

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 per confronto di 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

MHSSIVLATVLFVAIASASKTRELCKMSLEHAKVGTSEAKQDGLD
YKHMFEHYAPAMKKYFKHRENYTPADVQKDPFFIKQGQNILACHV
LCATYDDRETFDAYVYVGLMARHERDHVKVPNDVWNHFWHEHIEF
LGSK

TTLDEPTKHAWQEIGKFSHEISHHGRHSVRDHCMNSLEYIAIGDK
EHQKQNGIDLYKHMFEHYPHMRKAFKGFENFTKEDVQKDAFFVN
KDRFCWPFVCCDSSYDDEPTFDYFVDALMDRHKDDIHLPEQ
WHEFWK

LFAEYLNEKSHQHLTEAEKHAWSTIGEDFAHEADKHAKAEKDHHE
GEHKEEHHKLGGENAMKAAVPLFYKKVLADERVKHFFKNTDMDH
QTKQQTDFLTMLLGGPNHYKGNMT

EAHKGMNLQNLHFDAIENLAATLKGVTDAVINEAAKVIEHTRK
DMLGKTINIIKATVPVLKEHGVTTTTFYKNLFAKHPEVRPLFDMGR
QESLEQPKALAMTVLAAQNIEN

LPAILPAVKKIYVKKHCAQVAAAHYPIVQELLGAKEVLGDAATD
DILDWKGKAYGIADVFIQVEADLYAQAVEQELLGAKEVLGDAAT
DDILDWKGKAYGIADVFIQVEADLAQAVE

EPTKHAWQEIGKFSHEISHHGRHSVRDHCMNSLEYIAIGDKEHQ
KQNGIDLYKHMFEHYPHMRKAFKGFENFTKEDVQKDAFFVNKDT
PFCWVPCVCCSSYDDEPTFDYFVDALMDRHKDDIHLPEQW

Best alignment

METQGKKLKRNHGPPGLFVDRTYNSTWKKSVLGQILW

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

MESQGKRLIVNNGPPGIFVDRTFNSTWKYIKLGQILW

query
subject

Identità e similarità di sequenza vs omologia – il ruolo dei GAP



Come si determina la similarità – fattori genotipici e fenotipici

		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
	C	CUU CUC Leu CUA CUG	CCU CCC Pro CCA CCG	CAU His CAC CAA Gln CAG	CGU CGC Arg CGA CGG
	A	AUU AUC Ile AUA AUG Met	ACU ACC Thr ACA ACG	AAU Asn AAC AAA Lys AAG	AGU Ser AGC AGA Arg AGG
	G	GUU GUC Val GUA GUG	GCU GCC Ala GCA GCG	GAU Asp GAC GAA Glu GAG	GGU GGC Gly GGA GGG

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 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

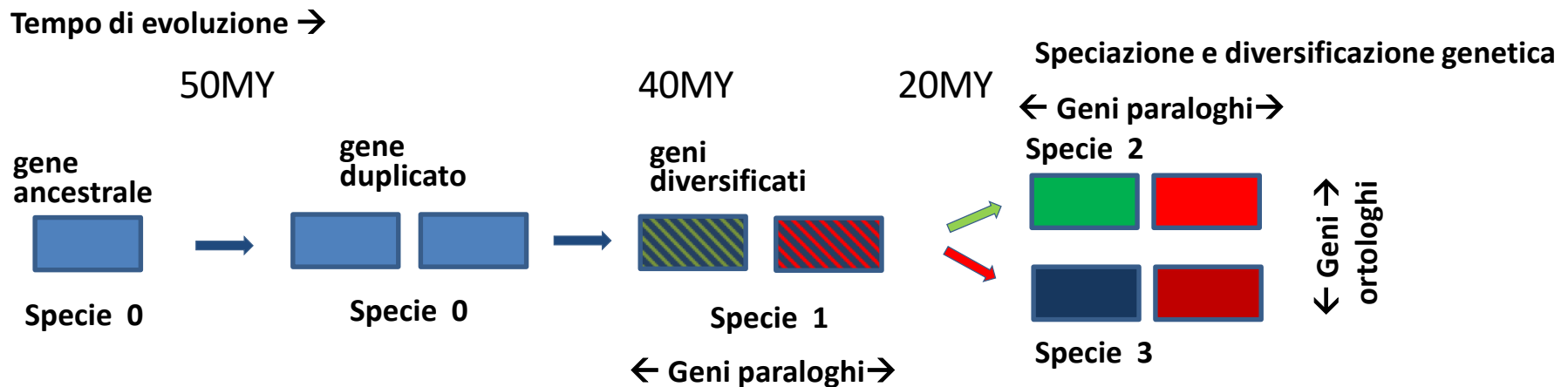
Metodi per misurare la similarità di sequenza

Sequenze proteiche

Distanze “genetiche” basandosi su modelli di come si evolvono le sequenze amminoacidiche

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

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 duplicate, che continua a diversificarsi in nuove specie** (in genere la proteina codificata mantiene strutture e funzione simili nelle diverse specie)

“Pairwise alignment” and “Substitution matrix”

PSHLQYHERHTHTGKPYECHQCGQATFKKCSLLQRHKRTHHTGKPYE-CGQCGKAFAQ-
 HSHLQCHKRTHHTGKPYECNQCGKA-FSQHGLLQRHKRTHHTGKPYMNVINMVKPLHNS
 **** * : ***** : *** : * . : ***** : * . : :

BLOSUM (BLOCKS SUBSTITUTION MATRIX). Matrici statistiche di frequenza intrinseca dei residui
 e probabilità che un residuo si trasformi in un altro (fra le 210 possibili paia di sostituzioni per
 i 20 amino acidi proteinogenici).

BLOSUM62 → matrice
 per protein fino al 62%
 di identità

Più alto il numero di
 matrice più simili sono
 le protein confrontate

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

“Pairwise alignment” and “Substitution matrix”

PSHLQYHERHTHTGEKPYECHQCGQATFKKCSLLQRHKHTHTGEKPYE-CGQCG#AFAQ-
 HSHKQCHKRHTHTGEKPYECNQCCKA-FSQHGLLQRHKRHTHTGEKPYMNVINMVKPLHNS
 ** * * :*****:***:* * : . *****:***** : . : :

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.

(# = STOP)



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

PAM70 basata sul confronto di
 sequenze con
 70 mutazioni / 100 residui

Più alto il numero PAM
 più distanti evolutivamente
 sono le sequenze confrontate

	G	A	V	L	I	P	S	T	D	E	N	Q	K	R	H	F	Y	W	M	C	B	Z	X	*	
G	5																								G
A	1	2																							A
V	-1	0	4																						V
L	-4	-2	2	6																					L
I	-3	-1	4	2	5																				I
P	0	1	-1	-3	-2	6																			P
S	1	1	-1	-3	-1	1	2																		S
T	0	1	0	-2	0	0	1	3																	T
D	1	0	-2	-4	-2	-1	0	0	4																D
E	0	0	-2	-3	-2	-1	0	0	3	4															E
N	0	0	-2	-3	-2	0	1	0	2	1	2														N
Q	-1	0	-2	-2	-2	0	-1	-1	2	2	1	4													Q
K	-2	-1	-2	-3	-2	-1	0	0	0	0	1	1	5												K
R	-3	-2	-2	-3	-2	0	0	-1	-1	-1	0	1	3	6											R
H	-2	-1	-2	-2	-2	0	-1	-1	1	1	2	3	0	2	6										H
F	-5	-3	-1	2	1	-5	-3	-3	-6	-5	-3	-5	-5	-4	-2	9									F
Y	-5	-3	-2	-1	-1	-5	-3	-3	-4	-4	-2	-4	-4	-4	0	7	10								Y
W	-7	-6	-6	-2	-5	-6	-2	-5	-7	-7	-4	-5	-3	-2	-3	0	0	17							W
M	-3	-1	2	4	2	-2	-2	-1	-3	-2	-2	-1	0	0	-2	0	-2	-4	6						M
C	-3	-2	-2	-6	-2	-3	0	-2	-5	-5	-4	-5	-5	-4	-3	-4	0	-8	-5	12					C
B	0	0	-2	-3	-2	-1	0	0	3	3	2	1	1	-1	1	-4	-3	-5	-2	-4	3				B
Z	0	0	-2	-3	-2	0	0	-1	3	3	1	3	0	0	2	-5	-4	-6	-2	-5	2	3			Z
X	-1	0	-1	-1	-1	-1	0	0	-1	-1	0	-1	-1	-1	-1	-2	-2	-4	-1	-3	-1	-1	-1		X
*	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	1	*
	G	A	V	L	I	P	S	T	D	E	N	Q	K	R	H	F	Y	W	M	C	B	Z	X	*	

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 ($s_{i,j}$)

gap opening -4

gap extension -1

end gap 0

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4	-1	-2	-2	0	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	1	0	-1	-1	0
Arg	-1	5	-2	-2	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-1	-1	-1	-1	-1
Asn	-2	0	5	-2	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-1	-1	-1	-1	-1
Asp	-2	-2	1	5	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-1	-1	-1	-1	-1
Cys	0	-3	-3	-3	5	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-1	-1	-1	-1	-1
Gln	-1	1	0	0	-3	5	-3	-3	-3	-3	-3	-3	-3	-3	-3	-1	-1	-1	-1	-1
Glu	-1	0	0	0	-4	2	5	-3	-3	-3	-3	-3	-3	-3	-3	-1	-1	-1	-1	-1
Gly	0	-2	0	-1	-3	-2	-2	5	-3	-3	-3	-3	-3	-3	-3	-1	-1	-1	-1	-1
His	-2	0	1	-1	-3	0	0	-2	5	-3	-3	-3	-3	-3	-3	-1	-1	-1	-1	-1
Ile	-1	-3	-3	-3	-3	-1	-3	-3	-4	5	-3	-3	-3	-3	-3	-1	-1	-1	-1	-1
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	5	-3	-3	-3	-3	-1	-1	-1	-1	-1
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-3	-3	-3	-1	-1	-1	-1	-1
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	-3	-3	-1	-1	-1	-1	-1
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-3	0	0	-3	0	5	-3	-1	-1	-1	-1	-1
Pro	-1	-2	-2	-3	-3	-1	-1	-2	-2	-3	-3	-3	-3	-3	5	-1	-1	-1	-1	-1
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	5	-1	-1	-1	-1
Thr	0	-1	0	-1	-1	-1	-2	-2	-1	-1	-1	-1	-1	-1	-1	-1	5	-1	-1	-1
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-3	-3	-3	-3	-3	-3	-3	-3	-3	5	-1	-1
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	5	-1
Val	0	-3	-3	-3	-1	-2	-3	-3	-3	3	1	1	1	-1	-2	-2	0	-3	-1	4

SCORE - punteggio grezzo

Bit-SCORE - punteggio normalizzato

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

Sequence alignment tools

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

LOCAL ALIGNMENT confronta una piccola parte della sequenza con le sequenze presenti in una banca dati. Poi estende dai lati della sequenza locale (veloce ma meno preciso, adatto a metodi euristici)

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

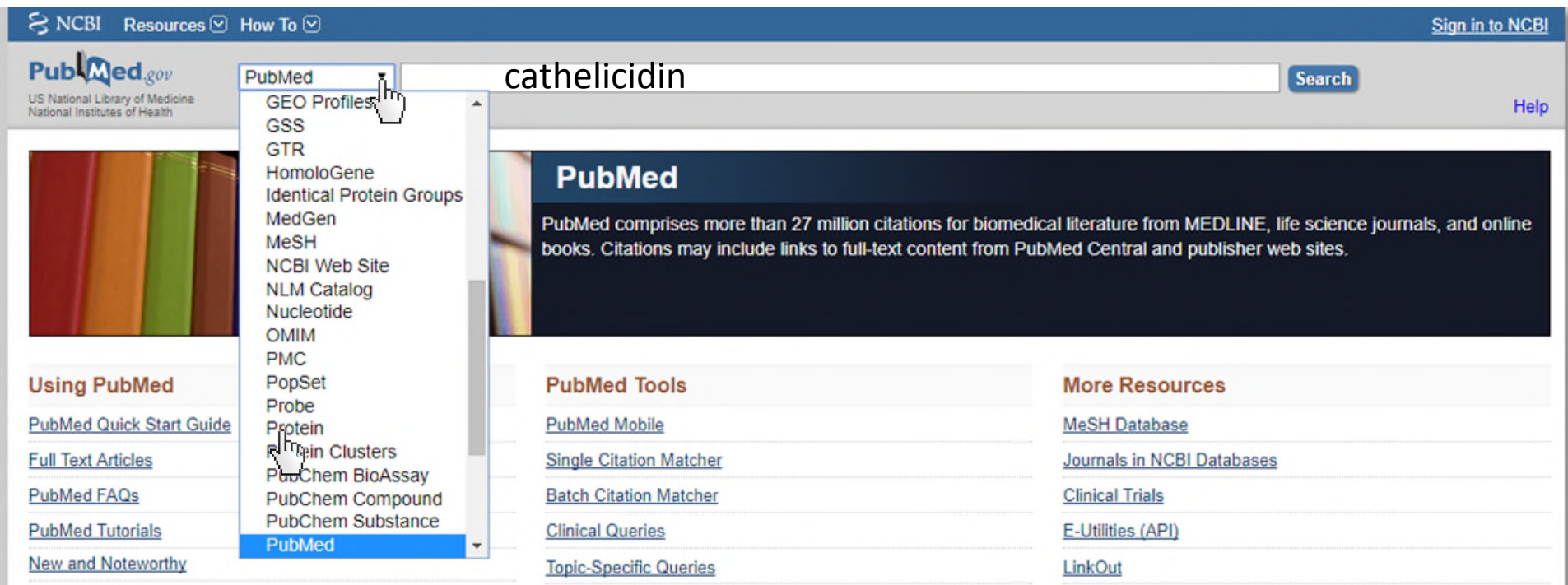
BLAST (metodo euristico) (**Basic Local Alignment Search Tool**) confronta una sequenza proteica o nucleotidica nota (query sequence) con tutte le sequenze presenti in una banca dati e trova quelle più simili. Usa matrici di similarità fra AA per decidere il grado di similarità. Usa brevi sequenze, tipicamente di 3 AA o 11 nucleotide per allineamenti locali (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>)

FASTA (altro metodo euristico) usa un approccio approssimativo per determinare il grado di similarità, sempre in base a matrici di similarità fra amminoacidi (<https://www.ebi.ac.uk/Tools/sss/fasta/>)

CLUSTAL (Cluster analysis) programma per l'allineamento di sequenze multiple (<https://www.ebi.ac.uk/Tools/msa/clustalo/>)

Scegliamo una sequenza proteica

1) Google [Pubmed]



NCBI Resources ▾ How To ▾ Sign in to NCBI

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

PubMed

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

cathelcidin

Search

Help

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](#)

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

1. 148 aa protein

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

2. 154 aa protein

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

3. 144 aa protein

cathelcidin [Bubalus bubalis]

GenBank: CAH23217.1

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

Go to: ▾

LOCUS CAH23217 144 aa
DEFINITION cathelcidin [Bubalus bubalis].

cathelcidin [Bubalus bubalis]

GenBank: CAH23217.1

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

>CAH23217.1 cathelcidin [Bubalus bubalis]

MQTQRASLSLGRMSLWLLLLGLVVPSSASAQQLSYREAVLRVDQLNERSSEANLYRLLVLDPLKDDADL
GTRKPVSFYKTVCPRTTQDAEQCDFKEKGRVKQCVGTVTLDPNDQFDLNCNALQSVGLPWILLRWL
FFRG



Seleziona e copia

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

Basic Local Alignment Search Tool

Web BLAST

DNA tradotto → PROTEINA



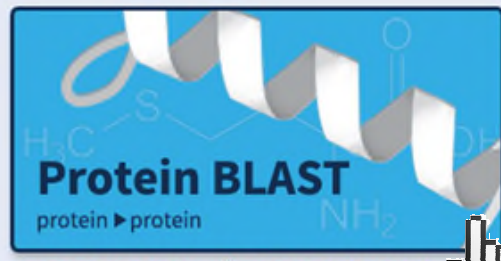
Nucleotide BLAST
nucleotide ► nucleotide

DNA → DNA

blastx
translated nucleotide ► protein

tblastn
protein ► translated nucleotide

PROTEINA → DNA tradotto



Protein BLAST
protein ► protein

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)

MQTQRASLSLGRWSLWLLLLGLVVPASQAQLSYREAVLRAVDQLNERSSEANLYRLL
VLDPPLKDDADLGRKPVSVFTVKETVCPRTTQQPAEQCDFKEKGRVKQCVGTVTLDP
NDQFDLNCNALQSVGLPWILLRWLFFRG

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

- Bovid
- Bovidae (taxid:9895)
- Bovidae sp. Adi No (taxid:9913)

Include +

BLAST

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

Job Title	Protein Sequence
RID	XHTT4PVN01N Search expires on 12-16 23:41 pm 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 100

☒ 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 [Bubalus bubalis]	water buffalo	285	285	100%	1e-99	97.92%	144	AIZ93891.1
☒	cathelicidin 4 [Bubalus bubalis]	water buffalo	259	259	100%	9e-90	93.75%	144	AIZ93880.1

Bit-score

E-value

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

PREDICTED: cathelicidin-3-like [Pantholops hodgsoni]

Identity (%)

Similarity (%)
Positive score in the substitution matrix

Gaps (%)

Sequence ID: [XP_005966417.1](#) Length: 192 Number of Matches: 1

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

[Next Match](#)
[Previous Match](#)

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 ↑

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	MOTORASLSLGRWSLWLLLLGLVVPASASAQDLSYREAVLRAVDQLNERSSEANLYRLLVL				60
	M+TORASL LGR SLWLLLLGL +PSASAO L YREAVLRAVDQLNE+SSEANLYRLL L				
Sbjct 1	METQRASLLGRCSLWLLLLGLALPSASAQALGYREAVLRAVDQLNEKSSEANLYRLLLEL				60
Query 61	DPPLKDDADLGTRKPVSFYKQVETVCPRTTQPAEQCDFKEKGRVKQCVGTVTLDPSNDQF				120
	DPP KD D G +KPVSF VKETVCPRTTQPAEQCDFKE G +KQCVGTV LDPS+D+F				
Sbjct 61	DPPPKDVEDRGAQKPVSFYKQVETVCPRTTQPAEQCDFKENGMLKQCVGTVNLDPSDDEF				120
Query 121	DLNCNALQSV 130				
	DLNCN LQSV				
Sbjct 121	DLNCNELQSV 130				

Download

GenPept

Graphics

Bac7.5 protein precursor [Capra hircus]

Identity (%)

Similarity (%)

Gaps (%)

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

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	MOTORASLSLGRWSLWLLLLGLVVPASASAQDLSYREAVLRAVDQLNERSSEANLYRLLVL				60
	M+TORASLSLGR SLWLLLLGL++PSASAO LSYREAVLRAV QLNE+SSE NLYRLL L				
Sbjct 1	METQRASLSLGRCSLWLLLLGLLLPSASAQALS YREAVLRAVGQLNEKSSEVNLYRLLLEL				60
Query 61	DPPLKDDADLGTRKPVSFYKQVETVCPRTTQPAEQCDFKEKGRVKQCVGTVTLDPSNDQF				120
	DPP KD D G +KPVSF VKETVCPRT+QOP EQCDFKE G VKQCVGTV+LDPSNDQF				
Sbjct 61	DPPPKDVEDQGAQKPVSFYKQVETVCPRTSQQPPEQCDFKENGVLKQCVGTVSLDPSNDQF				120
Query 121	DLNCNALQSV 130				
	DLNCN LQSV				
Sbjct 121	DLNCNELQSV 130				

CLUSTALL (<https://www.ebi.ac.uk/Tools/msa/clustalo/>)

Allineamento e confronto di sequenze multiple

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

Scegli e apri una scheda (entry) e poi all’interno selezionare la sequenza in formato FASTA e copia tutto

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

1. 170 aa protein

Accession: NP_004336.4 GI: 1543376666

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

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

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

MGTMTKTQRDGHSLGRWSLYLLLLGLVMPLAIIAQVLSYKEAVLRAIDGINQRSSDANLYRLLDLDPRPTMDGDPDTPKPVS
FTVKETVCPRTTQQSPEDCDFKKDGLVKRCMGTVTNLNARGSFDISCDKDNKRFBALLGDFFRKSKEKIGKEFKRIVQRIKD
FLRNLVPRTES (copia tutto in una file di testo)

Poi fai lo stesso per pig, rabbit, dog, bovine

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

MGTMTKTQRDGHSLGRWSLYLLLLGLVMPLAIIAQVLSYKEAVLRAIDGINQRSSDANLYRLLDLDPRPTMDGDPDTPKPVSFTVKETVCPRTTQQSPEDCDFKKDGLVKRCM
GTVTLNARGSFDISCDKDNKRFBALLGDFFRKSKEKIGKEFKRIVQRIKDFLRNLVPRTES

>CAA73261.1 cathelcidin [Bos taurus]

METQRASFSLGRSSLWLLLLGLVVPASAQDLSYREAVLRAVDQFNERSSEANLYRLLDLDPPPEQDVEHPGARKPVSFVKETVCPRTTPQPPEQCDFKENGVLVKQCVGT
TRYWIRGDFDITCNIQSAGLFRRLRDSIRRGQQKILEKARRIGERIKDIFRG

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

METQRASLCLGRWSLYLLLLGLVVPASTQALSREAVLRAVDRLNEQSSEANLYRLLDLDQPPKADED
GTPKPVSFVKETVCPRTTQRPPELCDFKENGVRVKQCVGTVTNLNPSNDPLDISCNEIQSVRRRPRPPYLP
RPRPPFFPPRLPPRIPPGFPPRFPFPGKR

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

METHKHGPSLAWWSLLLLLLGLLMPPAIAQDLTYREAVLRAVDAFNQSSSEANLYRLLSMDPQQLEDAP
YTPQPVSFVKETECPRTTWKLPEQCDFKEDGLVKRCVGTVTRYQAWDSFDIRCNAQESPEPTGLRKRL
RKFRNKIKEKLLKIGQKIQGFVPKLAPRTDY

Apri CLUSTAL OMEGA e incolla tutte le sequenze nella finestra

Multiple Sequence Alignment

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](#).

Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of

PROTEIN

sequences in any supported format:

```
>CAA73261.1 cathelacidin [Bos taurus]
METQRASFSGLGRSSLWLLLLGLVVPASASQDLSYREAVLRAVDQFNERSSSEANLYRLLLEDPPEQDVEHPGARKPVSFTVKETVCPRTTPQPPEQ
CDFKENGLVKQCVGTVTRYWIRGDFDITCNNIQSAGLFRRLRDSIRRGQQKILEKARRIGERIKDIFRG
>NP_004336.3 cathelacidin antimicrobial peptide preproprotein [Homo sapiens]
MGTMKTQRDGHSLGRWSLVLLLLGLVMPLAIIAQVLSYKEAVLRAIDGINQRSSDANLYRLLDLDPRTMDGDPDTPKPVSTVKETVCPRTTQQS
PEDCDFKKDGLVKRCMGTVTLNQARGSFDISCDKDNKRFBALLGDFFRKSKEKIGKEFKRIVQRIKDFLRNLVPRTES
```

Or, upload a file: Nessun file selezionato

STEP 2 - Set your parameters

OUTPUT FORMAT

Clustal w/o numbers

The default settings will fulfill the needs of most users.

(Click here, if you want to view or change the default settings.)

STEP 3 - Submit your job

☐ Be notified by email (Tick this box if you want to be notified by email when the results are available)

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

Alignments Result Summary Phylogenetic Tree Submission Details

Download Alignment File Show Colors End to Simple_Phylogeny

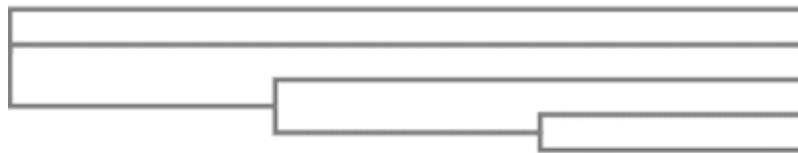
CLUSTAL O(1.2.4) multiple sequence alignment

```

NP_004336.3      MGTMTQRDGHSLGRWSLVLLLLGLVMP LAIIAQVLSYKEAVLRAIDGINQRSSDANLYR
NP_001075774.1  --METHKHGPSLAWMSLLLLLGLMPPA-IAQDLTYREAVLRAVDAFNQSSSEANLYR
AAR26245.1      --METQKDSPLGRWSLLLLLGLVITPA-ASRALSYREAVLRAVNGFNQSSSEANLYR
CAA73261.1      --METQRASFSLGRSSLWLLLLGLVVP SA-SAQDLSYREAVLRAVDQFNERSSSEANLYR
P49932.1        --METQRASLCLGRWSLWLLLLLALVPSA-SAQALSYREAVLRAVDRLNEQSSSEANLYR
                *:::: . ,*  ** ****,*:: * :: *::*****: :*::: ****

NP_004336.3      LLDLDPRTMD-GDPDTPKPVSTVKETVCPRTTQQSPEDCDFKKGDLVKRCMGTVTLNQ
NP_001075774.1  LLSMDPQQLED-AKPYTPQPVSTVKETECPRTTWKLPQCDFKEDGLVKRCVGTVTRYQ
AAR26245.1      LLQLNSQPKGD-EDPNIPKPVSTVKETVCPKTTQQPLEQCGFKDNGLVKQCEGTVILDE
CAA73261.1      LLELDPPPEQDVEHPGARKPVSTVKETVCPRTTQPPEQCDFKENGVLKQCVGTVTRYW
P49932.1        LLELDQPPKAD-EDPGTPKPVSTVKETVCPRTWRPPELCDFKENGVLKQCVGTVTLDQ
                **,::  * ,* :***** **: * : * ,*,,* **: ****

NP_004336.3      ARGSFDISCDKN--KRFALLGDFRKSKEKIGKEFKRIVQRIKDFLRNLVPRTES-
NP_001075774.1  AWDSDIRCNRAQESPEPTGLRKRIRKFRNKIKEKKIGQKIQQGFVPKLPARTDY-
AAR26245.1      DTGYFDLNCDSILQVKKIDRLKELITGGQKIGEKIRRIQRIKDFFKNLQPREEKS
CAA73261.1      IRGDFDITCNIIQSAGLFRRLRDSIRRGQQKILEKARRIGERIKDIFRG-----
P49932.1        IKDPLDITCNEIQSVGLLSRLRDFLSRGRRLGEKIERIGQIKDLSEFFQS-----
                . :*: *:      * . : .:: :. :* :*:::
    
```



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 DNA IN PROTEINA

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
CCTGGTGATGCCTCTGGCCATCATTGCCAGGTCCTCAGCTACAAGGAAGCTGTGCTTCGTGCTATAGATG
GCATCAACCAGCGGTCTCGGATGCTAACCTCTACCGCCTCCTGGACCTGGACCCCAGGCCACGATGGAT
GGGGACCCAGACACGCCAAAGCCTGTGAGCTTCACAGTGAAGGAGACAGTGTGCCCCAGGACGACACAGCA
CGGAAATCTAAAGAGAAGATTGGCAAAGAGTTTAAAGAATTGTCCAGAGAATCAAGGATTTTTTGCGGAA
TCTTGTACCCAGGACAGAGTCCTAG
```

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

DNA strands

- ☒ forward
- ☒ reverse

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

Standard



reset

TRANSLATE!



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

DNA or RNA sequence

```
atggggaccatgaagacccaaagggatggccactccctggggcgggtggtcactggtgctcctgctgctgggctggtgatgctctggccatcattgccaggtcctcagctacaa  
ggagctgtgcttcgtgctatagatggcatcaaccagcggctcctggatgctaacctctaccgcctcctggacctggacccagggccacgatggatggggaccagacacgcca  
agcctgtgagcttcacagtgaaggagacagtgtgccccaggacgacacagcaggaaatctaaagagaagattggcaaagagtttaaagaattgtccagagaatcaaggattttt  
tgcggaatcttgtaccaggacagagtcctag
```

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 LAIIAQVLSYKEAVLRAIDGINQRSSDANLYRLDLDPRPTMDGDPDTPKPVSTVKETVCPRTTQHGNLKRRLAKSLKELSRESRIFCGILYPGQSP

5'3' Frame 2

WGP-RPKGMATPWGGGHWCCCWAW-CLWPSPRSSATRKLCFVL-MASTSGPRMLTSTASWTWTPGPRWMGTQTRQSL-ASQ-RRQCAPGRHSTEI-REDWQRV-KNCPENQGGFFAESCTQDRVL

5'3' Frame 3

GDHEDPKGWPLPGA VVTGAPAAPGPDASGHHC PGPQLQGSCASCYRWHQPAVLGC-PLPPP GPGPQAHDGWGPRHAKACELHSEGDSPVQDDTARKSKEKIGKEFKRIVQRIKDFLANLVPRTES-

3'5' Frame 1

LGLCPGYKIPQKILDSL DNSFKLFANLLFRFPCCVVLGHTVSFTVKLTGFGVSGSPSIVGLGSRRRR-RLASEDRWLMP SIARSTASL-LRTWAMMARGITRPSSRSTSDHRPREWPSLWVFMVP

3'5' Frame 2

-DSVLGTRFRKKS LILWTILLNSLP IFSLDFRAVSSWGTLSPSL-SSQALACLGP HPSWAWGPGGGGRG-HPRTAG-CHL-HEAQLPCS-GPGQ-WPEASPGPAAGAEVTTAPGSGHPFGSSWSP

3'5' Frame 3

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

TRE cornici di lettura per il filamento 1 (5' → 3') e tre cornici di lettura per il filamento complementare (3' ← 5')

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

DNA or RNA sequence

```
atggggaccatgaagacccaaagggatggccactccctggggcgggtgctcactgggtgctcctgctgctgggcctgggtgatgcctctggccatcattgccaggtcctcagctacaa  
ggaagctgtgcttctgtctatagatggcatcaaccagcggtcctcgggatgctaactctaccgctcctggacctggacccagggcccagatggatggggacccagacacgcaa  
agcctgtgagcttcacagtgaaggagacagtgtgtcccgaggacgacacagcacggaaatctaaagagaagattggcaaagagtttaaaagaattgtccagagaatcaaggatttt  
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

```
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 tcg 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
```

Si può anche vedere la sequenza proteica allineata a quella nucleotidica