

Strumenti formali per la Bioinformatica

Corso di Laurea Magistrale in Informatica
a.a. 2024-2025

Docenti: Proff. C. De Felice, **R. Zizza**, R. Zaccagnino

Lezione 1
(23.09.2024)

1

Strumenti formali per la Bioinformatica

Corso di Laurea Magistrale in Informatica
a.a. 2024-2025

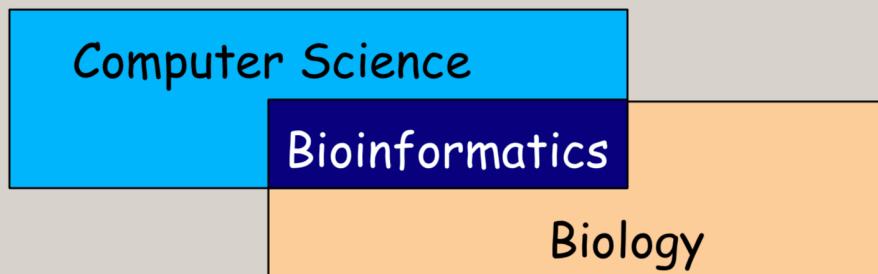
Docenti: Proff. C. De Felice, **R. Zizza**, R. Zaccagnino

Lezione 0: Presentazione del corso
(23.09.2024)

2

Strumenti formali per la Bioinformatica

What is bioinformatics? Application of methods from computer science to biology.



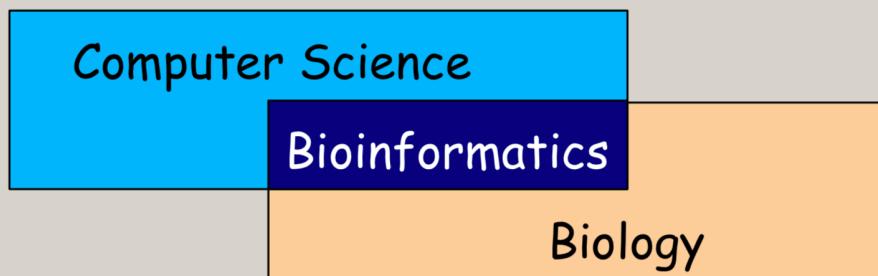
Why is it interesting?

- Important problems.
- Massive quantities of data.
- Great need for efficient solutions.

3

Strumenti formali per la Bioinformatica

What is bioinformatics? Application of methods from computer science to biology.



Why is it interesting?

- Important problems.
- Massive quantities of data.
- Great need for efficient solutions.

4

Strumenti formali per la Bioinformatica

Quali metodi?

Quali strumenti formali?

- Linguaggi e combinatoria delle parole (MMI, ETC, ...)
- Algoritmi e strutture dati

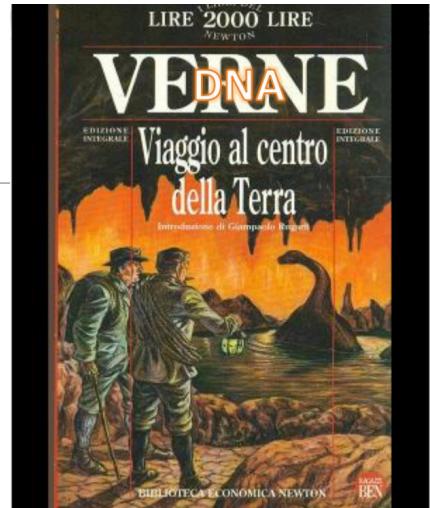
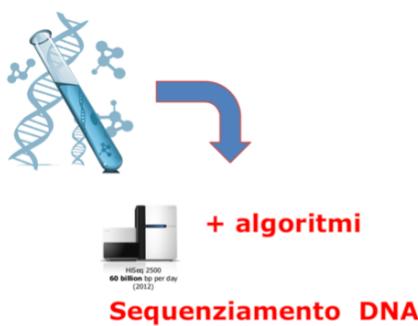
DNA USES A CODING LANGUAGE SIMILAR TO COMPUTER PROGRAMMING



DNA is similar to computer programming code. Computer coding is binary which means it is written with 2 numbers: 0 and 1. DNA uses a quaternary code consisting of 4 letters: A, C, T, G. From this code 20 different amino acid letters are used to write protein chains into living flesh. It is understood that any coding is derived only by an intelligent source, through an intelligent act, as to accomplish a designed purpose. It is absurd to reason that such code could arise by purely naturalistic processes.

5

Viaggio al centro del DNA



Viaggio al centro del DNA



Sequenziamento DNA

+ algoritmi



Terapia medica

7

Viaggio al centro del DNA



Sequenziamento DNA

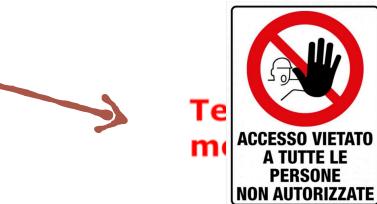
+ algoritmi



Terapia medica

8

Viaggio al centro del DNA



9

Viaggio al centro del DNA



Terapia medica

sono stringhe...

Viaggio al centro del DNA

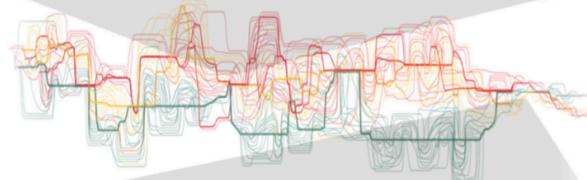
genoma è una sequenza?

Aligned sequences
Human A C A T T A G G G C A G G T G A T A A A A A A C A T A T
Chimpanzee A C A T T A G G G C A G G T G A T A A A A A A C A T A T
Macaque A T T A C A T T G C A C A G G A A G T A A A A A C A T A T

Aligned sequences
Human A C A T T A G G G C A G G T G A T A A A A A C A T A T
Chimpanzee A T T A C A T T G C A C A G G A A G T A A A A A C A T A T
Macaque A T T A C A T T G C A C A G G A A G T A A A A A C A T A T



grafo - Pan-genome



sono grafi

+ algoritmi ...

Terapia medica



11

Organizzazione del corso

Strumenti formali
per la
Bioinformatica

Orario del corso

- Lunedì 13:30-15:30 (aula P6)
- Mercoledì 13:30-15:30 (aula P6)
- Giovedì 13:30-15:30 (aula P6)

Organizzazione del corso

Orario del corso

Argomenti

13

Formazione attesa sugli “strumenti”...

1. Nella prima parte: stringhe e tecniche combinatoriche
2. Nella seconda parte, le usiamo, insieme alle strutture dati, per due grosse problematiche “in pratica”:
 1. confronto tra sequenze - allineamento
 2. assemblaggio
3. Nella terza parte, si cambia approccio... ma anche no!

Materiale didattico

Testi di riferimento [dalla scheda dell'insegnamento]

1. M. Sipser, Introduzione alla Teoria della Computazione, Apogeo, 2016
2. J. Hopcroft, R. Motwani, J. Ullman, Automi, Linguaggi e Calcolabilità, Addison Wesley Pearson Education Italia s.r.l, 3ed., 2009.
3. M. Lothaire, Applied Combinatorics on words,
4. M. Lothaire, Algebraic Combinatorics on words
5. N.C. Jones, P.A. Pevzner. An Introduction to Bioinformatics Algorithms. <http://bioinformaticsalgorithms.com>
6. D. Gusfield. Algorithm on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge Press.
7. Introduction to Machine Learning, *Lecture notes*, MIT, 2019.
8. Ian Goodfellow, Yoshua Bengio, Aaron Courville, Deep Learning, MIT Press, 2016.
9. Z.R. Yang, Machine learning approaches to bioinformatics, Science, Engineering, and Biology Informatics.
10. Habib Izadkhah, Deep Learning in Bioinformatics Techniques and Applications in Practice, 1st Edition, Elsevier.
11. *Selezione di articoli scientifici*

Saranno esplicitamente forniti i
capitoli dei libri di riferimento
a fine lezione

15

Strumenti formali
per la
Bioinformatica

Organizzazione del corso

Orario del corso

Argomenti

“Spirito del corso”

proporre tematiche attuali di
ricerca e tecniche innovative tra
le quali lo studente selezionerà
quelle più interessante da
approfondire

Organizzazione del corso

Orario del corso	Argomenti
“Spirito del corso”	<p>Esame finale</p> <p>sviluppo di un progetto (teorico o pratico), individuale o in coppia.</p> <p>Seminario di presentazione alla classe.</p>

Quando?

17

Organizzazione del corso

Orario del corso	Argomenti
“Spirito del corso”	<p>Esame finale</p> <p>sviluppo di un progetto (teorico o pratico), individuale o in coppia.</p> <p>Seminario di presentazione alla classe.</p>

<https://github.com/strumenti-formali-per-la-bioinformatica>

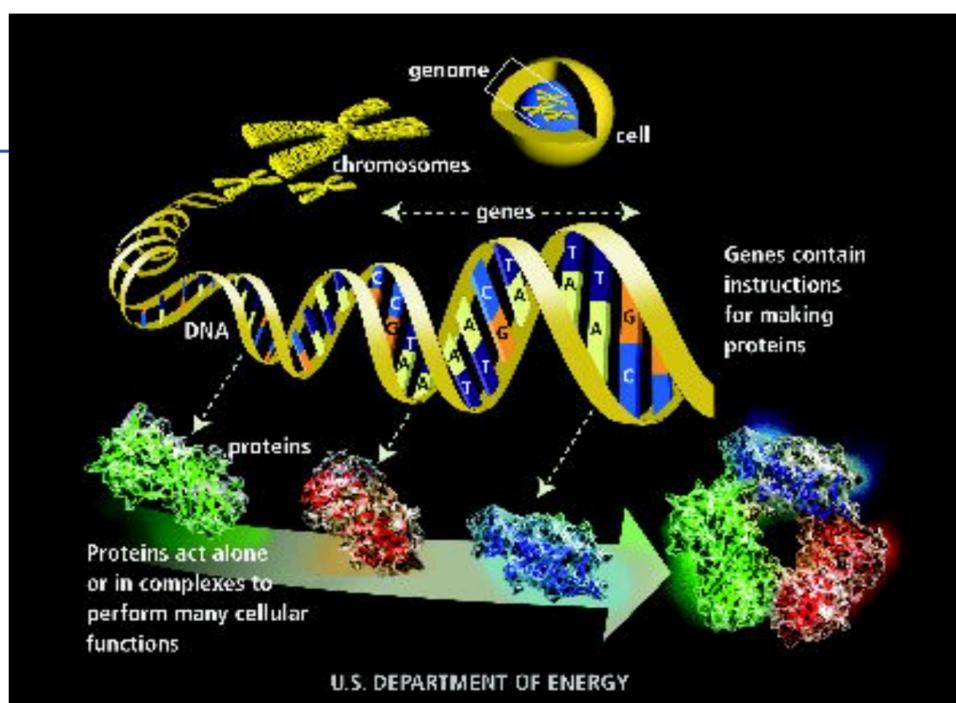
18

<https://github.com/strumenti-formali-per-la-bioinformatica>

- ...
- Studio di varianti Covid/Ebola
- Ricostruzione di alberi filogenetici
- Classificazione i specie virali
- Analisi/sperimentazione di tool per l'assemblaggio di reads
- Analisi di funzioni hash per strutture dati
- Predizione di proteine
- Classificazioni di tumori
- Rilevazione di virus da dati metagenomici
- Progettazione di molecole con proprietà farmacologiche

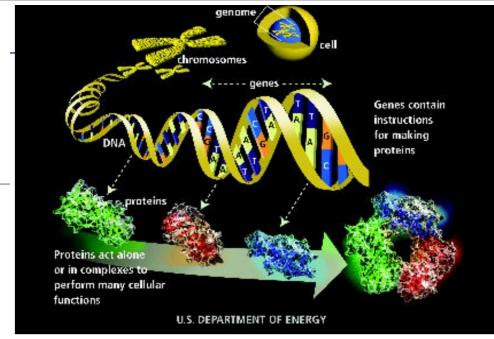
19

Un po' di nomenclatura è necessaria...
saranno i nostri dati “reali” (ovviamente “rappresentati”)



Human Genome Program, U.S. Department of Energy, Genomics and Its Impact on Medicine and Society: A 2001 Primer, 2001

Poche competenze...



Human Genome Program, U.S. Department of Energy, Genomics and Its Impact on Medicine and Society: A 2001 Primer, 2001

- **Genoma:** collezione di cromosomi presenti nell'individuo
- **Cromosomi:** all'interno della cellula (nel nucleo se eucarioti) [noi siamo organismi diploidi: 46 cromosomi, ogni cromosoma in coppia (23 padre, 23 madre), quindi 2 chr1, 2chr2, ...]
 - Ognuno è rappresentato da filamenti di DNA “attorcigliati”. In particolare, due si uniscono nella parte centrale (centromero).
 - DNA: contiene il nostro patrimonio genetico...
 - Sui cromosomi sono distribuiti i **geni, che producono le proteine.**

21

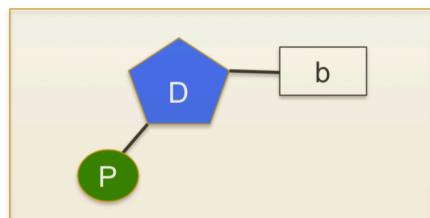
Esempio (Treccani)

Ogni molecola di **emoglobina**, la proteina che dà il colore rosso al sangue e che trasporta l'ossigeno dai polmoni ai tessuti, è un'unità funzionale fatta di quattro sottounità, chiamate *globine*, due di tipo α (*alfa*) e due di tipo β (*beta*). La formazione di ognuna di queste globine è diretta **da uno specifico gene**. Nell'uomo il gene per la globina α risiede nel cromosoma 16 e quello per la globina β nel cromosoma 11.

Il gene è una sequenza di nucleotidi capace di produrre la catena di amminoacidi che costituisce una proteina. Nel caso dell'emoglobina, **due geni diversi**, localizzati in **cromosomi diversi**, ognuno contenente il codice per una catena amminoacidica diversa, **cooperano** a permettere la precisa funzione di legare l'ossigeno contenuto nell'aria che respiriamo per fornirlo alle cellule.

22

DNA (acido deossiribonucleico) è una molecola composta da nucleotidi

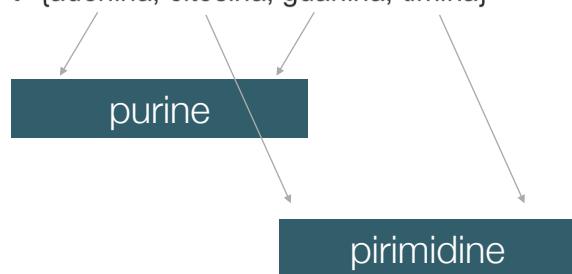


mattoncini fondamentali della nostra vita
che si differenziano per la
base azotata (A,C,G,T)

D: zucchero → deossiribosio

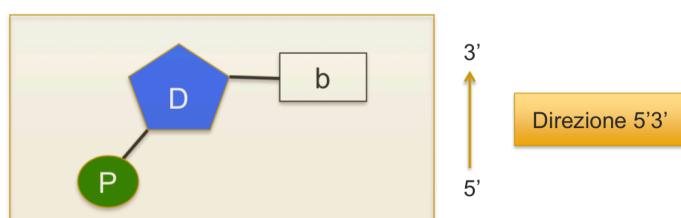
P: gruppo fosfato

b: base azotata → {adenina, citosina, guanina, timina}



23

DNA (acido deossiribonucleico) è una molecola composta da nucleotidi



D: zucchero → deossiribosio (5 atomi di carbonio)

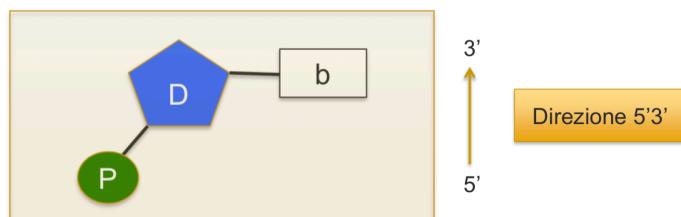
P: gruppo fosfato

b: base azotata → {adenina, citosina, guanina, timina}

DIREZIONE: in base ai legami biochimici che si instaurano nel nucleotide (che li fa legare l'uno all'altro): dal gruppo fosfato allo zucchero

24

DNA (acido deossiribonucleico) è una molecola composta da nucleotidi



D: zucchero → deossiribosio

P: gruppo fosfato

b: base azotata → {adenina, citosina, guanina, timina}

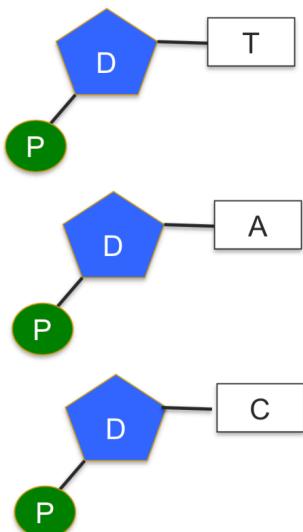
DIREZIONE: in base ai legami biochimici che si instaurano nel nucleotide (che li fa legare l'uno all'altro): dal gruppo fosfato allo zucchero

chi compone la catena è il legame tra D e P (legame fosfodiesterico)

leggo la sequenza primaria (le basi) dal 5' al 3'

25

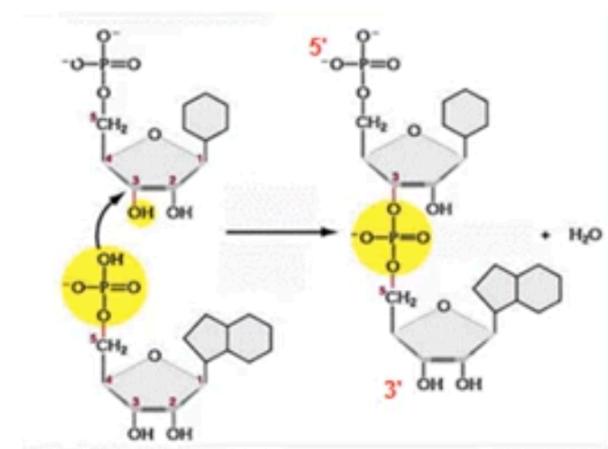
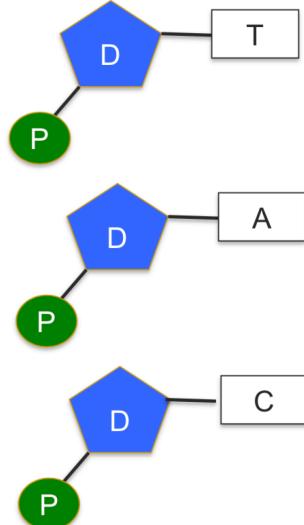
DNA (acido deossiribonucleico) è una molecola composta da nucleotidi - **DIREZIONE**



chi compone la catena è il legame tra D e P (legame fosfodiesterico)

26

DNA (acido deossiribonucleico) è una molecola composta da nucleotidi - **DIREZIONE**

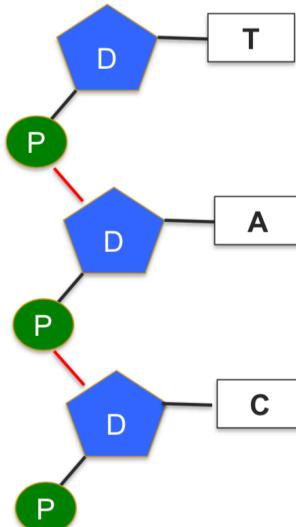
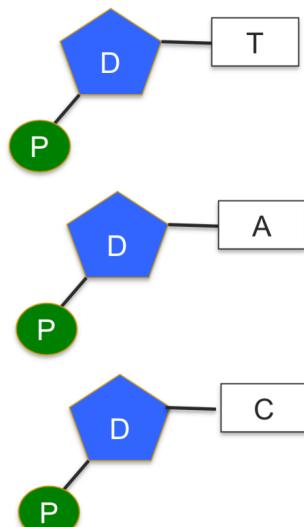


chi compone la catena è il legame tra D e P (legame fosfodiesterico)

Processo di condensazione:
il fosfato in 5' con l'OH (ossidrile) in 3'

27

DNA (acido deossiribonucleico) è una molecola composta da nucleotidi - **DIREZIONE**



3'
5'

L'unità di misura della lunghezza di una molecola di DNA è il **base pair (bp)**

Direzione 5'3'

CAT

sequenza primaria
(solo le basi, senza considerare la struttura)

C'è sempre una estremità 5' libera (primo fosfato della fila che non ha potuto reagire con un OH) e una estremità 3' libera (l'ultimo OH che non è stato attaccato ad un gruppo fosfato)

28

DNA (acido deossiribonucleico) è una molecola composta da nucleotidi

Perchè bp?

Perchè la molecola di DNA fondamentale di un individuo è il GENOMA, che è una **doppia** catena che si appaia

29

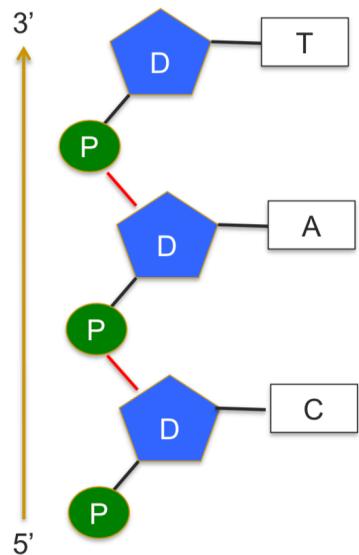
GENOMA

Il genoma è la lunga molecola di DNA che regola la funzione e lo sviluppo di un organismo vivente ed è:

- ✓ situata nel nucleo cellulare
- ✓ avvolta a doppia elica (Watson-Crick, 1953)
- ✓ frammentata in cromosomi

30

GENOMA

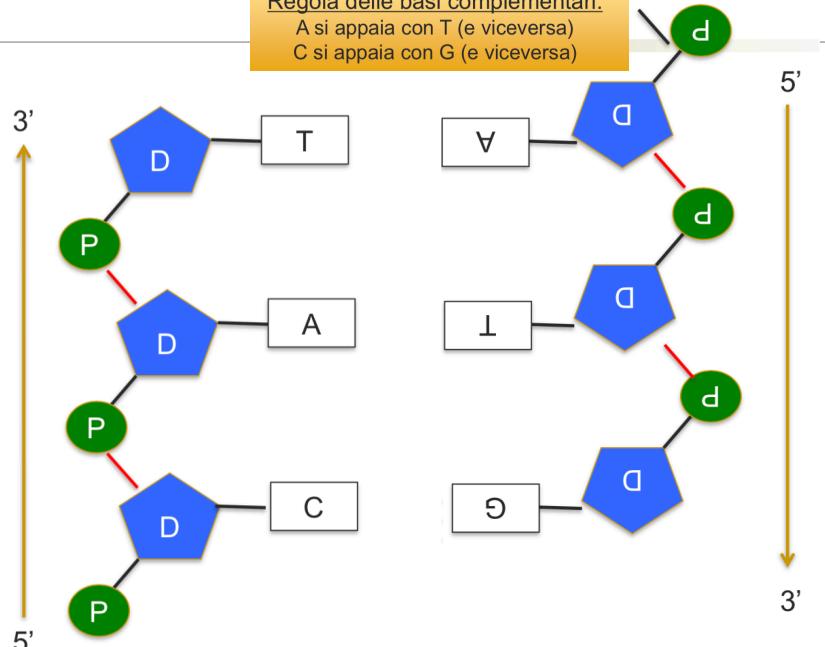


La catena che si appaia
è in senso opposto...

31

GENOMA

Regola delle basi complementari:
A si appaia con T (e viceversa)
C si appaia con G (e viceversa)



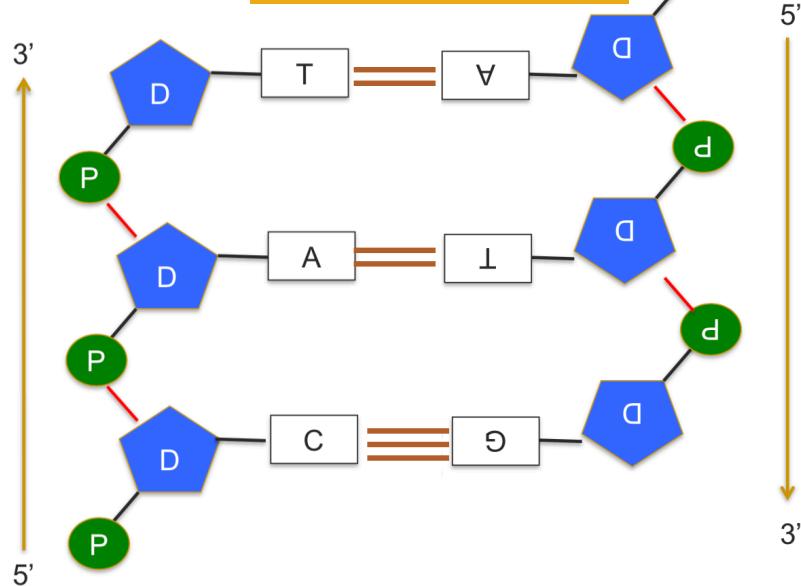
CAT
sequenza primaria

ATG
sequenza primaria

32

GENOMA

Regola delle basi complementari:
A si appaia con T (e viceversa)
C si appaia con G (e viceversa)



A,T
generano 2 legami di
idrogeno e si
appaiano

C,G
generano 3 legami di
idrogeno e si
appaiano

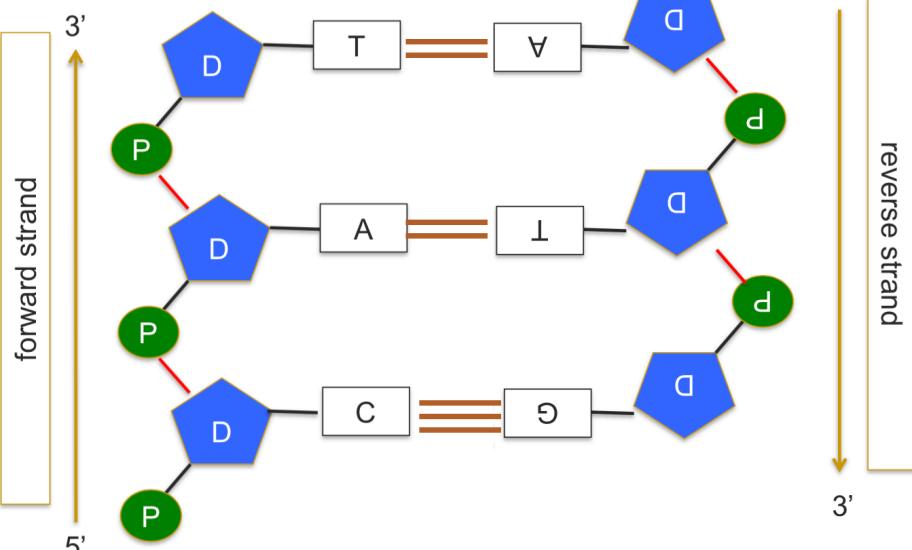
CAT
sequenza primaria

ATG
sequenza primaria

33

GENOMA

Regola delle basi complementari:
A si appaia con T (e viceversa)
C si appaia con G (e viceversa)



CAT
sequenza primaria

ATG
sequenza primaria

quando accediamo ad
una banca dati per
ottenere il genoma di un
organismo,
chiede se vuoi la catena
+ (forward, diretta) o
- (reverse, inversa)

34

DNA

La sequenza primaria di una catena di DNA è una stringa di simboli sull'alfabeto $\Sigma = \{A,C,G,T\}$
(o $\Sigma = \{a,c,g,t\}$)

Data la sequenza primaria di una delle due catene del DNA genomico, la sequenza primaria della catena appaiata è ottenuta per mezzo dell'operazione di **"reverse and complement"**:

ACGTAGGATGGACGATGACGATGACGAT

35

DNA

La sequenza primaria di una catena di DNA è una stringa di simboli sull'alfabeto $\Sigma = \{A,C,G,T\}$
(o $\Sigma = \{a,c,g,t\}$)

Data la sequenza primaria di una delle due catene del DNA genomico, la sequenza primaria della catena appaiata è ottenuta per mezzo dell'operazione di **"reverse and complement"**:

ACGTAGGATGGACGATGACGATGACGAT



TAGCAGTAGCAGTAGCAGGTAGGATGCA

la leggiamo al contrario

36

DNA

La sequenza primaria di una catena di DNA è una stringa di simboli sull'alfabeto $\Sigma = \{A,C,G,T\}$ (o $\Sigma = \{a,c,g,t\}$)

Data la sequenza primaria di una delle due catene del DNA genomico, la sequenza primaria della catena appaiata è ottenuta per mezzo dell'operazione di **"reverse and complement"**:

ACGTAGGATGGACGATGACGATGACGAT

TAGCAGTAGCAGTAGCAGGTAGGATGCA

“complementiamo”

37

DNA

La sequenza primaria di una catena di DNA è una stringa di simboli sull'alfabeto $\Sigma = \{A,C,G,T\}$ (o $\Sigma = \{a,c,g,t\}$)

Data la sequenza primaria di una delle due catene del DNA genomico, la sequenza primaria della catena appaiata è ottenuta per mezzo dell'operazione di **"reverse and complement"**:

ACGTAGGATGGACGATGACGATGACGAT

ATCGTCATCGTCATCGTCCATCCTACGT

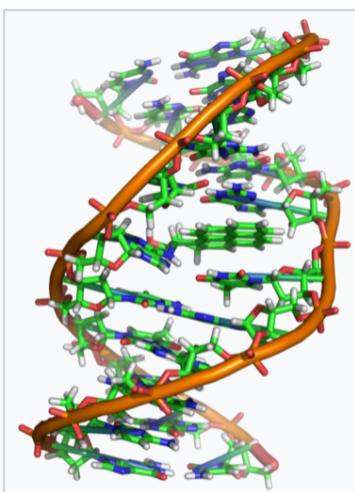
TAGCAGTAGCAGTAGCAGGTAGGATGCA

paired strands

ovviamente
letta da sinistra a destra...

38

Non tutto è perfetto: DNA e mutazioni



Benzopirene: agente presente nel tabacco, viene a formarsi anche durante la carbonizzazione dei cibi nelle cotture alla griglia,

Il tipo di danno causato al DNA dipende dal tipo di agente: gli UV, i raggi X...

inibiscono sia la trascrizione che la replicazione del DNA e aumentano la possibilità di insorgenza di mutazioni.

Esempio: gli intercalanti (molecole che si inseriscono nella doppia elica) sono considerati molecole cancerogene, come dimostrato da numerosi studi su molecole come il benzopirene.

In ogni caso, proprio grazie alla loro capacità di inibire trascrizione e replicazione, tali molecole sono anche utilizzate in chemioterapia per inibire la rapida crescita delle cellule neoplastiche.

39

GENOMA

Il genoma è la lunga molecola di DNA che regola la funzione e lo sviluppo di un organismo vivente ed è:

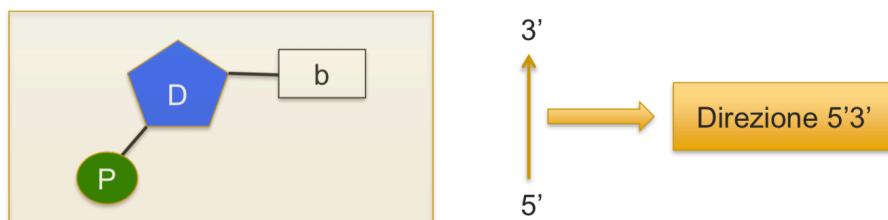
- ✓ situata nel nucleo cellulare
- ✓ avvolta a doppia elica (Watson-Crick, 1953)
- ✓ frammentata in cromosomi

Una regione di DNA genomico che ha una certa funzione prende il nome di locus

La sequenza primaria di un *locus* è chiamata sequenza genomica

40

RNA (acido ribonucleico) è una molecola composta da nucleotidi



D: zucchero → ribosio

P: gruppo fosfato

b: base azotata {adenina, citosina, guanina, **uracile**}

La sequenza primaria di una molecole di RNA è una stringa di simboli sull'alfabeto $\Sigma = \{A,C,G,U\}$ (o $\Sigma = \{a,c,g,u\}$)

L'RNA si trova in catene singole

41

PROTEINE (responsabili della struttura e delle attività di un organismo)

Una proteina è una catena di amminoacidi e la sua sequenza primaria è una stringa su alfabeto di 20 simboli (che rappresentano i 20 amminoacidi presenti in natura).



ogni proteina è costruita a partire da un **gene**

42

GENOMA E GENE

- Il **genoma governa** la vita di un organismo attraverso la *produzione* di proteine
- Un gene è *distribuito* sul genoma (quindi è frammentato) ed è una regione (locus) del genoma che esprime una proteina.
- Ogni proteina è costruita a partire da un gene (non vale il viceversa!!!)

esistono anche geni che “non riescono” ad esprimere una proteina (non-coding): si interrompe la trascrizione...

43

GENOMA E GENE

- Il **genoma governa** la vita di un organismo attraverso la *produzione* di proteine
- Un gene è distribuito sul genoma (quindi è frammentato) ed è una regione (locus) del genoma che esprime una proteina.
- Ogni proteina è costruita a partire da un gene (non vale il viceversa!!!)

Noi siamo un complesso di GENI
Noi siamo un complesso di PROTEINE

44

GENOMA E GENE

- Il genoma governa la vita di un organismo attraverso la produzione di proteine
- Un gene è distribuito sul genoma (quindi è frammentato) ed è una regione (*locus*) del genoma che esprime una proteina.
- Ogni proteina è costruita a partire da un gene (non vale il viceversa!!!)
- Un gene viene identificato tramite HUGO NAME (Human Genome Organization) univoco per ogni banca dati, ad es., ATP6AP1...

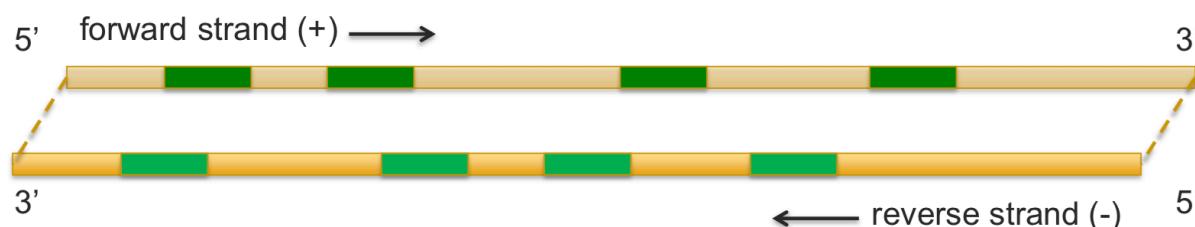
45

GENOMA E GENE

Un gene è una regione (*locus*) del genoma che esprime una proteina

Entrambe le catene del genoma (forward e reverse) contengono geni

Il genoma umano contiene circa 20000 geni codificanti proteine



i geni verde scuro “trascrivono” da sinistra a destra,
gli altri da destra a sinistra

46

ESPRESSIONE (STRUTTURA) DEL GENE

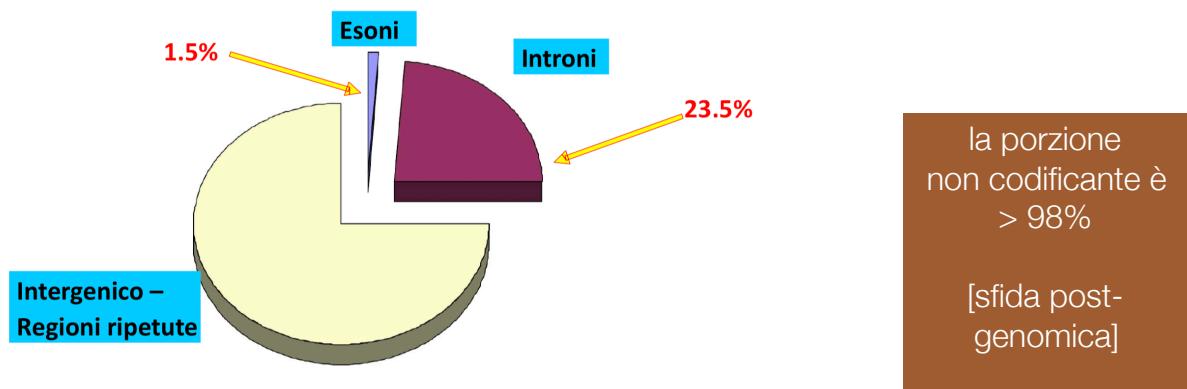


Eson = regione codificante

Intron = regione non codificante

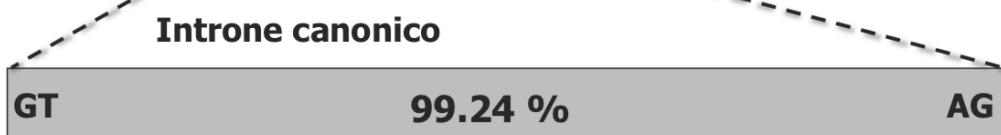
Quanti esoni?
Quanti introni?

Il Contenuto del genoma umano



47

ESPRESSIONE (STRUTTURA) DEL GENE

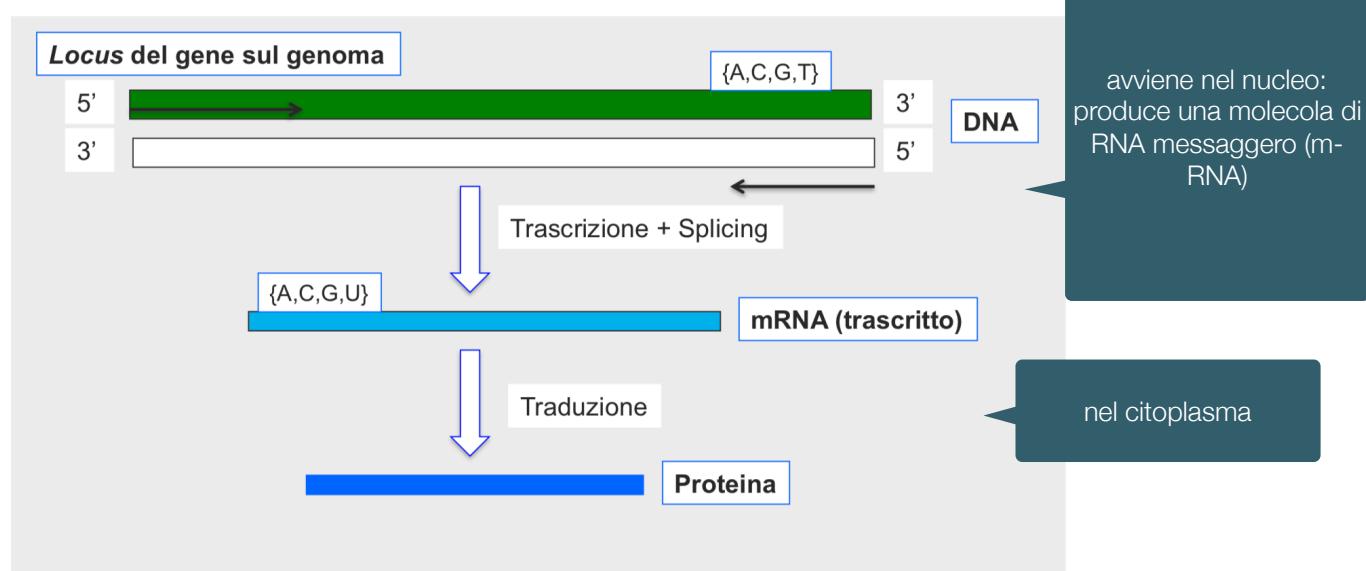


Introni non-canonicali



48

Dal GENE alla PROTEINA

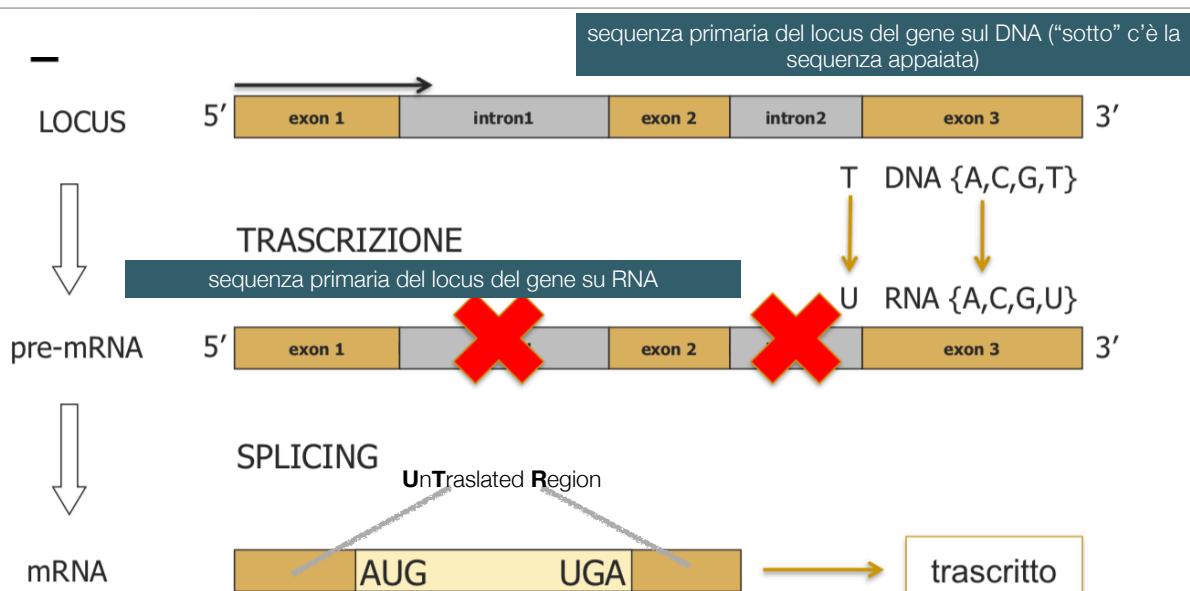


...ma se tutte le cellule hanno nel nucleo lo stesso genoma, perchè le cellule del fegato sono diverse da quelle del polmone?

Perchè cambia il profilo di espressione [non tutti i geni sono espressi in tutte le cellule, e per lo splicing alternativo vengono prodotte proteine diverse]

49

ESPRESSIONE (STRUTTURA) DEL GENE



La lunghezza della CDS è un multiplo di 3

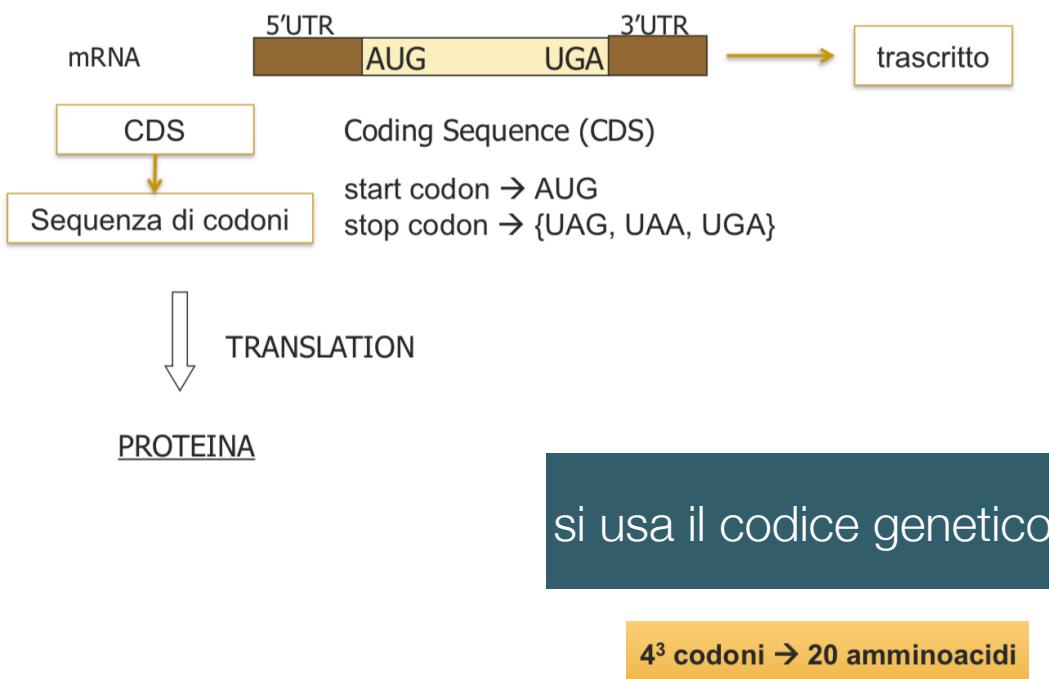
Coding Sequence (CDS)

tripletta di inizio → AUG

tripletta di fine → {UAG, UAA, UGA}

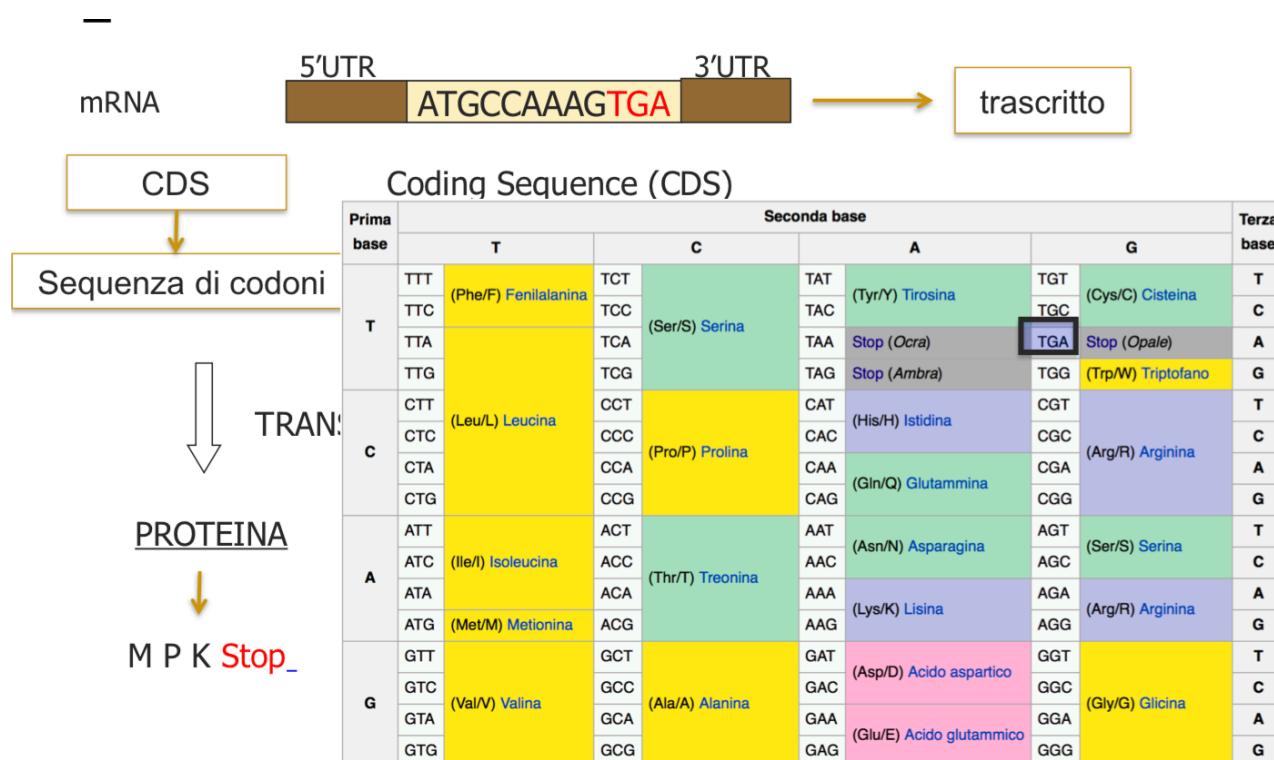
50

ESPRESSIONE (STRUTTURA) DEL GENE



51

ESPRESSIONE (STRUTTURA) DEL GENE



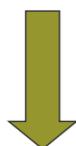
52

ESPRESSIONE (STRUTTURA) DEL GENE

Geni umani → circa 20,000

Proteine umane → centinaia di migliaia

Perché la corrispondenza 1:1 tra geni e proteine non è rispettata?



ALTERNATIVE SPLICING (AS)

Un gene è in grado di combinare i suoi esoni in modi diversi al fine di esprimere una molteplicità di trascritti (e quindi di proteine)

stesso ordine, ma uno si e uno no, oppure...

Malattie genetiche: cellula SANA ha un gene che esprime una proteina che la fa essere SANA in quella malata, genera la malattia

53

PERCHE' SERVE SAPERLO?

- Dobbiamo sapere se stiamo lavorando sulle sequenze genomiche o sui trascritti (quindi dopo trascrizione e splicing)

GENE TRANSFER FORMAT

Il formato standard GTF permette di annotare un gene su una sequenza genomica di riferimento e fornisce (in termini di coordinate):

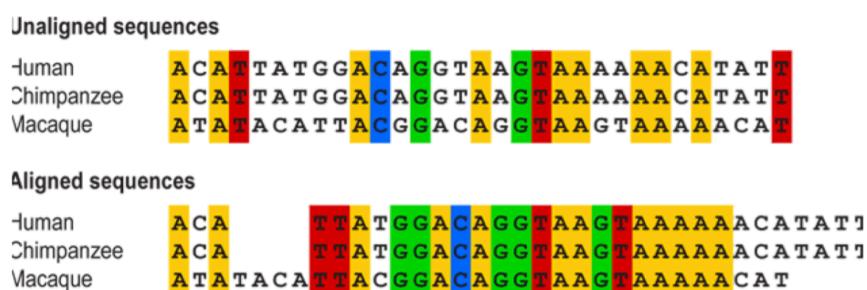
- ✓ la composizione in esoni dei suoi trascritti
- ✓ la composizione delle coding sequences (CDS) dei trascritti
- ✓ la composizione delle 5' UTR dei trascritti
- ✓ la composizione delle 3' UTR dei trascritti
- ✓ la presenza di start e stop codon delle CDS

ne ripareremo

55

Il genoma è una sequenza?

- DNA: filamento costituito da 4 nucleotidi (che differiscono per le basi azotate: A,C,G,T)



- 3 porzioni di un gene di un uomo, di uno scimpanzè, di un macaco, sono legati a livello evolutivo.
- Sovrapposizione senza spazi o con spazi (gap, delezioni, ...)

Bioinformatici hanno tool che fanno questi allineamenti

56

Viaggio al centro del DNA

Prof.ssa P. Bonizzoni - R. Rizzi (Univ. Milano-Bicocca, AlgoLab)



+ algoritmi

Sequenziamento DNA



+ algoritmi ...

Terapia medica

sono stringhe...

57

FORMATO FASTA

cromosoma

id del genoma di riferimento e la sua localizzazione

Formato standard per sequenze nucleotidiche

- ✓ plain text
 - ✓ sequenza primaria + informazioni addizionali
 - ✓ nato come formato di input del software FASTA
 - ✓ estensione → fa oppure fasta

>X dna:chromosome chromosome:GRCh38:X:154428200:154438516:1
GGATTGGCCGGGTGCAGCGATGCGCCCTGTAGTCCCACCTTCGGAGGCTGAGGCCGG
AGGATCGCTTGAGCCCAGGGACTCGAGACCAGCCTGGCCAACATAGCTAGACCCGTCTC
TAAAGAAAAAAAAAGACAA
AATAAAAACGGACGGCGAA
CCTCTTGTCTCGCCTAGTC
AGCGCCTGGAGACACGTACA
TGGTCCAATTACCTGCGGCC
CGGGGCTCGGGCGGGGGCAAA
GGAGGCTGAGGCTATGATGG
CGCCCAGGCGCTCTGGCGCATGCCGTGGCTGCCGGTGT
GGCAGGCTGGCTGGTGTGCTGGTGTGAGT
GGGCCGGGTGGGATGCGCTGTGGCGGCTGAGGCGCCCTCGCCCGACTCCGGCGCTGTCC
TAGGCAGGGGTGGTGGACTGGACTGTTCTTGCTCGAGCGAA
TCTGCCGGCGACAGAGCTCAGTCCACATGCCCCCCGCTGACAGCACCTCTGTGC
CCTGCCAGGGACTTGTGGCTCCTGCGGCCGACACTCATGAAGGCCACATCACCAGCGAC
TTGCAGCTCTCACCTACTTAGATCCCGCCCTGGAGCTGGTCCAGGAATGTGCTGTC
TTCCTGCAGGACAAGGTGCGCCGCCAGCCCACCTCTCCCCGGTCATCGGGAGGCAGC
[...]

FASTA header:

- ✓ primo simbolo → >
- ✓ unica riga

FASTA header:

- ✓ primo simbolo → >
 - ✓ unica riga

catena
diretta

Formato standard per sequenze nucleotidiche

FORMATO FASTA

- ✓ plain text
- ✓ sequenza primaria + informazioni addizionali
- ✓ nato come formato di input del software FASTA
- ✓ estensione → `fa` oppure `fasta`

```
>X dna:chromosome chromosome:GRCh38:X:154428200:154438516:1
```

```
GGATTGGCCGGGTGCGCGATGCGCGCCTGTAGTCCCACCTTTCCGGAGGCTGAGGCAGGG  
AGGATCGCTTGAGCCGGGGACTCGAGACCAGCCTGGCCAACATAGCTAGACCCCGTCTC  
TAAAGAAAAAAAAAGACAAAAGAAGAGAACACAACAAAAACCAAAATAAAATAAA  
AATAAAAACGGACGGCGAACCGCTCGGGTGGCGTAGGGTCCCGAAGCCTCCCTGTC  
CCTCTGTCTCGTCCTAGTCGGTCTAGCTGGGCCCATCCCTTCCACTCGGAGCGTG  
AGCGCTGGAGACACGTACAGCCAACCAGTGAGAAGGAGTGGCGCGAGTGGCATGCACT  
TGGTCCAATTACCTCGGCCCTGCCGGTGGGCCGCTGGGCCAATGGAGGTGCGAGG  
CGGGGCTCGGGCGGGCAACGGTCACCTGATCTGGCGCTGTCGAGGCCGCTGAGGCAGT  
GGAGGCTGAGGCTATGATGGCGGCATGGCGACGGCTCGAGTGCAGTGGATGGGCCGCGT  
CGCCCAGGCCTCTGGCGATGCCGTGGCTGCCGGTTTGTGTTGGCGGGCGCG  
GGCAGGCGCAGCGGCGGAGCAGCAGGTCCCGCTGGTGCTGTGGTCGAGTGACCGGTGAGC  
GGGCCGGGTGGGATGCCGTGTGGCGCTGAGGCCCTGCCCGACTCCGGCGCTGTCC  
TAGGCGAGGGGTGGTGAGGCCGGAGGTGGACTGTCCTGCTCGGGGCTCGCAGCGAA  
TCTGCCGGCGACAGAGCTCCAGTCCACATGCCCGTCTGACAGCACCTCTGTGC  
CCTGCCAGGGACTTGTGGCTCCTGCGGCCGACACTCATGAAGGCCACATCACCAGCGAC  
TTGCAGCTCTACCTA  
TTCCTGCAGGACAAGG  
[...]
```

Sequenza separata in righe di 60/80bp simboli

59

ENSEMBL www.ensembl.org

ne ripareremo

Ensembl è un Genome Browser

- ✓ Database + Interfaccia di esplorazione
- ✓ Iniziato nel 1999 da EMBL + Wellcome Trust Sanger Center
- ✓ Dal 2017 fa capo a EMBL-EBI
- ✓ Accesso all'informazione raccolta in relazione a genomi di 250 specie di cordati (Aprile 2020). Specie chiave:
 - uomo, topo, ratto e danio rerio (zebrafish)
- ✓ Accesso a diversi livelli di informazione (genoma, gene, proteina)
- ✓ Integrazione e consistenza rispetto a progetti e analisi genomiche

ENSEMBL www.ensembl.org

Le “cose” principali a cui si può accedere sono:

- ✓ la sequenza del genoma di una data specie o dei singoli cromosomi
- gene con informazioni
- ✓ le sequenze dei geni annotati su di un genoma e i relativi trascritti espressi

Interrogazione tramite genomic location
X:154428200:154438516

61

come si arriva a questi file

sequenziamento

62

Sequenziamento

- Non possiamo leggere nucleotide per nucleotide
- Quindi come possiamo sequenziare il genoma?
- I biologi possono identificare piccoli frammenti di DNA (100 nucleotidi), chiamati **reads** o frammenti più lunghi (10K, cioè 10000) chiamati **long reads**

63

Fonti

- <https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genome-project>
- [pdf sul sito]
 - Setubal, Meidanis, Introduction to Computational Molecular Biology, Cap. 1
 - Jones, Pevzner, Bioinformatics Algorithms, Cap. 1-2-6
- <https://www.focus.it/scienza/salute/vaccini-mrna-covid-storia-scoperte>

64