

Modelli Computazionali per lo Studio di Sistemi di Interesse Biologico (Bioinformatica)

Testo



Anna Tramontano
Bioinformatica
Ed. Zanichelli

**Lezioni *on-line* (<http://www.chem.unisa.it/luigicavallo/>)
(derivate da lezioni tenute dal gruppo BioComputing
<http://cassandra.bio.uniroma1.it>)**

1. Introduzione alla bioinformatica
2. Evoluzione e informazione
3. Ricerca di geni in genomi di procarioti ed eucarioti

I processi della vita sono dovuti all'azione concertata di (macro) molecole biologiche, principalmente proteine.

Le istruzioni che dirigono la sintesi proteica sono contenute nel genoma.

Dobbiamo quindi

- identificare le istruzioni e decifrarle
- comprendere quando e quanto sono "lette" (regolazione)
- capire quali modifiche subiscono nell'ambiente cellulare
- individuare le loro interazioni

Bioinformatica: scienza multi-disciplinare, al crocevia tra biologia, chimica, matematica, fisica ed informatica, che analizza l'informazione biologica con metodi computazionali al fine di formulare ipotesi sui processi della vita.

Anna Tramontano

BIOinformatica = geni + proteine + informatica
(biologia computazionale, biocomputing)

GENE: segmento di DNA che codifica per una proteina specifica e determina UN carattere ereditario

PROTEINA: prodotto di espressione di un gene ed EFFETTORE della funzione biochimica di cui il gene stesso contiene l'INFORMAZIONE

Raccolta, archiviazione,
organizzazione e
interpretazione di dati biologici.

J.M. Thornton

Bioinformatica

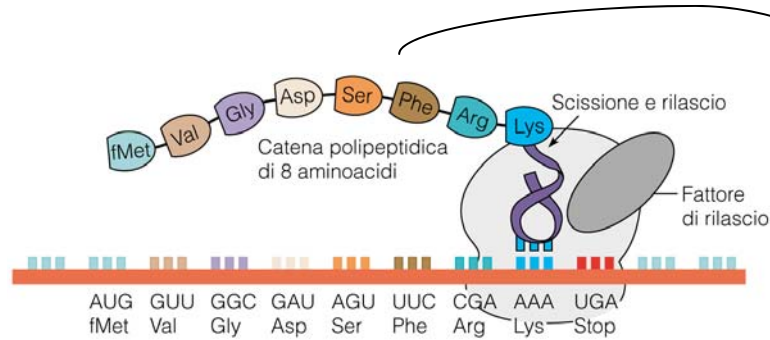
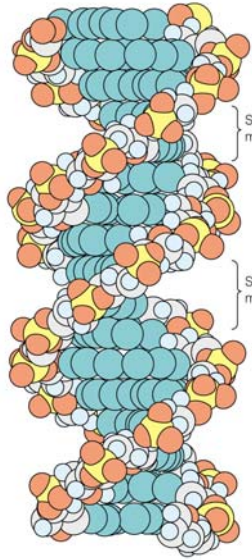
La bioinformatica e' cio' che la bioinformatica fa

Eric C. Snowdeal III

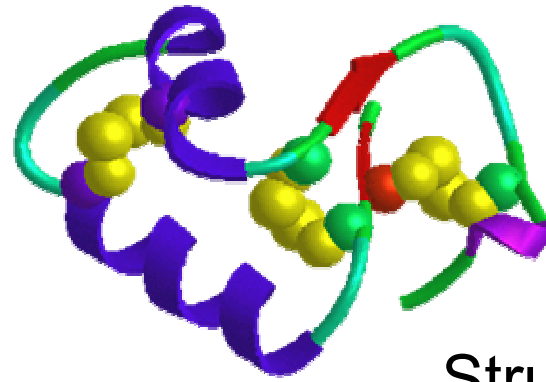
- Analisi di sequenze
- Analisi di strutture
- Predizione di strutture
- Disegno di molecole proteiche
- Disegno di inibitori
- Disegno di librerie combinatoriali
- Sviluppo di tools
- Gestione di dati

Campi d'applicazione della bioinformatica

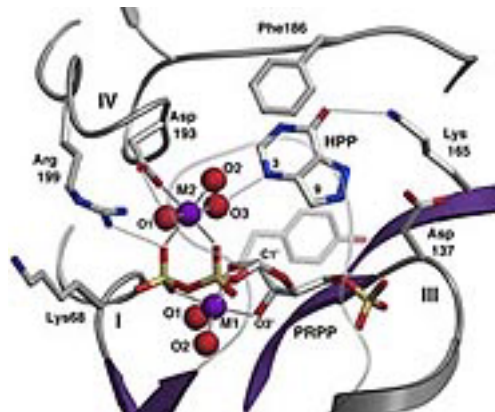
DNA



Struttura
secondaria



Struttura
terziaria



Attività enzimatica

Cosa ci si aspetta che impariate:

Analizzare una sequenza

Ricerca omologie

Costruire un modello tridimensionale di una proteina

Capire i problemi legati all'analisi di dati genomici e proteomici

Conoscere e usare i principali siti bioinformatici

Conoscere l'affidabilità dei metodi

Un po' di storia...

1951 Pauling: alfa e beta

1953 Doppia elica DNA

1955 Sequenza insulina

1959 Struttura mioglobina

1960 Anfinsen

1967 collezione Dayoff

1968 PAM

1970 Needleman and Wunsch

1977 PDB

1977 Chou and Fasman

1977 sequenza DNA

1980 Wutrich

1981 Greer

1985 FASTP

1986 Chothia e Lesk

1990 Blast

1991 Fold recognition

1993 PhD

1994 CASP

Un po' di storia...

1951 Pauling: alfa e beta

1953 Doppia elica DNA

1955 Sequenza insulina

1959 Struttura mioglobina

1960 Anfinsen

1967 collezione Dayoff

1968 PAM

1970 Needleman and Wunsch

1977 PDB

1977 Chou and Fasman

1977 sequenza DNA

1980 Wutrich

1981 Greer

1985 FASTP

1986 Chothia e Lesk

1990 Blast

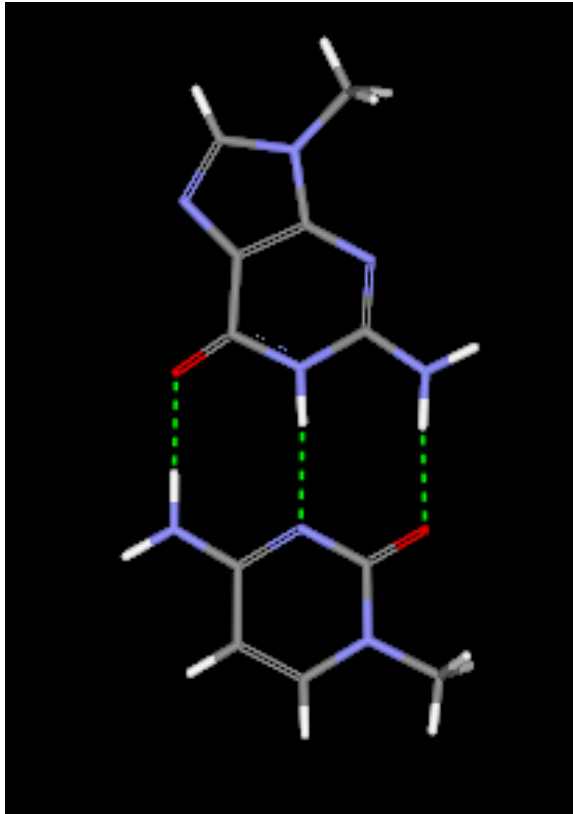
1991 Fold recognition

1993 PhD

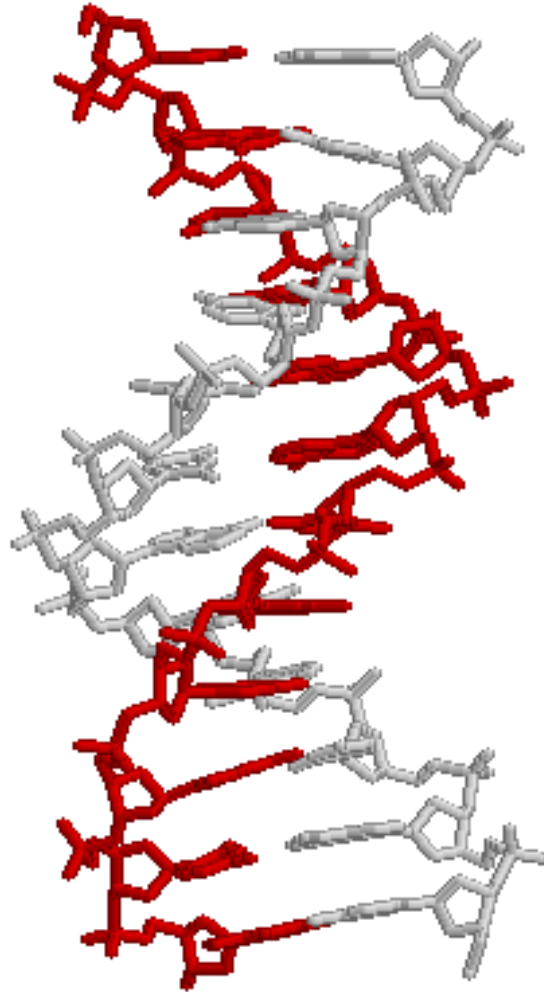
1994 CASP

DNA

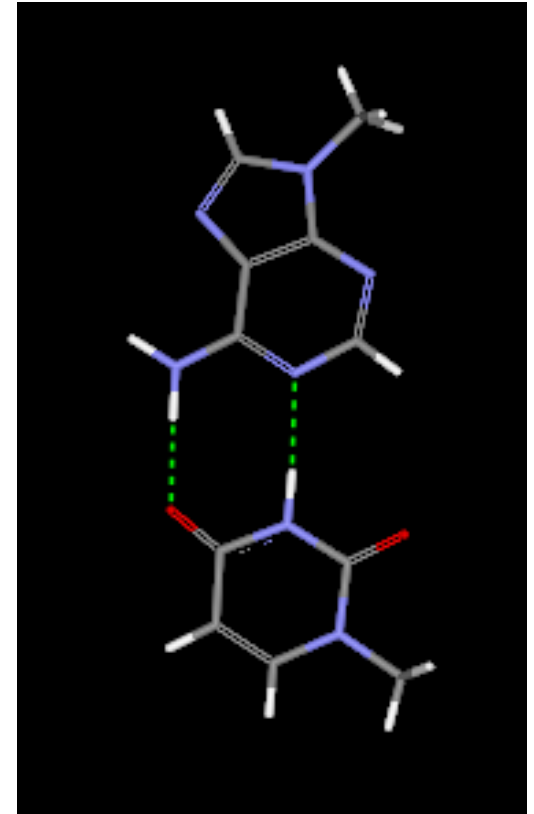
Guanina



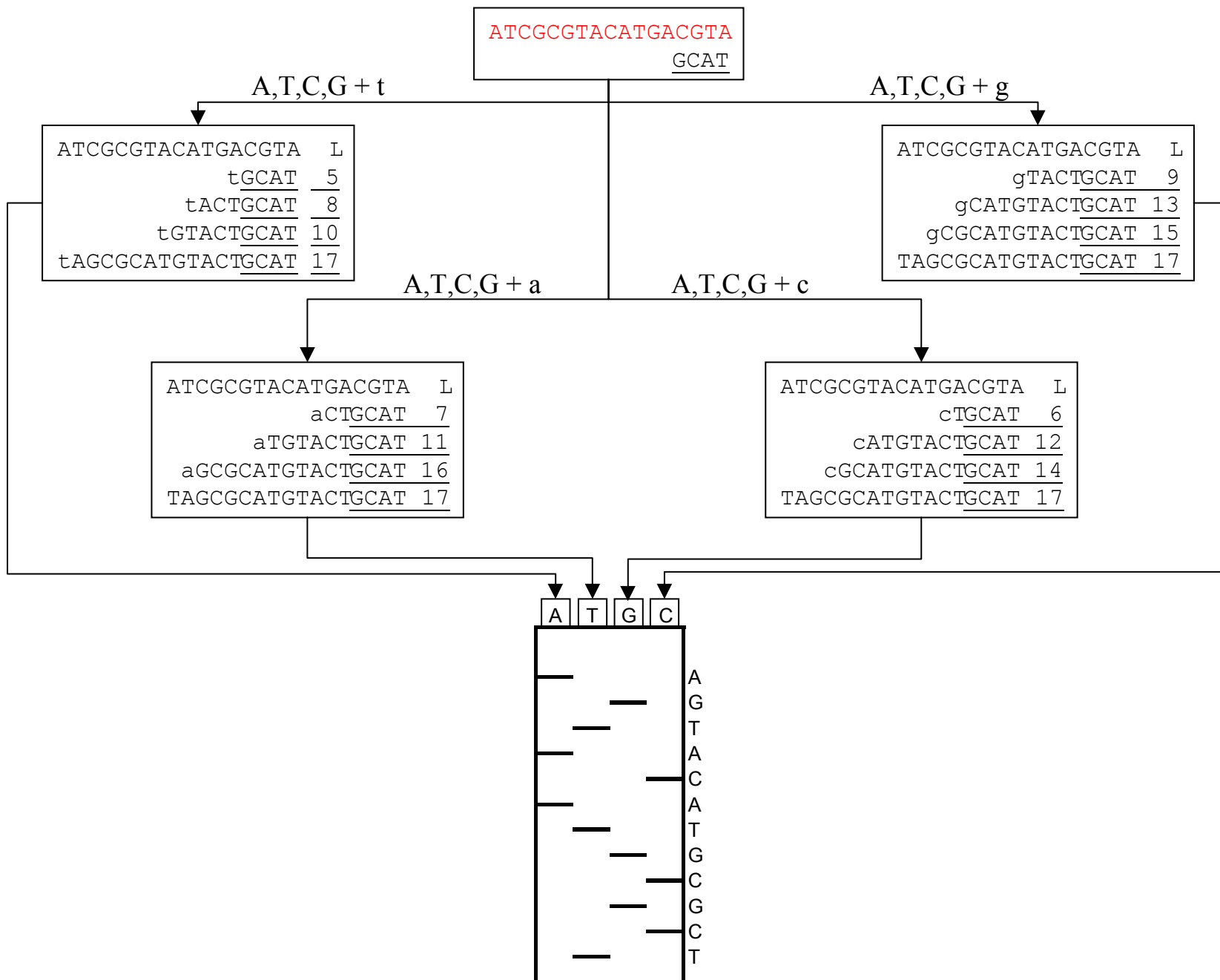
Citosina



Adenina



Timina



ATCGCGTACATGACGTA
GCAT

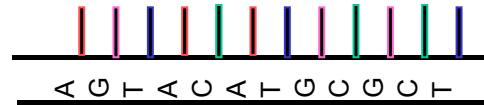
A,T,C,G + **tacg**

ATCGCGTACATGACGTA	L
t GCAT	<u>5</u>
t ACTGCAT	<u>8</u>
t GTACTGCAT	<u>10</u>
t AGCGCATGTACTGCAT	<u>17</u>

ATCGCGTACATGACGTA	L
g TACTGCAT	<u>9</u>
g CATGTACTGCAT	<u>13</u>
g CGCATGTACTGCAT	<u>15</u>
TAGCGCATGTACTGCAT	<u>17</u>

ATCGCGTACATGACGTA	L
a CTGCAT	<u>7</u>
a TGTACTGCAT	<u>11</u>
a GCGCATGTACTGCAT	<u>16</u>
TAGCGCATGTACTGCAT	<u>17</u>

ATCGCGTACATGACGTA	L
c TGCAT	<u>6</u>
c ATGTACTGCAT	<u>12</u>
c GCATGTACTGCAT	<u>14</u>
TAGCGCATGTACTGCAT	<u>17</u>



Il processo e' automatizzabile...

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 28, 2000

Web posted at: 12:09 p.m. EDT (1609 GMT)

In this story:

~~understanding~~
[Knowledge can help treat causes of diseases](#)

[Advances could come quickly](#)

[RELATED STORIES, SITES](#) ↓

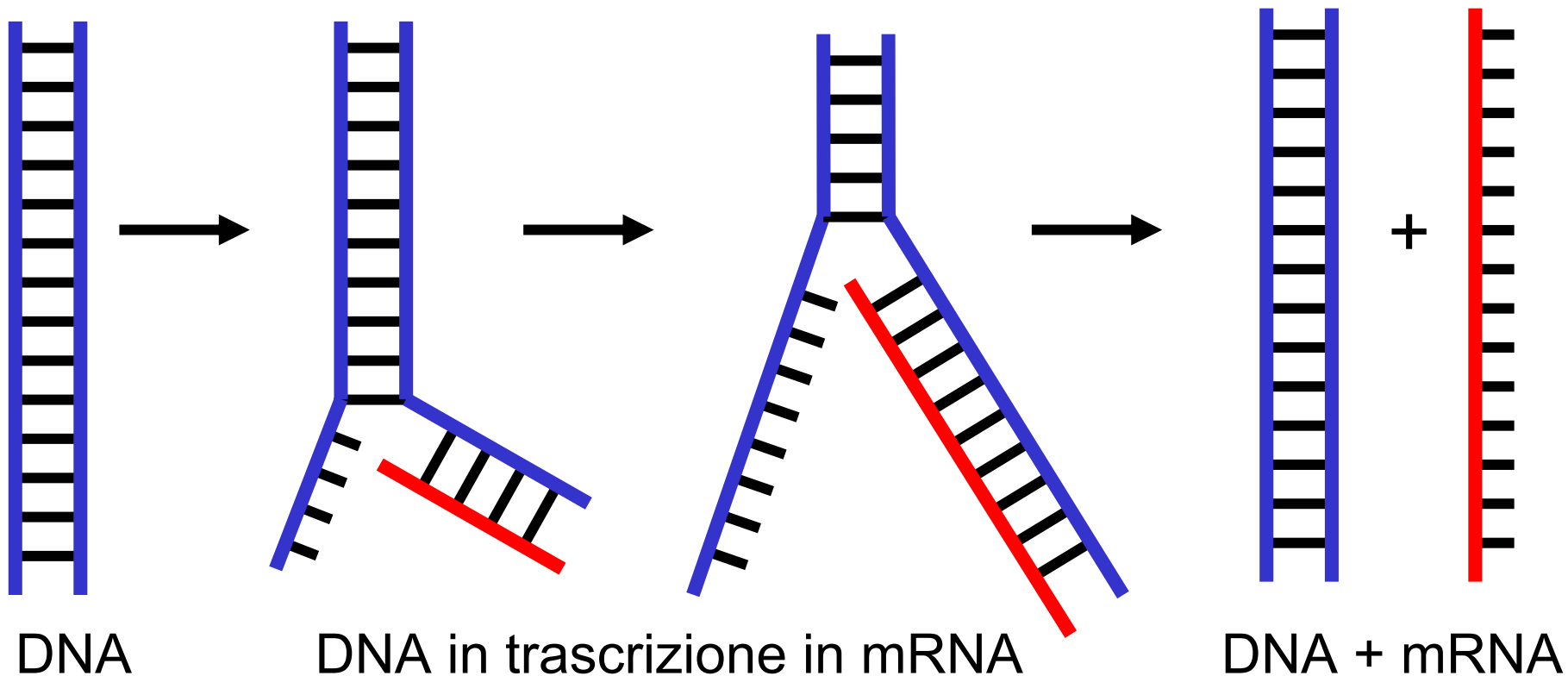
From staff and wire reports

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.



La sintesi proteica richiede la traduzione di sequenze di nucleotidi in sequenze di amminoacidi

1) Trascrizione dell'informazione del gene dal DNA allo mRNA da parte di una RNA-polimerasi



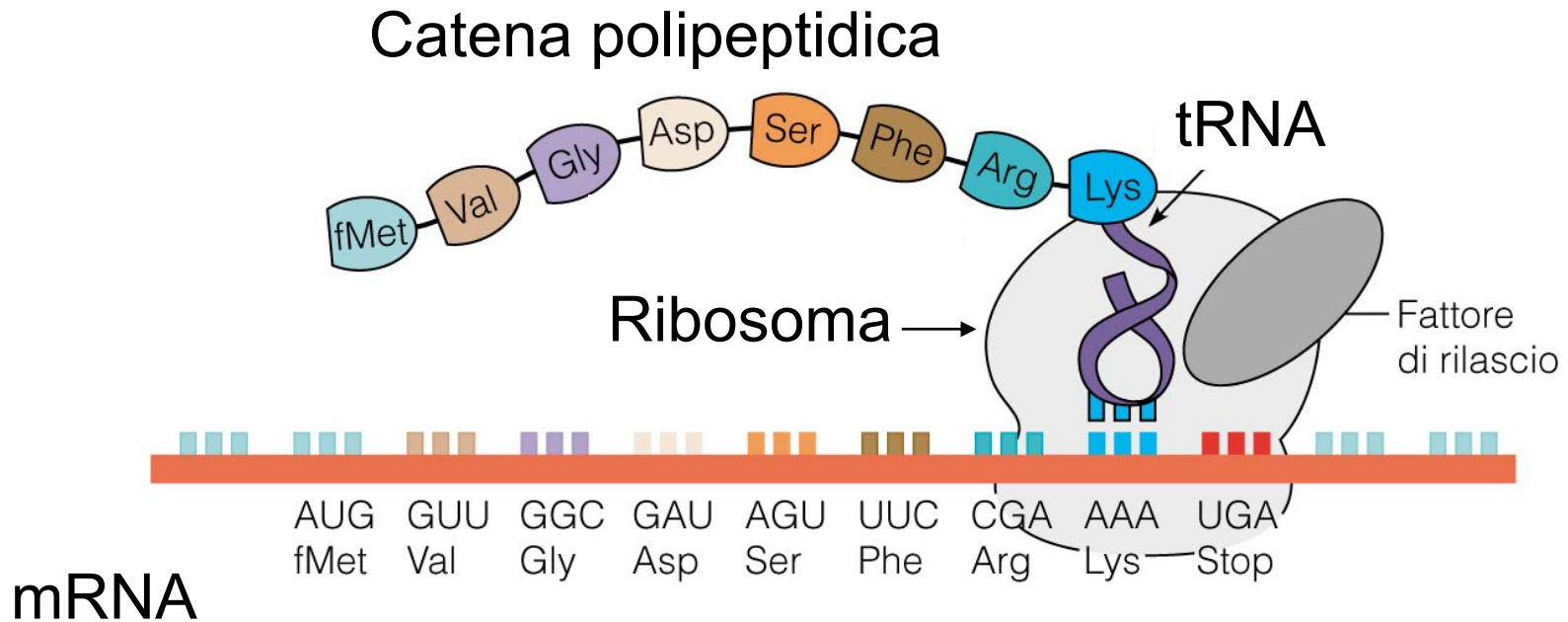
Il codice genetico

La sequenza dei nucleotidi (presi a triplette) dell'mRNA (e del DNA da cui deriva) contiene l'informazione su quali amminoacidi concatenare a formare la proteina

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

La sintesi proteica richiede la traduzione di sequenze di nucleotidi in sequenze di amminoacidi

2) Traduzione dell'mRNA in proteina da parte dei ribosomi



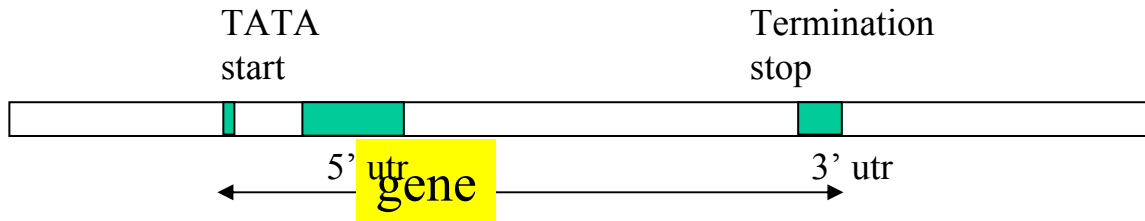
Dove e' il gene?

>cD0826Q1_425-22425 Main

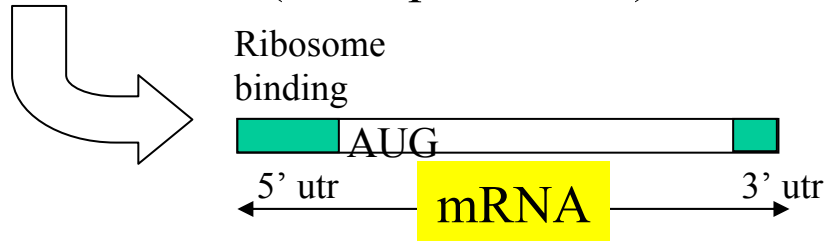
```
ggcataagaatgatacaatggactttggggacctgagaggaaaggtgggagggggcaagg
gatactgctcaggtgataggtgcacccaaaatctcacaaatcatcactaaagaaacttactc
atgtaaccaaatactacctgtaccactataacctacggggaaaaaagcaacataacctat
gaaccaactaataaaaaaacaccttgccctcagtcctgcacatccctagagacactctc
tctgtgtcctcacacttgagctaaagcttctgacttttgtctccagtacacccctgagga
tcctctcatcacggccatcagaacctctgtagaaggtcaaatccagtggggttctgtca
gtgcctctgacttgagttactgataatatttgaccataatccacttctttctaagtgc
tactctgtccttatttttctcctatttactgaatcctccttatcatcctttgaaatctcc
tcttaattattatgttctctcatcatcacccctgagatccctgcatttctgatttttgccac
tcttcttggaagctcatctaacctgcacctatgcttgatgactctcagttctctggct
taaactcctctactgagaccacccatcatacaaaatgtttacatatatttttcccttag
ataacttttagatattctaagtcaatagccccacactgaactcagtccttctctcagtc
caggctgtcttctctcattaccctttttaatgaatggaatcaagatgtttgcattgggtt
ggggagatgttggtcaaaggatacatccatttcatttcatttaggatcatttcaaaaga
tacatttcatattagattggaggaataattttaagagttttattgtataacatggactata
gttgctaacaatgtattgttgaaaattgctaaaaggtggattttaagtggtctcaccac
aaaaataaagtatgtgaggtgagccataagttcttagcttgatgtagccggtccatgat
gtacatacatttcaaaacaacataattatacatgataaataataaatttttgtcaatca
aaataatttagaaaagtgacacacacttacacacacacacacaaaagagatgttgcat
ggccagtcctaggaataagagttatctgggagttttctaagtcggatgccaccgacatcac
tcaccaataatcccttttaagtcaatcaaattaaagtcctcttcttccatcattttactcc
tatgccatttccctcactctttgttcaggcactattagtccttgccctctgaaaccaactc
tttactcatgctgccactgttgccgtagtgatcttcttaaatgcaaatgcgccatca
cttccctgcttaaaatccttcaatgattccttatgacttccaggacagagtaggccactcc
tgagctttgcatgtaacatctgtcatgatccagccctgcctgtctatttttctttttt
cttgctgtgtgtccacatccaaagctggctccattcactgaagcagctgaagttcttc
agatatgtcattggcacactgggcccacacttttgaacctgttccctctgtgtgagaag
tggcttctgcctgttttcggactgcctacattgaagccatctgttccccaggaagcctt
ccctgatgccttgacagcagcatcttgctgcctgccccatatctgacattatccatctggg
cctgctgttgcttctgtcacttctgtgttcttctctgtgaactgtaaacatcaggaggacaag
acctatgtcttacttttatttgaaatatttagcatctaacaatgttcgacatatagtaggc
ttttgatactattttttactatgacattgtagtatatgttaatatccagtaggacatag
gatatattctctctgttttcaatttttcatgttttacacacatttataattctatctata
aggatttacaattatttcatgaaatgaatgaataaatagagaatgttagatattaaga
gacagtggtggaagccaggctgggactagggatgcacttaccttagtgctgcaaaatttagg
aggataccaaaagaactcagtaataaaaagtcaatcatattttaatgaaatatcttaagaa
atctaaattaatggaaaatatataatgaacaaaatgtcaaaagagaactattcaagaaa
atggagaagcagagaggcagaagaattagtagaatatactggcacataagccaaggaggt
aaagatttccaggaaggaggaagtagagtgagtcagaagttcaacagaagtcatttcag
aaatcttacccttggttttgaaatcctttcagagagcagttttacataatgtgagcaatta
tttctccttcatccccatcattccagaattgagcttcttctctggcttcagaaatgtggc
ccttccccttgctcaggatattgttgccgacatgatgcatgaggatgcctcaaagtcagct
ggggttgggggtgaaattaattgactttagggaaactccttgaatgtaagttctgttca
cctggaggaccagagagggcacagagatgaccacctagcttctgcctgggacctaaacag
ggcagagaaaataggaggatcaggtataaaggagcagggaagatgggtctgggcttacag
```

circa il 97% del DNA umano è non-coding !

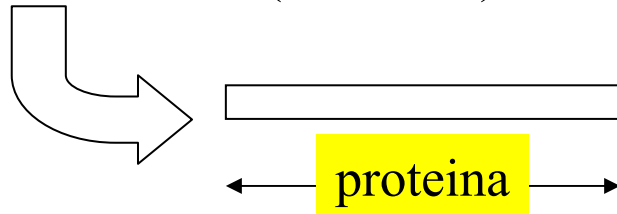
Struttura di un gene procariota



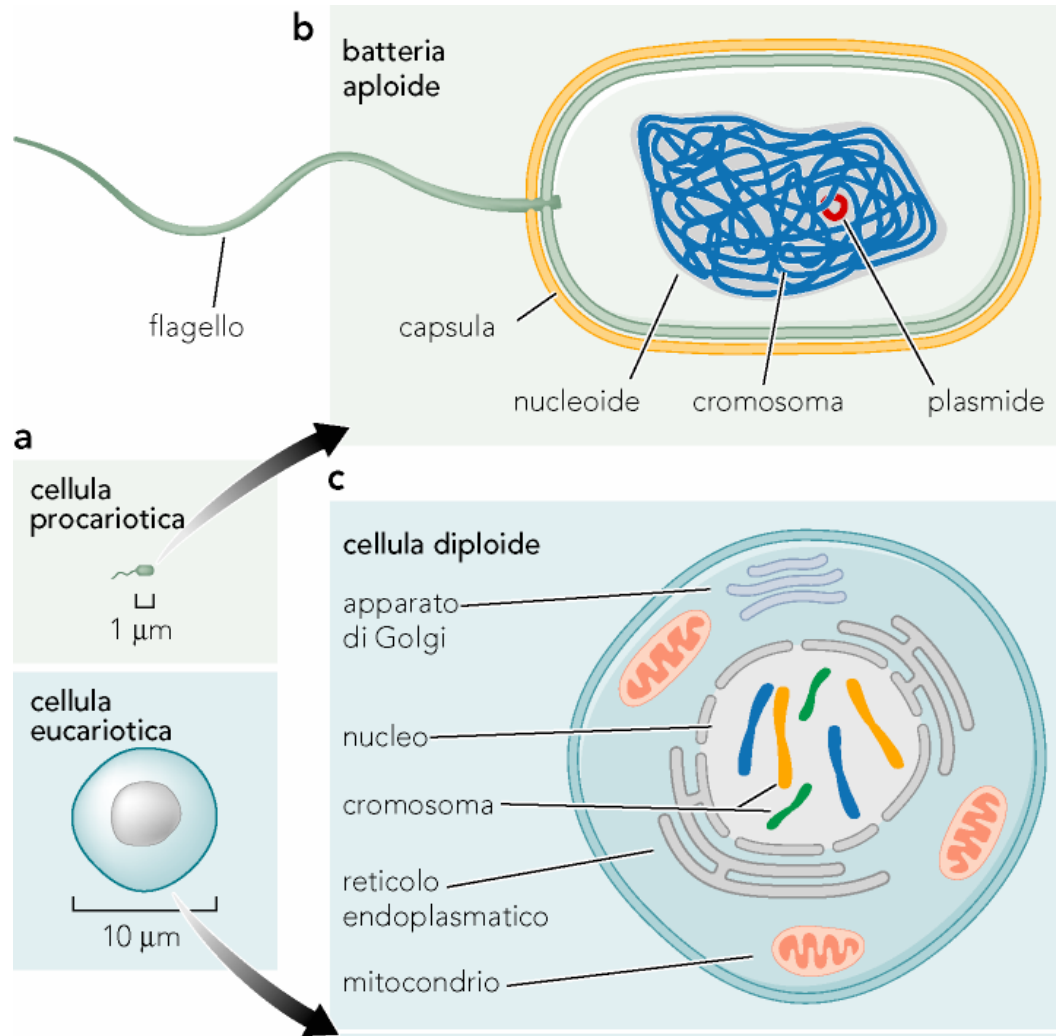
Trascrizione (RNA polimerasi)



Traduzione (ribosoma)

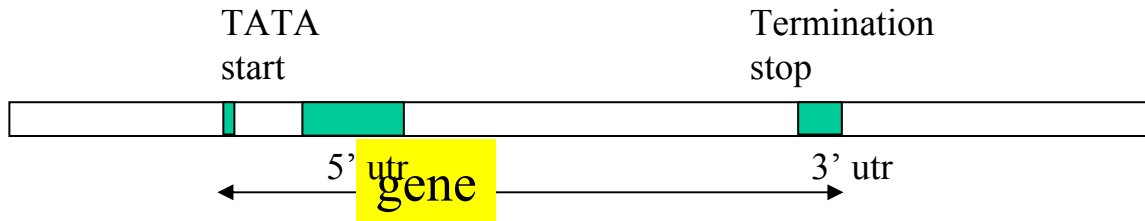


Cellula procariotica

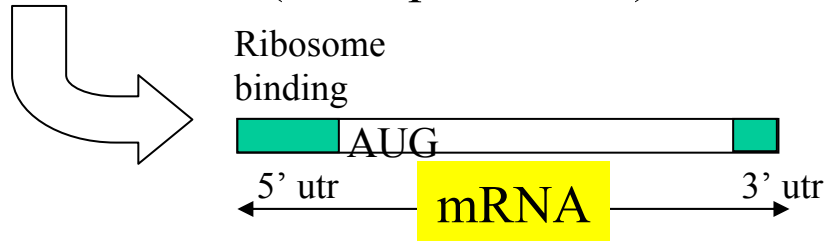


Cellula eucariotica

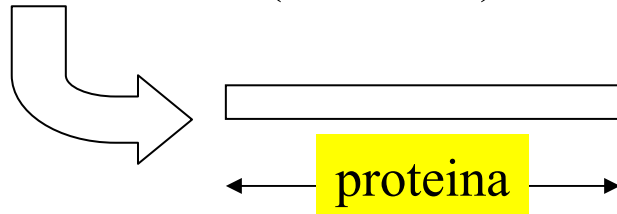
Struttura di un gene procariota



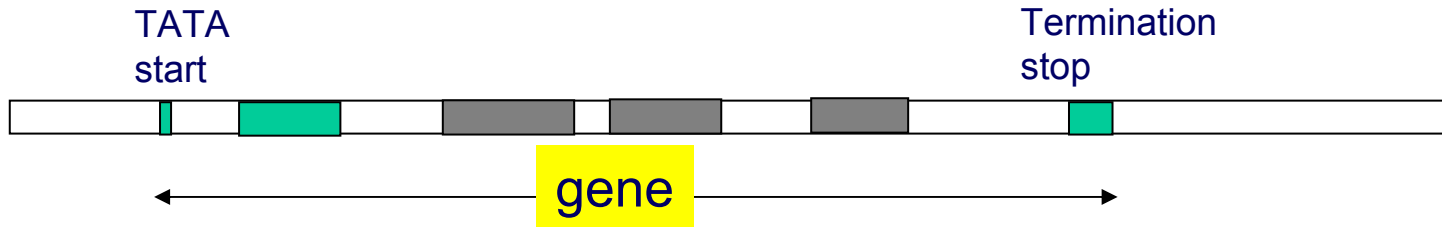
Trascrizione (RNA polimerasi)



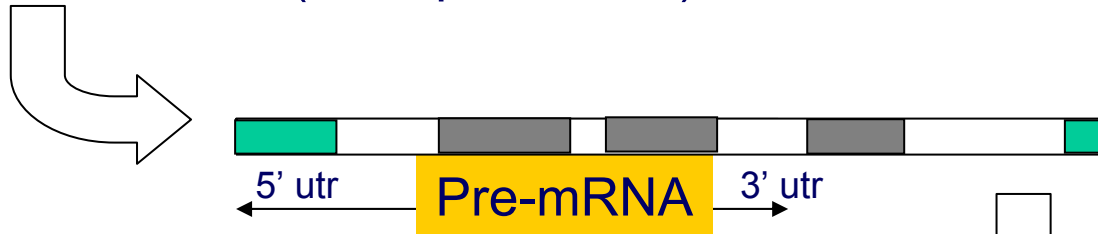
Traduzione (ribosoma)



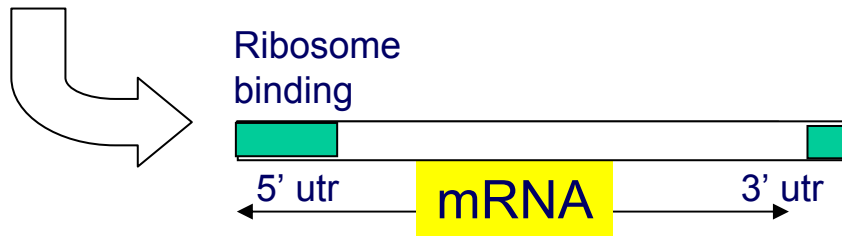
Struttura di un gene eucariota



Trascrizione (RNA polimerasi)

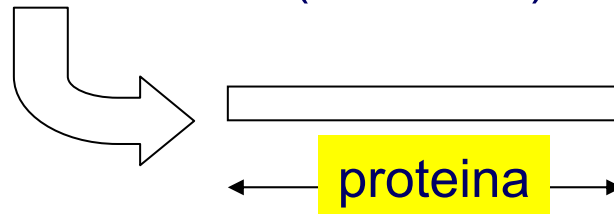


Splicing (Spliceosoma)

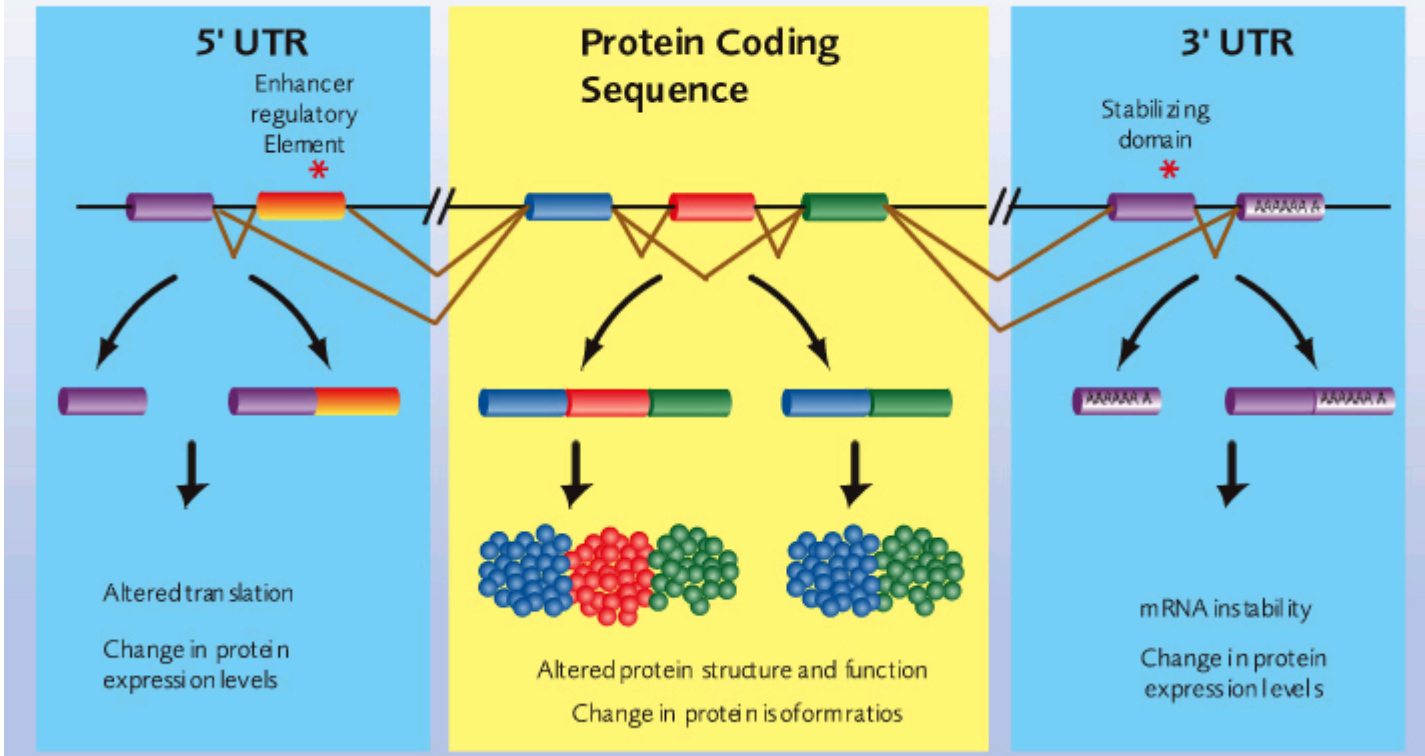


Splicing alternativo

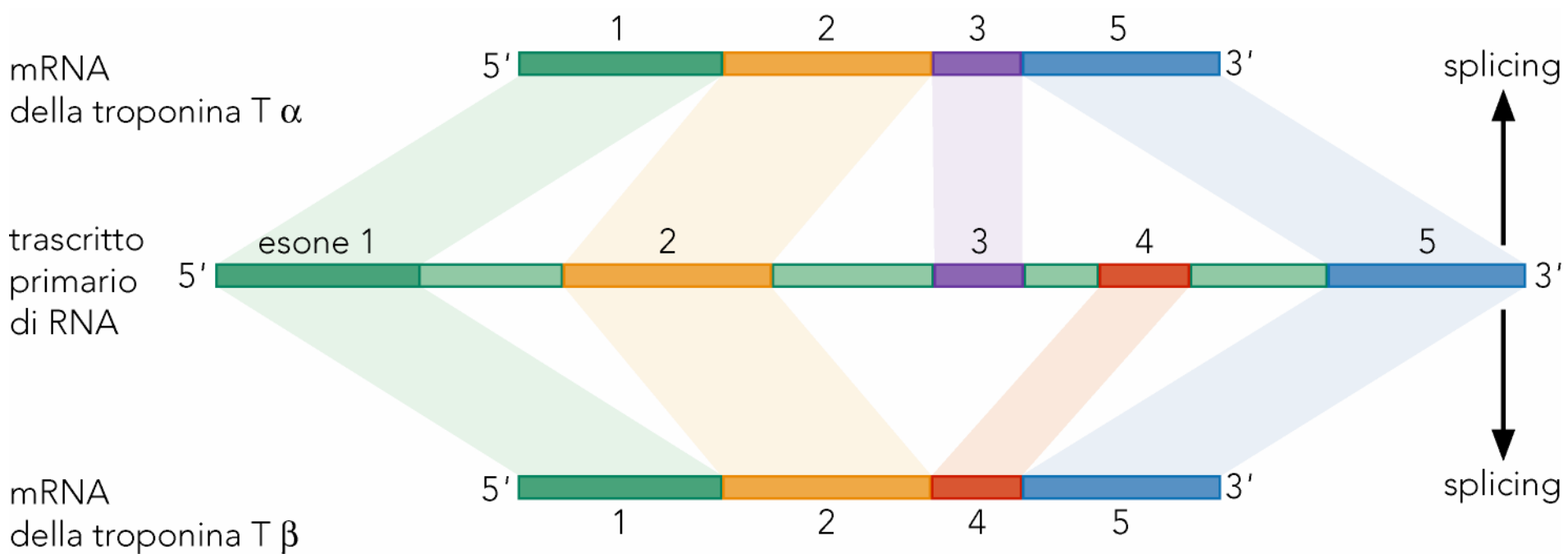
Traduzione (ribosoma)



The impact of alternative RNA splicing

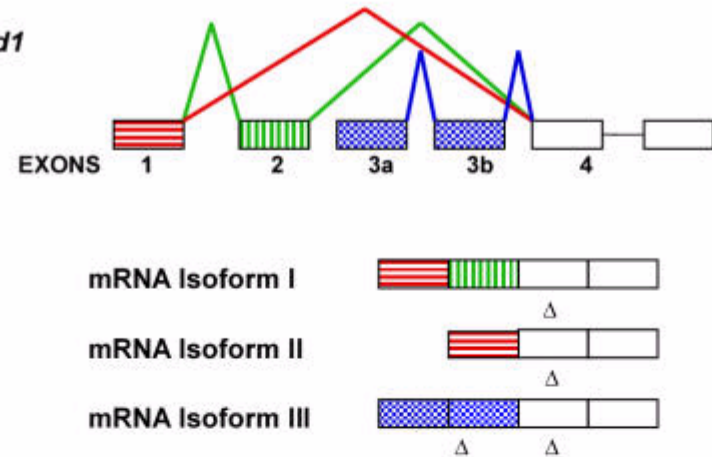


Un esempio: le immunoglobuline possono presentare un dominio che le tiene ancorate alla membrana cellulare, oppure essere solubili

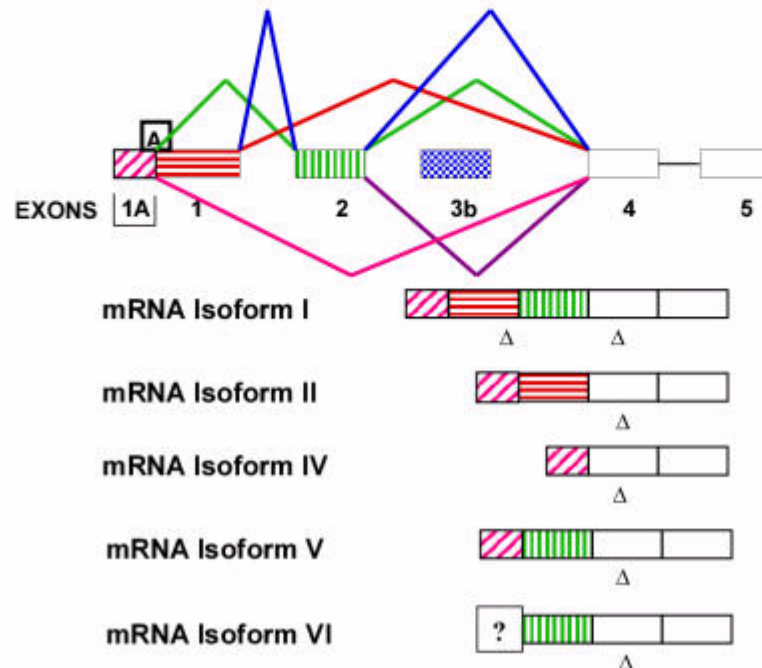


L'organizzazione
del gene della
**tiorredossina
reduttasi 1**
umana e da
topo
e' altamente
conservata ed
entrambi I geni
danno luogo a
splicing
alternativo

Mouse *Txnrd1*



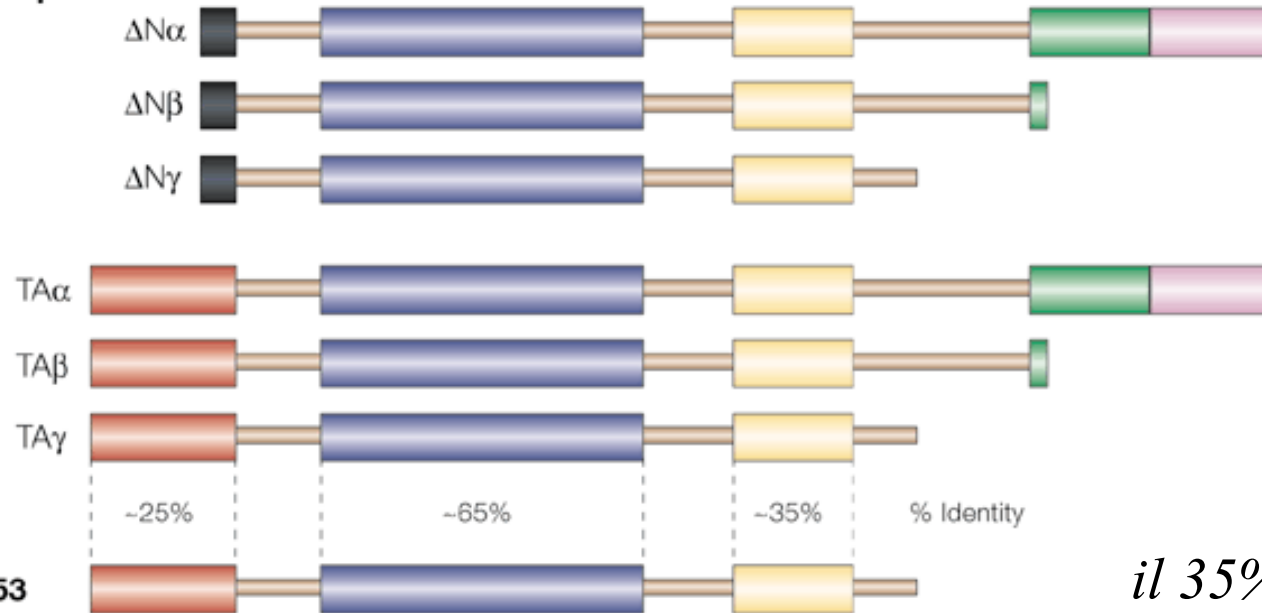
Human *TXNRD1*



TP63/TP73



p63/p73



p53



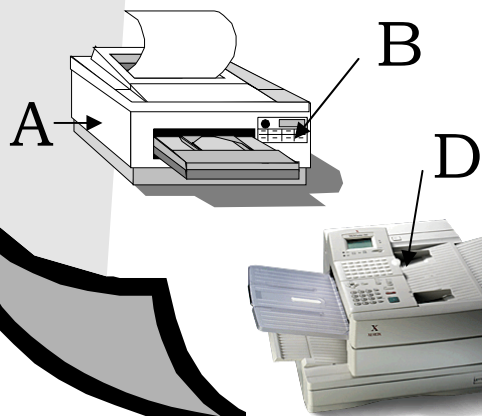
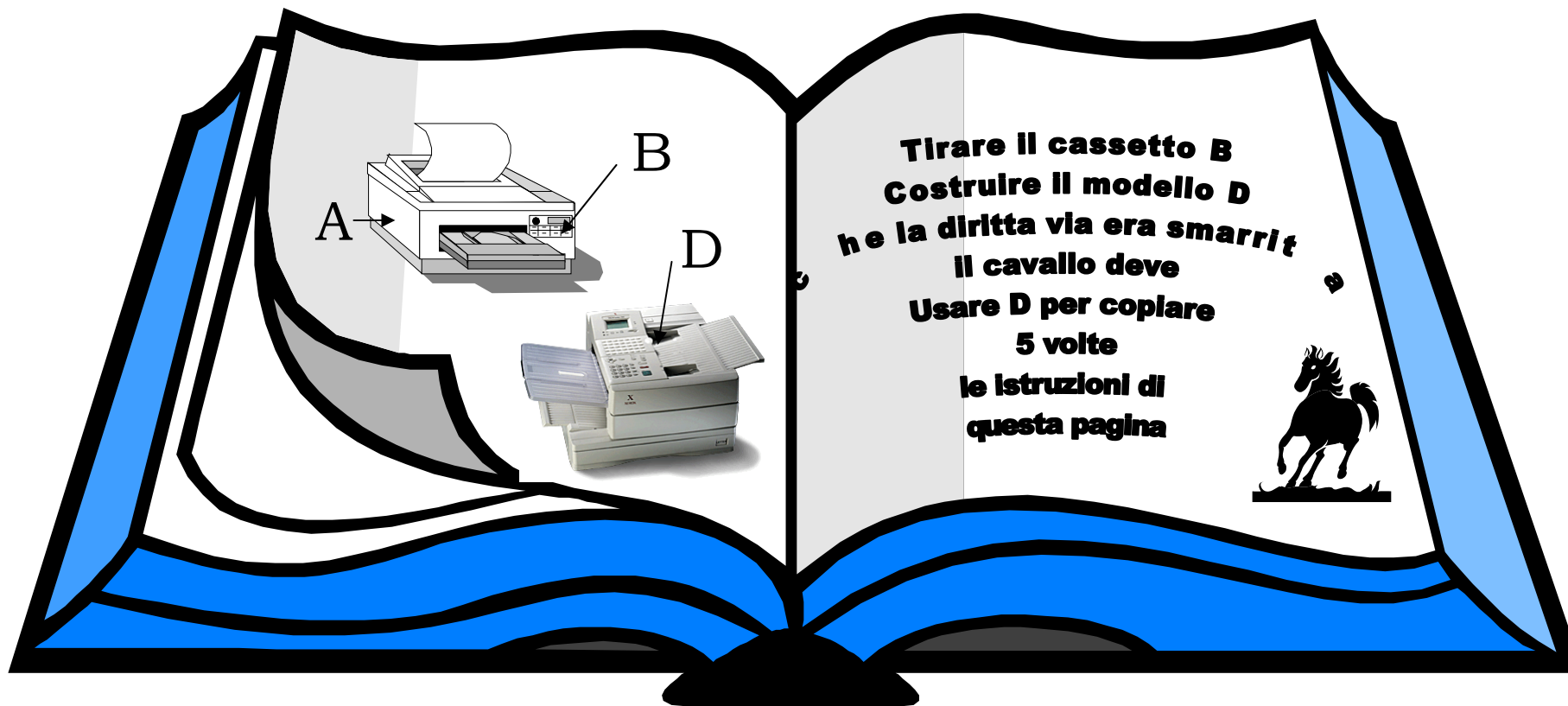
*il 35% dei
geni umani
da' luogo a
splicing
alternativo*



**Aprire lo sportello A
e il mezzo del cammin d
nostra vita,
mi ritrovai**

**Imagie all the people
Per sellare il cavallo
occorre prima di**





Tirare il cassetto B
Costruire il modello D
he la dritta via era smarrit
Il cavallo deve
Usare D per copiare
5 volte
le istruzioni di
questa pagina



Un manuale d'uso:

- non include quasi mai l'informazione che cercate
- nei rari casi in cui la include, e' incomprensibile

Ma le cose potrebbero andare peggio...

Se le istruzioni fossero anche
divise in 27 paragrafi diversi
interrotti da
lunghe pagine di informazioni
irrilevanti, avremmo qualcosa
di molto simile al
gene del retinoblastoma
umano (e a molti altri geni)

Complessita' dei geni

- Geni codificanti per 100Kb
 - E. coli 87
 - S. cerevisiae 52
 - C. elegans 22
 - H. sapiens 5

- | Numero medio | | Lunghezza media | |
|-------------------|---|-----------------|----------|
| di introni per Kb | | geni(Kb) | mRNA(Kb) |
| – S. cerevisiae | 0 | 1.6 | 1.6 |
| – C. elegans | 3 | 4.0 | 3.0 |
| – Drosophila | 3 | 11.3 | 2.7 |
| – Gallus | 8 | 13.9 | 2.4 |
| – H. sapiens | 6 | 16.6 | 2.2 |

La lunghezza approssimativa del DNA del genoma umano e' **3 miliardi di basi**.

Questa sequenza di nucleotidi (A, G, T, C) non è casuale !!!

Teoria dell'informazione (Claude Shannon)

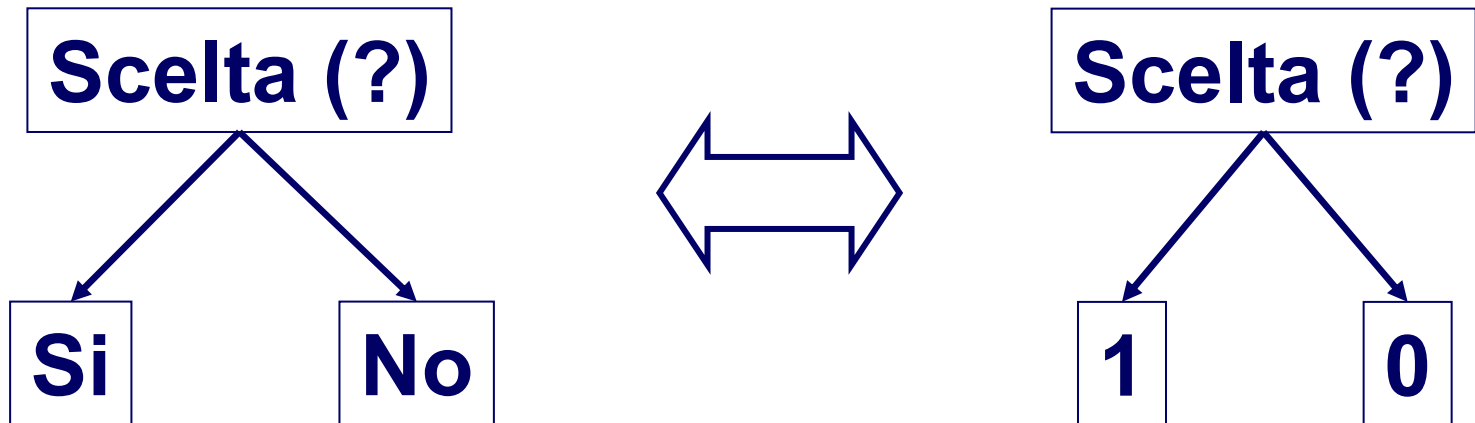
L'**informazione** e' una **misura universale** dell'ordine e puo' essere applicata a qualunque struttura o sistema.

L'**ordine** si riferisce alla disposizione strutturale del sistema.

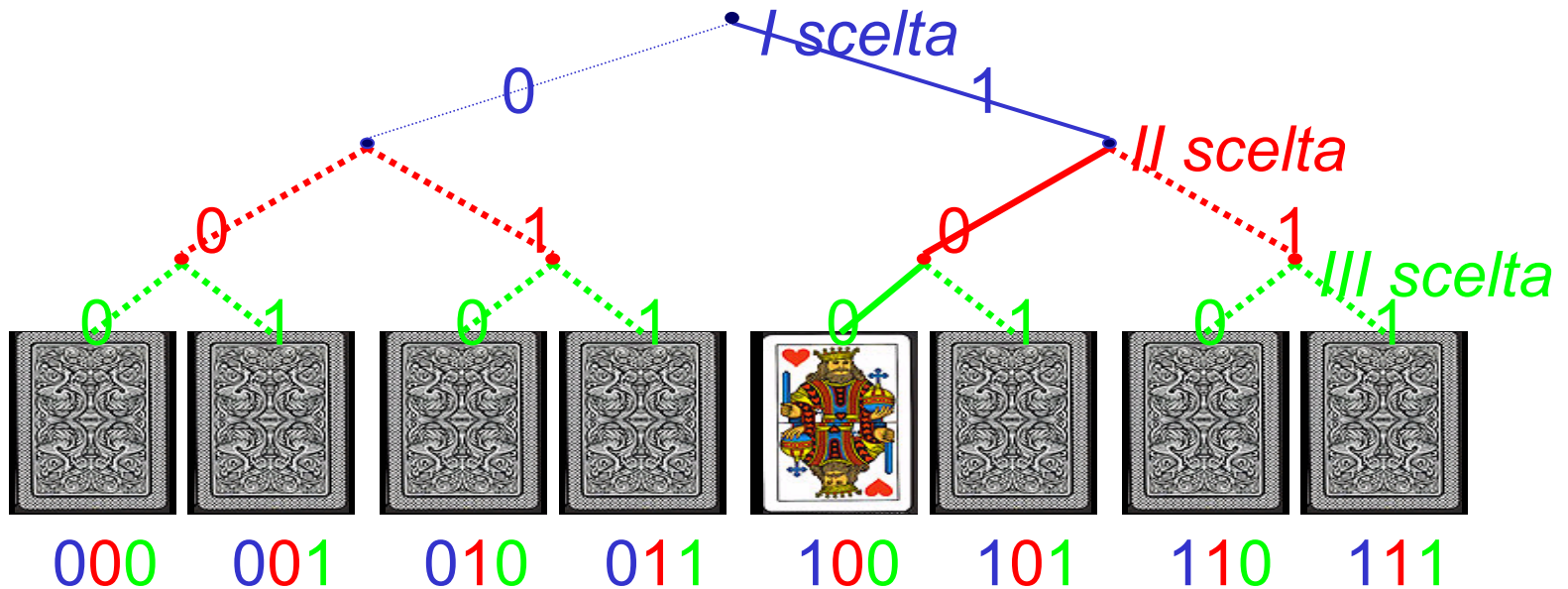


MSKGPVAVGIDLGTTYSCVGVFQHG
KVEIIANDQGNRTTPSYVAFTDTE
RLIGDAAKNQVAMNPTNTVFDAGR
LIGRRFDDAVVQSDMKHWPFMVVN
DAGRPKVQVEYKGETKSFYPEEVS
SMVLTMKKEIAEAYLGKTVTNVAVV
TVPAYFND SQRQATKDAGTIAGLN
VLRINEPTAAAIAYGLDKKVGAE
RNVLI FDLGGGT F DVSILTIEDGI
FEVKSTAGDTHLGGEDFDNRMVNH
FIAEFKRKHKKDISENKRAVRRLR

L'**informazione** quantifica le istruzioni necessarie a produrre una determinata forma di organizzazione e puo' essere raggiunta in termini di **scelte binarie** ed espressa in bit.

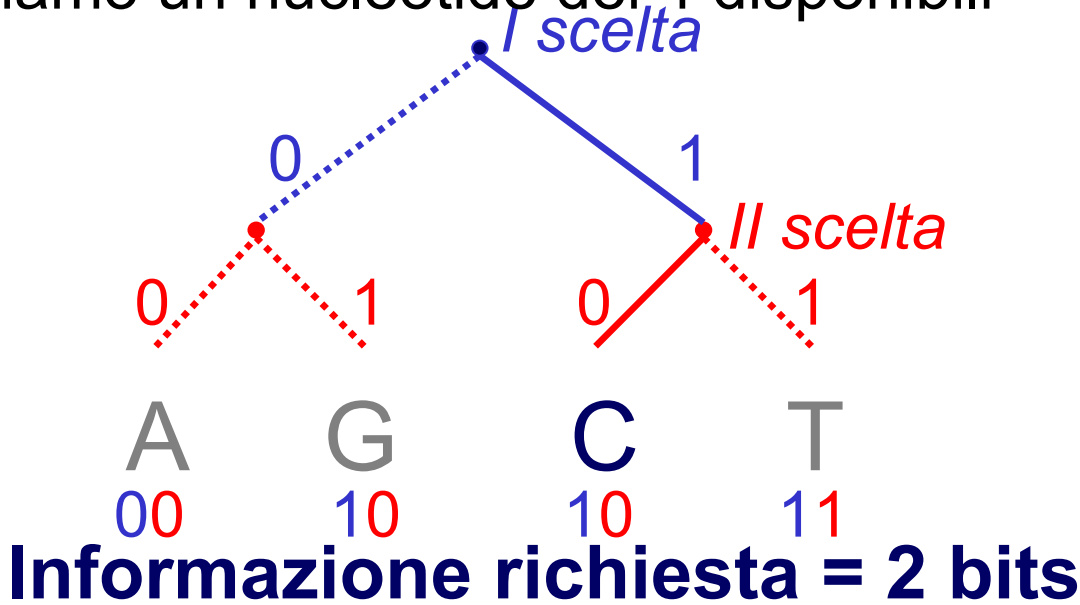


Scegliamo una carta da una pacchetto di 8...



Informazione richiesta = 3 bits

Scegliamo un nucleotide dei 4 disponibili



Quanta informazione
contiene una sequenza
dinucleotidica, es. 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

$$N_{\text{(tot)}} = y^x = 4^2$$

$N_{\text{(tot)}} = n$ di stati possibili

$y = n$ di possibili scelte per ogni posizione
(A, C, G, T)

$x = n$ di posizioni disponibili (1, 2)

La probabilità che la sequenza nucleotidica GC
si generi spontaneamente e' quindi

$$1/4^2 = 1/16 = 0.0625$$

?? Quanto e' ordinato il genoma umano ??

La lunghezza approssimativa del DNA del genoma umano e' $3 \cdot 10^9$.

Trascurando le mutazioni somatiche spontanee, si puo' dire **APPROSSIMATAMENTE** che le molecole di DNA di ogni individuo mostrano la stessa sequenza e calcolarne la capacita' di informazione:

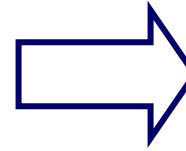
$$N_{(\text{tot})} = 4^{3.000.000.000} \quad \begin{array}{l} x = 3 \cdot 10^9 \\ y = 4 \text{ (A, G, T, C)} \end{array}$$

Il numero di possibilita' stimato $N(\text{tot})$ e' maggiore del numero stimato di particelle presenti nell'universo!!!

Ma la natura ne sceglie **SOLTANTO UNA...**

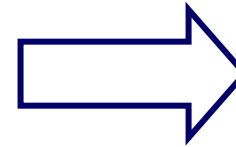
- Il *contenuto informativo* del genoma degli organismi appartenenti alle varie specie e' *altissimo*.
- L'informazione riguardante i sistemi biologici in natura si accumula in maniera graduale attraverso i processi di *variazione casuale del genotipo* e *selezione naturale* (Charles Darwin, *Origin of Species*)

Mutazione genotipica neutrale
o deleteria per il fenotipo



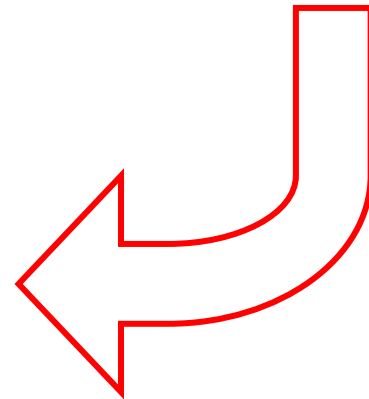
Selezione
negativa

Mutazione genotipica
vantaggiosa per il fenotipo

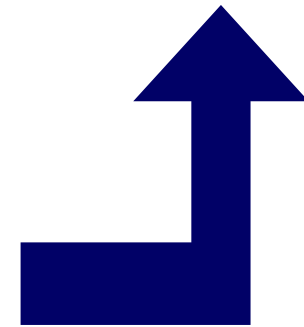
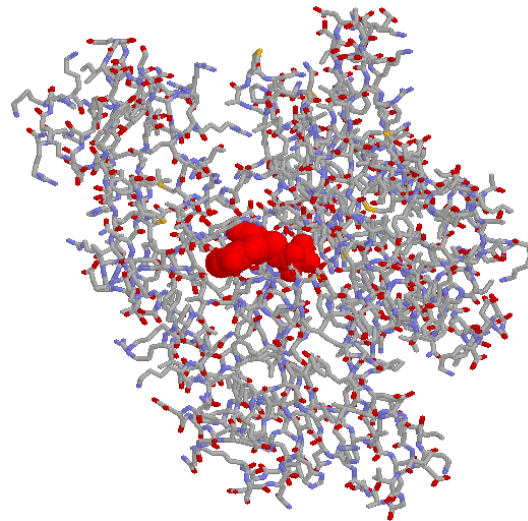
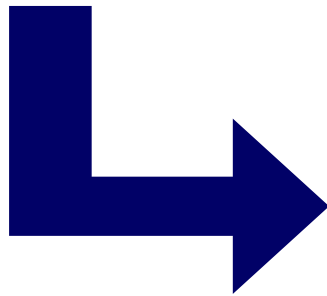
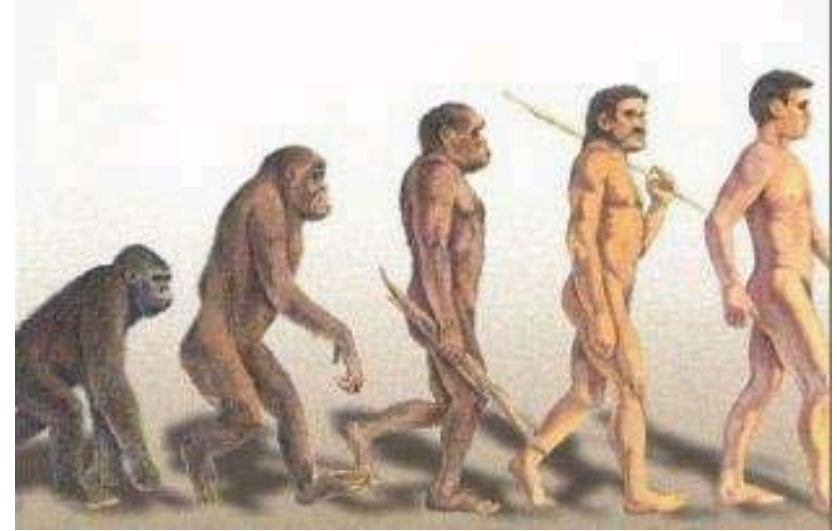
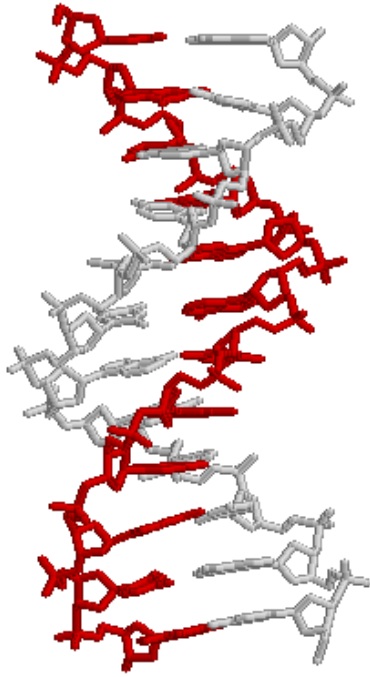


Selezione
positiva

NUOVA FUNZIONE
BIOLOGICA



EVOLUZIONE MOLECOLARE



I principi dell'evoluzione

Tutte le specie viventi si sono evolute da altre specie

Tutte le specie viventi sono legate le une alle altre a vari gradi attraverso progenitori comuni

Tutte le forme di vita sulla terra hanno una origine comune. E' esistita una forma di vita originale che ha dato luogo a tutte le forme successive (*LUCA: Last Universal Common Ancestor*)

Il processo attraverso cui una specie evolve in un'altra coinvolge mutazioni casuali, le mutazioni che risultano in vantaggio di sopravvivenza si diffondono e persistono piu' di quelle neutre o svantaggiose

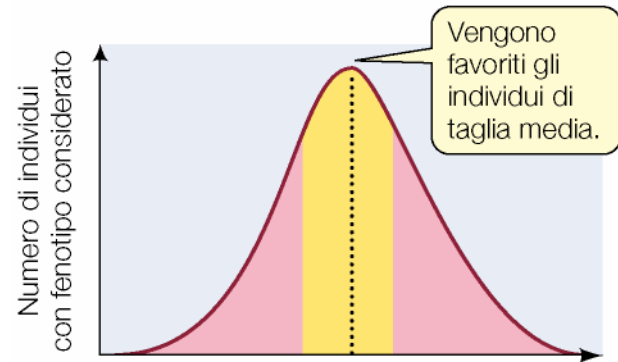
Archeobatteri



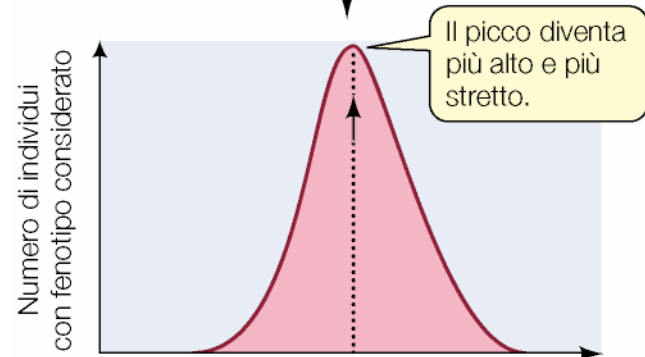
Evoluzione

Progenitore Ancestrale Comune

(a) **La selezione stabilizzante** riduce la variabilità, ma non influisce sulla media.

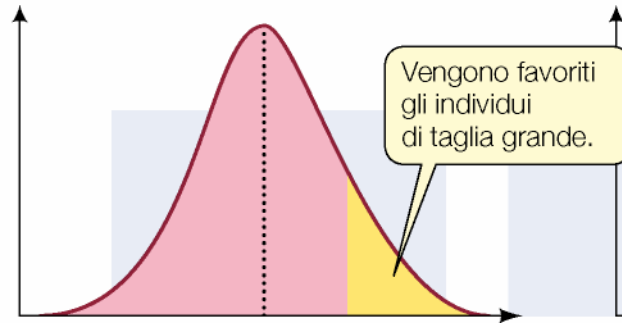


Selezione



Dimensioni corporee

(b) **La selezione direzionale** cambia il valore medio del carattere in questione (in questo caso verso una taglia più grande).

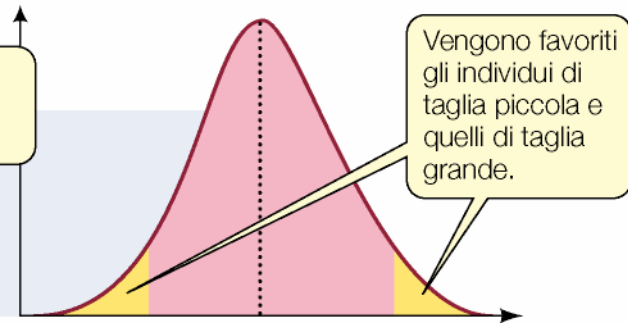


Selezione

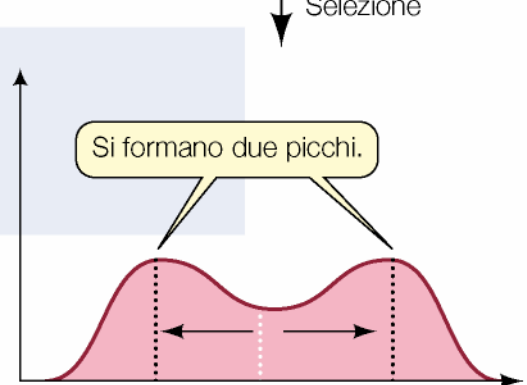


Dimensioni corporee

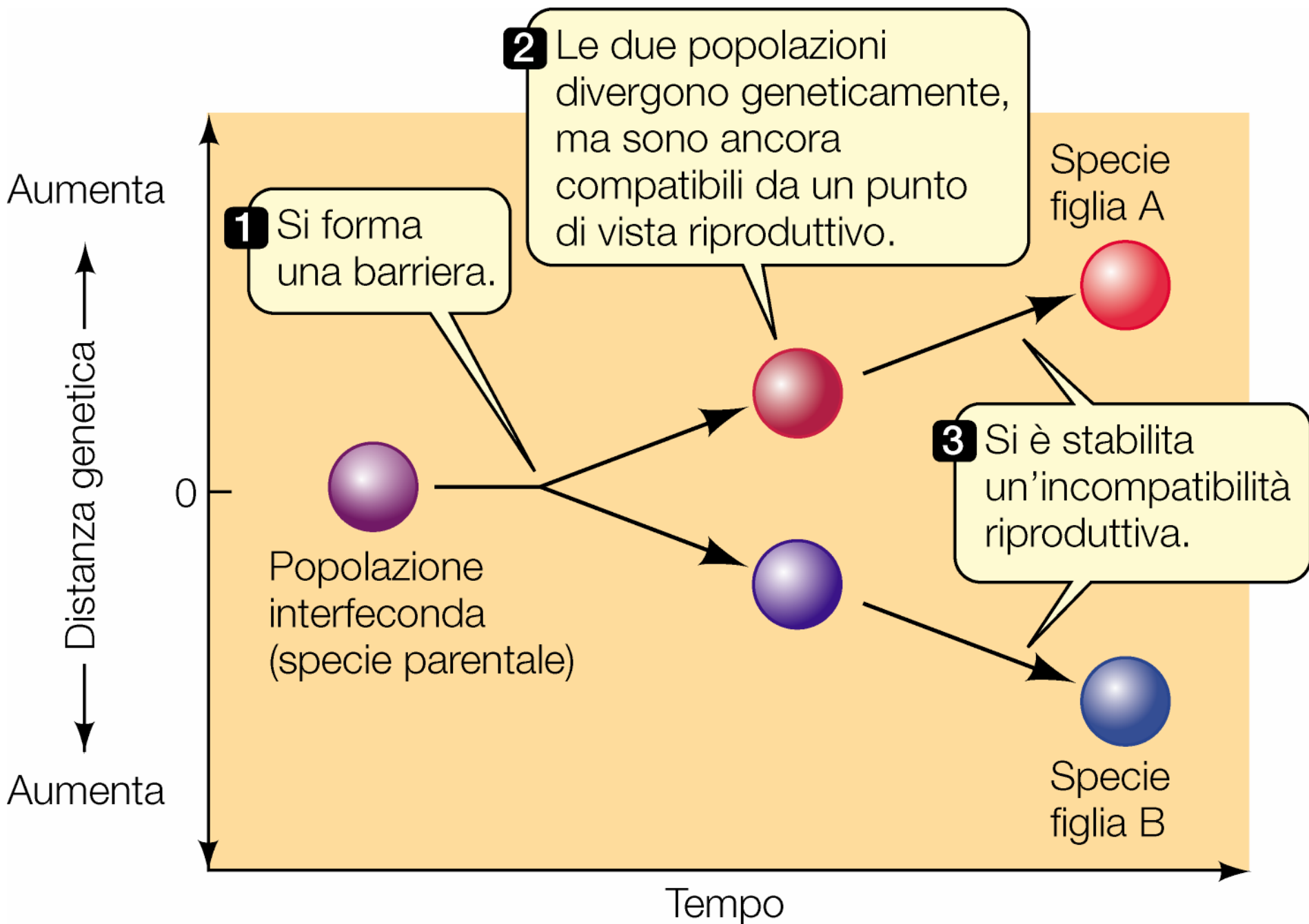
(c) **La selezione diversificante** favorisce entrambi i valori estremi e produce due picchi nella distribuzione del carattere in questione.



Selezione



Dimensioni corporee



Tempo

Una singola specie possiede un'ampia distribuzione geografica.



Il livello del mare aumenta, separando due popolazioni. Le popolazioni si adattano ai diversi ambienti creati 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.



Percentuale della concentrazione attuale di ossigeno

Trascorrono centinaia di milioni di anni prima che le concentrazioni di ossigeno diventino sufficientemente elevate da permettere l'evoluzione degli eucarioti.

Presente

Prime Angiosperme

Colonizzazione delle terre emerse

Primi cordati

Esoscheletri

Organismi pluricellulari

Primi eucarioti

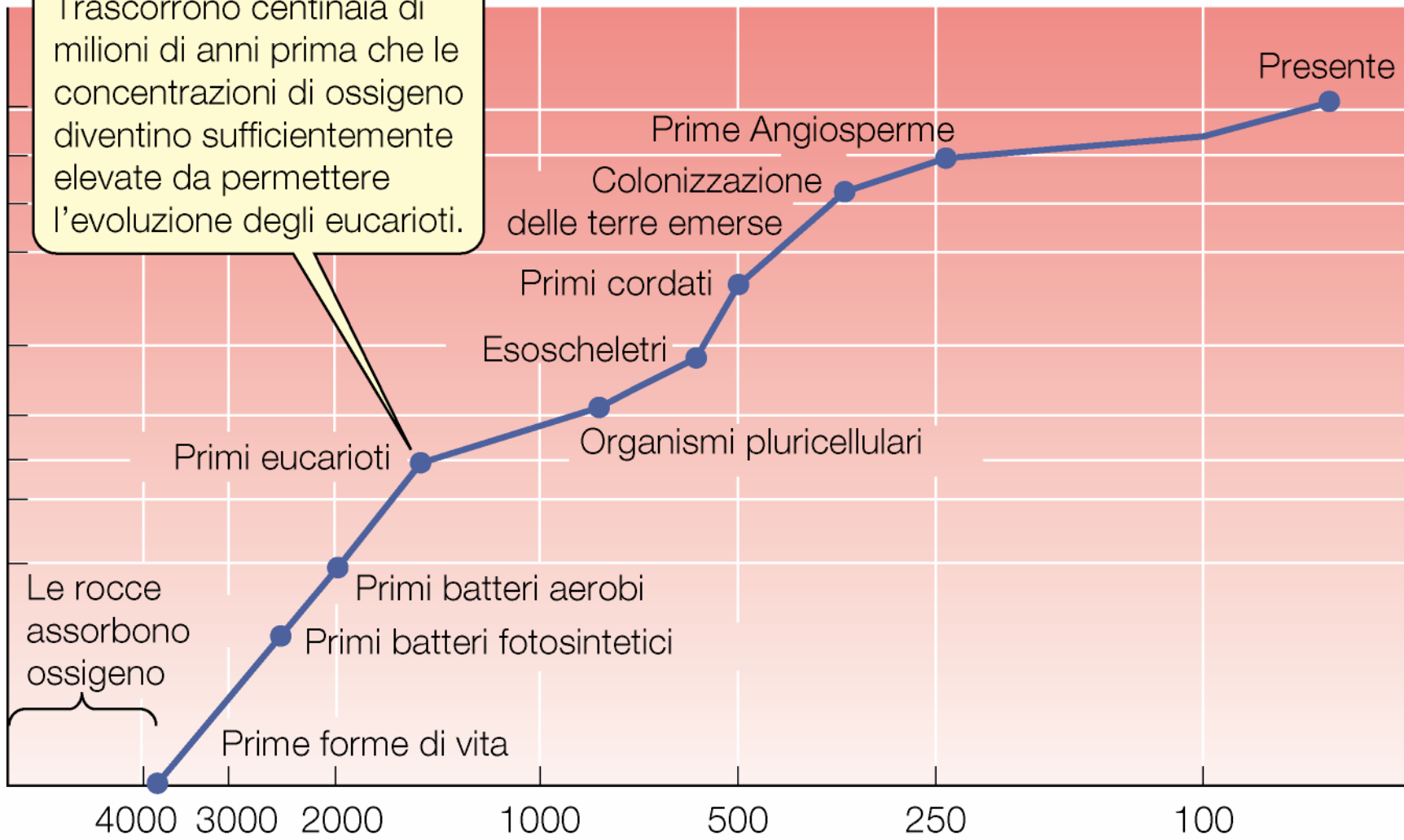
Primi batteri aerobi

Primi batteri fotosintetici

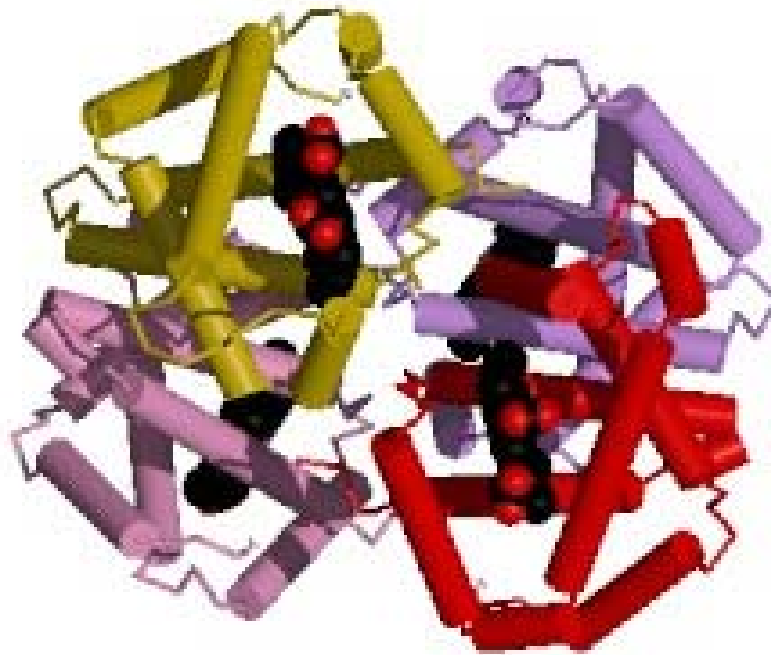
Prime forme di vita

Le rocce assorbono ossigeno

Milioni di anni fa



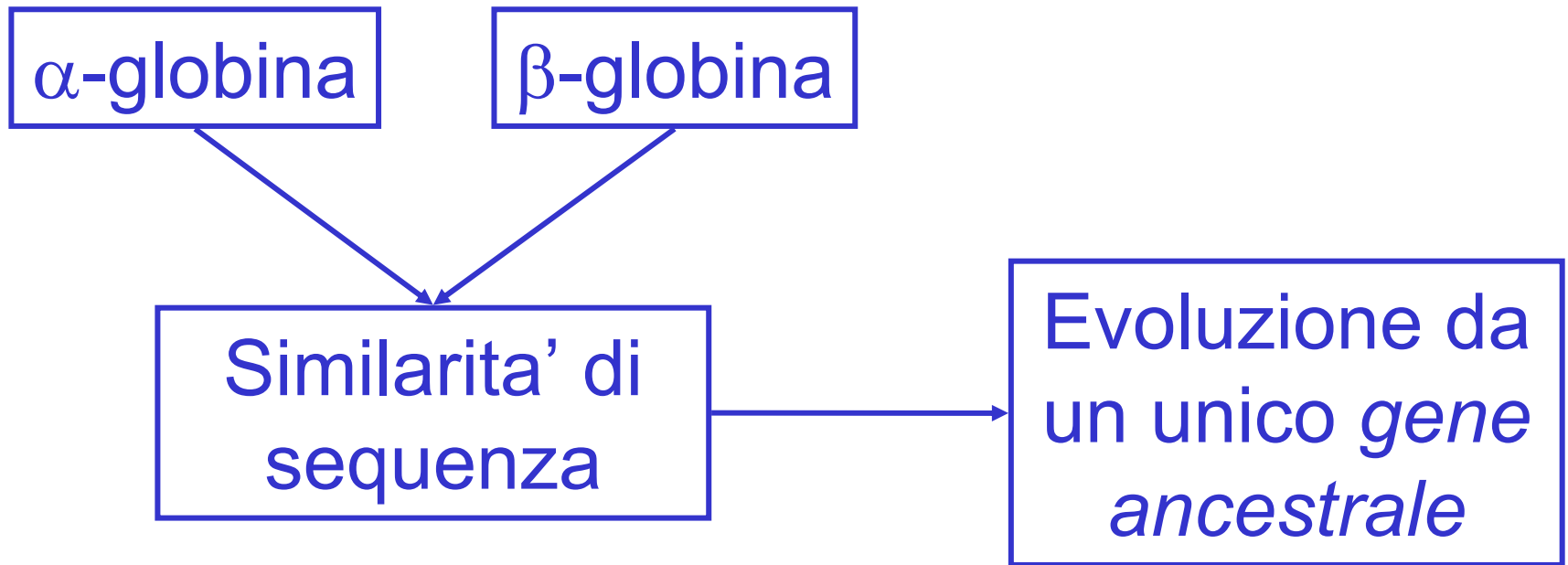
Un esempio di evoluzione a livello molecolare: la famiglia delle globine



L'*emoglobina* e' composta da 4 catene polipeptidiche, chiamate *globine*, necessarie al trasporto delle molecole di O_2 a due a due uguali.

Un esempio: la famiglia delle globine

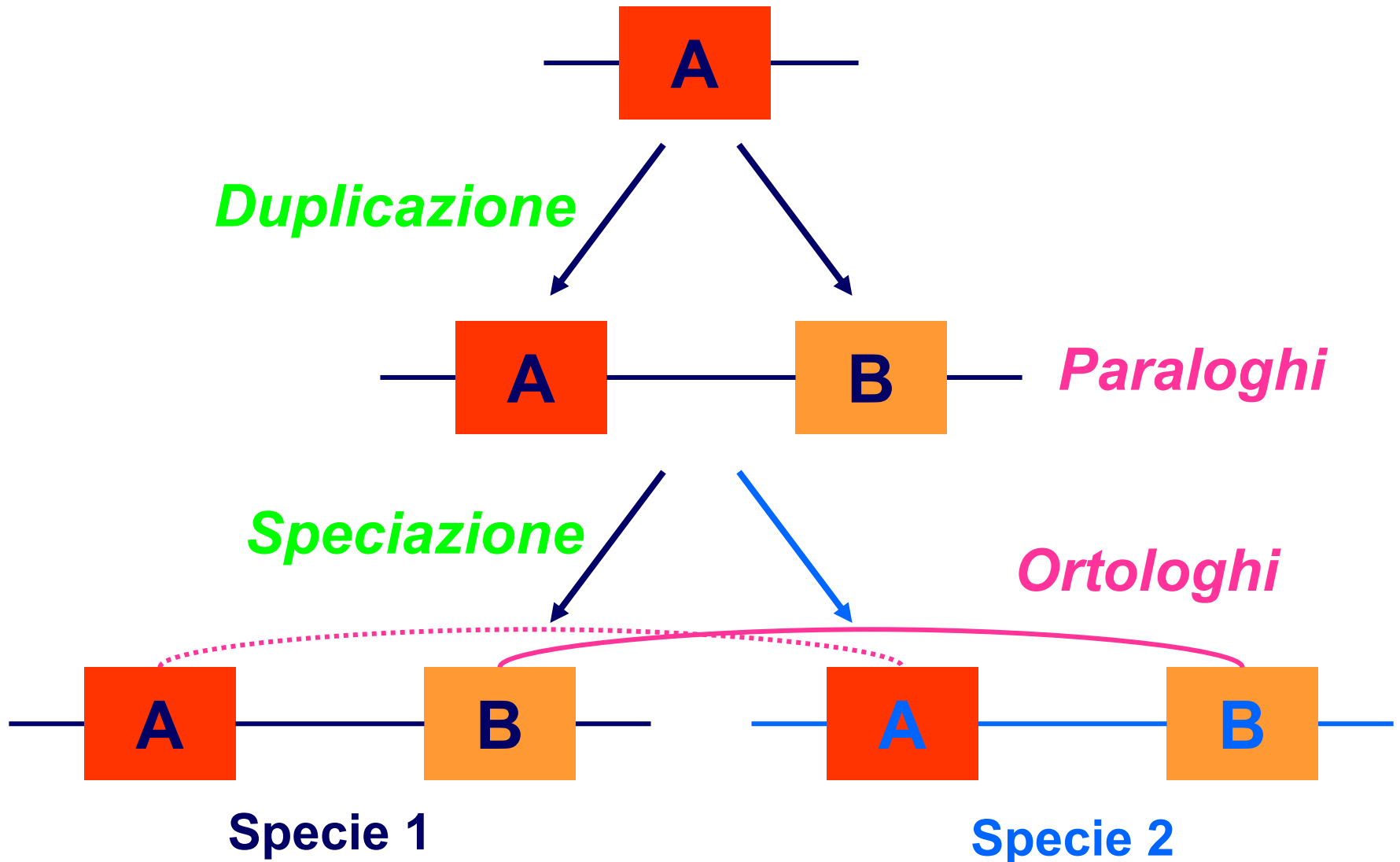
Le globine emoglobiniche possono essere di tipo α o di tipo β .



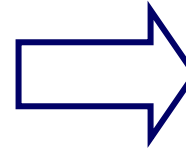
Similarita' e omologia

- Due o piu' proteine possono essere definite omologhe se derivano da un progenitore comune
- L'omologia tra proteine non puo' essere direttamente osservata ma si deduce dalla loro similarita' in sequenza o funzione.
- Due sequenze sono simili se possono essere allineate in modo che molti ammino acidi corrispondenti sono identici o simili.

Due proteine si definiscono omologhe se derivano da un progenitore comune:

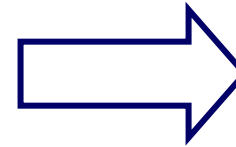


Mutazione genotipica neutrale
o deleteria per il fenotipo



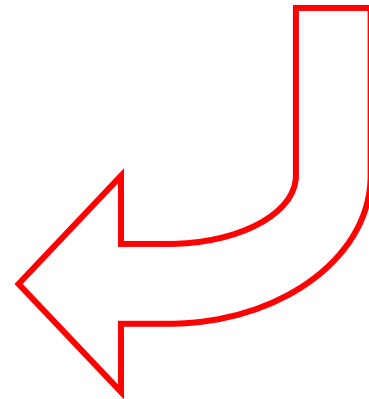
Selezione
negativa

Mutazione genotipica
vantaggiosa per il fenotipo



Selezione
positiva

NUOVA FUNZIONE
BIOLOGICA



Cosa causa le mutazioni genetiche ?

Mutazioni spontanee

Danneggiamenti nel DNA avvengono continuamente

Errori “editoriali” durante la replicazione

Mutazioni chimiche

Agenti chimici dall'ambiente (mutagenesi, carcinogenesi)

Modificazione di basi

Inserzioni tra basi

UV, radiazioni ionizzanti

Cross-linking tra coppie di basi

Aperture di anelli

Rottura di filamenti di DNA

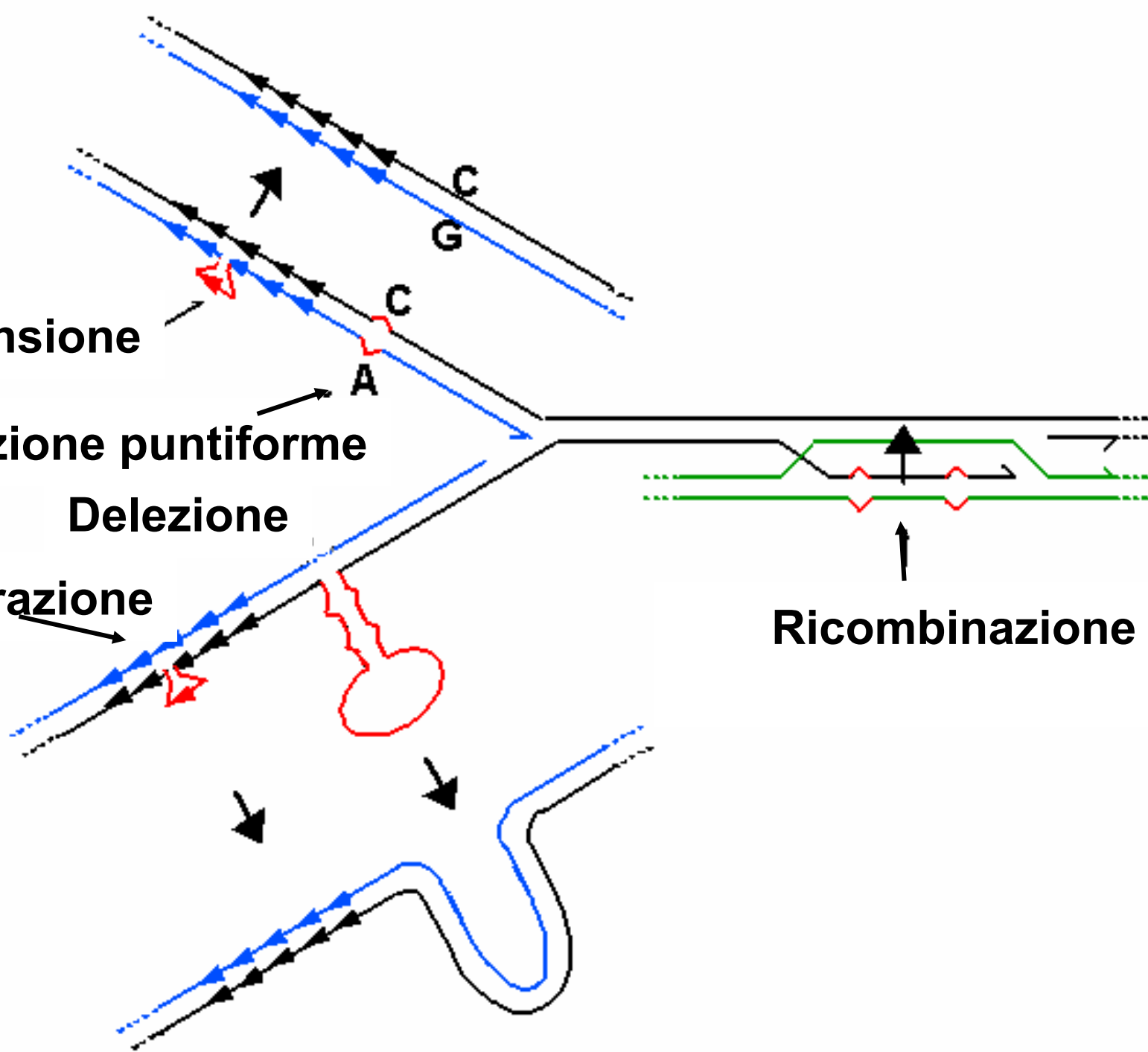
Espansione

Mutazione puntiforme

Delezione

Contrazione

Ricombinazione



con chi vai nel bus

von chi vai nel bus

coc hiv ain elb us

Classificazione delle sostituzioni nucleotidiche

Transizione: sostituzione di una purina con una purina ($A \Leftrightarrow G$) o di una pirimidina con una pirimidina ($C \Leftrightarrow T$)

Trasversione: sostituzione di una purina con una pirimidina ($A \text{ o } G \Leftrightarrow C \text{ o } T$) o viceversa ($C \text{ o } T \Leftrightarrow A \text{ o } G$)

Sinonima: non si verifica variazione dell'amminoacido codificato ($TCT \Leftrightarrow TCC$, entrambe codificano la L)

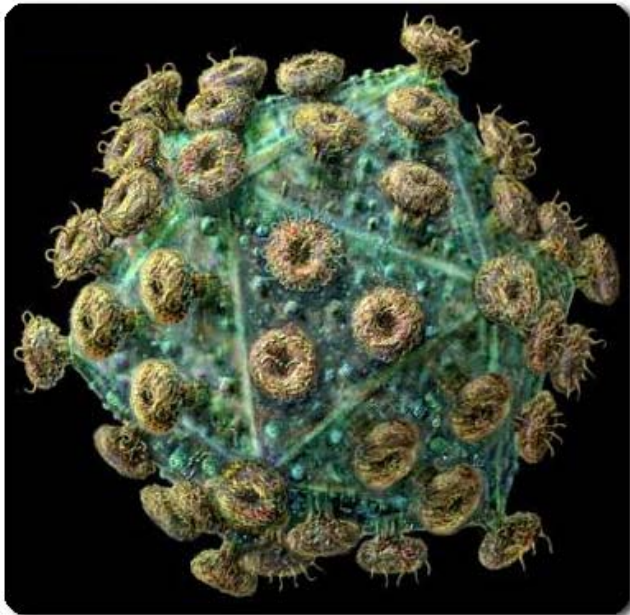
Non-sinonima: si verifica variazione dell'amminoacido codificato ($CTT \Leftrightarrow CCT$, $L \Leftrightarrow P$)

Errori di copia di materiale genetico (DNA/RNA)

Sistema	Tasso d'errore stimato (Mut/N _(Pos))	
Reazione chimica	0.05-0.1 (5-10/100)	
RNA virus (influenza, HIV)	10 ⁻³ -10 ⁻⁵	→ RNA-polimerasi
Procarioti (<i>E. coli</i>)	10 ⁻¹⁰ -10 ⁻¹¹	→
Eucarioti (<i>H. sapiens</i>)	3*10 ⁻⁸	→ DNA-polimerasi & meccanismi riparazione

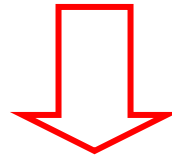
La *variabilità intraspecie* e' responsabile della *sopravvivenza* della specie stessa.

Maggiore e' la variabilità intraspecie, maggiore e' la probabilità che si verifichino mutazioni positive e che la specie riesca ad adattarsi a nuove condizioni ambientali e quindi a sopravvivere piu' a lungo.



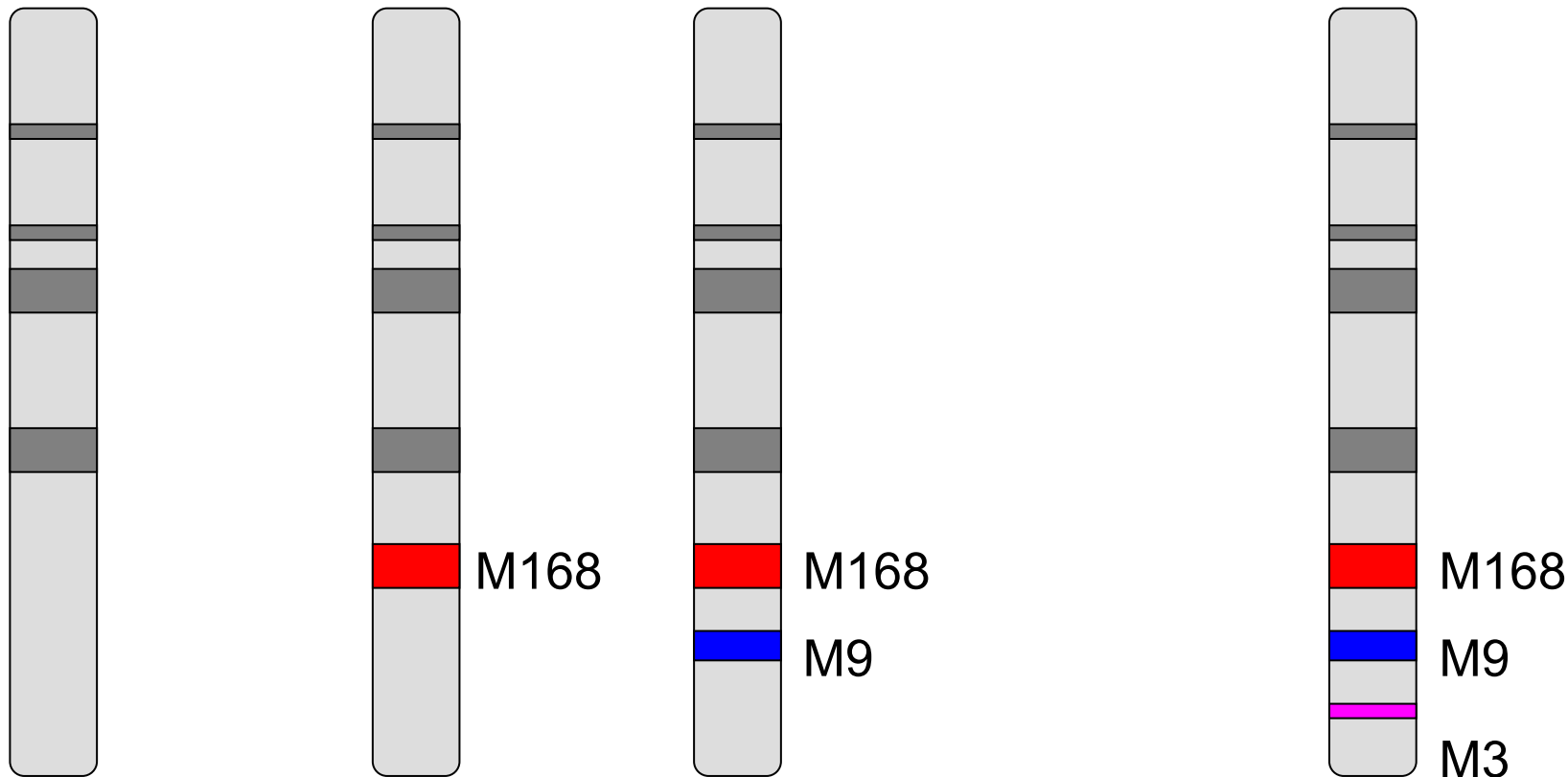
Gli **RNA-virus** (influenza, HIV, polio) sono tra gli organismi geneticamente piu' variabili e per questo sono difficilmente trattabili farmacologicamente

- I risultati delle variazioni genotipiche forniscono la *traccia* molecolare dell'evoluzione;
- Quanto piu' due specie sono evolutivamente vicine tanto piu' simili sono le corrispondenti sequenze genomiche ed i loro prodotti di espressione (cioe' le proteine);
- Laddove le sequenze hanno subito variazioni cosi' grandi da non essere piu' riconoscibili, le corrispondenti strutture 3D proteiche possono aver conservato una similarita' informativa.



Relazioni filogenetiche tra organismi

Il cromosoma Y permette di viaggiare nel tempo



Configurazione
di origine africana

Non-africani

Eurasiatici

Amerindi

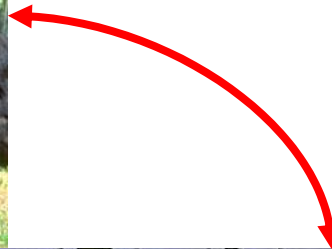
50.000

40.000

10.000



**Variabilita' inter-specie
uomo/scimpanze'
= ~1 %**



**Variabilita' intraspecie
scimpanze' = ~0,4 %**

**Variabilita'
intraspecie uomo
= ~0,1 %**



Popolazioni geneticamente isolate in Europa

lapponi
islandesi
finlandesi
gallesi
baschi





Dove e' il gene?

Metodi per la ricerca di geni

Obiettivo principale della ricerca di geni e' riuscire a distinguere nel genoma tra DNA-genico e DNA-non genico e a localizzare i geni.

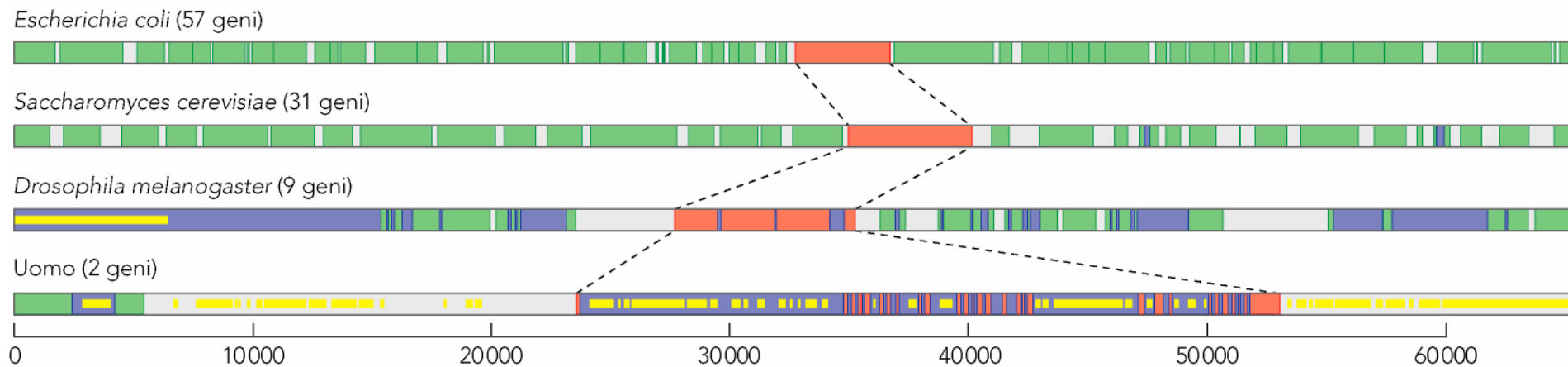
1. Lunghezza degli ORFs (Open Reading Frames)
2. Ricerca di *segnali* nel gene (es. promotori)
3. Differenze nella *composizione nucleotidica*
4. Ricerca di *omologia* (similarita' di sequenza)

Complessita' dei geni

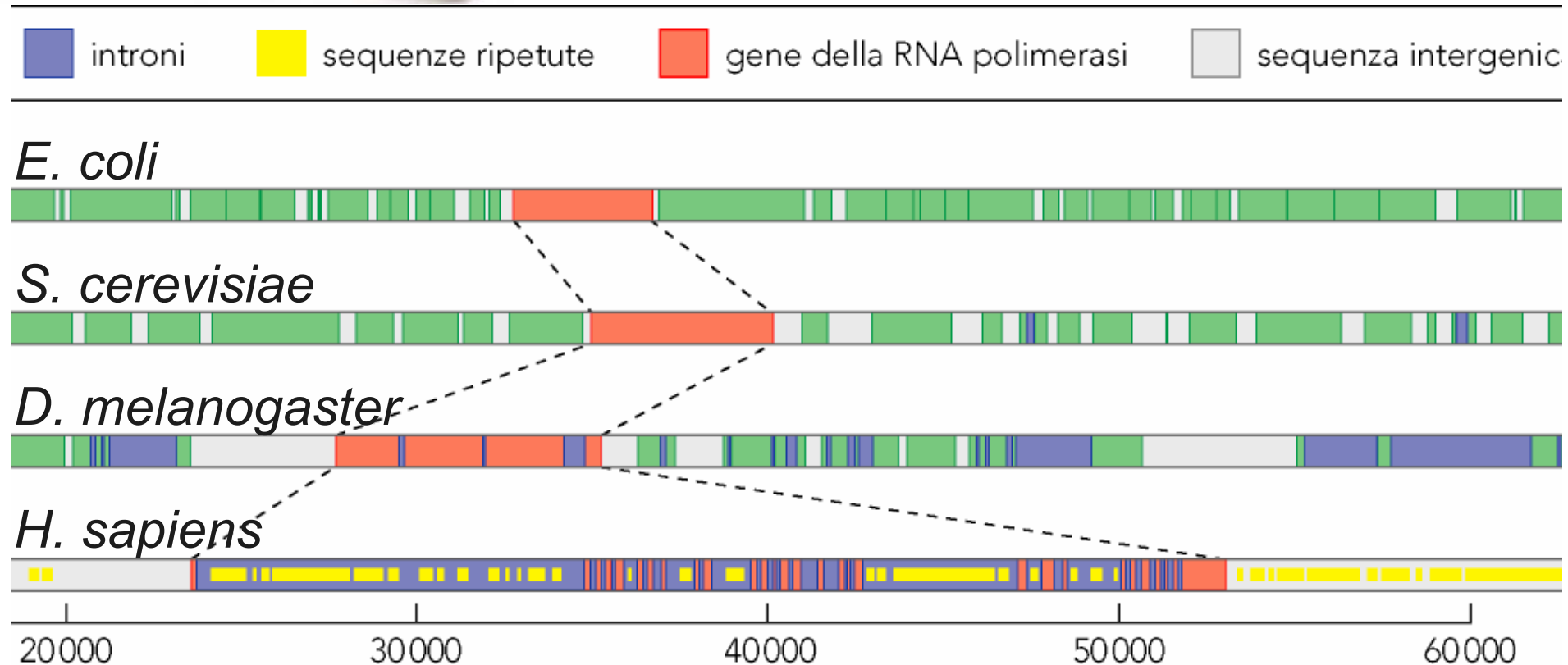
- *Geni codificanti* per 100Kb
 - E. coli 87
 - S. cerevisiae 52
 - C. elegans 22
 - H. sapiens 5
- *N medio di introni*

		gene Kb	mRNA Kb
– S. cerevisiae	0	1.6	1.6
– C. elegans	3	4.0	3.0
– Drosophila	3	11.3	2.7
– Gallus	8	13.9	2.4
– H. sapiens	6	16.6	2.2

Guardiamo in dettaglio una regione di ~ 65 Kb
contenente il gene della RNA polimerasi in 4
diversi genomi: *E. coli*, *S. cerevisiae*, *D.*
melanogaster e *H. sapiens*



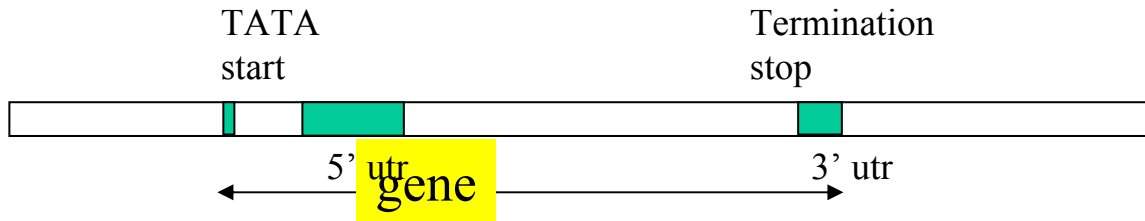
Guardiamo in dettaglio una regione di ~ 65 Kb
contenente il gene della RNA polimerasi in 4
diversi genomi: *E. coli*, *S. cerevisiae*, *D.*
melanogaster e *H. sapiens*



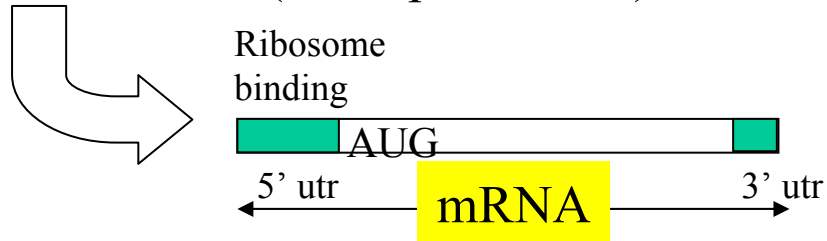
Genomi procariotici

- Piccole dimensioni (< 10 Mb)
- Alta densita' genica (>85% del genoma e' coding)
- Scarsa ridondanza nucleotidica
- Introni (quasi) assenti
- Scarsa complessita' dell'architettura genica e genomica
- Facilmente sequenziabili

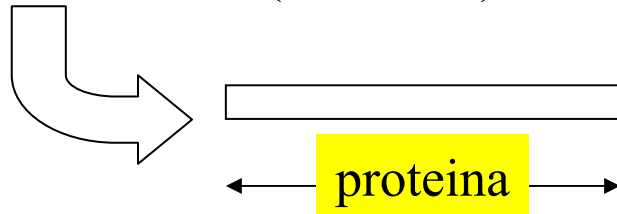
Struttura di un gene procariota



Trascrizione (RNA polimerasi)



Traduzione (ribosoma)



ORFs lunghi nei procarioti

Se assumiamo che i codoni sono distribuiti uniformemente ci aspettiamo uno stop codon ogni 3/64 codoni.

Le proteine sono in media lunghe circa 1000 bp e ogni regione codificante deve contenere solo uno stop codon.

La probabilità casuale di trovare un ORF lungo N codoni preceduto da un ATG e seguito da un codone di terminazione è:

$$P = (1/64) \times (61/64)^N \times (3/64)$$

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

ORFs lunghi nei procarioti

Se assumiamo che i codoni sono distribuiti uniformemente ci aspettiamo uno stop codon ogni 3/64 codoni.

Le proteine sono in media lunghe circa 1000 bp e ogni regione codificante deve contenere solo uno stop codon.

La probabilità casuale di trovare un ORF lungo N codoni preceduto da un ATG e seguito da un codone di terminazione è:

$$P = (1/64) \times (61/64)^N \times (3/64)$$

Per N=10

$$P = (1/64) \times (61/64)^N \times (3/64) = 0.045$$

Per N=100

$$P = (1/64) \times (61/64)^N \times (3/64) = 0.0006$$

Genomi eucariotici

- Dimensioni piu' grandi: da 13 Mb per i funghi piu' semplici a 10.000 Mb per alcune piante superiori
- Architettura genica e genomica complessa in proporzioni alla complessita' dell'organismo. DNA coding: 70% *S. cerevisiae*, 25% *D. melanogaster*, 1-3% vertebrati e piante superiori
- Forte presenza di introni
- Forte ridondanza nucleotidica

Geni eucariotici

I geni eucarioti sono piu' complessi:

Promotore - 5'UTR - Esone - Introne - Esone - ... 3'UTR - *PolyA*



Sito di poli-
adenilazione

Il segnale per un introne e' di sole 2 bp (AG....GU)

In media un gene di vertebrati e' ~ 30Kb, la regione codificante solo 1-3Kb. In media contiene 6 esoni (~ 150 bp ognuno)

La regione 5'UTR e' lunga ~ 750bp

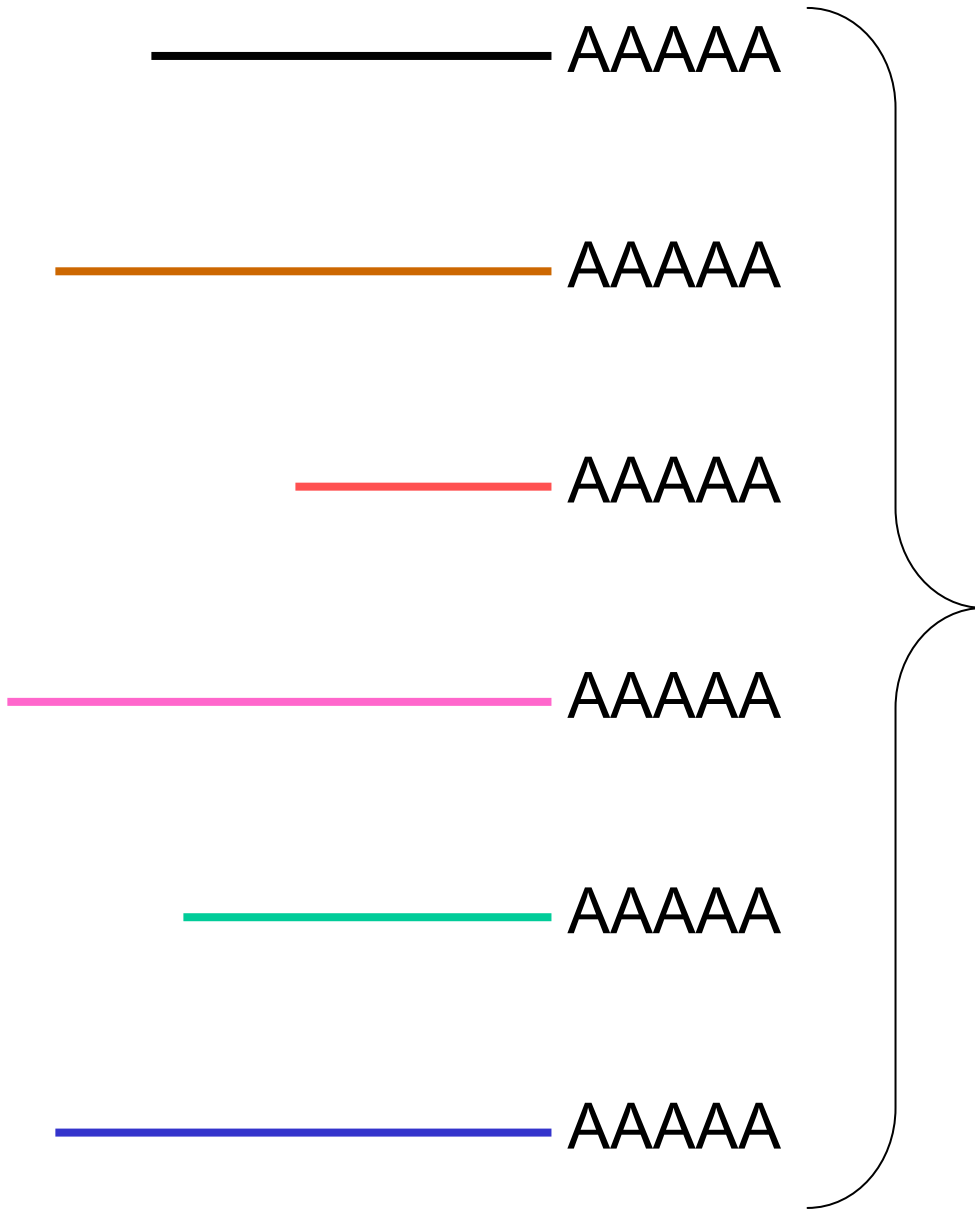
La regione 3'UTR e' lunga ~450bp

Metodi basati su similarita'

- Confronto con il database di EST (cDNA).
- Confronto della sequenza genomica tradotta con i database di proteine.
- Confronto della sequenza genomica con sequenze genomiche omologhe
- Confronto della sequenza genomica con sequenze di cDNA
- Spliced alignment.

BANCA DATI EST
(Expressed Sequence Tag)

Seleziona e ottieni la
sequenza di tutti gli mRNA
contenenti polyA



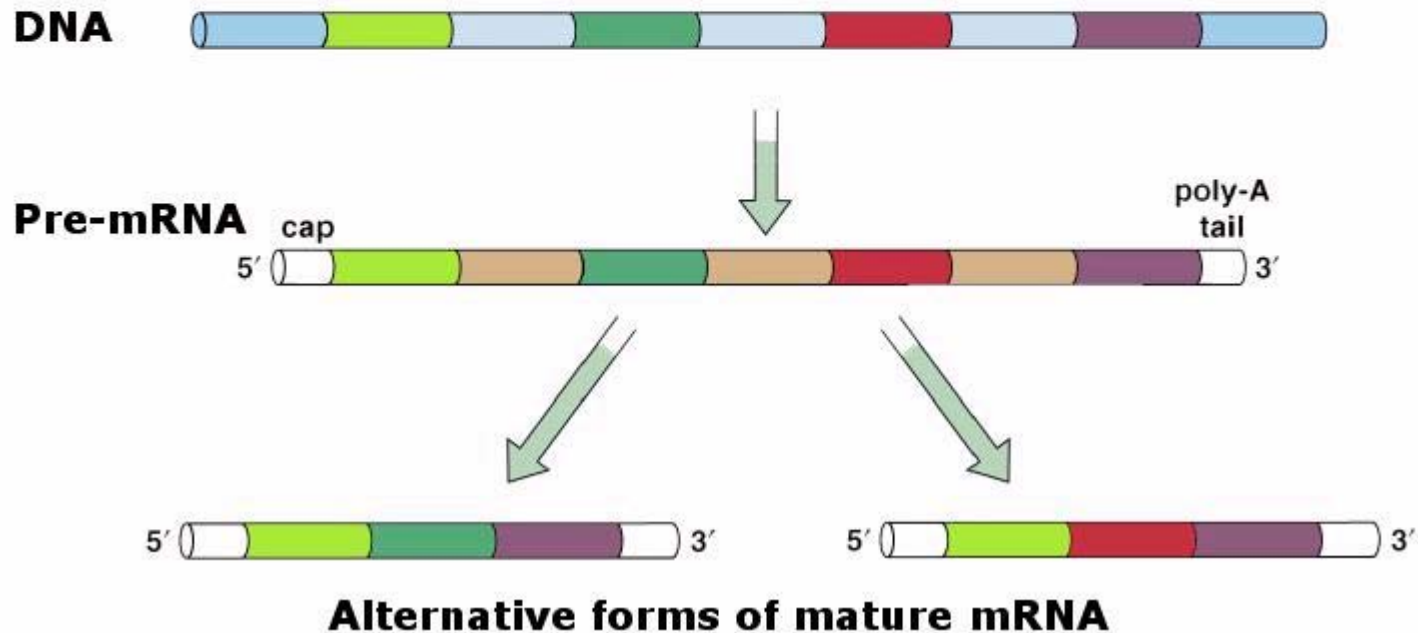
Spliced Alignment

Dati una sequenza genomica ed un insieme di esoni possibili (cioe' tutti i possibili frammenti fra i siti donatori e accettori di splicing):

- costruzione di tutte le possibili catene di esoni
- traduzione del corrispondente mRNA nella corrispondente sequenza proteica
- confronto con una serie di sequenze proteiche fino a che non si rileva una proteina simile.

Alternative RNA splicing

RNA structure can also be rearranged by alternative splicing.



Differenze di composizione nucleotidica

Le frequenze di ricorrenza dei codoni (cioe' le frequenze con cui i codoni si trovano nelle regioni codificanti) possono costituire un indice di *genicità* del DNA.

In che rapporto di frequenza si trovano alanine e triptofani nelle sequenze di proteine note?

Quante volte troviamo A o T nella terza posizione di un codone?
E quante volte G o C?

Differenze di composizione nucleotidica

Le frequenze di ricorrenza dei codoni (cioe' le frequenze con cui i codoni si trovano nelle regioni codificanti) possono costituire un indice di *genicità* del DNA.

Esempio 1: L, A e W (leucina, alanina e triptofano) sono codificati rispettivamente da 6, 4 e 1 codoni diversi e quindi, nella traduzione di un DNA casuale questi amminoacidi si dovrebbero trovare nel rapporto 6:4:1. Tuttavia, in una proteina di solito si osserva il rapporto 6:5:1 → DNA-coding contiene i codoni in rapporto 6:5:1

Quante volte troviamo A o T nella terza
posizione di un codone?
E quante volte G o C?

Differenze di composizione nucleotidica

Le frequenze di ricorrenza dei codoni (cioe' le frequenze con cui i codoni si trovano nelle regioni codificanti) possono costituire un indice di *genicita'* del DNA.

Esempio 1: L, A e W (leucina, alanina e triptofano) sono codificati rispettivamente da 6, 4 e 1 codoni diversi e quindi, nella traduzione di un DNA casuale questi amminoacidi si dovrebbero trovare nel rapporto 6:4:1. Tuttavia, in una proteina di solito si osserva il rapporto 6:5:1 → DNA-coding contiene i codoni in rapporto 6:5:1

Esempio 2: in un gene, A o T sono nella terza posizione di un codone il 90% delle volte, G e C solo il 10% (le proporzioni esatte dipendono dalla specie), mentre in un DNA casuale tale rapporto sara' 50%-50%.

Ricerca di *segnali* nel gene

I ***segnali*** genici sono unita' discrete con una sequenza consenso riconoscibile (promotori, siti di poliadenilazione, siti donatori o accettori di splicing)

Gli algoritmi utilizzati per ricercare segnali genici sono chiamati sensori di segnali (***signal sensors***).

I sensori di segnali possono ricercare il segnale ***isolatamente o in maniera contestuale*** (cioe' relativamente ad altri segnali locali)

Ricerca di promotori

Le sequenze codificanti sono fiancheggiate da regioni promotore, che non hanno una sequenza specifica.

Si possono definire sequenze consenso a cui tutti i promotori somigliano.

Per esempio, in *E. Coli* dall'analisi di 263 promotori e' stato derivato il consenso:

5' ---⁻³⁵**TTGACA**--17bp--⁻¹⁰**TATAAT**--10bp--**trascrizione**

Di solito si utilizzano **misure statistiche:**

5' --- **TTGACA** ---17 bp--- **TATAAT** ---10 bp--- **trascrizione**

	1	2	3	4	5	6
A	0.1	0.1	0.1	0.6	0.1	0.5
T	0.7	0.7	0.1	0.1	0.1	0.3
G	0.1	0.1	0.6	0.1	0.1	0.1
C	0.1	0.1	0.2	0.2	0.7	0.1
	T	T	G	A	C	A
Sequenza consenso (<i>consensus</i>)						

Come possiamo utilizzare un
consensus derivato da casi noti
per **predire** nuovi casi?

Sequenza consensus per un segnale es. promotori, sito di splicing, ...

	1	2	3	4	5	6
A	0.30	0.50	0.10	0.45	0.65	0.32
T	0.20	0.15	0.05	0.20	0.05	0.25
C	0.45	0.20	0.80	0.30	0.25	0.40
G	0.05	0.15	0.05	0.05	0.05	0.03

La probabilità che la sequenza **GGTCAQ**AACCCAGCG appartenga alla regione usata per derivare la matrice precedente è il prodotto delle probabilità cioè :

se consideriamo la prima G come corrispondente alla prima posizione

$$0.05 * 0.15 * 0.05 * 0.30 * 0.65 * 0.40 = 2.92 * 10^{-5}$$

Sequenza consensus per un segnale es. promotori, sito di splicing, ...

	1	2	3	4	5	6
A	0.30	0.50	0.10	0.45	0.65	0.32
T	0.20	0.15	0.05	0.20	0.05	0.25
C	0.45	0.20	0.80	0.30	0.25	0.40
G	0.05	0.15	0.05	0.05	0.05	0.03

La probabilità che la sequenza **GGTCACA**ACCCAGCG appartenga alla regione usata per derivare la matrice precedente è il prodotto delle probabilità cioè :

se consideriamo la prima G come corrispondente alla prima posizione

$$0.05 * 0.15 * 0.05 * 0.30 * 0.65 * 0.40 = 2.92 * 10^{-5}$$

se invece partiamo dalla seconda G :

$$0.05 * 0.15 * 0.80 * 0.45 * 0.25 * 0.32 = 2.59 * 10^{-5} \text{ e così via.}$$

La sequenza AACCCA ha una probabilità di

$$0.30 * 0.50 * 0.80 * 0.30 * 0.25 * 0.32 = 2.88 * 10^{-3}$$

TUTTE da confrontare con la probabilita' casuale:

$$0.25 * 0.25 * 0.25 * 0.25 * 0.25 * 0.25 = 2.16 * 10^{-4}$$

Dividiamo ciascun valore nella matrice sito specifica precedente per 0.25 (**valore casuale**), ne facciamo il log in base 2 e otteniamo:

	1	2	3	4	5	6
A	0.26	1.00	-1.32	0.85	1.38	0.36
T	-0.32	-0.74	-2.32	-0.32	-2.32	0.00
C	0.85	-0.32	1.68	0.26	0.00	0.68
G	-2.32	-0.74	-2.32	-2.32	-2.32	-3.06
Caso 1 = -3.06	G -2.32	G -0.74	T -2.32	C +0.26	A +1.38	C +0.68
Caso 2 = -0.17	G -2.32	T -0.74	C +1.68	A +0.85	C +0.00	A +0.36
Caso 3 = +3.56	A +0.26	A +1.00	C +1.68	C +0.26	C +0.00	A +0.36

Problema dello 0!!!

PSEUDOCOUNTS

E' una probabilità significativa?

Se $F=0.25 \rightarrow F/0.25=1 \rightarrow \ln(F/0.25) = 0$

Se $F<0.25 \rightarrow F/0.25<1 \rightarrow \ln(F/0.25) < 0$ neg

Se $F>0.25 \rightarrow F/0.25>1 \rightarrow \ln(F/0.25) > 0$ pos

dove *F* è la frequenza di un evento, ad es. di osservare un dato nucleotide in una data posizione
e *1/2* è la probabilità casuale di osservare quell'evento

E' una probabilità significativa?

Se $F=P_{\text{caso}} \rightarrow F/ P_{\text{caso}} =1 \rightarrow \ln(F/ P_{\text{caso}}) = 0$

cioe' se la frequenza con cui osserviamo qualcosa e' uguale a quella che ci aspettiamo per caso il valore corrispondente in una matrice log-odd e' "0"

***Se lanciamo un dado molte volte
e otteniamo "6"***

1/6 delle volte



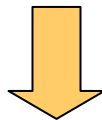
E' una probabilità significativa?

Se $F < P_{\text{caso}} \rightarrow F / P_{\text{caso}} < 1 \rightarrow \ln(F / P_{\text{caso}}) < 0$

cioe' se la frequenza con cui osserviamo qualcosa e' piu' bassa di quella che ci aspettiamo per caso il valore corrispondente in una matrice log-odd e' **negativo**

***Se lanciamo un dado molte volte
e otteniamo "6"***

1/10 delle volte



Il dado e' truccato

La **possibilita'** di numeri alti e' **sfavorita** rispetto al caso



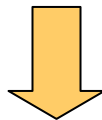
E' una probabilità significativa?

Se $F > P_{\text{caso}} \rightarrow F / P_{\text{caso}} > 1 \rightarrow \ln(F / P_{\text{caso}}) > 0$

cioe' se la frequenza con cui osserviamo qualcosa e' piu' alta di quella che ci aspettiamo per caso il valore corrispondente in una matrice log-odd e' **positivo**

Se lanciamo un dado molte volte e otteniamo "6"

1/2 delle volte



Il dado e' truccato

La **possibilita'** di numeri alti e' **favorita** rispetto al caso



1. Introduzione alla bioinformatica
2. Evoluzione e informazione
3. Ricerca di geni in genomi di procarioti ed eucarioti