Le simmetrie del DNA

Piano Nazionale Lauree Scientifiche.
Progetto Math-en-Jeans.
Dipartimento di Matematica
Università di Padova.

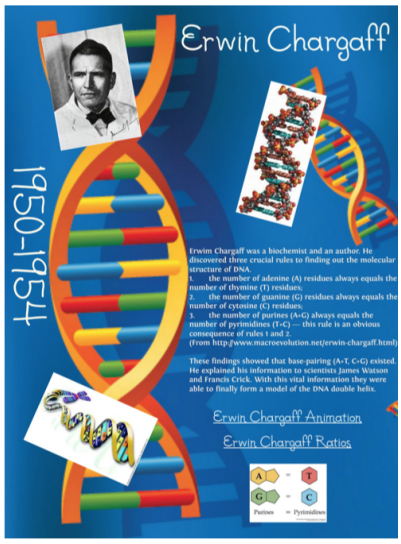
Mario Puppi, Valentina Novello

gennaio 2019

1. Introduzione

1.1. Abstract I Biologi hanno scoperto delle regolarità nelle sequenze del DNA e vorrebbero spiegare come queste emergano e si conservino nel processo dell'Evoluzione.

- Strutture regolari nel DNA, a diverse scale (ordine di grandezza delle sequenze genetiche), suggeriscono che possa esistere un'organizzazione complessa nel genoma.
- La simmetria è lo strumento matematico con cui possiamo osservare e interpretare l'organizzazione in un sistema complesso.
- Le osservazioni sperimentali permetteranno di formulare delle ipotesi di simmetria.
- L'analisi statistica del DNA ci consentirà di formulare un semplice modello matematico con cui potremo verificare le ipotesi di simmetria e spiegare l'esistenza delle simmetrie.

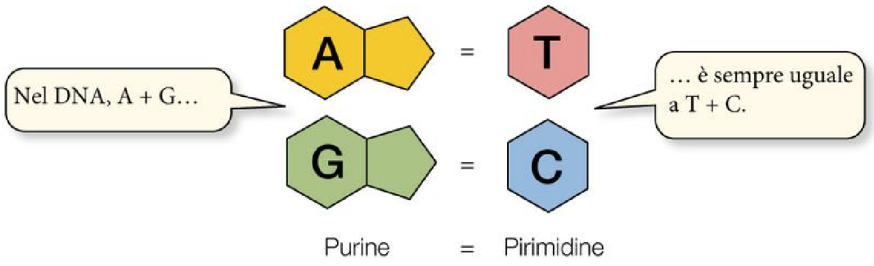


Chargaff's Rules

• Erwin Chargaff showed that the percentages of guanine and cytosine in DNA and adenine and thymine are almost equal.
•A=T (2 hydrogen bonds hold them together)
•G≡C (3 hydrogen bonds hold them together)

Percentages of Bases in Four Organisms				
Source of DNA	A	T	G	C
Streptococcus	29.8	31.6	20.5	18.0
Yeast	31.3	32.9	18.7	17.1
Herring	27.8	27.5	22.2	22.6
Human	30.9	29.4	19.9	19.8

1.2. La 1° legge di Chargaff Nel 1950, il chimico di origine austriaca Erwin Chargaff scopre delle regolarità nel DNA. Riguardano le concentrazioni dei nucleotidi A, C, T, G nel DNA di cellule di uno stesso individuo. La prima regola di parità di Chargaff verifica l'esistenza di un rapporto 1:1 tra la quantità di basi puriniche A+G e quella di basi pirimidiniche T+C contenute in una moldecola di DNA di una cellula.



Il rapporto tra la percentuale di A e quella di G varia da una specie all'altra ma è invariante per organismi della stessa specie.

1.3. La 2° legge di Chargaff Un'osservazione sperimentale, statistica, ben nota è la simmetria "Seconda regola di parità di Chargaff" che sembra essere condivisa dalla quasi totalità dei genomi di organismi esistenti. Su un intervallo del genoma la frequenza di un nucleotide (A, C, T oppure G) è approssimativamente uguale alla frequenza del suo complementare. Questa legge sperimentale è stata estesa, in seguito, a oligonucleotidi di lunghezza

fino a 10, usando la nozione di inversa del complementare (di un oligonucleotide)

1.4. Spiegazioni delle due leggi di Chargaff La prima legge di Chargaff riguarda la doppia elica del DNA ed è stata usata come prova per indirizzare alla scoperta della struttura ad elica del DNA, della quale è ora una banale conseguenza.

La seconda legge di Chargaff rimane di origine misteriosa. Non si capisce il suo ruolo funzionale. I biologi hanno proposto differenti meccanismi nel tentativo di spiegare la sua origine.

1.5. Obiettivi della ricerca. Studieremo alcuni risultati noti sulle simmetrie statistiche del DNA umano, in particolare: (1) l'estensione della legge di parità di Chargaff a brevi sequenze di oligonucleotidi (2) l'esistenza di una gerarchia di simmetrie del DNA, a differenti scale.

Formuleremo un modello per spiegare le osservazioni sperimentali.

L'ingrediente chiave del nostro modello sarà la simmetria *reverse-complement*.

2. Osservabili e simmetrie.

2.1. Genoma umano nel Wolfram Language. Nel Wolfram Language le sequenze del genoma umano sono stringhe di testo con le sole 4 lettere A, C, G, T (nucleotidi) I cromosomi umani sono specificati con un numero intero (da 1 a 22) oppure con le lettere "X", "Y"

Esempio 1. Definiamo la sequenza iniziale L del cromosoma 9, dal nucleotide in posizione 1 al nucleotide nella posizione 10^7:

```
L = GenomeData[{"9", {1, 10^7}}]
```

```
TAACCCTAACCCTAACCCTAACCCAACCCCACCCCAACCCCAACCCCAACCCA
    ACCCTAACCCTAACCCTAACCCAACCCTAACCCTAACCCTAACCCAAC
    CCTCACCCTCACC ...
 AGACCTCTGTTTTAATGGTAATGCTGGTCAGTTGTGCCTGAATTCCAAAGGG
    AGGAGAGTATAATGAAGCATATCCAATTCTCCCTTCCCATCATGGCCT
    GAATTAGCTTTTCAG
```

large output | show less | show more | show all | set size limit...

L'espressione

```
StringCount[L, "A"]
```

dà la *frequenza assoluta* del carattere **A** nella stringa **L**

Esercizio 1. Verificare la 2° Legge di Chargaff sulla sequenza L.

2.2. Frequenza di un osservabile. Esploriamo proprietà statistiche di sequenze genetiche $s = s_1 \dots s_N$ dove $s_i \in \{A, C, T, G\}$

quantificando la frequenza in s di un pattern (od osservabile) X . Vediamo due esempi di problemi relativi a frequenze di pattern.

Esempio 1. Studiare la frequenza del codone ACT in un intervallo genetico s del genoma dato.

La *lunghezza* di una stringa **L** è data dall'epressione
`StringLength[L]`

La forma decimale di una frazione **f** è data dall'epressione
`N@f`

Esercizio 2. Definire la frequenza assoluta e relativa di una parola w nella sequenza genetica L.

Esempio 2. Misurare la frequenza dell'osservabile $X = (AC\_G)$, nella sequenza genetica $s = GGACCGGCCACAGGAA$.

La sequenza s ha lunghezza 16, il numero dei suoi caratteri (nucleotidi), mentre la *lunghezza* dell'osservabile X è $l_X = 4$.

Ci sono 2 occorrenze dell'osservabile X tra le 13 sottosequenze di s di lunghezza 4:

$$s = GGACCGGCCACAGGAA$$

Si definisce dunque la frequenza $P(X) = \frac{2}{13}$.

Esercizio 3. Definire la frequenza assoluta e relativa dell'osservabile $X = (AC\_G)$ nella sequenza genetica iniziale L del cromosoma 1.

2.2. Simmetria di un osservabile.

Esempio 3. Consideriamo la simmetria *reverse-complement* S. Per l'osservabile $X = (AC\_G)$ il pattern simmetrico $S(X)$ è $X = (C\_GT)$.

Diciamo che una sequenza genetica s ha la simmetria S alla scala l se per ogni osservabile X di lunghezza $l_X = l$ si ha $P(X) = P(S(X))$, dove $S(X)$ è l'osservabile simmetrico di X.

Problema 1. Sia S la simmetria *reverse-complement*. Studiare la validità della legge $P(X) = P(S(X))$, a scale diverse, per pattern di grandezza l_X variabile fino all'ordine di milioni di basi.

La validità della legge $P(X) = P(S(X))$ è stata confermata ampiamente in letteratura per oligonucleotidi di lunghezza $l_X \leq 10$.

Per il cromosoma umano 1, ad esempio, sono state misurate le frequenze:

$$P(AG) = 7.14\% \approx P(CT) = 7.13\%$$

$$P(GA) = 6.01\% \approx P(TC) = 6.01\%$$

Osserviamo che la validità della legge P(X) = P(S(X)) per scale piccole, come ad esempio $l_X \leq 2$, non è abbastanza significativa per credere che la legge sia valida a scale maggiori come $l_X \gg 100$