Sviluppi teorici e applicativi delle metriche entropiche di Rohlin

Dawid Crivelli

Relazione di fine tirocinio

Lo scopo iniziale del lavoro era approfondire l'utilizzo delle misure entropiche di Rohlin su sequenze biologiche, corrispondenti alla sequenza di amminoacidi della proteina HA dei virus dell'influenza. Differente dalla solita misura che differenzia mutazioni puntuali nei simboli, la distanza di Rohlin lavora nello spazio delle partizioni sulle sequenze, permettendo di evidenziare strutture e differenze globali, con alto potere predittivo per quanto riguarda il riconoscimento dell'emergenza di nuovi ceppi virali a partire dalla distanza su un insieme di sequenze campionato nel tempo.

La metrica entropica misura la distanza tra le partizioni costruite a partire dai simboli sulle sequenze. La scelta più semplice, che ha dimostrato ottime proprietà, corrisponde a prendere partizioni formate da segmenti omogenei di simboli. Abbiamo inoltre generalizzato il criterio, rendendolo assolutamente arbitrario. Per migliorare la sensibilità delle metriche utilizzate, è risultato anche in questo caso cruciale definire un metodo di amplificazione delle distanze, detto *riduzione*, su cui abbiamo concentrato i nostri sforzi al fine di ottenere un'implementazione ottimale. La riduzione consiste nella semplificazione dei fattori che non influiscono sulla distanza: si calcola l'intersezione tra le due partizioni e si ricostruisce la partizione evitando di includere i fattori che risultano comuni.

Abbiamo sperimentato anche la definizione di diversi processi di riduzione, sempre perfettamente generali, utilizzabili su partizioni non aventi una geometria predefinita sottostante: ad esempio usando un criterio meno sensibile a variazioni sui singoli siti, tralasciando direttamente fattori anche nel caso in cui non sono comuni, ma differiscono in misura meno di una soglia ϵ. Per fare ciò si è dovuto definire operazioni binarie tra partizioni usando algoritmi di *clustering* e *sorting*, per potere rimuovere anche i limiti sull'insieme dei simboli ammesso nelle sequenze. Questo rende possibile definire lo studio delle distanze tra partizioni di qualunque forma e dimensione, generalizzando il problema allo studio di partizioni su reticoli anche con un elevato numero di siti ($N \sim 10^6$) prima non attaccabile, esibendo un andamento asintotico $\mathcal{O}(N \ln N)$ per singola operazione.

Nel caso di sequenze biologiche abbiamo studiato variazioni nei possibili tipi di misura. La complessità nelle sequenze è solitamente misurata con l'entropia metrica di Shannon, che dipende dalla lunghezza dei segmenti individuati. Non essendoci un criterio fondamentale per pesare le lunghezze, abbiamo sperimentato anche l'utilizzo dell'entropia topologica, definita solo in base al numero di atomi in cui è stata partizionata la sequenza. Abbiamo inoltre verificato cosa accade quando si formano partizioni nonconnesse, per evitare che mutazioni puntuali nascondano possibili invarianze su un piano più astratto, permettendo quindi di collegare amminoacidi uguali anche saltando un numero massimo

predefinito di simboli diversi. Si è studiato come l'informazione fornita dalla sequenza viene filtrata dall'ignorare mutazioni su scale sempre più lunghe o utilizzando alfabeti ridotti di amminoacidi. Tutto per svincolarsi da supposizioni ad hoc ed esplorare il panorama delle possibili misure, per verificare quali meglio mettono in evidenza la funzionalità biologica in gioco nel processo evolutivo, senza mai fare assunzioni richiedenti la conoscenza di caratteristiche biologiche non direttamente inferibili dalle sequenze considerate. Si è anche analizzato diversi algoritmi per raggruppare sequenze simili, notando che l'emergenza dei *cluster* di notevole interesse biologico evidenziati grazie alla misura di Rohlin non dipende dal metodo scelto. Lo studio di tutte le distanze ha selezionato il miglior tipo di misura, in cui la *riduzione* ha un ruolo fondamentale, mentre un'eccessiva aggregazione di siti e mutazioni maschera l'essenziale funzione del drift genetico.

La caratteristica di riconoscere e differenziare solo strutture formate da più simboli, ne permette l'applicazione laddove i valori presi singolarmente non contengono informazione sulla fisica del sistema. Così accade in certi problemi di meccanica statistica, dove le simmetrie rendono equiprobabili i simboli di una sequenza: la distanza di Rohlin si è rivelata importante nel catturare l'effettiva differenza tra realizzazioni dello stesso sistema. Per sistemi di Ising monodimensionali, anche generalizzando l'accoppiamento al caso disordinato, la distanza di Rohlin media al crescere della lunghezza delle sequenze ha permesso di definire e misurare una lunghezza di correlazione efficace in funzione della temperatura. Lo studio è stato esteso alla misura di un sistema bidimensionale di Ising in diversi istanti temporali, sia all'equilibrio che nell'evoluzione temporale da una configurazione casuale, per verificare la possibilità dello studio in casi disordinati oltre al reticolo monodimensionale.