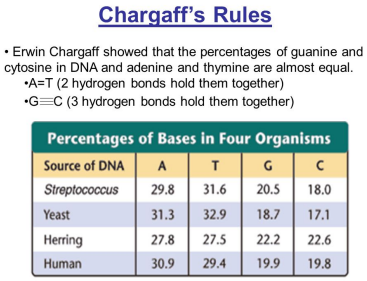
Le simmetrie del DNA.

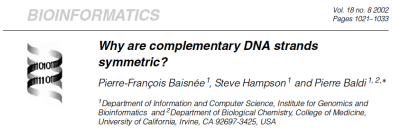
MARIO PUPPI, VALENTINA NOVELLO.

1. Misurare la simmetria.



Esistono diversi approcci per misurare la simmetria *reverse-complement* del DNA. Per lo più fanno uso di strumenti della statistica. Essi sono stati messi a punto da vari gruppi di ricerca che si sono dedicati alla valutazione quantitativa del fenomeno osservato della simmetria.
Comune a tutti gli studi sono i dati osservati ottenuti misurando la frequenza degli oligonucleotidi (parole del genoma) di lunghezza data. Ci concentreremo quindi su alcuni approcci e strumenti utilizzati nella ricerca, derivata per lo più dalla statistica, allo scopo di misurare la simmetria del genoma.

2. L'indice di Simmetria S^1



Uno dei primi lavori di ricerca sulla misura della simmetria è quello di Baisnée, Hampson e Baldi, apparso sulla rivista Bioinformatics nel 2002. In questo articolo vengono analizzate sequenze del genoma di alcuni organismi studiando le frequenze di tutte le 4^k parole genetiche di lunghezza $k = 1, \ldots, 9$.
Dai dati empirici (un lungo tratto di genoma), si ricavano le frequenze delle parole genetiche di lunghezza k, quindi si misura la *simmetria di ordine k* usando il seguente *indicatore di simmetria*:

$$S^1 = 1 - \frac{\sum_i |f_i - f_i^*|}{\sum_i (f_i + f_i^*)}$$

f_i, f_i^* rappresentano le frequenze relative dell'oligonucleotide ω_i e del suo reverse-complement ω_i^* rispettivamente.
Osserviamo che il denominatore vale 2 se la somma è estesa a tutti i 4^k oligomeri di lunghezza k. per un certo k fissato.
L'indice S^1 generalizza le misure classiche date dagli indicatori $\frac{A-T}{A+T}, \frac{G-C}{G+C}$ introdotti per misurare la validità della Seconda Legge di Chargaff.
Per studiare la simmetria di un dato ordine k vengono contate le frequenze di tutti i 4^k oligonucleotidi di lunghezza k su un tratto lineare del genoma. Otteniamo così una distribuzione della forma $f(x_1x_2 \ldots x_k)$.
Diciamo che si ha una simmetria perfetta quando $f(x_1x_2 \ldots x_k) = f(\hat{x}_k\hat{x}_{k-1} \ldots \hat{x}_1)$
Poichè la grandezza 4^k della distribuzione cresce in modo esponenziale con l'ordine k, occorrono insiemi di dati molto grandi perchè un numero significativo di oligonucleotidi sia presente ad un certo ordine k e si possano fare delle stime accurate della frequenza.

3. Come si comporta l'indice di simmetria?

	AA	TT	AC	GT	AG	CT	CA	TG	CC	GG	GA	TC	S^1
Chr. 1	10.4	10.5	5.4	5.6	5.9	5.7	6.6	6.8	4.0	4.1	6.3	6.1	98.75
Chr. 2	10.7	10.9	5.3	5.2	5.8	5.9	6.6	6.4	4.0	3.9	6.1	6.3	98.93
Chr. 3	10.9	10.3	5.6	5.2	5.7	5.8	6.8	6.3	4.1	3.8	6.1	6.3	97.33
Chr. 4	10.9	10.9	5.2	5.2	5.9	5.8	6.5	6.5	3.8	3.8	6.3	6.2	99.61
Chr. 5	10.6	10.8	5.2	5.4	5.8	5.8	6.4	6.6	3.9	4.0	6.2	6.1	98.86
Chr. 6	10.7	10.6	5.3	5.3	5.8	5.9	6.4	6.6	4.0	4.0	6.3	6.2	99.37
Chr. 7	10.9	10.9	5.3	5.2	5.8	5.8	6.5	6.4	3.9	3.8	6.2	6.2	99.63
Chr. 8	10.9	10.6	5.4	5.3	5.8	5.8	6.6	6.5	4.0	3.9	6.2	6.2	99.04
Chr. 9	10.6	10.6	5.3	5.4	5.9	5.9	6.5	6.5	4.0	4.0	6.2	6.2	99.80
Chr. 10	10.9	10.6	5.3	5.3	5.9	5.8	6.5	6.5	3.8	3.9	6.3	6.2	99.17
Chr. 11	10.9	10.9	5.2	5.2	5.8	5.9	6.5	6.4	3.9	3.8	6.2	6.3	99.45
Chr. 12	10.7	10.8	5.2	5.3	5.9	5.9	6.5	6.5	4.0	3.9	6.3	6.3	99.55
Chr. 13	10.9	10.8	5.3	5.3	5.8	5.8	6.5	6.4	3.9	3.9	6.2	6.2	99.68
Chr. 14	10.7	10.6	5.4	5.3	5.9	5.8	6.5	6.5	4.0	3.9	6.3	6.2	99.52
Chr. 15	11.0	10.8	5.3	5.2	5.8	5.8	6.5	6.4	3.9	3.9	6.2	6.2	99.26
Chr. 16	10.9	10.9	5.2	5.2	5.9	5.9	6.4	6.4	3.9	3.8	6.3	6.2	99.79

L'indice S^1 varia da 0 (assenza di simmetria) a 1 (simmetria perfetta). Esso calcola la percentuale di oligonucleotidi di una data lunghezza k che hanno la stessa frequenza del loro reverse-complement. Si verifica che S^1 è funzione decrescente della lunghezza k degli oligonucleotidi. Infatti, la frequenza $P(x) = P(x_1x_2 \dots x_k)$ di un oligonucleotide di lunghezza k è legata alla frequenza deigli oligonucleotidi di lunghezza $k+1$ dalla relazione:

$$P(x) = P(x_1x_2 \dots x_k) = \sum_Y P(x_1x_2 \dots x_kY)$$

con la somma estesa all'alfabeto genetico $Y \in \{A, C, T, G\}$. Per la disuguaglianza triangolare, ne segue che

$$\sum_x P(x) - P(x^*) = \sum_x (P(xA) + P(xC) + P(xT) + P(xG)) - (P(Tx^*) + P(Gx^*) + P(Ax^*) + P(Cx^*))$$

cresce passando da k a $k+1$.

4. Modello di Markov del genoma.

	AA	TT	AC	GT	AG	CT	CA	TG	CC	GG	GA	TC	S^1
Chr. 1	10.4	10.5	5.4	5.6	5.9	5.7	6.6	6.8	4.0	4.1	6.3	6.1	98.75
Chr. 2	10.7	10.9	5.3	5.2	5.8	5.9	6.6	6.4	4.0	3.9	6.1	6.3	98.93
Chr. 3	10.9	10.3	5.6	5.2	5.7	5.8	6.8	6.3	4.1	3.8	6.1	6.3	97.33
Chr. 4	10.9	10.9	5.2	5.2	5.9	5.8	6.5	6.5	3.8	3.8	6.3	6.2	99.61
Chr. 5	10.6	10.8	5.2	5.4	5.8	5.8	6.4	6.6	3.9	4.0	6.2	6.1	98.86
Chr. 6	10.7	10.6	5.3	5.3	5.8	5.9	6.4	6.6	4.0	4.0	6.3	6.2	99.37
Chr. 7	10.9	10.9	5.3	5.2	5.8	5.8	6.5	6.4	3.9	3.8	6.2	6.2	99.63
Chr. 8	10.9	10.6	5.4	5.3	5.8	5.8	6.6	6.5	4.0	3.9	6.2	6.2	99.04
Chr. 9	10.6	10.6	5.3	5.4	5.9	5.9	6.5	6.5	4.0	4.0	6.2	6.2	99.80
Chr. 10	10.9	10.6	5.3	5.3	5.9	5.8	6.5	6.5	3.8	3.9	6.3	6.2	99.17
Chr. 11	10.9	10.9	5.2	5.2	5.8	5.9	6.5	6.4	3.9	3.8	6.2	6.3	99.45
Chr. 12	10.7	10.8	5.2	5.3	5.9	5.9	6.5	6.5	4.0	3.9	6.3	6.3	99.55
Chr. 13	10.9	10.8	5.3	5.3	5.8	5.8	6.5	6.4	3.9	3.9	6.2	6.2	99.68
Chr. 14	10.7	10.6	5.4	5.3	5.9	5.8	6.5	6.5	4.0	3.9	6.3	6.2	99.52
Chr. 15	11.0	10.8	5.3	5.2	5.8	5.8	6.5	6.4	3.9	3.9	6.2	6.2	99.26
Chr. 16	10.9	10.9	5.2	5.2	5.9	5.9	6.4	6.4	3.9	3.8	6.3	6.2	99.79

Assumiamo l'ipotesi che le parole di lunghezza k del DNA siano un modello di Markov. Esso è determinato

- dai valori delle probabilità di transizione: $P(x_k|x_1x_2 \dots x_{k-1}) = \frac{P(x_1x_2 \dots x_k)}{P(x_1x_2 \dots x_{k-1})}$
- dalla distribuzione di frequenze iniziali $f(x_1x_2 \dots x_k)$

Poichè le dimensioni crescono in modo esponenziale si può usare questi modelli solo per valori dell'oridne k piccoli.