Sommario:

Introduzione

L'omologia fra sequenze aminoacidiche e nucleotidiche può essere **globale** o **locale**.
Il programma **FASTA** [1], [2], [3], [4], [5] nella sua versione classica, è un programma euristico in grado di ricercare **omologie globali** di sequenza. Due sue varianti, LFASTA e PLFASTA, sono in grado di ricercare **omologie locali** di sequenza.
Il nome **FASTA** che sta per **FAST-All**, è la versione migliorata di **FASTP** e di **FASTN** [6]. Questa versione del programma introduce uno step intermedio fra la seconda e la terza fase di analisi utilizzate da FASTP e FASTN, per cui le fasi dell'analisi diventano quattro.

FASTA rappresenta uno dei diversi "sotto programmi" che sono stati creati e che sono sotto elencati:

FAST-A	Paragona una sequenza di aminoacidi o di nucleotidi con una banca dati di sequenze di aminoacidi o di nucleotidi, rispettivamente. Riconosce le sequenze di aminoacidi da quelle di nucleotidi sulla base di una percentuale maggiore dell'85% di A+C+G+T.
TFASTA	Paragona una sequenza ad una banca dati di DNA traducendo "al volo" le sequenze di DNA usando tutti 6 i possibili *frame* di lettura del DNA. Ha una maggiore sensibilità di FASTA anche se impiega lo stesso tempo.
LFASTA	Paragona due proteine o due sequenze di nucleotidi riguardo ad **omologie locali** e mostra gli allineamenti locali di sequenza.
PLFASTA	Paragona 2 sequenze di nucleotidi o di proteine per **omologie locali** e ne fornisce una visualizzazione grafica (su terminali grafici).

Basi concettuali del programma

Wilbur & Lipman (1983) [7] hanno descritto un algoritmo che permette rapide ricerche su database di proteine e di sequenze nucleotidiche, focalizzandosi solo su *gruppi di identità fra le sequenze*.
Nella prima fase di esecuzione del programma vengono volutamente omesse ricerche relative a mutazioni, delezioni od inserzioni di basi o aminoacidi.
Originariamente l'algoritmo di *programmazione dinamica* [8] descritto da Needleman & Wunsch (1970) prevedeva il confronto di ciascun aminoacido o nucleotide con tutti gli aminoacidi o nucleotidi disponibili nel database consultato. Per esempio, una sequenza di 200 aminoacidi doveva essere confrontata (nel giugno 1994) con 14.000.000 aminoacidi presenti nel database e ciò comportava la necessità di effettuare 200 x 14.000.000 = **28 x 10^8** confronti.
Il programma FASTP [6] sviluppato da Lipman & Pearson (1985) ha cercato di ovviare a ciò passando ad un sistema "*euristico*" di controllo dell'omologia fra sequenze.
Il sistema che si utilizza è quello noto con il nome di **"LOOKUP TABLE"** con il quale si allineano le sequenze in esame e se ne ricercano le regioni di identità.

Prima Fase: DEFINIZIONE DELL'OFFSET (HASHING)

Inizialmente si crea una tabella contenente tutte le posizioni per ciascun tipo di aminoacido (o nucleotide) nell'ambito di ciascuna delle sequenze presenti nel database. Per esempio, se esiste nel database una sequenza come quella sotto riportata:

Posizione	**1**	**2**	**3**	**4**	**5**	**6**	**7**
	-	**-**	**-**	**-**	**-**	**-**	**-**
Sequenza "A"	**F**	**L**	**W**	**R**	**T**	**W**	**S**

FASTA costruisce una tabella del tipo:

F = 1
L = 2
W = 3, 6
R = 4
T = 5
S = 7

Supponiamo di dover paragonare a tale sequenza, la nostra sequenza che è del tipo:

Posizione	**1**	**2**	**3**	**4**	**5**	**6**
	-	**-**	**-**	**-**	**-**	**-**
Sequenza "B"	**S**	**W**	**R**	**T**	**W**	**T**

Anche per questa sequenza viene costruita una tabella delle posizioni, analoga a quella precedente.

Essa sarà del tipo:

S = 1
W = 2, 5
R = 3
T = 4, 6

N.B.: La tabella posizionale può essere costruita tenendo conto della posizione degli aminoacidi singolarmente presi (**ktup=1**) (come è il caso sopra riportato) o presi a coppie (**ktup=2**). Usando questa seconda modalità si otterrà una *velocizzazione* del processo a scapito della precisione del dato finale. Ma l'approssimazione, ad ogni buon conto, è comunque valida in quanto si può supporre che le omologie tra due sequenze siano significative solo qualora si possano considerare coppie di aminoacidi e non singoli aminoacidi. Nel caso di sequenze nucleotidiche *ktup* vale **4** o **6**.

A questo punto si deve eseguire la **differenza** matematica dei valori posizionali degli aminoacidi dello stesso tipo tra la sequenza che viene paragonata (**query sequence**) e le sequenze presenti nel database. Questa differenza viene anche definita **OFFSET**.
In questo modo il numero dei confronti da eseguire diventa più simile alla *somma delle lunghezze delle sequenze* piuttosto che al loro prodotto (la qual cosa è invece tipica dell'uso degli algoritmi originali di programmazione dinamica).
Nell'esempio riportato, gli aminoacidi comuni alle due sequenze sono **S**, **W** (presente in due posizioni), **R** e **T** e il confronto dei valori posizionali di questi aminoacidi nelle due sequenze dà luogo a questi valori di OFFSET:

AA		Delta		OFFSET	
S	=	**1-7**	=	**-6**	
W	=	**3-2**	=	**1**	**oppure**
W	=	**3-5**	=	**-2**	
W	=	**6-5**	=	**1**	**oppure**
W	=	**6-2**	=	**4**	
R	=	**4-3**	=	**1**	
T	=	**5-4**	=	**1**	**oppure**
T	=	**5-6**	=	**-1**	

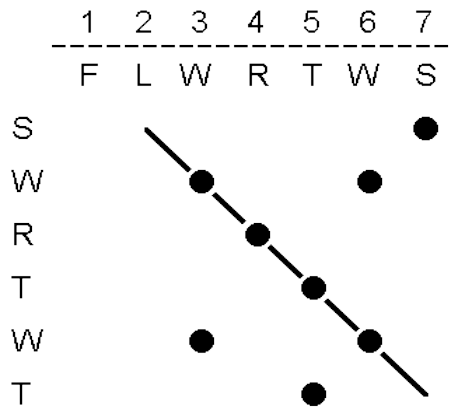

Illustrazione 1: LOOKUP TABLE

Si può quindi costruire una matrice nella quale porre le sequenze oggetto del confronto, cercando di individuare una o più diagonali che definiscano i segmenti di omologia tra le sequenze confrontate. Nell'esempio sovra riportato l'offset migliore è quello con valore **1** che consente l'allineamento di 4 aminoacidi. Gli altri offset consentono l'allineamento di un solo aminoacido o di nessun aminoacido. Il punteggio (*score*) in un offset viene aumentato per ogni identità e ridotto per ogni allineamento scorretto (*mismatch*). Quest'ultimo è consentito a differenza delle inserzioni/delezioni che invece sono proibite.

Utilizzando, assieme alla **Lookup Table**, il **metodo delle diagonali** che scandisce le sequenze da paragonare dall'inizio alla fine, si possono definire le *regioni locali di similarità* (*local region of similarity*) fra due sequenze. Esse sono quelle che hanno l'offset con il punteggio più elevato (*highest score*). Le 10 migliori regioni di similarità vengono "memorizzate" per ulteriori analisi.

N.B.: Il termine "*diagonale*" si riferisce alla linea diagonale che si determina quando, su una *dot-plot matrix*, una sequenza viene paragonata con se stessa.

Seconda Fase: VALUTAZIONE DELLA SOSTITUZIONE FRA AMINOACIDI

Un nuovo algoritmo valuta eventuali sostituzioni intervenute fra aminoacidi nelle 10 migliori regioni di similarità selezionate nella prima fase, utilizzando a questo scopo delle matrici di punteggio (**scoring matrix**) fra cui le più note e diffuse sono le **PAM** [9] e **BLOSUM** [10], e la più usata delle quali è la **PAM 250**. Tali matrici definiscono un punteggio per ogni possibile sostituzione di aminoacidi, basato sulla frequenza delle mutazioni intervenute nel corso del tempo. In particolare nel caso di matching di aminoacidi rari si dà un punteggio elevato, al contrario, nel caso di matching di aminoacidi frequentemente rappresentati nell'ambito delle proteine il punteggio è basso per la possibilità che l'omologia definita da un loro matching sia solo casuale.
In questo modo il punteggio che ne consegue tiene conto anche di questa evenienza.

Qualora non vi sia omologia, queste matrici definiscono dei punteggi diversi a seconda che il mismatch avvenga fra aminoacidi che hanno caratteristiche fisico-chimiche simili o tra aminoacidi completamente diversi. Per PAM 250 il punteggio minimo si ha nel caso della transizione *Trp* <--> *Cys* che ha un punteggio di **-8**.

A ciascuna delle 10 migliori regioni di similarità viene quindi assegnato un punteggio in base all'analisi effettuata con la *scoring matrix* e per ciascuna di esse si identificano quei residui che contribuiscono a definire il punteggio massimale. La subregione così definita viene denominata *regione iniziale* (**initial region**) e il suo punteggio è detto *punteggio iniziale* (**initial score**) o **INIT1**. Questo punteggio viene utilizzato come parametro per definire l'omologia fra le due sequenze (**similarity score**). Per calcolare il *punteggio iniziale* si può usare un *ktup = 1* o *2*. Come già detto il metodo che utilizza *ktup = 2* è molto più veloce (circa 5 volte) rispetto a quello che usa *ktup = 1*, ma fa aumentare anche il grado di imprecisione che, se risulta accettabile per proteine e sequenze nucleotidiche lunghe, è invece inaccettabile nel caso di oligonucleotidi o di oligopeptidi. La *regione iniziale* con il punteggio migliore viene utilizzata per creare una classifica (*rank*) delle sequenze di confronto presenti nella banca dati in modo di definire quali fra di esse sono le più omologhe alla sequenza in studio.

Terza Fase: COLLEGAMENTO DI DIVERSE REGIONI INIZIALI (JOINING)

FASTA effettua una valutazione relativamente alla possibilità di *collegare* (**join**) fra loro diverse *regioni iniziali*. I vincoli per la creazione del collegamento sono:

- Esclusione di eventuali aree di overlapping fra regioni
- Punteggio superiore ad un "*valore-soglia*"
- Introduzione di un punteggio di penalizzazione (**-16**) per ciascun gap (*gap penalty*)

Data la localizzazione delle *regioni iniziali* e i loro rispettivi punteggi, FASTA assegna un *punteggio di penalizzazione* alle regioni intermedie prive di omologia. Un algoritmo si occupa di valutare se la penalizzazione introdotta abbassa il punteggio al di sotto di un certo "*valore-soglia*". Se ciò non avviene, FASTA calcola un allineamento ottimale delle *regioni iniziali* definito dal collegamento (**join**) di più *regioni iniziali* a punteggio massimale. Conseguentemente viene ricalcolato anche il punteggio iniziale **INIT1** che viene ridefinito sulla base dei join creati e che viene denominato **INITN**. Viene anche rideterminato il *rank* delle sequenze di confronto presenti nel database.
Questo step aumenta la *sensibilità* del metodo a spese della *specificità*. La riduzione della specificità è in qualche modo compensata dall'inserimento nel *join* solo di quelle *regioni iniziali* che hanno un punteggio superiore ad un *valore-soglia* (**OPTCUT**) che è approssimativamente una *deviazione standard* al di sopra il punteggio medio che ci si aspetta per le sequenze non correlate presenti nel database.

Quarta Fase: VALUTAZIONE DELLE DELEZIONI E DELLE INSERZIONI

Le sequenze a maggiore omologia vengono allineate alla sequenza in esame utilizzando un procedimento che si basa su una modifica del metodo di *programmazione dinamica* di Needleman e Wunsch (allineamento globale) [8]. Esso tiene conto di possibili delezioni o inserzioni di aminoacidi (o nucleotidi). Nella variante del metodo proposto da Lipman e Pearson non si tiene conto delle parti di sequenza che non sono oggetto di omologia. Ciò consente di ottenere un **punteggio ottimizzato** (**OPT**). Il paragone finale considera tutti i possibili allineamenti della sequenza in esame trovati nella *seconda fase*, con le sequenze omologhe estratte dalla banca dati che cadono in un range di 32 aminoacidi, centrato attorno alla regione iniziale a più elevato punteggio. In questa fase si usa un *gap penalty* di **-12** per il primo residuo mancante e di **-4** per ogni ulteriore residuo.

Pertanto FASTA definisce 3 diversi punteggi:

1. **INIT1**: iniziale di vecchio tipo (dopo il confronto con le *scoring matrix*, nella seconda fase)
2. **INITN**: iniziale di nuovo tipo (dopo l'introduzione del **join**, nella terza fase)
3. **OPT**: ottimizzato (dopo la valutazione delle inserzioni e delezioni, nella quarta fase)

In conclusione il programma **FASTA è SPECIFICO ma non del tutto SENSIBILE**.

Valutazione statistica dei risultati: RDF & RDF2

La valutazione del significato delle omologie riscontrate (sono reali o no?) si fonda sull'utilizzo dell'analisi MonteCarlo che consente una valutazione statistica dei risultati ottenuti.
Si prende la sequenza di paragone (quella verso cui è stata trovata omologia) e la si permuta 100-200 volte in maniera randomizzata.
I programmi **RDF** e **RDF2** (versione migliorata di RDF) si occupano di effettuare tali permutazioni e di calcolare lo score iniziale ed ottimizzato per ciascuna coppia di sequenze da paragonare.
Viene calcolato un parametro **z** che è un indice della distribuzione statistica dei punteggi iniziali ed ottimizzati, in quanto tale distribuzione statistica non è normale (gaussiana).

Il parametro **z** è dato da:

$$z = \frac{X - Y}{W}$$

ove:
X = score di omologia delle sequenze non randomizzate
Y = media degli score nei confronti randomizzati
W = deviazione standard degli score nei confronti randomizzati

Per **z < 3** l'omologia NON è statisticamente significativa
Per **3 < z < 6** l'omologia è POSSIBILMENTE statisticamente significativa
Per **6 < z < 10** l'omologia è PROBABILMENTE statisticamente significativa
Per **z >10** l'omologia è statisticamente significativa

Una sequenza veramente e non casualmente omologa presenta un valore di **INITN** molto inferiore a quello iniziale, dopo la randomizzazione. Viceversa una sequenza solo casualmente omologa, dopo la randomizzazione ha il valore di **INITN** che rimane pressapoco costante.

DIFFERENZE FRA FASTA E BLAST

FASTA si differenzia da BLAST che è un altro programma euristico per il confronto fra sequenze comunemente usato, in quanto ricerca il migliore allineamento fra l'*intera sequenza* sottoposta ad indagine e il database di sequenze usato come riferimento a differenza di BLAST che invece ricerca solo **omologie locali** di sequenza.

FASTA inoltre usa una *scoring matrix* [9], [10] solamente durante la fase di estensione del confronto mentre BLAST usa una *scoring matrix* durante tutte le fasi della ricerca (scansione ed estensione).

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**).

Sia l'algoritmo che governa le prime fasi di **FASTA**, che quello che governa BLAST *non* ammetteno la presenza di *gap* all'interno di ciascun segmento di sequenza preso in considerazione. A differenza di BLAST però, l'algoritmo di FASTA contempla la possibilità di inserzioni e delezioni nell'ultima fase dell'allineamento.

	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 sequenze nucleotidiche	Migliore per il confronto di sequenze proteiche

© 1996 BioPD - CIV - Università di Padova - Autore: Leopoldo Saggin
Per messaggi: lsaggin@civ.bio.unipd.it - Versione **1.1** - Ultima Revisione: *8 luglio 1996*