

Sviluppi Teorici e Applicativi delle Metriche Entropiche di Rohlin

Dawid Crivelli

26 Aprile 2012

Distanza di Rohlin

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

Requisiti:

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

Ogni sequenza, reticolo, grafo \Rightarrow array con relazioni non locali



da $\mathcal{C}(\mathbf{M})$ a $\mathcal{Z}(\mathbf{M})$



Complessità di una partizione

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

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

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

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

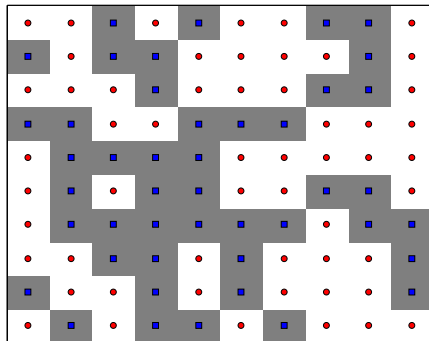
Entropia di Shannon: misura della complessità di una partizione

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

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

Partizionamento

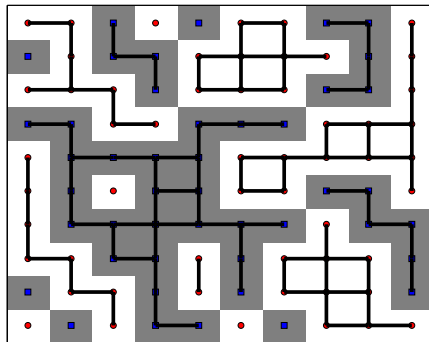
un partizione è una relazione di equivalenza, $i \sim j \iff i, j \in A_k$



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, $i \sim j \iff i, j \in A_k$

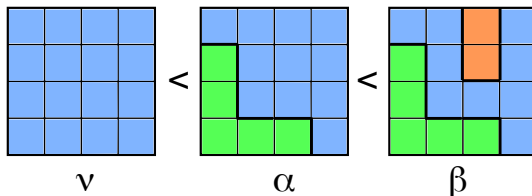


relazione locale(tra vicini) \Rightarrow partizione globale
 \Rightarrow colorazione di grafi, algoritmo Hoshen-Kopelman $\mathcal{O}(N \log(N))$

Ordinamento parziale e fattori

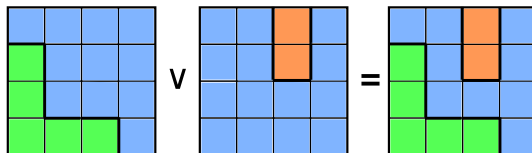
Ordinamento

- α è **fattore** di β
- β è più fine di α
- $H(\alpha) < H(\beta)$



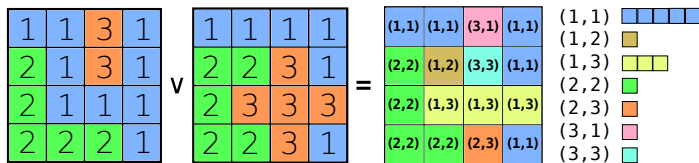
Prodotto

- proprietà associativa
- elemento neutro ν
- ogni partizione è prodotto di fattori
- “**minimo comune multiplo**”

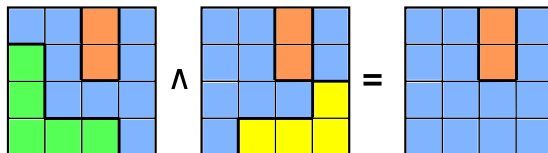


Prodotti tra partizioni

Partizione prodotto $\gamma = \alpha \vee \beta$, più fine: **unione** dei bordi



Partizione intersezione $\sigma = \alpha \wedge \beta$, meno fine: **intersezione** dei bordi



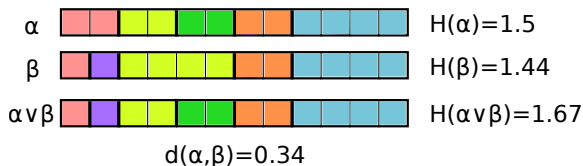
è il “**massimo comune divisore**”

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:



Funziona perché:

- prodotto idempotente $\alpha \vee \alpha = \alpha$
- l'entropia del prodotto è crescente $H(\alpha \vee \beta) \geq H(\alpha), \forall \beta$

Distanza piccola per partizioni estremamente frammentate...

Riduzione e amplificazione della distanza

Ridurre le partizioni: eliminare il più possibile fattori comuni

Definiamo una mappa dalle partizioni alle ridotte

$$\alpha \otimes \beta \xrightarrow{\text{riduzione}} \hat{\alpha}(\alpha, \beta) \otimes \hat{\beta}(\alpha, \beta)$$

Algoritmo

- scomposizione delle due partizioni in fattori
- confronto dei fattori tra le due partizioni
- scelta e scarto
- ricomposizione di ciascuna

La distanza è sempre amplificata:

$$R = \frac{d_R(\hat{\alpha} \otimes \hat{\beta})}{d_R(\alpha \otimes \beta)} \geq 1$$

Confronto tra diverse riduzioni

Fattori lineari

- **solo** partizioni lineari connesse
- ottimale come riduzione
- semplicissimo da implementare

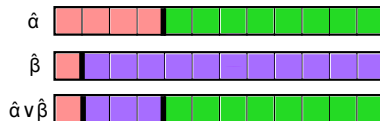
Fattori dicotomici semplici

- ovunque applicabile
- oneroso computazionalmente
- peggiore nel caso lineare

Partizioni non ridotte



Riduzione tramite fattori lineari



Riduzione tramite fattori dicotomici



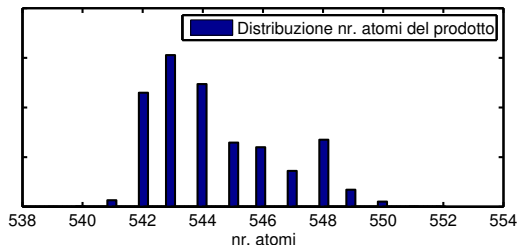
Definizione topologica della distanza

Funzionale entropia indipendente dalla misura μ :

$$H_{\text{top}} = \log(n) \quad n \text{ è il numero di atomi}$$

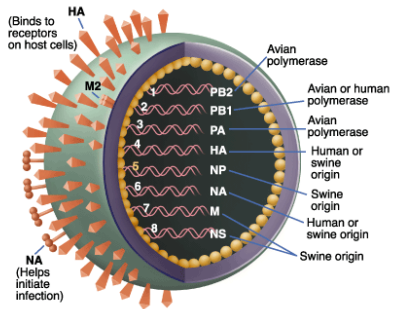
Una distanza di Rohlin opportunamente definita:

$$d_{\text{top}}(\alpha, \beta) = 2 \log(n_{\alpha \vee \beta}) - \log(n_\alpha) - \log(n_\beta)$$



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={GHHAVPNGT**L**VKTITTG**R**ICGDP**H**CDGFQNK**E**W}

B={GHHAVPNGT**I**VKTITTG**E**ICGDP**Q**CDGFQNK**K**W}

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

$d_H(A, B) = 4$

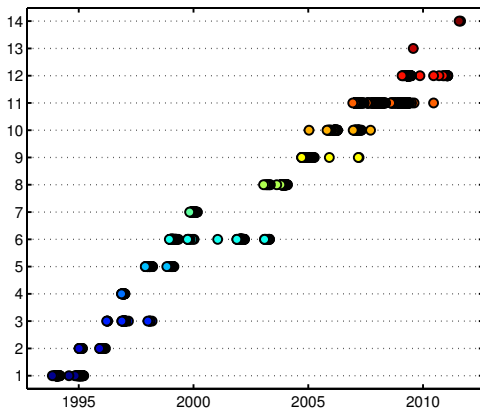
Hamming è poco adatto

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

$d_H \propto t$

Utilizzo misure entropiche

- 1 Selezione N sequenze da un database (FluDB, NCBI) e allineamento
- 2 Partizionamento delle sequenze
- 3 Calcolo matrice d_{ij} delle $N(N - 1)/2$ distanze tra partizioni
- 4 N punti, distanti tra di loro d_{ij} — grafo completo tra le sequenze

Utilizzo misure entropiche

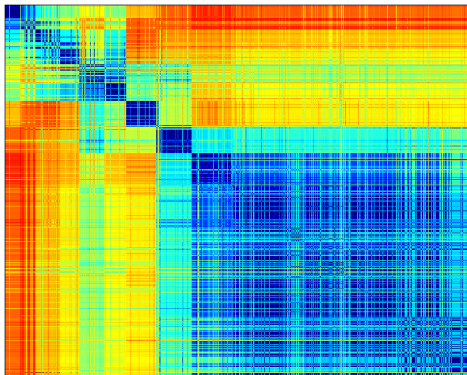
- 1 Selezione N sequenze da un database (FluDB, NCBI) e allineamento
- 2 Partizionamento delle sequenze
- 3 Calcolo matrice d_{ij} delle $N(N - 1)/2$ distanze tra partizioni
- 4 N punti, distanti tra di loro d_{ij} — grafo completo tra le sequenze

Utilizzo misure entropiche

- 1 Selezione N sequenze da un database (FluDB, NCBI) e allineamento
- 2 Partizionamento delle sequenze
- 3 Calcolo matrice d_{ij} delle $N(N - 1)/2$ distanze tra partizioni
- 4 N punti, distanti tra di loro d_{ij} — grafo completo tra le sequenze

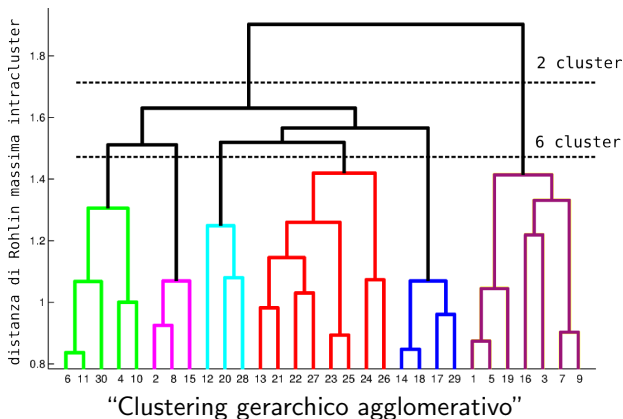
Utilizzo misure entropiche

- 1 Selezione N sequenze da un database (FluDB, NCBI) e allineamento
- 2 Partizionamento delle sequenze
- 3 Calcolo matrice d_{ij} delle $N(N - 1)/2$ distanze tra partizioni
- 4 N punti, distanti tra di loro d_{ij} — grafo completo tra le sequenze



Clustering di sequenze

Suddivisione di N sequenze in p **clusters**



Altri algoritmi: risultati qualitativamente indistinguibili

Distanza e partizionamento ottimale

Partizioni di segmenti omogenei

L A T G M R N V P P E K T R G I F G A I



L A T **M** **R** **C** V P P E K **R** G I F G A I



$$d_R = 0.13$$

Partizioni ridotte

Distanza e partizionamento ottimale

Partizioni di segmenti omogenei

L A T G M R N V P P E K T R G I F G A I
□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □

L A T **M** M R **C** V P P E K **R** R G I F G A I
□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □

$$d_R = 0.13$$

Partizioni ridotte

L A T G M R N V P P E K T R G I F G A I
□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □

L A T **M** M R **C** V P P E K **R** R G I F G A I
□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □

$$d_R = 1.07$$

Distanza e partizionamento ottimale

Partizioni di segmenti omogenei

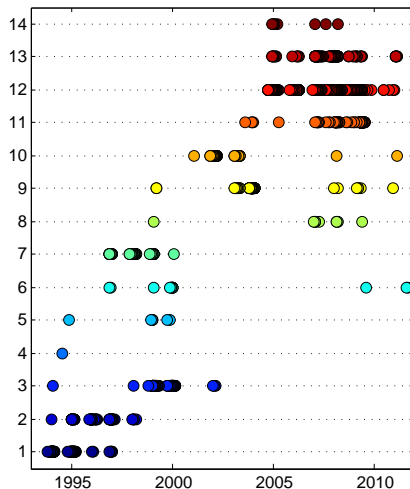


$$d_R = 0.13$$

Partizioni ridotte



$$d_R = 1.07$$



Partizionamento alternativo

Sequenze lineari di spin (Ising 1D)

Per una sequenza aperta di spin, lunga L

$$H = - \sum_i J_{i,i+1} \sigma_i \sigma_{i+1} \quad i \in \{1, \dots, L\}$$

variabili di **link**

$$l_i = \sigma_i \sigma_{i+1} \quad i \in \{1, \dots, L-1\}$$

i gradi di libertà si disaccoppiano:

$$H = - \sum_i J_i l_i$$

è possibile tornare alle variabili originali $l_i = \sigma_i \sigma_{i+1} \implies \sigma_i = l_i \sigma_i$

$$Z_{\text{catena}} = Z_{\text{link}}^L = (2 \cosh \beta J)^L$$

Clusters di spin \Longleftrightarrow Clusters di link

$$\langle \sigma_i \sigma_{i+r} \rangle = \frac{1}{Z} \sum_{\sigma} \sigma_i \sigma_{i+r} \exp \left(\beta \sum_k l_k(\sigma) \right) \quad (1)$$

$$= \exp \left(-\frac{r}{\xi} \right) \quad \text{con } \xi = -\frac{1}{\log(\tanh \beta J)} \quad (2)$$

Lunghezza di correlazione tra partizioni

Variazione in temperatura

Tipi di disordine

Ising 2D, reticolo quadrato