Che dati si producono e analizzano ?

Spoiler sui dati biologici

I Dati Biologici

Esistono svariate tipologie di dati biologici che la bioinformatica assiste, derivati soprattutto dallo svilluppo di nuove tecnologie. Alcuni esempi:

- Pubblicazioni di manuscritti, articoli scientifici
- Sequenze di acidi nucleici (DNA, RNA) e proteine
- Metaboliti risultanti da processi biochimici
- Archivi di esperimenti relativi a geni e/o genomi, RNA-seq, interazioni proteiche
- Dati relativi ad organismi modello o malattie genetiche

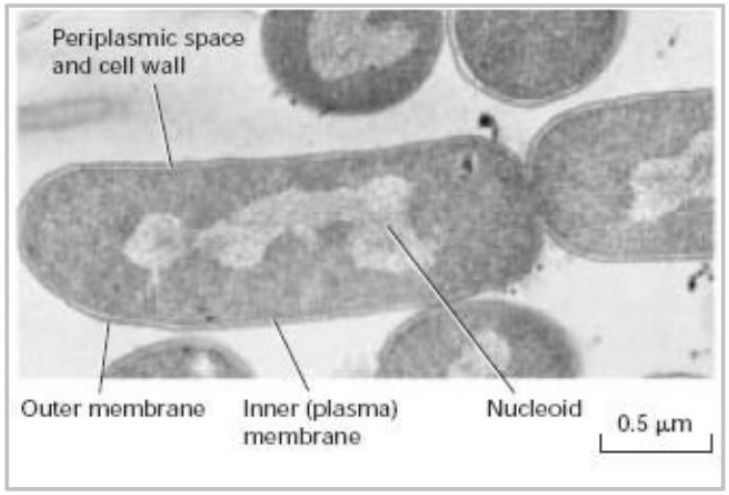
Biologia di base

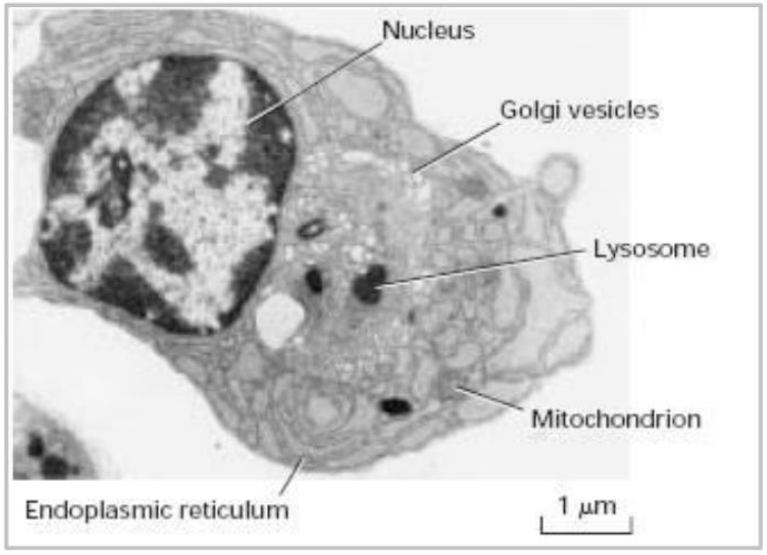
Riprendiamo alcune basi della biologia che sono utili per capire più facilmente dei concetti della Bioinformatica.

- Gli organismi viventi possiedono tutti un **genoma**, nel quale è presente **l'informazione biologica** necessaria alla costruzione e alla sopravvivenza di un individuo (il programma).
- L'informazione biologica contenuta in un genoma è codificata nell'**Acido DeossiriboNucleico (DNA)**, suddivisa in unità discrete chiamate **geni**.
- I geni codificano le **proteine** (e non solo!) di un organismo, comprese quelle che si legano in appropriati punti del genoma, in un processo chiamato **regolazione dell'espressione genica**.

Procarioti

Semplici organismi unicellulari, generalmente privi di membrane endocellulari e organelli. Si riproducono principalemnte attraverso scissione binaria. Il genoma (DNA circolare) si colloca libero nel citoplasma.





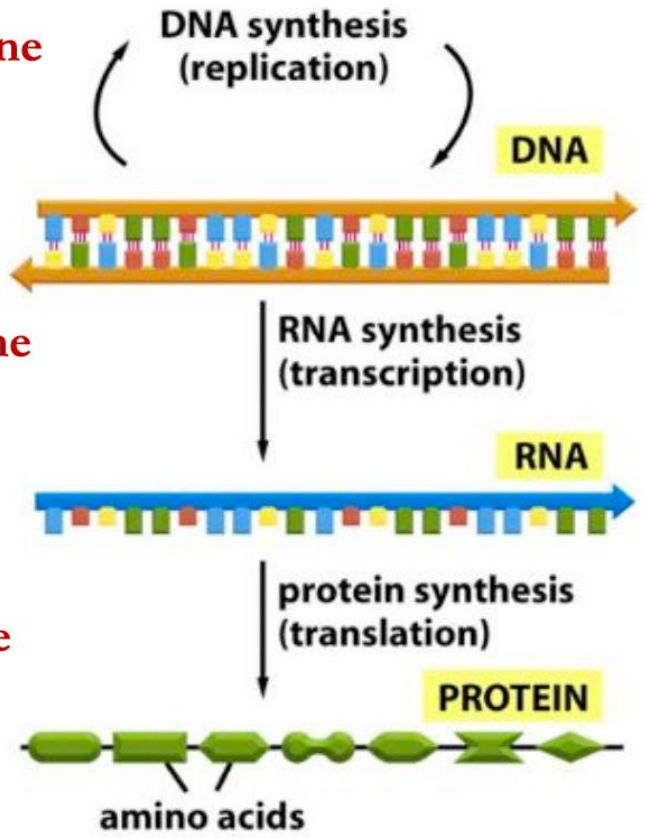
Eucarioti

Organismi che presentano un nucleo, dove è contenuto il DNA, e con dei compartimenti interni racchiusi da membrane, gli organelli, che svolgono particolari compiti biologici.

Alcuni di questi organelli presentano uno specifico genoma.

Dogma centrale della biologia

Replicazione



Trascrizione

Traduzione

DNA: an information archive

RNA: an information carrier
* these are functional molecules sometimes too

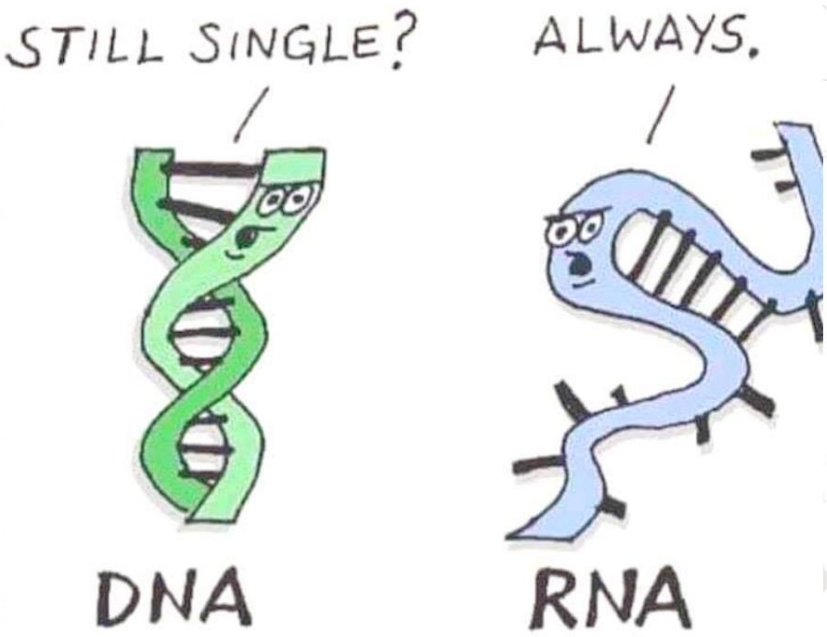
protein: functional molecules

Molecular Biology of the Cell, Fifth Edition (Garland Science 2008)

Acidi nucleici

Gli acidi nucleici **DNA** (acido deossiribonucleico) e **RNA** (acido ribonucleico) sono dei polimeri organici costituiti da monomeri chiamati **nucleotidi**.

1°



Gli acidi nucleici sono formati solo da quattro tipi di **basi azotate**:

- **Adenina**, **Guanina**, **Citosina** (A, G, C) sono comuni al DNA e all'RNA;
- la **Timina** (T) presente solo nel DNA;
- l'**Uracile** (U) solo nell'RNA.

Il DNA esiste prevalentemente in forma di doppia elica antiparallela, in cui due filamenti sono appaiati e avvolti tra loro.

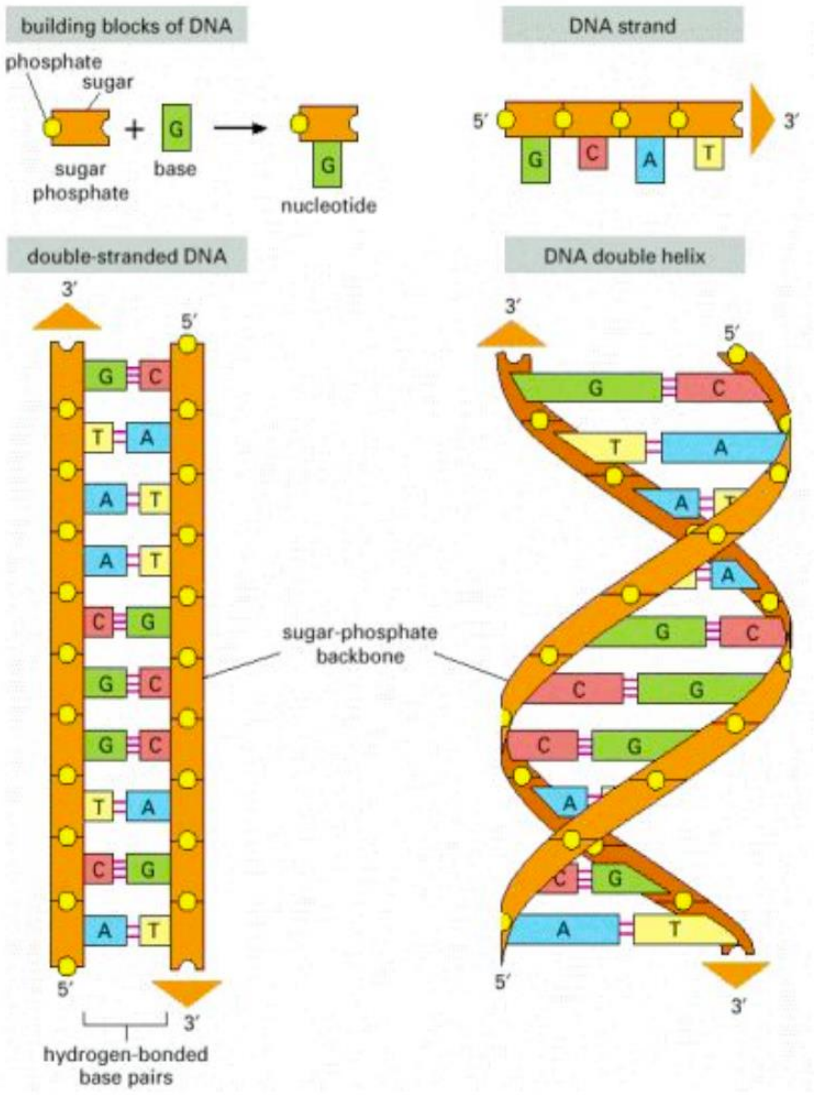
Le basi si appaiano all'interno della doppia elica secondo la regola:

Adenina con **Timina** (A-T)

Guanina con **Citosina** (G-C)

Le basi in grado di appaiarsi vengono dette **complementari**.

Ne consegue che se si conosce la sequenza di un'elica si può ricavare anche la sequenza dell'elica complementare.



La direzionalità

Una rappresentazione in formato testo di una doppia elica di DNA

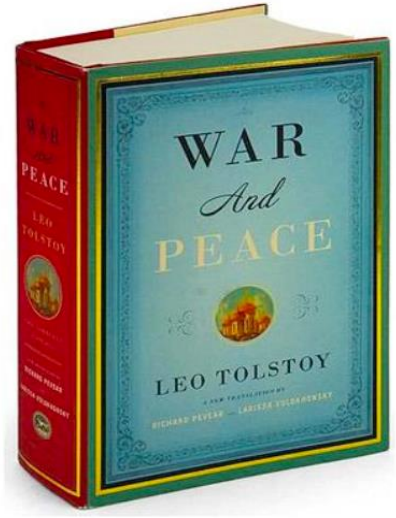
```
5'    nnCGATGCTAGTAGTTGTACGCAnn -> 3'OH
        ||||||||||||||||||||||
HO-3'<- nnGCTACGATCATCAACATGCGTnn - 5'
```

La complementarietà delle basi consente di poter archiviare solo una delle due eliche di una molecola di DNA, che viene in genere rappresentata in direzione 5' → 3'

```
5'-CGATGCCACCAAGTTGTACGCA-> 3'OH
```

Nota: il DNA può essere costituito da catene lunghissime di basi. Il genom umano, nei suoi 23 cromosomi, contiene circa 3.2 miliardi di basi.

La dimensione





Guerra e pace: 3 milioni di lettere

Il genoma umano: 3 miliardi di basi

Il genoma umano ha un numero di lettere pari a mille copie di guerra e pace, che se impilate una sull'altra formerebbero un edificio di 10 piani!!!

Oshlack, A., *Genome Biology 2013, 14:104.*

Geni e struttura genica

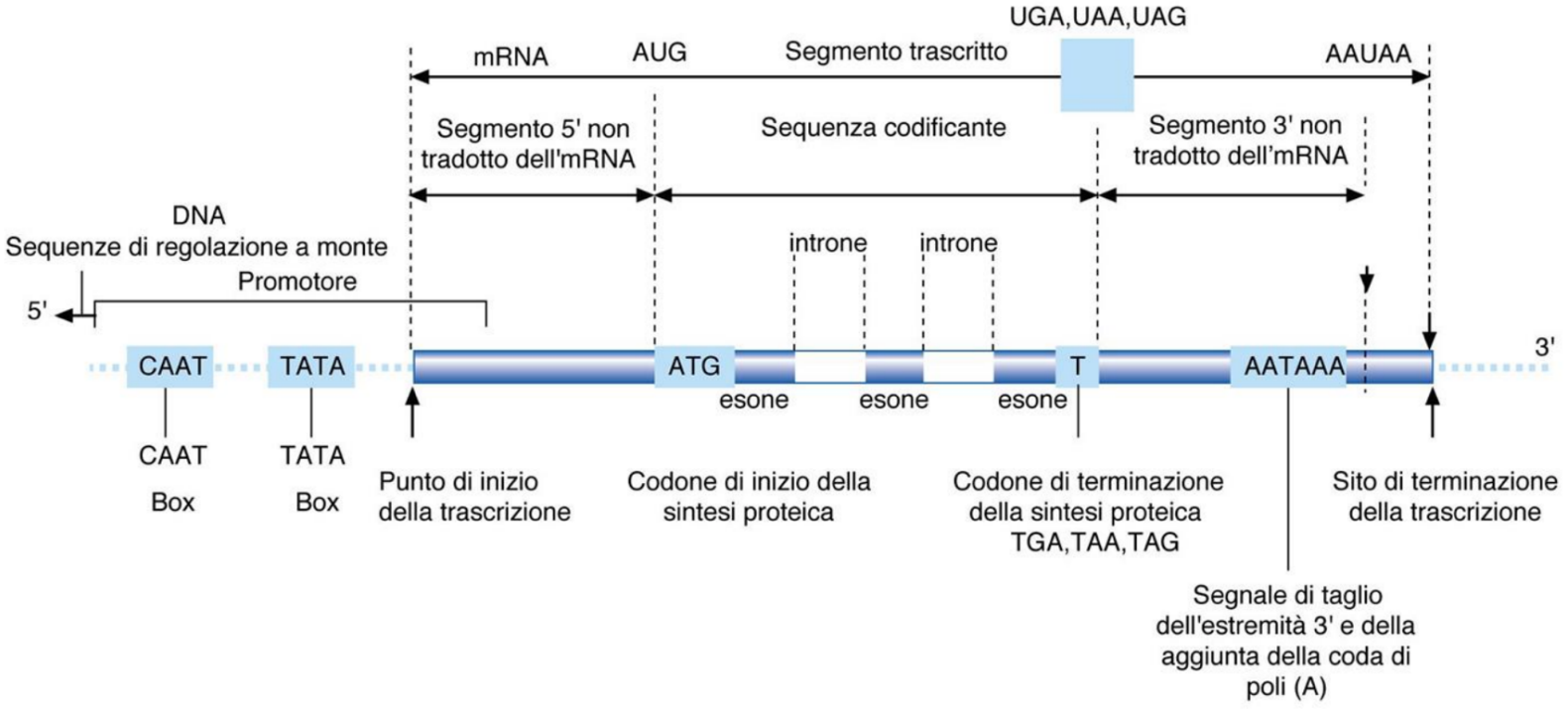
La definizione di **gene** nel corso degli anni è cambiata (e continua a cambiare):

- l'unità fondamentale degli organismi viventi responsabile della trasmissione dei caratteri ereditari
- unità funzionale di DNA che viene trascritto in RNA e tradotto in proteina
- sequenza di nucleotidi che codifica per una proteina delimitata da un codone di inizio ed uno di fine (**Open Reading Frame**) includendo anche sequenze di regolazione

Esistono sostanziali differenze tra i geni dei procarioti e quelli degli eucarioti.

Nei **procarioti** il gene corrisponde **all'intera sequenza nucleica** presente nell'mRNA mentre generalmente negli **eucarioti,** in particolare negli eucarioti superiori, i geni sono **interrotti da introni.**

La struttura dei geni

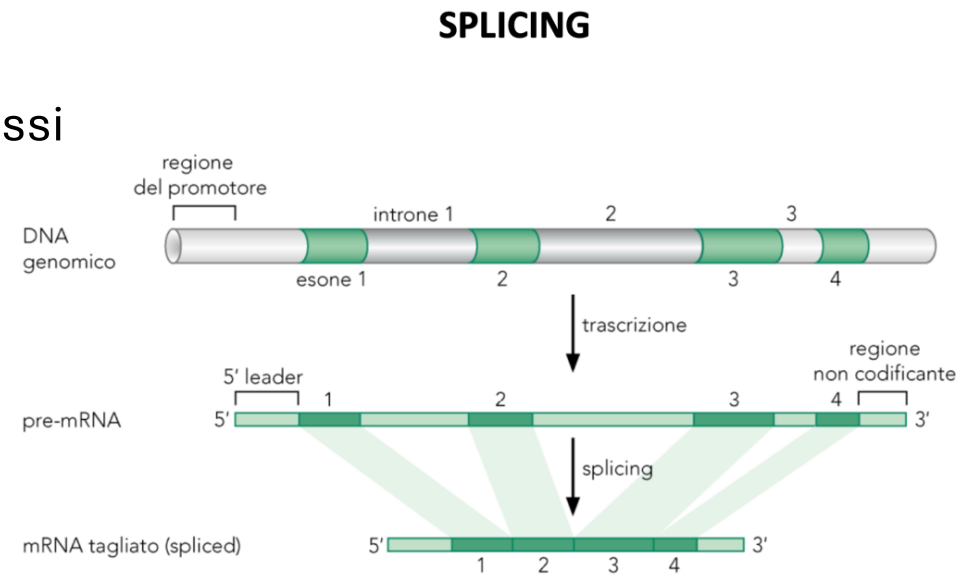


G. De Leo, S. Fasano, E. Ginelli
Biologia e Genetica, IV ed.
EdiSES Università

- porzioni codificanti: *ESONI*
- porzioni non codificanti: *INTRONI*
- sequenze regolatrici: *PROMOTORE* e *TERMINATORE*
- codone di inizio: *ATG*
- codone di STOP: *UAG, UAA, UGA*

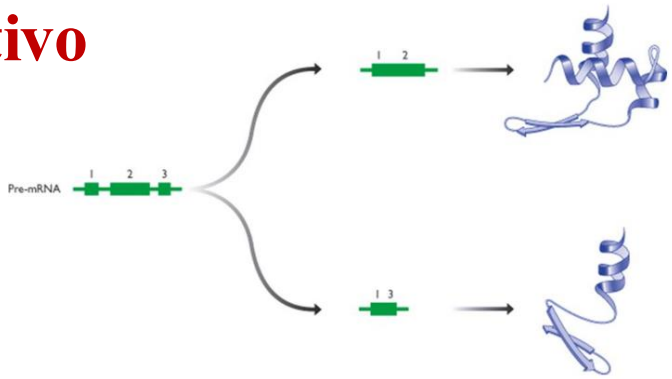
La trascrizione

Gli introni vengono rimossi



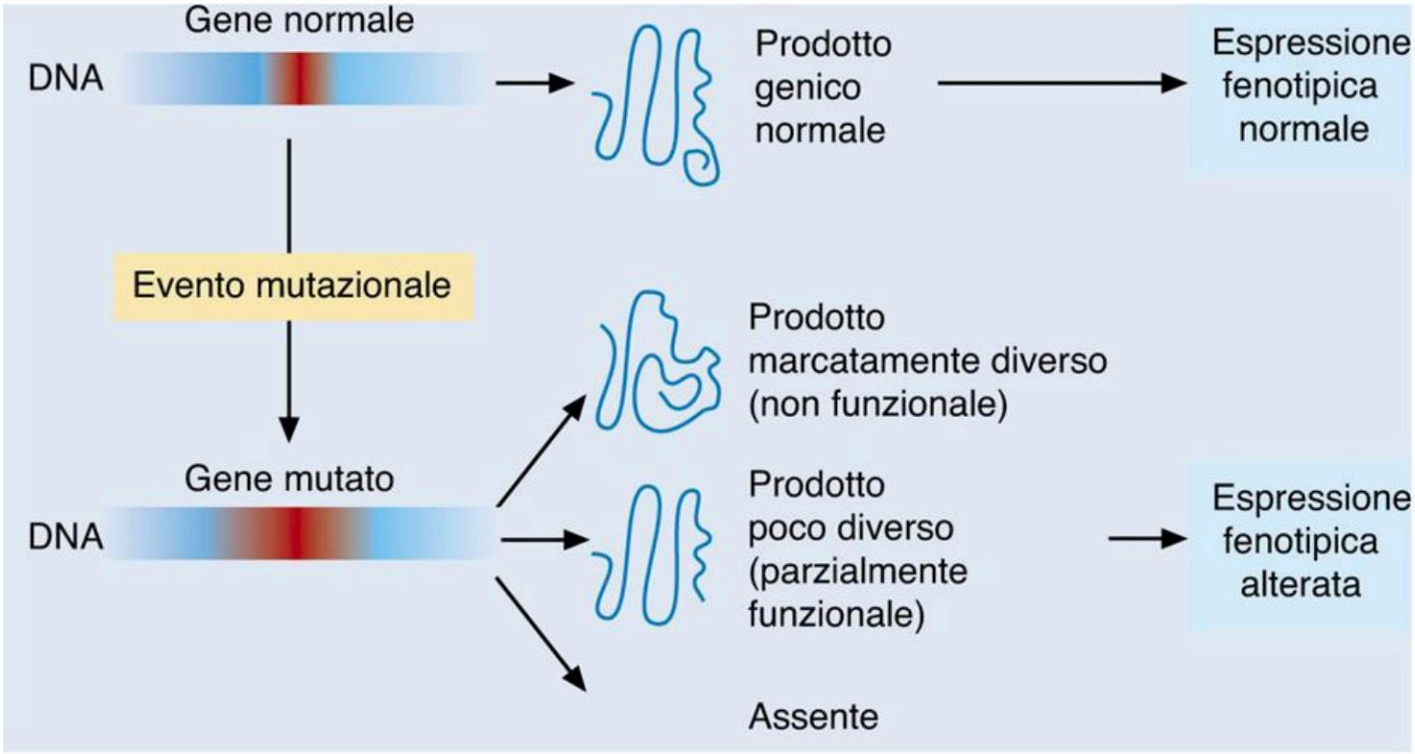
La splicing alternativo

La stessa sequenza di DNA può dare origine a più trascritti alternativi



G. De Leo, S. Fasano, E. Ginelli
Biologia e Genetica, IV ed.
EdiSES Università

La mutazione: modifica della sequenza di DNA



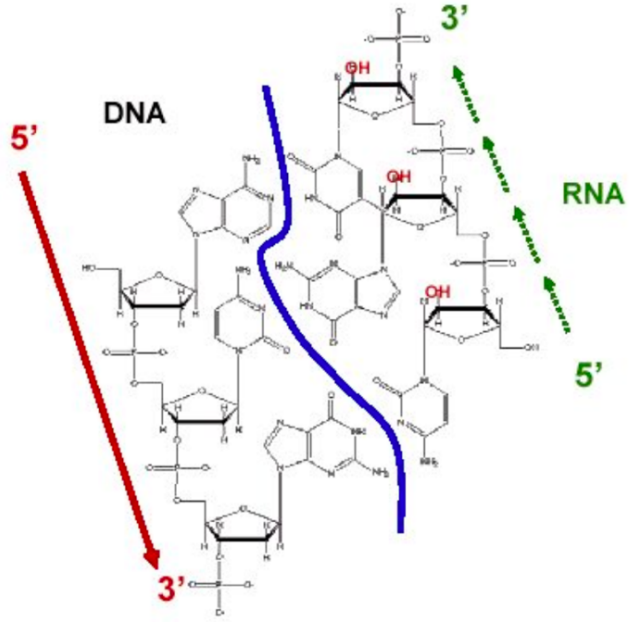
Trascrizione

Processo nel quale **l'RNA** (acido ribonucleico) è sintetizzato a partire dal DNA stampo.
La sintesi dell'RNA avviene in direzione 5' → 3' ad opera delle **RNA polimerasi**, che generano una copia ad RNA in singola elica a partire dal DNA di stampo.

Le molecole prodotte sono di tipologie essere diverse e si distinguono principalmente in:

- RNA messaggero o **mRNA**
- RNA transfer o **tRNA**
- RNA ribosomali o **rRNA**

ognuna delle quali è coinvolta, con ruoli diversi, nella sintesi proteica.



Traduzione

Processo finale in cui **dall'RNA** si arriva alla sintesi delle **proteine**.
L'informazione codificata nell'mRNA, utilizzando solo 4 tipi di nucleotidi, deve poter generare ben 20 differenti aminoacidi. **Come?**

Consideriamo una sequenza lineare di RNA **5'-AUGAUCAGAAUCG -3'** se leggessimo:

- 1 base alla volta (A, U, G, A, U, C,…) esisterebbero solo 4 aminoacidi
- 2 basi alla volta (AU, GA, UC, AG,…): 4^2 combinazioni = 16 aminoacidi, non basta!
- 3 basi alla volta (AUG, AUC, AGA,…): 4^3 combinazioni = 64 aminoacidi, **anche troppi**, ma è proprio così!

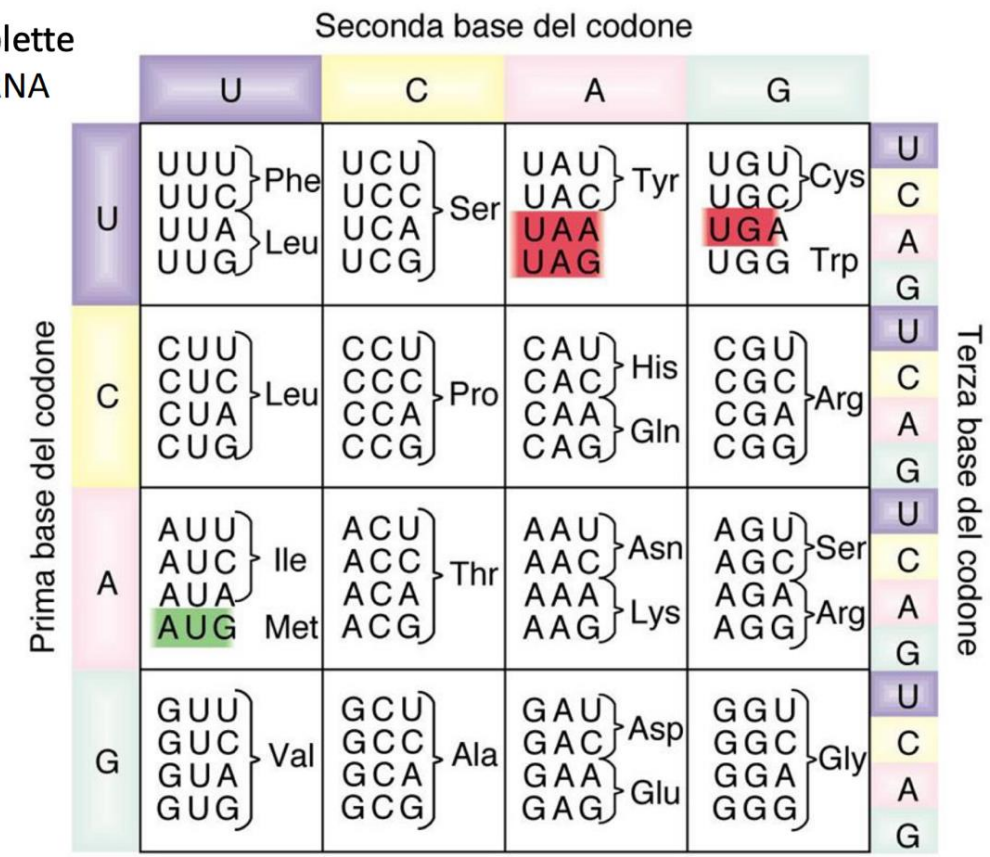
La cellula possiede un sistema di interpretazione di questo **codice genetico**. La regione codificante è letta tre basi alla volta, senza sovrapposizioni:
ogni gruppo di tre basi è chiamato tripletta o più propriamente **codone**.

Tutti gli organismi hanno essenzialmente lo **stesso codice genetico** con qualche piccola eccezione in casi molto particolari (ad esempio i mitocondri), che è perciò definito **universale**.

Il Codice Genetico

Il codice genetico è UNIVERSALE

CODONI
combinazioni a triplette
delle basi dell'mRNA



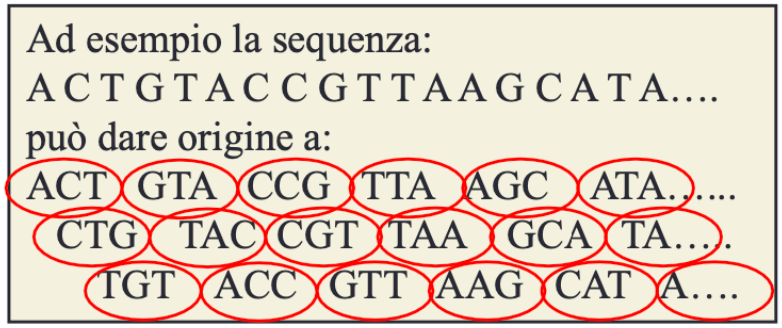
Codone di inizio

Codone di stop

Data la sequenza di un trascritto esistono tre possibili modi di tradurre tale sequenza in proteina, a seconda del punto di inizio:

queste tre possibilità sono chiamate ***reading frame*** (**fasi di lettura**).

Ad esempio la sequenza:
A C T G T A C C G T T A A G C A T A….
può dare origine a:

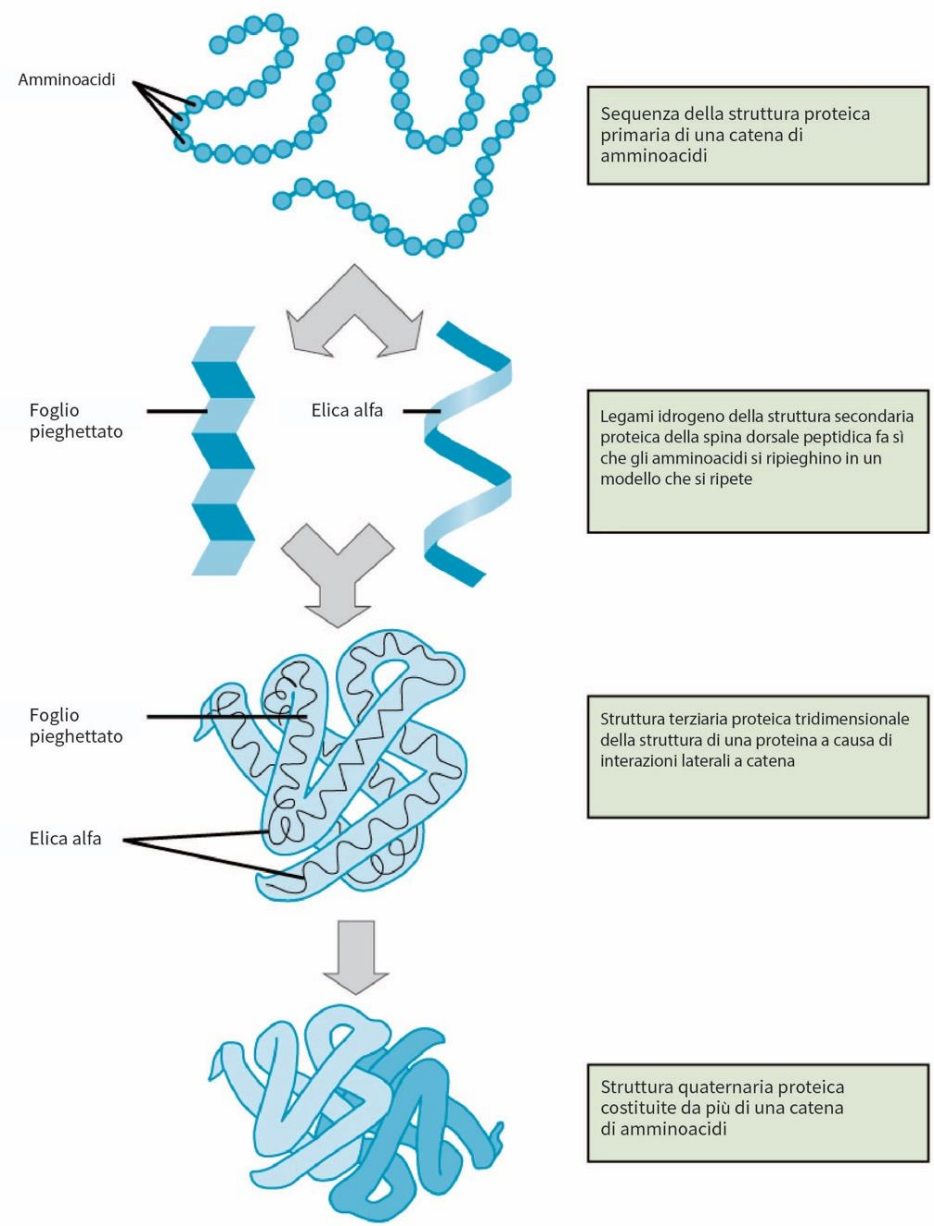


Quando la fase di lettura è costituita esclusivamente da codoni che rappresentano aminoacidi si parla di ***open reading frame (ORF)***.
Una sequenza tradotta in proteina ha una fase di lettura che comincia con un codone di inizio (in genere AUG) e si estende lungo i codoni finché raggiunge uno dei **tre codoni di terminazione.**
In genere, solo una delle tre possibili fasi di lettura è aperta in una data sequenza di DNA.

Una lunga *open reading frame* è difficile che sia presente per caso.

La struttura delle proteine

Conformazione dello spazio. Modo in cui gli aminoacidic si avvicinano creando anse, siti di legame e altro…



Le Discipline Biologiche

La gestione e l'analisi di dati prodotti in modo sistematico nelle numerose e più disparate discipline biologiche, definite come **-OMICs**.

Le principali discipline che si occupano di produrre, gestire ed analizzare i dati biologici sono:

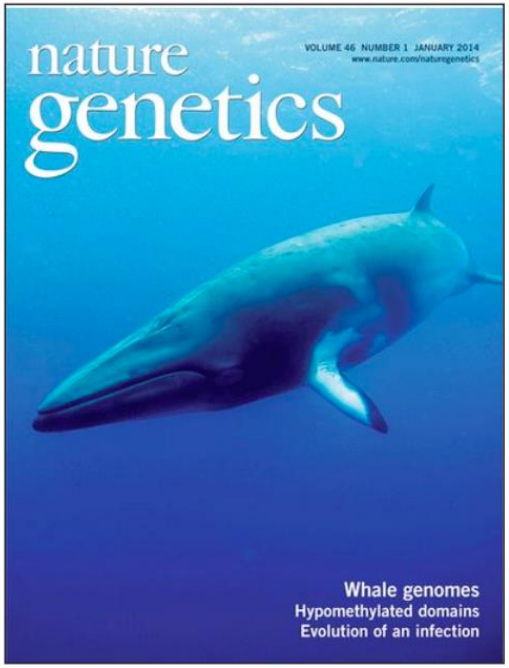
Genomics	⇨	Genome	=	DNA
Metagenomics	⇨	Metagenome	=	DNA
Transcriptomics	⇨	Transcriptome	=	RNA
Proteomics	⇨	Proteome	=	Protein
Metabolomics	⇨	Metabolome	=	Metabolite
Bibliomics	⇨	Bibliome	=	Article

Computational Biology

Memorizzazione accurata, organizzazione, indicizzazione e mantenimento di **informazioni biologiche.** Alcuni esempi di analisi bioinformatiche:

- ricerce bibliografiche
- ricerca di similarità tra sequenze, omologia funzionale
- predizione di geni nelle sequenze di DNA, predizione genica
- predizione di motivi funzionali nel DNA (e.g. siti di binding per fattori di trascrizione), nell'RNA (e.g. strutture secondarie) e nelle proteine (e.g. domini funzionali)
- analisi di comparazione genomica
- analisi filogenetica, tassonomia
- analisi di dati strutturali 3D e predizione della struttura di proteine
- analisi di comunità microbiche

Un esempio...



Articolo *open* (consultabile gratuitamente) intitolato:

"Minke whale genome and aquatic adaptation in cetaceans"

The shift from terrestrial to aquatic life by whales was a substantial evolutionary event. Here we report the whole-genome sequencing and *de novo* assembly of the minke whale genome, as well as the whole-genome sequences of three minke whales, a fin whale, a bottlenose dolphin and a finless porpoise. Our comparative genomic analysis identified an expansion in the whale lineage of gene families associated with stress-responsive proteins and anaerobic metabolism, whereas gene families related to body hair and sensory receptors were contracted. Our analysis also identified whale-specific mutations in genes encoding antioxidants and enzymes controlling blood pressure and salt concentration. Overall the whale-genome sequences exhibited distinct features that are associated with the physiological and morphological changes needed for life in an aquatic environment, marked by resistance to physiological stresses caused by a lack of oxygen, increased amounts of reactive oxygen species and high salt levels.

- sequenziamento ed assemblaggio del genoma di una balena Minke (e cetacei);
- analisi comparativa dei genomi di 8 specie di mammiferi;
- specifici adattamenti associabili ai cambiamenti fisiologici e morfologici necessari per la vita nell'acqua.

Sequenziamento di DNA e genomica comparata

Supplementary Table 2 | Raw read filtering statistics. The raw reads of six cetaceans were generated by Illumina Hiseq2000.

Sample	The number of raw read pairs	The number of proper read pairs	The percentage of proper read pairs	Estimated sequencing depth from raw read pairs	Estimated sequencing depth from proper read pairs
Minke whale01	544,765,050	516,266,141	94.77	39.43	37.37
Minke whale02	512,157,463	496,051,180	96.86	37.07	35.90
Minke whale03	438,408,187	413,853,352	94.40	31.73	29.95
Fin whale	619,941,025	578,742,088	93.35	44.87	41.89
Bottlenose dolphin	598,122,699	580,006,136	96.97	43.29	41.98
Finless porpoise	553,390,610	523,612,238	94.62	40.05	37.90

Supplementary Table 25 | Statistics for the orthologous gene families. Unclustered genes refer to those specific to the current species; Unique families refer to those specific to the current species.

Species	#Total genes	#Genes in familes	#Unclustered genes	#Families	#Unique families	Ave. genes per family
B. acutorostrata	20,605	17,950	2,655	12,675	46	1.42
T. truncatus	16,515	16,276	239	12,678	2	1.28
B. Taurus	19,970	19,795	175	13,288	8	1.49
S. scrofa	21,577	20,029	1,548	13,192	79	1.52
C. familiaris	19,839	19,022	817	13,256	17	1.43
H. sapiens	22,292	21,463	829	13,520	95	1.59
M. domestica	21,299	20,273	1,026	12,813	65	1.58
O. anatinus	21,663	16,121	5,542	13,669	530	1.18

- Come poter gestire e confrontare questa enorme quantità di dati?
- Come poter individuare, a partire da tali dati, quali caratteristiche sono specifiche dei cetacei?

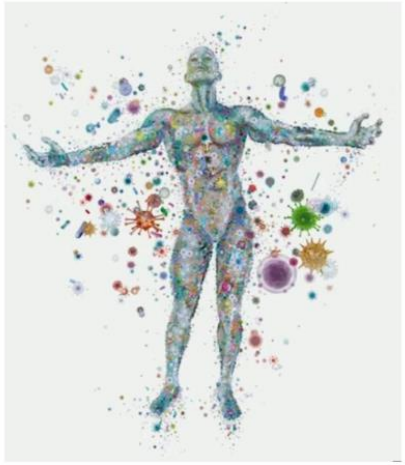
- Ma, sono poi così diversi??



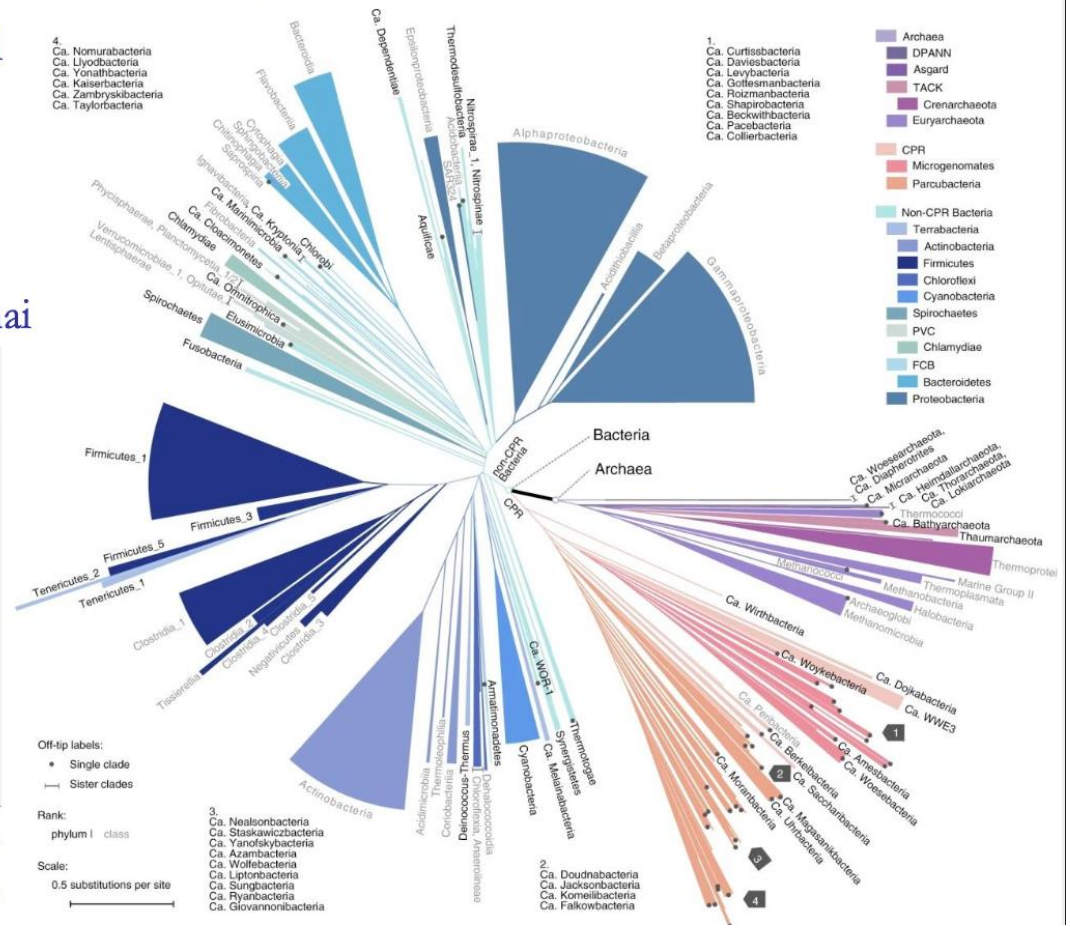
Cetartiodattili

... alcuni numeri della microbiologia...

- Ci sono **100 milioni** di batteri nell'oceano per ogni stella nota nell'universo
- Ci sono **10 cellule** batteriche nel corpo umano per ogni cellula umana
- Un grammo di placca dentale contiene **1 x 10^{11}** batteri: circa lo stesso numero di esseri umani mai vissuti

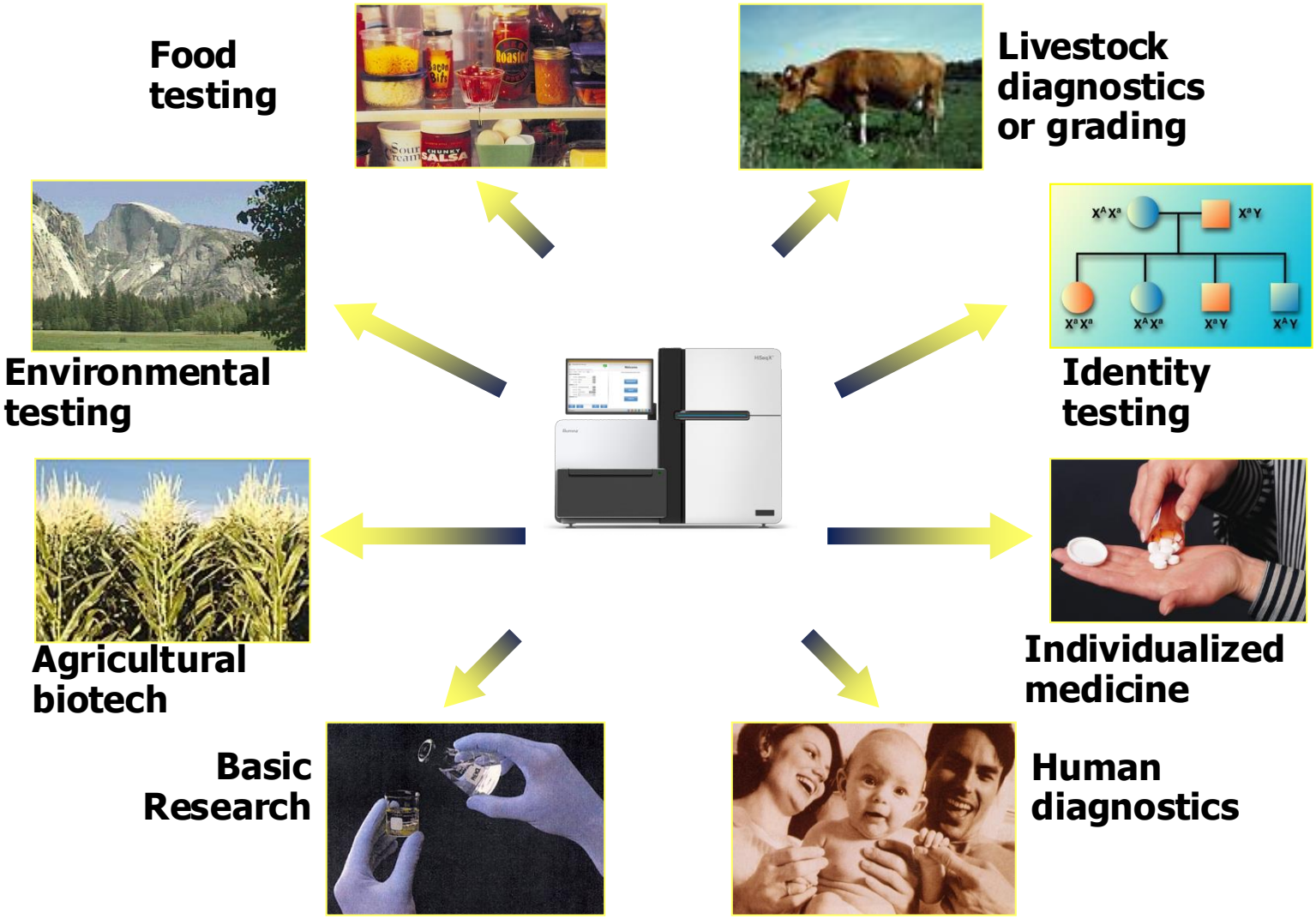


Microbiology by numbers, Nature Reviews Microbiology, Editorial September 2011



Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea, Nature Communications, 2019

Questi dati in che ambiti vengono utilizzati ?



Qual’è stato il punto di svolta ?



The human genome project

Margareth Dayhoff

Robert Ledley