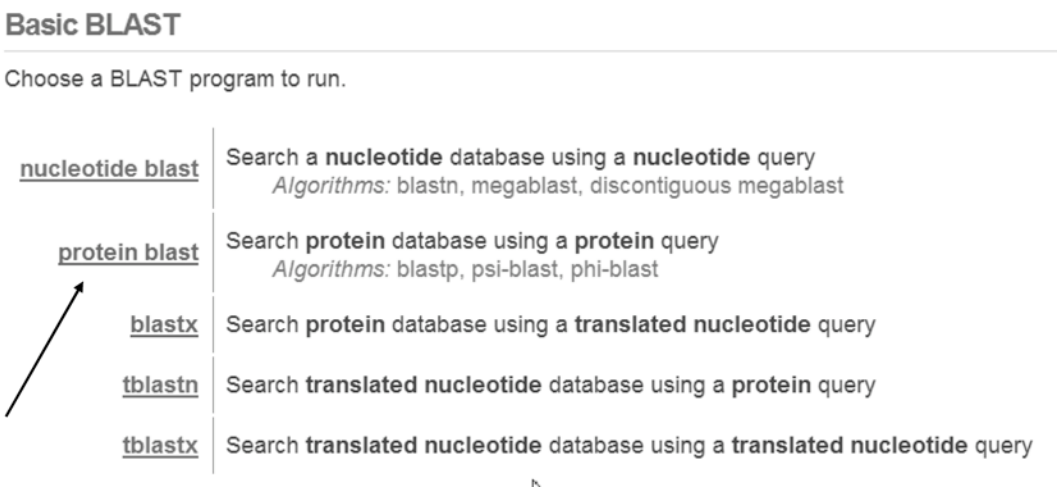
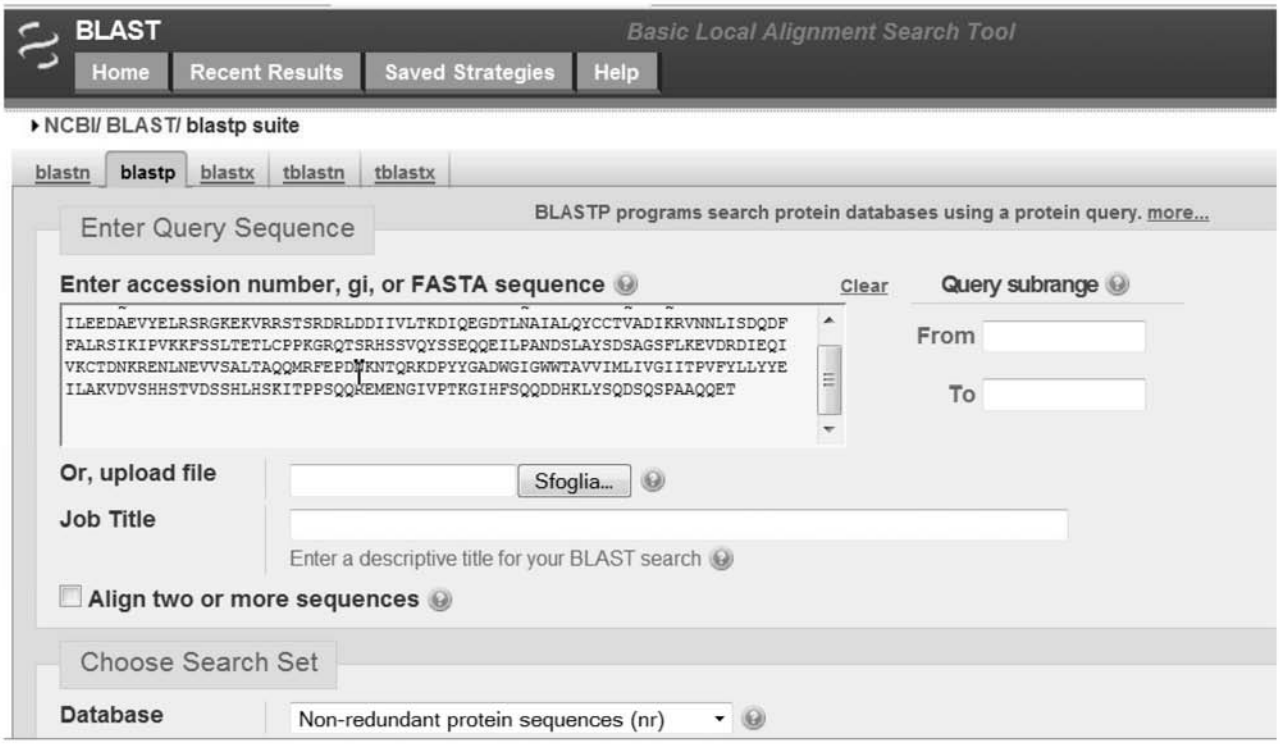
4. Ricerca di sequenze in banche dati e allineamento multiplo

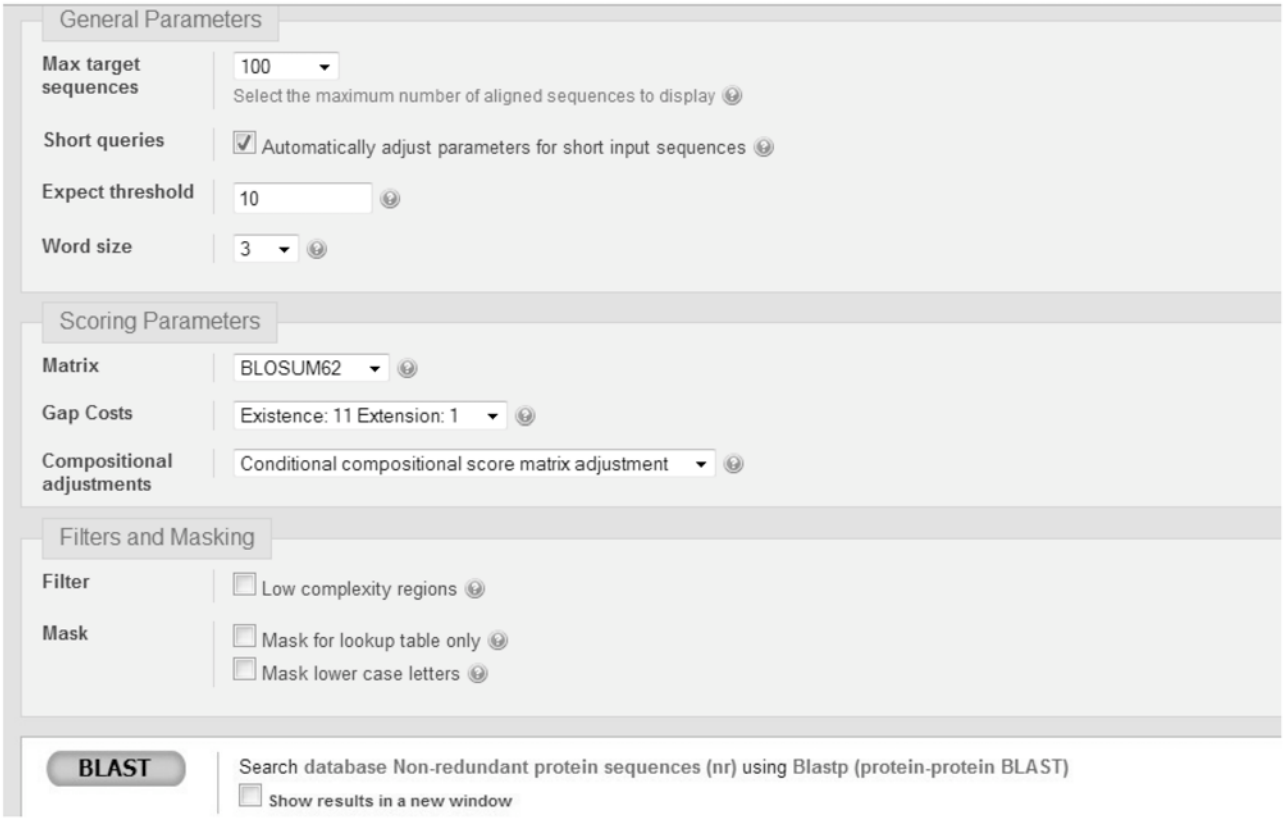
- Collegatevi al sito www.ncbi.nlm.nih.gov/BLAST. Apparirà una pagina nella quale le versioni di BLAST disponibili sono organizzate in base al tipo di ricerca che si desidera effettuare. Selezionate *protein-protein BLAST* (BLASTP), come mostrato nella seguente figura (freccia):



Aprite ora una nuova finestra del browser e collegatevi al sito dell'NCBI (www.ncbi.nlm.nih.gov) ed entrate nella sezione relativa alla banca dati di proteine. Cercate la proteina con codice CAD97936. Come si vedrà, questa è una proteina ipotetica di *Homo sapiens*. Incollate la sequenza in formato FASTA nel campo *Search* della pagina di BLAST (la prima riga di annotazione non va inserita. Alternativamente, si può inserire nel campo direttamente il codice GenInfo o il codice di accessione della proteina):



In fondo alla pagina c'è un collegamento (*Algorithm parameters*) attraverso il quale si apre la seguente interfaccia con i parametri utilizzati da BLAST che possono essere modificati all'occorrenza dall'utente (per l'interpretazione dei parametri si veda il Capitolo 5).



Lanciate il programma (cliccando sul pulsante *BLAST*). Dopo un breve lasso di tempo appariranno i risultati della ricerca. Il primo risultato che viene riportato riguarda la ricerca di domini conservati lungo la sequenza sonda (Figura 1) utilizzando la banca dati Conserved Domains (CDD).

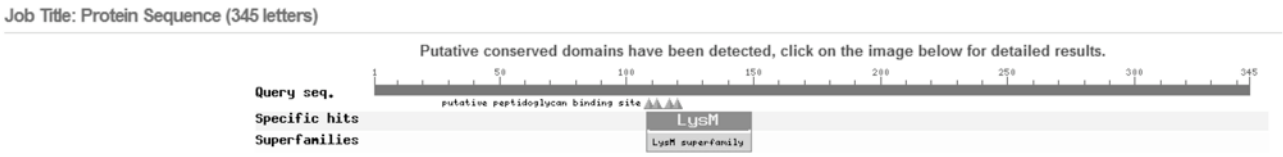


Figura 1 Risultato della ricerca nella banca dati CDD. La figura indica la presenza lungo la sequenza di un dominio noto nella banca dati come "LysM". L'icona del dominio può essere attivata attraverso mouse e accedere in questo modo a ulteriori informazioni sul dominio.

La ricerca utilizza RPS-BLAST (Capitolo 8). Il risultato che appare successivamente riporta l'elenco delle sequenze significativamente simili alla sequenza sonda. L'interpretazione della schermata può essere effettuata seguendo la descrizione riportata nel Capitolo 5. L'elenco delle proteine (Figura 2) contiene i collegamenti alle informazioni relativa a

ciascuna proteina contenute nella banca dati e all'allineamento con la sequenza sonda (Figura 3).

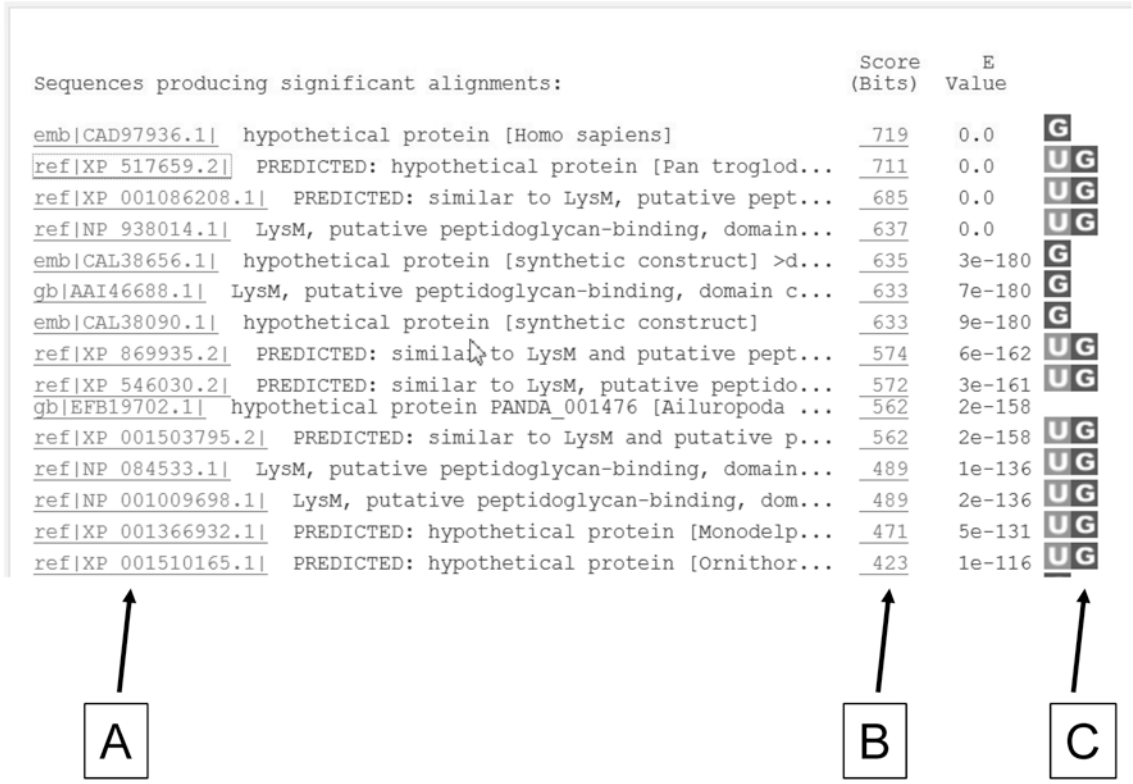
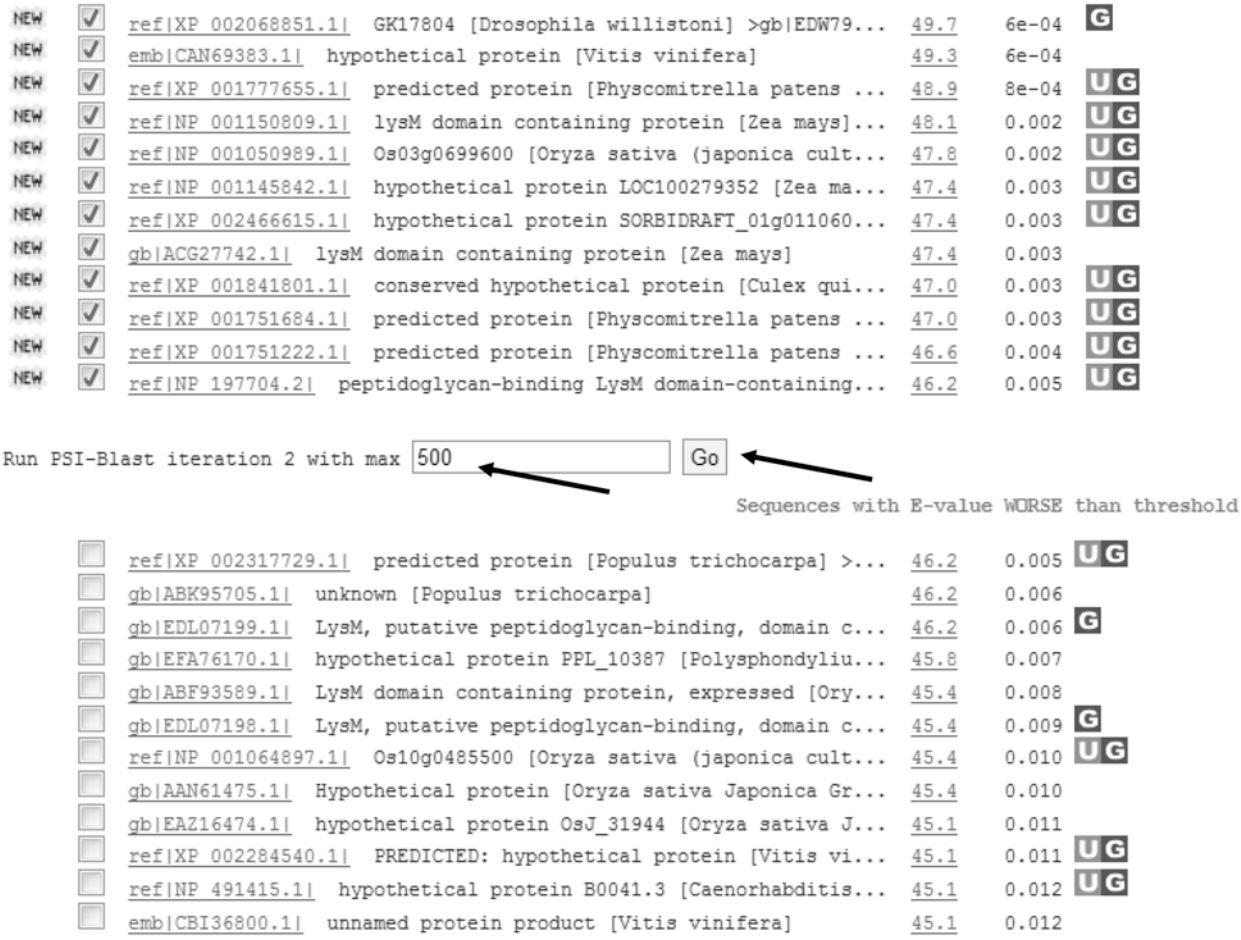


Figura 2 Parte della lista di sequenze risultate significativamente simili a quella sonda. Attivando i collegamenti dei codici (freccia A) si accede alla scheda della sequenza nella banca dati di proteine. Il collegamento del punteggio (freccia B) rimanda invece all'allineamento fatto da BLAST tra la sequenza sonda e la sequenza bersaglio. I collegamenti indicati dalla freccia C in corrispondenza di ciascuna sequenza rimandano ad altre banche dati contenenti informazioni sulla sequenza del gene (Entrez Gene) o del trascritto (UNIGENE)

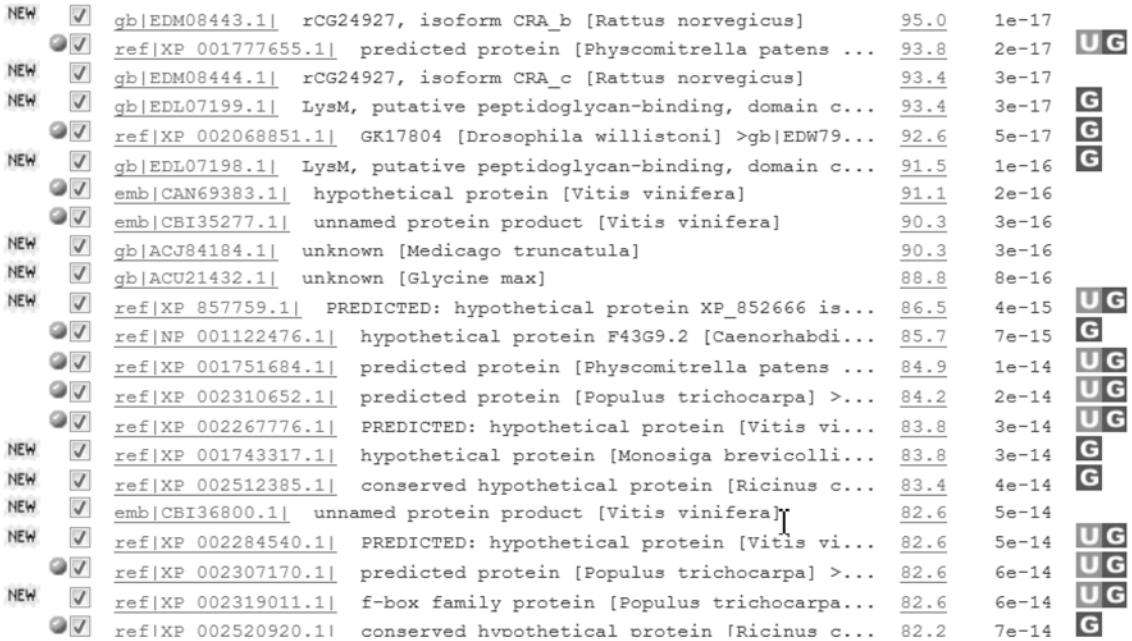
```
>ref|XP_001510165.1| PREDICTED: hypothetical protein [Ornithorhynchus anatinus]
Length=381

 GENE ID: 100079181 LOC100079181 | hypothetical protein LOC100079181
[Ornithorhynchus anatinus]

 Score =  423 bits (1088),  Expect = 1e-116, Method: Compositional matrix adjust.
 Identities = 215/328 (65%), Positives = 252/328 (76%), Gaps = 6/328 (1%)

Query  18   GGGPFGHLLAECSLLTGTDFNIMAGRHQNRSFPLPGVQS--SGQVHAFGNCSDSDILEED  75
            GG    HL A+       +F +MAGR QNRSF    VQ   +  ++ FGN +D DI EED
Sbjct  55   GGTKKTHLFAQAFW---EEFKMMAGRSQNRSFHGAAVQPVVNSHMYPFGNNTDPDISEED  111

Query  76   AEVYELRSRGKEKVRRSTSRDRLDDIIVLTKDIQEGDTLNAIALQYCCTVADIKRVNNLI  135
             EVYELR RG+EK RRS+SRDR DDI++LTKDIQEGDTL AIALQYCC+VADIKRVNNLI
Sbjct  112  GEVYELRPRGREKNRRSSSRDRCDDIVLLTKDIQEGDTLIAIALQYCCSVADIKRVNNLI  171

Query  136  SDQDFFALRSIKIPVKKFSSLTETLCPPKGRQTSRHSSVQYSSEQQEILPANDSLAYSDS  195
            SDQDFFALRS+KIPVKKFS LTET   PKGR     ++   +   Q+  PA+D  + +++
Sbjct  172  SDQDFFALRSVKIPVKKFSVLTETHYSPKGRPPLHPAAAADAPGPQDAAPASDPSSPNET  231

Query  196  AGSFLKEVDRDIEQIVKCTDNKRENLNEVVSALTAQQMRFEPDNKNTQRKDPYYGADWGI  255
            AG FLKEVDRDIEQIV+CTD K+ENLNEVVSAL  QQ+ FEP+ K+ +RKDPYYGADWGI
Sbjct  232  AGGFLKEVDRDIEQIVRCTDTKKENLNEVVSALATQQVCFEPEGKSVRRKDPYYGADWGI  291

Query  256  GWWTAVVIMLIVGIITPVFYLLYYEILAKVDVSHHSTVDSSHLHSKITPPSQQREMENGI  315
            GWWTAVVIMLIVGIITPVFYLLYYE+L KVDVSHHSTV+SS  HS +TPPS QRE+ NG
Sbjct  292  GWWTAVVIMLIVGIITPVFYLLYYEVLVKVDVSHHSTVESSQSHSGVTPPSPQREVGNGP  351
```

Figura 3 Parte dei risultati contenenti gli allineamenti tra la sequenza sonda (Query) e la sequenza della banca dati (Sbjct). La riga centrale marcata dalla freccia indica le posizioni in cui i residui delle due sono identici (viene visualizzato il residuo) o affini (in questo caso viene riportato un «+»). Le posizioni non conservate rimangono vuote.

- Tornate alla pagina di BLASTP, ma questa volta selezionate PSI-BLAST nel campo *Program selection* e attivate la ricerca cliccando sul solito pulsante BLAST. Dopo un breve lasso di tempo apparirà il risultato che si interpreta in modo del tutto simile a quanto visto per BLAST nel punto precedente. La differenza principale risiede nel fatto che adesso alcune sequenze sono etichettate con un bollino giallo contente la parola *new*. Le sequenze che mostrano un *E-value* superiore alla soglia stabilita o che sono oltre il numero massimo prefissato (nell'esempio 500) nei parametri del programma (e sotto il controllo dell'utente) invece non sono etichettate, come mostra la seguente figura:



Le prime verranno utilizzate per costruire la PSSM per la seconda iterazione di ricerca che si attiva con il pulsante *Go* (frecce nella figura precedente). La lista risultante da questa ricerca contiene alcune sequenze marcate con un bollino verde (che sono le stesse trovate nella prima ricerca e sono state utilizzate per la PSSM) e altre con un bollino giallo (si veda la figura successiva). Queste ultime sono sequenze "nuove", sequenze cioè che nella prima iterazione ottenevano un *E-value* superiore alla soglia di significatività e ora invece risultano al di sotto. Tutte le sequenze riportate al di sotto della soglia ma comprese nel numero massimo prestabilito saranno utilizzate per calcolare la nuova PSSM dell'iterazione successiva.

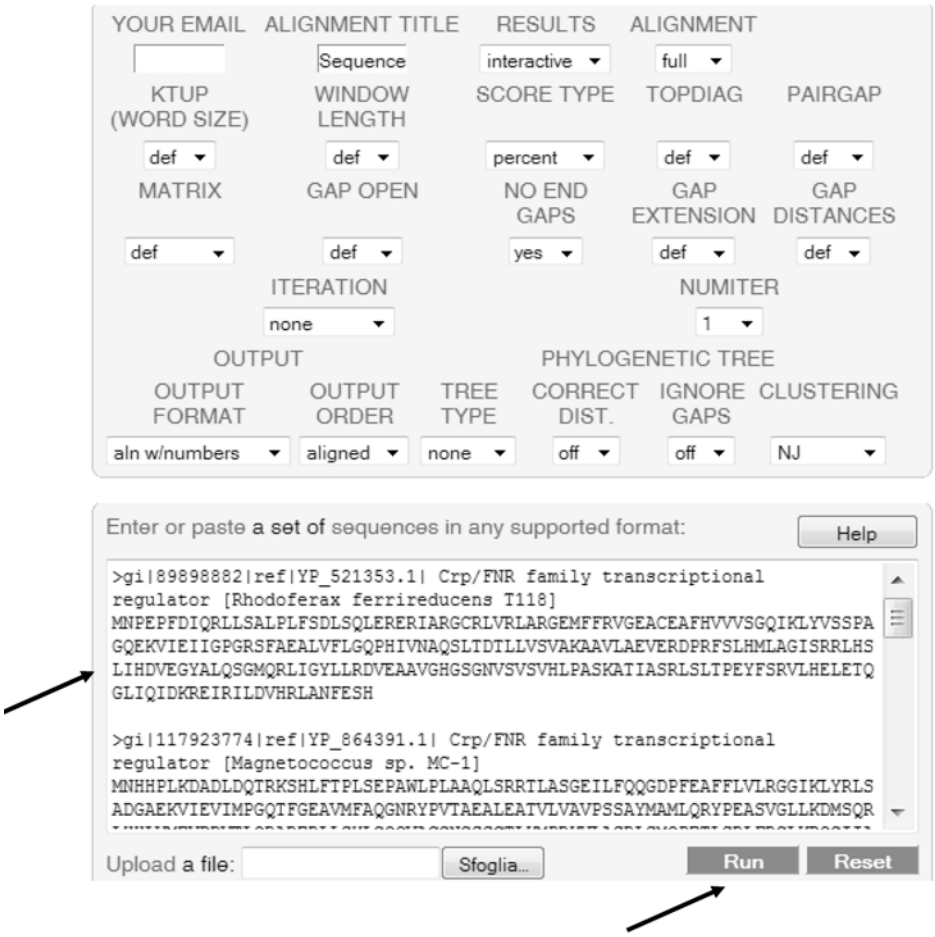


NEW	gb|EDM08443.1|	rCG24927, isoform CRA_b [Rattus norvegicus]	95.0	1e-17
	ref|XP_001777655.1|	predicted protein [Physcomitrella patens ...	93.8	2e-17
NEW	gb|EDM08444.1|	rCG24927, isoform CRA_c [Rattus norvegicus]	93.4	3e-17
NEW	gb|EDL07199.1|	LysM, putative peptidoglycan-binding, domain c...	93.4	3e-17
	ref|XP_002068851.1|	GK17804 [Drosophila willistoni] >gb|EDW79...	92.6	5e-17
NEW	gb|EDL07198.1|	LysM, putative peptidoglycan-binding, domain c...	91.5	1e-16
	emb|CAN69383.1|	hypothetical protein [Vitis vinifera]	91.1	2e-16
	emb|CBI35277.1|	unnamed protein product [Vitis vinifera]	90.3	3e-16
NEW	gb|ACJ84184.1|	unknown [Medicago truncatula]	90.3	3e-16
NEW	gb|ACU21432.1|	unknown [Glycine max]	88.8	8e-16
NEW	ref|XP_857759.1|	PREDICTED: hypothetical protein XP_852666 is...	86.5	4e-15
	ref|NP_001122476.1|	hypothetical protein F43G9.2 [Caenorhabdi...	85.7	7e-15
	ref|XP_001751684.1|	predicted protein [Physcomitrella patens ...	84.9	1e-14
	ref|XP_002310652.1|	predicted protein [Populus trichocarpa] >...	84.2	2e-14
	ref|XP_002267776.1|	PREDICTED: hypothetical protein [Vitis vi...	83.8	3e-14
NEW	ref|XP_001743317.1|	hypothetical protein [Monosiga brevicolli...	83.8	3e-14
NEW	ref|XP_002512385.1|	conserved hypothetical protein [Ricinus c...	83.4	4e-14
NEW	emb|CBI36800.1|	unnamed protein product [Vitis vinifera]	82.6	5e-14
NEW	ref|XP_002284540.1|	PREDICTED: hypothetical protein [Vitis vi...	82.6	5e-14
	ref|XP_002307170.1|	predicted protein [Populus trichocarpa] >...	82.6	6e-14
NEW	ref|XP_002319011.1|	f-box family protein [Populus trichocarpa...	82.6	6e-14
	ref|XP_002520920.1|	conserved hypothetical protein [Ricinus c...	82.2	7e-14

- Recuperate dalla banca dati NCBI le seguenti sequenze di regolatori trascrizionali in formato FASTA, i cui codici sono:

  YP_521353.1, YP_864391.1, YP_286398.1, NP_249218.1
  YP_316351.1, YP_284886.1, ZP_00942609.1

  Collegatevi al sito www.ebi.ac.uk/Tools/clustalw2/ e incollare le 7 sequenze in formato FASTA (compresa la prima linea di annotazione che inizia con ">") nell'apposito campo:



Eseguite il programma cliccando su *Run*. In poco tempo comparirà la seguente pagina dei risultati:

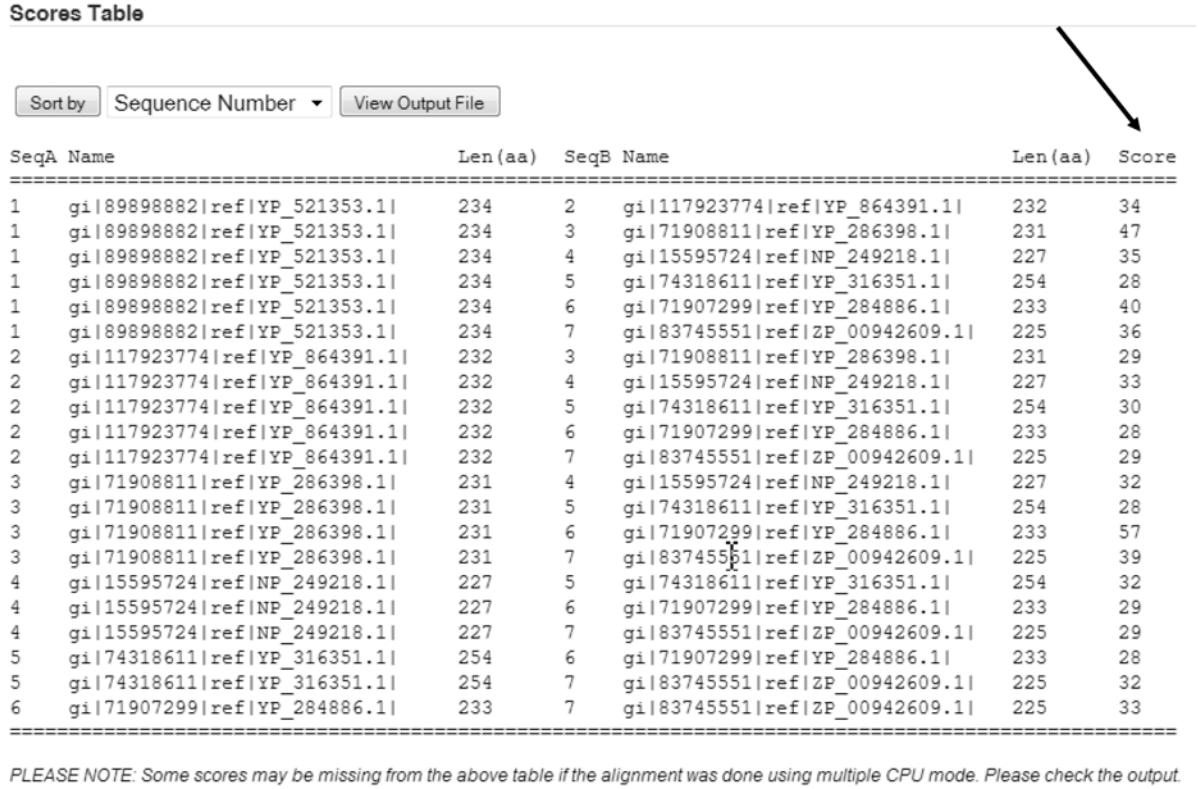
ClustalW2 Results

Results of search	
Number of sequences	7
Alignment score	8685
Sequence format	Pearson
Sequence type	aa
JalView	Start Jalview
Output file	clustalw2-20091222-1207153770.output
Alignment file	clustalw2-20091222-1207153770.aln
Guide tree file	clustalw2-20091222-1207153770.dnd
Your input file	clustalw2-20091222-1207153770.input

SUBMIT ANOTHER JOB

To save a result file right-click the file link in the above table and choose "Save Target As".
If you cannot see the JalView button, reload the page and check your browser settings to enable Java Applets.

Nella prima parte della pagina sono riportati i collegamenti ai file contenenti i risultati di ClustalW in formato testuale. Nella seconda parte (figura seguente) è riportata la matrice delle distanze in cui a ciascuna coppia di sequenze è associato il punteggio che ne misura la distanza:



Scores Table

Sort by | Sequence Number | View Output File

SeqA	Name	Len(aa)	SeqB	Name	Len(aa)	Score
1	gi|89898882|ref|YP_521353.1|	234	2	gi|117923774|ref|YP_864391.1|	232	34
1	gi|89898882|ref|YP_521353.1|	234	3	gi|71908811|ref|YP_286398.1|	231	47
1	gi|89898882|ref|YP_521353.1|	234	4	gi|15595724|ref|NP_249218.1|	227	35
1	gi|89898882|ref|YP_521353.1|	234	5	gi|74318611|ref|YP_316351.1|	254	28
1	gi|89898882|ref|YP_521353.1|	234	6	gi|71907299|ref|YP_284886.1|	233	40
1	gi|89898882|ref|YP_521353.1|	234	7	gi|83745551|ref|ZP_00942609.1|	225	36
2	gi|117923774|ref|YP_864391.1|	232	3	gi|71908811|ref|YP_286398.1|	231	29
2	gi|117923774|ref|YP_864391.1|	232	4	gi|15595724|ref|NP_249218.1|	227	33
2	gi|117923774|ref|YP_864391.1|	232	5	gi|74318611|ref|YP_316351.1|	254	30
2	gi|117923774|ref|YP_864391.1|	232	6	gi|71907299|ref|YP_284886.1|	233	28
2	gi|117923774|ref|YP_864391.1|	232	7	gi|83745551|ref|ZP_00942609.1|	225	29
3	gi|71908811|ref|YP_286398.1|	231	4	gi|15595724|ref|NP_249218.1|	227	32
3	gi|71908811|ref|YP_286398.1|	231	5	gi|74318611|ref|YP_316351.1|	254	28
3	gi|71908811|ref|YP_286398.1|	231	6	gi|71907299|ref|YP_284886.1|	233	57
3	gi|71908811|ref|YP_286398.1|	231	7	gi|83745551|ref|ZP_00942609.1|	225	39
4	gi|15595724|ref|NP_249218.1|	227	5	gi|74318611|ref|YP_316351.1|	254	32
4	gi|15595724|ref|NP_249218.1|	227	6	gi|71907299|ref|YP_284886.1|	233	29
4	gi|15595724|ref|NP_249218.1|	227	7	gi|83745551|ref|ZP_00942609.1|	225	29
5	gi|74318611|ref|YP_316351.1|	254	6	gi|71907299|ref|YP_284886.1|	233	28
5	gi|74318611|ref|YP_316351.1|	254	7	gi|83745551|ref|ZP_00942609.1|	225	32
6	gi|71907299|ref|YP_284886.1|	233	7	gi|83745551|ref|ZP_00942609.1|	225	33

PLEASE NOTE: Some scores may be missing from the above table if the alignment was done using multiple CPU mode. Please check the output.

Segue l'allineamento multiplo e l'albero guida codificato sia in formato testuale che grafico:

Lo stesso sito mette a disposizione alcuni programmi alternativi di allineamento multiplo. In particolare sono disponibili T-COFFEE (www.ebi.ac.uk/Tools/t-coffee), MAFFT (www.ebi.ac.uk/Tools/mafft/) e MUSCLE (www.ebi.ac.uk/Tools/muscle/). L'interfaccia utente è molto simile a quella utilizzata per Clustal.
Ricalcolate l'allineamento delle sequenze elencate con i tre programmi e confrontate i risultati con quelli ottenuti con Clustal. Noterete che le maggiori differenze sono localizzate nelle regioni N-terminali e C-terminali dell'allineamento e nelle zone in cui sono presenti inserzioni e delezioni.