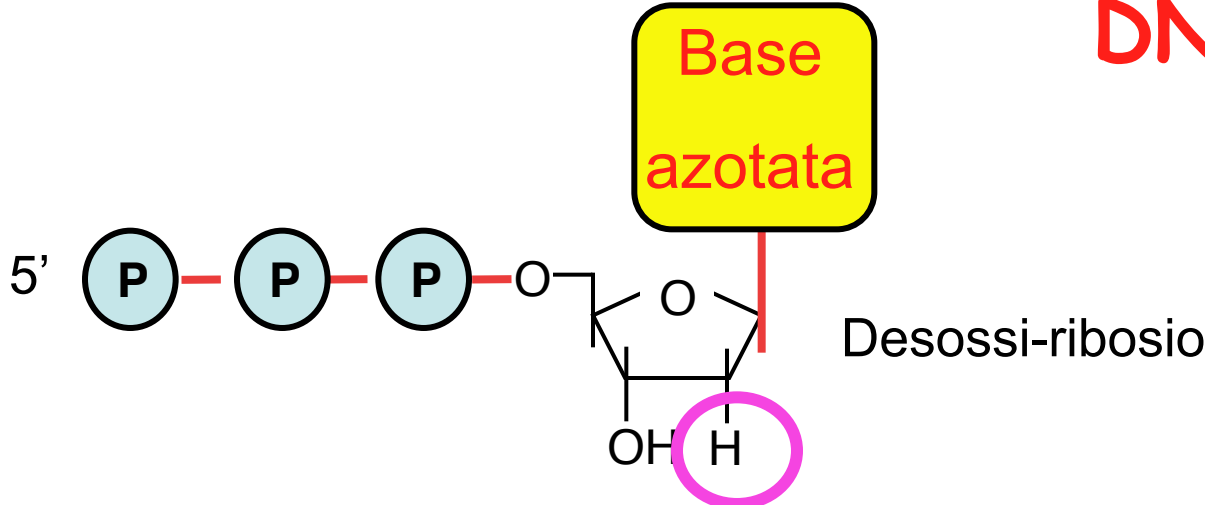
L' RNA è a **singola elica**, rispetto alla doppia elica del DNA, ma può strutturarsi in conformazioni a doppia elica molto complesse

Differenze tra DNA e RNA

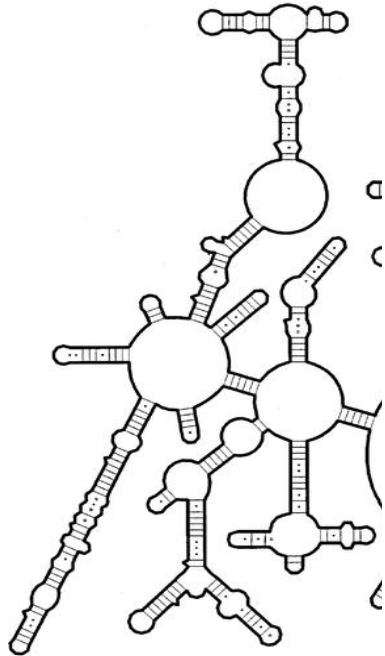
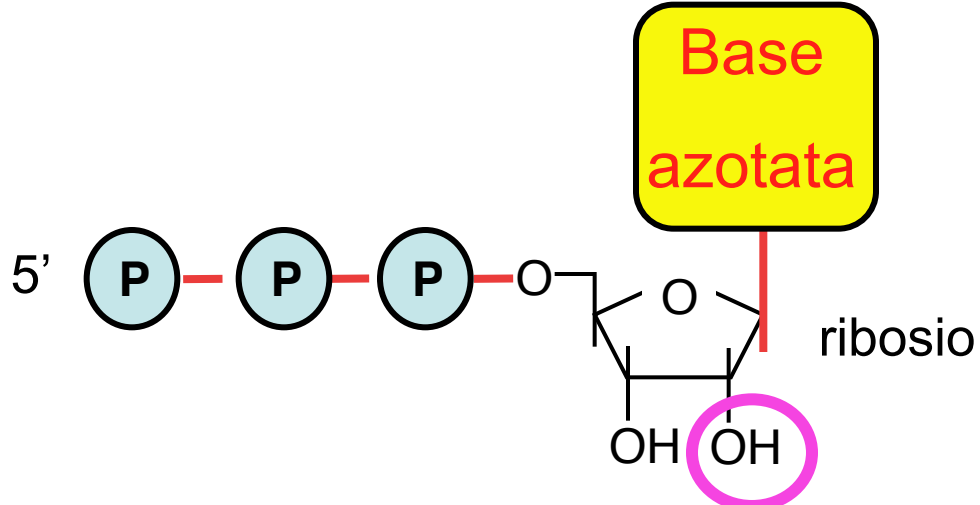
L' RNA ha l'uracile al posto della timina



DNA



RNA



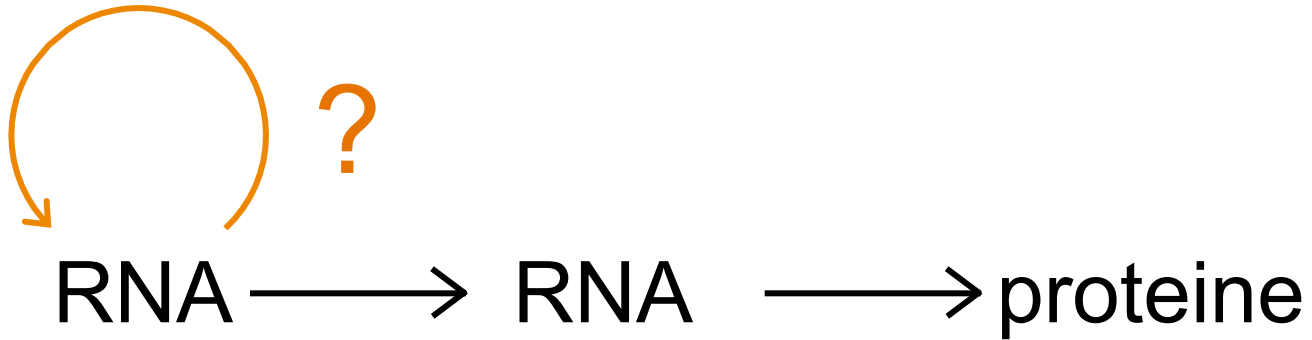
Queste differenze hanno delle implicazioni funzionali ed evolutive

Le prime molecole “viventi” dovevano essere in grado di replicare e svolgere funzioni

Agli inizi degli anni '80 furono scoperti RNA che erano in grado di ricopiare altre molecole di RNA e che svolgevano funzioni catalitiche simili a quelle svolte dagli enzimi proteici

- mRNA splicing e la traduzione sono ribo-enzimi
- parte del genoma deriva da elementi ad RNA
- l'RNA regola la stabilità di altre molecole di RNA

L'RNA è in grado di replicare se stesso

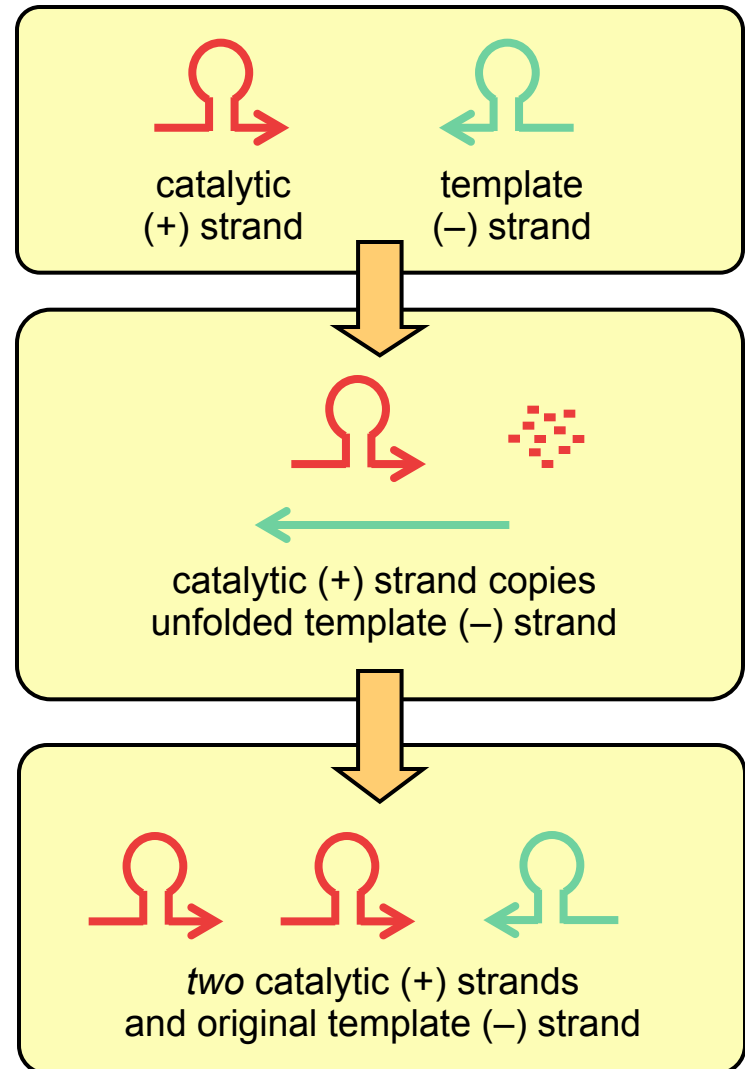


Mondo ad RNA

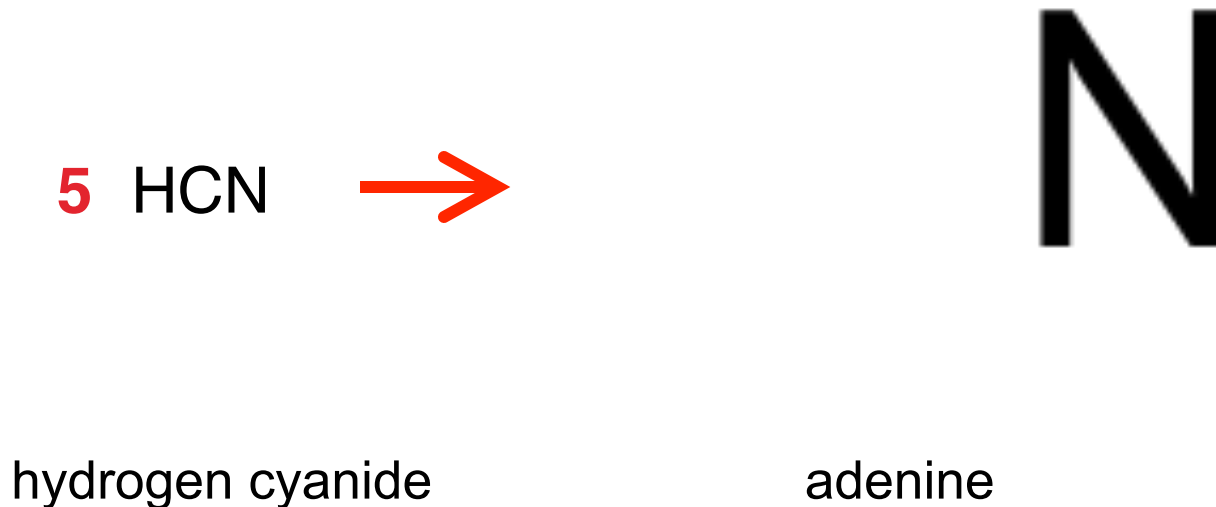
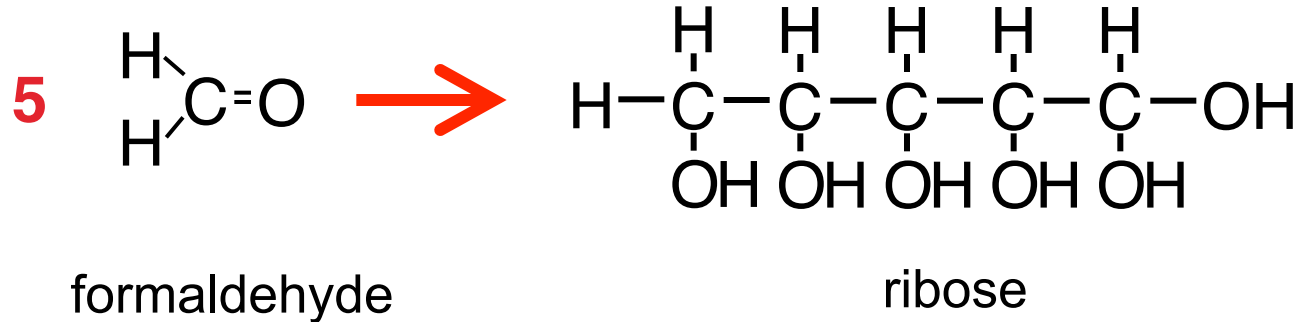
Cech dimostrò che l' RNA poteva polimerizzare legami fosfodiesterici ricopiando un template di RNA

Solo i polimeri di RNA hanno mostrato capacità replicativa.

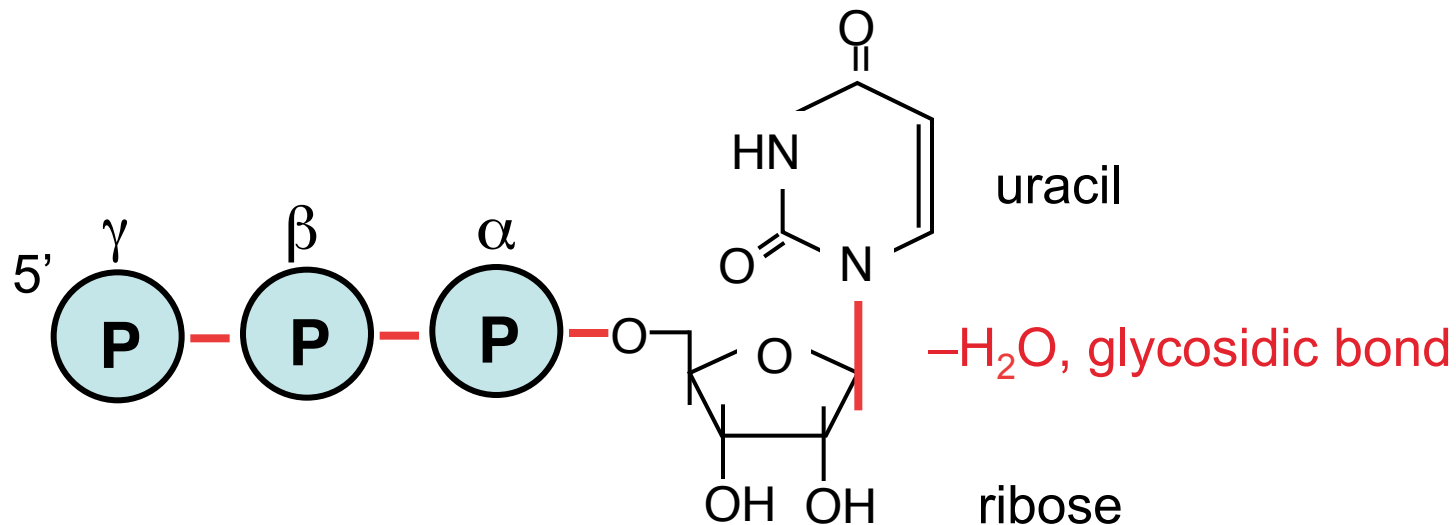
Se l'RNA può essere sia enzima che stampo, e l'RNA può replicare se stesso....ecco la prima **molecola "vivente"!**



Sintesi di precursori di RNA, DNA e proteine potevano essere ugualmente prodotti in un'atmosfera primitiva



Gli acidi nucleici possono originare da condensazione in seguito a deidratazioni.



-H₂O, phosphoester bond (ribose, α phosphate)

-H₂O, phosphoanhydride bonds (α, β and β, γ)

Inizialmente le molecole di RNA si replicavano in modo **lento e casuale** semplicemente fungendo da stampi per il legame di oligonucleotidi complementari che polimerizzavano spontaneamente.

Processo poco accurato - produzione di una gran varietà di sequenze di RNA

Casualmente si potrebbero essere formate molecole di RNA con proprietà catalitiche quali la capacità di **replicare più velocemente e accuratamente**

La **selezione** avrebbe favorito il predominio di molecole con maggior efficienza replicativa

Aumento di lunghezza senza perdita di specificità fornendo il potenziale per altre attività più sofisticate fino a raggiungere strutture molto complesse

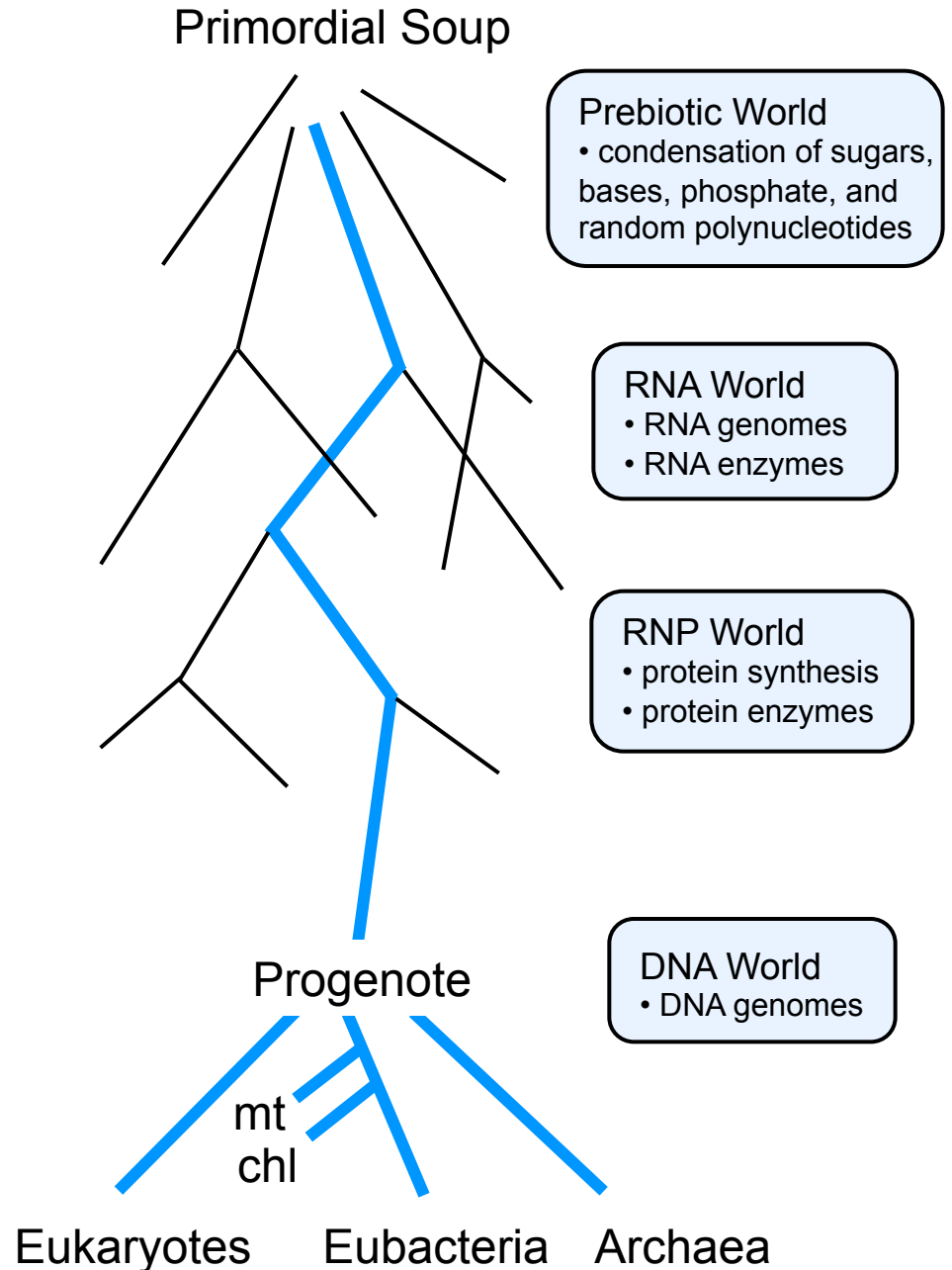
Protogenoma
molecole di RNA
autoreplicanti capaci di
dirigere semplici reazioni
biochimiche

The RNA World

If RNA is catalytic, it could function as both genome and replicase, replicating itself, and perhaps also encoding ribozymes that would carry out intermediary metabolism to make more RNA precursors.

We can debate whether an RNA World existed, or how complex it might have been, and whether RNA may have been preceded by a simpler RNA-like polymer.

However, if life did begin in an RNA-like World, it may *still* be an RNA World today only slightly disguised by a veil of DNA!



Nuove funzioni - *protoribosoma*

I primi polipeptidi (AA- basici) avrebbero potuto stabilizzare strutture secondarie dell' RNA e facilitare la formazione di strutture cataliticamente attive.

Selezione di quegli RNA che codificavano per proteine che ne **facilitavano la funzione replicativa**

Tra queste reazioni quelle del metabolismo energetico con rilascio di energia libera (idrolisi di P-P da ATP e GTP)

Compartimentalizzazione all' interno di membrane lipidiche cellulare avrebbe concentrato i vari componenti e reso più efficaci le diverse reazioni

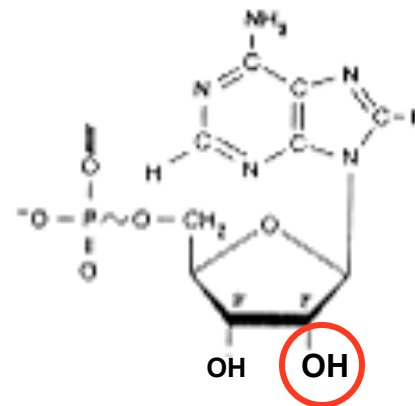
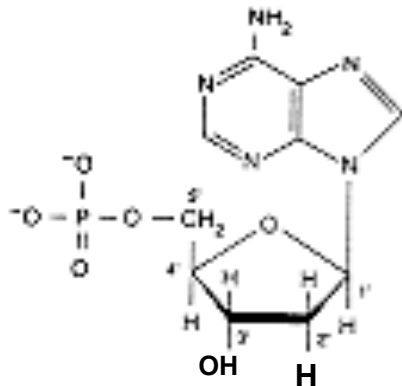
protocellula

Verso un mondo a DNA

sintesi di deossiNTP da riboNTP -

sintetasi mutanti che incorporavano dNTP
(retrotrascrittasi) - prime sintesi di **DNA**

vantaggi del DNA - desossiribonucleotidi più stabili
perché manca l' OH in 2' sullo zucchero

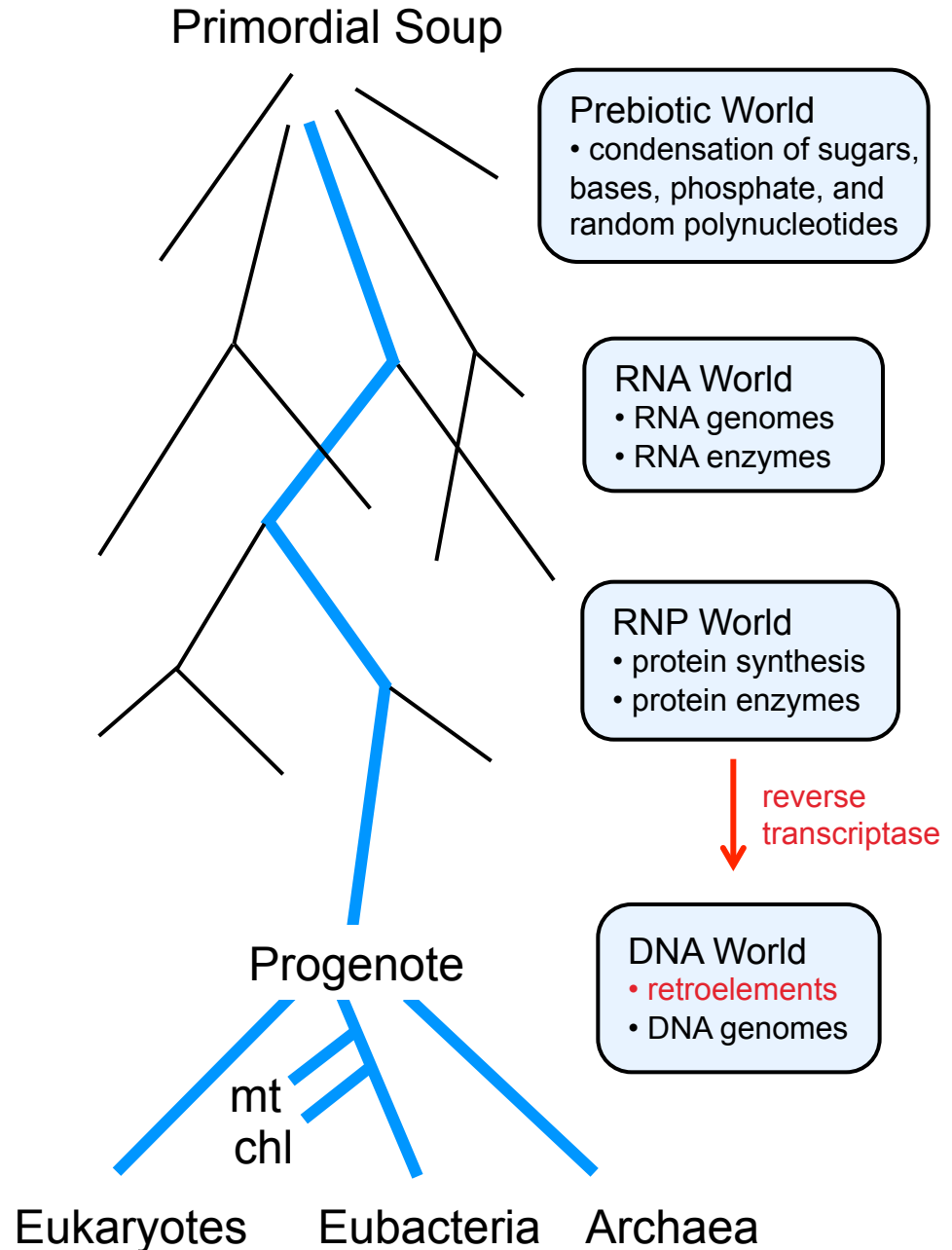


The RNA World

If RNA is catalytic, it could function as both genome and replicase, replicating itself, and perhaps also encoding ribozymes that would carry out intermediary metabolism to make more RNA precursors.

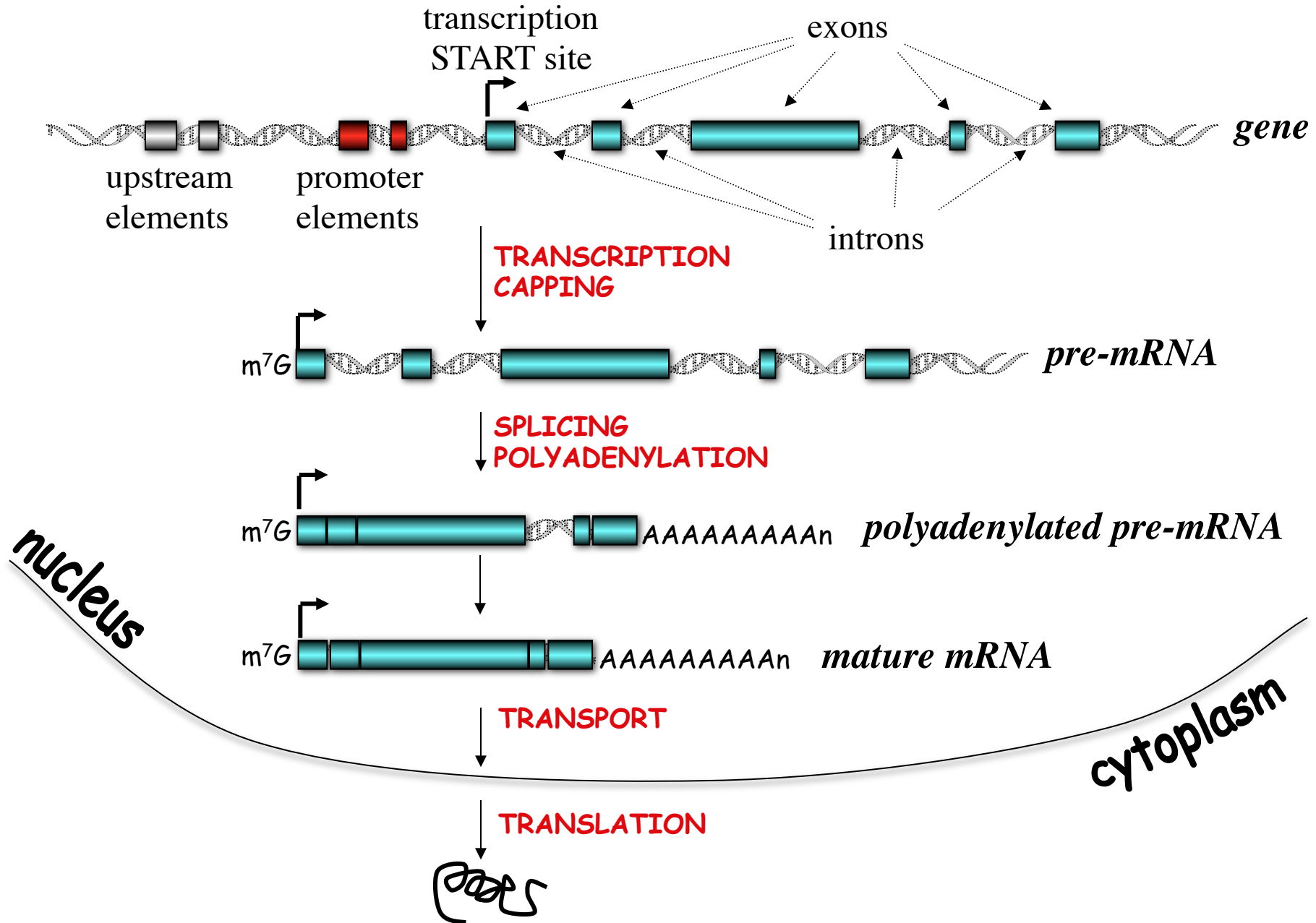
We can debate whether an RNA World existed, or how complex it might have been, and whether RNA may have been preceded by a simpler RNA-like polymer.

However, if life did begin in an RNA-like World, it may *still* be an RNA World today only slightly disguised by a veil of DNA!



Protogenoma
molecole di RNA
autoreplicanti capaci di
dirigere semplici reazioni
biochimiche

Espressione genica negli eucarioti



La regolazione post-trascrizionale

Per tanti anni lo studio della regolazione dell' espressione genica si è concentrato sul controllo trascrizionale.

Il motivo risiede nel fatto che i primi geni isolati codificavano per prodotti espressi nel differenziamento terminale (prevalentemente sotto controllo trascrizionale - globine, immunoglobuline, ovalbumina)

Quando si sono isolati i geni housekeeping o quelli finemente modulati nel differenziamento cellulare ci si è accorti che molta parte della regolazione dell' espressione genica avveniva a livello post-trascrizionale

Perché studiare l'RNA?

Regolazione post-trascrizionale

nucleus

splicing/processing *sn-snoRNAs*

poliadenilazione/formazione 3' *snRNAs*

modificazioni (CH_3 , ψU) *snoRNAs*

trasporto

cytoplasm

traduzione *miRNAs*

editing *gRNAs*

stabilità *siRNAs*

La maggior parte dei trascritti del genoma sono RNA non codificanti per proteine

RNA non codificanti

large - rRNA 18+28S

Xist

.....

.....

...?...

.....

small - 5S rRNA

tRNA

snRNAs

snoRNAs

scRNAs

gRNAs

piRNAs

miRNAs

siRNAs

raRNAs

...?....

traduzione

traduzione

splicing

modificaz./process. rRNA

controllo traduzionale

editing

stabilità del genoma

controllo traduzionale

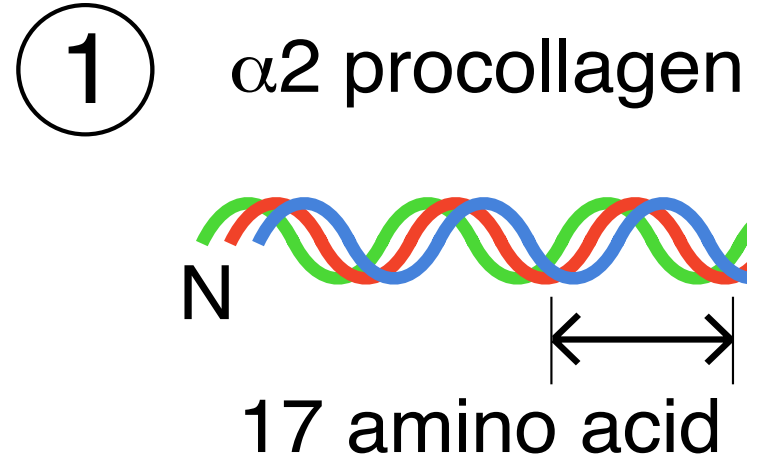
stabilità dell'RNA

silenziamento trascriz.

La teoria esonica dei geni: geni complessi si sarebbero formati da corti esoni che sarebbero stati assemblati insieme dallo splicing

Ricombinazioni accidentali sono random e possono creare perdita del registro di lettura...

lo splicing dell'mRNA è intelligente ed unisce senza errori ciascuna regione codificante a quella successiva



The exon structure of the collagen IV gene provides a striking example for collagen evolution and the role of introns in gene evolution. Collagen IV, a major component of basement membranes, differs from the fibrillar collagens in that it contains numerous interruptions in the triple helical Gly-X-Y repeat domain. We have characterized all 47 exons in the mouse alpha 2(IV) collagen gene and find two 36-, two 45-, and one 54-bp exons as well as one 99- and three 108-bp exons encoding the Gly-X-Y repeat sequence. All these exons sizes are also found in the fibrillar collagen genes.

J Mol Evol. 1990 Jun;30(6):479-88. **Evolution of collagen IV genes from a 54-base pair exon: a role for introns in gene evolution.** Butticè G, Kaytes P, D'Armiento J, Vogeli G, Kurkinen M.