

Modulo B4

Analisi bioinformatica delle sequenze amminoacidiche e nucleotidiche



2023-24

Similarità, omologia e identità fra sequenze proteiche

Identità grado di corrispondenza esatta fra due sequenze proteiche allineate (tenendo anche conto di **GAP**). (gap = lacuna)

$I \geq 25\%$ implica simili funzioni $18 \leq I \leq 25$ (grey area) possibile similarità
(NB 2 sequenze casuali con > 100 residui possono avere un'identità > 25%)

Similarità grado di somiglianza generale fra due sequenze (tiene conto di **gap** e AA simili)

Omologia discendenza da una sequenza ancestrale comune

identità vs. similarità

identity → exact match fra due sequenze nucleotidiche o peptidiche (considerando gaps)

similarity → assomiglianza fra due sequenze **maggiore di quella possibilmente casuale**.

Metodi di confronto sequenze basati su identità e similarità

- L'allineamento di sequenze è un metodo *bioinformatico* per confrontare due o più sequenze amminoacidiche o nucleotidiche.

METQGKKLKRNHGPPGLFVDRTYNSTWKKSVLGQILW

| | : | | | : | | | | | | : | | | | : | | | | | | | | |

MESQGKRLIVNNGPPGIFVDRTFNSTWKYIKLGQILW

- Permette di **individuare regioni identiche o simili** che potrebbero avere **relazioni filogenetiche** (evolutive) e/o **funzionali e strutturali**.
- Questo metodo è anche molto utile per determinare se una sequenza di interesse è presente nelle **database di sequenze conosciute**, oppure se esiste una sequenza simile in un altro organismo.

Query sequence

METQGKKLKRNHGPPGLFVDRTYNSTWKKSVLGQILW

Database

MHSSIVLATLVFAIASASKTRELCKMSLEHAKVGTSKEAQDGDIL
YKHMFEHYAPAMKKYFKHRENYTPADVQKDPFFIKQGNILLACHV
LCATYDDRETFDAYVYGELMARHERDHVKVPNDVWNHFWHEHFIEF
LGSK

TTLDEPTKHAWQEIGKFSHEISHHGRHSVRDHCMSLEYIAIGDK
EHQKQNGIDLKMHFEHYPHMRKAFKGFENFTKEDVQKDAFFVN
KDTRFCWPVFCCSSYDDEPTFDYFVDALMDRIKDDIHLPEQ
WHEFWK

LFAEYLNEKSHQHLTEAEKHAWSTIGEDFAHEADKHAKAEKDHE
GEHKEEHHKLGGENAMKAAVPLFYKKVLADERVKHFFKNTDMDH
QTKQQTDFLTMLLGGPNHYKGKNNMT

EAHKGMNLQNLHFDAIENLAATLKELGVTDAVINEAAKVIEHTRK
DMLGKTINIIKATVPVLKEHGVTTITTFYKNLFAKHPEVRPLFDMGR
QESLEQPKALAMTVLAAQNIEN

LPAILPAVKKIADVQAGVAAAHYPIVQELLGAKEVLGDAATD
DILDWAGKAYGVADVFQVEADLYAQAVEQELLGAKEVLGDAAT
DDILDWAGKAYGVADVFQVEADLAQAVE

EPTKHAWQEIGKFSHEISHHGRHSVRDHCMSLEYIAIGDKEHQ
KQNGIDLKMHFEHYPHMRKAFKGFENFTKEDVQKDAFFVNKDT
RCIWRPVCSSYDDEPTFDYFVDALMDRIKDDIHLPEQFW

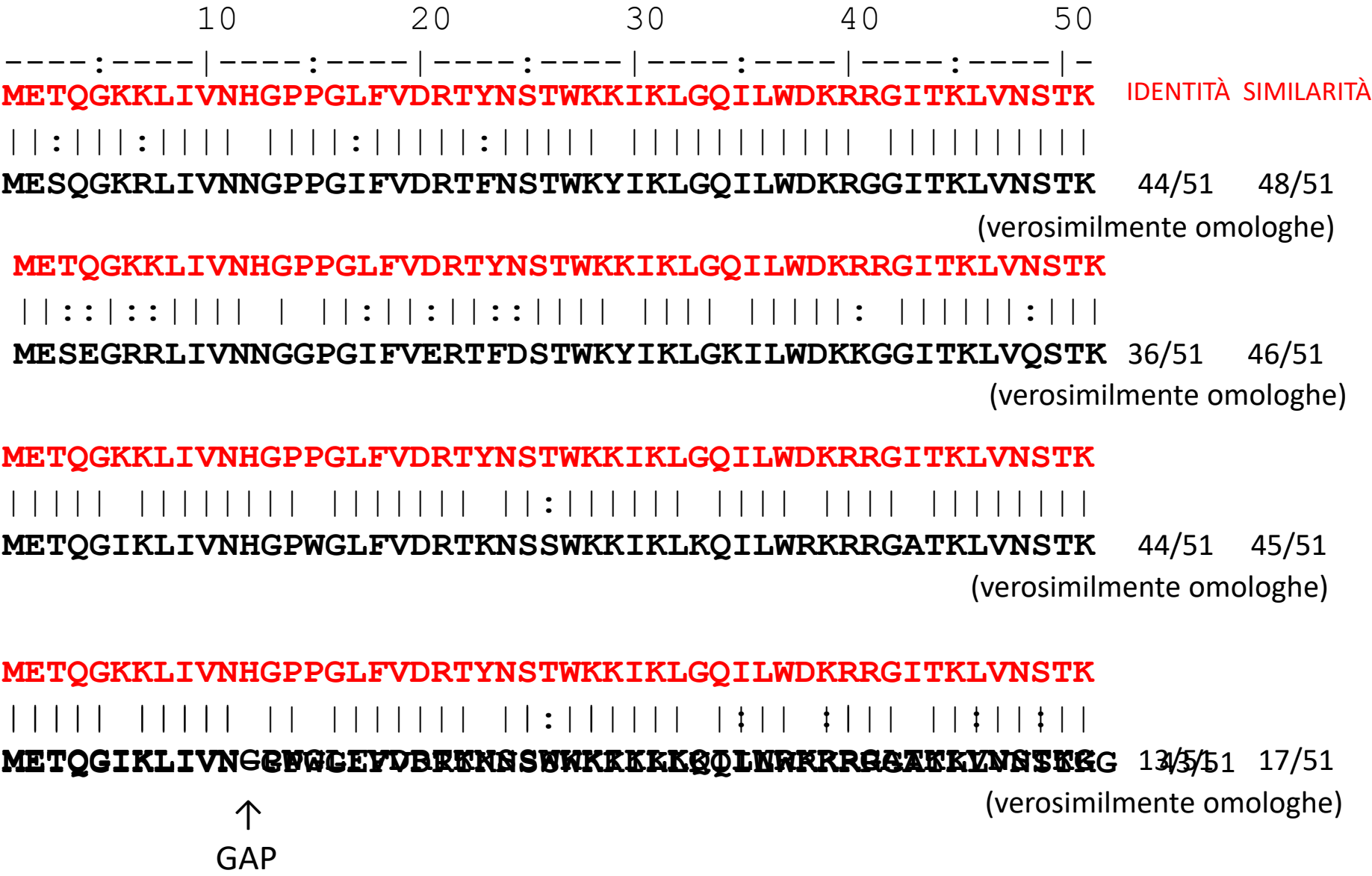
Best alignment

METQGKKLKRNHGPPGLFVDRTYNSTWKKSVLGQILW

| | : | | | : | | | | | | : | | | | : | | | | | | | | |
MESQGKRLIVNNGPPGIFVDRTFNSTWKYIKLGQILW

query
subject

Identità e similarità di sequenza vs omologia – ruolo dei GAP



Fattori genotipici e fenotipici che determinano la similarità fra AA

		Second Letter					
		U	C	A	G		
1st letter	U	UUU Phe UUC UUA Leu UUG	UCU UCC Ser UCA UCG	UAU Tyr UAC UAA Stop UAG Stop	UGU Cys UGC UGA Stop UGG Trp	U C A G	3rd letter
	C	CUU CUC Leu CUA CUG	CCU CCC Pro CCA CCG	CAU His CAC CAA Gln CAG	CGU CGC Arg CGA CGG	U C A G	
	A	AUU AUC Ile AUA AUG Met	ACU ACC Thr ACA ACG	AAU Asn AAC AAA Lys AAG	AGU Ser AGC AGA Arg AGG	U C A G	
	G	GUU GUC Val GUA GUG	GCU GCC Ala GCA GCG	GAU Asp GAC GAA Glu GAG	GGU GGC Gly GGA GGG	U C A G	

Lys (K) → Arg (R) probabilità
Gen Fen

AAA CGA CGC
AAG CGU CGG
 AGA AGG ✓ ✓

Phe (F) → Trp (W)
 → Tyr (Y)

UUU UGG ✗ ~✓
UUC UAU ✓ ~✓
 UAC ✓ ~✓

Phe (F) → Gly (G)

UUU → GGU GGC ✗ ✗
UUC GGA GGG

Ile (I) → Lys (K)

AUU → AAA ✓ ✗
AUC AUG
AUA

Major forms of mutation are:

Point mutation

Silent (synonymous) substitutions - change nucleotide sequence (genotype) but not the amino acid sequence (phenotype)

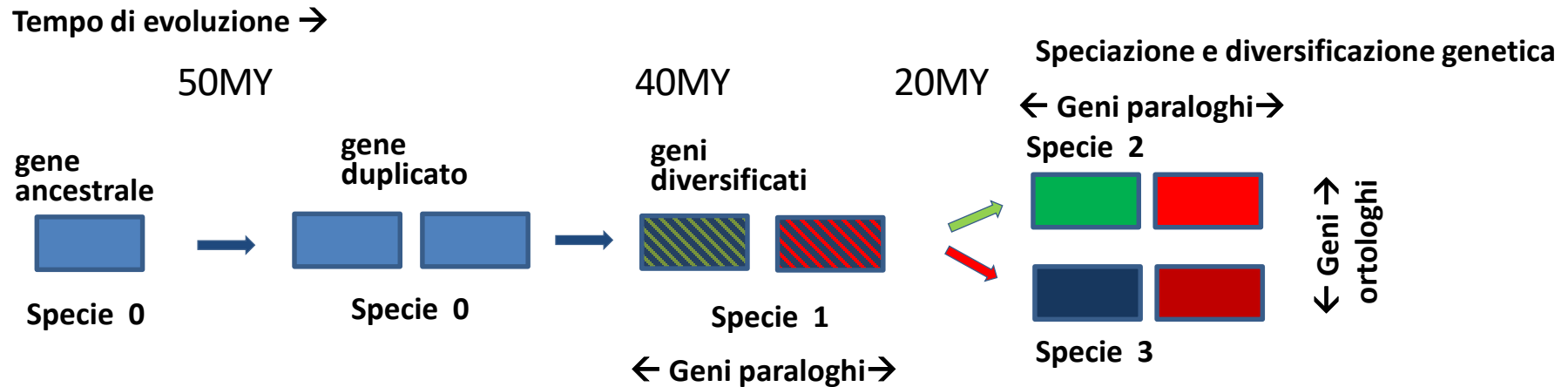
Nonsilent (nonsynonymous, missense) substitutions - change AA seq

 (**nonsense**) substitutions - AA → stop

Insertion/deletion (indel) – addition/removal of whole triplet

Fattori che determinano similarità fra sequenze

Gene duplication (fattore complicante) un organismo può avere due o più sequenze geniche derivate da un unico gene ancestrale che si è duplicato (geni paraloghi).



Paralogo: geni nello stesso organismo che derivano da un **gene duplicato e poi diversificato** (le proteine codificate possono avere funzioni diverse avendo seguito un percorso evolutivo diverso)

Ortologo: dovuto a speciazione - gene in diverse specie che deriva da **uno dei due geni duplicati, che continua a diversificarsi nelle nuove specie** (spesso la proteina codificata mantiene strutture e funzione simili nelle diverse specie)

Metodi per determinare la similarità di sequenze proteiche - si basano su "*Distanze genetiche*" calcolate utilizzando modelli di come si evolvono le sequenze amminoacidiche, utilizzando:

- matrici "*Point Accepted Mutation*" (PAM) (anni 1970)
- matrici BLOSUM (*Blocks Substitution Matrix*)

BLOSUM - BLOcs SUbstitution Matrix

PHLQHRTHTGEKPYECHQCGQAFFKKCSLLRHKRTHHTGEKPYENINMVTPLHNS
 HHLQHRTHTGEKPYECNQCCKAIFSQHGLLRHKRTHHTGEKPYMNINMVKHLHNT
 *****:*****:*****:*****:*****:*****:*****

Matrici statistiche di frequenza intrinseca dei singoli residui e probabilità che un residuo si trasformi in un altro (per 20 AA, 210 possibili sostituzioni non sinonime; 20 sinonime ad ognuna assegnato un valore).

Più alta la somma dei valori per ogni
posizione, più simili le sequenze

BLOSUM 62 (sequenze $\geq 62\%$ identità)

Posizioni sinonime: Trp (W) è raro.

Se presente nella stessa posizione pesa molto (+ 11). Ser (S) è comune (+4).

Posizioni non sinonime: Alcune (es. W con D, -4, $Y \rightarrow C$, -2) costano di più che altre (es. W con Y, +2, $E \rightarrow K + 1$)

Gap: l'apertura di gap (-) ha un costo molto elevato (-11), la continuazione un costo minore (-1).

Stop codon (*): Se è nello stesso posto ha un valore lievemente positivo (1).

La sostituzione di un AA con un codone di stop, se porta ad un stop *prematureo* ha un costo molto elevato

[illegible]

PAM - Point Accepted Mutation matrix

```
PSHLQYHERTHTGEKPYECHQCGQATFKKCSLLWRHKRTHHTGEKPYEN INMVTPLHNS
HSHLQCHKRTHHTGEKPYECNQCCKA-FSQHGLLWRHKRTHHTGEKPYMN INMVKHLHNT
***** *:*****:****:* *.: .***** ***** ****
```

PAM (Point Accepted Mutation) Matrix: Probabilità che un residuo amminoacidico si trasformi in un altro mediante i normali processi di mutazione, tenendo conto anche di sostituzioni multiple in una posizione.

PAM30 statistica basata sul confronto di sequenze con 30 mutazioni/100 residui

PAM120 sequenze con 120 mutazioni /
100 residui (evolativamente più distanti)

Posizioni sinonime: Trp (W) ha sempre il peso maggiore.

Posizioni non sinonime: $W \rightarrow D$, -8 è sempre più sfavorevole di $Y \rightarrow C$, -1 ma $W \rightarrow Y$, F -1 sono peggio che $W \rightarrow R$ +1)

Gap: l'apertura di gap ha sempre un costo molto elevato (-8), la continuazione un costo minore (-1).

Stop codon (*): Sinonimo (1), sostituzione di un AA (-8)

PAM120

STOP
↓

[illegible]

Score, Bit-score, E-value

Score: A number used to assess the biological relevance of a finding.

In the context of sequence alignments, a score is a numerical value that describes the overall quality of an alignment. Higher numbers correspond to higher similarity. The score scale depends on the scoring system used (substitution matrix, gap penalty).

Example:

	R	L	A	S	V	-	E	T	D	M	W	T	P	L	T	L	R	Q	H
	.		.		:		:		.	:			.		.	.			
	T	L	T	S	L	A	Q	T	T	L	-	-	K	A	H	L	G	T	H
S_{ij}	-1	+4	+0	+4	+1	-4	+2	+5	-1	+2	-4	-1	-1	-1	-2	+4	-2	-1	+8

= 12

$$S = \sum_{i=1}^L S_{r_{1,i} r_{2,i}}$$

Substitution matrix

gap penalty (s_{gap})

gap opening -4

gap extension -1

end gap 0

(s_{ij})

Ala	A	4																	
Arg	R	-1	5																
Asn	N	-2	0	6															
Asp	D	-2	-2	1	6														
Cys	C	0	-3	-3	-3	9													
Gln	Q	-1	1	0	0	-3	5												
Glu	E	-1	0	0	2	-4	2	5											
Gly	G	0	-2	0	-1	-3	-2	-2	6										
His	H	-2	0	1	-1	-3	0	0	-2	5									
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4								
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4							
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5						
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5					
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	5				
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7			
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-1	4	5		
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	3	5	5	
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	-4	-3	-2	11	
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	-2	-1	-1	-1	-1	-2	2	7	7	
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	-3	-3	-1	-2	-2	-1	-1	4	

SCORE - punteggio grezzo

Bit-SCORE - punteggio normalizzato per uso di diverse matrici, lunghezze ecc.

E-value - probabilità che lo score sia dovuto a identità/similarità casuali

Strumenti per l'allineamento di sequenze

GLOBAL ALIGNMENT: confronta l'intera sequenza di una proteina (query) con le sequenze presenti in una banca dati (processo più preciso ma lento)

LOCAL ALIGNMENT Metodo euristico che confronta una piccola parte della sequenza con quelle presenti in banca dati. Se trova similarità, estende la finestra sui due lati. Più veloce ma meno preciso

MULTIPLE ALIGNMENT confronta sequenze multiple, portando anche ad una relazione filogenetica tra le sequenze

Algoritmi per l'allineamento di sequenze (proteiche o di acidi nucleici)

BLAST (Basic Local Alignment Search Tool - euristico) (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>)

Confronta una sequenza proteica o nucleotidica nota (query sequence) con tutte le sequenze presenti in una banca dati e trova quelle più simili.. Usa brevi sequenze, tipicamente 3 AA o 11 nucleotide per allineamenti locali utilizzando matrici di similarità

FASTA (euristico) (<https://www.ebi.ac.uk/Tools/sss/fasta/>) Trova rapidamente tratti brevi e identici tra una sequenza di query e quelle nel database, identificando potenziali regioni di similarità, poi esegue un allineamento più lento e su queste regioni per identificare e perfezionare le corrispondenze migliori

CLUSTAL (Cluster analysis) (<https://www.ebi.ac.uk/Tools/msa/clustalo/>)

programma per l'allineamento di sequenze multiple che fornisce anche relazioni filogenetiche

Trovando una sequenza su PUBMED

- 1 su Google cerca PubMed e accedi
- 2 In PubMed seleziona la banca dati «Proteins»

NCBI Resources ▾ How To ▾ Sign in to NCBI

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed

- GEO Profile
- GSS
- GTR
- HomoloGene
- Identical Protein Groups
- MedGen
- MeSH
- NCBI Web Site
- NLM Catalog
- Nucleotide
- OMIM
- PMC
- PopSet
- Probe
- Protein
- Clusters
- Gen BioAssay
- PubChem Compound
- PubChem Substance
- PubMed

2

cathelidicin

3 metti la parola «cathelidicin» nella finestra

PubMed

PubMed comprises more than 27 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

PubMed Tools

- PubMed Mobile
- Single Citation Matcher
- Batch Citation Matcher
- Clinical Queries
- Topic-Specific Queries

More Resources

- MeSH Database
- Journals in NCBI Databases
- Clinical Trials
- E-Utilities (API)
- LinkOut

Using PubMed

- PubMed Quick Start Guide
- Full Text Articles
- PubMed FAQs
- PubMed Tutorials
- New and Noteworthy

- 4 Fra i risultati seleziona Bubalus bubalis e copia la sequenza in formato FATSA

☐ [cathelidicin \[Coturnix coturnix\]](#)

1. 148 aa protein

☐ [cathelidicin \[Phasianus colchicus\]](#)

2. 154 aa protein

☐ [cathelidicin \[Bubalus bubalis\]](#)

3. 144 aa protein

3

cathelidicin [Bubalus bubalis]

GenBank: CAH23217.1

[Identical Proteins](#) [FASTA](#) [Graphics](#)

Go to: ▾

LOCUS CAH23217 144 aa
DEFINITION cathelidicin [Bubalus bubalis].

cathelidicin [Bubalus bubalis]

GenBank: CAH23217.1

[GenPept](#) [Identical Proteins](#) [Graphics](#)

>CAH23217.1 cathelidicin [Bubalus bubalis]
MQTQRASLSLGRMSLWLLLLGLVVPASAAQQLSYREAVLRVDQLNERSSEANLYRLLVLDPLPKDDADL
GTRKPVSFYKVTVCPRITQOPAEQCDFKEKGRVKQCVGTVTLDPNDQFDLNCNALQSVGLPWILLRWL
FFRG

Seleziona e copia

Istruzioni per utilizzare BLAST

(<https://blast.ncbi.nlm.nih.gov/Blast.cgi>)

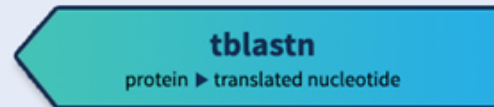
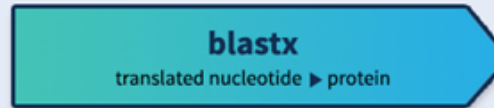
Basic Local Alignment Search Tool

Web BLAST

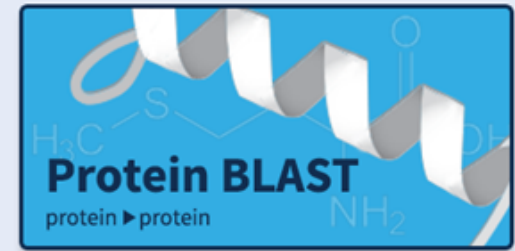


DNA → DNA

DNA tradotto → PROTEINA



PROTEINA → DNA tradotto



PROTEINA → PROTEINA



BLAST Genomes

Enter organism common name, scientific name, or tax id

Search

Human

Mouse

Rat

Microbes

Enter Query Sequence

BLASTP programs search protein databases using a protein query. [more...](#)

Enter accession number(s), gi(s), or FASTA sequence(s)

[Clear](#)

Query subrange

From

To

Or, upload file

Scegli file

Nessun file selezionato

Job Title

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database

Non-redundant protein sequences (nr)

Organism
Optional

Enter organism name or id—completions will be suggested

☐ Exclude

+

BLAST

Bovid

Bovidae (taxid:9895)

Bovidae sp. Adi Nefes (taxid:9913)

scegliere banca dati di sequenze

3



1

Accedi a
BLAST e
seleziona
Protein
Blast

2

Paste
sequence

3

Scegli la
Banca dati di
sequenze

4

Scegli
l'organismo

 Your search is limited to records that include: Bovidae (taxid:9895)

Job Title	Protein Sequence
RID	XHTT4PVN01N <i>Search expires on 12-16 23:41 pm</i> Download All ▾
Program	BLASTP  Citation ▾
Database	nr See details ▾
Query ID	lcl Query_89537
Description	None
Molecule type	amino acid
Query Length	144
Other reports	Distance tree of results Multiple alignment MSA viewer 

Filter Results

Organism

only top 20 will appear

☐ exclude

[+ Add organism](#)

Percent Identity

E value

Query Coverage

to

to

to

Filter

Reset

- Descriptions
- Graphic Summary
- Alignments
- Taxonomy

Sequences producing significant alignments

[Download ▾](#) New [Select columns ▾](#) [Show](#) 

☒ select all 100 sequences selected

[GenPept](#) [Graphics](#) [Distance tree of results](#) [Multiple alignment](#)

	Description	Common Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☒	cathelicidin-4-like isoform X1 [Bubalus bubalis]	water buffalo	291	291	100%	3e-102	100.00%	144	XP_025127734.1
☒	cathelicidin 4 [Bubalus bubalis]	water buffalo	290	290	100%	8e-102	99.31%	144	AIZ93856.1
☒	cathelicidin 4 [Bubalus bubalis]	water buffalo	290	290	100%	1e-101	98.61%	144	AIZ93859.1
☒	cathelicidin 4 [Bubalus bubalis]	water buffalo	288	288	100%	9e-101	98.61%	144	AIZ93860.1
☒	cathelicidin 4 [Bubalus bubalis]	water buffalo	285	285	100%	1e-99	97.22%	144	AIZ93889.1
☒	cathelicidin 4 [Bubalus bubalis]	water buffalo	285	285	100%	1e-99	97.22%	144	AIZ93858.1
☒	cathelicidin 4 [Bubalus bubalis]	water buffalo	285	285	100%	1e-99	97.92%	144	AIZ93901.1
☒	cathelicidin 4 [Bub	water buffalo	285	285	100%	1e-99	97.92%	144	AIZ93891.1
☒	cathelicidin 4 [Bub	water buffalo	259	259	100%	9e-90	93.75%	144	AIZ93880.1



Bit-score

E-value

bit score

S score normalized with respect to the scoring System

E value

represents the number of different alignments with scores equivalent to or better than S that Occur by chance.

E value ↓ significance ↑

[Pept Graphics](#)

licidin-3-like [Pantholops hodg:
166417.1 Length: 192 Number of I

Identity (%)

Similarity (%)
Positive score in the substitution matrix

Gaps (%)

[Pept Graphics](#)

Score	Expect	Method	Identities	Positives	Gaps
195 bits(496)	1e-63	Compositional matrix adjust.	107/130(82%)	114/130(87%)	0/130(0%)
Query 1	MQTQRASLSLGRWSLWLLLLGLVVPASAQDLSYREAVLRAVDQLNERSSEANLYRLLVL				60
Sbjct 1	M+TQRASL LGR SLWLLLLGL +PSASAO L YREAVLRAVDQLNE+SSEANLYRLL L				60
Query 61	DPPLKDDADLGRKPVSFYKQVETVCPRTTQOPAEQCDFKEKGRVKQCVGTVDLPSNDQF				120
Sbjct 61	DPPKD D G +KPVSF VKETVCPRTTQOPAEQCDFKE G +KQCVGTV LDPS+D+F				120
Query 121	DLNCNALQSV 130				
Sbjct 121	DLNCN LQSV 130				

[Download](#) [GenPept](#) [Graphics](#)

Bac7.5 protein precursor [Capra hircus]

Sequence ID: [NP_001272474.1](#) Length: 190 Number of Matches: 1

[See 1 more title\(s\)](#)

Range 1: 1 to 130 [GenPept](#) [Graphics](#)

▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
194 bits(493)	4e-63	Compositional matrix adjust.	108/130(83%)	115/130(88%)	0/130(0%)
Query 1	MQTQRASLSLGRWSLWLLLLGLVVPASAQDLSYREAVLRAVDQLNERSSEANLYRLLVL				60
Sbjct 1	M+TQRASLSLGR SLWLLLLGL++PSASAO LSYREAVLRV QNE+SSE NLYRLL L				60
Query 61	DPPLKDDADLGRKPVSFYKQVETVCPRTTQOPAEQCDFKEKGRVKQCVGTVDLPSNDQF				120
Sbjct 61	DPPKD D G +KPVSF VKETVCPRT+QOP EQCDFKE G VKQCVGTV+LDPSNDQF				120
Query 121	DLNCNALQSV 130				
Sbjct 121	DLNCN LQSV 130				

Allineamento e confronto di sequenze multiple

❶ In PUBMED seleziona la banca dati [Protein] e cerca “cathelicidin homo sapiens” (uomo)

☐ [cathelicidin antimicrobial peptide preproprotein \[Homo sapiens\]](#)

1. 170 aa protein

Accession: NP_004336.4 GI: 1543376666

[BioProject](#) [Nucleotide](#) [PubMed](#) [Taxonomy](#)

[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)



❷ Seleziona le sequenze in formato FASTA

>NP_004336.3 cathelicidin antimicrobial peptide preproprotein [Homo sapiens]

```
MGTMKTQRDGHSLGRWSLVLLLLGLVMPLAIIAQVLSYKEAVLRAIDGINQRSSDANLYRLLDLDPRPTMDGDPDTPKPVVS  
FTVKETVCPRTTQQSPEDCDFKKGDLVKRCMGTVTNLNQARGSFDISCDKDNKRFBALLGDFFRKSKEKIGKEFKRIVQRIKD  
FLRNLVPRTES
```

(copia tutto in una file di testo)

❸ Poi fai lo stesso per «pig, rabbit, dog, bovine» e aggiungile al file di testo

>NP_004336.3 cathelicidin antimicrobial peptide preproprotein [Homo sapiens]

```
MGTMKTQRDGHSLGRWSLVLLLLGLVMPLAIIAQVLSYKEAVLRAIDGINQRSSDANLYRLLDLDPRPTMDGDPDTPKPVVSFTVKETVCPRTTQQSPEDCDFKKGDLVKRCM  
GTVTLNQARGSFDISCDKDNKRFBALLGDFFRKSKEKIGKEFKRIVQRIKDFLRNLVPRTES
```

>CAA73261.1 cathelicidin [Bos taurus]

```
METQRASFSLGRSSLWLLLLGLVVPSASAQDLSYREAVLRAVDQFNERSSEANLYRLLELDPPPEQDVEHPGARKPVVSFTVKETVCPRTTPQPPEQCDFKENGVLVKQCVGTV  
TRYWIRGDFDITCNNIQSAGLFRRLRDSIRRGQQKILEKARRIGERIKDIFRG
```

>NP_999615.1 antibacterial protein PR-39 precursor [Sus scrofa]

```
METQRASLCLGRWSLVLLLLGLVVPSASTQALSREAVLRAVDRLNEQSSEANLYRLLELDQPPKADEDP  
GTPKPVVSFTVKETVCPRTTQRPPELCDFKENGVRVKQCVGTVTLNPSNDPLDISCNEIQSVRRRPRPPYLP  
RPRPPFFFPRLPPRIPPFPFRFPFRFPGR
```

>NP_001075774.1 antimicrobial protein CAP18 precursor [Oryctolagus cuniculus]

```
METHKHGPSLAWWSLLLLLGLLMPPAIAQDLTYREAVLRAVDAFNQSSSEANLYRLLSMDPQQLEDAKP  
YTPQPVVSFTVKETECPRTTWKLPQCDFKEDGLVKRCVGTVTRYQAWDSFDIRCNRAQESPEPTGLRKRL  
RKFRNKIKEKLKIGQKIQGFVVKLAPRTDY
```

4

Apri CLUSTAL OMEGA e incolla tutte le sequenze nella finestra

Clustal Omega

Multiple Sequence Alignment (MSA)

[Home](#) [Help & Privacy](#) [Recent Jobs](#) **Input Form**

[Feedback](#)

Welcome to the Job Dispatcher website! If you need assistance or have feedback, please [contact](#) us.

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools. This tool can align up to 4000 sequences or a maximum file size of 4 MB.

Input sequence ⓘ

Sequence type

☒ Protein ☐ DNA ☐ RNA

Paste your sequence here - or use the example sequence

```
|>CAA73261.1 cathelicidin [Bos taurus]
METQRASFSLGRSSLWLLLLGLVVPASQAQDLSYREAVLRAVDQFNERSSSEANLYRLLLELDPPEQDVEHPGARKPVSFTVKETVCPRTTPQPPEQ
CDFKENGLVKQCVGTVTRYWIRGDFDITCNNIQSAGLFRRLRDSIRRGQKILEKARRIGERIKDIFRG
>NP_004336.3 cathelicidin antimicrobial peptide preproprotein [Homo sapiens]
MGTMKTQRDGHSLGRWSLVLLLLGLVMPLAIIAQVLSYKEAVLRAIDGINQRSSDANLYRLLDLDPRTMDGDPDTPKPVSTVKETVCPRTTQQS
PEDCDFKKDGLVKRCMGTVTLNQARGSFDISCDKDNKRFALLGDFFRKSKEKIGKEFKRIVQRIKDFLRNLVPRTES
```

Scegli file Nessun file selezionato

Use the example

Clear sequence

More example inputs

Parameters

OUTPUT FORMAT ⓘ

ClustalW with character counts

More options ▼

Submit

Title

Clustal Omega's job

Submit



Results for job clustalo-I20171122-114750-0238-29794709-pg

Alignments Result Summary Phylogenetic Tree Submission Details

Download Alignment File Show Colors to Simple_Phylogeny

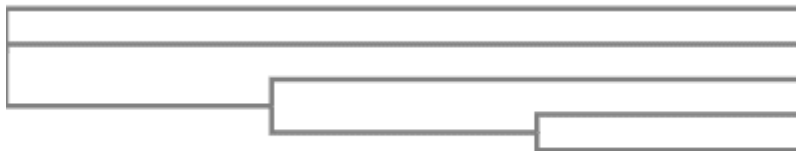
CLUSTAL O(1.2.4) multiple sequence alignment

```

NP_004336.3      MGMTKTQRDGHSLGRWSLVLLLLLGLVMP LAIIAQVLSYKEAVLRAIDGINQRSSDANLYR
NP_001075774.1  ---METHKHGPSLAWMSLLLLLLGLLMPPA-IAQDLTYREAVLRAVDAFNQSSSEANLYR
AAR26245.1      ---METQKDSPSLGRWSLLLLLLGLVITPA-ASRALSYREAVLRAVNGFNQRSSEENLYR
CAA73261.1      ---METQRASFSLGRSSLWLLLLLGLVVPSA-SAQDLSYREAVLRAVDQFNERSSSEANLYR
P49932.1        ---METQRASLCLGRWSLWLLLLLALVVPSA-SAQALS YREAVLRAVDRLNEQSSEANLYR
                *:::: . .*. ** ****.*:: * ::*:*****: :*:**: ****

NP_004336.3      LLDLDPRTMD-GDPDTPKPVSFVTKETVCPRTTQQSPEDCDFKKDGLVKRCMGTVTLNQ
NP_001075774.1  LLSMDPQQLED-AKPYTPQPVSFVTKETECPRTTWKLPEQCDFKEDGLVKRCVGTVTRYQ
AAR26245.1      LLQLNSQPKGD-EDPNIPKPVSFVTKETVCPKTTQQPLEQCGFKDGLVKQCEGTVILDE
CAA73261.1      LLELDPPPEQDVEHPGARKPVSFVTKETVCPRTTQPPEQCDFKENGVLKQCVGTVTRYW
P49932.1        LLELDQPPKAD-EDPGTPKPVSFVTKETVCPRTWRPPELCDFKENG RVKQCVGTVTL DQ
                **::: * . * :***** **: * : * *,*:* **:* **

NP_004336.3      ARGSFDISCDKDN--KRFALLGDFFRKSKEKIGKEFKRIVQRIKDFLRNLVPRTES-
NP_001075774.1  ANDSFDIRC NRAQESPEPTGLRKR LRKFRNKIKEKLKIGQKIQGFVPKLAPRTDY-
AAR26245.1      DTGYFDLNCDSILQVKKIDRLKELITTGGQKIGEKIRRIQRIKDFFKNLQPREEKS
CAA73261.1      IRGDFDITC NNIQSAGLFRRLRDSIRRGQQKILEKARRIGERIKDIFRG-----
P49932.1        IKDPLDITCNEIQSVGLLSRLRDFLSDRGRRLGEKIERIGQIKDLSEFFQS-----
                . :*: * : * . : .:: :. :* :*:::
    
```



NP_004336.3 0.18524
 NP_001075774.1 0.23488
 AAR26245.1 0.21227
 CAA73261.1 0.12484
 P49932.1 0.16784

Human	Euarchontoglires
Rabbit	Euarchontoglires
Dog	Laurasiatheria
Cow	Laurasiatheria
Pig	Laurasiatheria

Tradurre una sequenza di DNA in una sequenza proteica

1 apri il **Translate tool** <https://web.expasy.org/translate/>

Translate is a tool which allows the translation of a nucleotide (DNA/RNA) sequence to a protein sequence.

DNA or RNA sequence

Please enter a DNA or RNA sequence - numbers and blanks are ignored

```
ATGGGGACCATGAAGACCCAAAGGGATGGCCACTCCCTGGGGCGGTGGTCACTGGTGCTCCTGCTGCTGGG
CCTGGTGATGCCTCTGGCCATCATTGCCCAGGTCCTCAGCTACAAGGAAGCTGTGCTTCGTGCTATAGATG
GCATCAACCAGCGGTCTCGGATGCTAACCTCTACCGCCTCCTGGACCTGGACCCCAGGCCACGATGGAT
GGGGACCCAGACACGCCAAAGCCTGTGAGCTTCACAGTGAAGGAGACAGTGTGCCCCAGGACGACACAGCA
CGGAAATCTAAAGAGAAGATTGGCAAAGAGTTTAAAAGAATTGTCCAGAGAATCAAGGATTTTTTGC GGAA
TCTTGTACCCAGGACAGAGTCCTAG
```

2 Paste DNA sequence in window

Output format

- ☐ Verbose: Met, Stop, spaces between residues
- ☒ Compact: M, -, no spaces
- ☐ Includes nucleotide sequence
- ☐ Includes nucleotide sequence, no spaces

3 seleziona il formato dell'output

DNA strands

- ☒ forward
- ☒ reverse

Genetic codes - [See NCBI's genetic codes](#)

Standard

reset

TRANSLATE!

4

Translate is a tool which allows the translation of a nucleotide (DNA/RNA) sequence to a protein sequence.

DNA or RNA sequence

```
atggggaccatgaagacccaaagggatggccactccctggggcgggtggtcactggtgctcctgctgctgggacctggtgatgcctctggccatcattgccaggctcctcagctacaa  
ggaagctgtgcttctgctctatagatggcatcaaccagcgggtcctcggatgctaacctctaccgcctcctggacctggacccaggccacgatggatggggaccagacacgcca  
agcctgtgagcttcacagtgaaggagacagtgtgccccaggacgacacagcacggaatctaaagagaagattggcaagagatttaaaagaattgtccagagaatcaaggatttt  
tgcggaatctgtaccaggacagagtcctag
```

Output format

- ☐ Verbose: Met, Stop, spaces between residues
- ☒ Compact: M, -, no spaces
- ☐ Includes nucleotide sequence
- ☐ Includes nucleotide sequence, no spaces

DNA strands

- ☒ forward
- ☒ reverse

Genetic codes - [See NCBI's genetic codes](#)

Standard

reset

TRANSLATE!

Results of translation

- Open reading frames are highlighted in red
- Select your initiator on one of the following frames to retrieve your amino acid sequence

Download all the translated frames

5'3' Frame 1

MGTMKTQRDGHSLGRWSLVLLLLGLVMP LAIIAQVLSYKEAVLRAIDGINQRSSDANLYRLDLDPRPTMDGDPDTPKPVSFVTKETVCPRTTQHGNLKRRLAKSLKELSRRESRIFCGILYPGQSP

5'3' Frame 2

WGP-RPKGMATPWGGGHWCCWAW-CLWPSLPRSSATRKLCLFVL-MASTSGPRMLTSTASWTWTPGPRWMGTQTRQSL-ASQ-RRQCAPGRHSTEI-REDWQRV-KNCPENQGFFAESCTQDRV L

5'3' Frame 3

GDHEDPKGWPLPGA VVTGAPAGPGDASGHHCPGPQLQGSASCYRWHQPAVLGC-PLPPPGPGPQAHDGWGPRHAKACELHSEGDSVPQDDTARKSKEKIGKEFKRIVQRIKDFLRNLVPRTES-

3'5' Frame 1

LGLCPGYKIPQKILDSL DNSFKLFANLLFRFPCCVVLGHTVSFTVKLTGFGVSGSPSIVGLGSRSSRR-RLASEDRWLMP SIARSTASL-LRTWAMMARGITRPSSRSTSDHRPREWPSLWVFMVP

3'5' Frame 2

-DSVLGTRFRKKSLILWTILLNSLP IFSLDFRAVSSWGTLSPSL-SSQALACLGP HPSWAWGPGPGGGRG-HPRTAG-CHL-HEAQLPCS-GPGQ-WFEASPGPAAGAPVTTAPGSGHPFGSSWSP

3'5' Frame 3

RTL SWVQDSAKNP-FSGQFF-TLCQSSL-ISVLCRPGAHC LLHCEAHLRWVWVPIHRGPGVQVQEA VEVSIRGPLVD AIYSTKHSFLVAEDLGNDGQRHHQAQQQEHQ-PP PQGVAIPLGLHGPH

- Vedrai TRE cornici di lettura per il filamento 1 (5' → 3') e tre cornici di lettura per il filamento complementare (3' ← 5')
- La cornice di lettura corretta è verosimilmente quella più lunga che inizia con «M» (start codon) e finisce con – (stop codon), che corrisponde all'ORF

Translate is a tool which allows the translation of a nucleotide (DNA/RNA) sequence to a protein sequence.

DNA or RNA sequence

```
atggggaccatgaagacccaagggatggccactccctggggcgggtggtcactggtgctcctgctgctgggcctggtgatgcctctggccatcattgccaggtcctcagctacaa  
ggaagctgtgcttctgctatagatggcatcaaccagcggtcctcggatgctaactctaccgctcctggacctggacccccaggcccacgatggatggggaccagacacgcaa  
agcctgtgagcttcacagtgaaggagacagtgtccccaggacgacacagcaggaaatctaaagagaagattggcaagagtttaaaagaattgtccagagaatcaaggattttt  
tcggaatcttgtaccaggacagagtcctag
```

5

Output format

- ☐ Verbose: Met, Stop, spaces between residues
- ☐ Compact: M, -, no spaces
- ☒ Includes nucleotide sequence
- ☐ Includes nucleotide sequence, no spaces

DNA strands

- ☒ forward
- ☒ reverse

Genetic codes - [See NCBI's genetic codes](#)

Standard

reset

TRANSLATE!

Results of translation

- Open reading frames are highlighted in red
- Select your initiator on one of the following frames to retrieve your amino acid sequence

Download all the translated frames

5'3' Frame 1

```
atg ggg acc atg aag acc caa agg gat ggc cac tcc ctg ggg cgg tgg tca ctg gtg ctc  
M G T M K T Q R D G H S L G R W S L V L  
ctg ctg ctg ggc ctg gtg atg cct ctg gcc atc att gcc cag gtc ctc agc tac aag gaa  
L L L G L V M P L A I I A Q V L S Y K E  
gct gtg ctt cgt gct ata gat ggc atc aac cag cgg tcc tgg gat gct aac ctc tac cgc  
A V L R A I D G I N Q R S S D A N L Y R  
ctc ctg gac ctg gac ccc agg ccc acg atg gat ggg gac cca gac acg cca aag cct gtg  
L L D L D P R P T M D G D P D T P K P V  
agc ttc aca gtg aag gag aca gtg tgc ccc agg acg aca cag cac gga aat cta aag aga  
S F T V K E T V C P R T T Q H G N L K R  
aga ttg gca aag agt tta aaa gaa ttg tcc aga gaa tca agg att ttt tgc gga atc ttg  
R L A K S L K E L S R E S R I F C G I L  
tac cca gga cag agt cct  
Y P G Q S P
```

5 Cambiando il formato output si può anche vedere la sequenza proteica allineata a quella nucleotidica