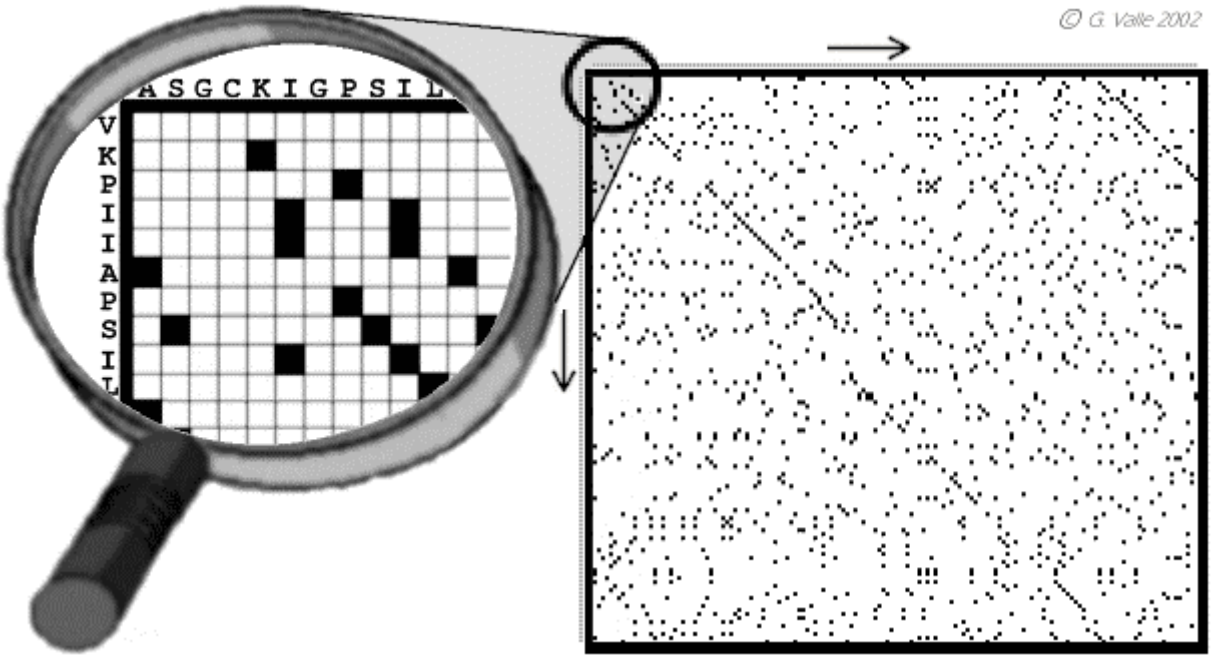
B I O I N F O R M A T I C A



Il metodo della dot matrix consiste nel creare una matrice in cui vengono confrontati tutti i possibili appaiamenti di ogni carattere delle due sequenze da allineare. In termini pratici, una sequenza viene scritta sul lato superiore della matrice, da sinistra a destra, ponendo ogni carattere in corrispondenza di ogni colonna. Chiameremo questa sequenza "sequenza orizzontale". Similmente, la seconda sequenza (sequenza verticale) viene scritta sul lato sinistro della matrice, dall'alto in basso ponendo ogni carattere in corrispondenza di ogni riga. Nella figura la direzione delle sequenze è indicata dalle frecce. I caratteri sono visibili solo nell'ingrandimento a sinistra, mentre nella matrice intera sono troppo piccoli per essere visti distintamente.

La fase succesiva consiste nel riempimento della matrice. Ogni casella dovrà essere analizzata e qualora le lettere a capo delle corrispondenti righe e colonne fossero identiche allora la casella sarà annerita. Si dovrà pertanto effettuare un numero di operazioni pari al numero di caselle, cioè al prodotto delle lunghezze delle due sequenze. Queste operazioni possono essere rapidamente completate da un computer per cui anche una matrice relativamente grande, ad esempio relativa a due sequenze di 1000 caratteri (un milione di caselle), viene tipicamente riempita in una frazione di secondo.

Una volta completata l'analisi il risultato può essere analizzato visivamente. In corrispondenza di allineamenti di vedrà una diagonale. Ad esempio nel riquadro ingrandito di destra si può vedere una diagonale di 4 caselle contigue corrispondente alla sequenza "PSIL" che è presente su entrambe le sequenze.

Consideriamo alcuni aspetti dell'analisi visuale di una dot matrix. Se analizzassimo due sequenze identiche (cioè usiamo la stessa sequenza sia come sequenza orizzontale che come sequenza verticale) allora otterremmo una diagonale continua che parte dall'angolo in alto a sinistra per arrivare a quello in basso a destra. Ovviamente oltre alla diagonale troveremmo molti altri puntini. Si consideri che ci sono 20 aminoacidi diversi, quindi in una sequenza casuale ci dovremmo aspettare una casella positiva ogni 20. Similmente, con acidi nucleici dovremmo aspettarci una casella positiva ogni 4, con un notevole rumore di fondo.

Cerchiamo ora di spiegare la piccola diagonale che si può distinguere in prossimità dell'angolo in alto a destra. Una diagonale posta in quella posizione indica che la parte finale della sequenza orizzontale è simile alla parte iniziale della sequenza verticale. Inoltre, considerando che la stessa parte della sequenza verticale è simile anche alla parte iniziale della sequenza orizzontale, possiamo dedurre che la prima parte e l'ultima parte della sequenza orizzontale sono simili, probabilmente

generate da una duplicazione parziale del gene. In generale, quando si osservano due diagonali parallele si deve pensare ad una porzione della sequenza ripetuta.
Un ultimo aspetto da spiegare è il fatto che la diagonale principale della figura mostra un'evidente interruzione nella parte centrale per poi continuare su una diagonale diversa, un po' più bassa. Questi salti di diagonale sono dovuti alla presenza di "gap", ossia di "buchi" in una delle due sequenze, cioè a segmenti di sequenza che sono presenti in una sequenza ma non nell'altra. Nel caso del nostro esempio abbiamo un gap nella sequenza orizzontale o, se si preferisce, un segmento di sequenza aggiuntivo nella sequenza verticale.

Matrici di Sostituzione

Le matrici di sostituzione sono matrici che assegnano a ciascuna delle possibili coppie di amminoacidi, un valore che indica il loro grado di similarità (informazione sulla probabilità che un amminoacido si sostitutisca ad un altro durante l'evoluzione).
Si ottengono con metodi statistici assegnando a ciascuna coppia un valore che riflette la frequenza con cui l'uno si sostituisce all'altro in gruppi di proteine omologhe. Per allineare sequenze proteiche vengono utilizzate, quasi esclusivamente, matrici basate sulla frequenza di sostituzioni in gruppi di proteine omologhe, ovvero le matrici PAM e le matrici BLOSUM

MATRICI PAM

Le matrici di sostituzione PAM furono proposte nel 1978 da Margaret Dayhoff e dai suoi collaboratori, sulla base di uno studio di filogenesi molecolare compiuto su 71 famiglie di proteine[1].
Queste matrici sono fondate sull'osservazione del conteggio del cambio degli aminoacidi in un gruppo di proteine fortemente in relazione tra di loro (85% identità). Esse sono costruite sulla base di relazioni evolutive evidenziano, inoltre, la probabilità di cambio di un aminoacido in un altro in proteine omologhe durante l'evoluzione.

MATRICI BLOSUM

Le matrici BLOSUM furono introdotte nel 1992 da S. Henikoff e J.G. Henikoff[2] per attribuire un punteggio alle sostituzioni nei confronti tra sequenze aminoacidiche. Il loro scopo era quello di sostituire le matrici della Dayhoff, facendo uso di una quantità maggiore di dati che si era resa disponibile successivamente al lavoro della Dayhoff. Le matrici BLOSUM sono basate sulla banca dati BLOCKS, che contiene una collezione di allineamenti multipli di segmenti proteici senza gap. Ogni blocco si riferisce normalmente ad un insieme di proteine in relazione tra di loro. Mediante delle tecniche di aggregazione, tutte le sequenze contenute in un blocco, vengono messe insieme in gruppi. All'interno di una famiglia viene determinata la *frequenza di sostituzione* e quindi le sostituzioni significative.
Il valore numerico associato alle matrici (es. BLOSUM62) rappresenta il valore di soglia applicato dal metodo di aggregazione. Un valore di 62 indica che le sequenze sono messe insieme nello stesso gruppo, se hanno un valore di identità uguale o maggiore al 62%.
Valori molto bassi di soglia (es. BLOSUM 45) stanno ad indicare che sono state raggruppate sequenze che sono più distanti evolutivamente.