

# Capitolo 1

## Allineamento multiplo di sequenze

### 1.1 Introduzione

Per ottenere informazioni su una sequenza genomica, questa viene sottoposta a un processo di analisi di sequenza, che può assumere forme diverse a seconda dell'origine della sequenza e degli obiettivi dello studio. La progettazione di metodi per il confronto delle sequenze è complessa a causa di numerose difficoltà, tra cui mutazioni puntiformi, variazioni di lunghezza, cambiamenti drastici, mascheramento e altri eventi evolutivi. Per superare tali difficoltà, il passaggio cruciale è l'allineamento delle sequenze, una tecnica fondamentale alla base di molti strumenti software utilizzati nella ricerca nei database biologici. L'allineamento consente di confrontare sequenze anche quando non sono identiche, permettendo di evidenziare regioni conservate nonostante la divergenza evolutiva. A causa delle mutazioni, anche le sequenze dello stesso gene o della stessa proteina in specie strettamente correlate sono raramente identiche. Il confronto tra sequenze è quindi essenziale per inferire l'omologia, ossia la derivazione da un antenato comune. Idealmente, l'obiettivo del confronto è allineare le sequenze in modo che basi o amminoacidi derivati dalla stessa posizione ancestrale risultino in corrispondenza. In assenza di informazioni aggiuntive, ciò si ottiene massimizzando la somiglianza tra le sequenze allineate. Quando le sequenze sono molto simili, l'allineamento è semplice e le regioni identiche risultano chiaramente visibili. Tuttavia, quando le sequenze sono più divergenti, ad esempio a causa di inserzioni o delezioni, un confronto diretto può nascondere la reale somiglianza e generare corrispondenze false tra posizioni non equivalenti. Per ovviare a questo problema, si introducono gap (spazi vuoti) in una o entrambe le sequenze, così da preservare il maggior numero possibile di corrispondenze significative. Poiché per una stessa coppia di sequenze esistono molti allineamenti possibili, non è sempre immediato individuare quello migliore, soprattutto quando il livello di somiglianza è basso. Per questo motivo, i metodi di confronto

delle sequenze utilizzano algoritmi basati su schemi di punteggio quantitativi, che valutano la qualità degli allineamenti e scartano quelli non soddisfacenti secondo criteri prestabiliti. Nel confronto tra sequenze biologiche, la questione fondamentale è stabilire se le somiglianze osservate siano dovute al caso oppure riflettano una relazione evolutiva, cioè una omologia. I termini similarità e omologia non sono equivalenti: la similarità indica semplicemente il grado di somiglianza tra sequenze, mentre l'omologia implica una discendenza comune e ha quindi un preciso significato evolutivo e biologico. I geni o le proteine omologhe derivano da un antenato comune e divergono nel tempo a causa dell'accumulo di mutazioni. Spesso l'omologia suggerisce anche una struttura o funzione simile, poiché la selezione naturale tende a conservare mutazioni che mantengono la funzione e la stabilità strutturale delle proteine. Tuttavia, residui identici o simili possono riflettere solo una divergenza evolutiva recente e non essere necessariamente essenziali per la funzione. Inoltre, un'elevata similarità di sequenza non garantisce sempre una funzione comune, poiché nuove funzioni possono emergere con modifiche minime della sequenza. Al contrario, proteine con bassa similarità di sequenza possono conservare la stessa struttura tridimensionale e la stessa funzione. In questi casi sono necessarie informazioni aggiuntive, come dati strutturali o biochimici. Esistono anche casi di evoluzione convergente, in cui sequenze simili non sono omologhe ma hanno acquisito somiglianze indipendentemente per svolgere una funzione analoga. Tuttavia, questo processo raramente produce lunghe regioni altamente simili. Un allineamento di sequenze rappresenta quindi una ipotesi evolutiva su quali residui derivino dallo stesso antenato, ma da solo non consente di distinguere tra omologia ed evoluzione convergente. Per questo motivo, i metodi di confronto delle sequenze devono tenere conto dei meccanismi evolutivi, delle proprietà chimico-fisiche degli amminoacidi e delle pressioni selettive, al fine di distinguere allineamenti casuali da quelli che riflettono una reale relazione evolutiva.