

Sviluppi Teorici e Applicativi delle Metriche Entropiche di Rohlin

Dawid Crivelli

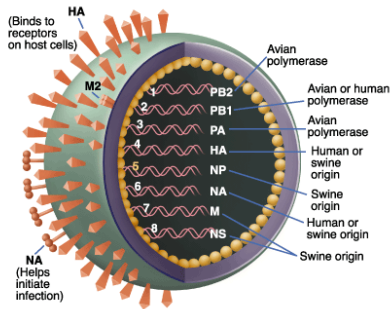
26 Aprile 2012

Sommario

① Distanze Entropiche

Proteine dell'influenza H3N2

- proteine come stringhe
- approccio *black box*
- sequenze lunghe 566
- alfabeto di 24 lettere
- solo 10% mutazioni
- **antigenic drift**



Sequenze a confronto:

```

PGNDNSMATLCLGHHAVPNGTLVKTIITNDQIEVTNATELVQSSSTGRICDSPHQILDGENCTLIDALLGDPHCDGFO
PGNDNSTATLCLGHHAVPNGTLVKTIITNDQIEVTNATELVQSSSTDRIICDSPHQILDGGNCTLIDALLGDPHCDGFO
PGNDNSTATLCLGHHAVPNGTIVKTIITNDQIEVTNATELVQSSSTGGICDSPHQILDGENCTLIDALLGDPQCDGFO
PGNDNSTATLCLGHHAVPNGTIVKTIITNDQIEVTNATELVQSSSTGGICDSPHQILDGENCTLIDALLGDPQCDGFO
PGNDNSTATLCLGHHAVPNGTLVKTIITNDQIEVTNATELVQSSSTDKICDSPHQILDGGNCTLIDALLGDPHCDGFO
PGNDNSTATLCLGHHAVPNGTIVKTIITNDQIEVTNATELVQSSSTGGICDSPHQILDGENCTLIDALLGDPQCDGFO
PGNDNSTATLCLGHHAVPNGTLVKTIITNDQIEVTNATELVQSSSTGRICDSPHQILDGENCTLIDALLGDPHCDGFO

```

Hamming è poco adatto

A={GHHAVPNGTLVKTITTGRICGDPHCDGFQNKWE}

B={GHHAVPNGTIVKTITTGEICGDPQCDGFQNKKW}

$d_H(A, B) = \# \text{differenze}$

Hamming è poco adatto

 $A = \{GHHAVPNGT\textcolor{red}{L}VKTITTG\textcolor{red}{R}ICGDP\textcolor{red}{H}CDGFQNK\textcolor{red}{E}W\}$ $B = \{GHHAVPNGT\textcolor{red}{I}VKTITTG\textcolor{red}{E}ICGDP\textcolor{red}{Q}CDGFQNK\textcolor{red}{K}W\}$ $d_H(A, B) = \text{\#differenze}$ $d_H(A, B) = 4$

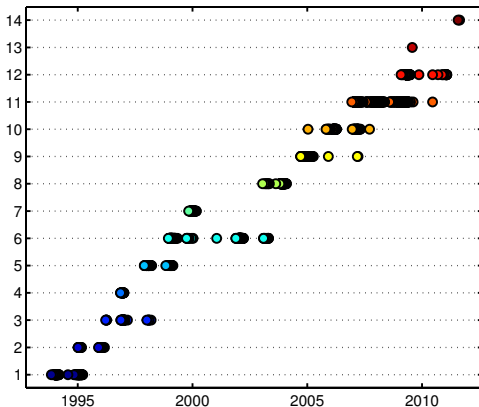
Hamming è poco adatto

A={GHHAVPNGT**L**VKTITT**G**RICGDP**H**CDGFQNK**E**W}

B={GHHAVPNGT**I**VKTITT**G**EICGDP**Q**CDGFQNK**K**W}

$d_H(A, B) = \text{\#differenze}$

$d_H(A, B) = 4$



Antigenic drift

$d_H \propto t$

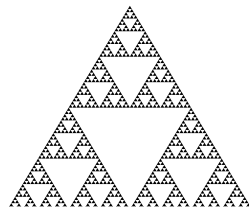
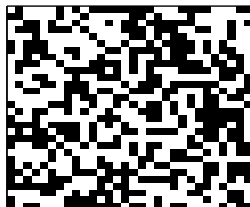
Distanza di Rohlin

Distanza non tra configurazioni, ma tra **partizioni**

Requisiti:

- uno spazio di probabilità: $(\mathbf{M}, \sigma, \mu)$
- un criterio per partizionare (relazione di equivalenza)
- usiamo \mathbf{M} discreto, μ è banale

Applicabile a molte strutture diverse:



Complessità di una partizione

Partizione \iff scomposizione in **atomi** disgiunti di *misura* $\mu(A_k)$

Rappresentazione associando ad ogni sito un'etichetta (atomo):

$$A = \{ \underbrace{(1, 2, 3, 4)}_{A_1}, \underbrace{(5, 6)}_{A_2}, \underbrace{(7, 8, 9)}_{A_3}, \underbrace{(10, 11, 12, 13)}_{A_4} \}$$

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 \\ \color{red}{1} & \color{red}{1} & \color{red}{1} & \color{red}{1} & \color{green}{2} & \color{green}{2} & \color{blue}{3} & \color{blue}{3} & \color{blue}{3} & \color{orange}{4} & \color{orange}{4} & \color{orange}{4} & \color{orange}{4} \end{bmatrix}$$

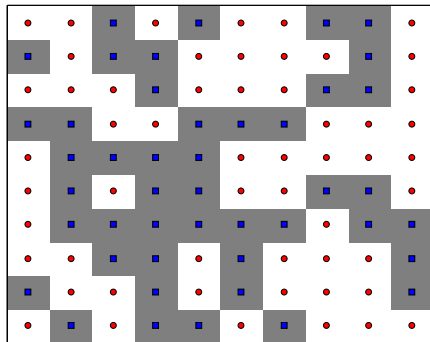
Entropia di Shannon: misura della complessità di una partizione

$$H(A) = \sum_k^n \mu(A_k) \log(\mu(A_k))$$

$H = \log(n)$ (max) \iff partizione con n atomi equivalenti
 $H = 0$ (min) \iff partizione banale ν

Partizionamento

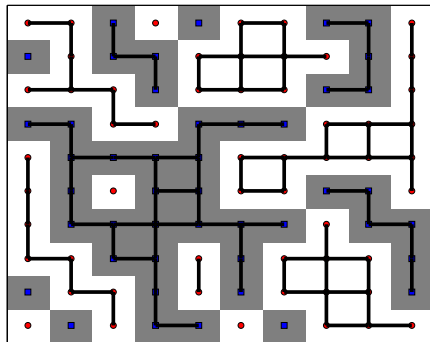
un partizione è una relazione di equivalenza, $i \sim j \iff i, j \in A_k$



relazione locale(trai vicini) \Rightarrow partizione globale
 \Rightarrow colorazione di grafi, algoritmo Hoshen-Kopelman $\mathcal{O}(N \log(N))$

Partizionamento

un partizione è una relazione di equivalenza, $i \sim j \iff i, j \in A_k$



relazione locale(tra vicini) \Rightarrow partizione globale
 \Rightarrow colorazione di grafi, algoritmo Hoshen-Kopelman $\mathcal{O}(N \log(N))$

Prodotti tra partizioni

Partizione prodotto $\gamma = \alpha \vee \beta$








- proprietà associativa
- elemento neutro ν
- ogni partizione è scrivibile come prodotto
- idempotente

1	1	3	1		1	1	1	1				
2	1	3	1		2	2	3	1				
2	1	1	1		2	3	3	3				
2	2	2	1		2	2	3	1				

$$\alpha \vee \alpha = \alpha$$

- l'entropia del prodotto è sempre maggiore

$$H(\alpha \vee \beta) \geq H(\alpha), \forall \beta$$

(1,1)	(1,1)	(3,1)	(1,1)	(1,1)	
(2,2)	(1,2)	(3,3)	(1,1)	(1,2)	
(2,2)	(1,3)	(1,3)	(1,3)	(1,3)	
(2,2)	(2,2)	(2,3)	(1,1)	(2,2)	
				(2,3)	
				(3,1)	
				(3,3)	

Distanza di Rohlin

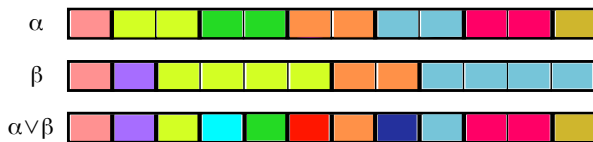
Distanza tra partizioni, tramite l'entropia del prodotto:

$$d_R(\alpha, \beta) = 2 H(\alpha \vee \beta) - H(\alpha) - H(\beta)$$

Partizioni simili hanno piccola distanza:



Cosa fare per partizioni molto diverse e frammentate?



Intersezione tra partizioni

Definiamo $\sigma = \alpha \wedge \beta$, la partizione **comune**

Riduzione e amplificazione della distanza

Definizione topologica della distanza