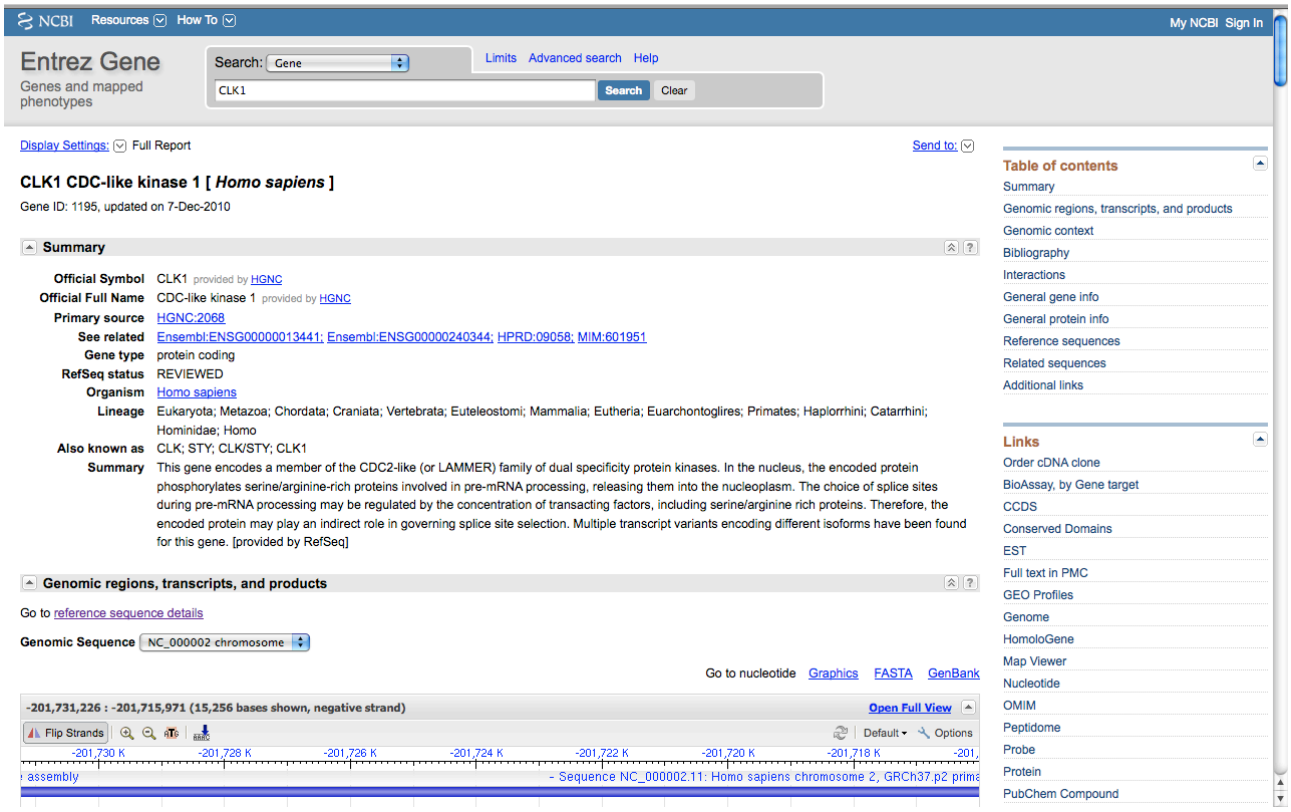
Peron Margherita matricola: 574821

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 gnomico 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.



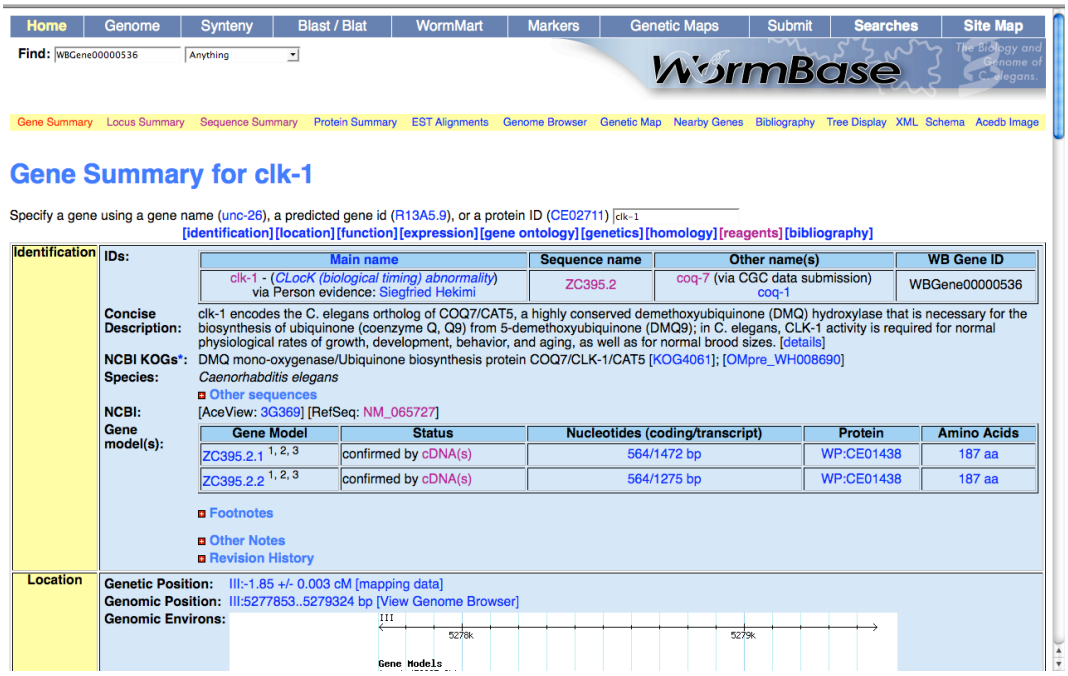
Si è notato in questo caso specifico che esistono due isoforme per questo gene che sono frutto di splicing alternativo: la prima isoforma è più corta (lenght:1455 pb) e presenta la seguente sequenza CDS:

```
   1 atgagacact caaagagaac ttactgtcct gattgggatg acaaggattg ggattatgga
  61 aaatggagga gcagcagcag tcataaaaga aggaagagat cacatagcag tgcccaggag
 121 aacaagcgct gcaaatacaa tcactctaaa atgtgtgata gccattattt ggaaagcagg
 181 tctataaatg agaaagatta tcatagtcga cgctacattg atgagtacag aaatgactac
 241 actcaaggat gtgaacctgg acatcgccaa agagaccatg aaagccggta tcagaaccat
 301 agtagcaagt cttctggtag aagtggaaga agtagttata aaagcaaaca caggattcac
 361 cacagtactt cacatcgtcg ttcacatggg aagagtcacc gaaggaaaag aaccaggagt
 421 gtagaggatg atgaggaggg tcacctgatc tgtcagagtg gagacgtact aagtgcaaga
 481 tatgaaattg ttgatacttt aggtgaagga gcttttggaa aagttgtgga gtgcatcgat
 541 cataaagcgg gaggtagaca tgtagcagta aaaatagtta aaaatgtgga tagatactgt
 601 gaagctgctc gctcagaaat acaagttctg gaacatctga atacaacaga ccccaacagt
 661 actttccgct gtgtccagat gttggaatgg tttgagcatc atggtcacat ttgcattgtt
 721 tttgaactat tgggacttag tacttacgac ttcattaaag aaaatggttt tctaccattt
 781 cgactggatc atatcagaaa gatggcatat cagatatgca agtctgtgaa ttttttgcac
 841 agtaataagt tgactcacac agacttaaag cctgaaaaca tcttatttgt gcagtctgac
 901 tacacagagg cgtataatcc caaaataaaa cgtgatgaac gcaccttaat aaatccagat
 961 attaaagttg tagactttgg tagtgcaaca tatgatgacg aacatcacag tacattggta
1021 tctacaagac attatagagc acctgaagtt attttagccc tagggtggtc ccaaccatgt
1081 gatgtctgga gcataggatg cattcttatt gaatactatc ttgggtttac cgtatttcca
1141 acacacgata gtaaggagca tttagcaatg atggaaagga ttcttggacc tctaccaaaa
1201 catatgatac agaaaaccag gaaacgtaaa tattttcacc acgatcgatt agactgggat
1261 gaacacagtt ctgccggcag atatgtttca agacgctgta aacctctgaa ggaatttatg
1321 ctttctcaag atgttgaaca tgagcgtctc tttgacctca ttcagaaaat gttggagtat
1381 gatccagcca aaagaattac tctcagagaa gccttaaagc atcctttctt tgaccttctg
1441 aagaaaagta tatag
```

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

```
   1 atggcggctg ggcggaggcc ggcttcggcc ctgtggccgg aaaggcgagg ctccccgttg
  61 agggggggatt tgctggggtt ccagaatgtg cgtgagccaa gcagctgtgg ggaaacgttg
 121 tctggaatga gacactcaaa gagaacttac tgtcctgatt gggatgacaa ggattgggat
 181 tatggaaaat ggaggagcag cagcagtcat aaaagaagga agagatcaca tagcagtgcc
 241 caggagaaca agcgctgcaa atacaatcac tctaaaatgt gtgatagcca ttatttggaa
 301 agcaggtcta taaatgagaa agattatcat agtcgacgct acattgatga gtacagaaat
 361 gactacactc aaggatgtga acctggacat cgccaaagag accatgaaag ccggtatcag
 421 aaccatagta gcaagtcttc tggtagaagt ggaagaagta gttataaaag caaacacagg
 481 attcaccaca gtacttcaca tcgtcgttca catgggaaga gtcaccgaag gaaaagaacc
 541 aggagtgtag aggatgatga ggagggtcac ctgatctgtc agagtggaga cgtactaagt
 601 gcaagatatg aaattgttga tactttaggt gaaggagctt ttggaaaagt tgtggagtgc
 661 atcgatcata aagcgggagg tagacatgta gcagtaaaaa tagttaaaaa tgtggataga
 721 tactgtgaag ctgctcgctc agaaatacaa gttctggaac atctgaatac aacagacccc
 781 aacagtactt tccgctgtgt ccagatgttg gaatggtttg agcatcatgg tcacatttgc
 841 attgtttttg aactattggg acttagtact tacgacttca ttaaagaaaa tggttttcta
 901 ccatttcgac tggatcatat cagaaagatg gcatatcaga tatgcaagtc tgtgaatttt
 961 ttgcacagta ataagttgac tcacacagac ttaaagcctg aaaacatctt atttgtgcag
1021 tctgactaca cagaggcgta taatcccaaa ataaaacgtg atgaacgcac cttaataaat
1081 ccagatatta aagttgtaga ctttggtagt gcaacatatg atgacgaaca tcacagtaca
1141 ttggtatcta caagacatta tagagcacct gaagttattt tagccctagg gtggtcccaa
1201 ccatgtgatg tctggagcat aggatgcatt cttattgaat actatcttgg gtttaccgta
1261 tttccaacac acgatagtaa ggagcattta gcaatgatgg aaaggattct tggacctcta
1321 ccaaaacata tgatacagaa aaccaggaaa cgtaaatatt ttcaccacga tcgattagac
1381 tgggatgaac acagttctgc cggcagatat gtttcaagac gctgtaaacc tctgaaggaa
1441 tttatgcttt ctcaagatgt tgaacatgag cgtctctttg acctcattca gaaaatgttg
1501 gagtatgatc cagccaaaag aattactctc agagaagcct taaagcatcc tttctttgac
1561 cttctgaaga 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 .
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*.



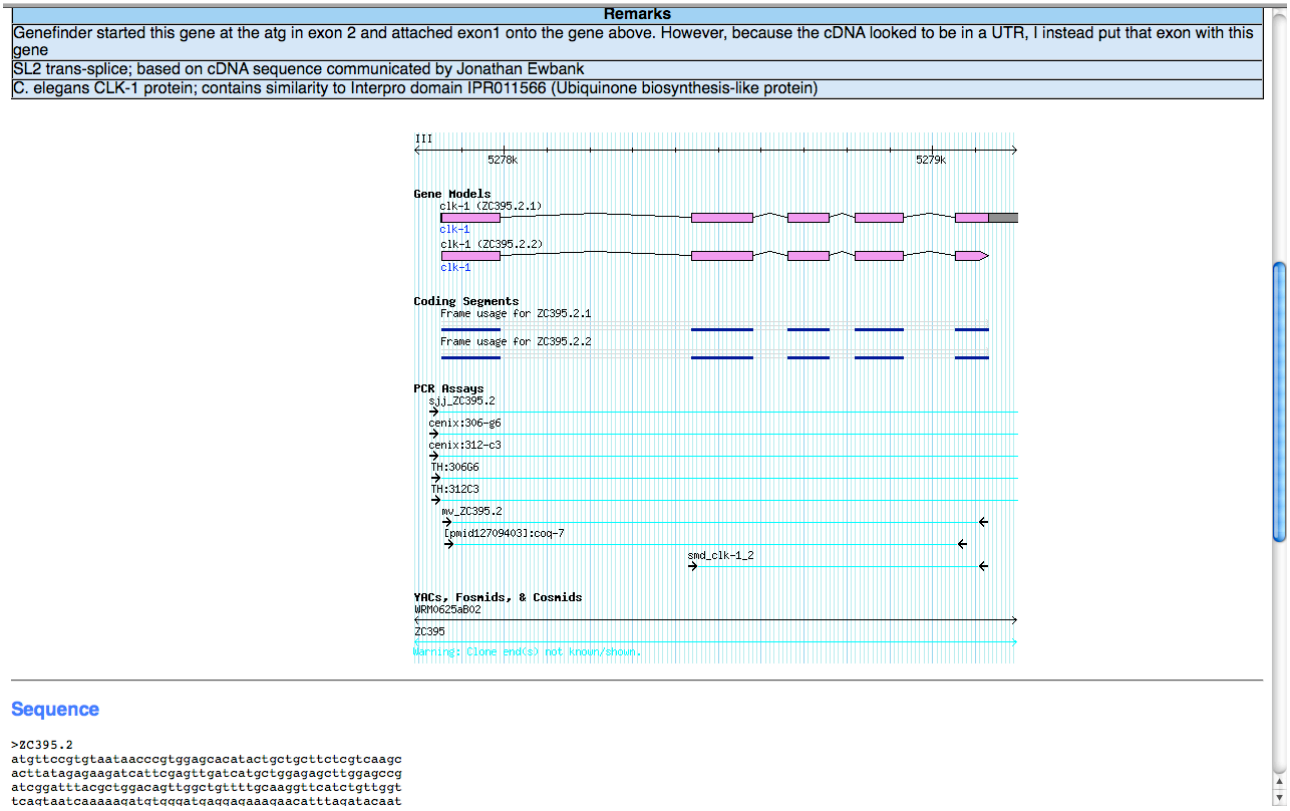

Il nome completo del gene è **CLocK (biological timing) abnormality.**
Al di sotto del campo IDs vi è una breve descrizione delle funzioni del gene ma per ottenere informazioni più dettagliate basta scorrere la facciata verso il basso fino al campo *Functions* dove sono riportati non solo gli effetti molecolari e fenotipici derivanti dall'espressione del gene, ma anche gli esperimenti che sono serviti per osservarli e dimostrarli.
Ciccando 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.
Ciccando 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 TGGAGCACAT ACTGCTGCTT CTCGTCAAGC ACTTATAGAG AAGATCATTC GAGTTGATCA
TGCTGGAGAG CTTGGAGCCG ATCGGATTTA CGCTGGACAG TTGGCTGTTT TGCAAGGTTC ATCTGTTGGT TCAGTAATCA
AAAAGATGTG GGATGAGGAG AAAGAACATT TAGATACAAT GGAAAGATTA GCTGCTAAAC ACAATGTACC TCATACTGTT
TTCTCTCCAG TTTTCAGTGT GGCTGCTTAT GCTCTCGGTG TCGGTTCAGC ACTTCTAGGA AAAGAAGGTG CAATGGCTTG
TACAATTGCA GTTGAAGAAC TCATTGGACA ACATTATAAT GATCAATTGA AAGAACTCCT TGCCGACGAT CCTGAAACAC
ACAAAGAATT GCTGAAAATT CTCACAAGAT TACGTGATGA GGAGCTTCAT CATCATGATA CTGGAGTAGA ACACGATGGA
ATGAAGGCTC CAGCCTACTC GGCTCTCAAA TGGATTATTC AAACTGGATG 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 ATCCTGAAAC
ACACAAAGAA TTGCTGAAAA TTCTCACAAG ATTACGTGAT GAGGAGCTTC ATCATCATGA TACTGGAGTA GAACACGATG
GAATGAAGGC TCCAGCCTAC TCGGCTCTCA AATGGATTAT TCAAACTGGA TGCAAGGGAG CTATTGCGAT TGCTGAGAAA
ATTTGAacaa tttcaataat gtgtcccttt tcttccatgg ttctttgcaa aatatcccca tagacgctgc tattccccag
attctagatg atcctcaact ctttgtaaaa acaacttttt tgaaaatatt ttctcgttag aaatgtccga 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.

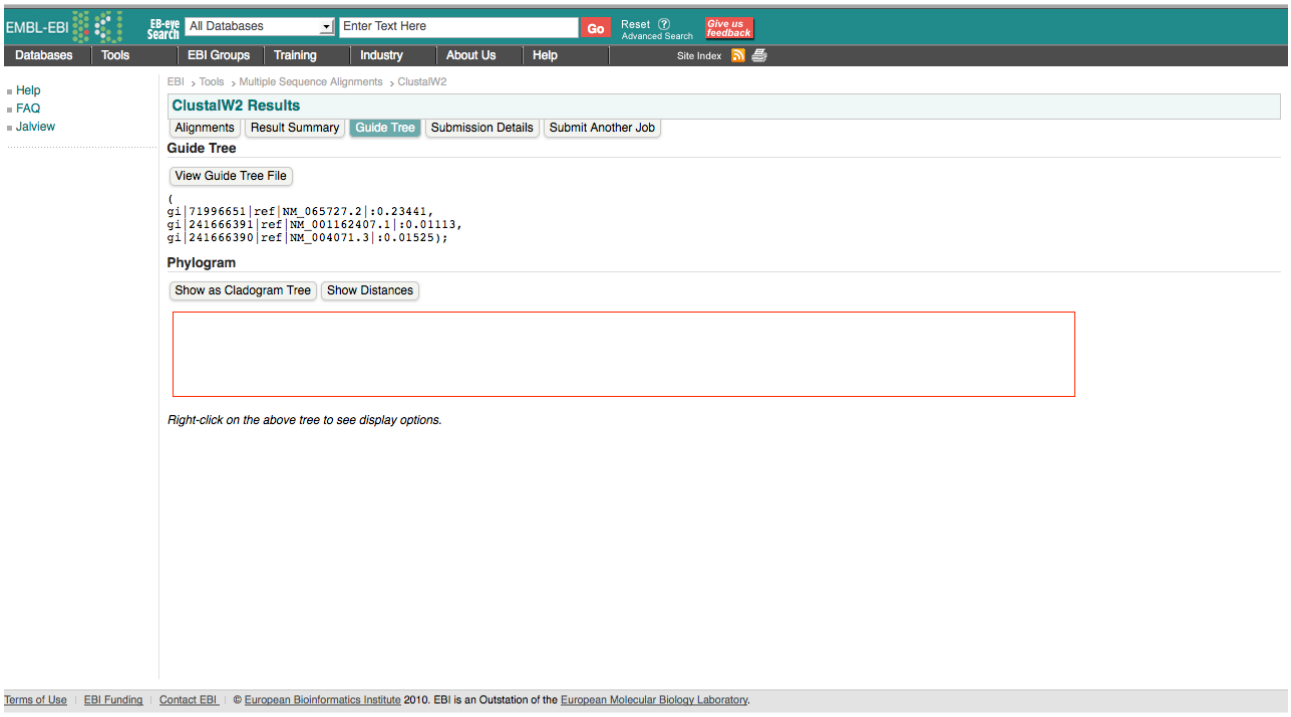



## *B- Allineamento delle sequenze nucleotidiche*

L'allineamento delle sequenze può essere fatto semplicemente utilizzando ClustalW in un multiallineamento 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 *Calcolate 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.




```
 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]
```