

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):

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms:</i> blastn, megablast, discontinuous megablast
protein blast	Search protein database using a protein query <i>Algorithms:</i> blastp, psi-blast, phi-blast
blastx	Search protein database using a translated nucleotide query
tblastn	Search translated nucleotide database using a protein query
tblastx	Search translated nucleotide database using a translated nucleotide query

Apriete 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):

The screenshot shows the NCBI BLASTP interface. At the top, there is a navigation bar with 'Home', 'Recent Results', 'Saved Strategies', and 'Help'. Below this, the 'blastp' option is selected under the 'NCBI/BLAST/blastp suite' section. The main area is titled 'Enter Query Sequence' and contains a text input field with the following FASTA sequence:

```
IILEEDAEVVELRSRGKEKVRVRSRDRLLDDIIVLTKDIQEGDTLNAIALQYCCTVADIKRVNLI SDQDF
FALRSIKIPVKFSSLTETLCPFKGRQTSRHSSVQYSSEQQEILPANDSLAYS SDSAGSFLKEVD RDIEQI
VKCTDNKRENLENEVVSALTAQMRFEFDYKNTQRKDPYYGADWGI GWTAVVIMLIVGLITPVFYL LYE
ILARVDVSHHSTVDSSHLHSKITPPSQQREMENGIVPTKGIHFSQQDDHRLYSQDSQPAAQQET
```

 To the right of the input field are 'Clear' and 'Query subrange' options. Below the input field, there are fields for 'Or, upload file' (with an 'Sfoggia...' button), 'Job Title' (with a descriptive title prompt), and a checkbox for 'Align two or more sequences'. At the bottom, the 'Choose Search Set' section shows the 'Database' dropdown set to 'Non-redundant protein sequences (nr)'.

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).

General Parameters

Max target sequences: 100
Select the maximum number of aligned sequences to display

Short queries: Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 3

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

Filter: Low complexity regions

Mask: Mask for lookup table only
 Mask lower case letters

BLAST Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
 Show results in a new window

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).

Sequences producing significant alignments:			Score (Bits)	E Value	
emb CAD97936.1 	hypothetical protein [Homo sapiens]		719	0.0	G
ref XP_517659.2 	PREDICTED: hypothetical protein [Pan troglod...		711	0.0	UG
ref XP_001086208.1 	PREDICTED: similar to LysM, putative pept...		685	0.0	UG
ref NP_938014.1 	LysM, putative peptidoglycan-binding, domain...		637	0.0	UG
emb CAL38656.1 	hypothetical protein [synthetic construct] >d...		635	3e-180	G
gb AAI46688.1 	LysM, putative peptidoglycan-binding, domain c...		633	7e-180	G
emb CAL38090.1 	hypothetical protein [synthetic construct]		633	9e-180	G
ref XP_869935.2 	PREDICTED: similar to LysM and putative pept...		574	6e-162	UG
ref XP_546030.2 	PREDICTED: similar to LysM, putative peptido...		572	3e-161	UG
gb EFB19702.1 	hypothetical protein PANDA_001476 [Ailuropoda ...		562	2e-158	
ref XP_001503795.2 	PREDICTED: similar to LysM and putative p...		562	2e-158	UG
ref NP_084533.1 	LysM, putative peptidoglycan-binding, domain...		489	1e-136	UG
ref NP_001009698.1 	LysM, putative peptidoglycan-binding, dom...		489	2e-136	UG
ref XP_001366932.1 	PREDICTED: hypothetical protein [Monodelp...		471	5e-131	UG
ref XP_001510165.1 	PREDICTED: hypothetical protein [Ornithor...		423	1e-116	UG



A



B



C

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| UG 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  GGGPFGHLLAECSSLTGTDFNIMAGRHNQRSFPLPGVQS--SGQVHAFGNCSDSDILEED 75
          GG  HL A+      +F +MAGR QNRSF  VQ  +  ++ FGN +D DI EED
Sbjct 55  GGTKKTHLFAQAFW---EEFKMMAGRSQNRSFHGAAVQPVVNSHMYPFGNNTDPDISEED 111

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

Query 136 SDQDFFALRSIKIPVKKFSSLTETLCPKGRQTSRHSSVQYSSEQQEILPANDSLAYSDS 195
          SDQDFFALRS+KIPVKKFS  LTET  PKGR  ++  +  Q+  PA+D  +  +++
Sbjct 172  SDQDFFALRSVKIPVKKFVLTETHTYSPKGRPPLHPAAAADAPGQDAAPASDPSSPNET 231

Query 196  AGSFLKEVDRDIEQIVKCTDNKRENLNEVVSALTAQQMRFEFDPNKNTQRKDPYYGADWGI 255
          AG  FLKEVDRDIEQIV+CTD  K+ENLNEVVSAL  QQ+  FEP+  K+  +RKDPYYGADWGI
Sbjct 232  AGGFLEKVEVDRDIEQIVRCTDTTKKENLNEVVSALATQQVCFEPEGKSVRRKDPYYGADWGI 291

Query 256  GWWTAVVIMLIVGIITPVFYLLYYEILAKVDVSHHSTVDSSHLHSKITPPSQREMENGI 315
          GWWTAVVIMLIVGIITPVFYLLYYE+L  KVDVSHHSTV+SS  HS  +TPPS  QRE+  NG
Sbjct 292  GWWTAVVIMLIVGIITPVFYLLYYEVLVKVDVSHHSTVESSQSHSGVTPPSPQREVGNP 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 contenente 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:

NEW	<input checked="" type="checkbox"/>	ref XP_002068851.1 	GK17804 [Drosophila willistoni] >gb EDW79...	49.7	6e-04	G
NEW	<input checked="" type="checkbox"/>	emb CAN69383.1 	hypothetical protein [Vitis vinifera]	49.3	6e-04	
NEW	<input checked="" type="checkbox"/>	ref XP_001777655.1 	predicted protein [Physcomitrella patens ...	48.9	8e-04	UG
NEW	<input checked="" type="checkbox"/>	ref NP_001150809.1 	lysM domain containing protein [Zea mays]...	48.1	0.002	UG
NEW	<input checked="" type="checkbox"/>	ref NP_001050989.1 	Os03g0699600 [Oryza sativa (japonica cult...	47.8	0.002	UG
NEW	<input checked="" type="checkbox"/>	ref NP_001145842.1 	hypothetical protein LOC100279352 [Zea ma...	47.4	0.003	UG
NEW	<input checked="" type="checkbox"/>	ref XP_002466615.1 	hypothetical protein SORBIDRAFT_01g011060...	47.4	0.003	UG
NEW	<input checked="" type="checkbox"/>	gb ACG27742.1 	lysM domain containing protein [Zea mays]	47.4	0.003	
NEW	<input checked="" type="checkbox"/>	ref XP_001841801.1 	conserved hypothetical protein [Culex qui...	47.0	0.003	UG
NEW	<input checked="" type="checkbox"/>	ref XP_001751684.1 	predicted protein [Physcomitrella patens ...	47.0	0.003	UG
NEW	<input checked="" type="checkbox"/>	ref XP_001751222.1 	predicted protein [Physcomitrella patens ...	46.6	0.004	UG
NEW	<input checked="" type="checkbox"/>	ref NP_197704.2 	peptidoglycan-binding LysM domain-containing...	46.2	0.005	UG

Run PSI-Blast iteration 2 with max

Sequences with E-value WORSE than threshold

<input type="checkbox"/>	ref XP_002317729.1 	predicted protein [Populus trichocarpa] >...	46.2	0.005	UG
<input type="checkbox"/>	gb ABK95705.1 	unknown [Populus trichocarpa]	46.2	0.006	
<input type="checkbox"/>	gb EDL07199.1 	LysM, putative peptidoglycan-binding, domain c...	46.2	0.006	G
<input type="checkbox"/>	gb EFA76170.1 	hypothetical protein PPL_10387 [Polysphondyliu...	45.8	0.007	
<input type="checkbox"/>	gb ABF93589.1 	LysM domain containing protein, expressed [Ory...	45.4	0.008	
<input type="checkbox"/>	gb EDL07198.1 	LysM, putative peptidoglycan-binding, domain c...	45.4	0.009	G
<input type="checkbox"/>	ref NP_001064897.1 	Os10g0485500 [Oryza sativa (japonica cult...	45.4	0.010	UG
<input type="checkbox"/>	gb AAN61475.1 	Hypothetical protein [Oryza sativa Japonica Gr...	45.4	0.010	
<input type="checkbox"/>	gb EAZ16474.1 	hypothetical protein OsJ_31944 [Oryza sativa J...	45.1	0.011	
<input type="checkbox"/>	ref XP_002284540.1 	PREDICTED: hypothetical protein [Vitis vi...	45.1	0.011	UG
<input type="checkbox"/>	ref NP_491415.1 	hypothetical protein B0041.3 [Caenorhabditis...	45.1	0.012	UG
<input type="checkbox"/>	emb CBI36800.1 	unnamed protein product [Vitis vinifera]	45.1	0.012	

Le prime verranno utilizzate per costruire la PSSM per la seconda iterazione di ricerca che si attiva con il pulsante *Go* (frecche 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	<input checked="" type="checkbox"/>	gb EDM08443.1	rCG24927, isoform CRA_b [Rattus norvegicus]	95.0	1e-17	
	<input checked="" type="checkbox"/>	ref XP_001777655.1	predicted protein [Physcomitrella patens ...	93.8	2e-17	UG
NEW	<input checked="" type="checkbox"/>	gb EDM08444.1	rCG24927, isoform CRA_c [Rattus norvegicus]	93.4	3e-17	
NEW	<input checked="" type="checkbox"/>	gb EDL07199.1	LysM, putative peptidoglycan-binding, domain c...	93.4	3e-17	G
	<input checked="" type="checkbox"/>	ref XP_002068851.1	GK17804 [Drosophila willistoni] >gb EDW79...	92.6	5e-17	G
NEW	<input checked="" type="checkbox"/>	gb EDL07198.1	LysM, putative peptidoglycan-binding, domain c...	91.5	1e-16	G
	<input checked="" type="checkbox"/>	emb CAN69383.1	hypothetical protein [Vitis vinifera]	91.1	2e-16	
	<input checked="" type="checkbox"/>	emb CBI35277.1	unnamed protein product [Vitis vinifera]	90.3	3e-16	
NEW	<input checked="" type="checkbox"/>	gb ACJ84184.1	unknown [Medicago truncatula]	90.3	3e-16	
NEW	<input checked="" type="checkbox"/>	gb ACU21432.1	unknown [Glycine max]	88.8	8e-16	
NEW	<input checked="" type="checkbox"/>	ref XP_857759.1	PREDICTED: hypothetical protein XP_852666 is...	86.5	4e-15	UG
	<input checked="" type="checkbox"/>	ref NP_001122476.1	hypothetical protein F43G9.2 [Caenorhabdi...	85.7	7e-15	G
	<input checked="" type="checkbox"/>	ref XP_001751684.1	predicted protein [Physcomitrella patens ...	84.9	1e-14	UG
	<input checked="" type="checkbox"/>	ref XP_002310652.1	predicted protein [Populus trichocarpa] >...	84.2	2e-14	UG
	<input checked="" type="checkbox"/>	ref XP_002267776.1	PREDICTED: hypothetical protein [Vitis vi...	83.8	3e-14	UG
NEW	<input checked="" type="checkbox"/>	ref XP_001743317.1	hypothetical protein [Monosiga brevicolli...	83.8	3e-14	G
NEW	<input checked="" type="checkbox"/>	ref XP_002512385.1	conserved hypothetical protein [Ricinus c...	83.4	4e-14	
NEW	<input checked="" type="checkbox"/>	emb CBI36800.1	unnamed protein product [Vitis vinifera]	82.6	5e-14	
NEW	<input checked="" type="checkbox"/>	ref XP_002284540.1	PREDICTED: hypothetical protein [Vitis vi...	82.6	5e-14	UG
	<input checked="" type="checkbox"/>	ref XP_002307170.1	predicted protein [Populus trichocarpa] >...	82.6	6e-14	UG
NEW	<input checked="" type="checkbox"/>	ref XP_002319011.1	f-box family protein [Populus trichocarpa...	82.6	6e-14	UG
	<input checked="" type="checkbox"/>	ref XP_002520920.1	conserved hvpothetical protein [Ricinus c...	82.2	7e-14	G

- 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:

YOUR EMAIL: ALIGNMENT TITLE: Sequence RESULTS: interactive ALIGNMENT: full

KTUP (WORD SIZE): def WINDOW LENGTH: def SCORE TYPE: percent TOPDIAG: def PAIRGAP: def

MATRIX: def GAP OPEN: def NO END GAPS: yes GAP EXTENSION: def GAP DISTANCES: def

ITERATION: none NUMITER: 1

OUTPUT: aln w/numbers OUTPUT ORDER: aligned TREE TYPE: none PHYLOGENETIC TREE: off CORRECT DIST.: off IGNORE GAPS: off CLUSTERING: NJ

Enter or paste a set of sequences in any supported format:

```
>gi|89898882|ref|YP_521353.1| Crp/FNR family transcriptional
regulator [Rhodospirillum rubrum T118]
MNPEPFDIQRLLSALPLFSDLSQLERERIARGCRLVRLARGEMFFRVGEACEAFHVSVSGQIKLYVSSPA
GQEKVIEIIGPGRSFAEALVFLGQPHVNAQSLDTLLVSVAKAAVLAEVERDPRFSLHMLAGISRRLHS
LIHDVEGYALQSGMQRLIGYLLRDVEAAVGHGSGNVSVSVHLPASKATIASRLSLTPEYFSRVLHELETQ
GLIQDKREIRILDVHRLANFESH

>gi|117923774|ref|YP_864391.1| Crp/FNR family transcriptional
regulator [Magnetococcus sp. MC-1]
MNHHPKLDADLDQTRKSHLFTPLSEPAWLPLAAQLSRRTLASGEILFQQGDPFEAFFLVLRRGGIKLYRLS
ADGAEKVIEVIMPQTFGEAVMFAQGNRYFVTAEALEATVLVAVPSSAYMAMLRYPEASVGLLKDMSQR
```

Upload a file:

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	<input type="button" value="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
<input type="button" value="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

Sequence Number ▾

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.