Sommario:

Introduzione

L'omologia fra sequenze aminoacidiche e nucleotidiche può essere **globale** o **locale**.
I programmi che sono in grado di analizzare il secondo tipo di omologia sono senz'altro più utili, specialmente qualora si debbano confrontare delle sequenze di DNA in quanto spesso vengono ricercate solo piccole zone di omologia e in quanto il DNA può contenere ampie zone non codificanti.

BLAST [1] [2] che è l'acronimo di **Basic Local Alignment Search Tool**, è un programma euristico per la ricerca di *omologie locali di sequenza* ed è in realtà costituito da un insieme di **5** programmi:

BLASTP	paragona una sequenza aminoacidica ad un database di sequenze proteiche.
BLASTN	paragona una sequenza nucleotidica ad un database di sequenze nucleotidiche
BLASTX	paragona una sequenza nucleotidica (traducendola in tutti 6 possibili frame di lettura) ad un database di proteine. Di tutti i progammi che fanno parte di **BLAST** è il più usato.
TBLASTN	paragona una sequenza aminoacidica ad un database di acidi nucleici tradotto dinamicamente nelle 6 possibili sequenze di aminoacidi che possono derivarne.
TBLASTX	paragona una sequenza nucleotidica letta secondo tutti i 6 possibili frame di lettura con un database di acidi nucleici anch'esso letto secondo tutti i 6 possibili frame di lettura. Poichè ne derivano 36 combinazioni, questo programma viene utilizzato solo per ricerche su database di tipo EST.

Basi concettuali del programma

BLAST [1] [2] è un programma euristico per la ricerca di *omologie locali di sequenza* basato sulla dimostrazione data da Karlin & Altschul (1990) [3] che un allineamento locale di sequenze *prive* di gap può essere valutato con metodi statistici.
In questo **BLAST** si differenzia da FASTA che è un altro programma euristico per il confronto fra sequenze comunemente usato. FASTA infatti ricerca il migliore allineamento fra l'*intera sequenza* sottoposta ad indagine e il database di sequenze usato come riferimento.

1. Compilazione di una lista di words a punteggio superiore ad un valore-soglia T

In **BLAST** la valutazione dell'omologia comincia con l'analisi della sequenza che deve essere sottoposta al confronto. Si crea un elenco di tutte le **words** che compongono tale sequenza.
Con questo termine si indicano i tratti di sequenza di lunghezza ***w*** (in genere di **3** aminoacidi o **12** nucleotidi) che rappresentano uno dei cardini sui quali si fonda l'algoritrmo di BLAST.
Il numero totale di *words* presenti in una sequenza da sottoporre a confronto, risulta essere:

$$n = l - w + 1$$

ove ***w*** è il numero degli aminoacidi che compongono una *word* ed ***l*** è la lunghezza della sequenza in esame.

Per ogni *word* della sequenza da esaminare viene costruita una lista di possibili *words* che, se confrontate con la sequenza in questione, abbiano un punteggio superiore ad un *valore-soglia* **T** (compreso fra 11 e 15) calcolato di volta in volta in base alla composizione e alla lunghezza della sequenza in esame e in base alla matrice di sostituzione utilizzata (normalmente **PAM 120** [4] o **BLOSUM 62** [5]). A tale scopo si usa una equazione *ad hoc* che considera i parametri **H** (entropia del target), e **lambda** (unità di informazione guadagnata per un allineamento). Quest'ultimo è funzione della matrice di sostituzione.

N.B.: Le matrici di sostituzione assegnano un punteggio positivo per ogni identità o per una sostituzione con aminoacidi dello stesso tipo (idrofobici con idrofobici, carichi positivamente con carichi positivamente ecc...) e negativo per una sostituzione con aminoacidi fra loro diversi (es. aminoacido basico con aminoacido acido ecc..). Tali matrici inoltre assegnano punteggi positivi di differente entità a seconda che gli aminoacidi coinvolti siano rari o frequenti. In questo secondo caso infatti si può pensare che l'omologia sia casuale.

Nel caso dei nucleotidi (**BLASTN**) il punteggio è di più semplice valutazione: viene assegnato un punteggio di **+5** ad una identità di residui e di **-4** per una mancata identità.
Dati questi presupposti, si è visto che la combinazione che è il miglior compromesso fra sensibilità, specificità del metodo e velocità di esecuzione del confronto fra le sequenze, è quella con **w=3** e **T=11-15**. Utilizzando questi valori, si ottengono delle liste di circa *50 words di confronto* denominate **neighbors** per ogni word della sequenza da testare, cioè circa 12.500 *words* nel caso di una sequenza di 250 aminoacidi.

Questo dato è ben diverso dalle 20^3 combinazioni possibili (per $w = 3$) per ciascuna word della sequenza da testare, che sarebbero necessarie se non venisse effettuata questa preselezione.

2. Scansione del database per ricercare le corrispondenze (hits)

In questa fase ciascuna delle word della lista compilata (12.500 circa nel caso di una sequenza di 250 aminoacidi), viene confrontata con il database delle sequenze.
Quando viene riscontrata una corrispondenza (*hit*), essa viene estesa a monte e a valle per vedere se è possibile definire un tratto di sequenza in grado di raggiungere un punteggio superiore ad un *valore-soglia* detto **S**.
Tale valore **S**, è funzione di un altro valore, detto **E**, che è il numero atteso (*Expected*) di tratti di sequenza *casualmente* omologhi, aventi punteggio superiore a **S**.
Come detto, c'è una relazione tra **E** ed **S**: tanto più elevato è **E**, tanto minore diventa **S**, per cui aumenta la sensibilità del risultato, ma si riduce del pari la specificità del metodo.
Per un dato valore di **E**, una certa matrice di sostituzione ed una certa sequenza da esaminare, **S** assume diverso valore a seconda dell'ampiezza del database con il quale si effettua il confronto.
Pertanto, per *normalizzare* la situazione è stato introdotto un ulteriore parametro denominato **Z**.

I tratti di sequenza omologhi aventi un punteggio (*score*) superiore al *valore-soglia* **S**, vengono denominati **HSP** (*High Score Segment Pair*). Essi possono essere anche più di uno all'interno di una medesima sequenza e definiscono una *zona locale di omologia*.
Un particolare tipo di HSP è il cosiddetto **MSP** (*Maximal Segment Pair*). Con questo termine si definisce la coppia di segmenti di identica lunghezza (presenti nelle sequenze confrontate) avente il punteggio più elevato. In sostanza l'**MSP** è l'**HSP** a punteggio massimo.

N.B.: Nel caso dei programmi BLASTP e TBLASTN, viene effettuata una seconda scansione per trovare ulteriori HSP aventi cutoff superiore ad un valore **S2** ove S2 < S. Questo secondo passaggio dà l'opportunità di trovare **HSP** aventi un basso punteggio ma che, comunque, possono rivestire un'importanza biologica.

3. Estensione della ricerca delle zone di corrispondenza (hits)

Si prosegue l'estensione del segmento di omologia in entrambe le direzioni fino a che si raggiunge un abbassamento del punteggio di tale segmento al di sotto di un certo valore ottenibile con sequenze più corte. Una coppia di segmenti viene definita essere *localmente massimale* qualora sia una sua estensione che un suo accorciamento non ne migliorino il punteggio.

Valutazione statistica dei risultati

Il punteggio degli HSP gode della proprietà di poter essere analizzato statisticamente (Karlin & Altschul, 1990) [3]
A questo scopo viene utilizzata la **distribuzione di Poisson**. Nell'output di **BLAST** compare infatti un valore **P(N)** che rappresenta la probabilità che il punteggio di tali HSP definisca una similarità *casuale*. Tanto più piccolo è questo valore, tanto maggiore è la probabilità che **non** si tratti di pura casualità.

Peculiarità di BLAST

- Nel caso di confronti fra sequenze di nucleotidi, la lista delle ***n*** words da comporre, risulta molto semplice in quanto, per una sequenza di lunghezza *l*, costituita da words lunghe *w* nucleotidi, è uguale a:

  $$n = l - w + 1$$

  poichè il confronto con le matrici di sostituzione è del tipo *presente/assente*.

- BLAST si fa carico di controllare ed escludere le sequenze di nucleotidi che contengono:
  1. Sequenze ricche in *A+T*
  2. Sequenze ripetentisi *(es. Alu, Kpn)*

  mediante l'uso di opportuni programmi-filtro. Tali particolari sequenze potrebbero essere infatti fonte di problemi in quanto potrebbero essere erroneamente interpretate come sequenze omologhe.

DIFFERENZE FRA FASTA E BLAST

FASTA ricerca il migliore allineamento fra l'*intera sequenza* sottoposta ad indagine e il database di sequenze usato come riferimento.

BLAST usa inoltre una *scoring matrix* [4] [5] durante tutte le fasi della ricerca (scansione ed estensione), a differenza di FASTA che usa una *scoring matrix* solo durante la fase di estensione del confronto.
Inoltre, mentre FASTA esamina gli aminoacidi a coppie (*ktup=2*) o singolarmente presi (*ktup=1*), **BLAST** utilizza per il confronto gruppi di 3-4 aminoacidi (**words**) il che consente una *velocizzazione* del processo. Per far fronte alla riduzione di specificità derivante dall'uso di questi gruppi piuttosto "*ampi*", **BLAST** prende in considerazione solo quei gruppi di 3-4 aminoacidi il cui punteggio è superiore ad un *valore-soglia* **T (CUTOFF)**, in modo che l'eventuale omologia identificata possa considerarsi probabile (su base statistica) già *a priori*.
Così come prevede l'algoritmo che governa le prime fasi di FASTA, anche **BLAST** *non* ammette la presenza di *gap* all'interno di ciascun segmento di sequenza preso in considerazione.
A differenza di FASTA che nell'ultima fase prende in considerazione eventuali inserzioni e delezioni nei segmenti allineati, **BLAST** *non* contempla tale possibilità in nessuna fase.

	FASTA	**BLAST**
OMOLOGIA	Globale Locale (LFASTA)	Locale
USO DELLA SCORING MATRIX	Durante la 2ª fase (estensione)	Fase di scansione Fase di estensione
K-TUPLE	1-2 aa / 4-6 nt	3 aa / 11-12nt
GAP	Consentiti nella 4ª fase	Mai consentiti
VELOCITA'	Da 1/2 ad 1/5 di BLAST	Da 2 a 5 volte maggiore di FASTA
SPECIFICITA'	Migliore per il confronto di sequenza nucleotidiche	Migliore per il confronto di sequenze proteiche