Sviluppi Teorici e Applicativi delle Metriche Entropiche di Rohlin

Dawid Crivelli

26 Aprile 2012

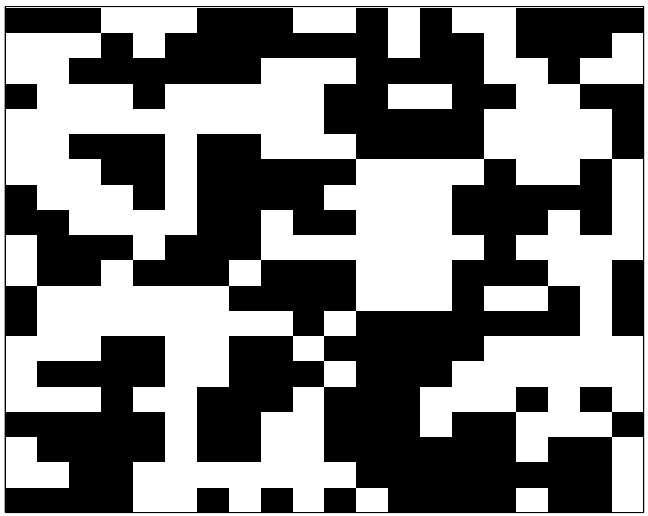
Distanza di Rohlin

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

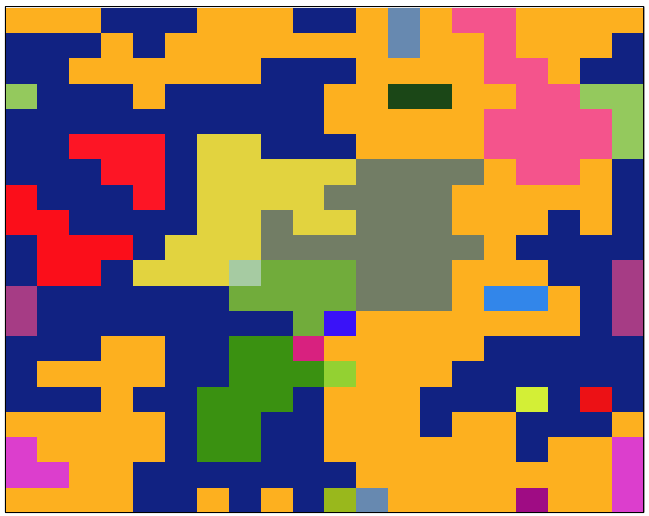
Requisiti:

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

Ogni sequenza, reticolo, grafo $=>$ array con relazioni non locali



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



Complessità di una partizione

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

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

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

$$\mathrm{A} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 \\ 1 & 1 & 1 & 1 & 2 & 2 & 3 & 3 & 3 & 4 & 4 & 4 & 4 \end{bmatrix}$$

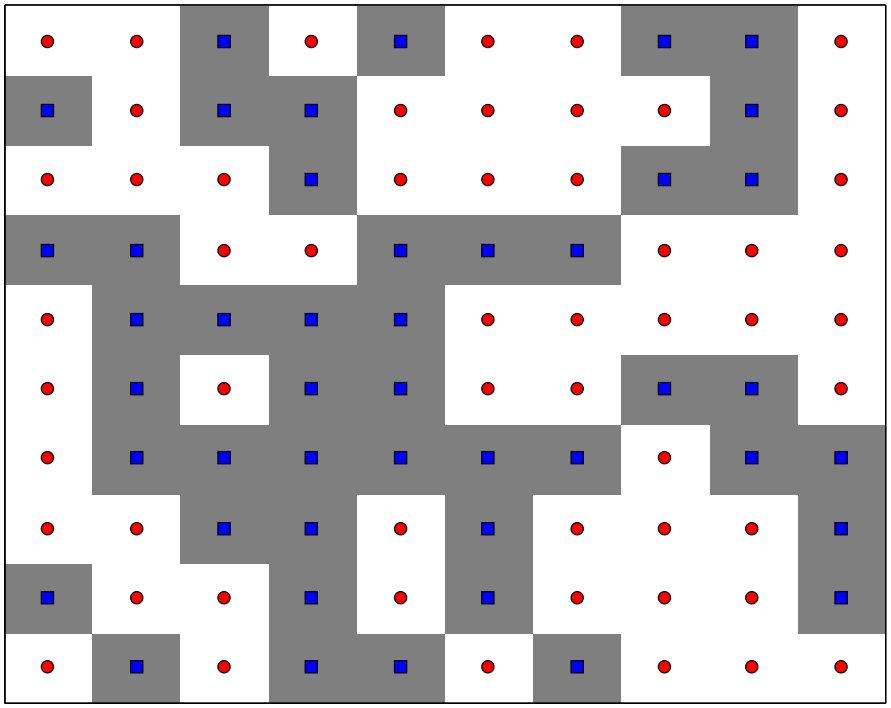
Entropia di Shannon: misura della complessità di una partizione

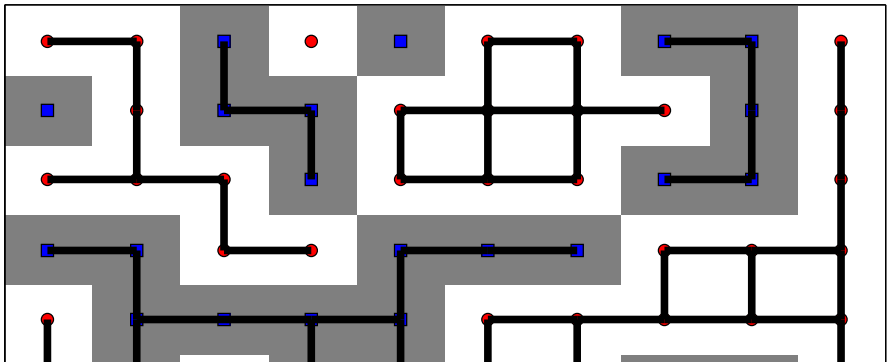
$$H(A) = \sum_{k}^{n} \mu(A_k) \log\left(\mu(A_k)\right)$$

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

Partizionamento

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

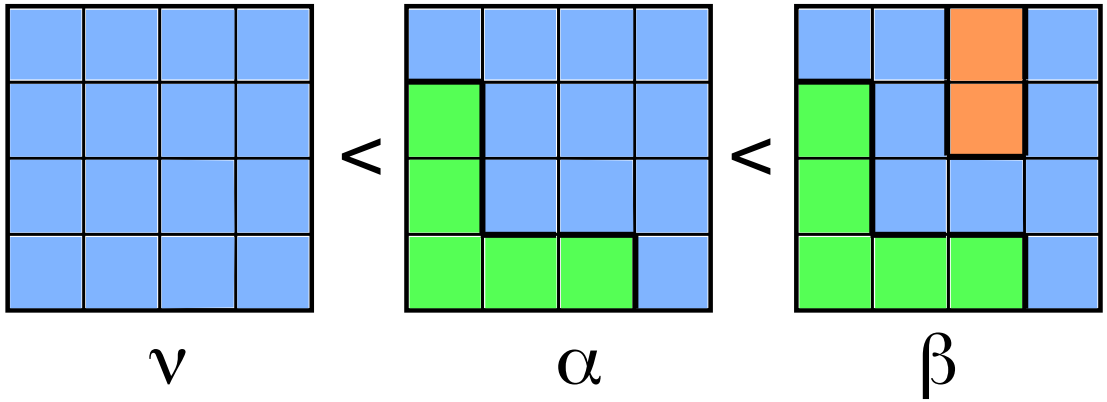




Ordinamento parziale e fattori

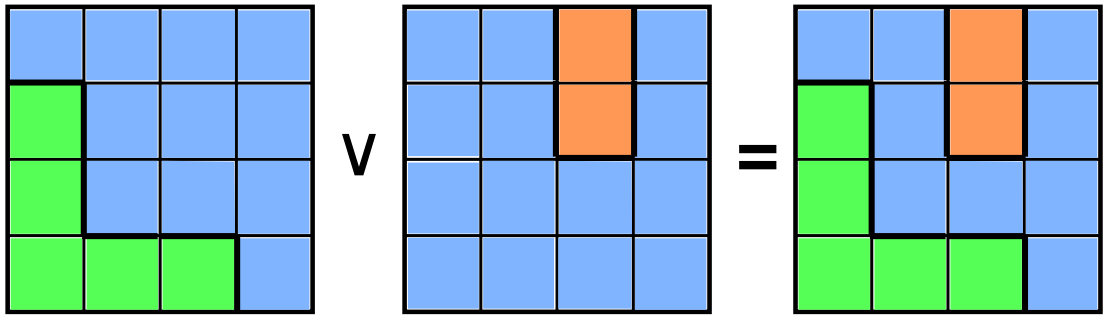
Ordinamento

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



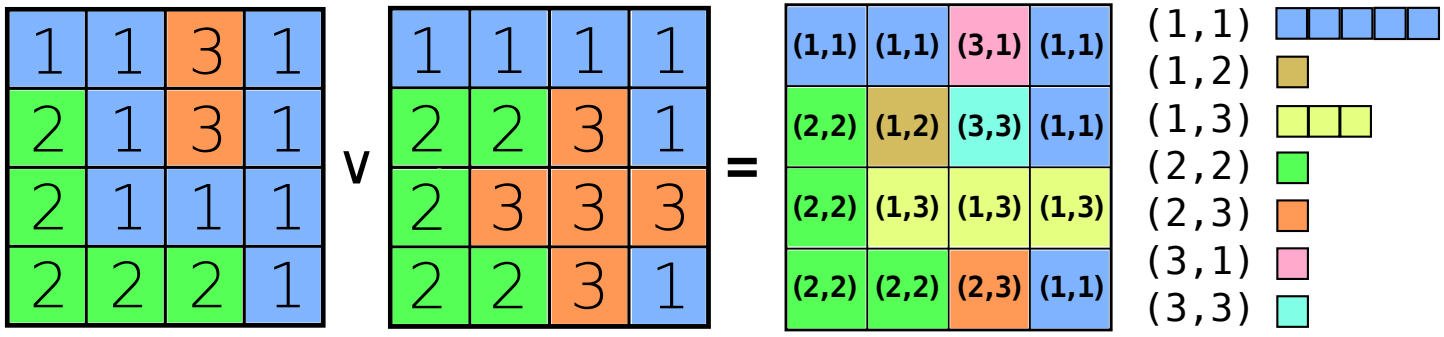
Prodotto

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

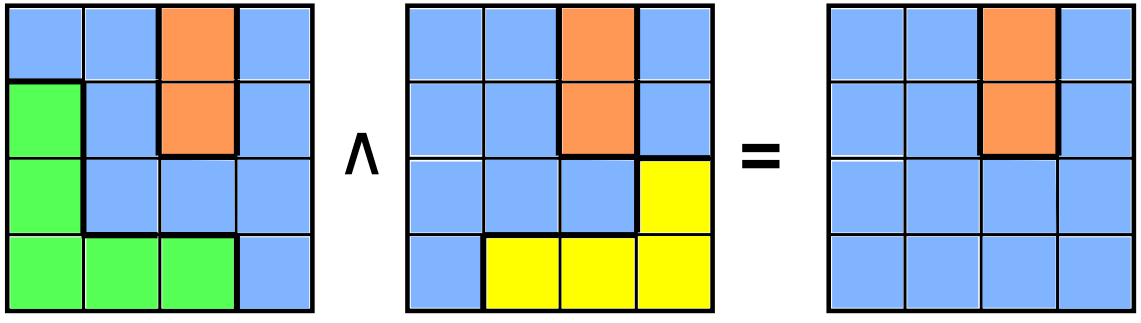


Prodotti tra partizioni

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



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



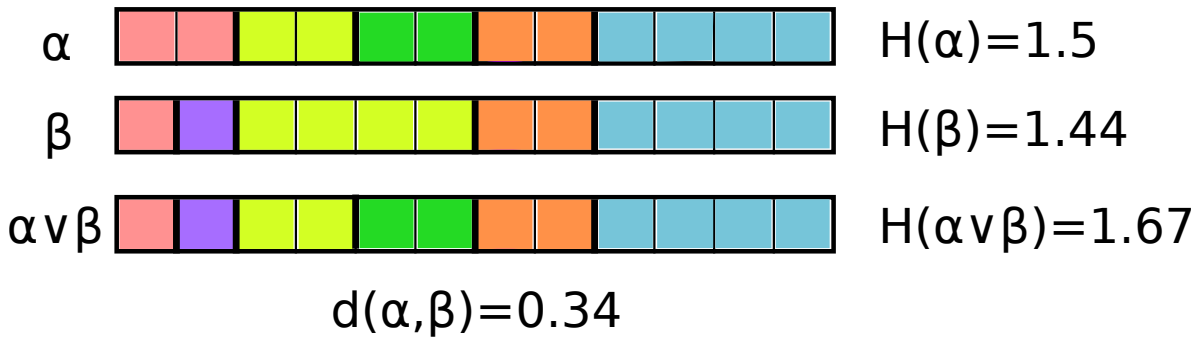
è il "**massimo comune divisore**"

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 estramemente 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

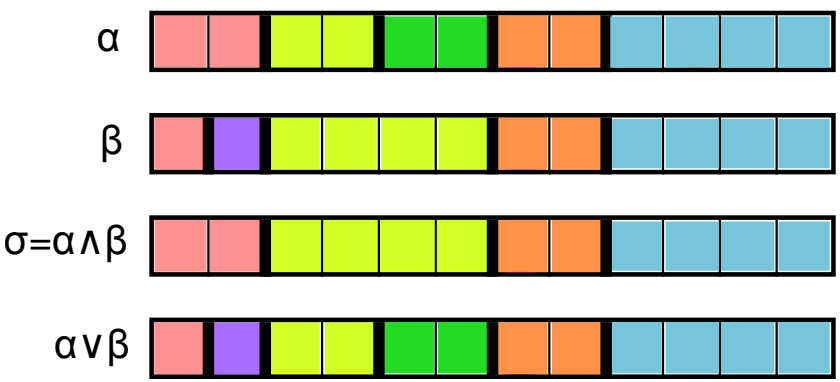
Fattori lineari

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

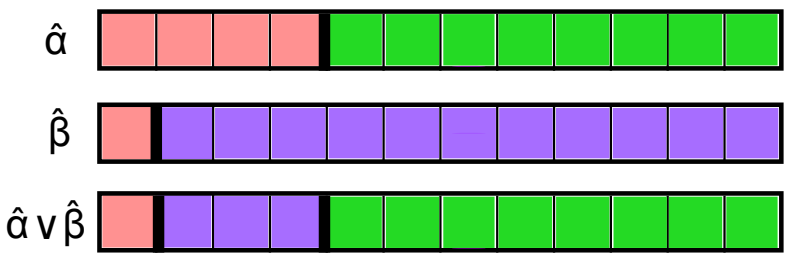
Fattori dicotomici semplici

- ovunque applicabile
- oneroso computazionalmente
- peggiore nel caso lineare

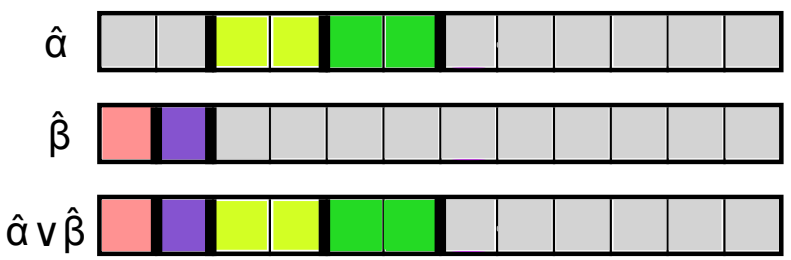
Partizioni non ridotte



Riduzione tramite fattori lineari



Riduzione tramite fattori dicotomici



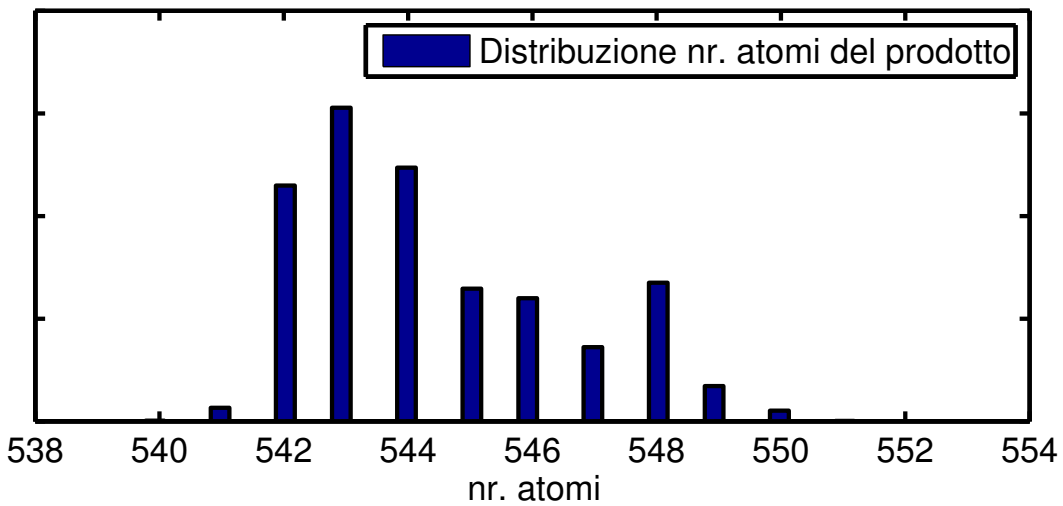
Definizione topologica della distanza

Funzionale entropia indipendente dalla misura μ:

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

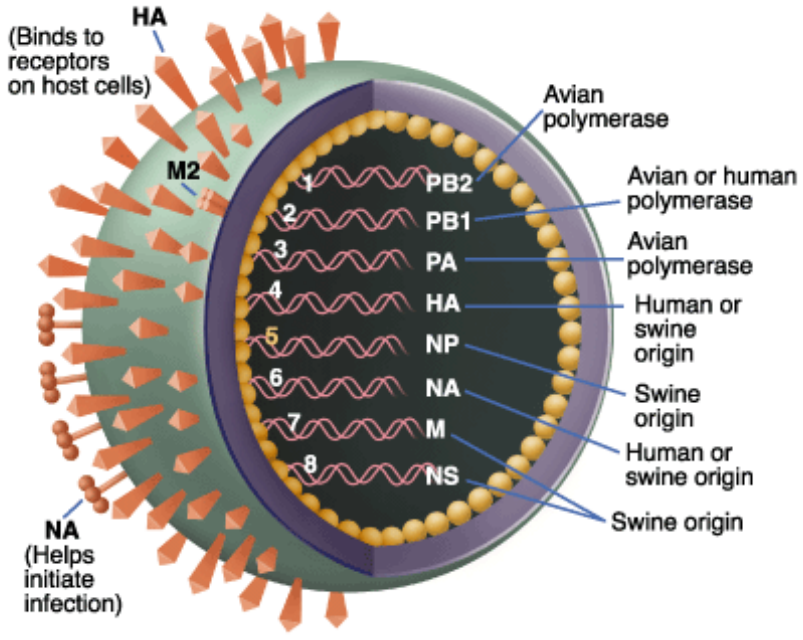
Una distanza di Rohlin opportunamente definita:

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

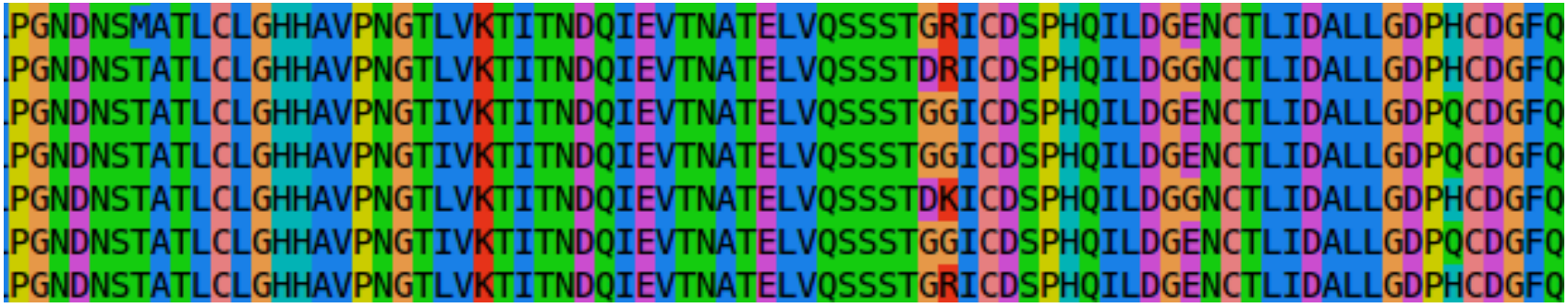


Proteine dell'influenza H3N2

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



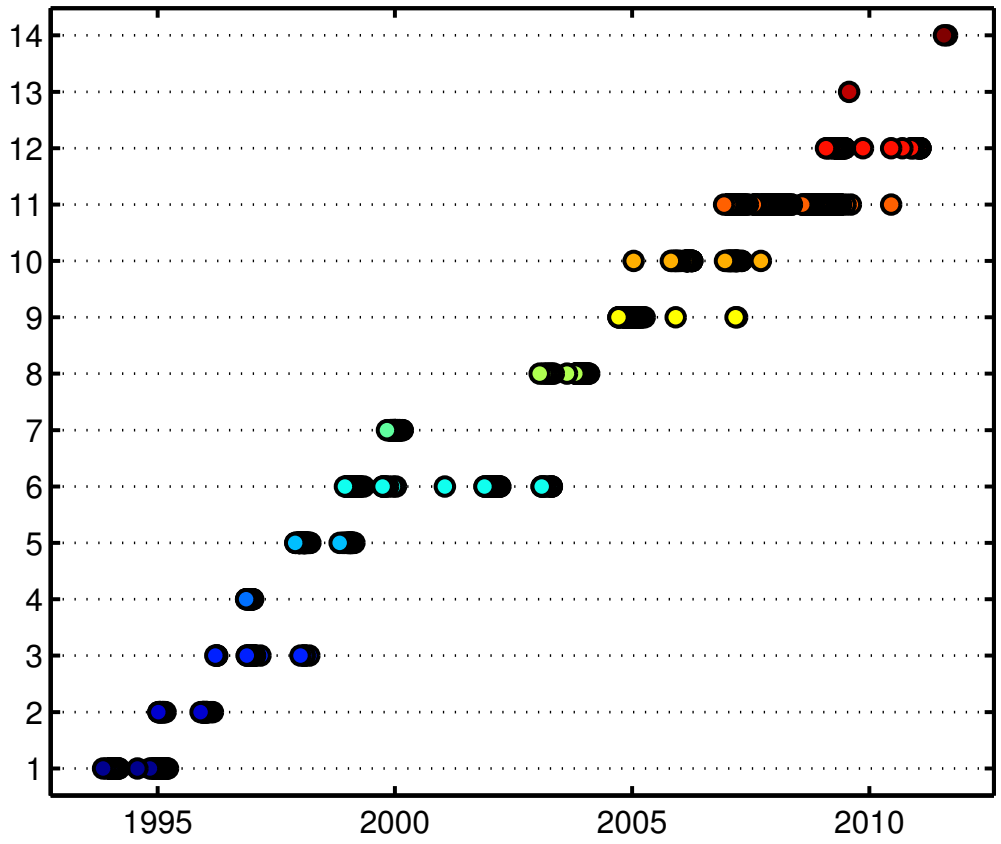
Sequenze a confronto:



Hamming è poco adatto

A={GHHAVPNGTLVKTITTGRICGDPHCDGFQNKEW}
B={GHHAVPNGTIVKTITTGEICGDPQCDGFQNKKW}

$d_H(A, B)$= #differenze
$d_H(A, B) = 4$

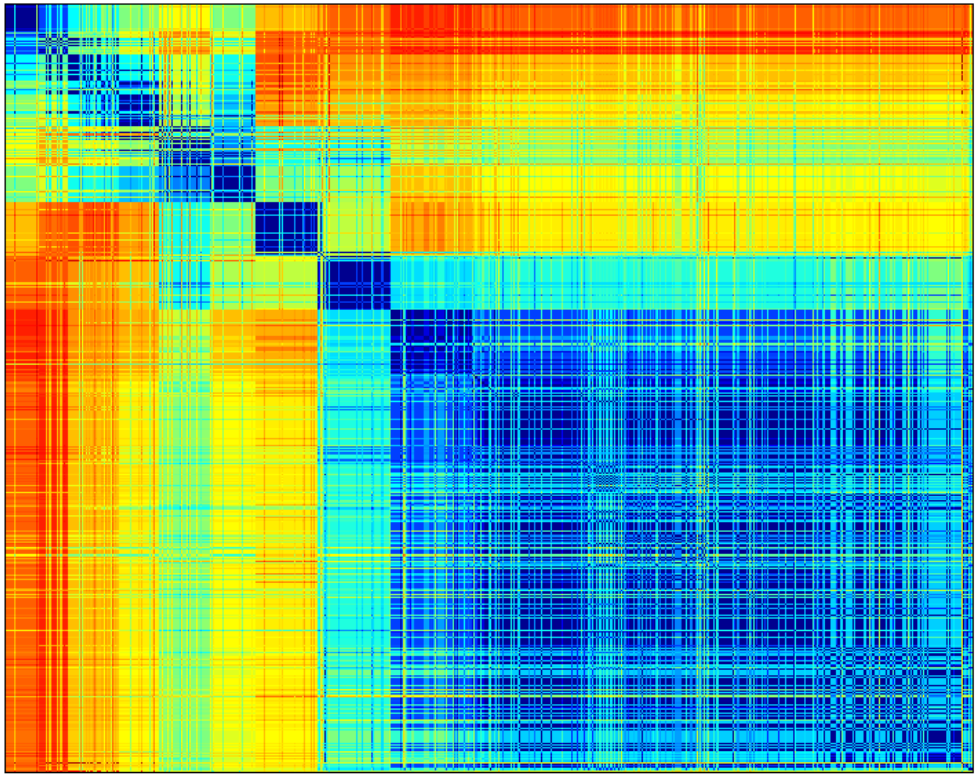


Antigenic drift

$d_H \propto t$

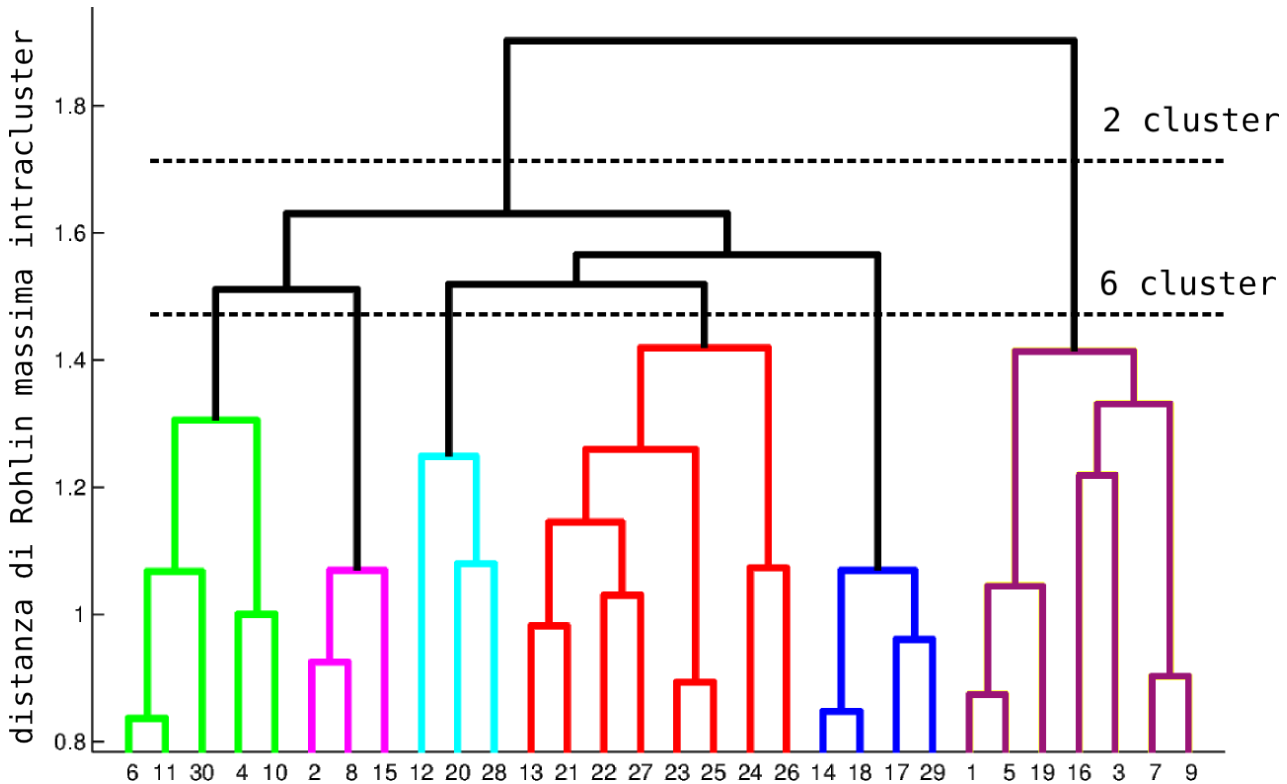
Utilizzo misure entropiche

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



Clustering di sequenze

Suddivisione di N sequenze in p **clusters**



"Clustering gerarchico agglomerativo"

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 M R C V P P E K R R G I F G A I

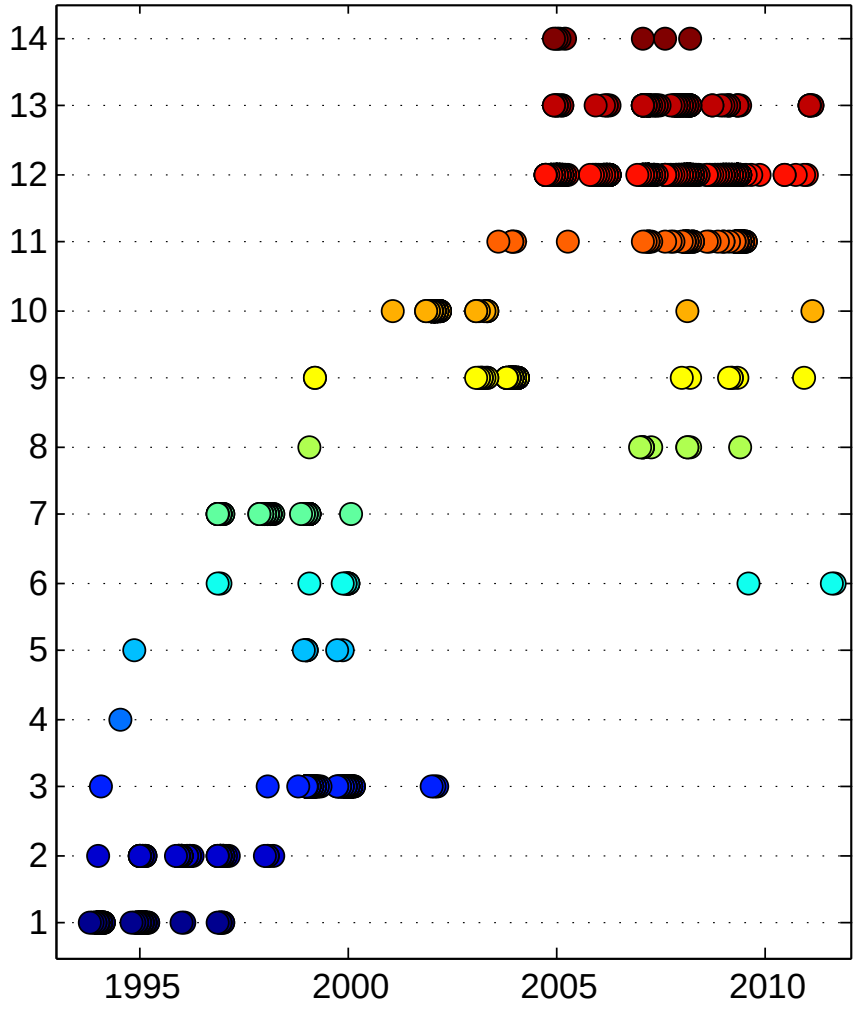
$d_R = 0.13$

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

$d_R = 1.07$



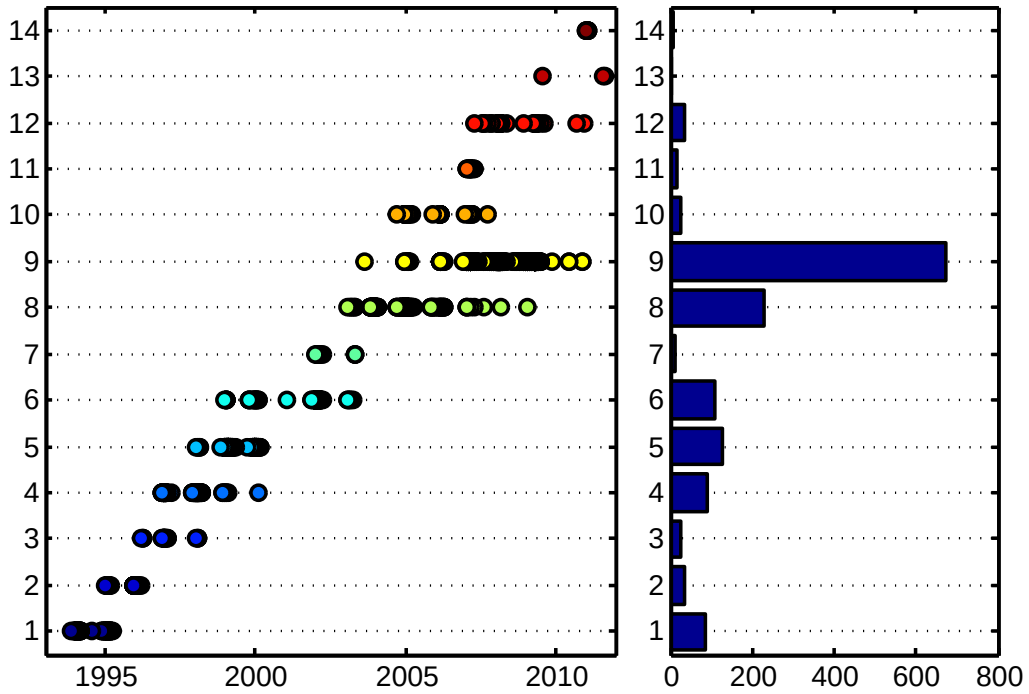
Partizionamento alternativo

Partizioni con “salto”

AAAA B AAA DDDD K DD CCCCCC

Alfabeti ridotti

LVIMC AG ST P FYW EDNO KR H



Sequenze lineari di spin (Ising 1D)

Modello di sequenza aperta di spin, lunga L:

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

variabili di **link**:

$$l_i = \sigma_i\ \sigma_{i+1}\ \mathrm{sgn}(J_i) \Longrightarrow \sigma_{i+1} = l_i\,\sigma_i\,\mathrm{sgn}(J_i) \qquad i \in \{1,\ldots,L-1\}$$

gradi di libertà indipendenti:

$$H = -\sum_i J_i l_i$$

generazione indipendente:

$$p(l_i = +) = (1 + e^{-2\beta|J_i|})^{-1}$$

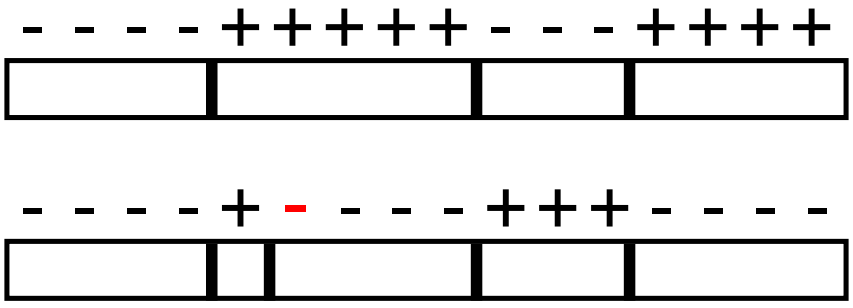
Lunghezza di correlazione tra partizioni

Se J = cost

$$\langle \sigma_i \, \sigma_{i+r} \rangle = \exp\left(-\frac{r}{\xi}\right) \qquad \text{con } \xi = -\frac{1}{\log\left(\tanh \beta J\right)}$$

quando J non positivo? $\langle \sigma_i \, \sigma_{i+r} \rangle = 0$

Distanza di Rohlin: sensibile solo ai bordi dei clusters



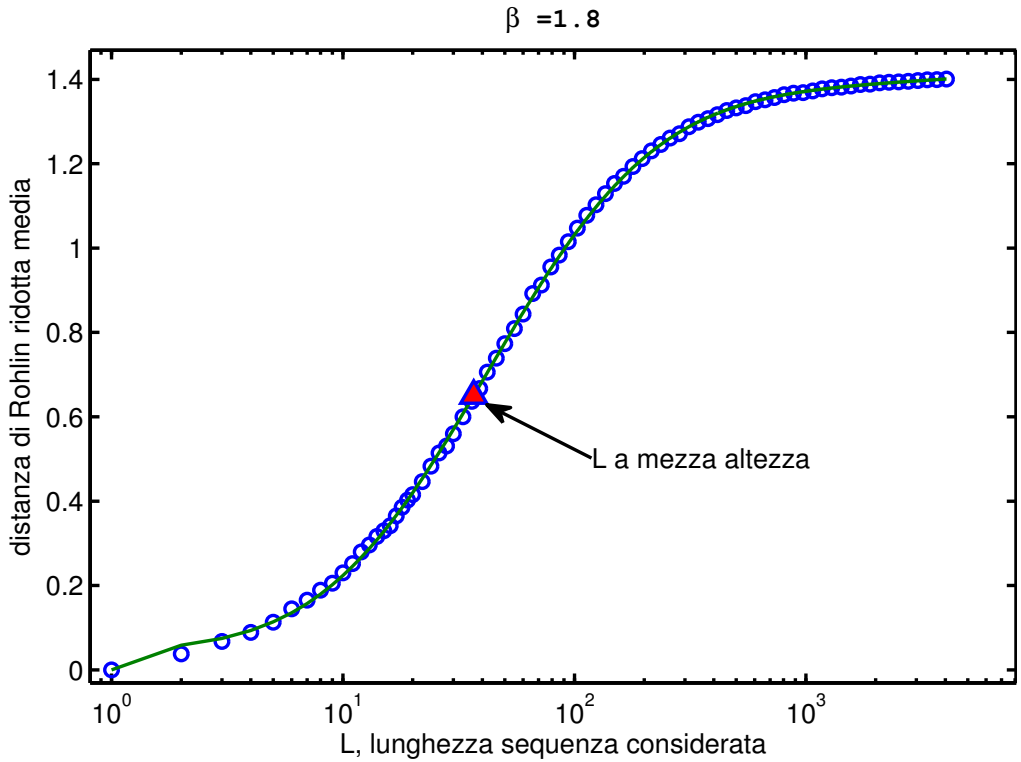
Variazione lunghezza

il numero dei cluster $n \propto p\,L = n(\beta, L) \Longrightarrow$ entropia crescente in L!

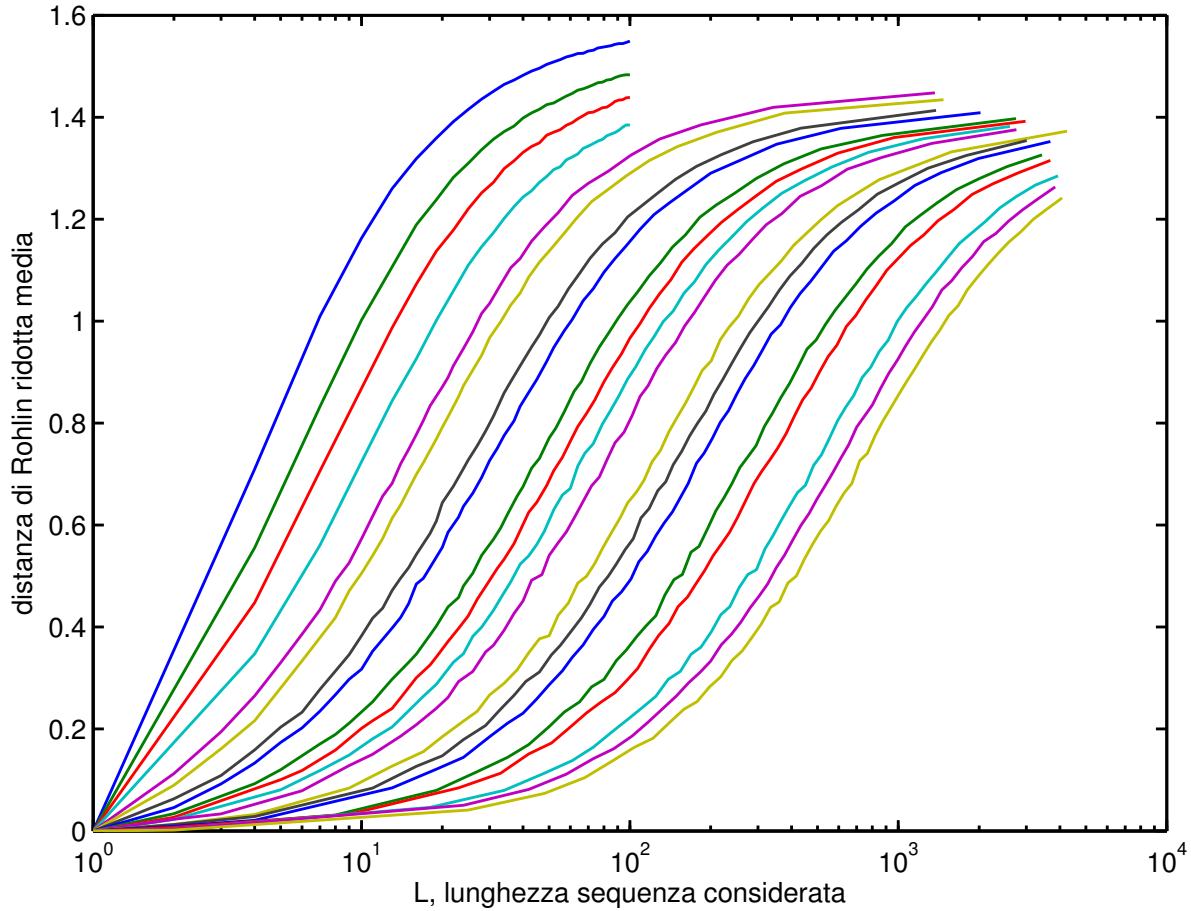
Maggiore complessità delle partizioni $\Longrightarrow$ maggiore distanza media

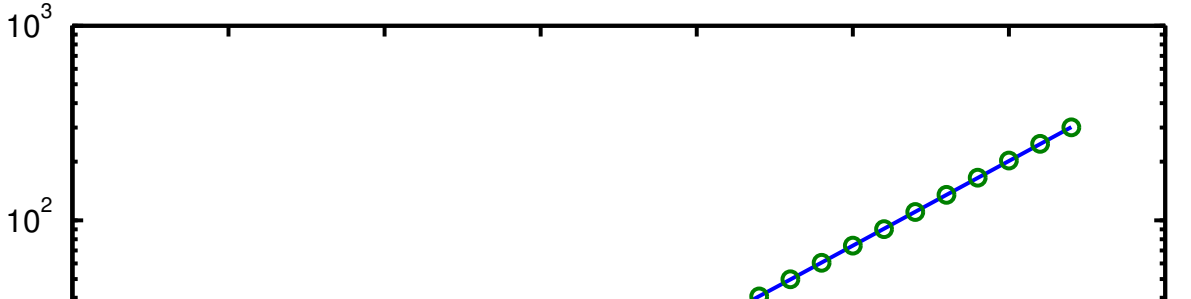
Estrazione lunghezza

1. si fissa β e L
2. generazione N sequenze a parametri fissati
3. distanza media tra N sequenze
4. $L_\xi(\beta) = L$ a metà altezza



Dipendenza dalla temperatura



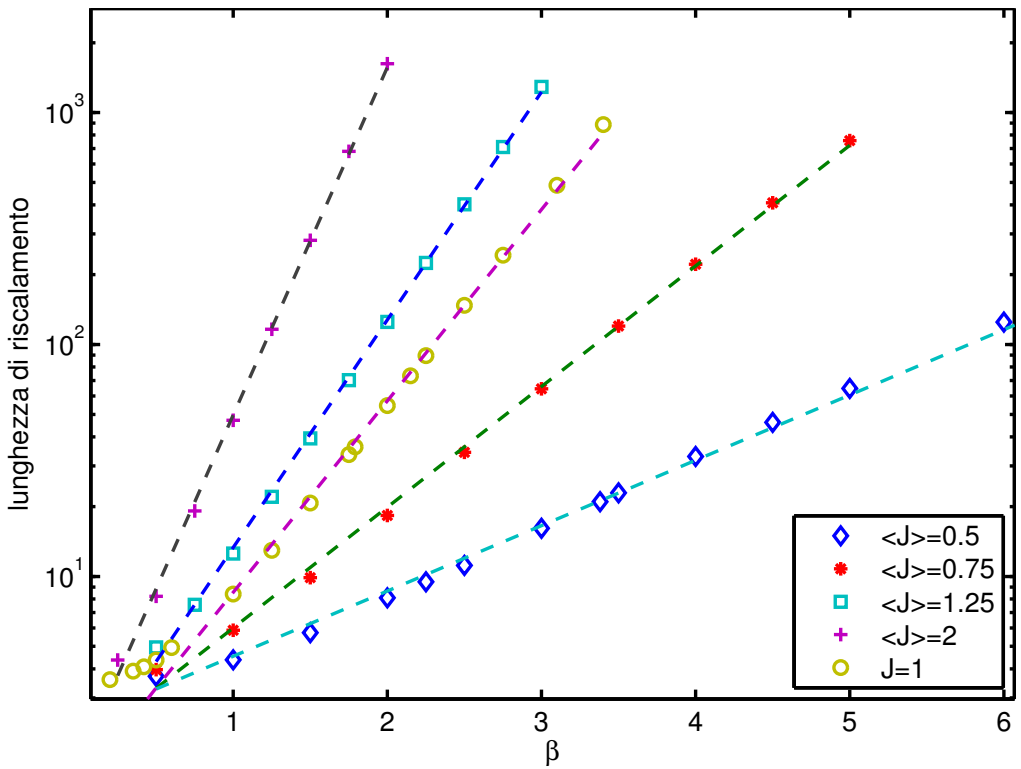


Tipi di disordine

Clusters di link

bordi dei cluster $\sigma_{i+1} = J_{i,i+1}\,\sigma_i \iff l_i = -1$

- se $J = \pm 1$ nessuna differenza, $L_\xi(\beta) \propto -(\log\tanh(\beta))^{-1}$
- se $J \in [a, b], a > 0$: $L_\xi(\beta) \propto -(\log\tanh(\langle J\rangle\beta))^{-1}$
- se $J \in [-1, 1]$, la possibilità di $J = 0$ cambia tutto: $L_\xi(\beta) \propto \beta$



Ising 2D, reticolo quadrato

- $T > T_C$: distanza arriva a un massimo, si definisce $L_\xi(\beta)$
- $T < T_C$: la distanza diverge come $\mathcal{O}(\log(N))$ – come l'integrale della correlazione sconnessa a due punti in 2D