ESERCITAZIONE 3

OBIETTIVO: Ricerca di omologhe mediante i programmi FASTA e BLAST

L'esercitazione prevede l'utilizzo di risorse web per effettuare *ricerche di similarità* con la proteina **GRB2** (growth factor receptor-bound protein 2, *Homo sapiens*) e con il messaggero che la codifica. Tale proteina è coinvolta nei processi di trasduzione del segnale.

La ricerca dovrà essere effettuata a partire dai seguenti link:

- <u>http://www.ncbi.nlm.nih.gov/Entrez</u> Entrez, NCBI
- <u>http://www.ncbi.nlm.nih.gov/BLAST</u> BLAST (programma per effettuare ricerche di similarità), NCBI
- <u>http://www.ebi.ac.uk/fasta33/</u> FASTA (programma per effettuare ricerche di similarità), EBI

1) RECUPERO DELLA SEQUENZA NUCLEOTIDICA DI GRB2

- Da <u>Entrez</u> recuperare la voce relativa all'RNA messaggero di GRB2, utilizzando i seguenti criteri:
 - selezionare il database nucleotidico ("nucleotide");
 - utilizzare come parole chiave "grb2" e "homo" (quest'ultima ristretta al campo "[Organism]" unite dall'operatore booleano AND;
 - utilizzare come limiti ("Limits"):
 - 1. "exclude all of the above",
 - 2. per l'opzione "Molecule" scegliere "mRNA".

Per i dettagli su come effettuare l'interrogazione di database con Entrez riferirsi al testo della <u>prima esercitazione</u>.

- Si dovrebbe ottenere in questo modo un elenco di circa 350 entries (solo 3 anni fa erano 61!). Analizzando l'elenco si trova un'enty che corrisponde alla nostra interrogazione: NM_002086. Scegliamo a questo punto l'entry NM_002086, <u>a pagina 8-9</u>, in quanto contiene maggiori informazioni. Cliccare sul link relativo per accedervi.
- Scegliere dal menu a tendina "display" posto in alto a sinistra l'opzione "FASTA", quindi "send to" "text".
- Selezionare con il mouse la sequenza nucleotidica: partire dal segno ">", con cui comincia l'intestazione della sequenza, e arrivare fino all'ultima base; copiare quanto evidenziato cliccando sul menu di Netscape "Edit" e poi sull'opzione "Copy" (alternativamente copiare digitando contemporaneamente i tasti "Ctrl" e "c"). In questo modo abbiamo recuperato la sequenza nucleotidica in formato FASTA per la successiva ricerca di similarità.

2) RICERCA DI SIMILARITA' CON LA SEQUENZA NUCLEOTIDICA DI GRB2 UTILIZZANDO IL PROGRAMMA BLASTN

- A questo punto collegarsi all'indirizzo: <u>http://www.ncbi.nlm.nih.gov/BLAST</u>.
 BLAST ("Basic Local Alignment Search Tool") consiste in un insieme di sottoprogrammi che effettuano diversi tipi di ricerche di similarità.
- Dalla pagina che si presenta scegliere nella box contrassegnata da Nucleotide: "Nucleotide-nucleotide BLAST (blastn)". Si arriva così alla pagina del sottoprogramma di BLAST BLASTN che effettua una ricerca di similarità utilizzando come "query" una sequenza nucleotidica e come database un database anch'esso nucleotidico.
- Incollare la sequenza nucleotidica (menu "Edit" di Netscape, opzione "Paste", oppure "Ctrl" e "v") nella casella di input denominata "Search"; come database su cui eseguire la ricerca ("Choose database") lasciamo "nr" (un database esteso di sequenze nucleotidiche Non Ridondanti). Anche gli altri paramentri non vanno modificati.
- Cliccare sul tasto "BLAST!", posto poco sotto la casella di input denominata "Search", per inoltrare la richiesta al server dell'NCBI.
- Si ottiene così una pagina di attesa in cui in piccolo, sotto il tasto "Format!", è indicato il tempo stimato per la ricerca ("The results are estimated to be ready in nn seconds but may be done sooner.").
- Dopo aver atteso approssimativamente quanto indicato, cliccare sul tasto "Format!" per ricevere i risultati.
- Si dovrebbe aprire una nuova finestra contenente i risultati. Se il server non ha ancora terminato, compare un messaggio di attesa che si aggiorna automaticamente. **NOTA:** il tempo di attesa varia a seconda del carico di lavoro del server.

INTERPRETAZIONE DEI RISULTATI DI BLASTN

La pagina dei risultati si può dividere in cinque parti (dall'alto in basso):

PRIMA PARTE La prima parte dà informazioni sul programma (in questo caso BLASTN), sulle banche dati nucleotidiche utilizzate e sulla sequenza "query" (quella con cui si effettua la ricerca), tra le quali la sua lunghezza (indicata come "nn letters", dove "letters" sta per basi).



SECONDA PARTE La seconda parte consiste in una immagine che illustra graficamente i risultati (vedi sotto).



Per prima cosa e' riportato il numero dei risultati (o "hits") ottenuti. Quindi:

- la linea rossa spessa e graduata rappresenta la sequenza "query";
- i numeri sotto di essa si riferiscono alla sua lunghezza in basi;
- ciascuna linea sottile sottostante, di vari colori, indica un allineamento della suddetta sequenza con una sequenza del database nucleotidico;
- il colore di tali linee indica la bontà dell'allineamento, in base alla scala colorimetrica posta all'inizio dell'immagine (rosso migliore, nero peggiore);
- passando su ciascuna linea sottile con la freccia del mouse, sul riquadro subito sopra l'immagine compare la descrizione della sequenza nucleotidica del database con cui la sequenza query si allinea in quel caso;
- cliccando su tali linee si va invece direttamente all'allineamento (situato nella quarta parte della pagina) tra la "query" e una data sequenza del database.

TERZA PARTE La terza parte consiste nell'elenco delle sequenze nucleotidiche del database scelto che producono allineamenti significativi con la sequenza "query" e comincia con la frase: "Sequences producing significant alignments:". Le sequenze sono ordinate in base all'E value (colonna a destra), parametro che misura la significatività degli allineamenti: minore è, più significativo è l'allineamento. Ciascuna sequenza contiene un link (parte sottolineata e in colore), da cui si arriva alla voce di Entrez relativo a quella sequenza. A destra, per una data sequenza sono inoltre eventualmente presenti i link a UniGene ("U", che raggruppa i geni in cluster non ridondanti e fornisce svariate

informazioni, quali l'espressione nei diversi tessuti etc etc), Geo profiles ("E", che riporta livelli di espressione del gene a diverse condizioni cellulari) e Gene ("G", che riporta la localizzazione genomica e la struttura del gene).

🔮 RID=11-	RID=1142852831-32225-37724636761.BLASTQ1, gi 45359857 ref NM_002086.3 Homo sapiens growth fact - Microsoft Internet Explorer								
File Edit View Favorites Tools Help									
😓 Back 👻	🔿 🔹 🗿 🚮 🧕 🛐 Favorites 🛞 Media 🧭 🛃 🗉 🗐								
Address 🧕	http://www.ncbi.nlm.nih.gov/blast/Blast.cgi				•	∂Go Links	»		
Searcl ×		Saoro	F				•		
Ć∜tNe ≫	Sequences producing significant alignments:	(Bits)	Value			1			
Search The Web Find a Web page containir Search Brought to you by	gi 45359857 ref NM_002086.3] Homo sapiens growth factor recep gi 21749282 dbj AK091010.1 Homo sapiens cDNA FLJ33691 fis, c gi 55732372 emb CR860761.1 Pongo pygmaeus mRNA; cDNA DKF2p45 gi 64579022 dbj AE220412.1 Pongo pygmaeus mRNA; cDNA DKF2p45 gi 84579022 dbj AE220412.1 Macaca fascicularis mRNA, clone Q gi 84579022 dbj AE200631.2 Macaca fascicularis mRNA, clone Q gi 33875666 gb EC000631.2 Homo sapiens growth factor recept gi 33877979 gb EC019082.2 Homo sapiens growth factor recept gi 33877979 gb EC019082.2 Homo sapiens growth factor recept gi 32512147 gb EC039627.1 Homo sapiens growth factor recept gi 14714383 gb EC039417.1 Homo sapiens growth factor recept gi 14714383 gb EC039417.1 Homo sapiens growth factor recept	3505 3273 3166 3073 3017 2446 2422 2369 2343 2335 2335 2202 2194 2165	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0						
MSN Search	gi 27695085 gb BC043007.1 Homo sapiens growth factor recepto gi 21707087 gb BC033644.1 Homo sapiens growth factor recepto gi 181975 gb M96995.1 HUMEGFGRBA Homo sapiens epidermal growt	2161 2125 2050	0.0 0.0 0.0	U G U G U E G					
Grant Strain Str	<pre>giiatososizigpi AF302079.1[AF302079.1] gii33337967[gb]AF171699.1] fomo sapiens MSTP084 (MST084) mPNA, c gii28875[emb]X62852.1]HSASH H.sapiens ash mNA gii73964903[ref]XM_550974.1] PREDICTED: Canis familiaris simi gii400667[gb]L29511.1[HUNGRB33X Human GRE2 isoform mNA gii49456838[emb]CR541942.1] Homo sapiens full open reading fr gii60655182[gb]AF398364.1] Synthetic construct Homo sapiens c gii6085182[gb]AF393202.1] Synthetic construct Homo sapiens c gii60812930[gb]AF393201.1] Synthetic construct Homo sapiens c gii60812930[gb]AF393201.1] Synthetic construct Homo sapiens c gii60812930[gb]AF393201.1] Synthetic construct Homo sapiens c gii60812930[gb]AF39980[a].1] Synthetic construct Homo sapiens c gii60812930[gb]AF39980[b].1] Synthetic construct Homo sapiens c gii60812930[gb]AF39980[b].1] Synthetic construct Homo sapiens c gii608559[gb]AF3980[b]390[b].1] Synthetic construct Homo sapiens c gii6335659[gb]AF3980[b].1] Synthetic construct Homo sapiens c</pre>	1588 1513 1509 1447 1413 1312 1296 1296 1296 1296 1296 1296 1296 1296 1292 1292	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0						
A http://ww	ww.nchi.nlm.nib.nov/hlast/Blast.coi#21749282				internet		-		

QUARTA PARTE La quarta parte visualizza gli allineamenti significativi della sequenza "query" con sequenze del database scelto e comincia con la frase: "Alignments".

🖉 RID=11	RID=1142852831-32225-37724636761.BLASTQ1, gi 45359857 ref NM_002086.3 Homo sapiens growth fact - Microsoft Internet Explorer												
File Edit	View Fa	avorites	Tools Help										
😓 Back 🕞	\Rightarrow \cdot \otimes	1	🕴 🔯 Search 🕼 Favorites 🐠 Media 🧭 🖏 - 🎒 🕅 - 📄										
Address 🧧	http://www	w.ncbi.nlm	n.nih.gov/blast/Blast.cgi#464004	•	€Go	Links »							
Searcl × ©t Ne »	> 🗌 gi Length	<u>46400</u> ⊫2362	4 gb U07617.1 MMU07617 UEG Mus musculus Grb2 adaptor protein (grb2) mRNA, complete cds			•							
Search The Web	Score = 1029 bits (519), Expect = 0.0 Identities = 822/913 (90%), Gaps = 10/913 (1%) Strand=Plus/Plus												
Find a Web	Query Sbjct	296 18	CTTCAGCAGGGAAGACTCCCTTCCCC-CTGCTTCAGGCTGCTGAGCAGCGGCGCCC 354										
page containir	Query	355 78											
Search Brought	Query	415	AAAAGGGGGGACATCCTCAAGGTTTTGAACGAAGAATGTGATCAGAACTGGTACAAGGCA 474										
you by MSN Search	Sbjct Query	138 475	AAAAGGGGGGACATCCTTAAGGTTTTGAATGAAGAATGTGACCAGAACTGGTATAAGGCA 197 GAGCTTAATGGAAAAGACGGCTTCATTCCCAAGAACTACATAGAAATGAAACCACATCCG 534										
	Sbjct	198	II II IIIII IIIII IIIIIIII IIIIIIII IIII										
Search for other	Query	535	TGGTTTTTTGGCAAAATCCCCAGAGCCAAGGCAAGAAATGCTTAGCAAACAGCGGCAC 594										
items: Files or	Sbjct	258	TGGTTTTTTGGCAAAATCCCCAGAGCCAAGGCAGAAGAAATGCTCAGCAAACAGCGGCAT 317										
Computer People	Query	595	GATGGGGCCTTTCTTATCCGAGAGAGTGAGAGCGCTCCTGGGGACTTCTCCCCTCTCTGTC 654										
0	Sbjct	318	GACGGGGCCTTCCTGATCCGAGAGAGCGGAGAGCGCTCCTGGGGGACTTCTCCCCTGTCCGTC 377										
2006 Microso <u>MSN</u> Privacu	Sbjct	655 378	ANGTITIGGAAATGATGTGCAGCACTTCAAGGTGCTCCGCGACGGAGCCGGGAAGTACTTC 437										
	Querv	715	CTCTGGGTGGTGAAGTTCAATTCTTTGAATGAGCTGGTGGATTATCACAGATCTACATCT 774										

Data una sequenza del database scelto che produce uno o più allineamenti significativi con la sequenza "query", come prima cosa viene visualizzata la descrizione della sequenza del database, che contiene un link (parte sottolineata e in colore), da cui si arriva alla voce di Entrez relativa a quella sequenza. La descrizione contiene eventualmente anche i link a Gene ("G"), UniGene ("U") e GEO ("E"). Sotto la descrizione è indicata la lunghezza in basi della sequenza. Successivamente vengono visualizzati il o gli allineamenti di questa sequenza con la sequenza "query". Per ciascun allineamento è indicato:

- "Score": è il punteggio dell'allineamento (data la matrice di scoring e le penalizzazioni per i gap);
- "Espect": è l'E value dell'allineamento;
- "Identities": è il numero di basi identiche (nell'esempio sono 822) diviso per la lunghezza dell'allineamento in questione (nell'esempio in figura è di 913 basi); tra parentesi è indicata la risultante percentuale di identità su quella lunghezza di allineamento (nell'esempio è il 90%);
- "Gaps": indica il numero di gap presenti nell'allineamento (nell'esempio sono 10) / la lunghezza dell'allineamento in questione; tra parentesi è indicata la risultante percentuale di gap su quella lunghezza di allineamento (nell'esempio è l'1%);
- "Strand": indica l'orientamento della sequenza "query" rispetto alla sequenza del database con cui si allinea.

Vale la relazione: lunghezza allineamento = identities + gaps + altri mismatch nell'allineamento (ad esempio "a" che sia allinea con "t", ...).

Successivamente si ha l'allineamento vero e proprio tra la sequenza "query" e la sequenza del database in questione, denominata "Sbjct". I numeri indicano la posizione delle basi all'interno delle rispettive sequenze: ad esempio, il numero 77 cerchiato in rosso sta ad indicare che la base "c", indicata con la freccia rossa, è la base numero 77 nella sequenza del database.

Se in una data posizione dell'allineamento la base nella sequenza "query" e la corrispondente della sequenza del database coincidono, nella riga tra le due sequenze compare il carattere "|". Quando tale carattere non è presente significa che, in quella posizione dell'allineamento, la base nella sequenza "query" e la corrispondente della sequenza del database sono diverse oppure che una delle due sequenze presenta un gap.

QUINTA PARTE La quinta parte contiene dettagli statistici della ricerca.

Prendere nota di eventuali ortologhi e del loro grado di similarita' con la sequenza query.

3) RECUPERO DELLA SEQUENZA AMINOACIDICA DI GRB2

• Dalla pagina web che contiene la sequenza dell'entry NM_002086 (utilizzata per la ricerca di similarità) passare alla proteina codificata da questo messaggero cliccando

sul link NP_002077, in corrispondenza di "protein_id", che porta alla entry relativa a tale proteina.

- Scegliere dal menu a tendina "display" posto in alto a sinistra l'opzione "FASTA", quindi "send to" "text".
- Selezionare con il mouse la sequenza aminoacidica: partire dal segno ">", con cui comincia l'intestazione della sequenza, e arrivare fino all'ultimo aminoacido; e copiare quanto evidenziato. In questo modo recuperiamo la sequenza aminoacidica in formato FASTA per la successiva ricerca di similarità.

4) RICERCA DI SIMILARITA' CON LA SEQUENZA AMINOACIDICA DI GRB2 UTILIZZANDO IL PROGRAMMA BLASTP

- A questo punto collegarsi all'indirizzo: <u>http://www.ncbi.nlm.nih.gov/BLAST</u> per effettuare tale ricerca.
- Dalla pagina che si presenta scegliere nella box contrassegnata da Protein: "Proteinprotein BLAST (blastp)". Si arriva così alla pagina del sottoprogramma di BLAST BLASTP che effettua una ricerca di similarità utilizzando come "query" una sequenza proteica e come database un database anch'esso proteico.
- Incollare la sequenza aminoacidica (menu "Edit" di Netscape, opzione "Paste") nella casella di input denominata "Search"; come database su cui eseguire la ricerca ("Choose database") lasciamo "nr" (uno dei database di proteine). Anche gli altri paramentri non vanno modificati.
- **NOTA**. alternativamente potete impostare uno o piu' parametri diversamente da un vostro collega dello stesso "gruppo di lavoro" e verificare se e come i risultati ottenuti cambino a seguito di queste diverse impostazioni (es. diversa matrice di scoring, diverse penalizzazioni per i gap, diverso database etc.).
- Cliccare sul tasto "BLAST!", posto poco sotto la casella di input denominata "Search", per inoltrare la richiesta al server dell'NCBI.
- Si ottiene così una pagina di attesa in cui in piccolo, sotto il tasto "Format!", è indicato il tempo stimato per la ricerca ("The results are estimated to be ready in nn seconds but may be done sooner."). In questa pagina gia' compare l'organizzazione in domini della proteina query. (Cliccando sullo schema per ciascun dominio e' possibile accedere ad informazioni funzionali relative al dominio stesso).
- Sopra il tasto "Format!", viene eventualmente visualizzata una immagine cliccabile che mostra i domini proteici individuati.
- Dopo aver atteso approssimativamente quanto indicato, cliccare sul tasto "Format!" per ricevere i risultati.
- Si dovrebbe aprire una nuova finestra contenente i risultati. Se il server non ha ancora terminato, compare un messaggio di attesa che si aggiorna automaticamente. **NOTA:** il tempo di attesa varia a seconda del carico di lavoro del server.

INTERPRETAZIONE DEI RISULTATI DI BLASTP

La pagina dei risultati si può dividere in cinque parti (dall'alto in basso), del tutto equivalenti a quelle dei risultati di BLASTN visti precedentemente:

PRIMA PARTE La prima parte dà informazioni sul programma (in questo caso BLASTP), sulla sequenza "query" (quella con cui si effettua la ricerca), tra le quali la sua lunghezza (indicata come "nn letters", dove "letters" sta per aminoacidi), e sul database utilizzati.

SECONDA PARTE La seconda parte consiste in una immagine che illustra graficamente i risultati. Per prima cosa e' riportato il numero dei risultati (o "hits") ottenuti. Quindi:

- la linea rossa spessa e graduata rappresenta la sequenza "query";
- i numeri sotto di essa si riferiscono alla sua lunghezza in aminoacidi;
- ciascuna linea sottile sottostante, di vari colori, indica un allineamento della suddetta sequenza con una sequenza del database di proteine;
- il colore di tali linee indica la bontà dell'allineamento, in base alla scala colorimetrica posta all'inizio dell'immagine (rosso migliore, nero peggiore);
- passando su ciascuna linea sottile con la freccia del mouse, sul riquadro subito sopra l'immagine compare la descrizione della proteina del database con cui la sequenza query si allinea in quel caso;
- cliccando su tali linee si va invece direttamente all'allineamento (situato nella quarta parte della pagina) tra la "query" e una data sequenza del database.

TERZA PARTE La terza parte consiste nell'elenco delle proteine del database scelto che producono allineamenti significativi con la sequenza "query" e comincia con la frase: "Sequences producing significant alignments:". Le sequenze sono ordinate in base all'E value (colonna a destra). Ciascuna sequenza contiene un link (parte sottolineata e in colore), da cui si arriva alla vocedi Entrez relativa a quella sequenza. A destra, per una data sequenza sono inoltre presenti eventuali link a Gene ("G") e al Protein Data Bank ("S").

QUARTA PARTE La quarta parte visualizza gli allineamenti significativi della sequenza "query" con sequenze del database scelto e comincia con la frase: "Alignments".

```
>gi|1174436|sp|P42686|SRK1 SPOLA TYROSINE-PROTEIN KINASE SRK1
gi|283505|pir||S24550 protein-tyrosine kinase (EC 2.7.1.112) 1 [similarity] - freshwat
          sponge (Spongilla lacustris)
gi|10150|emb|CAA43798.1| (X61601) src-type tyrosine kinase 1 [Spongilla lacustris]
         Length = 505
Score = 107 bits (267), Expect = 9e-23
 Identities = 57/141 (40%), Positives = 91/141 (64%), Gaps = 8/141 (5%)
Query: 4
         IAKYDFKATADDELSFKRGDILKVLNEECDONWY--KAELNGKDGFIPKNYI----EMKP 57
          + KYD+ + DD+LSFK+GD++ +++ + W+ + GK+G+IP NY+
                                                                 +
Sbjct: 60 VGKYDYDSRTDDDLSFKKGDLMYIISTDEGDWWFARSKDTAGKEGYIPSNYVAEYKSLDA 119
Query: 58 HPWFFGKIPRAKAEEMLSKQRHD-GAFLIRESESAPGDFSLSVKFGNDVQHFKVLRDGAG 116
           WF GKI R +AE+ML++ + G+FLIR+SE+ PGDFSLSVK + V+H++V R
                                                                   G
Sbjct: (120) EEWFLGKIKRVEAEKMLNQSFNQVGSFLIRDSETTPGDFSLSVKDQDRVRHYRVRRLEDG 179
Query: 117 KYFLW-VVKFNSLNELVDYHR 136
           F+ F L+ELVD+++
Sbjct: 180 SLFVTRRSTFQILHELVDHYK 200
```

Data una sequenza del database scelto che produce uno o più allineamenti significativi con la sequenza "query", come prima cosa viene visualizzata la descrizione della sequenza del database, che contiene un link (parte sottolineata e in colore), da cui si arriva alla entry di Entrez relativa a quella sequenza. La descrizione contiene eventualmente anche il link a Gene ("G") e, nel caso siano disponibili una o piu' strutture sperimentali per la proteina query, a Related Structures ("S"), da dove e' possibile accedere alle corrispondenti entries in CDs (Conserved Domain con descrizioni dei domini) e nel Protein Data Bank (PDB). Sotto la descrizione è indicata la lunghezza in aminoacidi della sequenza.

Per ciascun allineamento è indicato:

- "Score": è il punteggio dell'allineamento (data la matrice di scoring e le penalizzazioni per inserzioni/delezioni);
- "Espect": è l'E value dell'allineamento; •
- "Identities": è il numero di aminoacidi identici (nell'esempio sono 57) / la lunghezza dell'allineamento in questione (nell'esempio è 141 aminoacidi); tra parentesi è indicata la risultante percentuale di identità su quella lunghezza di allineamento (nell'esempio è il 40%);
- "Positives": è il numero di aminoacidi identici più il numero di aminoacidi simili (nell'esempio tale numero è 91) / la lunghezza dell'allineamento in questione; tra parentesi è indicata la risultante percentuale su quella lunghezza di allineamento (nell'esempio è il 64%); due aminoacidi vengono definiti simili in base alla matrice utilizzata per l'allineamento (nel nostro caso BLOSUM62);
- "Gaps": indica il numero di gap presenti nell'allineamento (nell'esempio sono 8) / la lunghezza dell'allineamento in questione; tra parentesi è indicata la risultante percentuale di gap su quella lunghezza di allineamento (nell'esempio è il 5%).

Vale la relazione: lunghezza allineamento = identities + gaps + altri mismatch nell'allineamento. Successivamente si ha l'allineamento vero e proprio tra la sequenza "query" e la sequenza del database in questione, denominata "Sbjct". I numeri indicano la posizione degli aminoacidi all'interno delle rispettive sequenze: ad esempio, il numero 120 cerchiato in rosso nella in alto sta ad indicare che l'aminoacido "E" (acido glutammico), indicato con la freccia rossa, è l'aminoacido numero 120 nella sequenza del database. Se in una data posizione dell'allineamento l'aminoacido nella sequenza "query" e il corrispondente della sequenza del database coincidono, la lettera che lo rappresenta viene ripetuta nella riga tra le due sequenze (vedi, ad esempio, la posizione dell'allineamento, l'aminoacido nella sequenza "query" e il corrispondente della sequenza del database sono simili. Quando nella riga tra le due sequenze carattere non è presente nessun carattere significa che, in quella posizione dell'allineamento, l'aminoacido nella sequenza "query" e il corrispondente della sequenza del database sono diversi oppure che una delle due sequenze presenta un gap.

QUINTA PARTE La quinta parte contiene dettagli statistici della ricerca.

Prendere nota di eventuali ortologhi (sono gli stessi identificati con il gene?) e delle sequenze proteiche piu' simili alla query per cui e' disponibile una struttura nel Protein Data Bank.

5. ALLINEAMENTO TRA LA SEQUENZA AMINOACIDICA DI GRB2 E SRK1_SPOLA

- Dalla pagina dei risultati di BLASTP individuare l'allineamento di GRB2 con la proteina **SRK1_SPOLA Tyrosine-protein kinase SRK1 di Spongilla lacustris**, che ha numero di accesso **P42686** (si trova più o meno verso la fine della pagina).
- Cliccare sul link che porta alla entry di Entrez relativo.
- Scegliere dal menu a tendina "display" posto in alto a sinistra l'opzione "FASTA" .
- Selezionare con il mouse la sequenza aminoacidica: partire dal segno ">", con cui comincia l'intestazione della sequenza, e arrivare fino all'ultimo aminoacido; copiare quanto evidenziato cliccando sul menu di Netscape "Edit" e poi sull'opzione "Copy". In questo modo recuperiamo la sequenza aminoacidica in formato FASTA per la successiva ricerca di similarità.
- A questo punto collegarsi all'indirizzo: <u>http://www.ncbi.nlm.nih.gov/BLAST</u>.
- Dalla pagina che si presenta nella box Special selezionare Align two sequences (bl2seq). Si arriva così alla pagina del sottiprogramma di BLAST BLAST 2 SEQUENCES che allinea due date sequenze (nucleotidiche o aminoacidiche).
- Dal menu a tendina denominato "Program" posto in alto a sinistra scegliere "blastp" (allineamento proteina).
- Incollare la sequenza aminoacidica di SRK1_SPOLA (menu "Edit" di Netscape, opzione "Paste") nella casella di input <u>GRANDE</u> sotto la scritta Sequence 2 (parte inferiore della pagina).

- Nella pagina web che contiene la sequenza dell'entry NP_002077 (GRB2) selezionare con il mouse la sequenza aminoacidica: partire dal segno ">", con cui comincia l'intestazione della sequenza, e arrivare fino all'ultimo aminoacido; copiare quanto evidenziato cliccando sul menu di Netscape "Edit" e poi sull'opzione "Copy".
- Incollare la sequenza aminoacidica di GRB2 (menu "Edit" di Netscape, opzione "Paste") nella casella di input <u>GRANDE</u> sotto la scritta Sequence 1 (parte centrale della pagina).
- Cliccare sul tasto "Align" posto in basso a sinistra per effettuare l'allineamento.
- L'allineamento risultante mette meglio in rilievo la distribuzione degli allineamenti locali tra le due sequenze con una immagine riassuntiva. In questo caso si può notare (vedi Figura che segue) come la sequenza SRK1_SPOLA ha una zona di allineamento locale tra gli amminoacidi 57-111 (circa) che si allinea su GRB2 con gli amminoacidi nella regione 1-57 (circa) e 158-210.
- L'interpretazione dell'allineamento e la statistica associata è uguale a quella di BLASTP.

Quale dei due allineamenti parziali ottenuti e' il piu' significativo?

S NCBI Blast 2 Sequences results										
PubMed Entrez BLAST OMIM Taxonomy Structure										
BLAST 2 SEQUENCES RESULTS VERSION BLASTP 2.2.2 [Dec-14-2001]										
Matrix BLOSUM62 🔽 gap open: 11 gap extension: 1 x_dropoff: 50 expect: 10.0 wordsize: 3 <u>Filter</u> 🔽 <u>Align</u>										
Sequence 1 gi 4504111 growth factor receptor-bound protein 2 [Homo sapiens] Length 217 (1217)										
Sequence 2 gi 1174436 TYROSINE-PROTEIN KINASE SRK1 Length 505 (1505)										
NOTE The statistics (bitscore and expect value) is calculated based on the size of nr database										
Score = 107 bits (267), Expect = $6e-23$ Identities = $57/141$ (40%), Positives = $91/141$ (64%), Gans = $8/141$ (5%)										
Query: 4 IAKYDFKATADDELSFKRGDILKVLNEECDQNWYKAELNGKDGFIPKNYIEMKP 57 + KYD+ + DD+LSFK+GD++ +++ + W+ + GK+G+IP NY+ + Sbjct: 60 VGKYDYDSRTDDDLSFKKGDLMYIISTDEGDWWFARSKDTAGKEGYIPSNYVAEYKSLDA 119										
Query: 58 HPWFFGKIPRAKAEEMLSKQRHD-GAFLIRESESAPGDFSLSVKFGNDVQHFKVLRDGAG 116 WF GKI R +MF+ML++ + G+FLIR+SF+ PGDFSLSVK + V+H++V R G										
Sbjct: 120 EEWFLGKIKRVEAEKMLNQSFNQVGSFLIRDSETTPGDFSLSVKDQDRVRHYRVRRLEDG 179										
Query: 117 KYFLW-VVKFNSLNELVDYHR 136										
Sbjet: 180 SLFVTRRSTFQILHELVDHYK 200										
Score = 42.0 bits (97), Expect = 0.003 Identities = 18/56 (32%), Positives = 31/56 (55%), Gaps = 4/56 (7%)										
Query: 158 PTYVQALFDFDPQEDGELGFRRGDFIHVMDNSDPNWWKGACHGQTGMFPRNYV 210 P YV +D+D + D +L F++GD ++++ + +WW G+ G P NYV Sbjct: 57 PIYV-GKYDYDSRTDDDLSFKKGDLMYIISTDEGDWWFARSKDTAGKEGYIPSNYV 111										
CPU time: 0.06 user secs. 0.03 sys. secs 0.09 total secs.										

6) RICERCA DI SIMILARITA' CON LA SEQUENZA AMINOACIDICA DI GRB2 UTILIZZANDO IL PROGRAMMA FASTA

Andare all'indirizzo <u>http://www.ebi.ac.uk/fasta33/</u>. Incollare la sequenza amminoacidica per la proteina **GRB2** (growth factor receptor-bound protein 2, *Homo sapiens*) nella finestra grande.

Lasciare tutti i valori di defaults per quanto riguarda il programma, il database (tutte le informazioni relative sono disponibili cliccando su PROGRAM e DATABASES), la matrice di scoring e le penalizzazioni per inserzioni/delezioni e "protein" in corrispondenza di "molecule type". Quindi cliccare su "Run Fasta3".

NOTA. alternativamente potete impostare uno o piu' parametri diversamente da un vostro collega dello stesso "gruppo di lavoro" e verificare se e come i risultati ottenuti cambino a seguito di queste diverse impostazioni (es. diversa matrice di scoring, diverse penalizzazioni per i gap, diverso database etc.).

Summary Table Sequence - Microsoft Internet Explorer													<u>_ 8 ×</u>				
File Edit View Favorites Tools Help													-				
🕀 Back 🔻	🔿 - 🔕 😰 🖄 🔕 Search	🗼 Favorite	s 🌚 Me	dia 🎯	B- 3	W •											
Address 🧧	http://www.ebi.ac.uk/cgi-bin/sumta	b?tool=fasta	&jobid=fas	sta-200603	20-14582919)									_	ୖ୶୕	Links »
Searcl × Ck Ne »	Help General Help															-	
Search	• Formats	SUBMISSION PARAMETERS															
The ₩eb	• Gaps	Title				Sequence Database											
	• Matrix	Sequence length				217		Sequence type			p	p					
Find a Web page	References	Program			fasta Ve		Vers	Version			3.4t2	3.4t25 Sept 2, 2005					
	• Fasta Help	Expectation upper value			10.0 Matrix			BL50									
	 MView Help 	Sequence range			1- Number of scores				50								
Search Brought	VisualFasta Help	Number of alignments			50 Word size					2							
		Open gap penalty			-10 Gap extension penalty			enalty	-2								
you by	UniProt	Histogram			false												
MSN Search	▶ UniParc	Show Annotation Fasta (Result MView VisualFasta X				ML	SUBMIT ANOTHER JOB						
		Show Alignments Clear all C						Check all	ck all Invert selection Reset								
Search for																	
other items:																	
Folders		Alignment DB:ID				Source Len					<u>Length</u>	Identity%	<u>Simi</u>	lar%	<u>Overlap</u>	<u>E0</u>	
People		1 🗹	UNIPROT:Q2PG25_MACFA			Hypothetical protein. Growth factor receptor bound			217	100.000	100.0	000	217	2.1e-97			
0		2 🗹							217	100.000	0.000 100.0		217	2.1e-97			
2006 Microso		3 🗹	UNIPROT:GRB2_HUMAN			Growth factor receptor-bound				217	100.000	100.0	000	217	2.1e-97		
MSN Privacu		4 🗹	UNIPROT:Q3T0F9_BOVIN			Hypothetical protein MGC12				217	100.000	100.0	000	217	2.1e-97		
- maby		5 V UNIPROT:GRB2_PONPY Growth factor recept				eptor-bound 217		100.000).000 100.000		217	2.1e-97					
			UNIPROT:GRB2_RAT			Growth factor receptor-bound p				217	100.000	100.0	000	217	2.1e-97		
		7 🔽	UNIPRO	T:GRB2	MOUSE	Growth	n factor rec	eptor-	-bound		217	99.539	100.0	000	217	4e-97	-
ど Done															🥝 Interr	net	

La pagina dei risultati si presenta come mostrato di seguito.

Analogamente ai risultati di BLAST, nella prima parte troviamo informazioni riguardo alle impostazioni usate per la ricerca di omologie e alla sequenza query (titolo, se impostato e lunghezza).

A seguire troviamo l'elenco delle sequenze nel database proteico Uniprot, impostato per la ricerca, identificate da FASTA come possibili omologhe della nostra query. Come in BLAST, le sequenze sono ordinate per E-value crescente, cioe' per significativita' dell'allineamento decrescente. Cliccando sul codice identificativo e' possibile accedere alle

voci corrispondenti alle relative proteine in Uniprot. Sono anche riportate brevemente, sotto Source, nome e origine delle diverse proteine trovate e, a seguire, la loro lunghezza in amminoacidi, la percentuale di identita' di sequenza e di similarita' con la query, la lunghezza dell'allineamento (sotto Overlap) e naturalmente l'E-value.

Cliccando su "Show alignments" e' possibile visualizzare gli allineamenti a coppie della sequenza query con ciscuna delle sue possibili omologhe.

Cliccando su "VisualFasta" e' possibile visualizzare gli stessi allineamenti a coppia, anticipati da una figura assai simile a quella gia' vista e commentata per BLAST.

Cliccando su "MView" e' anche possibile visualizzare un allineamento a colori di tutte le sequenze trovate rispetto alla query. Sotto ciascun blocco dell'allineamento e' inoltre riportata la corrispondente sequenza *consensus* a diverse soglie di percentuali di identita'. Notate quali sono gli amminoacidi piu' conservati in questo insieme di possibili proteine omologhe.

Cliccando su "Show Annotation" e' possibile visualizzare una annotazione dettagliata per ciascuna sequenza identificata da FASTA.

Prendere nota di eventuali ortologhi e del loro grado di similarita' con la sequenza query.