

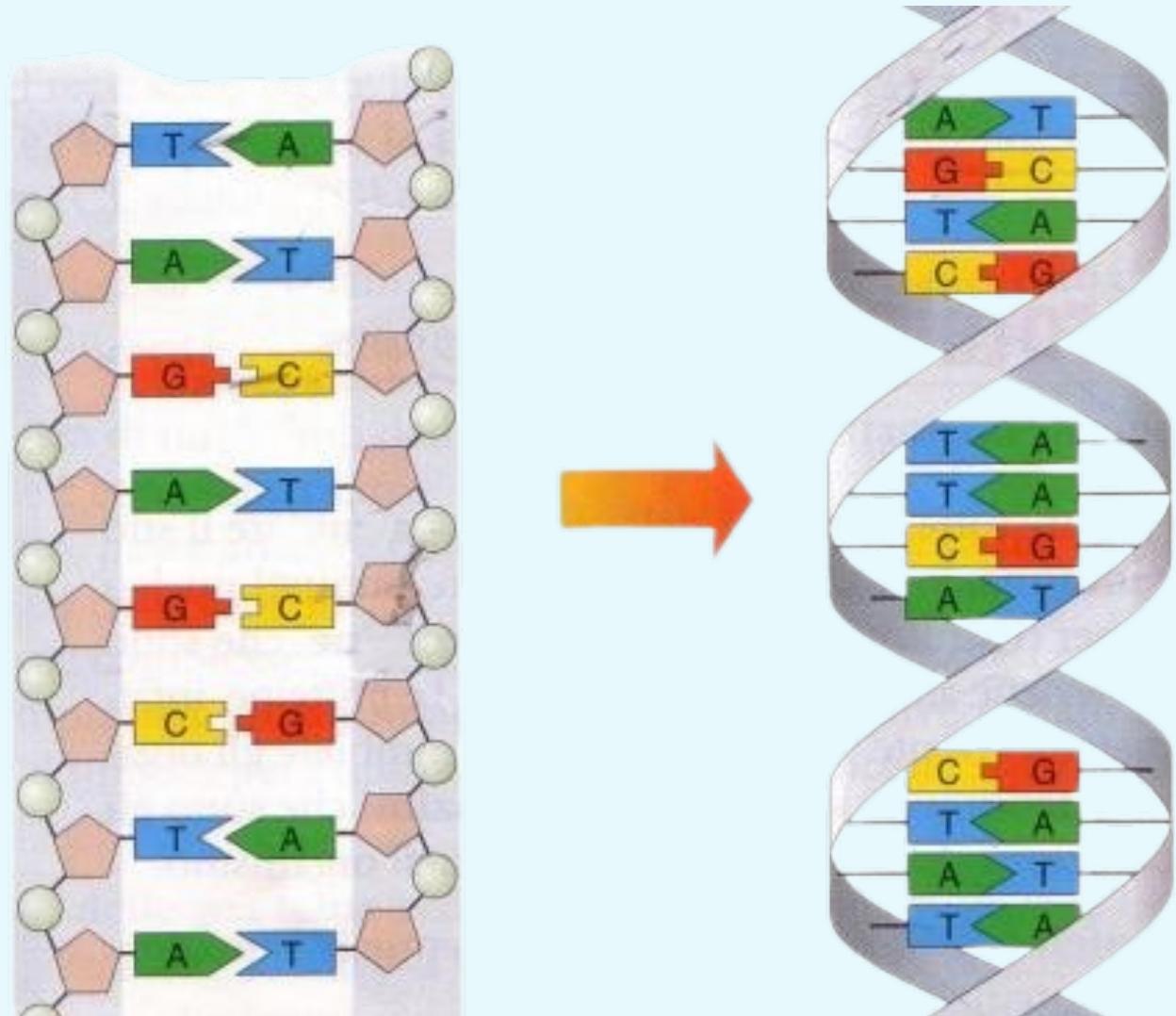
# Sense PUZZle

Siamese neural network for  
sequence embedding

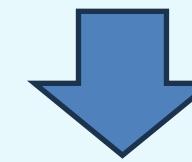
Progetto Strumenti Formali per la Bioinformatica  
Mattia d'Argenio - Davide Di Sarno

# INTRODUZIONE

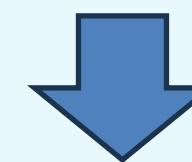
L'**allineamento di sequenze** è la procedura con cui vengono messe a confronto ed allineate due o più sequenze di aminoacidi, DNA o RNA. Permette di individuare regioni identiche o simili che possono avere relazioni funzionali, strutturali o evolutive.



L'analisi delle sequenze è essenziale in bioinformatica per applicazioni come ricerca nei database e predizione dei geni.



Uso di metodi tradizionali come l'algoritmo Needleman-Wunsch.



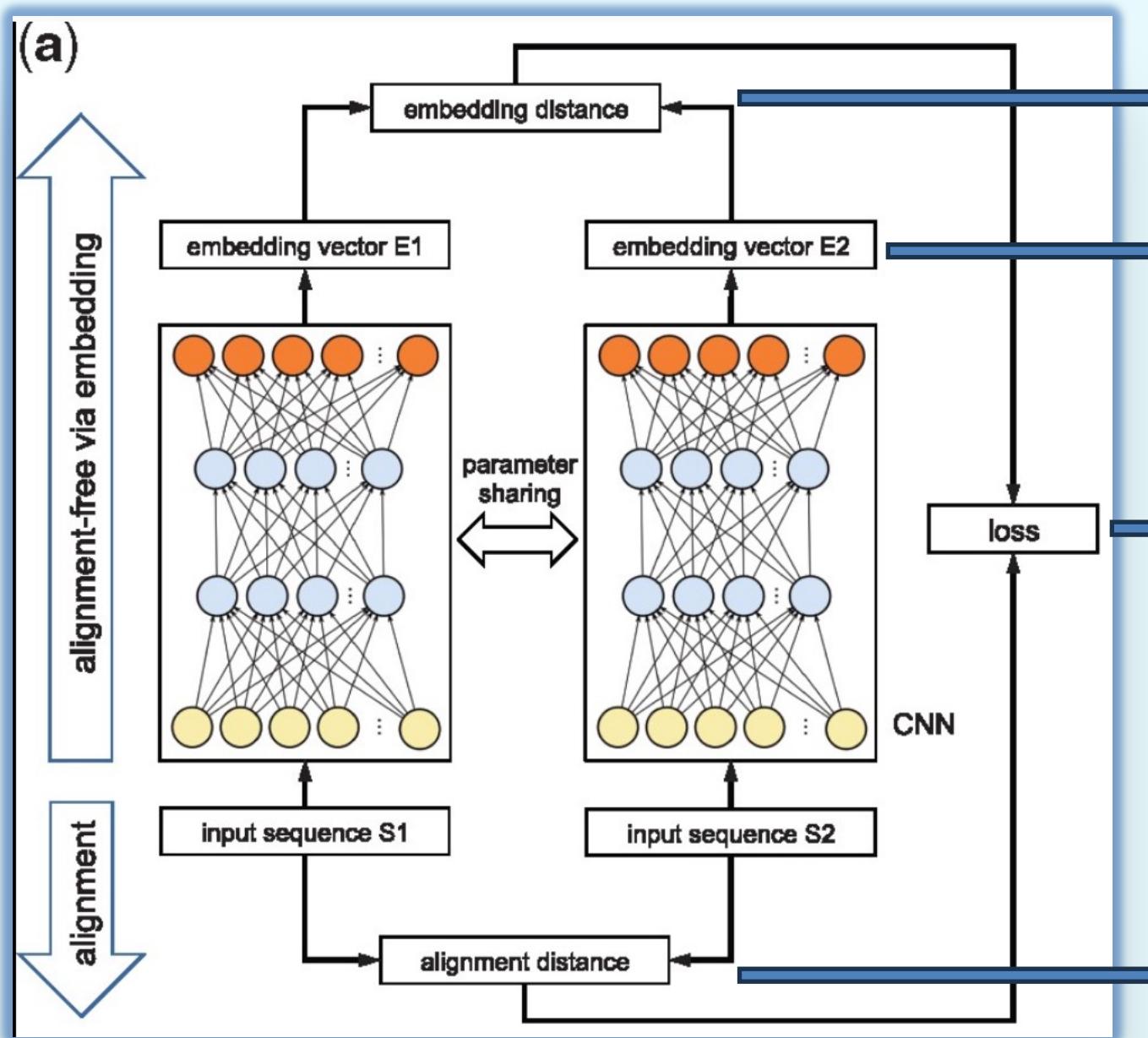
limiti computazionali



Uso di modelli intelligenti avanzati per un confronto efficiente e preciso delle sequenze, superando le limitazioni

# SENSE

## SiamEse Neural network for Sequence Embedding



Distanza di allineamento tra gli embedding calcolata con la formula di Jaccard generalizzata

Utilizzo di rappresentazioni vettoriali

La rete viene addestrata con back-propagation minimizzando l'errore quadratico medio tra le due distanze calcolate

Distanza di allineamento computata con l'algoritmo di NW

# Modello Siamese

Il modello Siamese utilizza una coppia di gemelli neurali per apprendere rappresentazioni comuni e consentire un confronto preciso tra sequenze di DNA.

## 1. Architettura Gemelli Siamesi:

Due bracci neurali identici elaborano simultaneamente coppie di sequenze di DNA, garantendo una valutazione coerente.

## 2. Elaborazione delle Coppie di Dati:

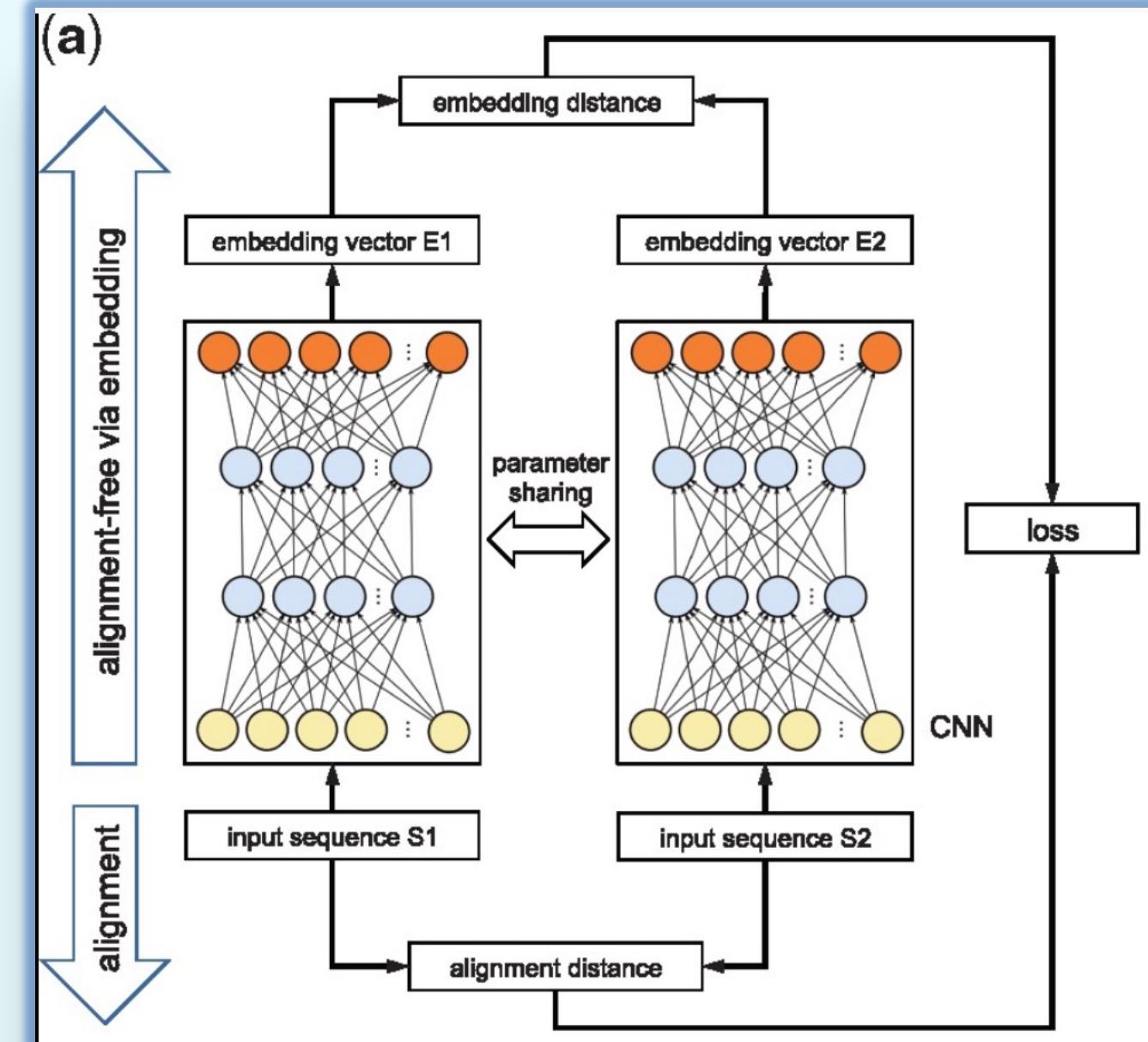
Coppie di sequenze, rappresentanti dati simili o distinti, sono processate dai gemelli Siamesi per estrarre rappresentazioni significative.

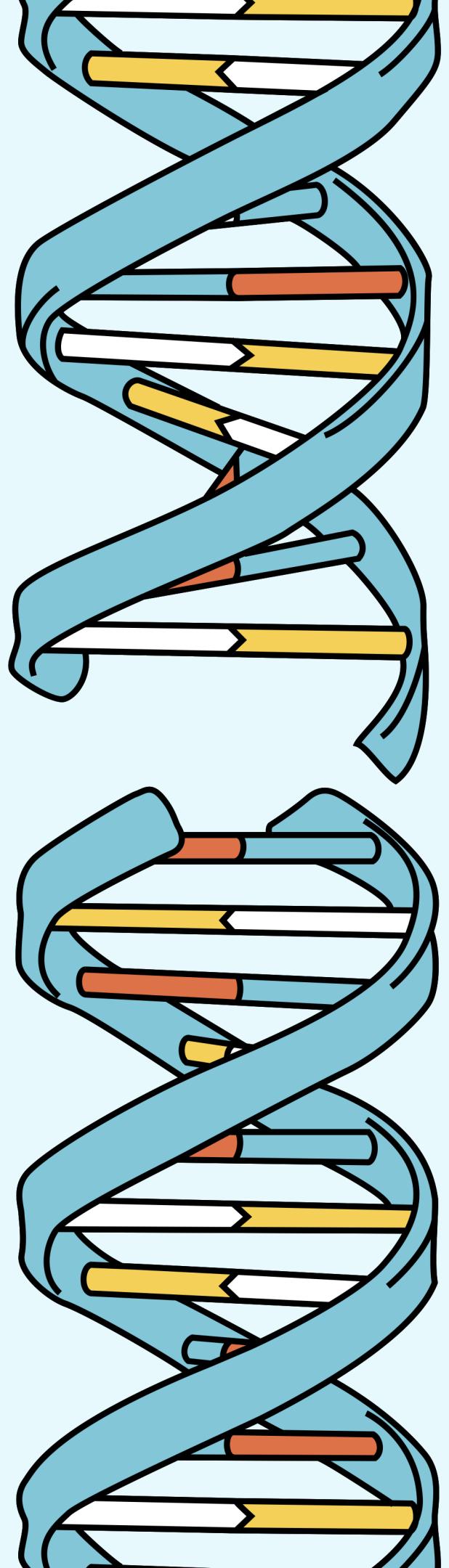
## 3. Apprendimento di Rappresentazioni Comuni:

L'obiettivo principale è acquisire rappresentazioni condivise, assicurando che sequenze simili siano rappresentate in modo simile.

## 4. Calcolo della Distanza:

Utilizzando le rappresentazioni apprese, il modello calcola la distanza tra le coppie di sequenze, fondamentale per il confronto e l'analisi.





## SENSE WORKFLOW

SENSE: <https://github.com/lh3/seqltk/tree/master>



file.fa → 10000 sequenze

sample.py → 500 sequenze

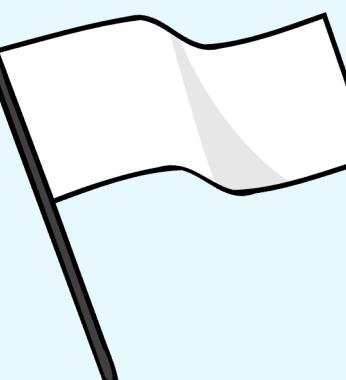
pair.py

nw (Needleman-Wunsch)

dist.py

select\_training\_data

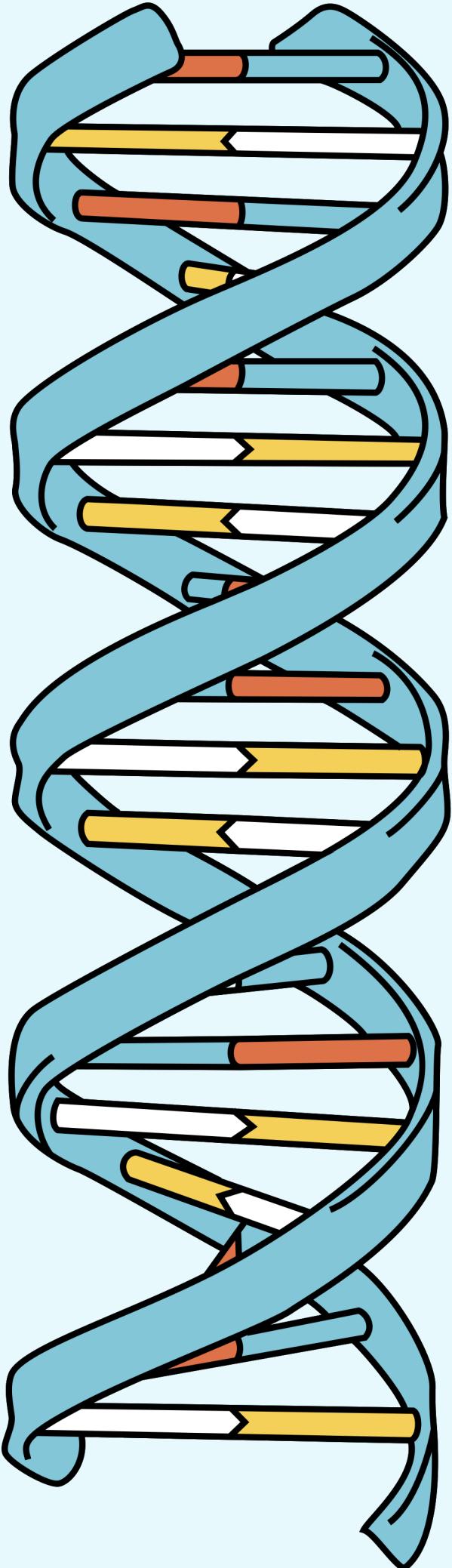
shuffle.py → txt

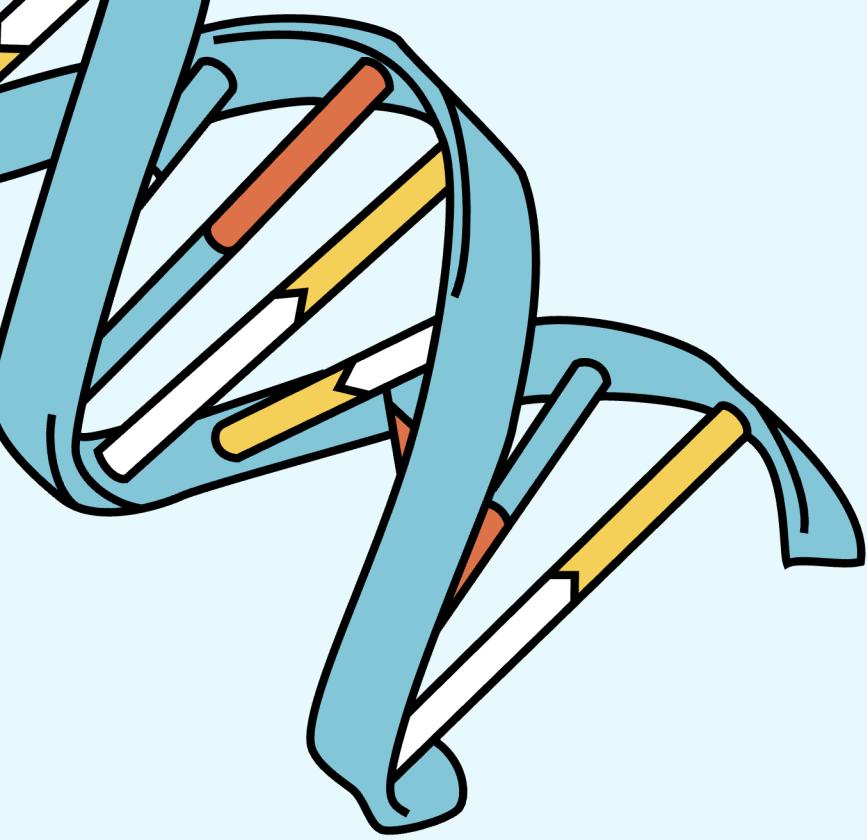


siamese.ipynb

## OBIETTIVI

- 1.Utilizzo del Modello Sense:** Utilizzare il modello Sense per ottenere un allineamento ottimizzato delle sequenze nucleotidiche, massimizzando efficacia e precisione.
- 2.Metodologie di Allineamento Miste:** Utilizzo di approcci "brute force" e algoritmi genetici per l'accoppiamento delle sequenze.
- 3.Innovazione con Algoritmo Genetico:** Tracciamento delle sottosequenze e valutazione accurata delle sequenze più allineabili.
- 4.Confronto e Ottimizzazione:** Analisi dell'efficienza dell'algoritmo genetico rispetto al "brute force"





## DATASET

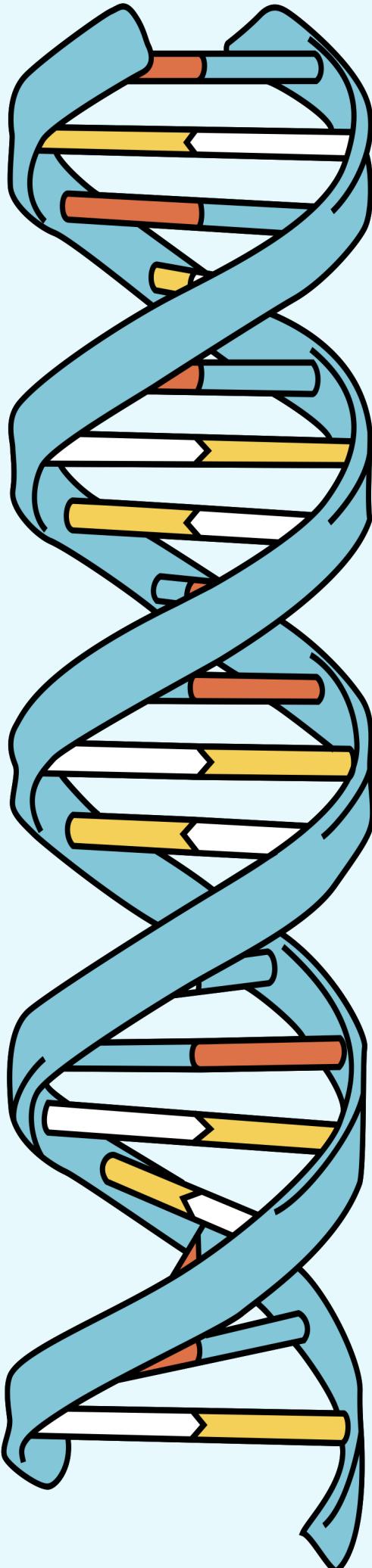
Nell'esperimento condotto, abbiamo utilizzato due dataset di sequenze biologiche: il dataset Qiita, composto da 66 campioni di pelle, saliva e feci con un totale di 6,734,572 sequenze di 151 bp, coprenti la regione ipervariabile V4 del gene 16S rRNA. Il secondo dataset veniva creato dall'algoritmo in modo artificiale rispettando i criteri del dataset originale, con lo scopo di testare la corretta efficacia di SENSE anche su sequenze piuttosto diverse.

## Preprocessing

- Conversione in FASTA
- Eliminazione caratteri malformati
- Eliminazione sequenze troppo corte
- Unificazione in un dataset uniforme

• • •

```
>MISEQ02:42:000000000-AAN9D:1:2114:11465:27347#3:N:0
CCTCTCGATCCCCATGTTGCCCGAGCGTCAGTAGTGCAGCGGATAGCTGCCTCGCTTGCGTCCGTAGATCTACGCTTCACCCCTACC
CCACGCACCTCGCTCTCCCTTTCTCCACTATCCTGTTCTC
>MISEQ02:42:000000000-AAN9D:1:2114:14871:27348#3:N:0
CCTGGTTGCTCCCCACGCTTCGAGCCTCAGCGTCAGTTACAAGCCAGAGGCCGCTTCGCCACCGGTGTTCCATATCTACGCATTCAACGCTACA
CATGGAATTCCACTCTCCCTTTGCACTCAAGTTAACCGTTCCAA
>MISEQ02:42:000000000-AAN9D:1:2114:11485:27348#3:N:0
CCTCTTGCTCCCTATGCTCCGCCCGAGCGCCGACGCACCCAGTAGTTGCTTGTCCCGGTCTCCTTCACAGCCCTACGTATCCACTGCTAAC
TGCGGAATTCCAATTCTCCGCTACCTCTAGTCTCCAGTTAGAC
>MISEQ02:42:000000000-AAN9D:1:2114:14474:27348#3:N:0
TCTAAATGTCGTCAGTGCTGCTGGTGTGGTGTATTGCTAACGCTGGCCGATACTGGAAACAATTGGAAAGACGGTAAAG
CTGATGGTATTGGCTCTTTGTCTACTCAAGAACGGTCGGTATTAA
>MISEQ02:42:000000000-AAN9D:1:2114:15120:27348#3:N:0
TGACTCGCAAGGTTAGTGCTGAGGTTGACTTAGTTCATCAGCAAACGCGAGAACGCGGTATGGCTTCTCATATTGGCGTACTGCAAAGGATATTCTAA
TGTGCCACTGATGCTGCTGGTGCAGGTGATATTCTCATGGTAT|
```



# Allineamento

# Metodo di Allineamento Globale: `pairwise2.align.globalxx`

Usa un sistema di punteggi che tiene conto di corrispondenze, mismatch e penalità per gap. Implementa l'algoritmo di programmazione dinamica, riempiendo una matrice di punteggi e ricostruendo successivamente l'allineamento.

# Metodo di Formattazione: pairwise2.format\_alignment

organizza l'allineamento tra le sequenze in un formato più leggibile. Divide l'allineamento in linee di lunghezza massima prestabilita, evidenziando le corrispondenze con il simbolo "|" e gli spazi per le differenze.

## Brute force

Metodo sistematico che cerca di identificare l'allineamento ottimale per ogni sequenza esplorando tutte le possibili configurazioni. Questo processo si articola in diverse fasi:

- **Generazione embedding:** le sequenze genomiche vengono trasformate in rappresentazioni vettoriali dense attraverso la rete siamese
- **Calcolo distanza:** viene calcolata la distanza tra gli embedding
- **Valutazione dell'allineamento:** ogni coppia ha un punteggio di allineamento (somiglianza) che permette di valutare la coppia in modo quantificabile
- **Identificazione dell'Allineamento Ottimale:** determinazione dell'allineamento che massimizza il punteggio di somiglianza.

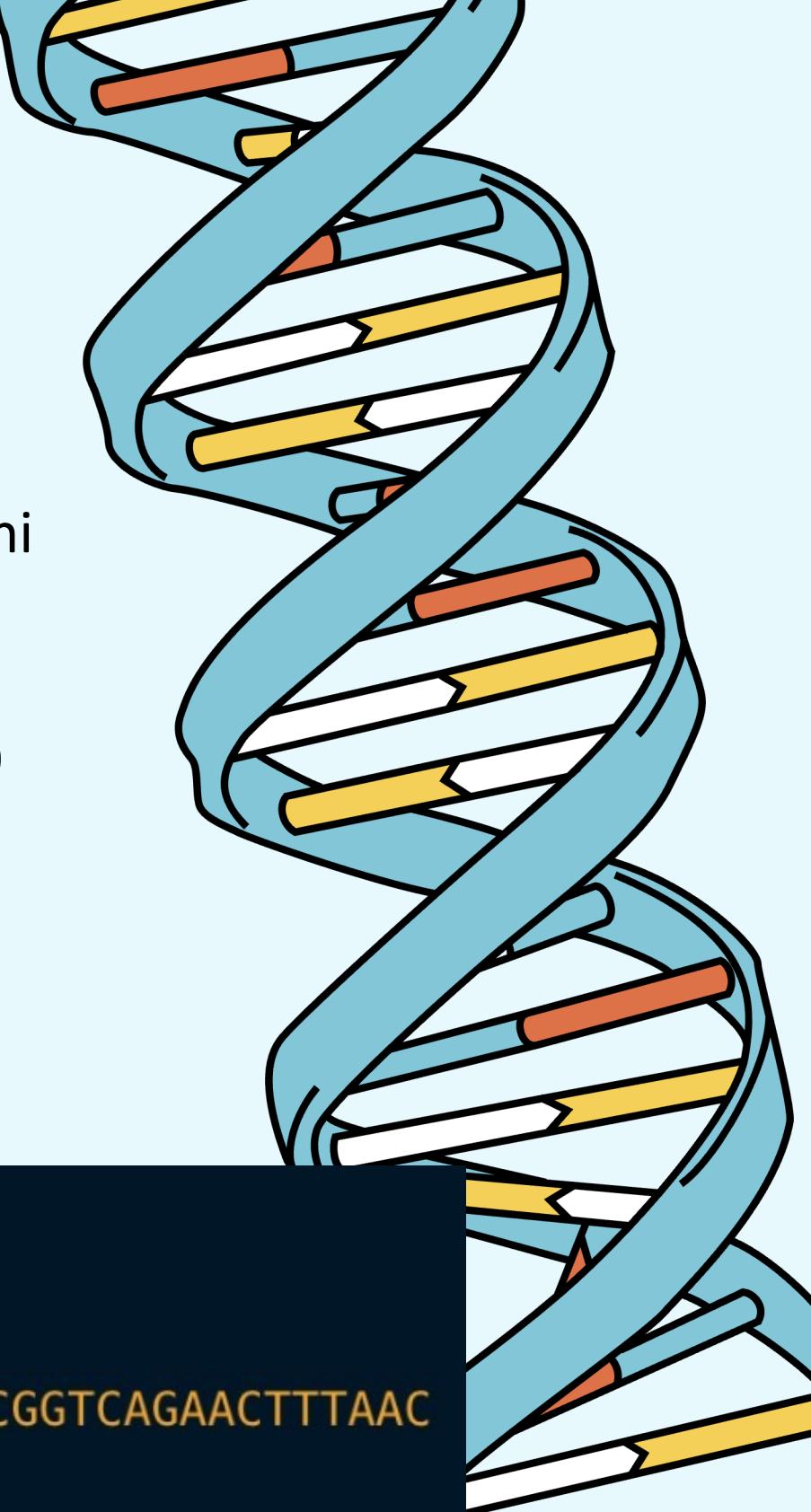
Per la sequenza 7:

Miglior punteggio di allineamento: 0.4636295437812805

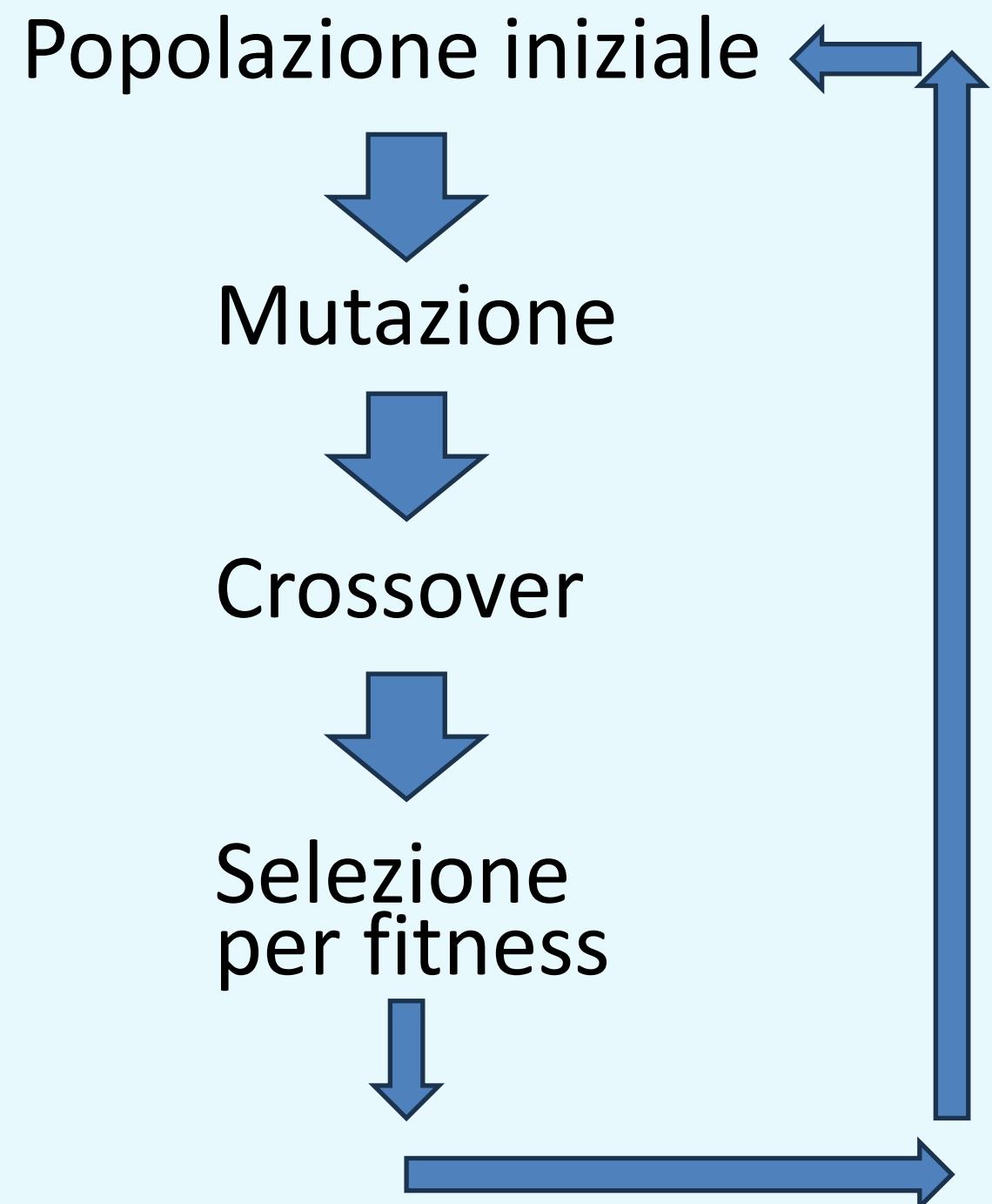
Sequenza corrispondente:

TAAGTGGCTAAAGATTGTAGATATCTAGCCAAGTGTACACGGCTTGTCAAGGCCACTCTAGTAGTCGTGGGTAGAGTCGGAGAGCACC GGTCAGAACTTAAC  
ATATGCTGGACGGCGGTGGCTGGAAGCAGTGTGATGGTGGTCTAGC

Numero di sequenza corrispondente: 8 su 10 (percentuale: 80.00%)



# Algoritmo Genetico

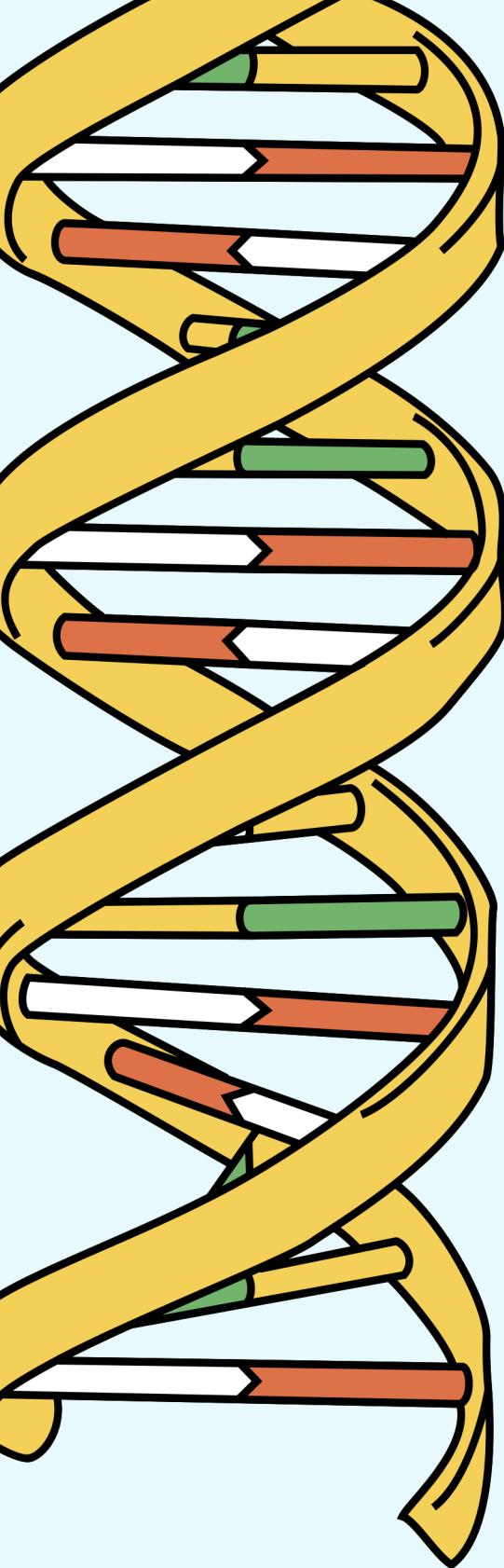


Un individuo della popolazione è composto da una coppia di sequenze casualmente selezionate dal dataset di partenza.

Ogni individuo è un possibile accoppiamento tra le sequenze che abbiamo a disposizione.

Lo scopo dell'algoritmo genetico è riuscire ad individuare gli accoppiamenti migliori tramite gli operatori applicati.





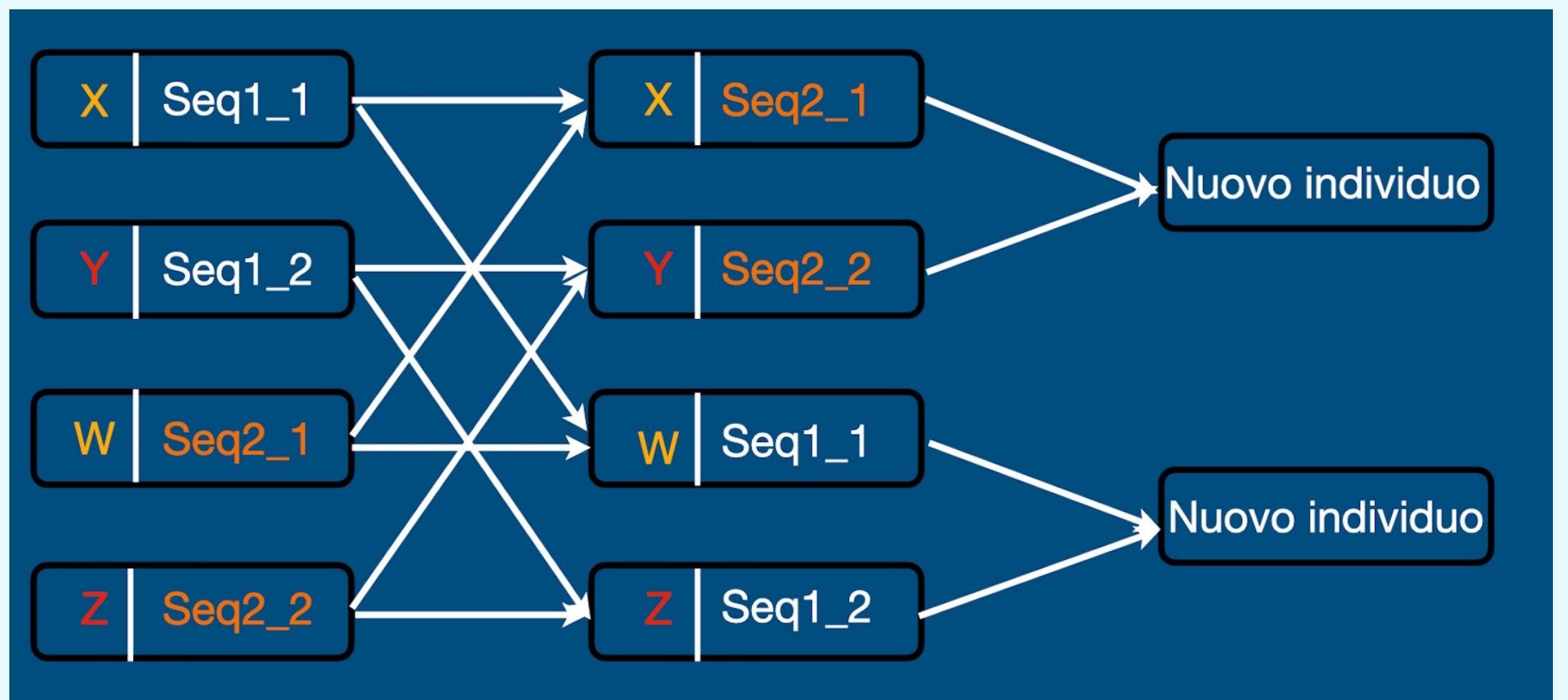
## CROSSOVER

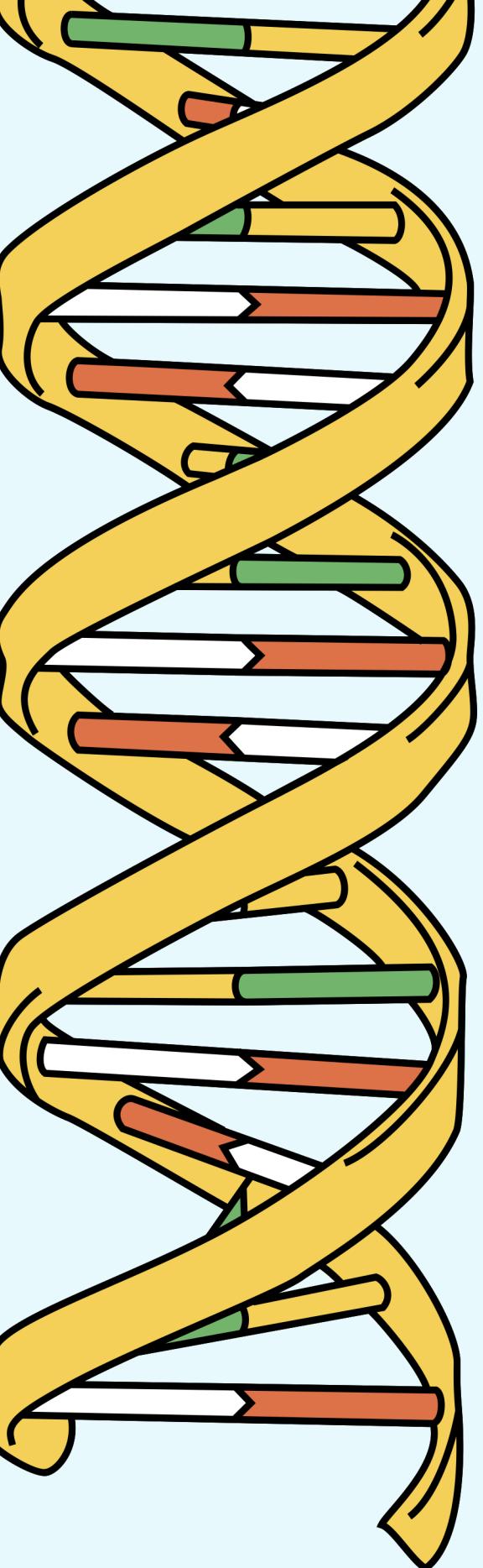
Popolazione iniziale

Mutazione

Crossover

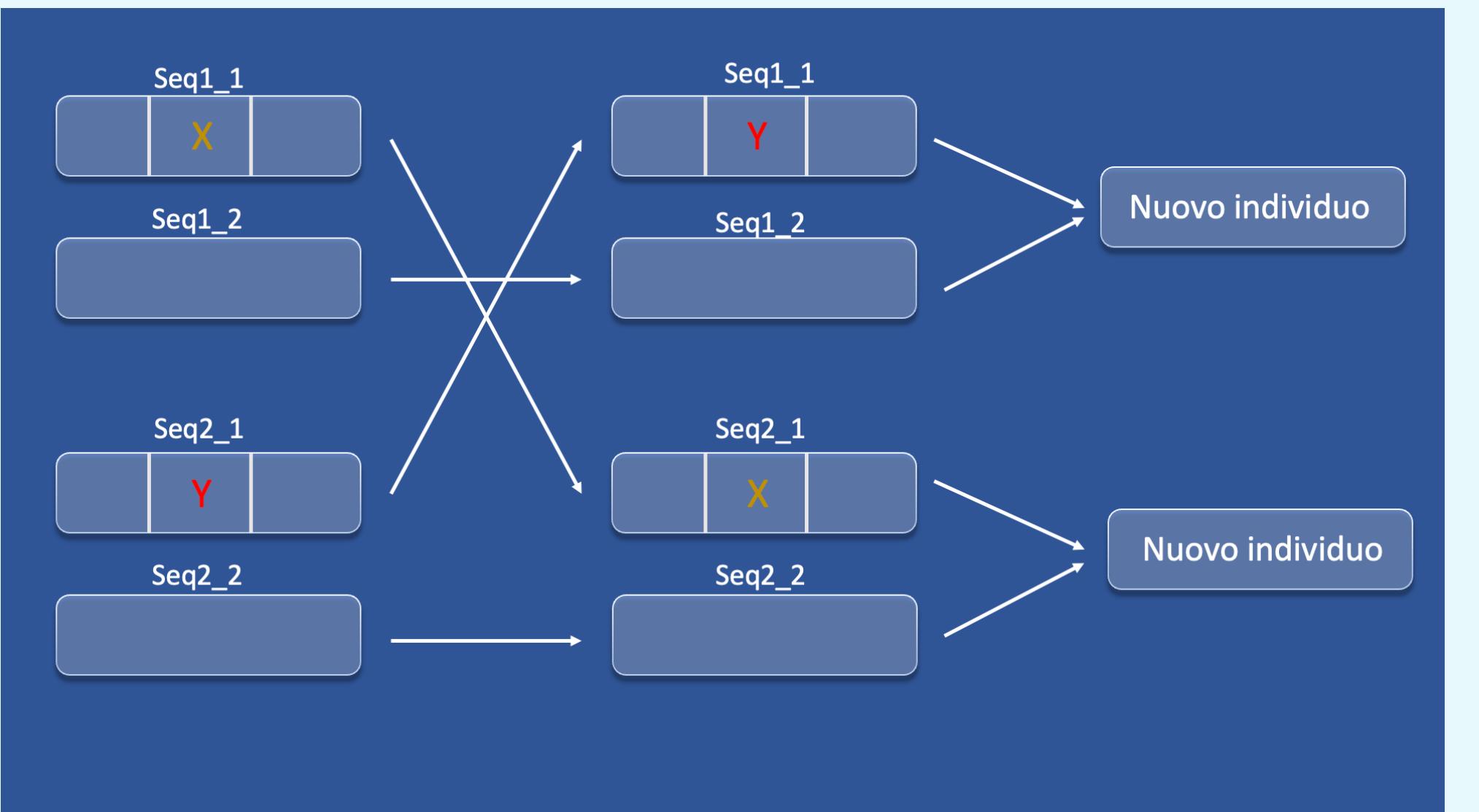
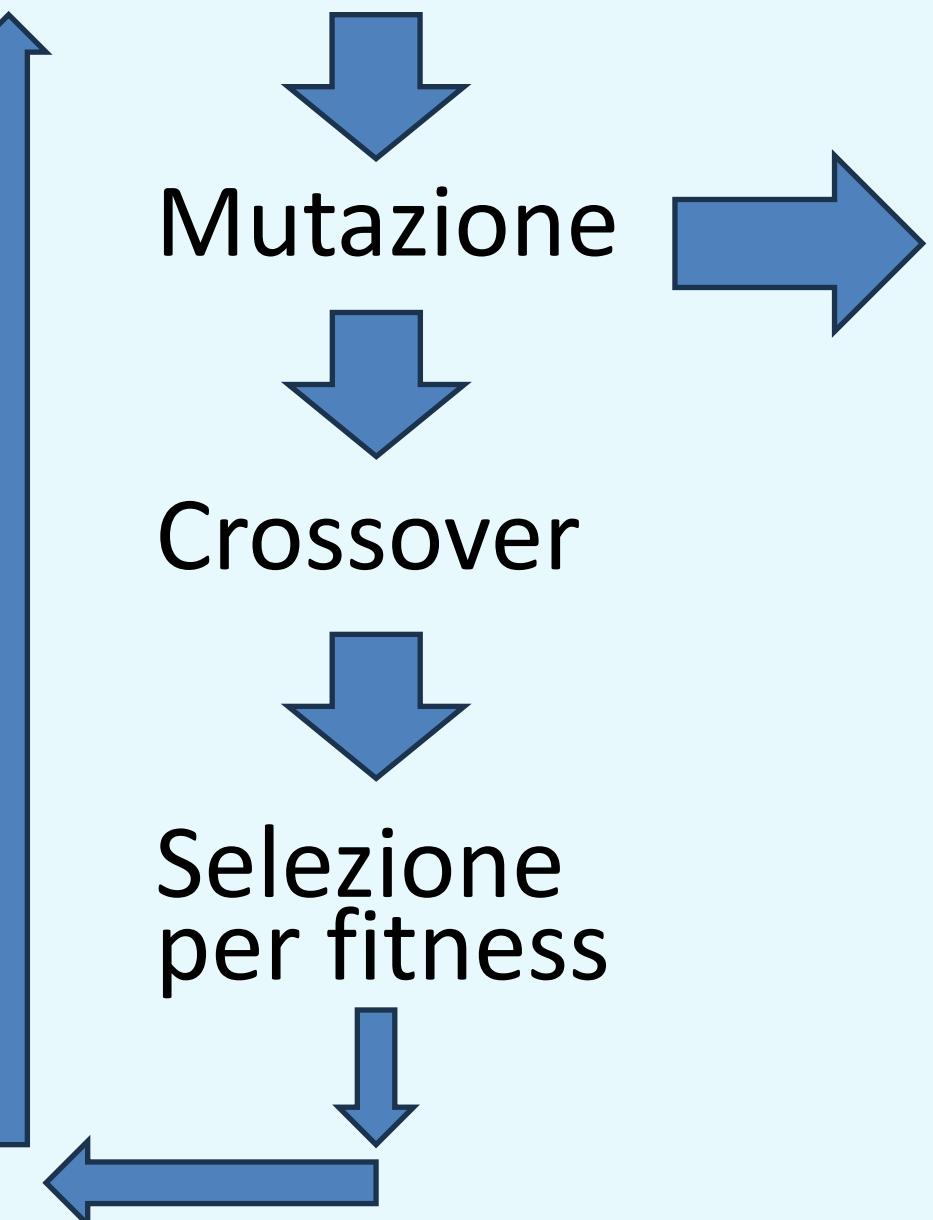
Selezione  
per fitness



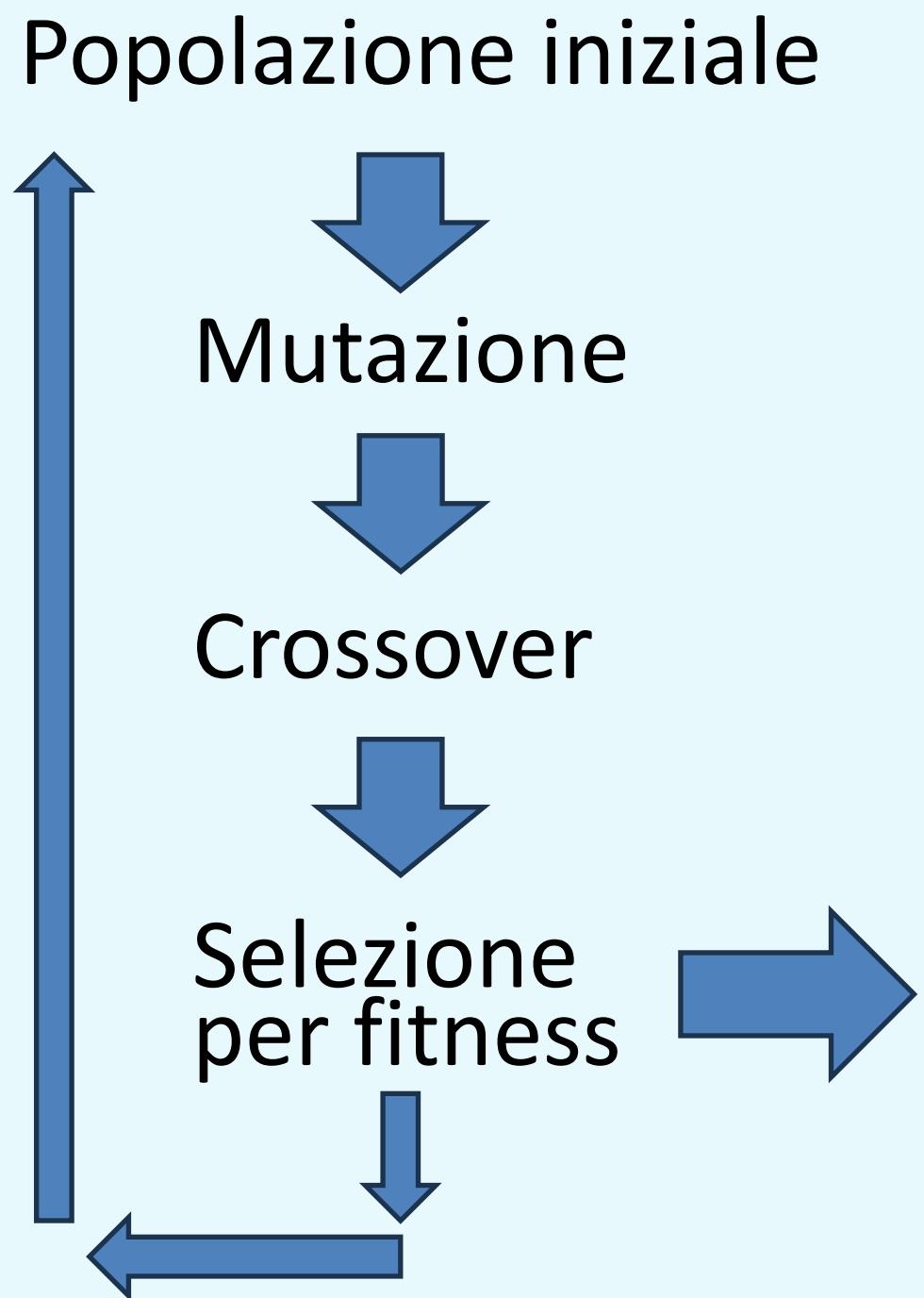


## MUTAZIONE

Popolazione iniziale

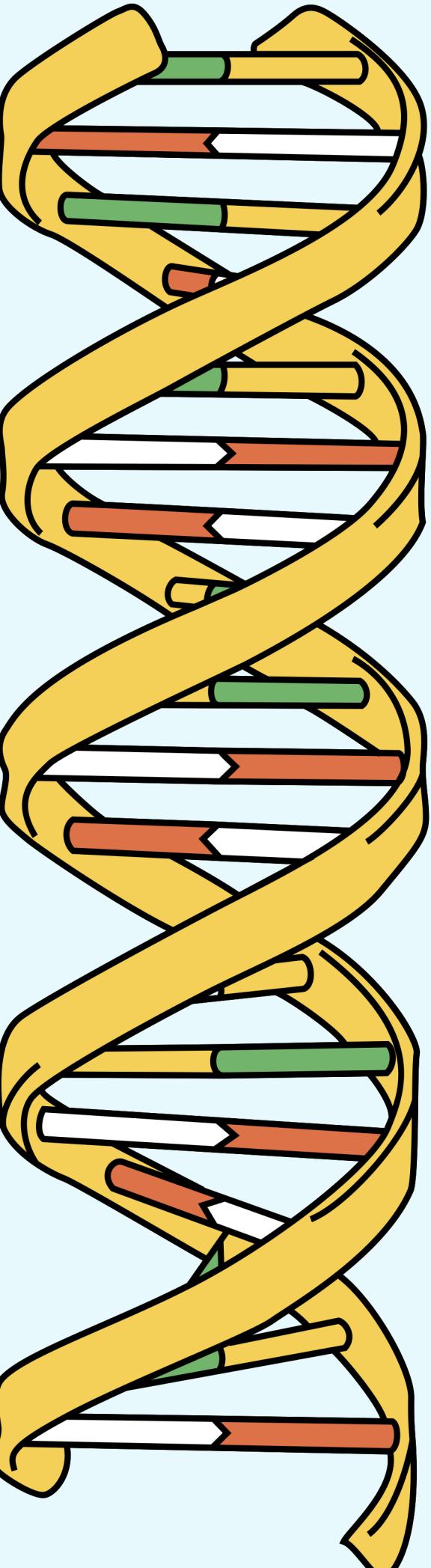


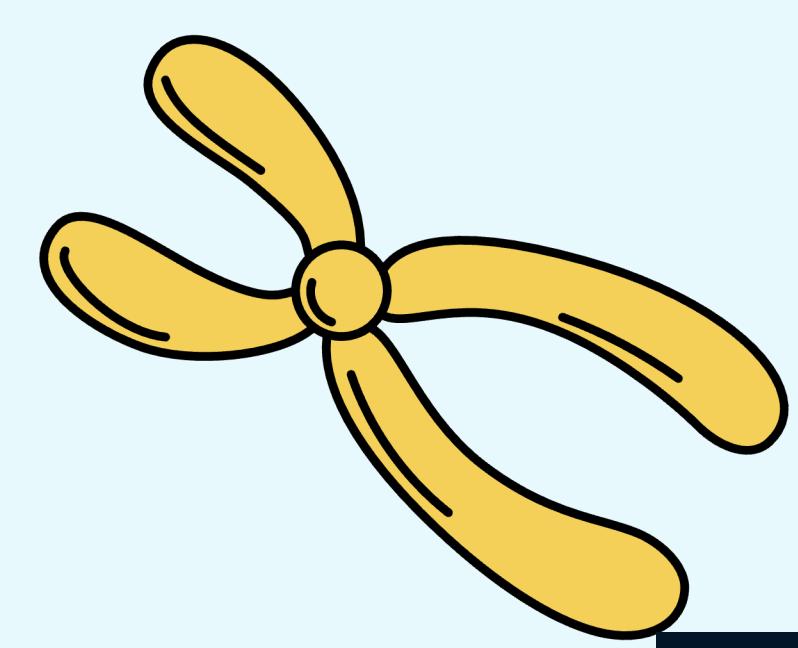
## FUNZIONE DI FITNESS



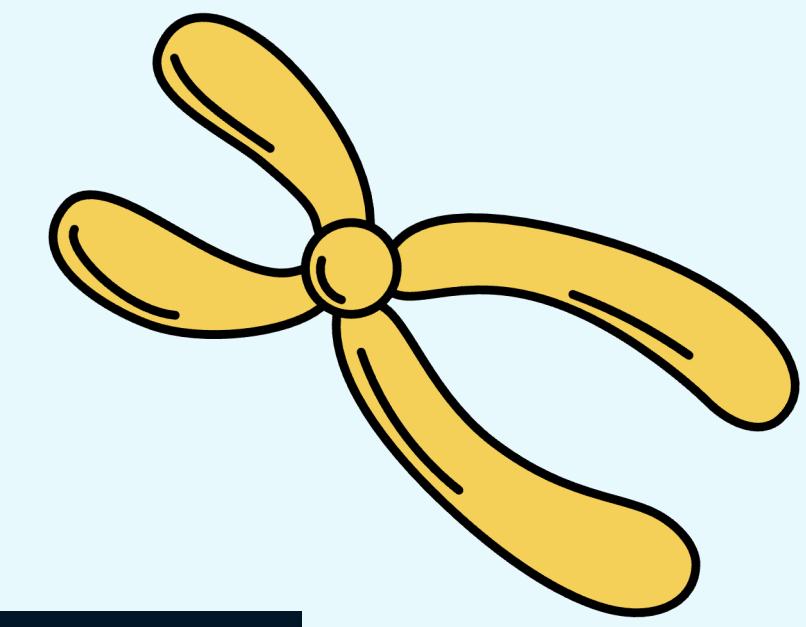
Una volta ottenuta la popolazione con gli individui mutati e crossoverati si vanno a selezionare i migliori accoppiamenti in ordine di fitness, ovvero quelli che hanno un punteggio di allineamento migliore. Questi saranno i genitori della popolazione successiva.

La funzione di fitness utilizzata è **Sense**, con la sua capacità di calcolare la distanza tra gli embedding delle sequenze analizzate.





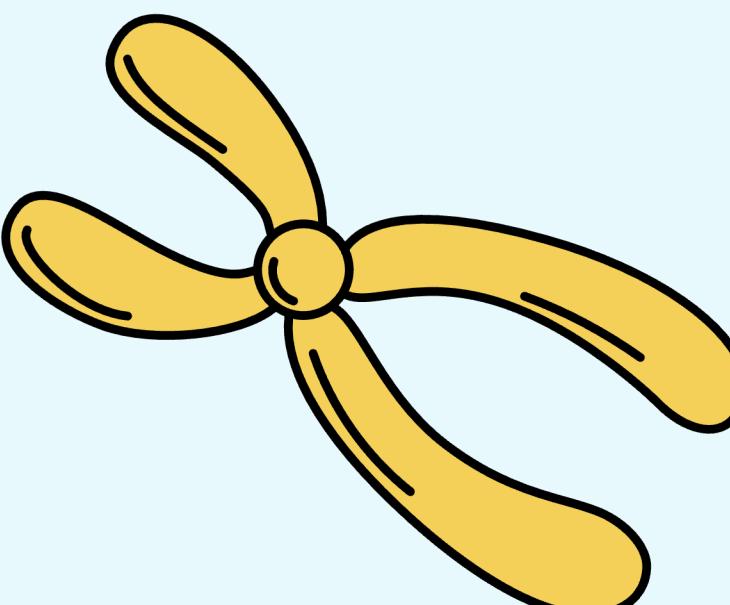
# **Analisi genetica**



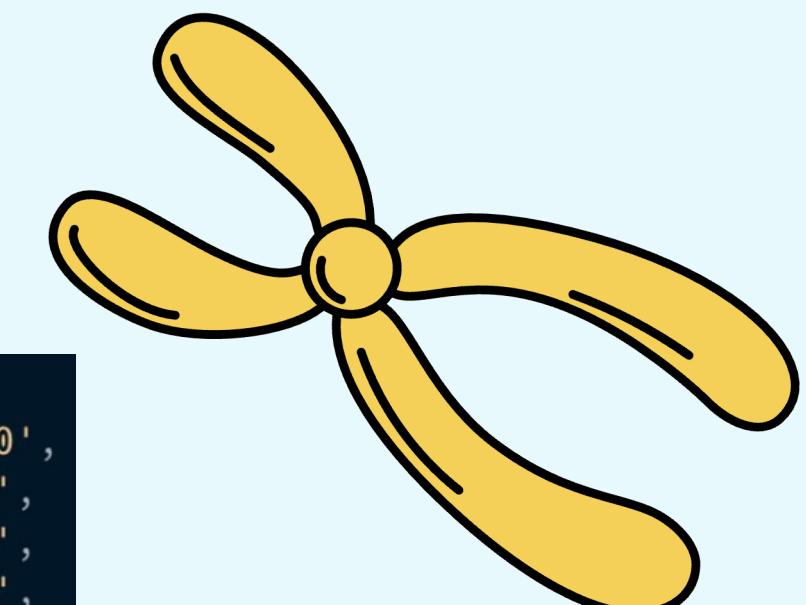
Sequenza 1: 9, Sequenza 2: 3

## Sequenza 1:

GAGAACAGCGGCTGGACAACGTCCATAGGCCTATCAGATTAAACAGCCTTATCAGATATTTTCCGGGATGACTATCGCGGCGACTGACTGTTGAGAAA  
GACGACTTGAAATTCTTGGCGGACATACAGGGCAGTTACGTAATAG, Sequenza 2:  
CCTTGTGTTATGGGAGGCAACTGCCGGTCTGAGAGGGTCGCCCCGTCTCATGACACAAACCCCCGCAGGTCGACTTGCTAGGAACGAGCCATTGCTA  
TTAATGCAAGGCCGGCACGCCGGACCACGTGATGGATCAGTGAA



# Analisi genetica



```
array originale 1_1
['0', '9', '9', '3', '3', '7', '1', '7', '7', '5', '3', '9', '8', '5', '1', '3', '9', '3', '9', '0',
'7', '9', '3', '0', '9', '7', '5', '8', '3', '9', '0', '8', '0', '0', '5', '0', '1', '3', '1', '9',
'7', '0', '1', '7', '1', '0', '0', '3', '0', '0', '8', '7', '9', '5', '5', '7', '3', '3', '0', '3',
'3', '8', '7', '5', '8', '9', '9', '5', '1', '9', '1', '3', '8', '3', '9', '9', '5', '9', '8',
'3', '7', '7', '0', '1', '8', '7', '5', '9', '9', '7', '7', '5', '9', '9', '5', '0', '9', '5', '0',
'5', '7', '1', '9', '1', '3', '7', '3', '5', '7', '9', '5', '3', '9', '5', '7', '5', '7', '9', '9',
'0', '7', '9', '3', '7', '7', '3', '3', '1', '1', '0', '0', '1', '8', '7', '3', '9', '9', '1',
'8', '0', '0', '9', '8', '7', '9', '9', '0', '1']

nuovo array 1_1
['0', '9', '9', '3', '3', '7', '1', '7', '7', '5', '3', '9', '1', '5', '9', '7', '7', '7', '0', '5',
'9', '5', '0', '8', '5', '7', '8', '1', '9', '0', '7', '0', '8', '0', '0', '7', '1', '7', '0', '5',
'3', '0', '0', '1', '8', '0', '0', '0', '3', '3', '8', '3', '9', '8', '7', '7', '7', '9', '0', '8',
'9', '9', '3', '7', '3', '9', '9', '3', '3', '3', '9', '9', '3', '7', '5', '8', '8', '9', '9', '9',
'1', '9', '5', '0', '8', '9', '5', '1', '3', '3', '9', '9', '5', '7', '1', '0', '1', '8', '0', '8',
'0', '5', '8', '1', '7', '0', '5', '8', '9', '7', '0', '1', '1', '1', '3', '0', '3', '9', '0', '5',
'1', '9', '7', '1', '1', '0', '9', '1', '8', '7', '5', '3', '0', '5', '7', '5', '5', '7', '9',
'3', '5', '5', '3', '1', '0', '0', '1', '7', '7']
```



Array numero 1:

Max Occurrences (Array 1):

```
{'0': 50, '96': 120, '26': 1, '16': 30, '66': 49, '94': 1, '40': 23, '50': 19, '60': 17, '82': 43,
'34': 6, '4': 24, '52': 17, '74': 33, '36': 1, '6': 31}
```

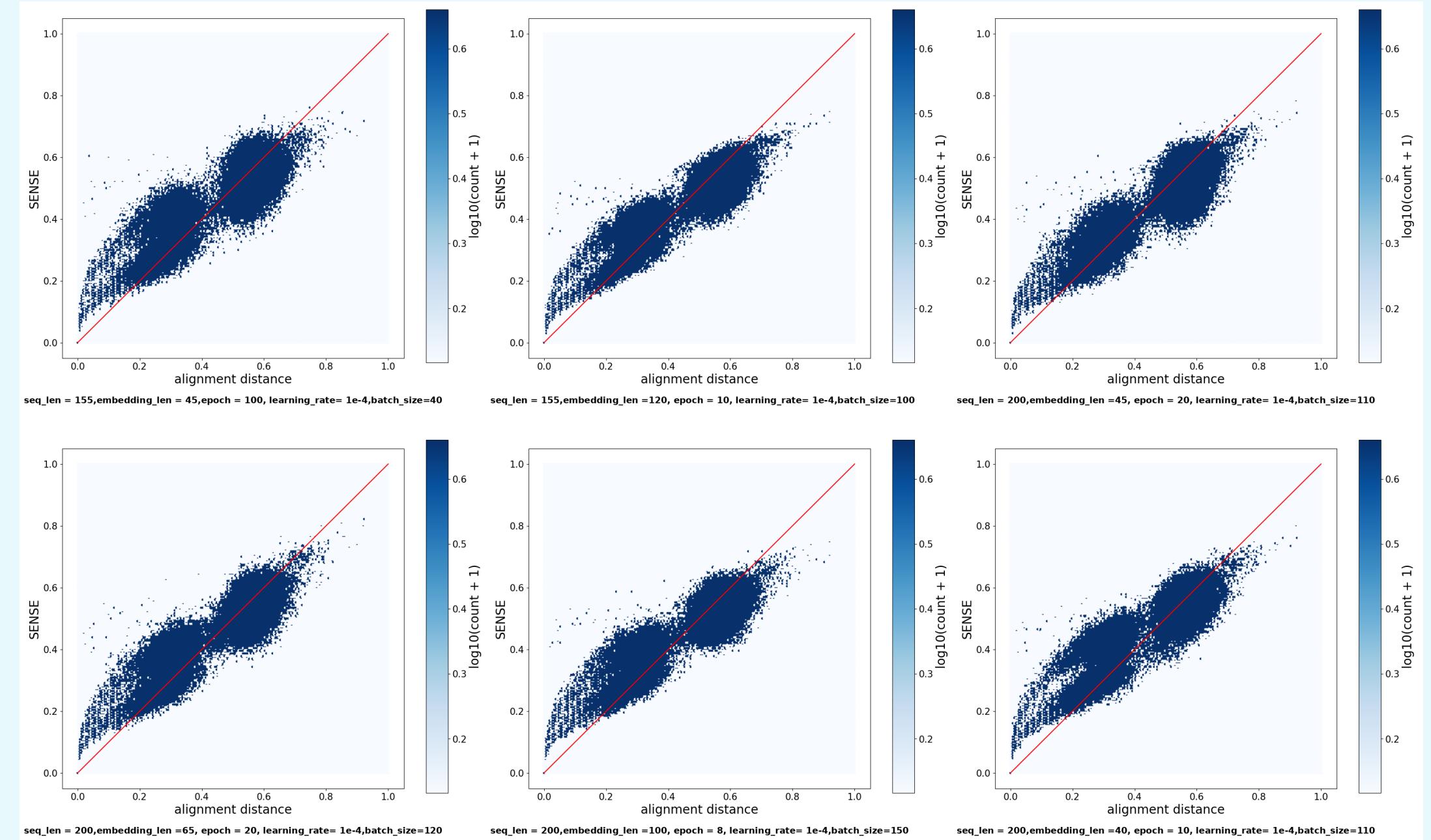
Indice con il valore massimo in Array 1: 96 (120/465 corrisponde al 25.806%)

# Risultati modello

La fase di sperimentazione ha puntato a ottimizzare il modello per allineare i risultati il più possibile alla linea di oracolo, indicatore di massima aderenza ai valori di riferimento.

Con le configurazioni appropriate, il modello mostra un'elevata capacità di emulazione dell'oracolo anche con un dataset complesso e ampio.

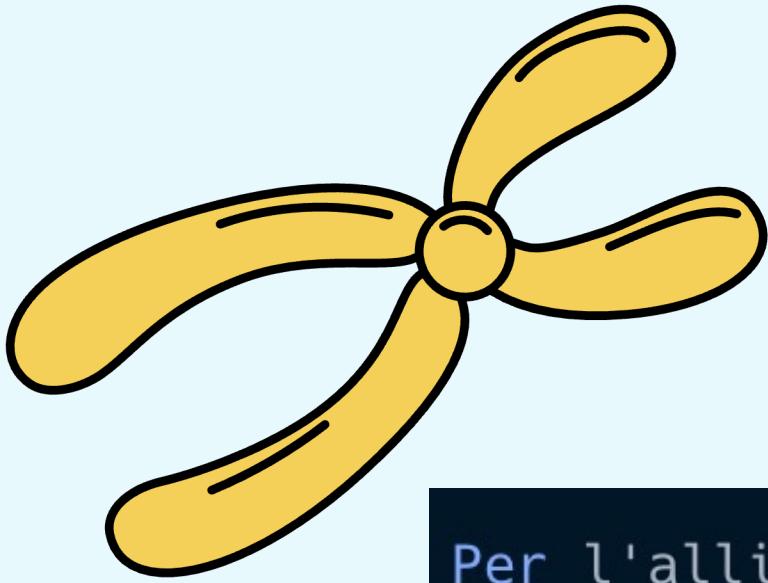
Una distribuzione dei punti vicina alla diagonale rossa indica una forte corrispondenza tra il comportamento del modello e l'allineamento oracolare.



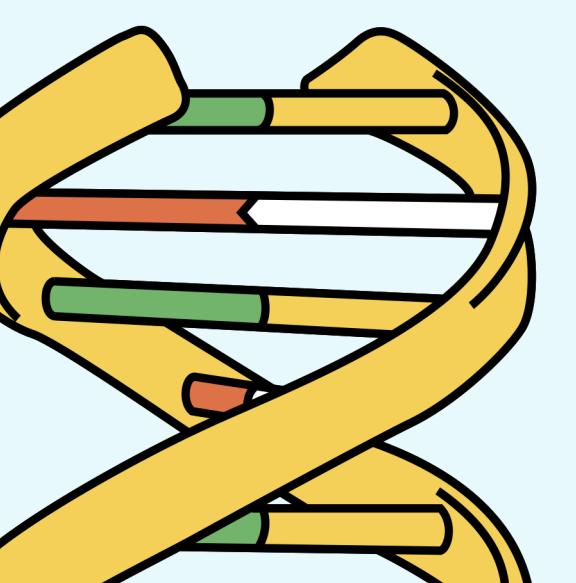
# Risultati Brute force

Per la sequenza 1:  
Miglior punteggio di allineamento: 0.46448516845703125  
Sequenza corrispondente:  
**GAGTCTTGCAGAACCAGCACTTGCCGTTGGGTGAACTGGACATAGGAACCTCGTATGGAAGTTCTCAGGTGATGATGATCATTTATGCTTCAACGTTGGAC**  
GCCCAAGCTGCGACATACTGTTGGTAACCCGGAGGTATTGGTT  
Numero di sequenza corrispondente: 3 su 10 (percentuale: 30.00%)  
Per la sequenza 2:  
Miglior punteggio di allineamento: 0.42194414138793945  
Sequenza corrispondente:  
**CCTTGTTATGGGAGGCAACTGCCGTTCTGAGAGGGTCGCCCCGTCTCATGACACAAACCCCGCAGGTCGACTTG** Per la sequenza 6:  
Miglior punteggio di allineamento: 0.443720281124115  
Sequenza corrispondente:  
**CCTTGTTATGGGAGGCAACTGCCGTTCTGAGAGGGTCGCCCCGTCTCATGACACAAACCCCGCAGGTCGACTTGCTAGGAACGAGCCATTGCTA**  
TTAATGCAAGGCCGCCACGCCGACAGTGATATGGATCAGTGAA  
Numero di sequenza corrispondente: 4 su 10 (percentuale: 40.00%)  
Per la sequenza 3:  
Miglior punteggio di allineamento: 0.46448516845703125  
Sequenza corrispondente:  
**GCGGGGAATCCCTAACGTTCCGAAATACCAAACCTGGATCACATGGTACCGAGGTGCGACAAACTGGCCAATTGTT**  
ACAATTGCCCATTGGCGGACAGTACTGAACTTAGCGCCTAACATCA  
Numero di sequenza corrispondente: 1 su 10 (percentuale: 10.00%)  
Per la sequenza 4:  
Miglior punteggio di allineamento: 0.42194414138793945  
Sequenza corrispondente:  
**CAGCACCGCCTAGTGAATCCGTTCCGAGGCGCAGATAGTCTAATGTGCTCCACTGCGCCTCTCGGTGCATGACTGGAAT**  
GAATATGGTCTTGTATTTGGCGCAGGGACATGGTCAGCTTTA  
Numero di sequenza corrispondente: 2 su 10 (percentuale: 20.00%)  
Per la sequenza 5:  
Miglior punteggio di allineamento: 0.4444347321987152  
Sequenza corrispondente:  
**CCTTGTTATGGGAGGCAACTGCCGTTCTGAGAGGGTCGCCCCGTCTCATGACACAAACCCCGCAGGTCGACTTG**  
TTAATGCAAGGCCGCCACGCCGACAGTGATATGGATCAGTGAA  
Numero di sequenza corrispondente: 4 su 10 (percentuale: 40.00%)

Per la sequenza 6:  
Miglior punteggio di allineamento: 0.443720281124115  
Sequenza corrispondente:  
**CCTTGTTATGGGAGGCAACTGCCGTTCTGAGAGGGTCGCCCCGTCTCATGACACAAACCCCGCAGGTCGACTTGCTAGGAACGAGCCATTGCTA**  
TTAATGCAAGGCCGCCACGCCGACAGTGATATGGATCAGTGAA  
Numero di sequenza corrispondente: 4 su 10 (percentuale: 40.00%)  
Per la sequenza 7:  
Miglior punteggio di allineamento: 0.4636295437812805  
Sequenza corrispondente:  
**TAAGTGGCTAAAGATTGTAGATATCTAGCCAAGTGTACACGGCTGTCAGGCCACTCTAGTAGTCGTGGTAGAGTCGGAGAGCACCGGTAGAACCTTAAC**  
ATATGCTGGACGGCGGTGGCTGGAAAGCAGTGGTGTAGGGTCTAGC  
Numero di sequenza corrispondente: 8 su 10 (percentuale: 80.00%)  
Per la sequenza 8:  
Miglior punteggio di allineamento: 0.424669086933136  
Sequenza corrispondente:  
**CAGCACCGCCTAGTGAATCCGTTCCGAGGCGCAGATAGTCTAATGTGCTCCACTGCGCCTCTCGGTGCATGACTGGAATAGTGTAGCTGCTAGTTGTCG**  
GAATATGGTCTTGTATTTGGCGCAGGGACATGGTCAGCTTTA  
Numero di sequenza corrispondente: 2 su 10 (percentuale: 20.00%)  
Per la sequenza 9:  
Miglior punteggio di allineamento: 0.43801960349082947  
Sequenza corrispondente:  
**CAGCACCGCCTAGTGAATCCGTTCCGAGGCGCAGATAGTCTAATGTGCTCCACTGCGCCTCTCGGTGCATGACTGGAATAGTGTAGCTGCTAGTTGTCG**  
GAATATGGTCTTGTATTTGGCGCAGGGACATGGTCAGCTTTA  
Numero di sequenza corrispondente: 2 su 10 (percentuale: 20.00%)  
Per la sequenza 10:  
Miglior punteggio di allineamento: 0.4360352158546448  
Sequenza corrispondente:  
**TAAGTGGCTAAAGATTGTAGATATCTAGCCAAGTGTACACGGCTGTCAGGCCACTCTAGTAGTCGTGGTAGAGTCGGAGAGCACCGGTAGAACCTTAAC**  
ATATGCTGGACGGCGGTGGCTGGAAAGCAGTGGTGTAGGGTCTAGC  
Numero di sequenza corrispondente: 8 su 10 (percentuale: 80.00%)



## Risultati Algoritmo genetico



Per l'allineamento 1-3 sono state trovate 18 corrispondenze in 14 array diversi.  
Percentuale di corrispondenza: 11.61%. Percentuale di array: **70.00%**.

Per l'allineamento 2-4 sono state trovate 69 corrispondenze in 18 array diversi.  
Percentuale di corrispondenza: 44.52%. Percentuale di array: **90.00%**.

Per l'allineamento 3-1 sono state trovate 0 corrispondenze in 0 array diversi.  
Percentuale di corrispondenza: 0.00%. Percentuale di array: **0.00%**.

Per l'allineamento 4-2 sono state trovate 21 corrispondenze in 16 array diversi.  
Percentuale di corrispondenza: 13.55%. Percentuale di array: **80.00%**.

Per l'allineamento 5-4 sono state trovate 0 corrispondenze in 0 array diversi.  
Percentuale di corrispondenza: 0.00%. Percentuale di array: **0.00%**.

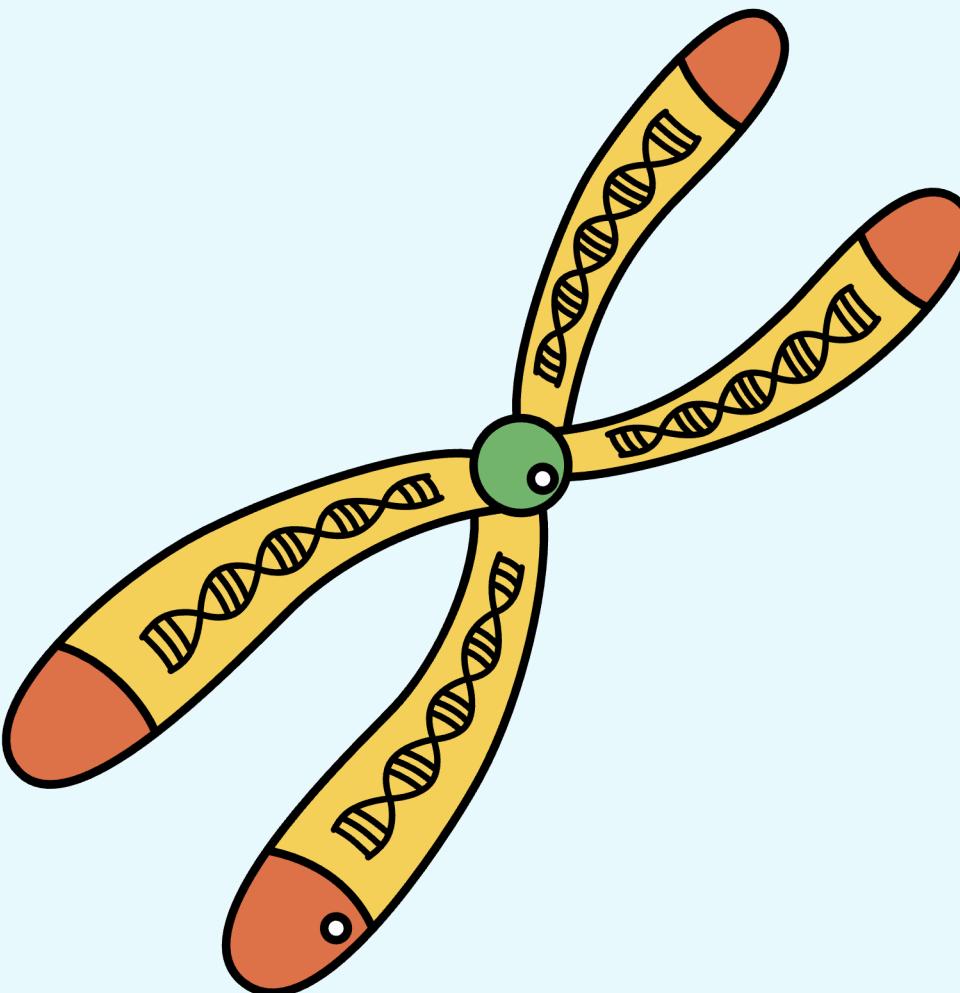
Per l'allineamento 6-4 sono state trovate 70 corrispondenze in 20 array diversi.  
Percentuale di corrispondenza: 45.16%. Percentuale di array: **100.00%**.

Per l'allineamento 7-8 sono state trovate 0 corrispondenze in 0 array diversi.  
Percentuale di corrispondenza: 0.00%. Percentuale di array: **0.00%**.

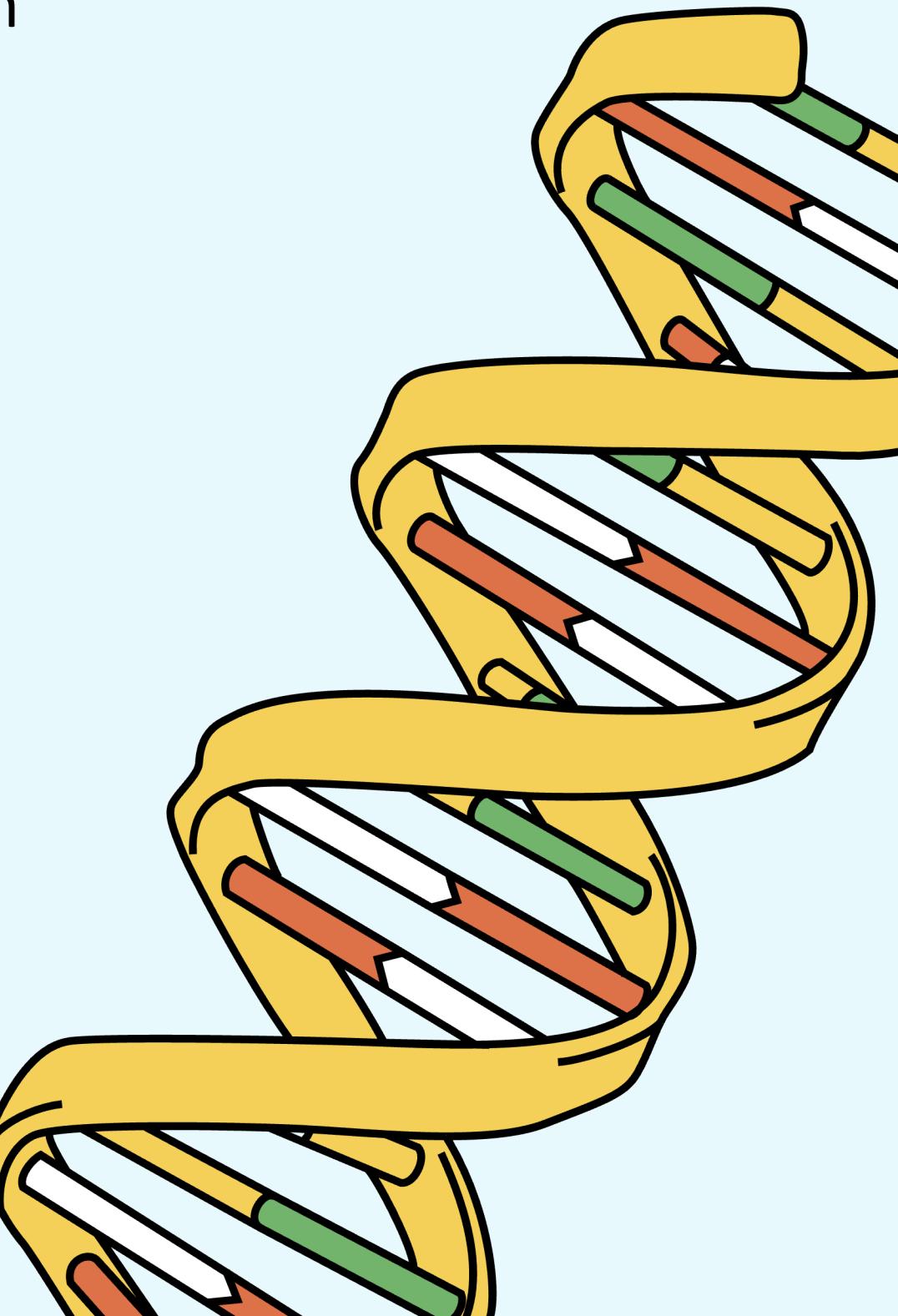
Per l'allineamento 8-2 sono state trovate 28 corrispondenze in 14 array diversi.  
Percentuale di corrispondenza: 18.06%. Percentuale di array: **70.00%**.

Per l'allineamento 9-2 sono state trovate 15 corrispondenze in 12 array diversi.  
Percentuale di corrispondenza: 9.68%. Percentuale di array: **60.00%**.

Per l'allineamento 10-8 sono state trovate 41 corrispondenze in 17 array diversi.  
Percentuale di corrispondenza: 26.45%. Percentuale di array: **85.00%**.

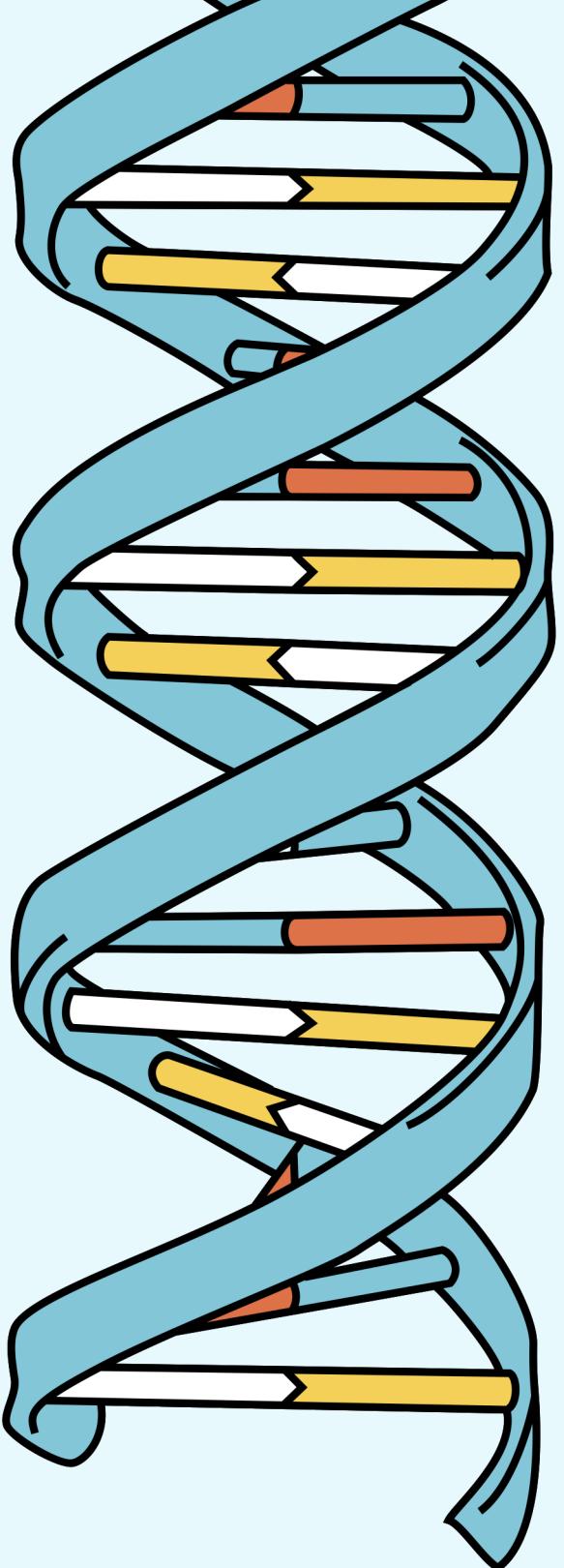


## Conclusioni



la sperimentazione proposta è stata effettuata su un campione piccolo di sequenze (10) e su un numero elevato di generazioni (1000). I risultati ottenuti sono soddisfacenti, dato che raggiungiamo l'70% degli accoppiamenti individuati, nonostante il campione utilizzato.





# Grazie per l'attenzione

- Mattia d'Argenio, matricola:0522501524
- Davide Di Sarno, matricola:0522501490

