RICERCA
in database di sequenze

TESTUALE

Ricerca dei record i cui campi soddisfano determinati criteri (hanno certi valori) = utilizzando parole-chiave e ricavando i nomi dei files che le contengono

SIMILARITÀ

Ricerca dei record che hanno le sequenze più "simili" ad una sequenza fornita come sonda/esca

QUERY

Acidi nucleici e proteine sono costituiti da sequenze lineari rispettivamente di nucleotidi e di aminoacidi; entrambi possono essere rappresentati da singole lettere. È quindi possibile rappresentare acidi nucleici e proteine come stringhe di lettere e perciò, usando programmi informatici, **trattarli come qualsiasi stringa di caratteri**. La stringa di caratteri è soltanto una rappresentazione semplificata del corrispondente acido nucleico o proteina.

Ricerca di SIMILARITÀ in database

Affinchè una sequenza risulti informativa dev'essere analizzata comparativamente al contenuto dei database. La comparazione e le informazioni che se ne ricavano permettono di formulare ipotesi sulle sue relazioni evolutive con sequenze simili o sulla sua funzione. In particolare il confronto permette:

- ✓ Identificazione di domini strutturali ➔ ipotesi funzionali
- ✓ Analisi molecolare comparata / Costruzione di alberi filogenetici
- ✓ Studiare l'evoluzione di specie o popolazioni
- ✓ Costruzione di modelli di struttura 3D per omologia
- ✓ Trovare le regione di sovrapposizione tra sequenze contigue
- ✓ Trovare la regione genomica codificante un trascritto
- ✓ Attribuzione di una possibile funzione a geni sconosciuti ➔ %

% ricerca di similarità in database

Il sequenziamento sistematico di interi organismi e di interi trascrittomi ha permesso di identificare migliaia di geni, molti dei quali codificano per proteine sconosciute.
L'analisi di similarità (e quindi l'allineamento) con proteine già note può fornire indicazioni sulla loro funzione.

un esempio ormai storico

Quello di *Saccharomyces cerevisiae* è considerato il "modello base" di genoma eucariotico.La sequenza genomica completa dei 16 cromosomi (12.052 kb) di S.cerevisiae è stata pubblicata nel 1997, risultato di una collaborazione scientifica internazionale basata su "nuove" metodologie biomolecolari, sul sequenziamento automatico e sulla bioinformatica.

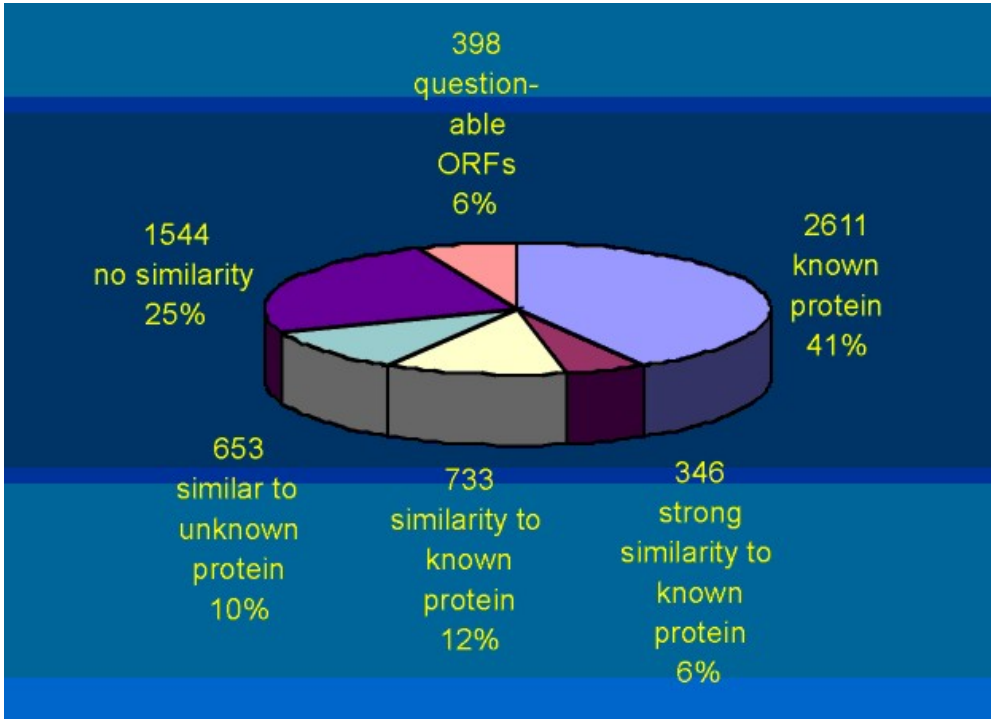
Overview of the yeast genome

H. W. Mewes, K. Albermann, M. Bähr, D. Frishman, A. Gleissner, J. Hani, K. Heumann, K. Kleine, A. Maierl, S. G. Oliver[1], F. Pfeiffer & A. Zollner

Max-Planck-Institut für Biochemie, D-82152 Martinsried, Germany
[1]University of Manchester Institute of Science And Technology (UMIST), Sackville Street, Manchester M60 1QD, UK

The collaboration of more than 600 scientists from over 100 laboratories to sequence the *Saccharomyces cerevisiae* genome was the largest decentralised experiment in modern molecular biology and resulted in a unique data resource representing the first complete set of genes from a eukaryotic organism. 12 million bases were sequenced in a truly international effort involving European, US, Canadian and Japanese laboratories. While the yeast genome represents only a small fraction of the information in today's public sequence databases, the complete, ordered and non-redundant sequence provides an invaluable resource for the detailed analysis of cellular gene function and genome architecture. In terms of throughput, completeness and information content, yeast has always been the lead eukaryotic organism in genomics; it is still the largest genome to be completely sequenced.

I programmi bioinformatici hanno individuato le ORFs maggiori di 99 aminoacidi.
Le ORFs <99 aa sono state individuate perché codificanti proteine note o per omologia con proteine note di altri organismi.
Complessivamente nel genoma nucleare sono state identificate 6275 ORFs, 87 delle quali <100 aa .



Una parentesi sulla terminologia …

Attenzione alla proprietà di linguaggio!

Similarità: é un dato che prescinde da eventuali ipotesi sulla causa della similarità stessa; "grado" di somiglianza tra sequenze, è un dato che si osserva e si può misurare.

≠

PUO' suggerire che le sequenze siano omologhe, cioe' evolutivamente correlate.

Omologia: due sequenze si dicono omologhe se condividono una stessa origine evolutiva. È una proprietà qualitativa, non si può misurare.

!!

~~Percentuale di omologia~~ → Percentuale di similarità

~~Ricerca di omologia~~ → Ricerca di similarità

... e ancora:

Strutture o sequenze **ortologhe** in due organismi sono sequenze omologhe che sono evolute dalla stessa caratteristica nel loro ultimo antenato comune ma che non necessariamente mantengono la loro funzione ancestrale.

Sequenze omologhe la cui evoluzione riflette invece eventi di duplicazione genica si definiscono **paraloghe**. Per esempio, la catena β dell' emoglobina e' un paralogo della catena α dell' emoglobina e della mioglobina, dal momento che ambedue si sono evolute dallo stesso gene ancestrale attraverso ripetuti eventi di duplicazione genica.

OMOLOGIA

ANTENATO COMUNE

ORTOLOGIA

PROCESSO DI SPECIAZIONE

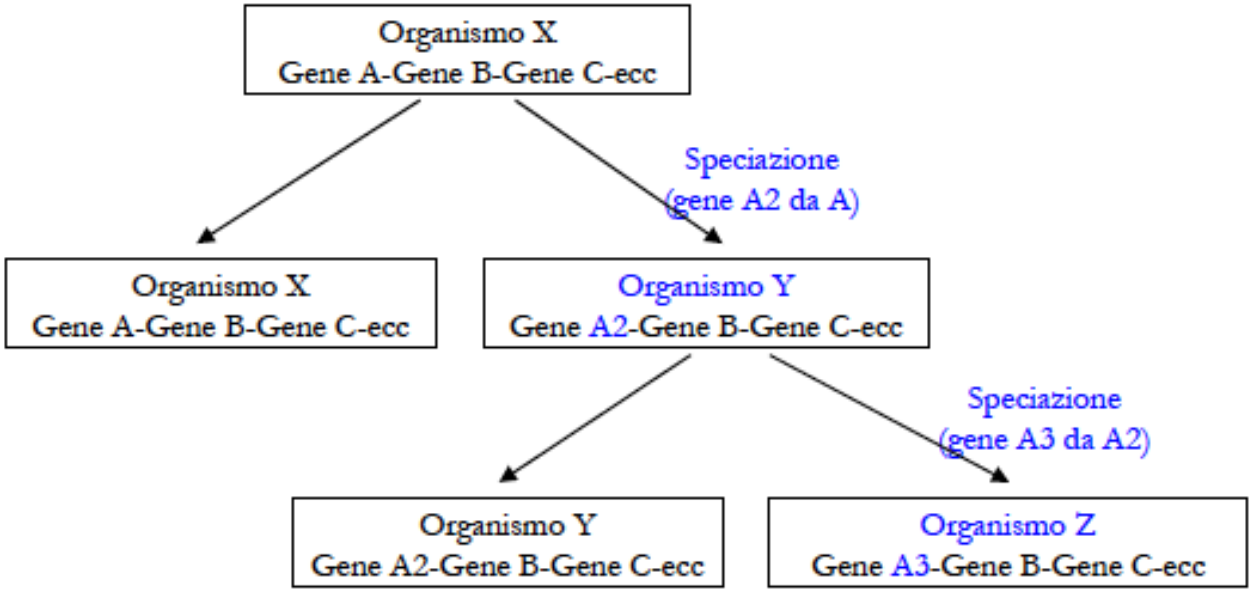
Descrive relazioni tra geni di una famiglia in diversi organismi

PARALOGIA

DUPLICAZIONE GENICA

Descrive relazioni tra geni di una famiglia intraorganismo

… e ancora: omologia ed evoluzione genica



Posto quindi che l'omologia presuppone l'esistenza di un **gene ancestrale** (un organismo ancestrale) da cui i geni omologhi (le strutture omologhe) si sono evoluti,

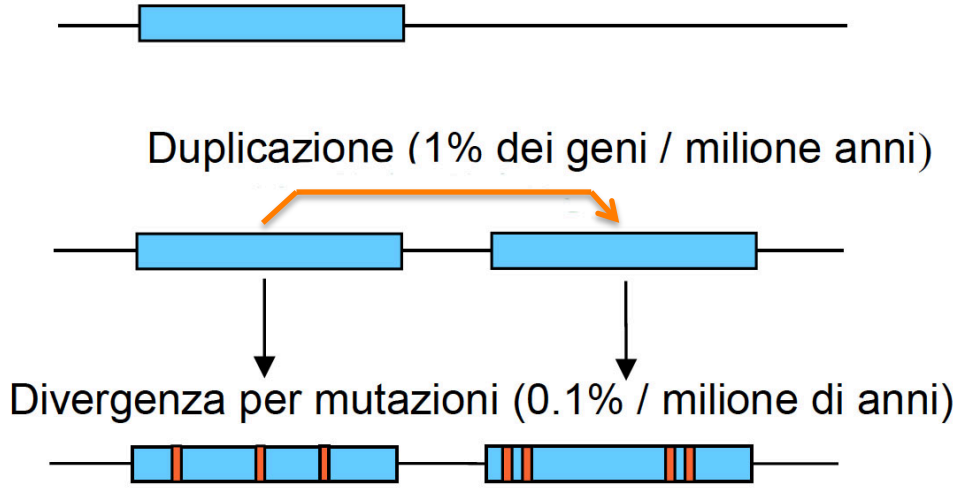
la **speciazione** (origine di una nuova specie da una già esistente) implica il cambiamento di alcune funzioni geniche le quali derivano dal 'cambiamento' dei rispettivi geni.

In figura: A è il gene ancestrale; A, A2, A3 sono geni omologhi.

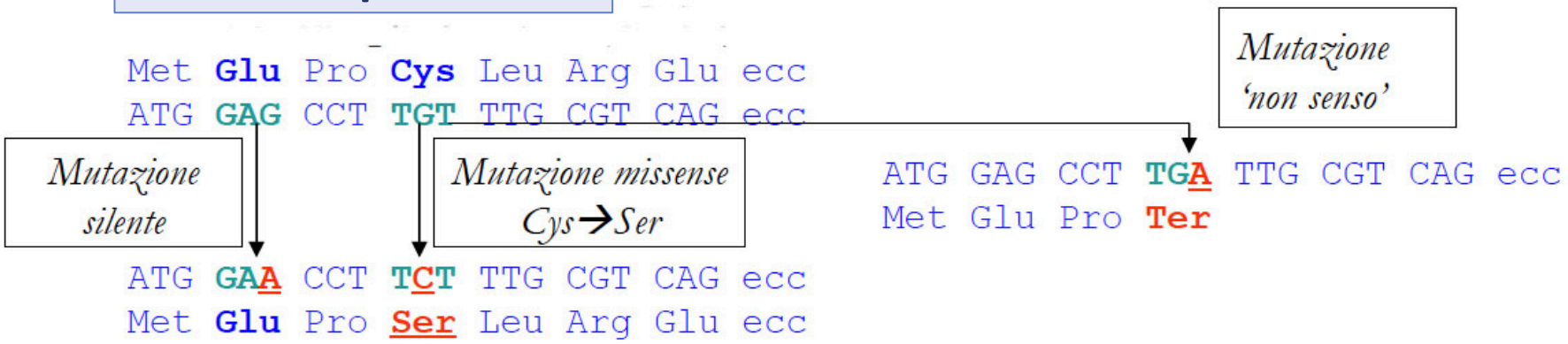
Come agisce l'evoluzione?

Principali motori responsabili dell'evoluzione genica

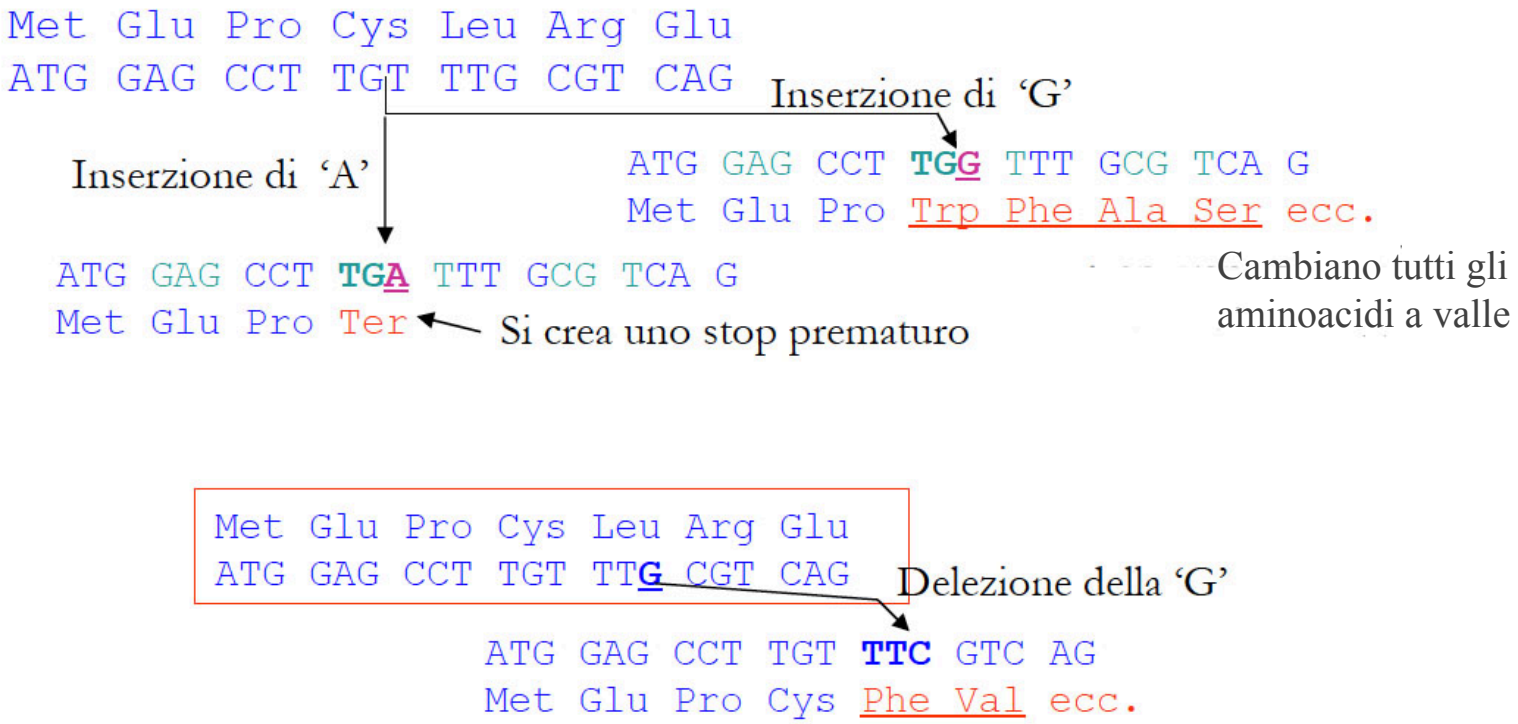
- La **duplicazione genica** è l'evento più frequente nell'evoluzione di nuovi geni o funzioni (Nelle cellule germinali, intere regioni genomiche possono essere duplicate. Inizialmente, l'organismo che si svilupperà, potrà avere due geni che producono la stessa proteina. A volte ciò può essere conveniente, spesso è letale.)
- Le **mutazioni del codice genetico** sono responsabili della divergenza (mutazioni puntiformi, inversioni, delezioni o inserzioni fanno variare il codice genetico dei singoli geni).



mutazioni puntiformi



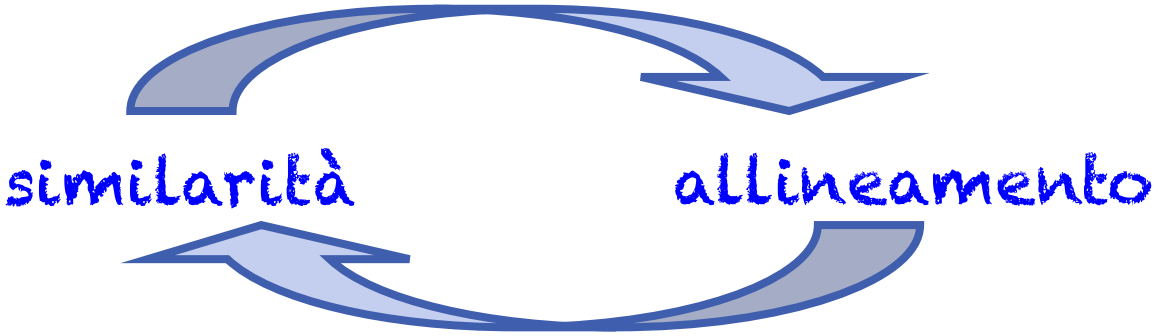
inserzioni o delezioni



Ricerca di similarità in database

Operativamente:

- Identificare all'interno di una banca dati di sequenze quelle sequenze che sono più simili ad una sequenza di nostro interesse.
- Allineare la sequenza di interesse ("**query**") a tutte le sequenze del DB (sequenze "**subject**") e individuare gli allineamenti migliori.



- Per **valutare la similarità** tra sequenze è necessario allinearle.
- Non è possibile allineare due sequenze senza **definire criteri si similarità**.

Ci occuperemo prevalentemente di COPPIE di sequenze, esaminando inizialmente i metodi di allineamento e poi i metodi più complessi di ricerca di similarità in banche dati.

Allineamento di due sequenze: allineamento pairwise

- La similarità tra due o più sequenza può essere definita in base a una funzione distanza: tanto più simili sono le sequenze, tanto meno distanti sono;
- Esistono diversi algoritmi di allineamento ciascuno dei quali definisce una funzione distanza;
- Dato un allineamento possiamo assegnare un punteggio, o **Score**, che indica il grado di similarità delle due sequenze.

- GLOBALE: Si cerca la corrispondenza ottimale tra tutti gli amminoacidi (o nucleotidi) di entrambe le sequenze.
- LOCALE: Si cerca di individuare regioni locali di similarità.

Globale

```
LTGARDWEDIPLWTDWDIEQESDFKTRAFGTANCHK
 ||.  | |  |  .|     .|  ||  || | ||
 TGIPLWTDWDLEQESDNSCNTDHYTREWGTMNAHKAG
```

Locale

```
LTGARDWEDIPLWTDWDIEQESDFKTRAFGTANCHK
         ||||||||.||||
       TGIPLWTDWDLEQESDNSCNTDHYTREWGTMNAHKAG
```

Due tipi di confronto che rispondono a domande diverse

- Qual è il miglior allineamento tra due intere sequenze?
- ✓ Risposta: Allineamento globale: vengono allineate le intere sequenze dall'N- al C-terminale (nel caso di proteine)
- Qual è la parte più somigliante tra due sequenze?
- ✓ Risposta: allineamento locale: vengono considerate solo le zone con la più alta densità di somiglianza; si possono avere uno o più suballineamenti

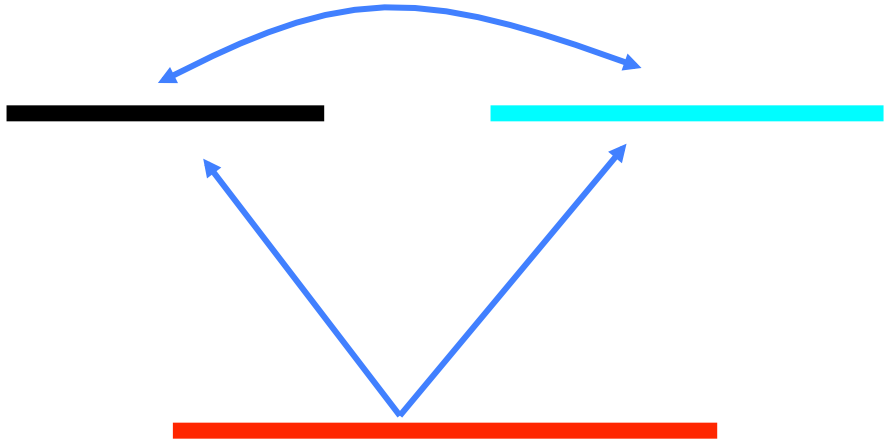
```
L G P S S K Q T G K G S — S R I W D N
L N - I T K S A G K G A I M R L G D A

- - - - - - - - T G K G - - - - - - - -
- - - - - - - - A G K G - - - - - - - -
```

NB: in generale, un allineamento locale non è necessariamente un sottoinsieme di un allineamento globale: è questa la cosa interessante!

Allineare due sequenze vuol dire:

- scriverle orizzontalmente in modo da avere il maggior numero di simboli identici o "simili" in registro verticale introducendo eventualmente intervalli (**gaps** = inserzioni/delezioni = **indels**)
- Una volta ottenuto l'allineamento, questo ci indica:
  - il grado di somiglianza complessivo tra le due sequenze
  - la posizione di queste somiglianze lungo la sequenza
- Possiamo pensare anche che, se le due sequenze sono omologhe, l'allineamento ci indichi il numero minimo di mutazioni che è necessario assumere per trasformare, durante l'evoluzione, una sequenza nell'altra



Allineare due sequenze vuol dire:

- scriverle orizzontalmente in modo da avere il maggior numero di simboli identici o "simili" in registro verticale introducendo eventualmente intervalli (**gaps** = inserzioni/delezioni = **indels**)
- Una volta ottenuto l'allineamento, questo ci indica:
  - il grado di somiglianza complessivo tra le due sequenze
  - la posizione di queste somiglianze lungo la sequenza
- Possiamo pensare anche che, se le due sequenze sono omologhe, l'allineamento ci indichi il numero minimo di mutazioni che è necessario assumere per trasformare, durante l'evoluzione, una sequenza nell'altra

In che modo si possono allineare due sequenze?

Le sequenze molto brevi e simili si possono allineare a occhio, ma non appena le sequenze diventano (realisticamente) più lunghe si presenta la necessità di utilizzare metodi automatici.

Tali metodi si possono ascrivere a tre gruppi:

① metodo della matrice a punti (**dot matrix**)

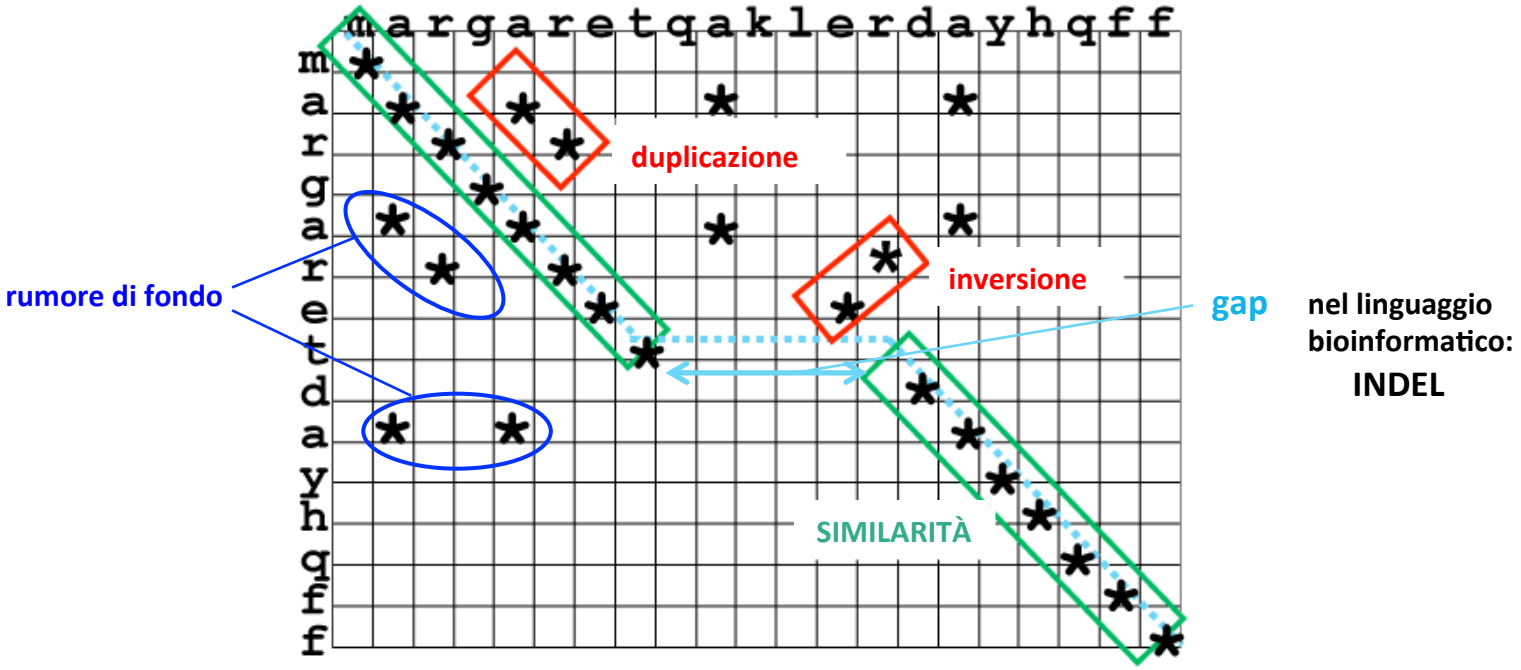
② programmazione dinamica (***dynamic programming***)

③ metodi basati su sistemi "euristici" (***Fasta, Blast***)

1 - Allineamento Pairwise mediante DOT MATRIX

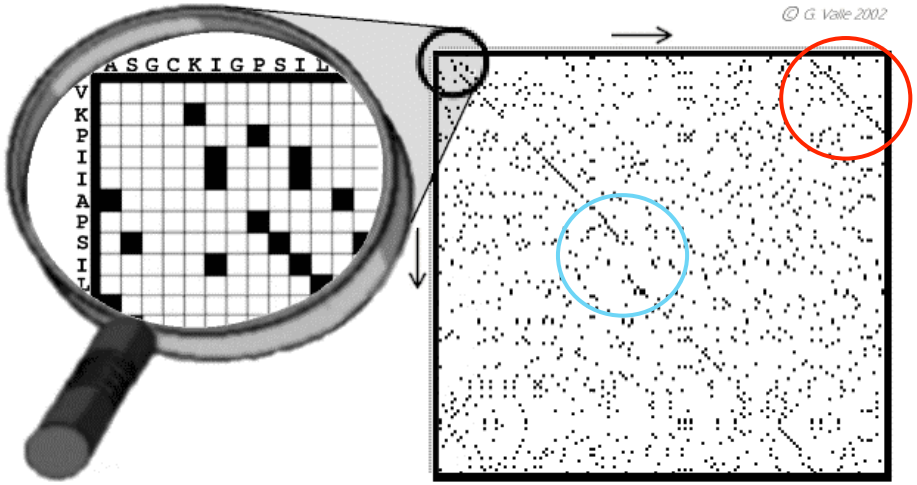
(o DOT PLOT, matrice a punti)

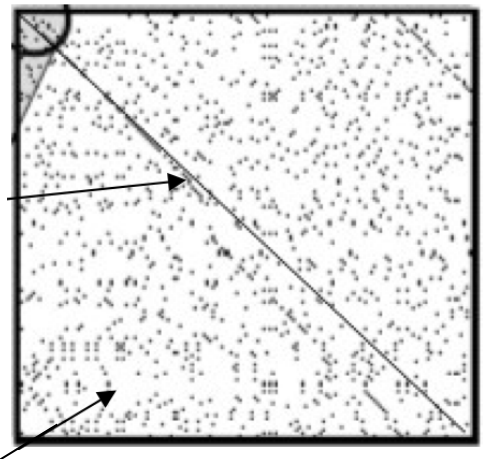
- Si crea una matrice in cui vengono confrontati tutti i possibili appaiamenti di ogni carattere delle due sequenze da allineare / stringhe di caratteri.
- Si riempie la matrice, annerendo le caselle che hanno nella corrispondente riga e colonna la stessa lettera.



- Il programma DOTLET (http://myhits.isb-sib.ch/cgi-bin/dotlet), date due sequenze in input permette di disegnare facilmente la relativa matrice Dot Plot.

Allineamento DOT MATRIX





Il rumore di fondo è più alto per INPUT nucleotidici che per input amminoacidici (una casella positiva ogni 4 vs 20).

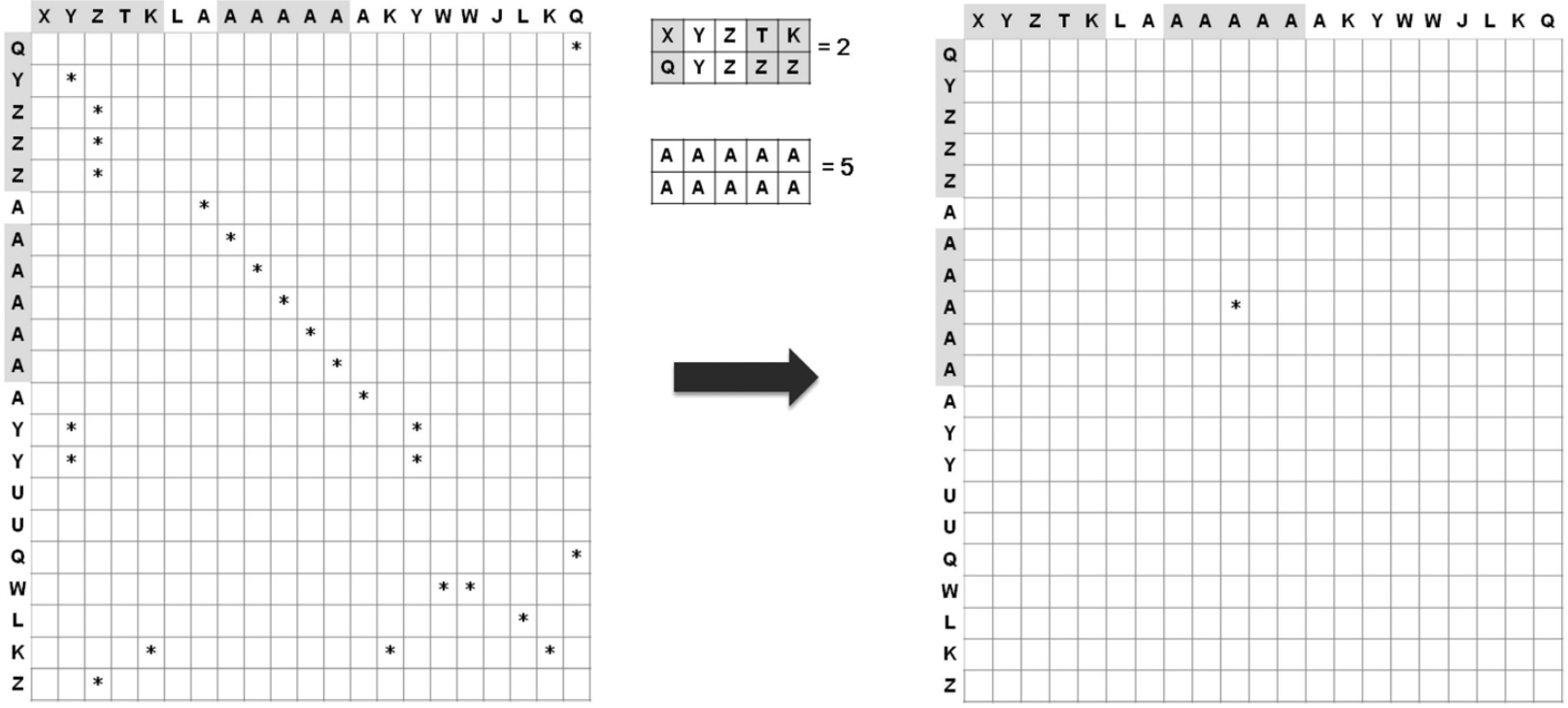
Con il metodo appena descritto le matrici tendo a essere molto "rumorose"

➔ Si possono applicare **filtri** per eliminare il rumore di fondo ed evidenziare il segnale significativo:

- ✓ Uso di finestre e di "stringenza" o "finestre scorrevoli"
- ✓ Uso di sistemi di punteggio più sofisticati, cioè di misura della somiglianza dei simboli confrontati.

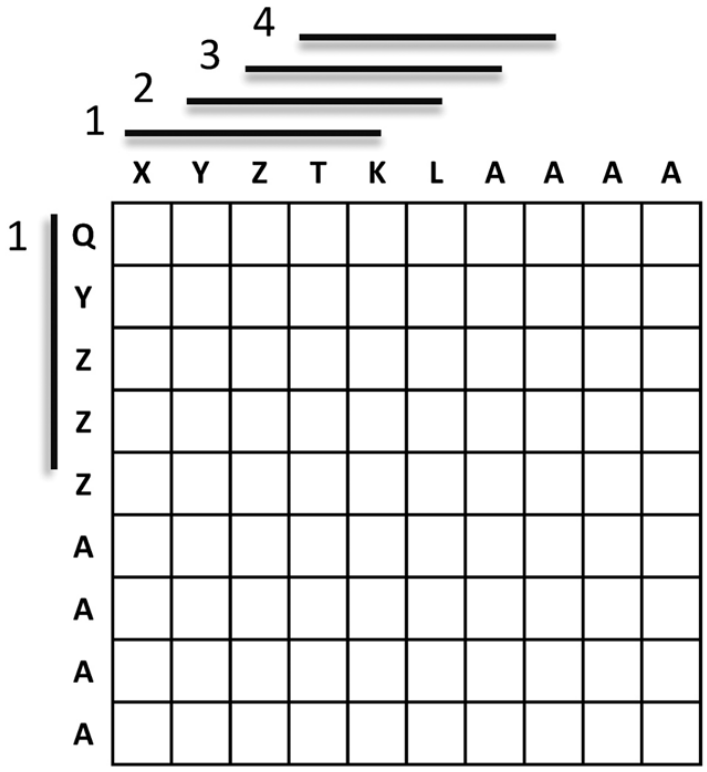
✓ Uso di finestre e di “stringenza” o “finestre scorrevoli”

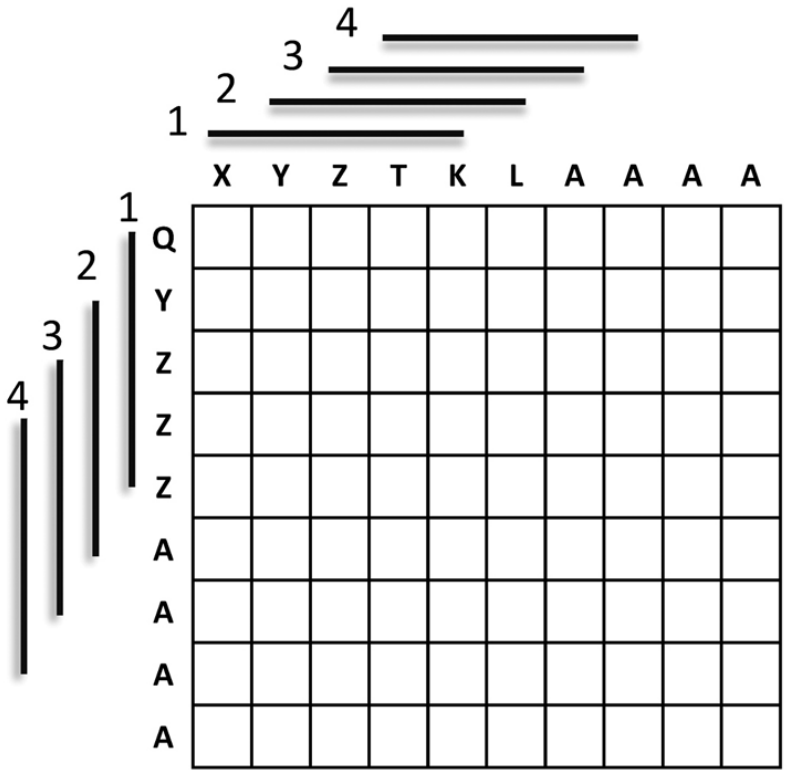
➔ Si considerano finestre di un numero ridotto di residui (5 ÷ 15) anziché i singoli residui. Ad es., usando una finestra di 5 residui, la casella corrispondente al centro delle due finestre viene riempita col simbolo “ ✖ “ solo se tutti i 5 residui sono identici.



✓ Uso di finestre e di “stringenza” o “finestre scorrevoli”

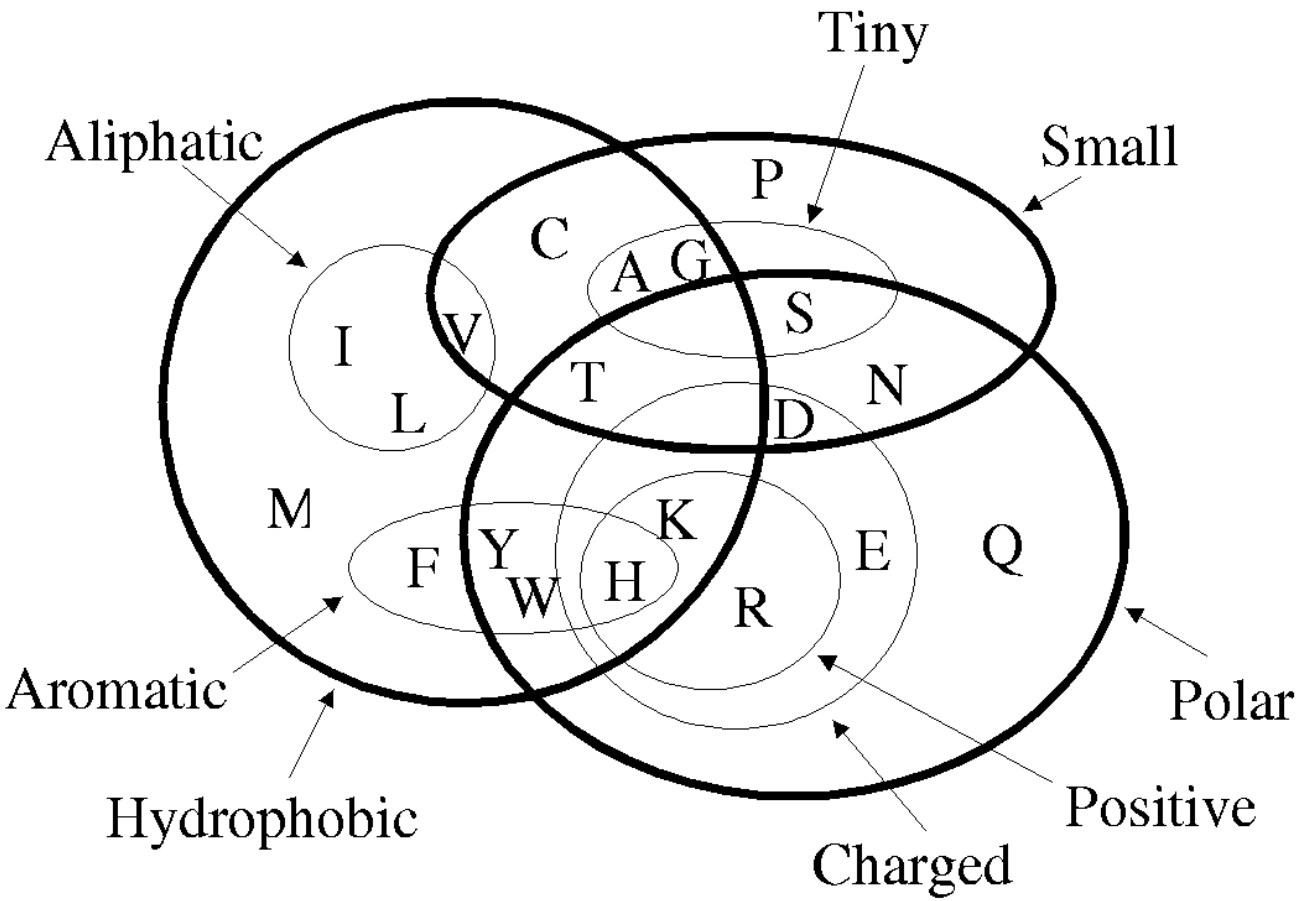
- ➔ Si considerano finestre di un numero ridotto di residui (5 ÷ 15) anziché i singoli residui. Ad es., usando una finestra di 5 residui, la casella corrispondente al centro delle due finestre viene riempita col simbolo “ ✖ “ solo se tutti i 5 residui sono identici.
- ➔ Si tiene fissa una finestra su una sequenza e si fa scorrere la finestra sull’altra…





✓ Un ulteriore raffinamento del filtraggio consiste nell'utilizzo di un sistema di "misura della somiglianza" tra i residui

➔ Per gli amminoacidi, ad es., si considerano le caratteristiche chimico-fisiche delle catene laterali



✓ Un ulteriore raffinamento del filtraggio consiste nell'utilizzo di un sistema di "misura della somiglianza" tra i residui

➔ Per gli amminoacidi, ad es., si considerano le caratteristiche chimico-fisiche delle catene laterali

➔ Si attribuisce un punteggio ad ogni possibile sostituzione, sicché al confronto tra due finestre si può assegnare un valore somma; solo se la somma è superiore ad un valore S (= soglia) prefissato si inserisce il simbolo " ✖ " nella casella corrispondente alla finestra

	A	R	N	K
A	5	-2	-1	-1
R	-1	7	-1	3
N	-1	-1	7	0
K	-1	-1	-1	6

Un esempio di matrice di punteggio

A K R A N R

K A A A N K

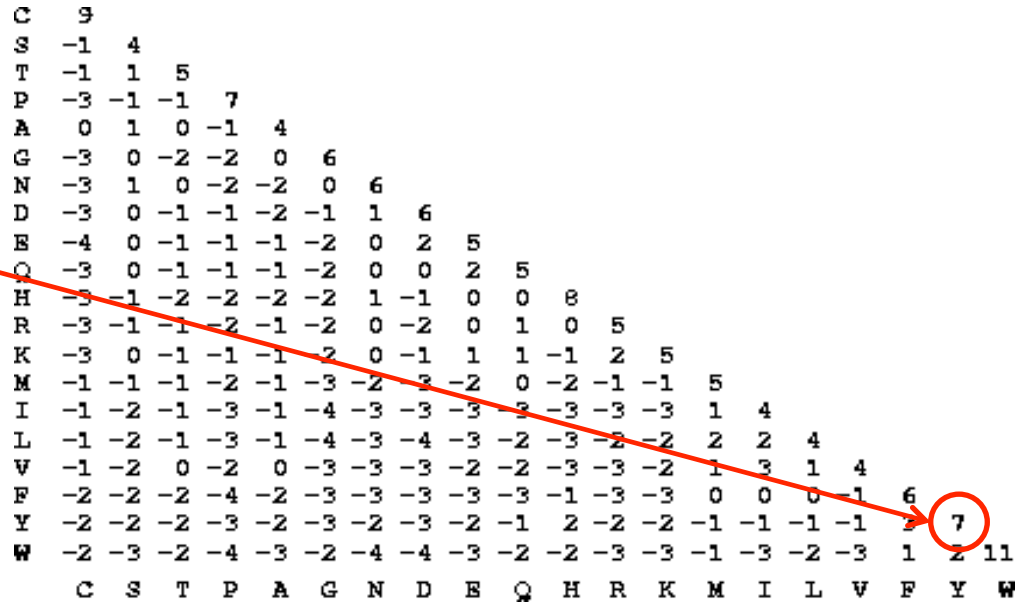
-1 +(-1) +(-2) + 5 + 7 + 3 = 11

allineamento pairwise – matrici di *score*

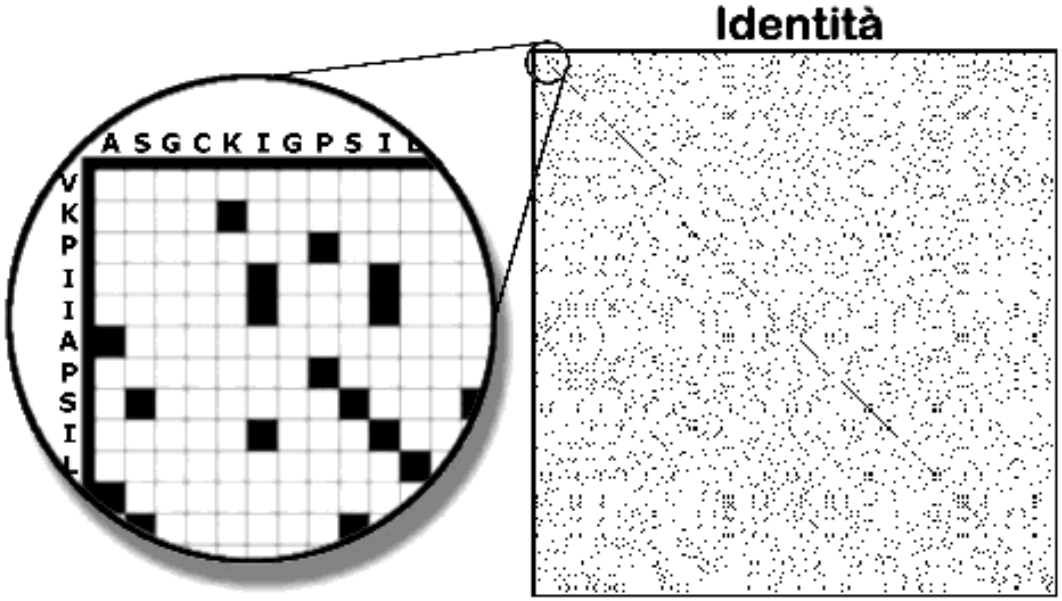
- Un modo per definire il punteggio o ***score*** è ad esempio quello di assegnare 1 in caso di caratteri uguali e 0 altrimenti.
- Nel caso di Nucleotidi questa definizione può andare bene;
- Nel caso di aminoacidi si introducono le matrici di score che assegnano ad ogni coppia di amino acidi un punteggio:
  - **Matrici PAM (Percent Accepted Mutations):** si basano su calcoli statistici;
  - **Matrici BLOSUM (Blocks Substitution Matrix) :** sono invece basate su una banca dati (BLOCKS) di allineamenti multipli di segmenti proteici.

Score allineamento: 15						
Seq1	V	D	S	-	C	Y
Seq2	V	E	S	L	C	Y
Score	4	2	4	-11	9	7

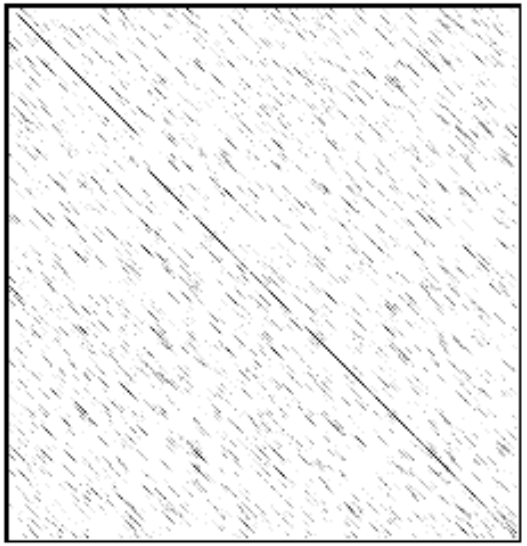
Penalità del gap



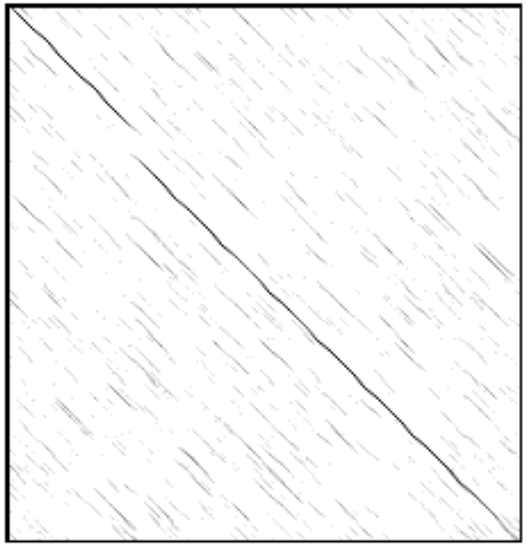
blosum62



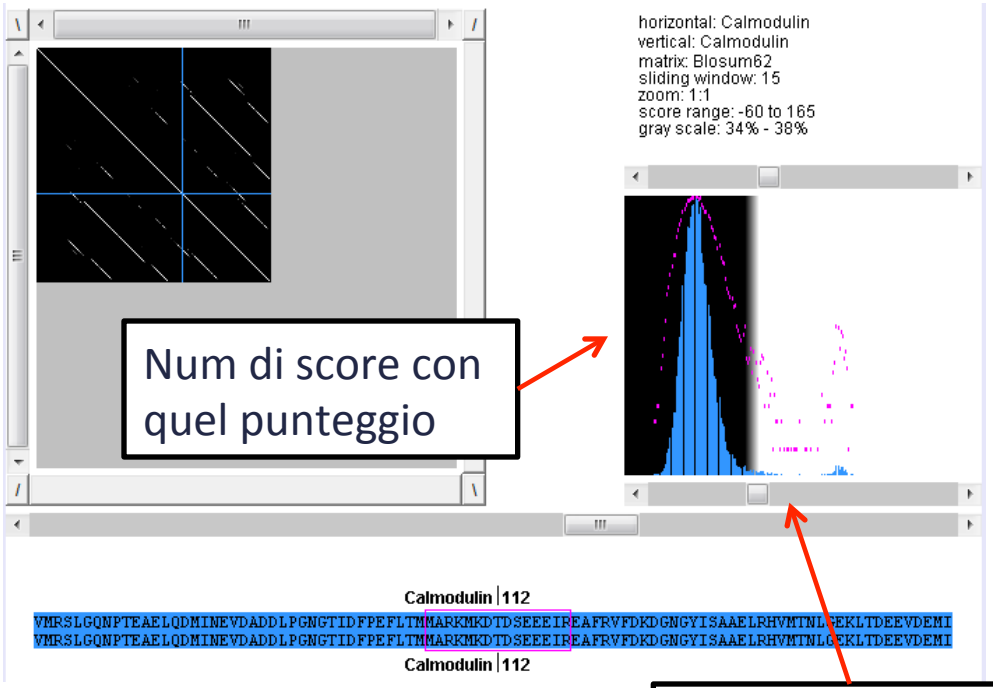
Blosum 62 - Window 5



Blosum 62 - Window 15

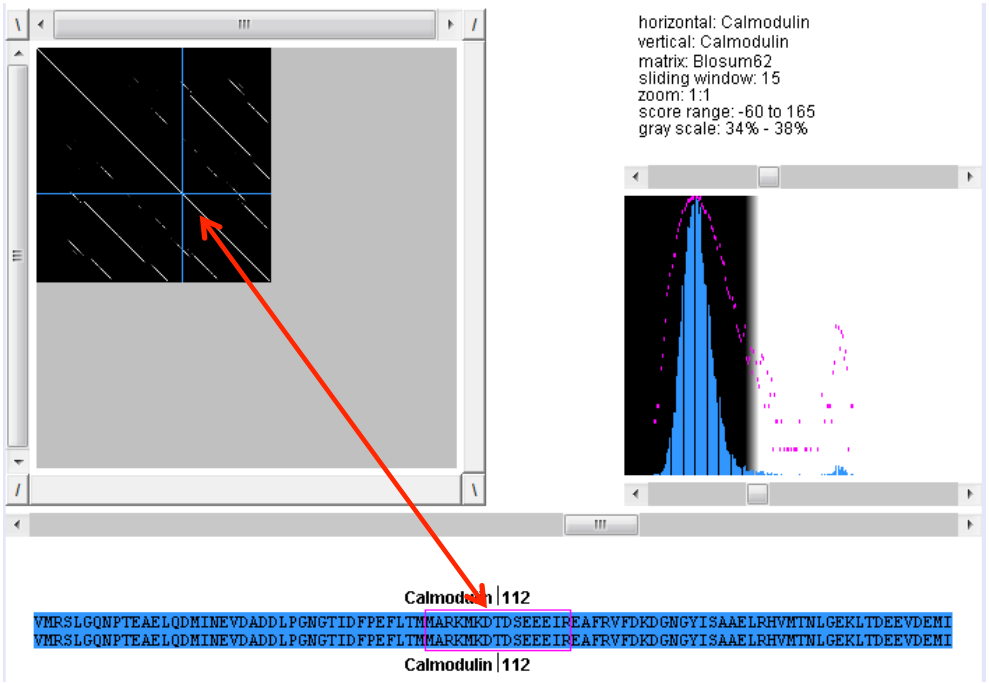


DotLet - http://myhits.isb-sib.ch/cgi-bin/dotlet



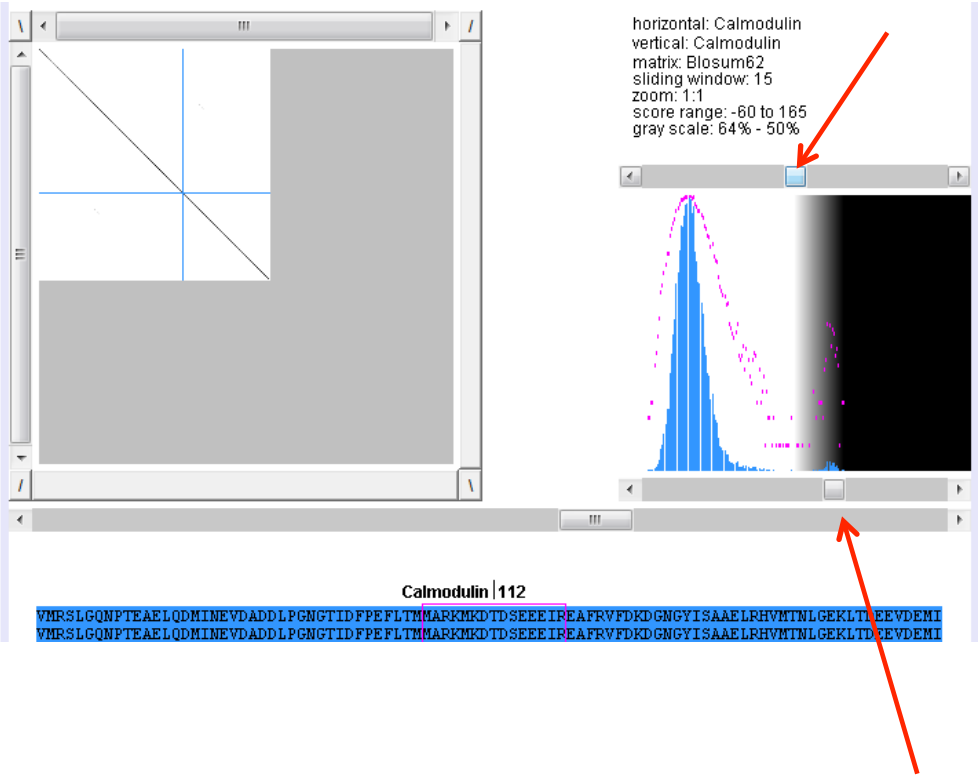
- Il grafico riporta la **distribuzione degli score** ottenuti da tutte le coppie di finestre di sequenza confrontate (usando le matrici di score).
- Si noti che la maggior parte dei punteggi ricade nella distribuzione a sinistra a basso punteggio, mentre una piccola popolazione a punteggio elevato si trova a destra.
- Spostando i cursori si variano i punteggi limite al di sotto dei quali la cella assume il colore nero e al di sopra il colore bianco. Tra i due limiti le celle assumono un tono di grigio proporzionale al punteggio che contengono.

DotLet - http://myhits.isb-sib.ch/cgi-bin/dotlet



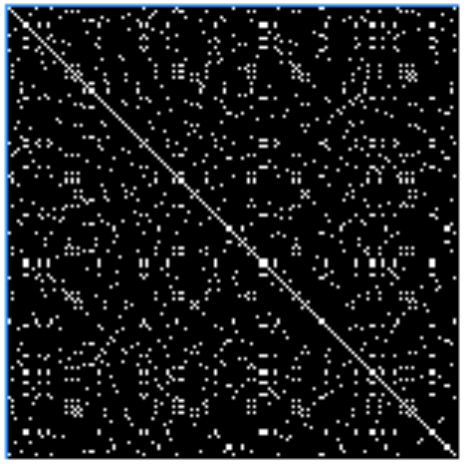
- Cliccando sulla matrice si attiva un reticolo che può essere spostato sulla superficie della matrice stessa con il puntatore del mouse;
- In basso viene riportato l'allineamento tra i due segmenti della proteina corrispondenti alla posizione del centro del reticolo sulla matrice;

DotLet - http://myhits.isb-sib.ch/cgi-bin/dotlet

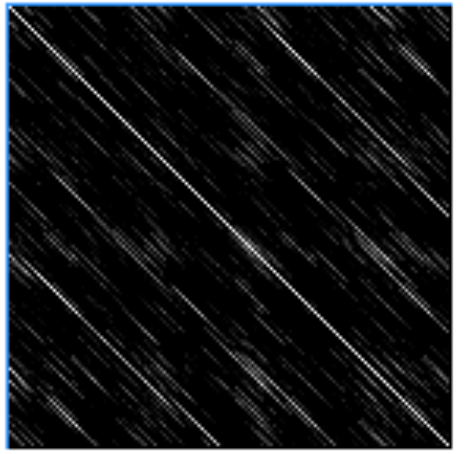


- Spostando i cursori in modo da posizionarci sulla piccola distribuzione a destra a punteggio elevato verranno visualizzati solo i punteggi elevati che ovviamente corrispondono alla diagonale principale;

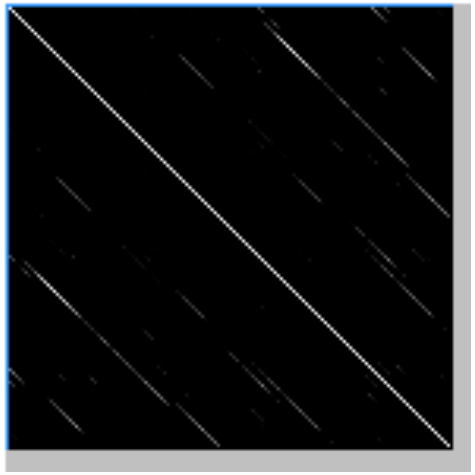
Singolo residuo

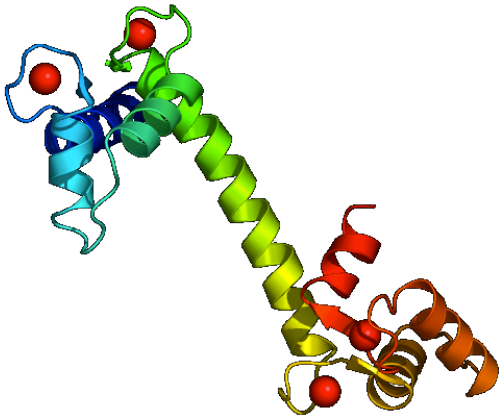


Finestra di 11 residui con soglia bassa e matrice Blosum62



Finestra di 11 residui con soglia alta e matrice Blosum62





Autoconfronto della calmodulina umana

2 - Allineamento Pairwise mediante Algoritmi Dinamici

- Forniscono l'allineamento *ottimale* tra due sequenze includendo anche le **inserzioni** e le **delezioni** che nella matrice a punti non sono considerate.

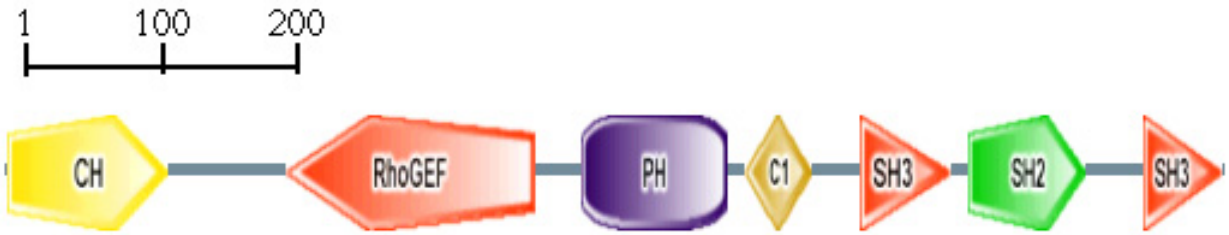
- Con semplici variazioni dell'algoritmo di base si possono produrre allineamenti globali o locali ma l'allineamento calcolato dipende dalla scelta di **alcuni parametri** iniziali da parte dell'utente.

- Innanzitutto ...

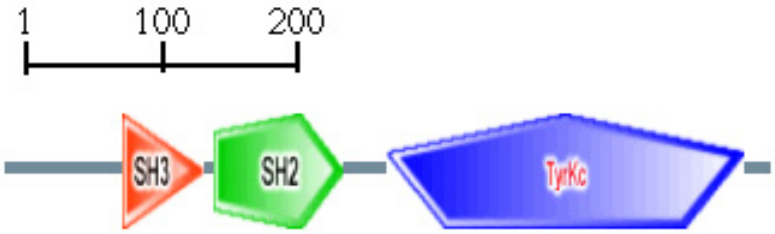
... siamo interessati ad un allineamento
globale o locale?

- similarità locali servono a identificare proteine anche diverse, ma che contengono lo stesso dominio

Protein Q8UUX6
Description GDP/GTP EXCHANGE FACTOR VAV3.
Species Gallus gallus

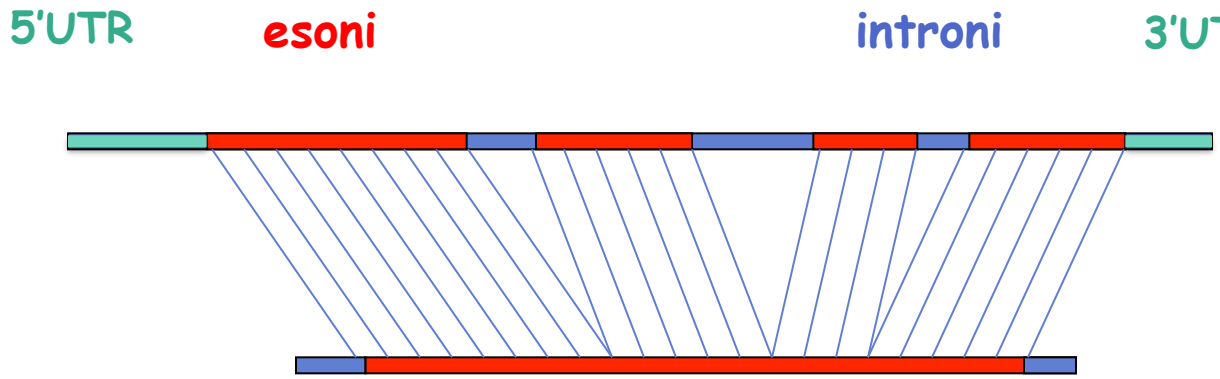


Protein SRC_CHICK
Description Proto-oncogene tyrosine-protein kinase SRC (EC 2.7.1.112) (P60-SRC) (C-SRC).
Species Gallus gallus



Allineamento globale o locale?

- a livello di DNA, troviamo regioni con similarità locali che riflettono situazioni interessanti: ad esempio introni/esoni, inserzioni/delezioni, transposoni, regioni promotore...



- gli allineamenti globali possono comunque essere utilizzati per confrontare accuratamente due sequenze la cui similarità sia estesa per tutta la lunghezza.

Allineamento mediante Algoritmi Dinamici

- ✓ Abbiamo definito un nuovo schema di punteggi per la valutazione della similarità tra due sequenze;
- ✓ una matrice di sostituzione per valutare l'appaiamento tra qualsiasi coppia di residui;
- ✓ possiamo associare un punteggio di penalizzazione (**gap penalty**, es. –1) per ogni gap aggiunto all'allineamento

```
IPLMTRWDQEQESDFGHKLP-IYTREWCTRG
||||||||| |||||||||| ||||||
CHKIPLMTRWDQ-QESDFGHKLPVIYTREW
```

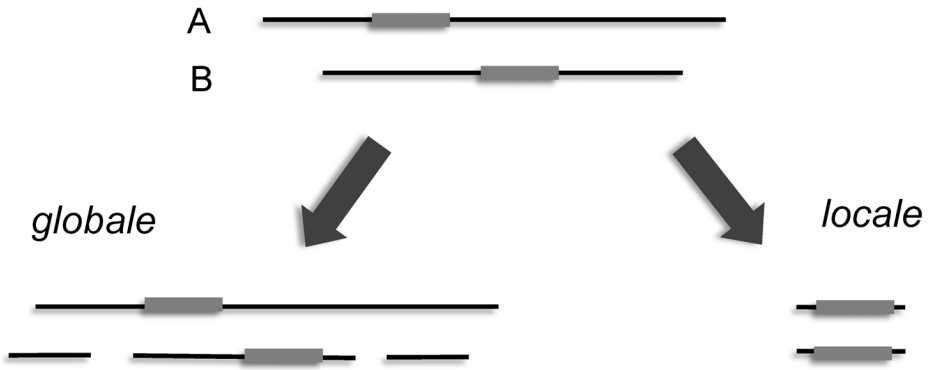
- ✓ o attribuire un punteggio di penalizzazione diverso per l'apertura di un gap nell'allineamento o per il suo allungamento (**gap extension penalty**, es. –0.1 per ogni ins/del successiva alla prima).

```
IPLMTRWDQEQESDFGHKLP----IYTREWCTRG
||||||||| |||||||||||    ||||||
CHKIPLMTRWDQ-QESDFGHKLPVGSSIYTREW
```

Allineamento mediante Algoritmi Dinamici

Algoritmi di allineamento che utilizzano una tecnica di programmazione dinamica:

- Needleman e Wunsch (1970)
  - ➔ allineamento globale
- Smith e Waterman (1981)
  - ➔ allineamento locale



- Entrambi i metodi permettono di determinare l'allineamento ottimale mediante un'**interpretazione computazionale della matrice dotplot**.
- L'allineamento ottimale viene **calcolato iterativamente** per sottosequenze via via piu' lunghe, cosa possibile in virtu' dell'indipendenza e dell'additivita' dei punteggi.
- Entrambi considerano matematicamente il principio della **massima parsimonia**: individuano l'allineamento che implica il minor numero possibile di mutazioni necessarie per trasformare una sequenza nell'altra o, in altre parole, che assume il più breve percorso evolutivo.

3 – Ricerca di similarità in banche dati

mediante sistemi "euristici": FASTA e BLAST

Quando si deve effettuare una ricerca per similarità di sequenza in una banca dati, l'operazione di confronto tra due sequenze deve inoltre essere ripetuta per ogni coppia di sequenze:

1. sequenza in input (**query sequence**)
2. **ognuna delle sequenze della banca dati**

Gli algoritmi descritti fin qui effettuano delle ricerche esaustive ed esplorano tutto lo spazio degli allineamenti possibili.

Si tratta comunque di algoritmi di ordine n^2, ovvero per allineare due sequenze lunghe ognuna 1000 residui, effettuano 1000x1000 = un milione di confronti.

L'algoritmo di Smith-Waterman impiega da centesimi a decimi di secondo per ogni allineamento. Se il DB contiene milioni di sequenze saranno necessarie decine o centinaia di migliaia di secondi, cioè alcuni giorni!!!

Per effettuare ricerche di similarità in banche dati, c'è comunque necessità di algoritmi più veloci

3 – Ricerca di similarità in banche dati

mediante sistemi "euristici": FASTA e BLAST

Programmi come FASTA e BLAST sono in grado di effettuare velocemente ricerche di similarità, grazie a soluzioni euristiche che sono basate su assunzioni non certe, ma estremamente probabili.

In pratica la ricerca è resa più veloce a scapito della certezza di avere veramente trovato la soluzione migliore.

FASTA velocizza la ricerca utilizzando un strategia di "**indicizzazione delle parole**": la sequenza *query* viene spezzetata in **parole** di 2 o 3 amminoacidi o 6 nucleotidi.

La lunghezza delle parole è definita "**k-tuple**" ("**ktup**").

Ad esempio se **ktup = 2** il numero di parole amminoacidiche è 20x20 cioè 400.
Se **ktup = 3** il numero di parole amminoacidiche è 20x20x20 cioè 8000.
Il programma crea un indice con tutte le "parole" contenute nella sequenza *query*.

DVVHKILLAPERDDKVLAFV

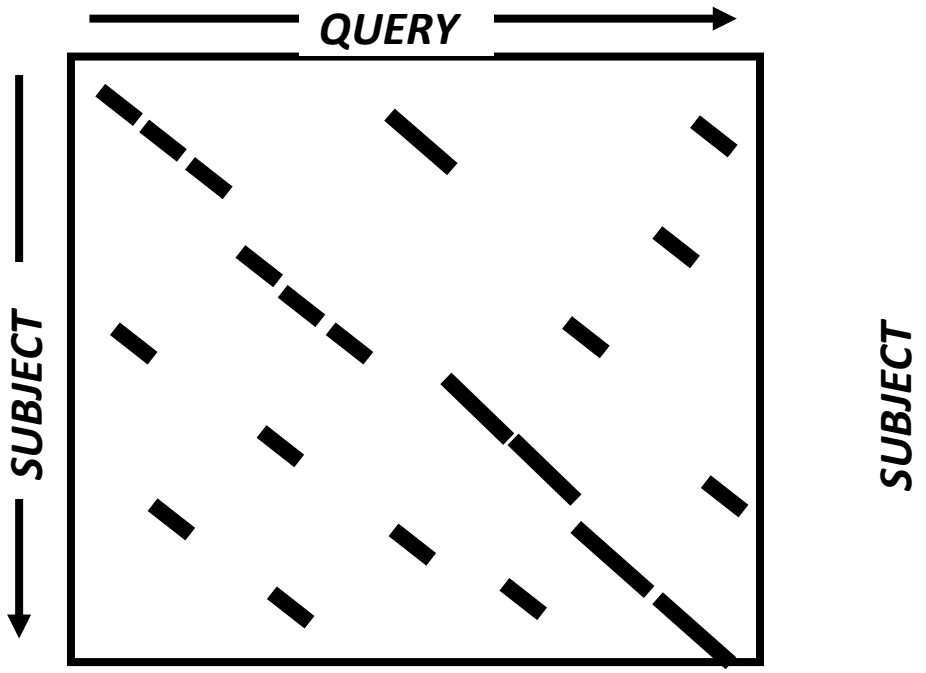
L'analisi di FASTA quindi è approssimata:

Se il valore di ktup è elevato l'analisi è più veloce ma alcune sequenze omologhe possono essere perse.
Ad esempio se ktup = 3 proteine omologhe alla sequenza *query* ma che non hanno conservato alcuni tripeptidi identici a quelli nella *query* verranno scartate fin dai primi stadi.
Pertanto l'analisi è più precisa usando **ktup = 2** o addirittura **ktup = 1** se la versione del programma lo consente.

L'allineamento fornito da FASTA non è il migliore in assoluto perché l'algoritmo di Smith Waterman viene applicato solo ad una fascia della matrice dot plot

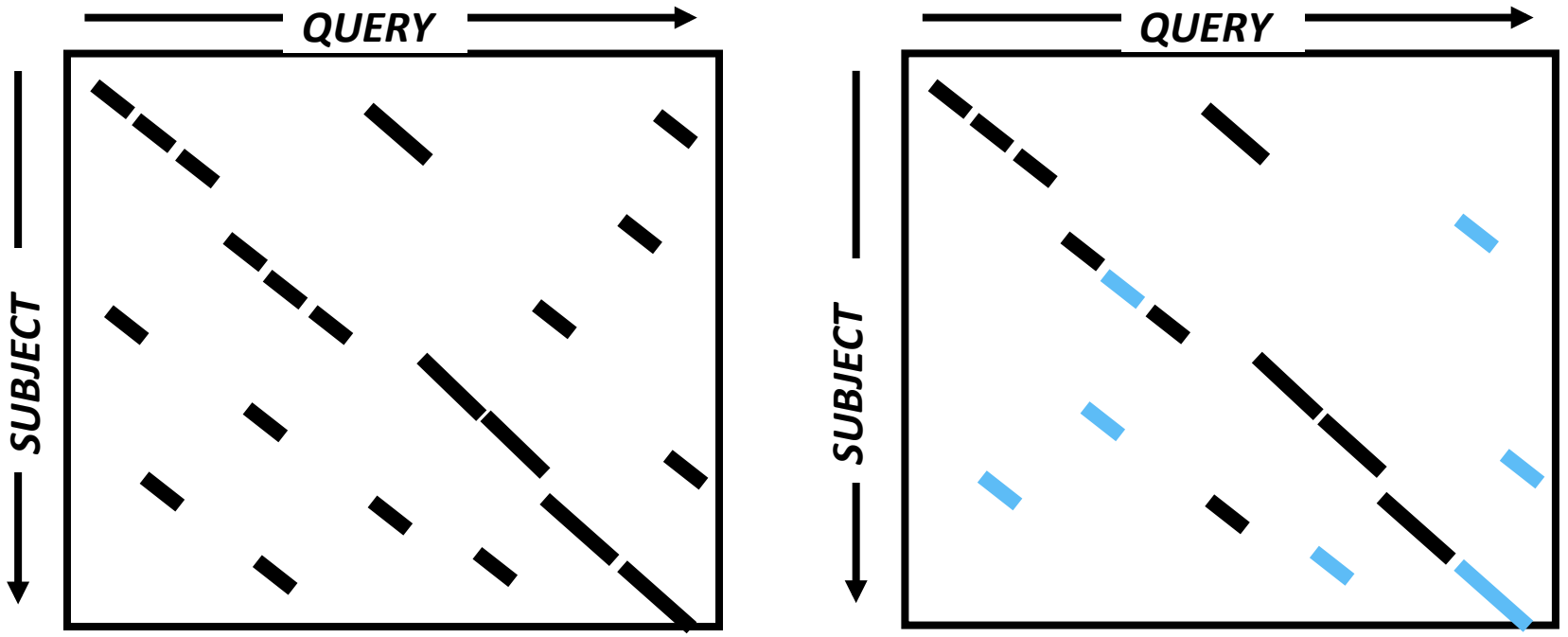
FASTA seleziona dal database solo le sequenze che contengono parole comprese nell'indice creato a partire dalla *query*.

Quindi vengono create matrici dot plot di identità



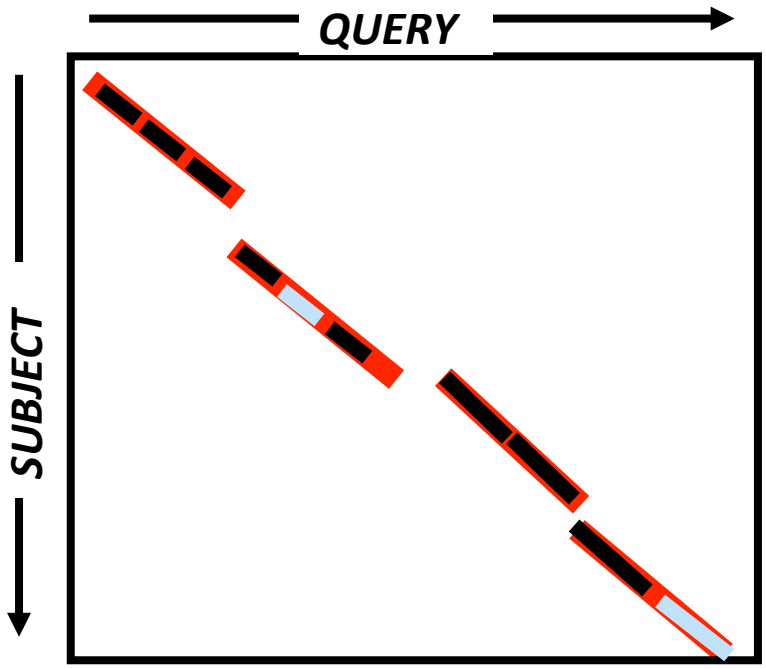
FASTA seleziona dal database solo le sequenze che contengono parole comprese nell'**indice creato a partire dalla *query***.

Quindi vengono create **matrici dot plot di identità**



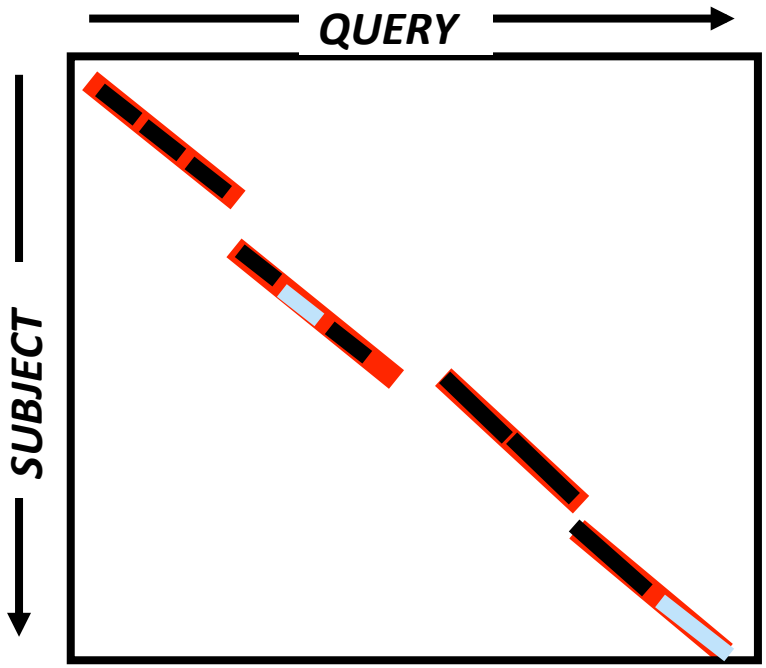
Poi seleziona le diagonali con più parole identiche quindi per queste diagonali calcola i punteggi utilizzando le **matrici PAM e BLOSUM**.
Vengono tenute solo le sequenze che danno i punteggi più alti.

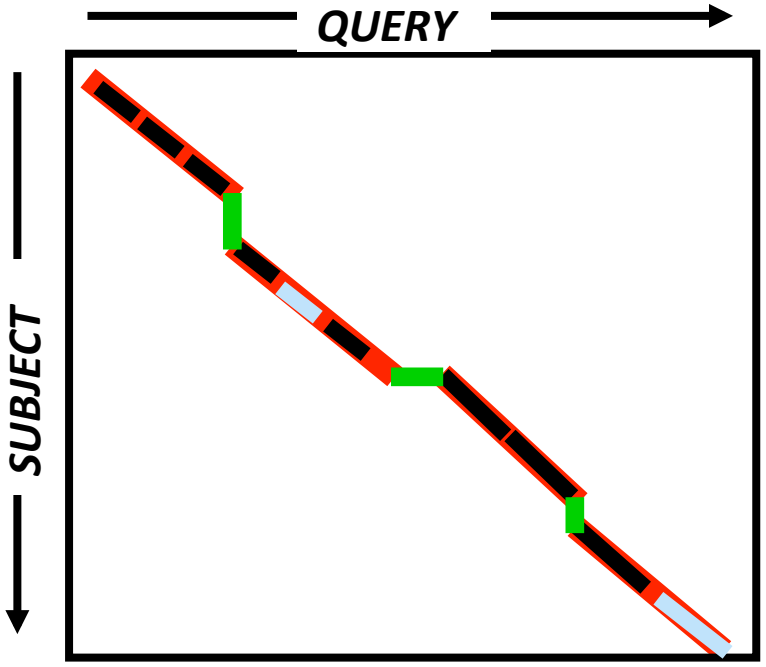
FASTA allunga le regioni in diagonale congiungendo le parole che si trovano sulla **stessa diagonale**.



FASTA allunga le regioni in diagonale congiungendo le parole che si trovano sulla **stessa diagonale**.

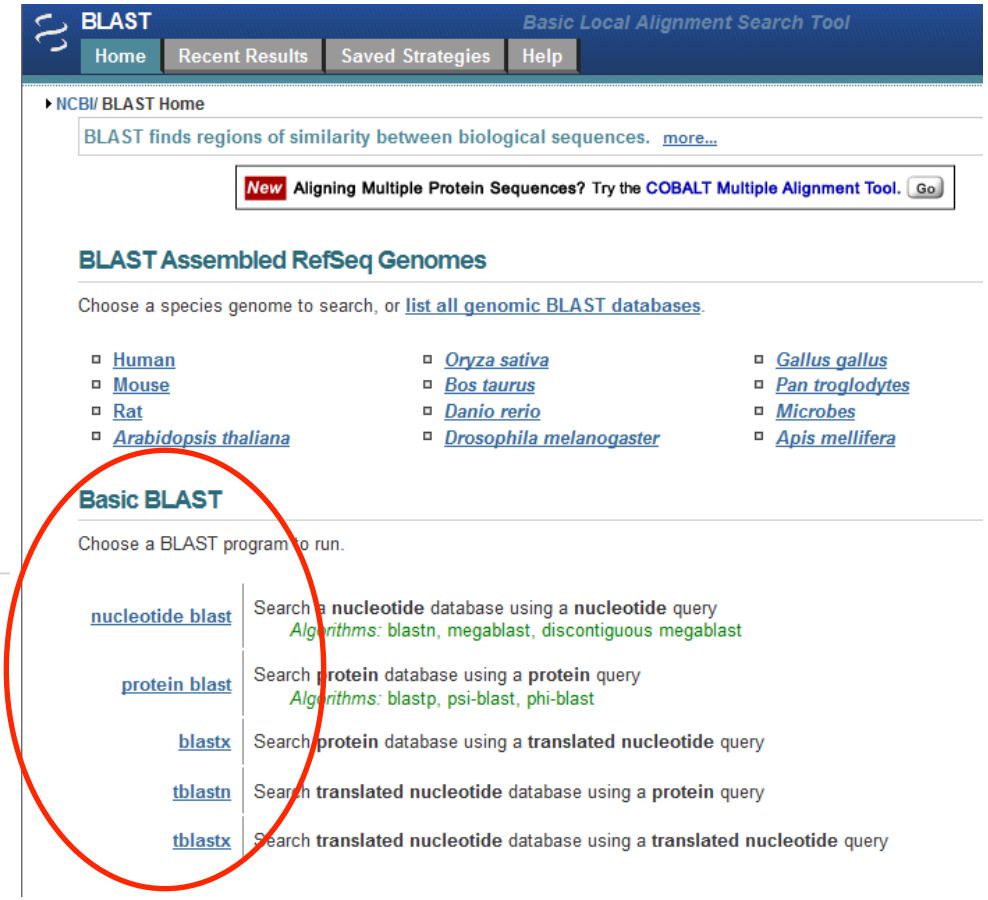
Congiunge le frammenti che si trovano su diagonali diverse mediante **gap** e che possono essere congiunti entro una soglia di accettabilità.

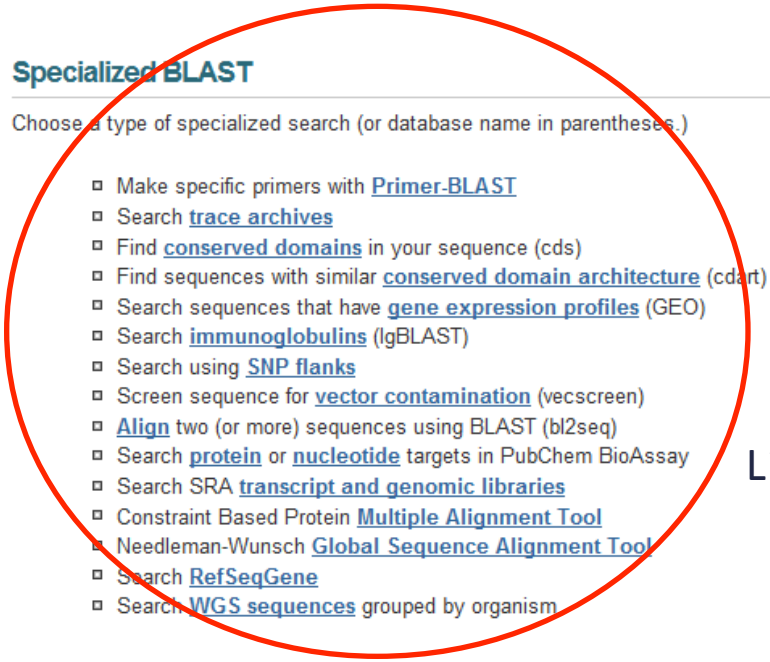




BLAST - http://www.ncbi.nlm.nih.gov/BLAST

BLAST =
(Basic Local Alignment Search Tool)
Permette di ricercare regioni di similarità locale tra una sequenza data e una collezione di sequenze in banca dati.





L'idea di base dell'algoritmo consiste nel procedere ad allineare passo dopo passo piccole sequenze (WORD e KTUPLE) e tentando di estendere poi l'allineamento.

BLAST - http://www.ncbi.nlm.nih.gov/BLAST

BLAST NUCLEOTIDICO

MEGABLAST
E' utilizzato per trovare efficientemente lunghi allineamenti tra sequenze molto simili tra loro o per identificare una sequenza di input sconosciuta.

Discontiguous MEGABLAST
E' utilizzato per trovare efficientemente lunghi allineamenti tra sequenze che hanno alcune differenze tra loro.

BLASTN
Utilizzato in tutti gli altri casi.

BLAST - http://www.ncbi.nlm.nih.gov/BLAST

BLAST PROTEICO

BLASTP
E' utilizzato per identificare una sequenza proteica di input nel DB o per ricercare sequenze proteiche simili;

PSI-BLAST
Position-Specific Iterata BLAST è il programma BLAST più sensibile, il che lo rende molto utile per trovare proteine poco correlate (molto distanti).

PHI-BLAST
Pattern-Hit Initiated BLAST è progettato per la ricerca di proteine che contengono un pattern specificato dall'utente e sono simili alla sequenza query in prossimità del pattern.

BLAST - http://www.ncbi.nlm.nih.gov/BLAST

ALTRI TOOL

BLASTX (Translated query vs protein database)

E' utilizzato per trovare proteine simili a quelle codificate da una query di nucleotidi;

TBLASTN (Protein query vs translated database)

E' utilizzato per trovare proteine omologhe a quella data in input. Le sequenze nucleotidiche del DB vengono tradotte in sequenze aminoacidiche utilizzando tutti e sei i frame di lettura e poi contfrontate con la query.

TBLASTX (Translated query vs translated database)

Prende in input una sequenza nucleotidica, la traduce in tutti e sei i frame di lettura e confronta queste sequenze tradotte con il DB di nucleotidi a sua volta tradotto in Aminoacidi.
Utile per trovare nuovi geni.

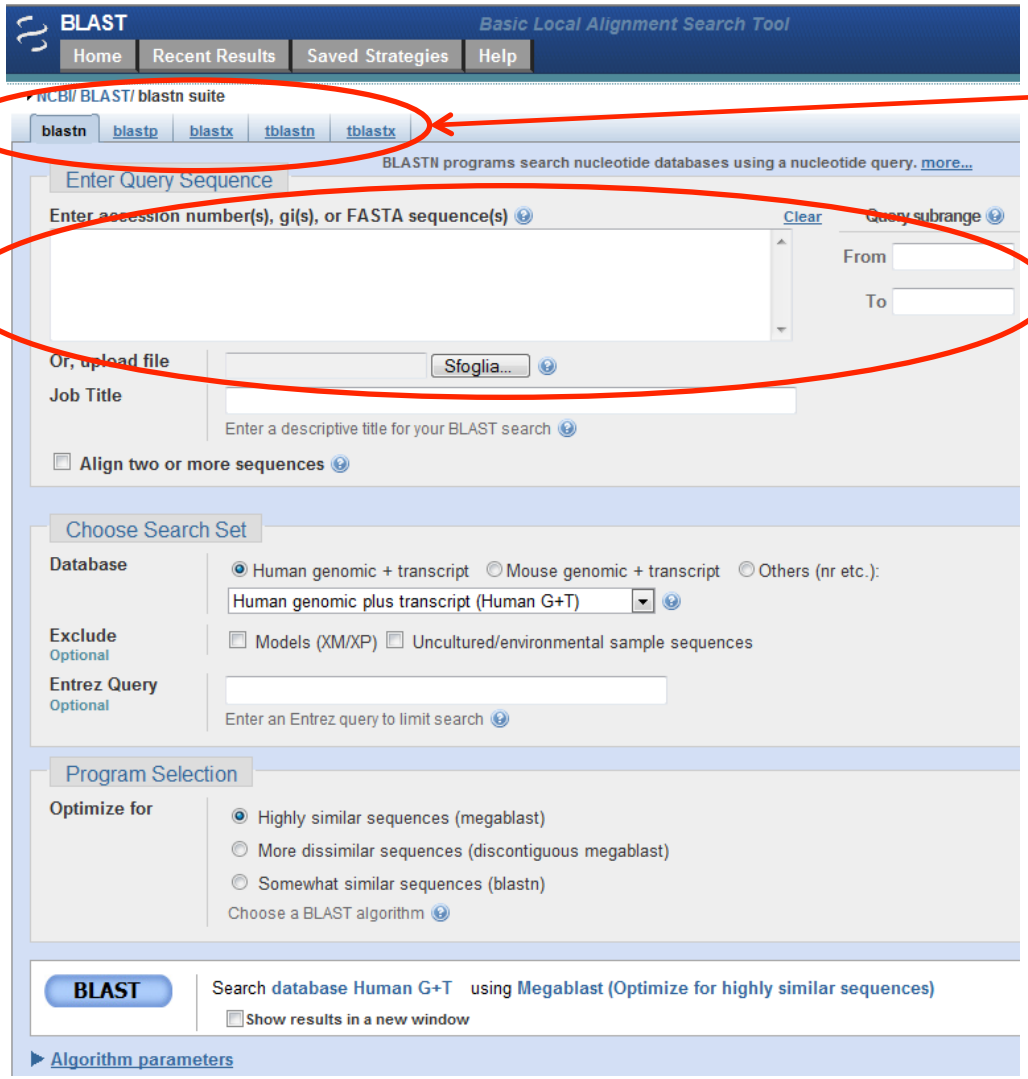
BLAST2SEQ

Utilizza BLAST per allineare due o più sequenze.

Program	Query	Database
BLASTP	aa	aa
BLASTN	nt	nt
BLASTX	nt (⇒ aa)	aa
TBLASTN	aa	nt (⇒ aa)
TBLASTX	nt (⇒ aa)	nt (⇒ aa)

Similarità nei DB - BLAST

http://www.ncbi.nlm.nih.gov/BLAST

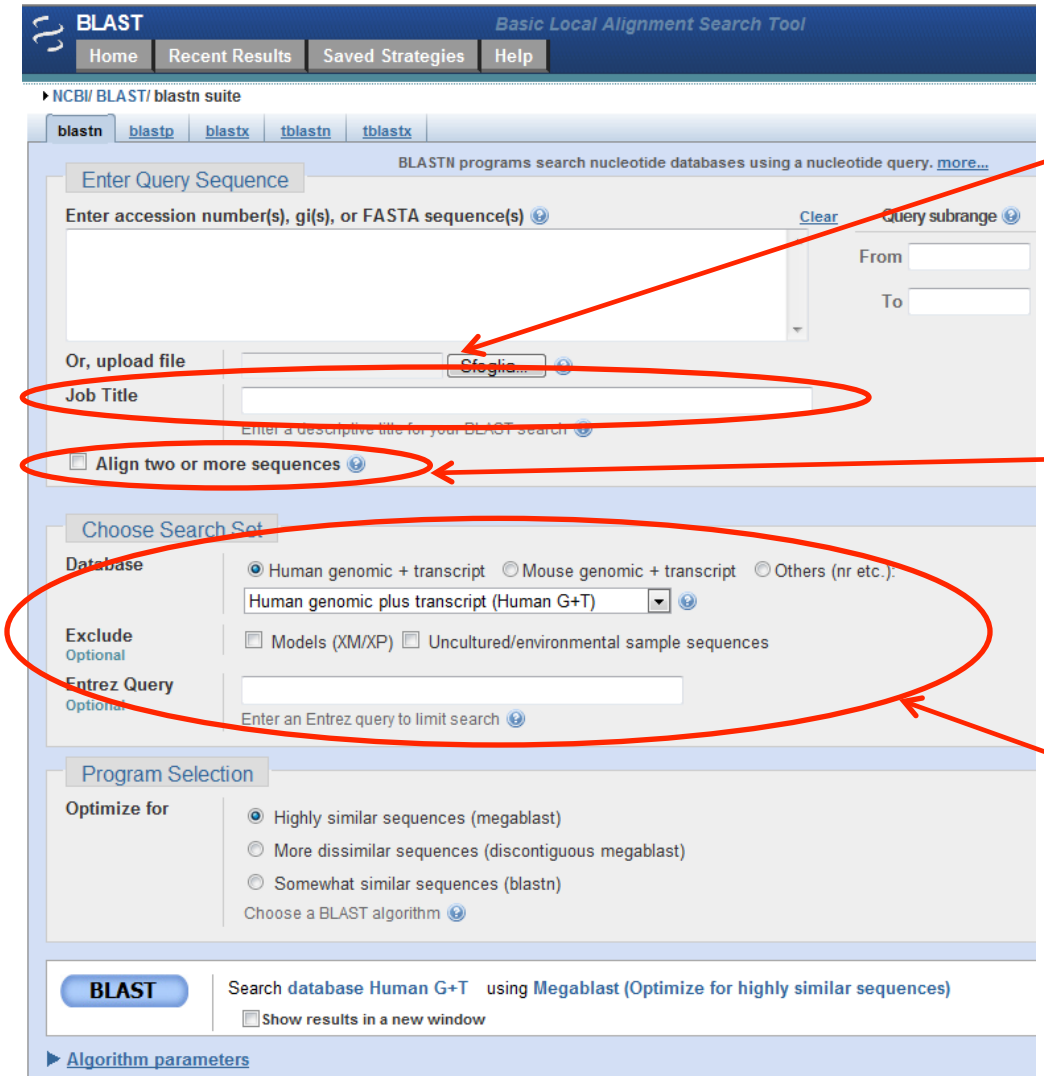


Scelta dei vari BLAST

Inserire la sequenza in formato FASTA (anche da file) oppure specificare l'Accession Number o il Gene ID. Specificare eventualmente l'intervallo di interesse.

Similarità nei DB - BLAST

http://www.ncbi.nlm.nih.gov/BLAST



Scegliere un nome descrittivo per la ricerca che apparirà nei risultati.

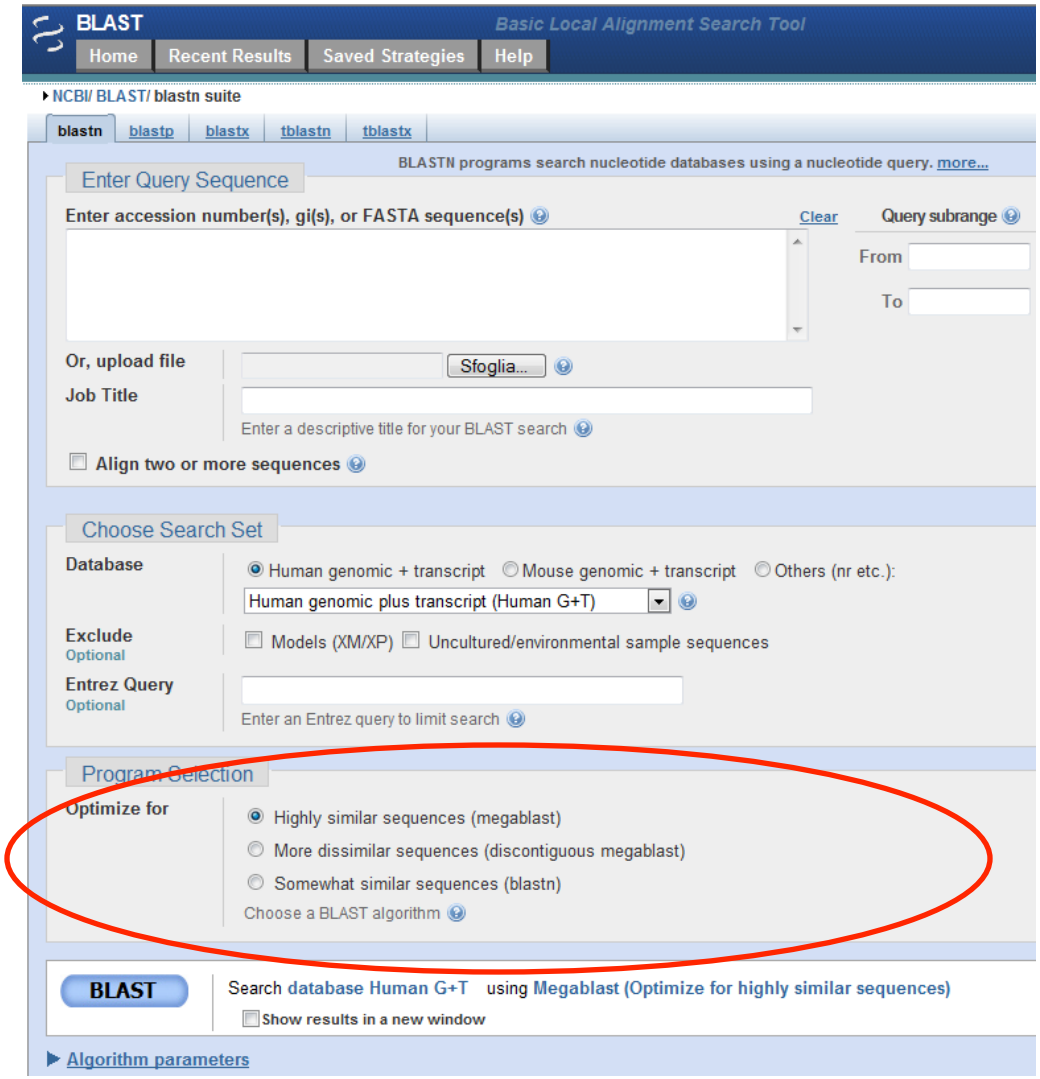
Selezionare se si vuole utilizzare BLAST per allineare due o più sequenze.

Campo di ricerca: DB, Organismo.

E' possibile usare la sintassi di entrez per filtrare i DB selezionati.

Similarità nei DB - BLAST

http://www.ncbi.nlm.nih.gov/BLAST

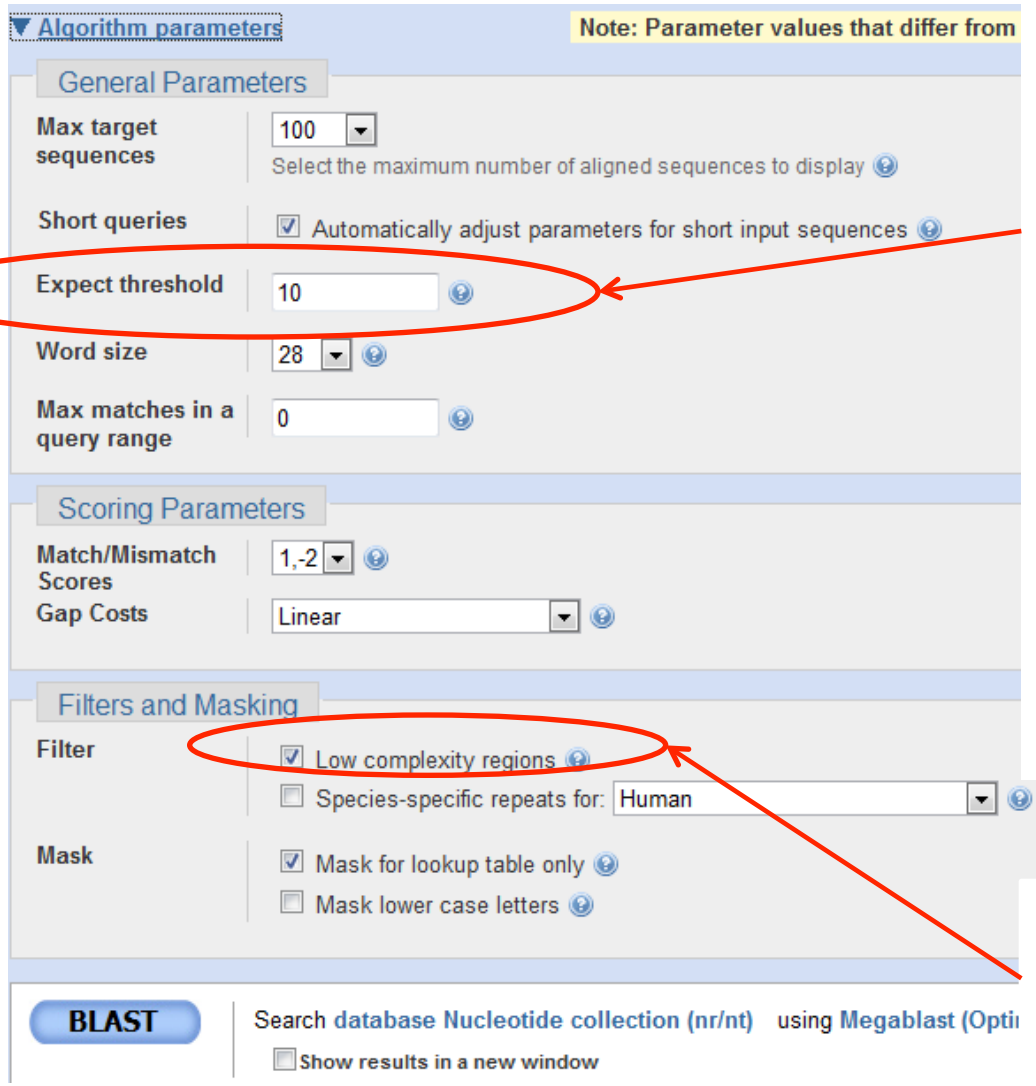


Ottimizza la ricerca per:

- Similarità;
- Dissimilarità;
- Ricerca generica;

Similarità nei DB - BLAST

http://www.ncbi.nlm.nih.gov/BLAST



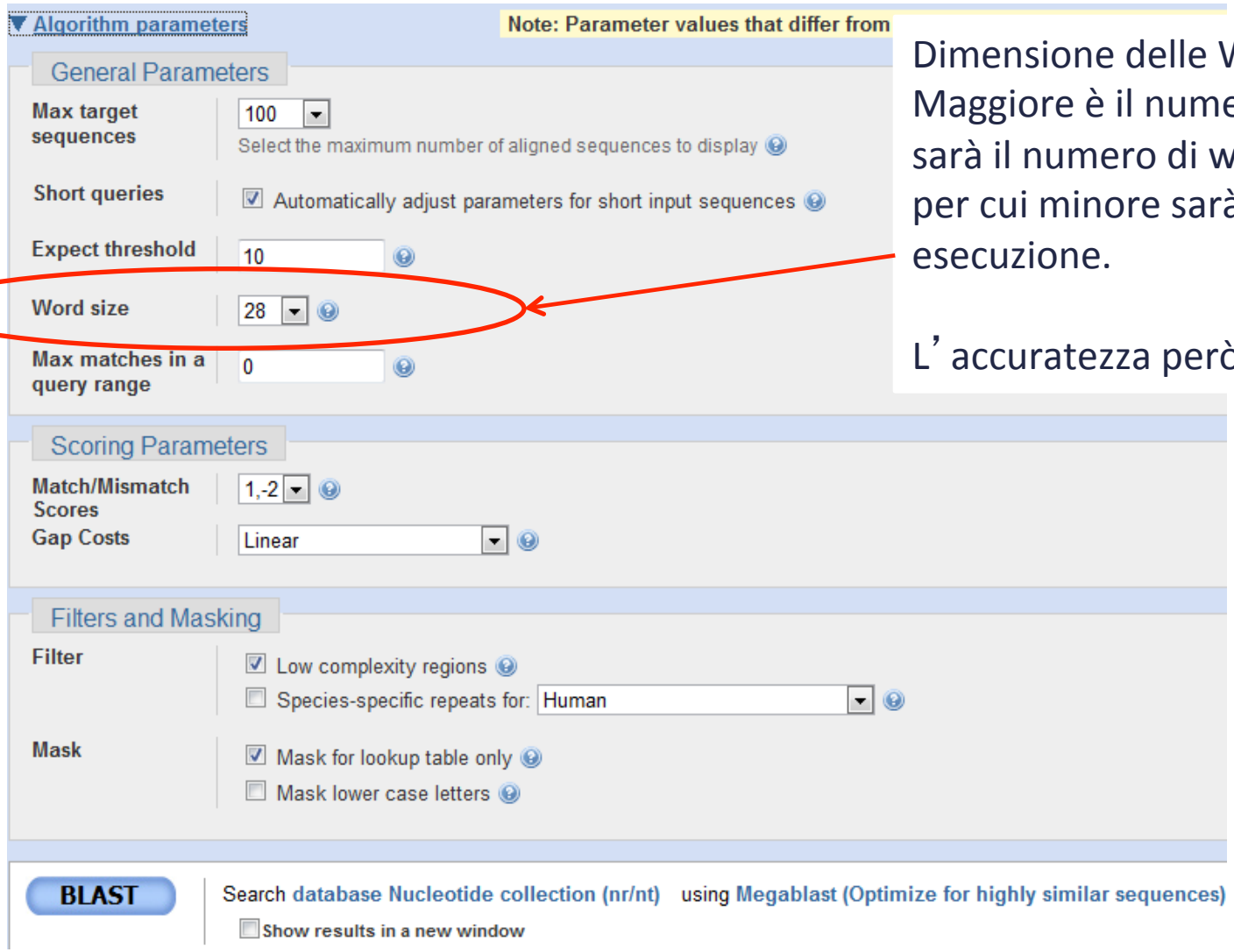
E' possibile cambiare la soglia di significatività statistica. Ogni match trovato ha un valore di significatività statistica, che indica quanto è statisticamente probabile che quel match sia casuale.

Minore è il numero, maggiore sarà il tempo di esecuzione.L' accuratezza però cresce.

Filtrare regioni il cui match avrebbe scarso significato biologico.

Similarità nei DB - BLAST

http://www.ncbi.nlm.nih.gov/BLAST



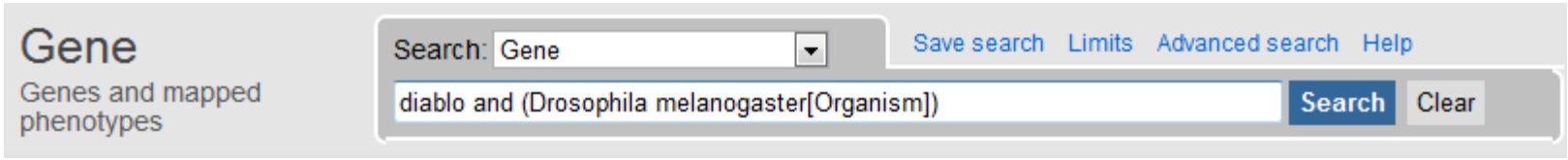
Dimensione delle Word: Maggiore è il numero, minore sarà il numero di word generate per cui minore sarà il tempo di esecuzione.

L'accuratezza però decresce.

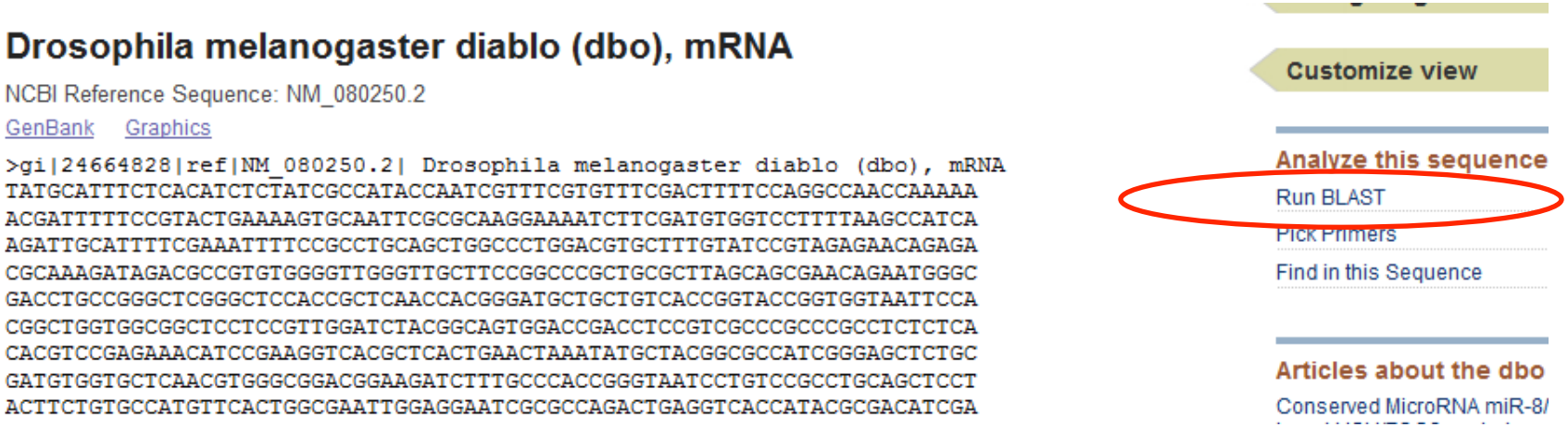
Similarità nei DB - BLAST

http://www.ncbi.nlm.nih.gov/BLAST

Esempio ricerchiamo il gene DIABLO in *Drosophila Melanogaster*

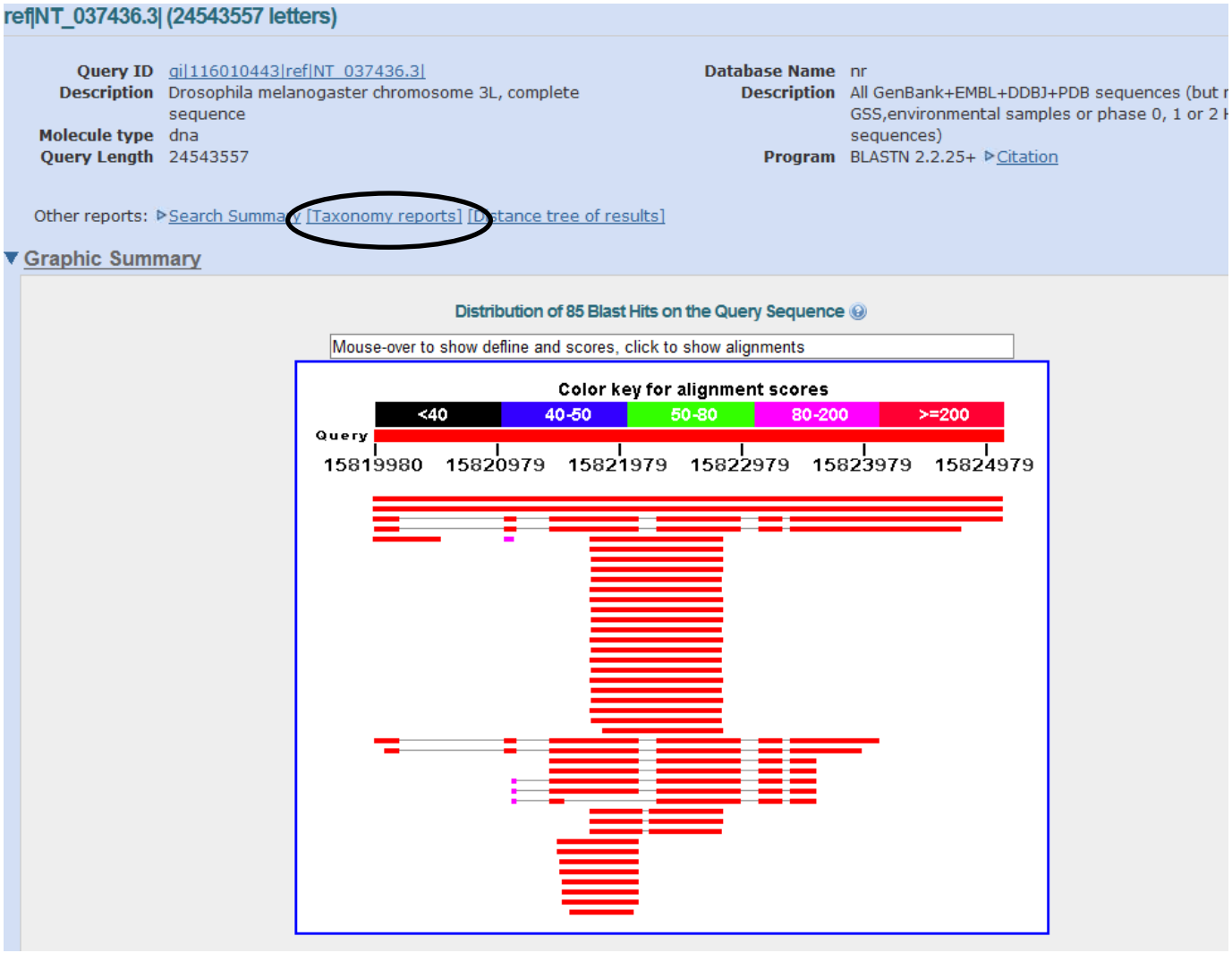


La prima voce che troviamo è il gene cercato. Selezioniamo la sequenza corrispondente di mRNA in formato FASTA e diamola in pasto a BLAST scegliendo come DB nt e tool Megablast.



Similarità nei DB - BLAST

http://www.ncbi.nlm.nih.gov/BLAST



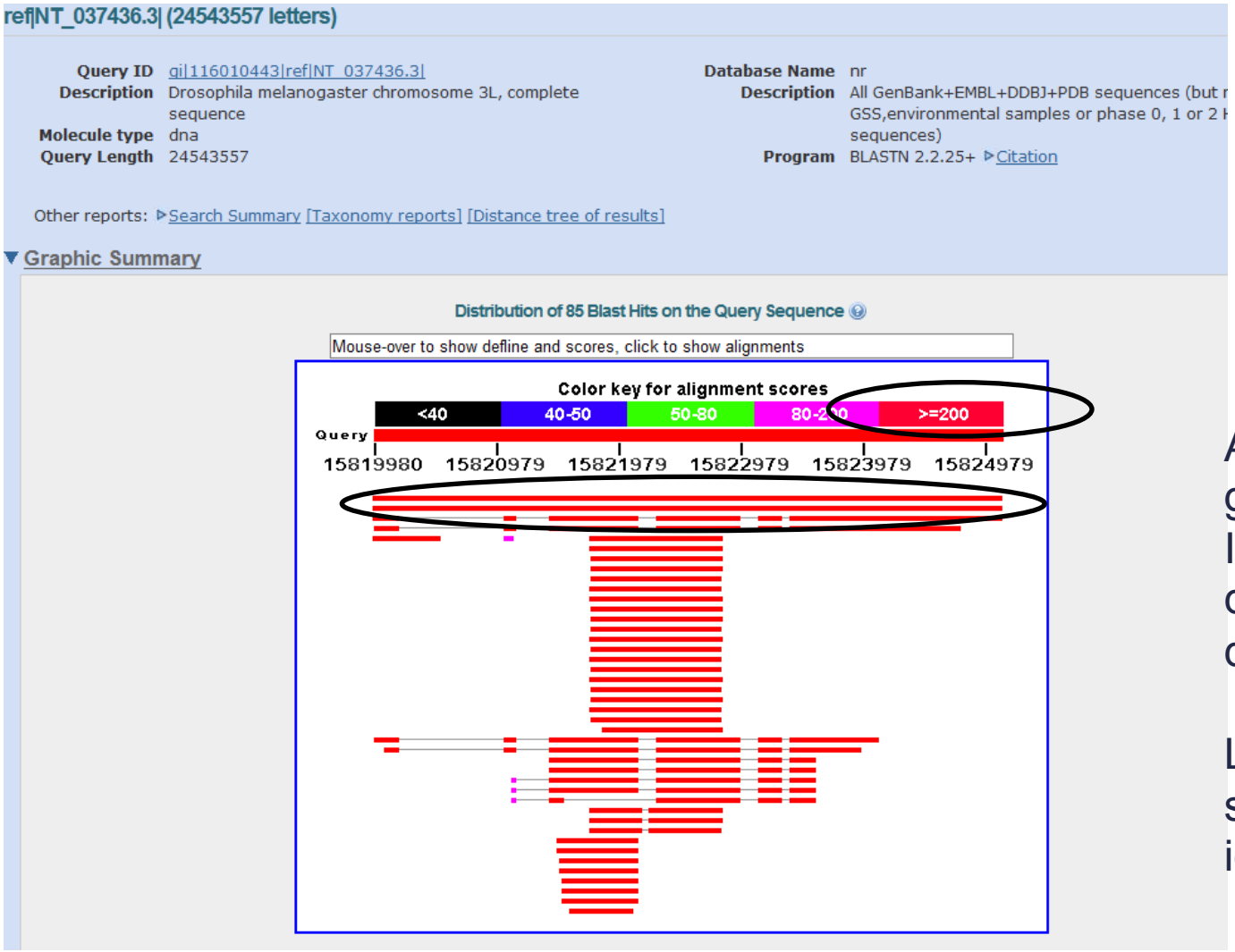
Dati generali

Taxonomy Report ci da informazioni sulle specie coinvolte nei risultati;

Può essere utile per verificare la presenza di sequenze ortologhi in altre specie;

Similarità nei DB - BLAST

http://www.ncbi.nlm.nih.gov/BLAST



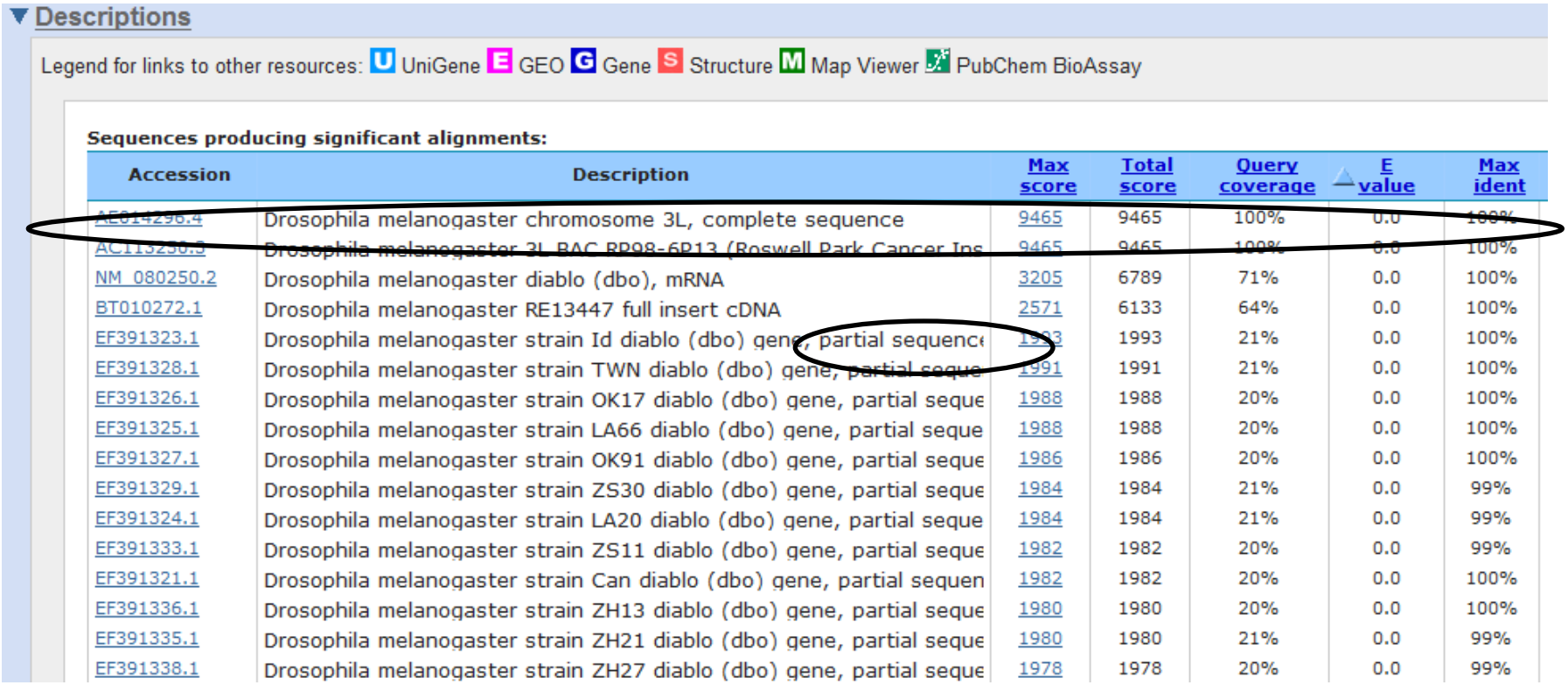
Dati generali

Allineamento grafico:
I colori indicano la qualità dell'allineamento.

Le prime due sequenze sono identiche.

Similarità nei DB - BLAST

http://www.ncbi.nlm.nih.gov/BLAST



Descriptions

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer PubChem BioAssay

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
AE014296.4	Drosophila melanogaster chromosome 3L, complete sequence	9465	9465	100%	0.0	100%
AC113236.3	Drosophila melanogaster 3L BAC RP98-6P13 (Roswell Park Cancer Ins	9465	9465	100%	0.0	100%
NM_080250.2	Drosophila melanogaster diablo (dbo), mRNA	3205	6789	71%	0.0	100%
BT010272.1	Drosophila melanogaster RE13447 full insert cDNA	2571	6133	64%	0.0	100%
EF391323.1	Drosophila melanogaster strain Id diablo (dbo) gene, partial sequence	1993	1993	21%	0.0	100%
EF391328.1	Drosophila melanogaster strain TWN diablo (dbo) gene, partial seque	1991	1991	21%	0.0	100%
EF391326.1	Drosophila melanogaster strain OK17 diablo (dbo) gene, partial seque	1988	1988	20%	0.0	100%
EF391325.1	Drosophila melanogaster strain LA66 diablo (dbo) gene, partial seque	1988	1988	20%	0.0	100%
EF391327.1	Drosophila melanogaster strain OK91 diablo (dbo) gene, partial seque	1986	1986	20%	0.0	100%
EF391329.1	Drosophila melanogaster strain ZS30 diablo (dbo) gene, partial seque	1984	1984	21%	0.0	99%
EF391324.1	Drosophila melanogaster strain LA20 diablo (dbo) gene, partial seque	1984	1984	21%	0.0	99%
EF391333.1	Drosophila melanogaster strain ZS11 diablo (dbo) gene, partial seque	1982	1982	20%	0.0	99%
EF391321.1	Drosophila melanogaster strain Can diablo (dbo) gene, partial sequen	1982	1982	20%	0.0	100%
EF391336.1	Drosophila melanogaster strain ZH13 diablo (dbo) gene, partial seque	1980	1980	20%	0.0	100%
EF391335.1	Drosophila melanogaster strain ZH21 diablo (dbo) gene, partial seque	1980	1980	21%	0.0	99%
EF391338.1	Drosophila melanogaster strain ZH27 diablo (dbo) gene, partial seque	1978	1978	20%	0.0	99%

Le prime due sequenze sono identiche alla query (per questo motivo BLAST può essere usato per ricercare sequenze sconosciute).

Le altre sono sequenze parziali.

Similarità nei DB - BLAST

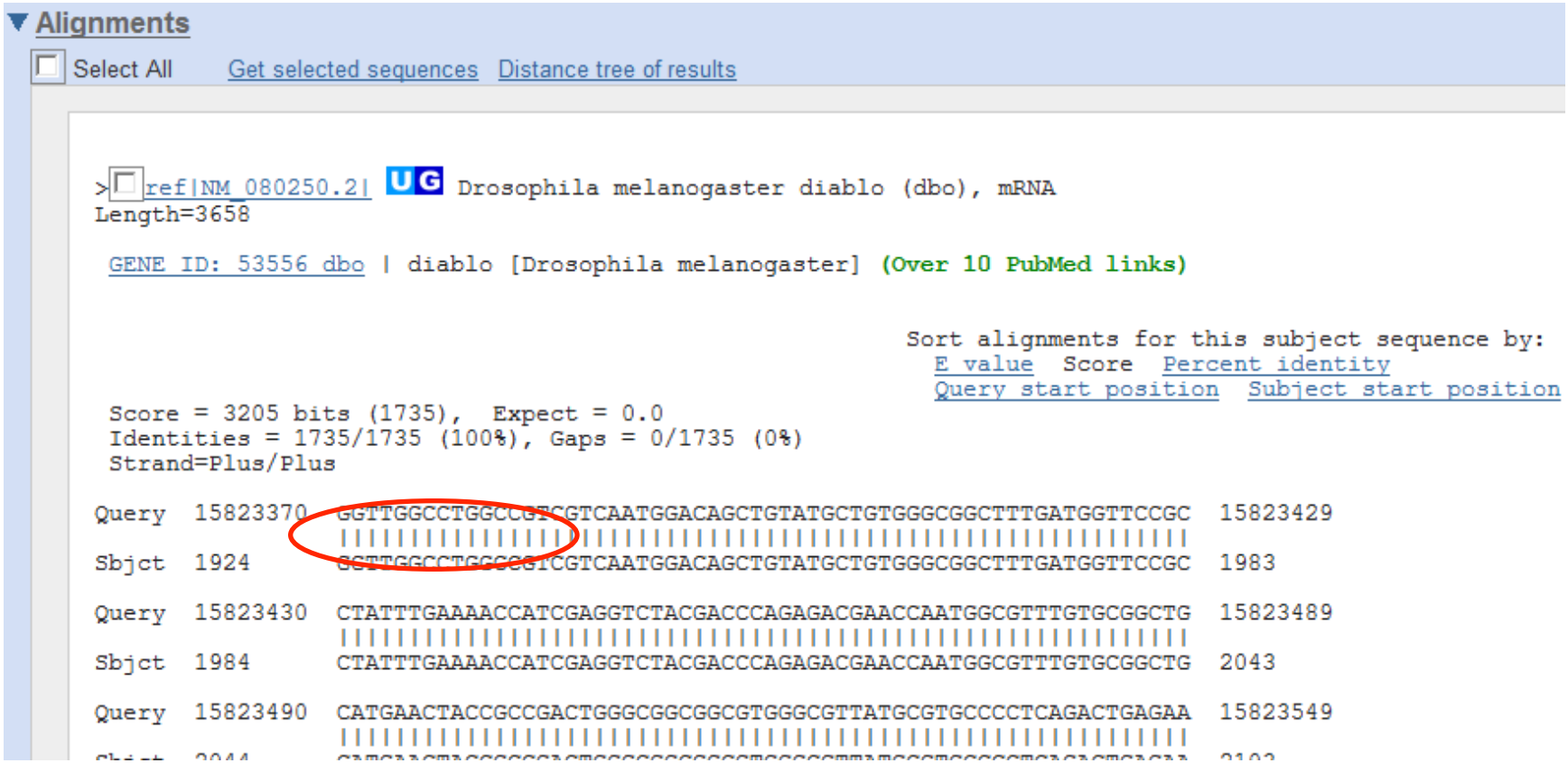
http://www.ncbi.nlm.nih.gov/BLAST

XM_002084980.1	Drosophila simulans GD14575 (Dsim\GD14575), mRNA	1269	3156	35%	0.0	99%
XM_002030609.1	Drosophila sechellia GM25560 (Dsec\GM25560), mRNA	1258	3140	35%	0.0	98%
XM_001973101.1	Drosophila erecta GG15931 (Dere\GG15931), mRNA	1208	3021	36%	0.0	100%
XM_002095190.1	Drosophila yakuba GE22281 (Dyak\GE22281), mRNA	1197	3015	36%	0.0	100%
XM_002086465.1	Drosophila yakuba GE23165 (Dyak\GE23165), mRNA	1011	2009	25%	0.0	100%
AF092016.1	Drosophila melanogaster microsatellite DM87	1009	1009	10%	0.0	99%
EF391340.1	Drosophila simulans strain S132 diablo (dbo) gene, partial sequence	983	1715	19%	0.0	97%
EF391341.1	Drosophila sechellia strain S9 diablo (dbo) gene, partial sequence	972	1706	20%	0.0	97%
EF391342.1	Drosophila mauritiana strain G105 diablo (dbo) gene, partial sequence	959	1691	19%	0.0	97%
XM_001963388.2	Drosophila pseudoobscura pseudoobscura GA19454 (Dpse\GA19454),	702	702	12%	0.0	85%
XM_002021185.1	Drosophila persimilis GL25213 (Dper\GL25213), mRNA	702	702	12%	0.0	85%
XM_001985454.1	Drosophila grimshawi GH17090 (Dgri\GH17090), mRNA	652	652	12%	0.0	84%
XM_002008626.1	Drosophila mojavensis GI11691 (Dmoj\GI11691), mRNA	632	632	12%	6e-177	84%
XM_309921.3	Anopheles gambiae str. PEST AGAP011587-PA (AgaP_AGAP011587) r	536	536	12%	5e-148	82%
XM_001861779.1	Culex quinquefasciatus ring canal kelch protein, mRNA	486	486	12%	5e-133	80%
EZ975260.1	TSA: Anopheles funestus Afun009205 mRNA sequence	411	411	12%	3e-110	78%
XM_002195997.1	PREDICTED: Taeniopygia guttata similar to kelch-like 20 (LOC100232	368	368	10%	2e-97	79%
XM_001957307.1	Drosophila ananassae GF24095 (Dana\GF24095), mRNA	134	134	1%	8e-27	97%

Scorrendo i risultati troviamo altre sequenze (anche parziali) in altri tipi di Drosophila.

Similarità nei DB - BLAST

http://www.ncbi.nlm.nih.gov/BLAST



Alignments

Select All Get selected sequences Distance tree of results

>ref|NM_080250.2| UG Drosophila melanogaster diablo (dbo), mRNA
Length=3658

GENE ID: 53556 dbo | diablo [Drosophila melanogaster] (Over 10 PubMed links)

Sort alignments for this subject sequence by:
E value Score Percent identity
Query start position Subject start position

Score = 3205 bits (1735), Expect = 0.0
Identities = 1735/1735 (100%), Gaps = 0/1735 (0%)
Strand=Plus/Plus

```
Query  15823370  GGTTGGCCTGGCCGTCGTCAATGGACAGCTGTATGCTGTGGGCGGCTTTGATGGTTCCGC  15823429
                 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1924      GGTTGGCCTGGCCGTCGTCAATGGACAGCTGTATGCTGTGGGCGGCTTTGATGGTTCCGC  1983

Query  15823430  CTATTTGAAAACCATCGAGGTCTACGACCCAGAGACGAACCAATGGCGTTTGTGCGGCTG  15823489
                 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1984      CTATTTGAAAACCATCGAGGTCTACGACCCAGAGACGAACCAATGGCGTTTGTGCGGCTG  2043

Query  15823490  CATGAACTACCGCCGACTGGGCGGCGGCGTGGGCGTTATGCGTGCCCCTCAGACTGAGAA  15823549
                 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
```

Infine troviamo i dettagli dei vari allineamenti.
I trattini indicano un match, la loro assenza indica un mismatch.

Similarità nei DB - BLAST

http://www.ncbi.nlm.nih.gov/BLAST

MAX SCORE
Punteggio dell'allineamento locale più significativo.
(punteggio alto → elevata similarità);

TOTAL SCORE
La somma dei punteggi di tutti gli allineamenti locali trovati tra la sequenza query e le sequenze del database.

QUERY COVERAGE
Percentuale della sequenza allineata

E-VALUE
Esprime la probabilità che l'allineamento trovato sia casuale. Più basso è, maggiore è la probabilità che NON sia casuale.
(dipende, oltre che dalla similarità, anche dalla numerosità delle sequenze in database e dalla lunghezza delle sequenze).

MAX INDENT
Percentuale di identità dell'allineamento locale più significativo.

Max score	Total score	Query coverage	E value	Max ident
9465	9465	100%	0.0	100%
9465	9465	100%	0.0	100%
3205	6789	71%	0.0	100%
2571	6133	64%	0.0	100%
1993	1993	21%	0.0	100%
1991	1991	21%	0.0	100%
1988	1988	20%	0.0	100%
1988	1988	20%	0.0	100%
1986	1986	20%	0.0	100%
1984	1984	21%	0.0	99%

Allineamento Pairwise – BLAST2SEQ

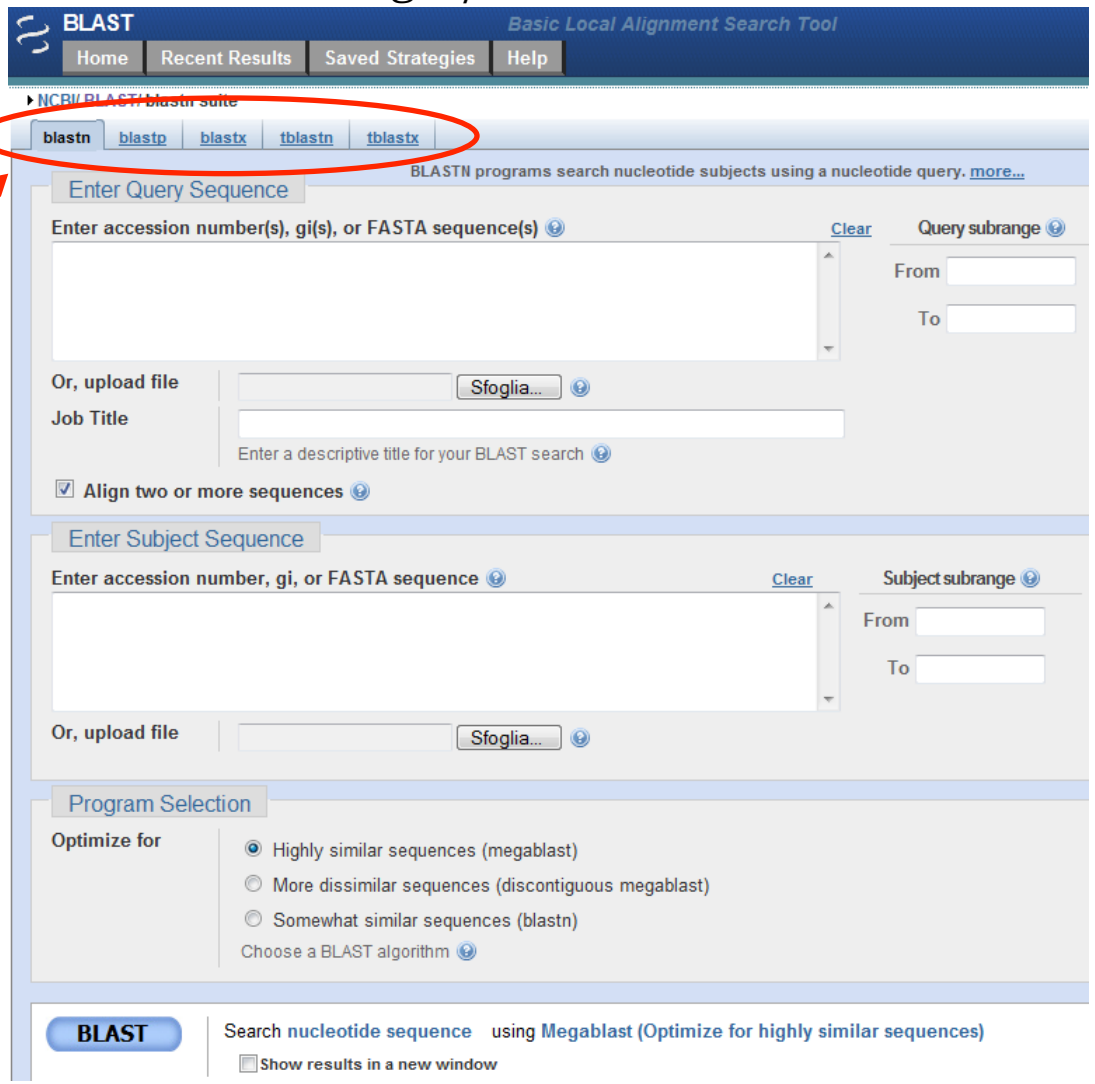
http://www.ncbi.nlm.nih.gov/BLAST

E' possibile usare BLAST per fare un allineamento di due sequenze.

In questo caso verranno evidenziate le similarità locali.

Si sceglie il programma adatto, si inseriscono le sequenze e si ottiene il risultato.

I parametri dell'interfaccia cambiano leggermente quanto si sceglie di allineare proteine piuttosto che nucleotidi (ad esempio le matrici di score).



Similarità nei DB - BLAST

http://www.ncbi.nlm.nih.gov/BLAST

Ritorniamo alla pagina principale di BLAST

C'è una sezione dedicata
ai genomi completi
(o in fase di completamento);

In questo modo è possibile fare
un BLAST su sequenze di una
data specie;