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 $18 \leq I \geq 25$ (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
```

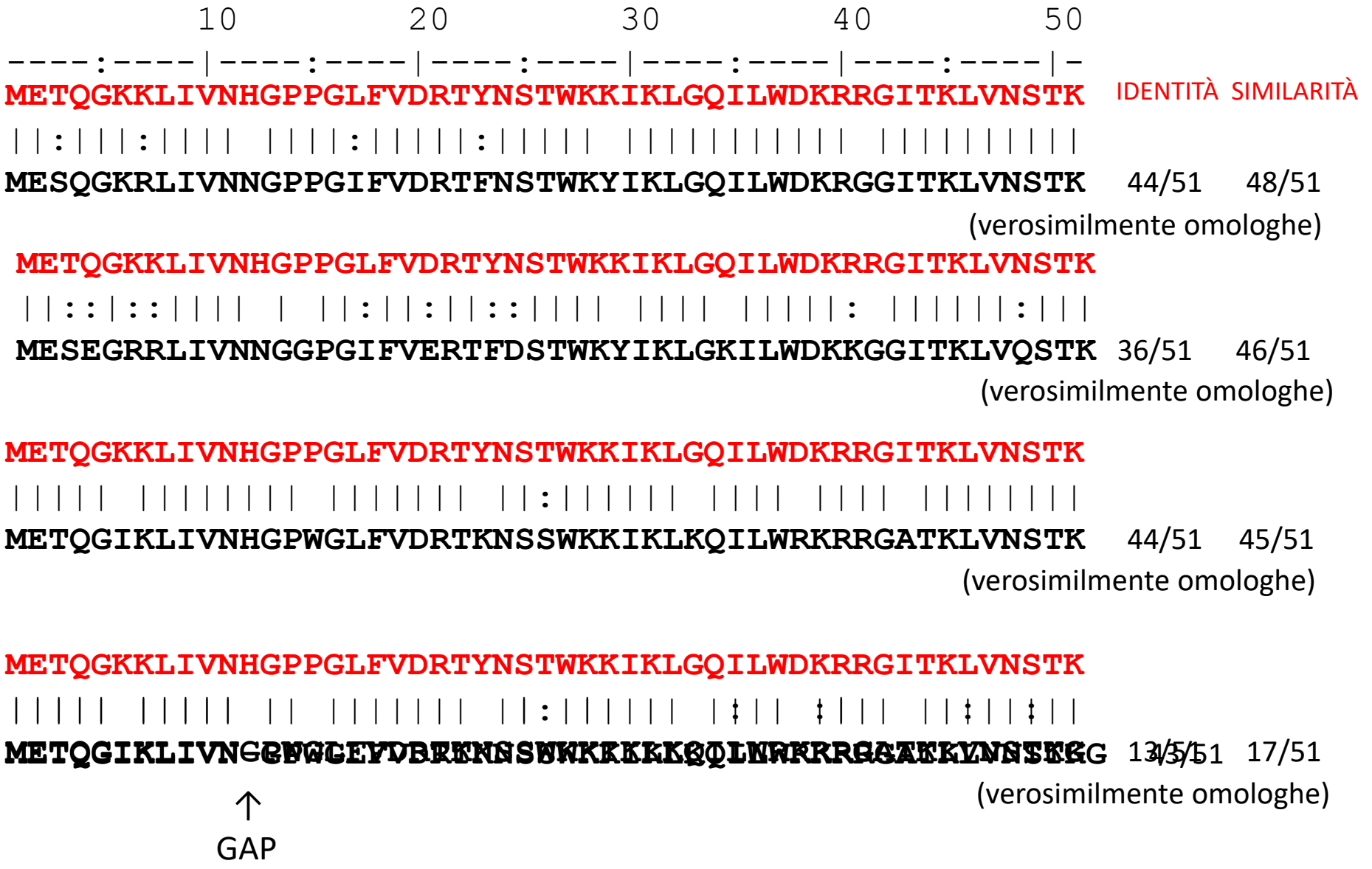
Database

```
MHSSIVLATVLFVAIASASKTRELCMKSLEHAKVGTSKEAKQDGIDL
YKHMFEHYPAMKKYFKHRENYTPADVQKDPFFIKQGQNILLACHV
LCATYDDRETFDAYVGELMARHERDHVKVPNDVWNHFWEHFIEF
LGSK

TTLDEPTKHAWQEIGKEFSHEISHHGRHSVRDHCMNSLEYIAIGDK
EHQKQNGIDLYKHMFEHYPHMRKAFKGRENFTKEDVQKDAFFVN
KDTRFCWPFVCCDSSYDDEPTFDYFVDALMDRHIKDDIHLPQEQ
WHEFWK

LFAEYLNEKSHQHLTEAEKHAWSTIGEDFAHEADKHAKAEKDHHE
GEHKEEHHKLGGENAMKAAVPLFYKKVLADERVKHFFKNTDMDH
QTKQQTDFLTMLLGGPNHYKGKNMT

EAHKGMNLQNLHFDAIIENLAATLKELGVTDAVINEAAKVIEHTRK
DMLGKTINIIKATVPVLKEHGVTITTTFYKNLFAKHPEVRPLFDMGR
QESLEQPKALAMTVLAAAQNIEN

LPAILPAVKKIAVKHCQAGVAAAHYPIVGQELLGAIKEVLGDAATD
DILDAWGKAYGVIADVFIQVEADLYAQAVEQELLGAIKEVLGDAAT
DDILDAWGKAYGVIADVFIQVEADLAQAVE

EPTKHAWQEIGKEFSHEISHHGRHSVRDHCMNSLEYIAIGDKEHQ
KQNGIDLYKHMFEHYPHMRKAFKGRENFTKEDVQKDAFFVNKDT
```

Best alignment

```
METQGKKLKRNHGPPGLFVDRTYNSTWKKSVLGQILW   query
  ||:|||:|   | ||||:|||||:|||||    ||||||
MESQGKRLIVNNGPPGIFVDRTFNSTWKYIKLGQILW   subject
```

Identità e similarità di sequenza vs omologia – ruolo dei GAP



Fattori genotipici e fenotipici che determinano la similarità fra AA

1st letter	Second Letter: U	Second Letter: C	Second Letter: A	Second Letter: G	3rd 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
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

Major forms of mutation are:

Point mutation

Silent (**synonimous**) 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**

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)

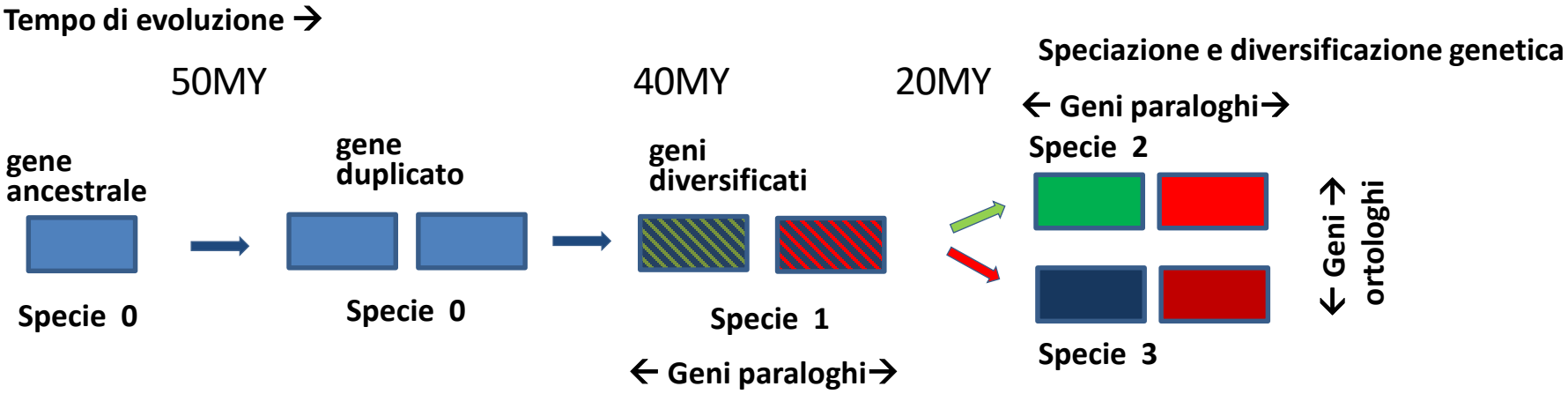
UUU → GGU GGC ☒ ☒
UUC GGA GGG

Ile (I) → Lys (K)

AUU → AAA ☑ ☒
AUC AUG
AUA

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).



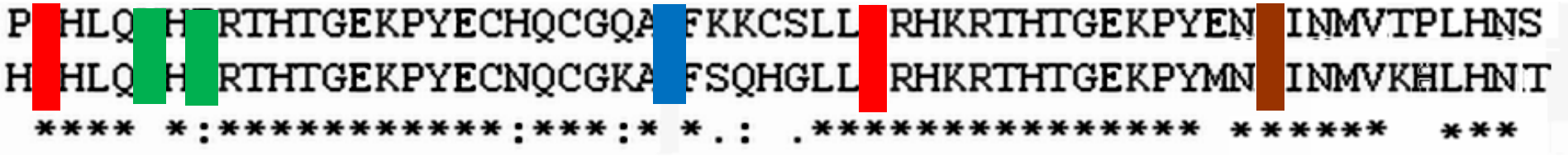
Paralogo: geni nello stesso organismo che derivano da un **gene duplicato e poi diversificato** (le proteine codificate possono avere funzioni diverse avendo seguito un percorso evolutivo diverso)

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



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

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→C, -2) costano di più che altre (es. W con Y, +2, E→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 *prematuro* ha un costo molto elevato

BLOSUM 62 (sequenze ≥ 62% identità)

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

> 3 = *

≥ 1 = .

0 = .

< 0 =

\- Gap init.

--- Gap cont.

PAM - Point Accepted Mutation matrix

```
PSHLQYHERTHTGEKPYECHQCGQATFKKCSLLWRHKRTHTGEKPYEN INMVTPLHNS
HSHLQCHKRTHTGEKPYECNQCGKA-FSQHGLLWRHKRTHTGEKPYMN 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 3O mutazioni/100 residui

PAM120 sequenze con 12O mutazioni / 100 residui (evolutivamente più distanti)

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

Posizioni non sinonime: W → D, -8 è sempre più sfavorevole di Y→C, -1 ma W → Y, F -1 sono peggio che W→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 ↓

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

Score, Bit-score, E-value

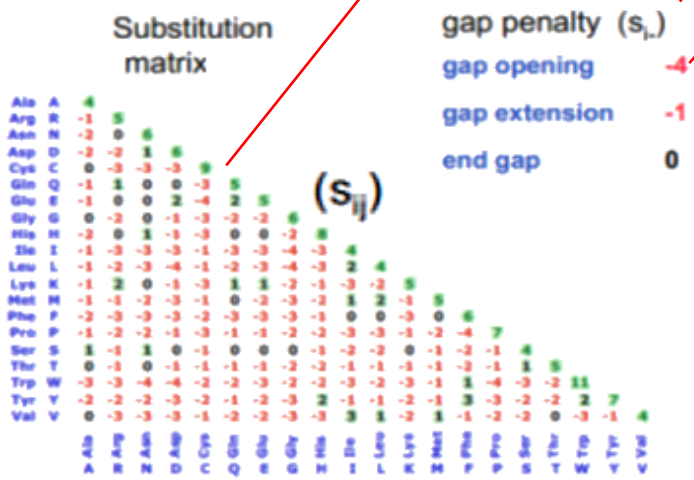
Score: A number used to assess the biological relevance of a finding.

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

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



gap penalty (s_i)

gap opening	-4
gap extension	-1
end gap	0

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 piccolo 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

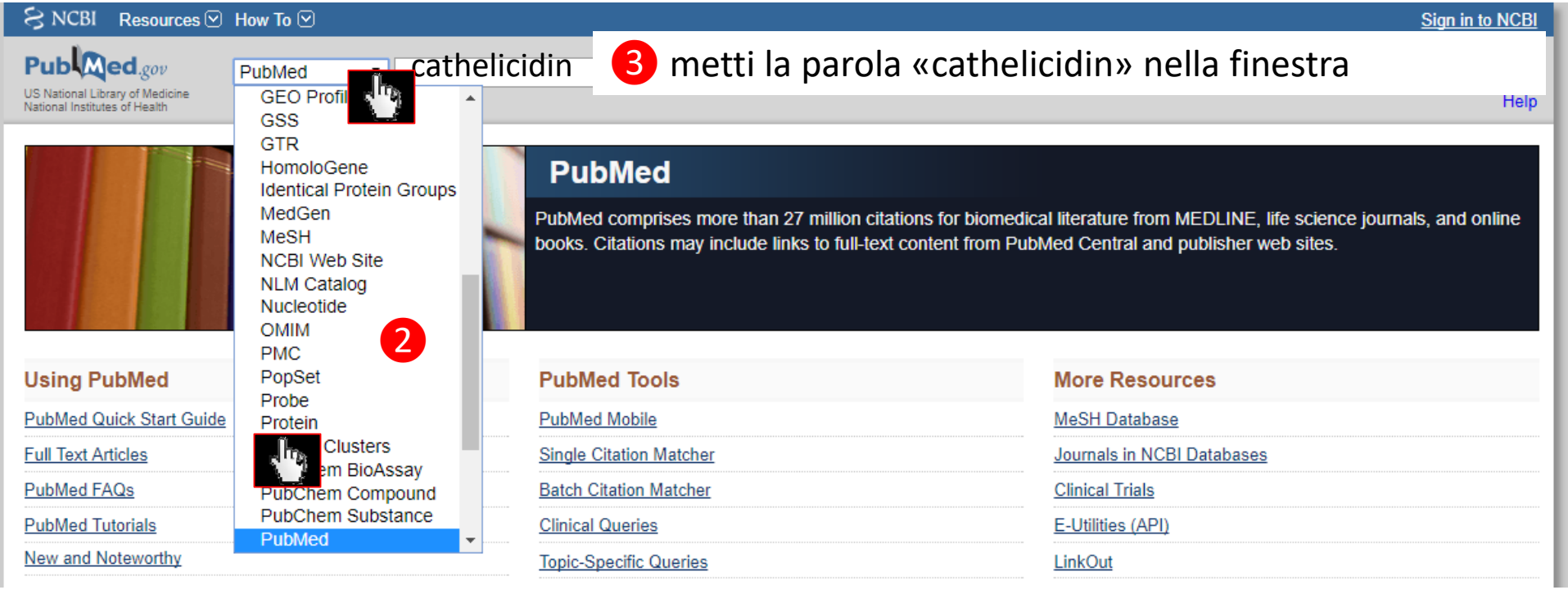
FASTA (euristico) (https://www.ebi.ac.uk/Tools/sss/fasta/) Trova rapidamente tratti brevi e identici tra una sequenza di query e quelle nel database, identificando potenziali regioni di similarità, poi esegue un allineamento più lento e su queste regioni per identificare e perfezionare le corrispondenze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

CLUSTAL (Cluster analysisis) (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»

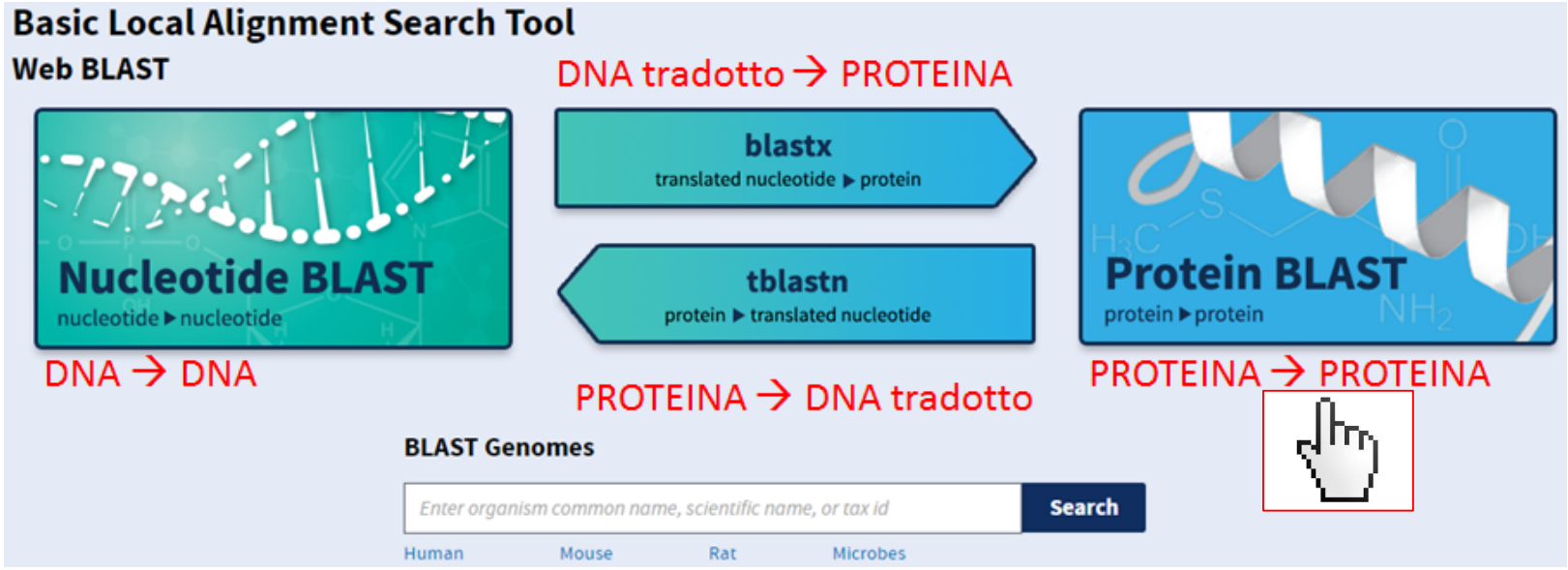


4 Fra i risultati seleziona Bubalus bubalis e copia la sequenza in formato FATSA

Istruzioni per utilizzare BLAST

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

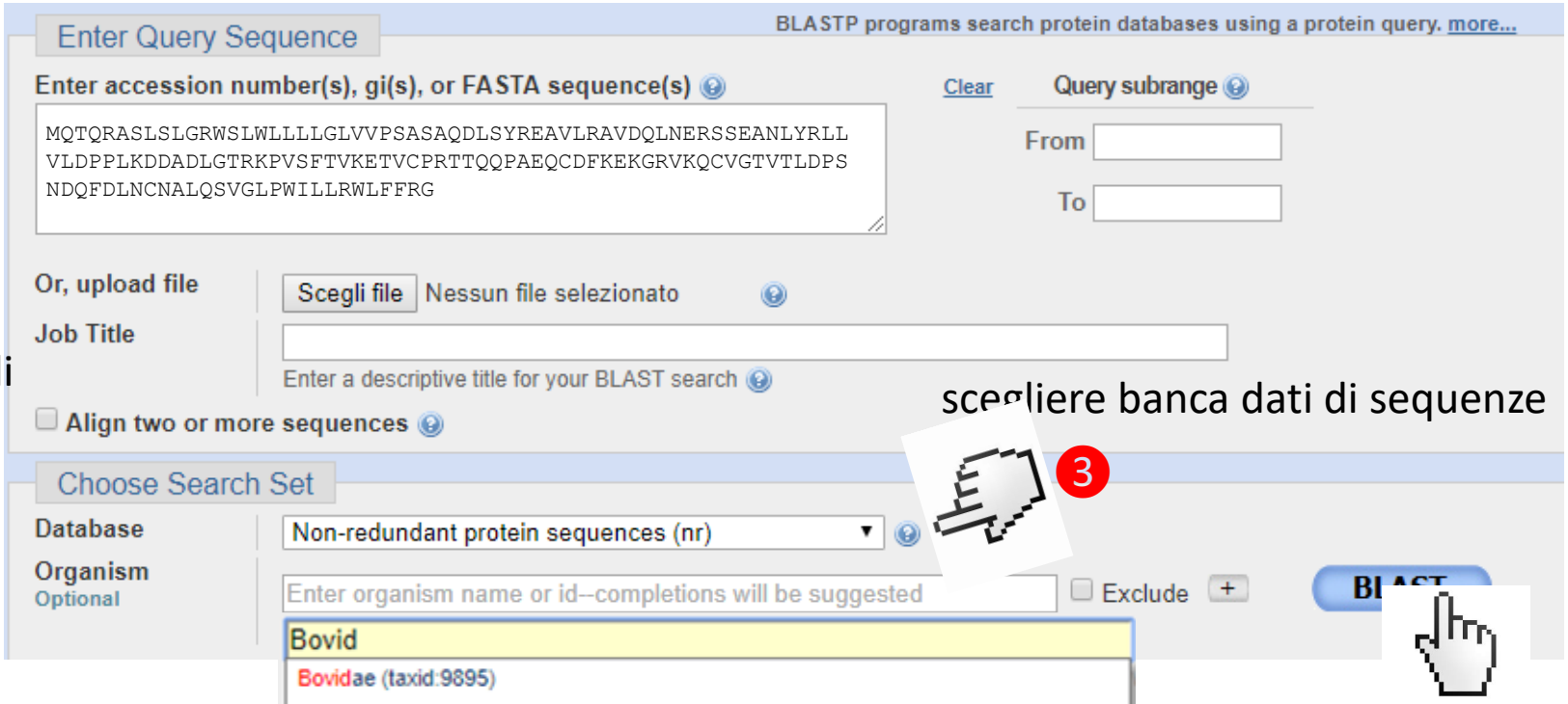
1 Accedi a BLAST e seleziona Protein Blast



2 Paste sequence

3 Scegli la Banca dati di sequenze

4 Scegli l'organismo



‹ Edit Search | Save Search | Search Summary

How to read this report? | BLAST Help Videos | Back to Traditional Results Page

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

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity to

E value to

Query Coverage 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 [But	water buffalo	285	285	100%	1e-99	97.92%	144	AIZ93891.1
cathelicidin 4 [But	water buffalo	259	259	100%	9e-90	93.75%	144	AIZ93880.1

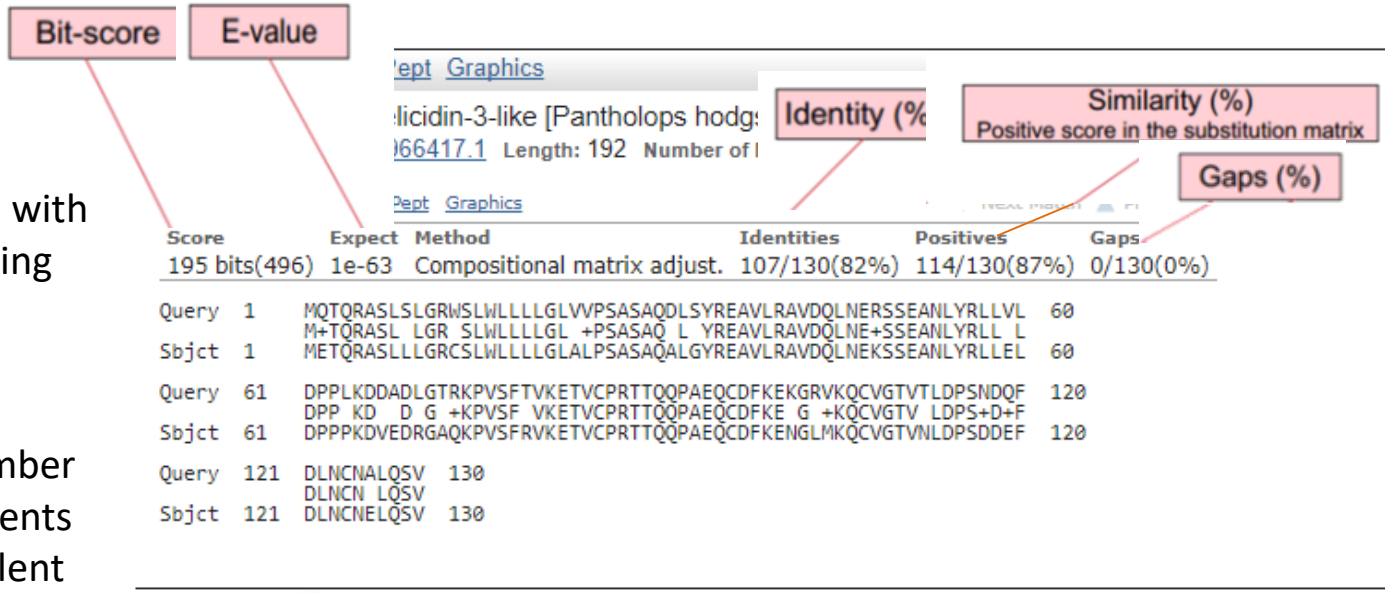
bit score

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 ↑



Istruzioni per utilizzare CLUSTAL https://www.ebi.ac.uk/jdispatcher/msa/clustalo

Allineamento e confronto di sequenze multiple

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

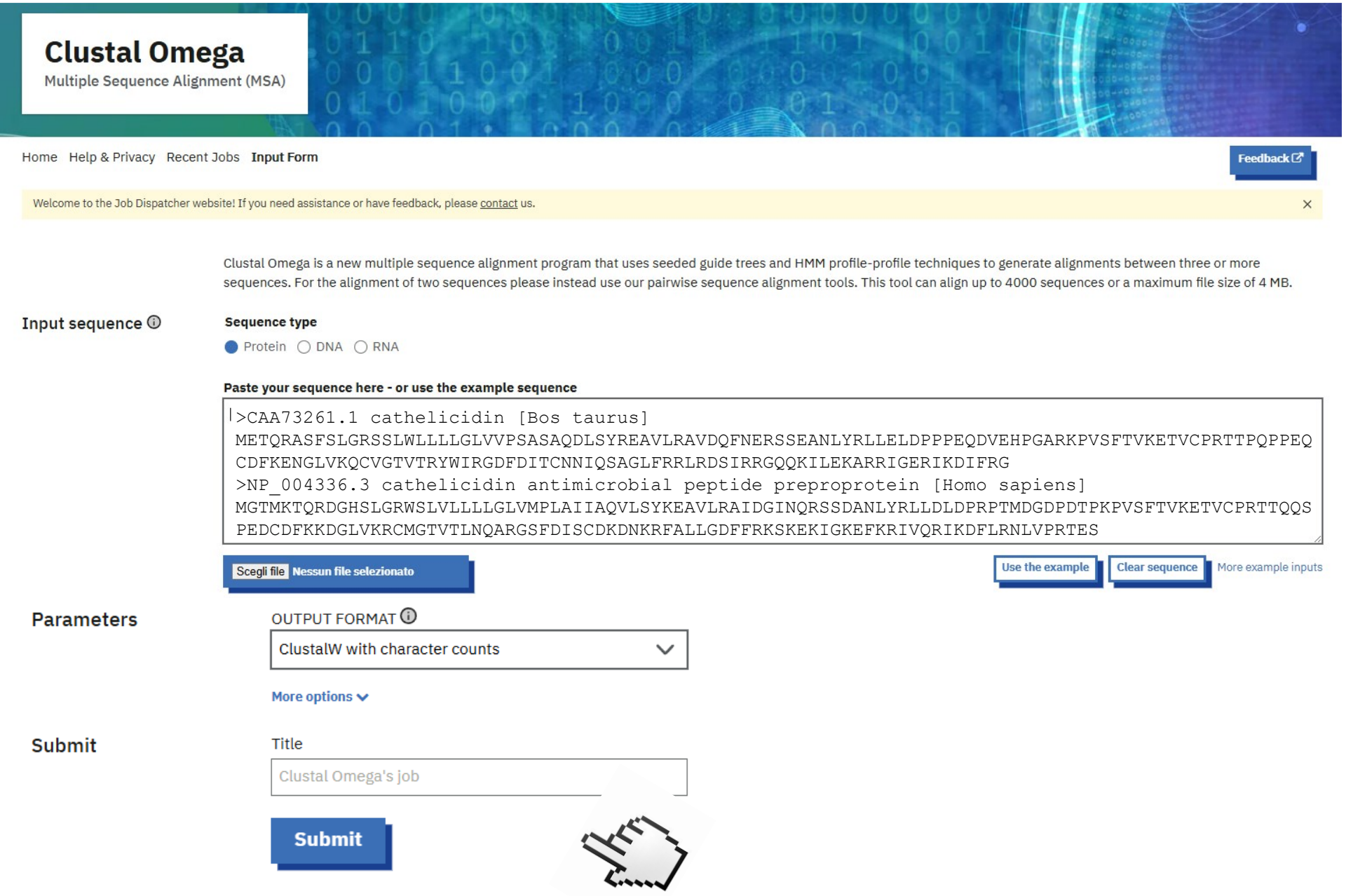
2 Seleziona le sequnza in formato FASTA

>NP_004336.3 cathelicidin antimicrobial peptide preproprotein [Homo sapiens]
MGTMKTQRDGHSLGRWSLVLLLLGLVMPLAIIAQVLSYKEAVLRAIDGINQRSSDANLYRLLDLDPRPTMDGDPDTPKPVS
FTVKETVCPRTTQQSPEDCDFKKDGLVKRCMGTVTLNQARGSFDISCDKDNKRFALLGDFFRKSKEKIGKEFKRIVQRIKD
FLRNLVPRTES (copia tutto in una file di testo)

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

```
>NP_004336.3 cathelicidin antimicrobial peptide preproprotein [Homo sapiens]
MGTMKTQRDGHSLGRWSLVLLLLGLVMPLAIIAQVLSYKEAVLRAIDGINQRSSDANLYRLLDLDPRPTMDGDPDTPKPVSFTVKETVCPRTTQQSPEDCDFKKDGLVKRCM
GTVTLNQARGSFDISCDKDNKRFALLGDFFRKSKEKIGKEFKRIVQRIKDFLRNLVPRTES
>CAA73261.1 cathelicidin [Bos taurus]
METQRASFSLGRSSLWLLLLGLVVPSASAQDLSYREAVLRAVDQFNERSSEANLYRLLELDPPPEQDVEHPGARKPVSFTVKETVCPRTTPQPPEQCDFKENGLVKQCVGTV
TRYWIRGDFDITCNNIQSAGLFRRLRDSIRRGQQKILEKARRIGERIKDIFRG
>NP_999615.1 antibacterial protein PR-39 precursor [Sus scrofa]
METQRASLCLGRWSLWLLLLGLVVPSASTQALSYREAVLRAVDRLNEQSSEANLYRLLELDQPPKADEDP
GTPKPVSFTVKETVCPRPTQRPPELCDFKENGRVKQCVGTVTLNPSNDPLDISCNEIQSVRRRPRPPYLP
RPRPPPFFPPRLPPRIPPGFPPRFPPRFPGKR
>NP_001075774.1 antimicrobial protein CAP18 precursor [Oryctolagus cuniculus]
METHKHGPSLAWWSLLLLLLGLLMPPAIAQDLTYREAVLRAVDAFNQQSSEANLYRLLSMDPQQLEDAKP
YTPQPVSFTVKETECPRTTWKLPEQCDFKEDGLVKRCVGTVTRYQAWDSFDIRCNRAQESPEPTGLRKRL
RKFRNKIKEKLKKIGQKIQGFVPKLAPRTDY
```

4 Apri CLUSTAL OMEGA e incolla tutte le sequenze nella finestra

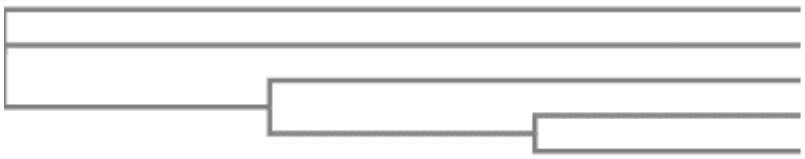


Results for job clustalo-I20171122-114750-0238-29794709-pg

Alignments | Result Summary | Phylogenetic Tree | Submission Details

Download Alignment File | Show Colors | d to Simple_Phylogeny

```
CLUSTAL O(1.2.4) multiple sequence alignment


NP_004336.3         MGTMKTQRDGHSLGRWSLVLLLLGLVMPLAIIAQVLSYKEAVLRAIDGINQRSSDANLYR
NP_001075774.1      ---METHKHGPSLAWWSLLLLLLGLLMPPA-IAQDLTYREAVLRAVDAFNQQSSEANLYR
AAR26245.1          ---METQKDSPSLGRWSLLLLLLGLVITPA-ASRALSYREAVLRAVNGFNQRSSEENLYR
CAA73261.1          ---METQRASFSLGRSSLWLLLLGLVVPSA-SAQDLSYREAVLRAVDQFNERSSEANLYR
P49932.1            ---METQRASLCLGRWSLWLLLLALVVPSA-SAQALSYREAVLRAVDRLNEQSSEANLYR
                       *:*:: . .*.  ** ****.*::  *  :: *:*:******:: :*::**: ****

NP_004336.3         LLDLDPRPTMD-GDPDTPKPVSFTVKETVCPRTTQQSPEDCDFKKDGLVKRCMGTVTLNQ
NP_001075774.1      LLSMDPQQLED-AKPYTPQPVSFTVKETECPRTTWKLPEQCDFKEDGLVKRCVGTVTRYQ
AAR26245.1          LLQLNSQPKGD-EDPNIPKPVSFTVKETVCPKTTQQPLEQCGFKDNGLVKQCEGTVILDE
CAA73261.1          LLELDPPPEQDVEHPGARKPVSFTVKETVCPRTTPQPPEQCDFKENGLVKQCVGTVTRYW
P49932.1            LLELDQPPKAD-EDPGTPKPVSFTVKETVCPRPTWRPPELCDFKENGRVKQCVGTVTLDQ
                    **.::     *  .*   :********* **: * :  * *.**.:* **:* ***

NP_004336.3         ARGSFDISCDKDN--KRFALLGDFFRKSKEKIGKEFKRIVQRIKDFLRNLVPRTES-
NP_001075774.1      AWDSFDIRCNRAQESPEPTGLRKRLRKFRNKIKEKLKKIGQKIQGFVPKLAPRTDY-
AAR26245.1          DTGYFDLNCDSILQVKKIDRLKELITTGGQKIGEKIRRIGQRIKDFFKNLQPREEKS
CAA73261.1          IRGDFDITCNNIQSAGLFRRLRDSIRRGQQKILEKARRIGERIKDIFRG--------
P49932.1            IKDPLDITCNEIQSVGLLSRLRDFLSDRGRRLGEKIERIGQKIKDLSEFFQS-----
                      . :*: *:          * . :    .:: :: .:* ::*:.:
```



NP_004336.3 0.18524 Human Euarchontoglires
NP_001075774.1 0.23488 Rabbit Euarchontoglires
AAR26245.1 0.21227 Dog Laurasiatheria
CAA73261.1 0.12484 Cow Laurasiatheria
P49932.1 0.16784 Pig Laurasiatheria

Tradurre una sequnza 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
GCATCAACCAGCGGTCCTCGGATGCTAACCTCTACCGCCTCCTGGACCTGGACCCCAGGCCCACGATGGAT
GGGGACCCAGACACGCCAAAGCCTGTGAGCTTCACAGTGAAGGAGACAGTGTGCCCCAGGACGACACAGCA
CGGAAATCTAAAGAGAAGATTGGCAAAGAGTTTAAAAGAATTGTCCAGAGAATCAAGGATTTTTTGCGGAA
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

```
atggggaccatgaagacccaaagggatggccactccctggggcggtggtcactggtgctcctgctgctgggcctggtgatgcctctggccatcattgcccaggtcctcagctacaa
ggaagctgtgcttcgtgctatagatggcatcaaccagcggtcctcggatgctaacctctaccgcctcctggacctggaccccaggcccacgatggatggggacccagacacgccaa
agcctgtgagcttcacagtgaaggagacagtgtgccccaggacgacacagcacggaaatctaaagagaagattggcaaagagtttaaaagaattgtccagagaatcaaggatttttt
tgcggaatcttgtacccaggacagagtcctag
```

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

MGTMKTQRDGHSLGRWSLVLLLLGLVMPLAIIAQVLSYKEAVLRAIDGINQRSSDANLYRLLDLDPRPTMDGDPDTPKPVSFTVKETVCPRTTQHGNLKRRLAKSLKELSRESRIFCGILYPGQSP

5'3' Frame 2

WGP-RPKGMATPWGGGHWCSCCWAW-CLWPSLPRSSATRKLCFVL-MASTSGPRMLTSTASWTWTPGPRWMGTQTRQSL-ASQ-RRQCAPGRHSTEI-REDWQRV-KNCPENQGFFAESCTQDRVL

5'3' Frame 3

GDHEDPKGWPLPGAVVTGAPAAGPGDASGHHCPGPQLQGSCASCYRWHQPAVLGC-PLPPPGPGPQAHDGWGPRHAKACELHSEGDSVPQDDTARKSKEKIGKEFKRIVQRIKDFLRNLVPRTES-

3'5' Frame 1

LGLCPGYKIPQKILDSLDNSFKLFANLLFRFPCCVVLGHTVSFTVKLTGFGVSGSPSIVGLGSRSRRR-RLASEDRWLMPSIARSTASL-LRTWAMMARGITRPSSRSTSDHRPREWPSLWVFMVP

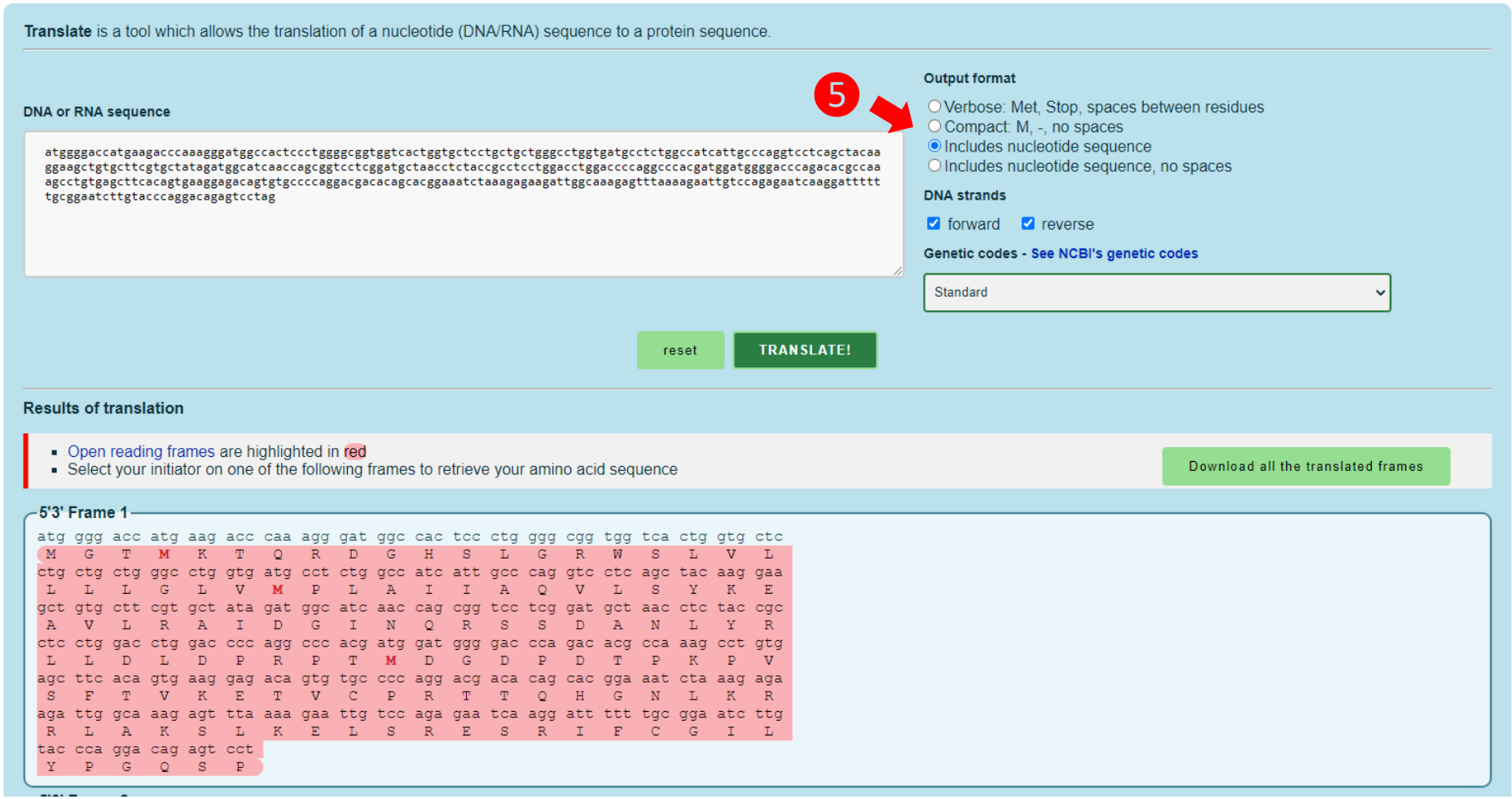
3'5' Frame 2

-DSVLGTRFRKKSLILWTILLNSLPIFSLDFRAVSSWGTLSPSL-SSQALACLGPHPSWAWGPGPGGGRG-HPRTAG-CHL-HEAQLPCS-GPGQ-WPEASPGPAAGAPVTTAPGSGHPFGSSWSP

3'5' Frame 3

RTLSWVQDSAKNP-FSGQFF-TLCQSSL-ISVLCRPGAHCLLHCEAHRLWRVWVPIHRGPGVQVQEAVEVSIRGPLVDAIYSTKHSFLVAEDLGNDGQRHHQAQQQEHQ-PPPQGVAIPLGLHGPH

- 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



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