Capitolo 1

Allineamento multiplo di sequenze

1.1 Contesto Generale e nascita sequence analysis

La rivoluzione nell'analisi genetica, iniziata negli anni 70 con lo sviluppo della tecnologia del DNA ricombinate e delle tecniche di sequenziamento, ha portato alla disponibilità di un'enorme quantità di dati biologici. Oggi esistono vasti database contenenti sequenze di nucleotidi e amminoacidi provenienti da una varietà di organismi. Nonostante questa abbondanza d'informazioni di sequenza, per una larga frazione dei geni e delle proteine annotate non sono note nè la struttura tridimensionale nè la funzione biologica. Questo ha portato alla nascita di un nuovo campo di ricerca, noto come sequence analysis, che si occupa di sviluppare metodi computazionali per l'analisi delle sequenze biologiche al fine di inferire informazioni strutturali e funzionali. La sequence analysis includono diverse fasi a seconda dell'origine della sequenza e degli obiettivi dell'analisi. Tra queste fasi, una delle più importanti e ricorrenti è **l'allineamento di sequenze**, utilizzato per individuare omologie e per confrontare sequenze nuove con quelle già presenti nei database.

1.2 Ruolo dell'Allineamento di sequenze

L'allineamento di sequenze è un passaggio chiave sia per: il confronto diretto tra due o più sequenze; sia per il database searching, ovvero la ricerca di sequenze simili in grandi archivi biologici. L'identificazione di sequenze simili ha numerose applicazioni:

- Permette di riconoscere geni codificanti all'interno di sequenze genomiche grezze

- Consente di assegnare una funzione putativa a un gene tramite il confronto con geni già caratterizzati
- Fornisce indicazioni sulla struttura tridimensionale delle proteine

1.3 Concetto di Allineamento e introduzione dei Gap

L'allineamento di sequenze consiste nel disporre due o più sequenze in modo tale che le residui derivati da uno stesso residuo ancestrale occupino la stessa posizione. Questo obiettivo viene generalmente preseguito massimizzando la similarità complessiva dell'allineamento. Quando le sequenze differiscono in lunghezza, un allineamento diretto produce:

- Posizioni in cui i residui non corrispondono (mismatches)
- Posizioni in cui una sequenza ha residui aggiuntivi rispetto all'altra (inserzioni o delezioni, note come indels)

L'introduzione di **gap** che rappresentano insezioni o delezioni evolutive, consente di migliorare l'allineamento tra le sequenze. Tuttavia non esistono regole rigide per l'inserimento dei gap, e la loro gestione rappresenta una sfida significativa nell'allineamento di sequenze. Per questo motivo, gli algoritmi di allineamento devono :

- generare possibili allineamenti
- assegnare a ciascuno un punteggio quantitativo
- scartare quelli non significarivi in base a criteri statistici

1.4 Similarità a Confronto con omologia

Un punto concettuale fondamentale è la distinzione tra similarità e omologia. La Similarità è un termine desccrittivo che indica un certo grado di corrispondenza tra sequenze; L'omologia implica invece una relziione evolutiva, ovvero la derivazione da un antenato comune Due sequenze omologhe possono aver divergenze significative a livello di sequenza, pur mantenendo :

- una struttura tridimensionale simile
- una funzione biologica analoga

Tuttavia: un elevata similarità tra sequenze non garantisce necessariamente un'origine evolutiva comune; ed una bassa similarità non esclude l'omologia. Esistono anche casi di evoluzione convergente, in cui sequenze non omologhe mostrano somiglianze locali dovute a vincoli funzionali simili. In tali casi, la similarità non riflette una comune origine evolutiva. Un allineamento deve quindi

essere interpretato come una ipotesi sulle corrispondenze evolutive tra residui, non come una dimostrazione definitiva di omologia. Esistono anche casi di **evoluzione convergente**, in cui sequenze non omologhe mostrano somiglianze locali dovute a vincoli funzionali simili. In tali casi, la similarità non riflette una comune origine evolutiva. Un allineamento deve quindi essere interpretato come una ipotesi sulle corrispondenze evolutive tra residui, non come una dimostrazione definitiva di omologia.

1.5 Modelli evolutivi, scoring e algoritmi

I metodi computazioni di confronto delle sequenze devono tenere conto di :

- diversi tipi di mutazione
- proprietà fisico-chimiche degli amminoacidi
- pressioni selettive che favoliscono o eliminano determinate variazioni

Questi fattori vengono incorporati in :

- **schemi di scoring** che assegnano punteggi a matches, mismatches e gap
- **algoritmi di allineamento** che cercano di ottimizzare il punteggio complessivo

Infine, è essenziale distiguere tra

- allineamenti apparentemente buoni ma dovuti al caso
- allinemaneti che riflettono una relazione evolutiva reale, utilizzando criteri statici adeguati.

1.6 Limiti della percent identity come misura di similarità

La **percentuale d'identità** rappresenta la misura più semplce e immediata della quantità di un allineamento, in quanto si ottiene semplicemente calcolando la frazione di posizioni allineate in cui i residui sono identici. Sebbene questa misura sia utile come **test preliminare rapido**, essa risulta concettualmente grossolana e insufficiente a descrivere in modo accurato il grado reale di somiglianza biologica tra due sequenze, in particolare nel caso delle **sequenze proteiche** Il principale limite della percent identity risiede nel fatto che essa assegna un valore binario alle posizioni allineate:

- **1** per un match identico
- **0** per un mismatches

Questo approccio ignora completamente la natura chimica e fisica degli amminoacidi coinvolti, trattando allo stesso modo sostituizioni biologicamente plausibili e sostituzioni estremamente improbabili dal punto di vista evolutivo.

1.7 Similarità funzionale tra amminoacidi non identici