

Peron Margherita

matricola: 574821

RELAZIONE DI BIOLOGIA MOLECOLARE

## ORGANISMI MODELLO: Confronto tra la sequenza del gene umano CLK1 e il gene clk-1 di C.elegans

### *A- Ottenimento delle sequenze nucleotidiche da confrontare*

In generale, il conseguimento di sequenze nucleotidiche per geni appartenenti a organismi sia modello che non, può avvenire semplicemente consultando il database genomico del NCBI.

Nella home principale è sufficiente aprire il menu a tendina del campo di ricerca *search* e selezionare la voce *gene* per restringere il campo di ricerca alle sole sequenze geniche note.

Poi se si è a disposizione del codice identificativo del gene (in questo caso CLK1 o clk-1), lo si inserisce nel campo di ricerca e si lancia la richiesta; automaticamente si aprirà la pagina con il risultato delle ricerche contro il database Entrez Gene.

Cliccando poi sul risultato corrispondente alla richiesta fatta che dovrebbe trovarsi al primo posto si apre una schermata che riporta le informazioni principali sul gene come il simbolo, la descrizione, la principale fonte di informazioni e l'organismo a cui appartiene.

Scorrendo la schermata verso il basso si trova invece un grafico che schematizza la posizione del gene nel cromosoma, la sua lunghezza e più sotto la sequenza di mRNA indicando la posizione di esoni ed introni.

Cliccando sul link sopra il grafico *Go to reference sequence details* si può infine accedere sia alla cds completa dell'mRNA, tramite un link al database "Nucleotide", sia alla sequenza della proteina tramite invece un link a "Protein"; le sequenze per le CDS si ottengono cliccando sul link CDS e scorrendo la pagina fino a *origin*.

Si è seguito quindi questo procedimento per ottenere la sequenza del trascritto processato per il gene CLK1 di homo sapiens.

The screenshot displays the NCBI Entrez Gene interface for the gene CLK1 (CDC-like kinase 1) in Homo sapiens. The search bar at the top shows "CLK1" entered. The main content area includes a summary section with the following details:

- Official Symbol:** CLK1 (provided by HGNC)
- Official Full Name:** CDC-like kinase 1 (provided by HGNC)
- Primary source:** HGNC:2068
- See related:** Ensembl:ENSG0000013441; Ensembl:ENSG00000240344; HPRD:09058; MIM:601951
- Gene type:** protein coding
- RefSeq status:** REVIEWED
- Organism:** Homo sapiens
- Lineage:** Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
- Also known as:** CLK; STY; CLK/STY; CLK1

The **Summary** section provides a brief description: "This gene encodes a member of the CDC2-like (or LAMMER) family of dual specificity protein kinases. In the nucleus, the encoded protein phosphorylates serine/arginine-rich proteins involved in pre-mRNA processing, releasing them into the nucleoplasm. The choice of splice sites during pre-mRNA processing may be regulated by the concentration of transacting factors, including serine/arginine rich proteins. Therefore, the encoded protein may play an indirect role in governing splice site selection. Multiple transcript variants encoding different isoforms have been found for this gene. [provided by RefSeq]"

Below the summary, the **Genomic regions, transcripts, and products** section is visible, showing the genomic location on chromosome 2 (NC\_000002) and a genomic map of the sequence. The map shows the gene structure with exons and introns, and the sequence coordinates are displayed as -201,731,226 to -201,715,971 (15,256 bases shown, negative strand).

Si è notato in questo caso specifico che esistono due isoforme per questo gene che sono frutto di splicing alternativo: la prima isoforma è più corta (length: 1455 pb) e presenta la seguente sequenza CDS:

```
1 atgagacact caaagagaac ttactgtcct gattgggatg acaaggattg ggattatgga
61 aaatggagga gcagcagcag tcataaaga aggaagagat cacatagcag tgcccaggag
121 aacaagcgt ccaaatacaa tcactctaaa atgtgtgata gccattattt ggaaagcagg
181 tctataaatg agaaagatta tcatagtcga cgctacattg atgagtacag aatgactac
241 actcaaggat gtgaactgg acatcgcaa agagaccatg aaagccggtg tcagaacat
301 agtagcaagt cttctggtag aagtggaaaga agtagttata aaagcaaaca caggattcac
361 cacagtact cacatcgtcg ttcacatggg aagagtcacc gaaggaaaaa aaccaggagt
421 gttagagatg atgaggaggg tcacctgac tgtagagtg gagacgtact aagtgcaaga
481 tatgaaattg ttgatactt aggtgaagga gctttggaa aagtgtgga gtgcatcgt
541 cataaagcgg gaggtagaca ttagcagta aaaatagta aaaatgtgga tagatactgt
601 gaagctgctc gctcagaaa acaagtctg gaacatctga atacaacaga cccaacagt
661 actttccgtg gtgcccagat gttggaatg tttgagcctc atggtcacat ttgactgtt
721 tttgaaactg tgggacttag tactacgac ttcataaag aaaatggtt tctaccatt
781 cgactgac atacagaaa gatggcatat cagatagca agtctgtgaa tttttgcac
841 agtaataagt tgactcacac agacttaag cctgaaaaca tctatttgt gcagtctgac
901 tacacagagg cgtataatc caaaataaa cgtgatgac gcacctaat aaatccagat
961 attaaagtg tagactttg tagtgaaca tatgatgac aacatcacag tacattgga
1021 tctacaagac attatagagc acctgaagt attttagccc tagggtggtc ccaacctagt
1081 gatgtctgga gcataggatg cattctatt gaatactatc ttgggttac cgtatttcca
1141 acacacgata gtaaggagca ttagcaatg atggaaagga ttctggacc tctacaaaa
1201 catatgatac agaaaaccag gaaacgtaa ttttacc acgatcgtat agactgggat
1261 gaacacagtt ctgcccgcag atatgttca agacgctgta aacctctgaa ggaattatg
1321 ctttcaag atgtgaaaca tgagcgtctc ttgacctca ttcagaaaaa gttggagat
1381 gatccagcca aaagaattac tctcagagaa gccttaaac atccttct tgaccttgc
1441 aagaaaagta tata
```

La seconda è invece più lunga (length: ) e presenta la seguente sequenza CDS:

```
1 atggcggctg ggcggaggcc ggcttggcc ctgtggccgg aaaggcgagg ctcccgttg
61 aggggggatt tgctggggtt ccagaatgtg cgtgagccaa gcagctgtgg ggaacactg
121 tctggaatga gacactcaa gagaacttac tgcctgatt gggatgaca ggattggat
181 tatggaaaat ggaggagcag cagcagcat aaaagaagga agagatcaca tagcagtgcc
241 caggagaaca agcgtgcaa atacaatcac tctaaaatgt gtgatagcca ttattggaa
301 agcaggctta taaatgagaa agatattcat agtcgacgct acattgatga gtacagaaa
361 gactacactc aaggatgta acctggacat cgccaagag acctgaaag ccggtatcag
421 aacctagta gcaagtctc tggtagaagt ggaagaagta gttataaag caaacacagg
481 attcaccaca gtacttaca tctctgtca catgggaaga gtcaccgaag gaaaagaacc
541 aggagtgtag aggatgatga ggagggtcac ctgatctgc agagtggaga cgtactaagt
601 gcaagatag aaattgtga tactttagg gaaggagctt ttggaaaagt tgggagtg
661 atcgatcata aagcgggagg tagacatgta gcagtaaaaa tagtaaaaa ttggataga
721 tactgtgaag ctgctcgtc agaaataca gttctggaac atctgaatac aacagacc
781 aacagtact tccgtgtgt ccagatgtt gaatggttg agcatatg tcacattgc
841 attgtttg aactattgg acttagtact tacgactca ttaagaaaa tggtttca
901 ccatttcgac tggatcatat cagaaagatg gcatatcaga tatgcaagtc tgtgaattt
961 ttgacagta ataatgtgac tcacacagac taaagcctg aaaacatctt atttgtcag
1021 tctgactaca cagaggcgt taatccaaa ataaaactg atgaacgcac ctaataaat
1081 ccagatata aagttgtaga ctttgtagt gcaacatag atgacgaaca tcacagata
1141 ttggtatcta caagacatta tagagcacct gaagtatt tagccctagg gtgtccca
1201 ccattgtatg tctggagcat aggatgcatt ctattgaat actatcttg gttaccgta
1261 tttccaac acgatagta ggagcattt gcaatgatg aaaggattc tggacctca
1321 ccaaacata tgatacagaa aaccaggaaa cgtaaatatt taccacga tctgattag
1381 tgggatgaac acagtctgc cggcagatg gttcaagac gctgtaaac tctgaaggaa
1441 tttatcctt ctaagatg tgaacatgag cgtctcttg acctatca gaaaatgtg
1501 gagtatgac cagccaaaag aattactctc agagaagcct taagcatcc tttcttgac
1561 cttcgaaga aaagtatata g
```

Si potrebbe quindi applicare lo stesso procedimento per ottenere la sequenza di clk-1 di C.Elegans.

Per ogni organismo modello sono tuttavia stati creati dei database specifici, di facile e veloce consultazione, che racchiudono tutte le informazioni sia sui trascritti che sulle proteine da essi codificati.

Per Caenorhabditis Elegans questo database è Wormbase: [www.wormbase.org](http://www.wormbase.org) .

Inserendo nello spazio per la ricerca il codice del gene si ottiene subito la seguente schermata che riassume le informazioni generali sul gene: *Gene Summary*.

Home Genome Synteny Blast / Blast WormMart Markers Genetic Maps Submit Searches Site Map

Find:

WormBase The Biology and Genome of *C. elegans*.

Gene Summary Locus Summary Sequence Summary Protein Summary EST Alignments Genome Browser Genetic Map Nearby Genes Bibliography Tree Display XML Schema AceDb Image

## Gene Summary for clk-1

Specify a gene using a gene name ([unc-26](#)), a predicted gene id ([R13A5.9](#)), or a protein ID ([CE02711](#)) [[clk-1](#)]

[[identification](#)] [[location](#)] [[function](#)] [[expression](#)] [[gene ontology](#)] [[genetics](#)] [[homology](#)] [[reagents](#)] [[bibliography](#)]

Identification	IDs:	Main name	Sequence name	Other name(s)	WB Gene ID
		clk-1 - ( <i>CLock (biological timing) abnormality</i> ) via Person evidence: Siegfried Hekimi	ZC395.2	coq-7 (via CGC data submission) coq-1	WBGene0000536

**Concise Description:** clk-1 encodes the *C. elegans* ortholog of COQ7/CAT5, a highly conserved demethoxyubiquinone (DMO) hydroxylase that is necessary for the biosynthesis of ubiquinone (coenzyme Q, Q9) from 5-demethoxyubiquinone (DMQ9); in *C. elegans*, CLK-1 activity is required for normal physiological rates of growth, development, behavior, and aging, as well as for normal brood sizes. [[details](#)]

**NCBI KOGs:** DMO mono-oxygenase/Ubiquinone biosynthesis protein COQ7/CLK-1/CAT5 [[KOG4061](#)]; [[OMPpre\\_WH008690](#)]

**Species:** *Caenorhabditis elegans*

**NCBI:** [Other sequences](#)  
[[AceView: 3G369](#)] [[RefSeq: NM\\_065727](#)]

Gene Model	Status	Nucleotides (coding/transcript)	Protein	Amino Acids
ZC395.2.1 1, 2, 3	confirmed by cDNA(s)	564/1472 bp	WP:CE01438	187 aa
ZC395.2.2 1, 2, 3	confirmed by cDNA(s)	564/1275 bp	WP:CE01438	187 aa

[Footnotes](#)  
[Other Notes](#)  
[Revision History](#)

**Location**

**Genetic Position:** III:-1.85 +/- 0.003 cM [[mapping data](#)]  
**Genomic Position:** III:5277853..5279324 bp [[View Genome Browser](#)]  
**Genomic Environs:**

Il nome completo del gene è **CLock (biological timing) abnormality**.

Al di sotto del campo IDs vi è una breve descrizione delle funzioni del gene ma per ottenere informazioni più dettagliate basta scorrere la facciata verso il basso fino al campo *Functions* dove sono riportati non solo gli effetti molecolari e fenotipici derivanti dall'espressione del gene, ma anche gli esperimenti che sono serviti per osservarli e dimostrarli.

Cliccando su Locus Summary si ottengono invece informazioni sulla posizione del gene: clk-1 si trova sul cromosoma 3 in posizione -1.85 +/- 0.003 cM (centiMorgan) ovvero dalla posizione 5,277,853 alla 5,279,324.

Sul fondo della schermata in corrispondenza del campo *Allels* si trovano i nomi dei possibili alleli per questo gene, in tutto 7.

Cliccando quindi su *Sequence Summary* si ottengono infine tutte le informazioni utili sulla sequenza nucleotidica, compresa la sequenza completa dei due possibili trascritti di questa zona che si trova cliccando sui link con nome *Transcripts in this region*.

In questo caso tuttavia non sembra trattarsi di splicing alternativi ma di due distinti modelli per il gene di cui uno che si allunga nella parte C-terminale ma che mantiene lo stesso numero e la stessa posizione per gli esoni rispetto all'altro.

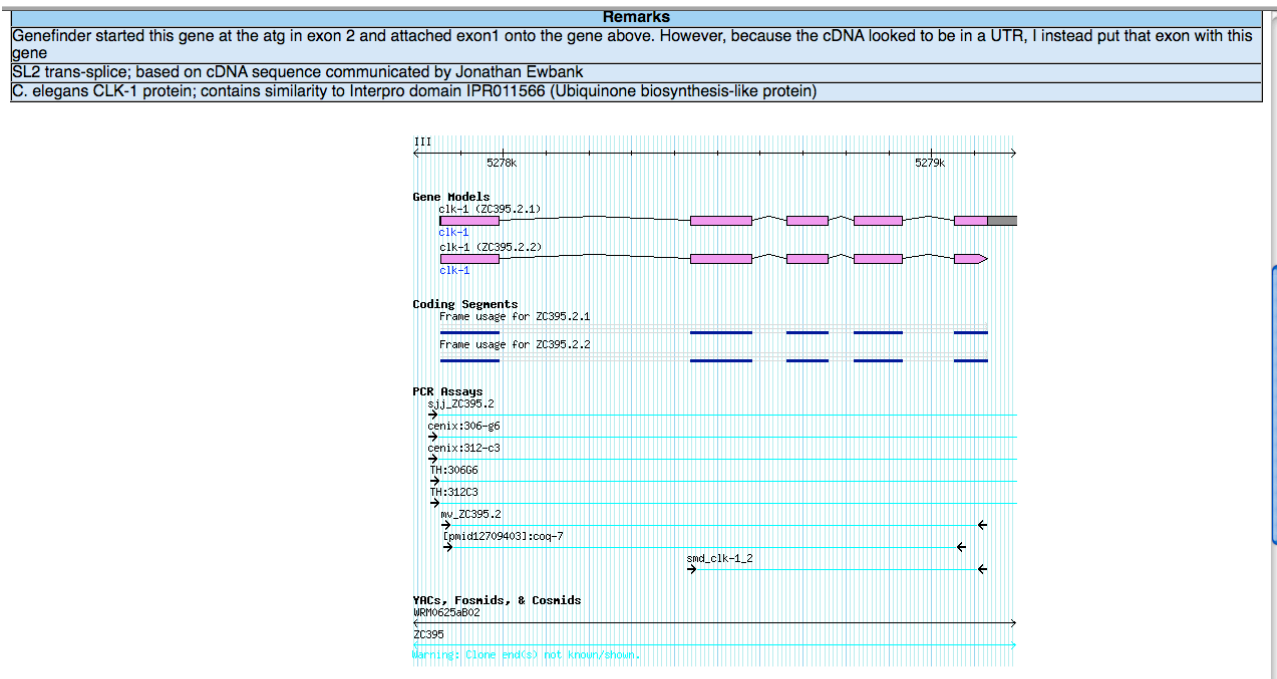
La sequenza più corta ZC395.2.2, è lunga 564 bp:

```
ATGTTCCGTG TAATAACCCG TGGAGCAT ACTGCTGCTT CTCGTCAGC ACTTATAGAG AAGATCATTG GAGTTGATCA
TGCTGGAGAG CTTGGAGCCG ATCGGATTTA CGCTGGACAG TTGGCTGTTT TGCAAGGTTT ATCTGTTGGT TCAGTAATCA
AAAAGATGTG GGATGAGGAG AAAGAACATT TAGATACAAT GGAAAGATTA GCTGCTAAAC ACAATGTACC TCATACTGTT
TTCTCTCCAG TTTTCAGTGT GGCTGCTTAT GCTCTCGGTG TCGGTTTCAGC ACTTCTAGGA AAAGAAGGTG CAATGGCTTG
TACAATTGCA GTTGAAGAAC TCATTGGACA ACATTATAAT GATCAATTGA AAGAACTCCT TGCCGACGAT CCTGAAACAC
ACAAAGAATT GCTGAAAATT CTCACAAGAT TACGTGATGA GGAGCTTCAT CATCATGATA CTGGAGTAGA ACACGATGGA
ATGAAGGCTC CAGCCTACTC GGCTCTCAA TGGATTATC AAAGTGGATG CAAGGGAGCT ATTGCGATTG CTGAGAAAAT
TTGA
```

La più lunga invece ZC395.2.1 è lunga 746 bp:

```
agATGTTCCG TGTAATAACC CGTGGAGCAC ATACTGCTGC TTCTCGTCAA GCACTTATAG AGAAGATCAT TCGAGTTGAT
CATGCTGGAG AGCTTGGAGC CGATCGGATT TACGCTGGAC AGTTGGCTGT TTTGCAAGGT TCATCTGTTG GTTCAGTAAT
CAAAAAGATG TGGGATGAGG AGAAAGAACA TTTAGATACA ATGGAAAGAT TAGCTGCTAA ACACAATGTA CCTCATACTG
TTTTCTCTCC AGTTTTCAGT GTGGCTGCTT ATGCTCTCGG TGTCGGTTCA GCACTTCTAG GAAAAGAAGG TGCAATGGCT
TGTACAATTG CAGTTGAAGA ACTCATTGGA CAACATTATA ATGATCAATT GAAAGAACTC CTTGCCGACG CCTGAAACAC
ACACAAAGAA TTGCTGAAAA TTCTCACAAG ATTACGTGAT GAGGAGCTTC ATCATCATGA TACTGGAGTA GAACACGATG
GAATGAAGGC TCCAGCCTAC TCGGCTCTCA AATGGATTAT TCAAAGTGGG TGCAAGGGAG CTATTGCGAT TGCTGAGAAA
ATTTGAacaa ttcaataat gtgcccttt tctccatgg tctttgcaa aatataccca tagacgctgc tattccccag
attctagatg atctcaact cttgtgtaaaa acaacttttt tgaaaatatt ttctglttag aaatgcccga ctgaccactg
cagataaaca atattctagt aaaaaa
```

Scendendo nella pagina un grafico descrive la disposizione di esoni e introni sulla sequenza di DNA del cromosoma: nel caso di *clk-1* ci sono cinque esoni.



#### Sequence

```
>ZC395.2
atgttccggtgtaataaacccggtggagcacatactgctgctctcgtcaagc
acttatagagaagatcattccagttgatcatgctggagagctggagocg
atcgatttaagctggacagttgctgttttgaagggtcactcgttgggt
tcaatatacaaaaatctcqqatcaqcaaaaqaacattatagatacaat
```

### B- Allineamento delle sequenze nucleotidiche

L'allineamento delle sequenze può essere fatto semplicemente utilizzando ClustalW in un multi-allineamento comprendente i due diversi splicing del gene CLK1 e il modello ZC395.2.2 di *clk-1*.

Quest'ultimo è stato scelto a favore dell'altro modello sulla base di un confronto incrociato con il database NCBI che restituisce solo una delle due sequenze CDS.

Dall'allineamento multiplo con ClustalW non si evince alcuna zona particolarmente conservata, e il grafico dell'albero filogenetico fornito dal tool *Calculate tree* conferma la lontananza evolutiva tra le sequenze esoniche dei due geni.

Il secondo screenshot mostra lo score dei singoli allineamenti a coppie delle sequenze: in entrambi gli allineamenti con *clk-1* lo score è di 75 contro invece uno score di 97 delle due isoforme di CLK1 umana.

### C- Conclusioni

Nonostante l'evidente somiglianza funzionale delle proteine codificate da questi due geni le sequenze di mRNA che codificano per esse, e di conseguenza i geni, non sembrano evidenziare particolari somiglianze o regioni conservate.

Non si è stati in grado, utilizzando dei programmi alternativi, di confrontare in maniera più efficace le sequenze nucleotidiche: sia l'allineamento delle CDS che quello dell'intera sequenza genica comprese le regioni istoniche danno gli stessi risultati poco significativi.

- Help
- FAQ
- Jalview

EBI > Tools > Multiple Sequence Alignments > ClustalW2

### ClustalW2 Results

[Alignments](#) | [Result Summary](#) | [Guide Tree](#) | [Submission Details](#) | [Submit Another Job](#)

#### Guide Tree

[View Guide Tree File](#)

```
{
gi|71996651|ref|NM_065727.2|:0.23441,
gi|241666391|ref|NM_001162407.1|:0.01113,
gi|241666390|ref|NM_004071.3|:0.01525);
```

#### Phylogram

[Show as Cladogram Tree](#) | [Show Distances](#)



Right-click on the above tree to see display options.

#### CLUSTAL 2.0.12 Multiple Sequence Alignments

```
Sequence type explicitly set to DNA
Sequence format is Pearson
Sequence 1: gi|71996651|ref|NM_065727.2| 729 bp
Sequence 2: gi|241666391|ref|NM_001162407.1| 2213 bp
Sequence 3: gi|241666390|ref|NM_004071.3| 1933 bp
Start of Pairwise alignments
Aligning...

Sequences (1:2) Aligned. Score: 75
Sequences (1:3) Aligned. Score: 75
Sequences (2:3) Aligned. Score: 97
Guide tree file created: [clustalw2-I20101219-173516-0962-5371536.dnd]

There are 2 groups
Start of Multiple Alignment

Aligning...
Group 1: Sequences: 2 Score:35518
Group 2: Sequences: 3 Score:9271
Alignment Score 16947

CLUSTAL-Alignment file created [clustalw2-I20101219-173516-0962-5371536.aln]
```