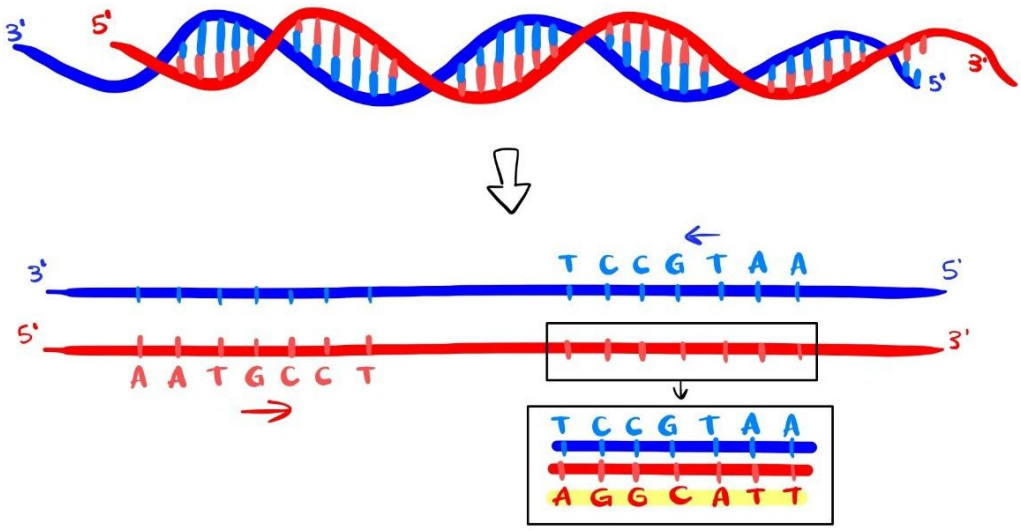
Piccolo disclaimer iniziale: questi calcoli sono basati sull'assunto teorico che tutte le basi abbiano la stessa probabilità di essere "selezionate" e che la sequenza del DNA sia completamente casuale. Queste sono approssimazioni che però si rendono necessarie per poter fare i calcoli.

Ricapitolando: qual è la probabilità che una specifica sequenza di 20 bp si trovi all'interno del DNA umano (3Gbp)?

Per prima cosa una considerazione sul DNA effettivamente "utile" per il nostro calcolo. Il nostro genoma contiene delle sequenze ripetitive che possiamo escludere dalla nostra ricerca, che sono circa il 50%.
Quindi per i nostri calcoli considereremo una lunghezza del DNA umano di 1,5 miliardi di base pair.

Poi c'è da considerare il fatto che, per il modo in cui è costruito il DNA, le sequenze che andrebbero bene sono in realtà due. Questo perché la sequenza che stiamo cercando potrebbe essere sull'altro filamento, quindi ci va bene anche la sequenza complementare.
Sembra complesso, ma una mia amica è stata così gentile da fornire una veloce illustrazione che penso renda tutto più chiaro.



Analizzando il filamento rosso, ipotizziamo che la sequenza che ci interessi sia quella a sinistra (il DNA si legge dall'estremità 5' alla 3' quindi AATGCCT).
Come potete vedere sul filamento blu, a destra, c'è un'altra sequenza identica (sempre leggendo da 5' a 3'). Questo si riflette sul filamento rosso in un'altra sequenza (AGGCATT) per via degli accoppiamenti delle basi del DNA (C-G e A-T).
Per questo motivo, cercando sul filamento rosso, possiamo dire che le sequenze utili siano in realtà due: AATGCCT e la sua complementare specchiata AGGCATT.

Passiamo adesso al calcolo vero e proprio.

Dal momento che il DNA può essere formato da quattro diverse basi, la probabilità che ognuna venga selezionata, in un dato momento, è 0,25 o ¼.

Quindi la probabilità di ottenere una specifica sequenza di 20 bp è ¼ × ¼ × ¼… così per 20 volte

$$\frac{1}{4^{20}}$$

Quante possibili sequenze da 20 bp ci sono nel DNA?
Ogni base del filamento può essere la prima di una nuova sequenza utile, tranne per le ultime 19.
Possiamo quindi approssimare il numero di sequenze possibili a 1,5 miliardi ($1{,}5 \times 10^9$).

Detto ciò, i calcoli sono più intuitivi se partiamo dalla probabilità che la sequenza che vogliamo non si presenti MAI.

La probabilità di questo evento è

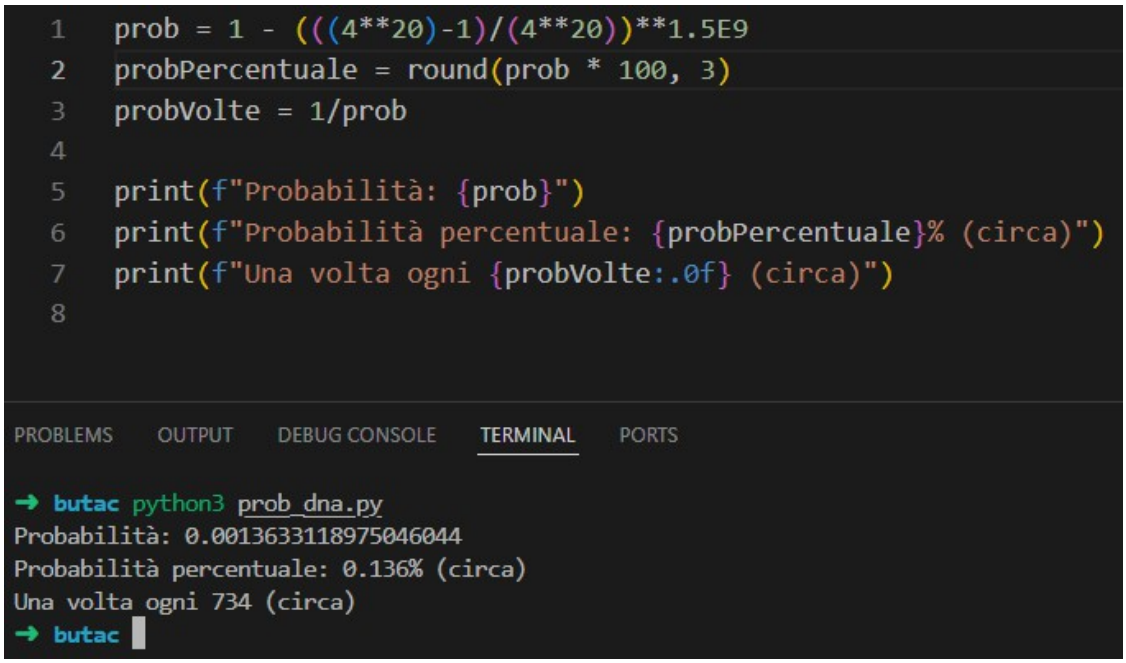
$$\left[\frac{(4^{20}-1)}{4^{20}}\right]^{1{,}5 \times 10^9}$$

Per avere invece la probabilità di trovare la sequenza che ci serve basterà fare 1 - il risultato del calcolo sopra.

La probabilità risultante andrà poi moltiplicata per due, visto che come abbiamo detto le sequenze utile sono due.

Il piccolo problema tecnico è che le calcolatrici normali non sono in grado di gestire calcoli come questo (elevare alla 1,5 miliardi), ma per fortuna mi sono venuti in aiuto.

A quanto pare Python può essere usato, oltre che come linguaggio di programmazione, anche come calcolatrice sotto steroidi.
Trascriviamo quindi il calcolo su Python. (** significa “elevato alla”, E significa “10 elevato alla”)



La risposta che emerge quindi è 0,136%.

Questa probabilità va poi moltiplicata per due, quindi 0,272%.