Annotation Structurale: Caractérisation d’un domaine THAP chez le Zebrafish

Présenté par Célia, Ema et Véronique

Contexte…

PubMed®

Mol Biol Evol. 2005 Apr;22(4):833-44. doi: 10.1093/molbev/msi068. Epub 2004 Dec 22.

Homologs of Drosophila P transposons were mobile in zebrafish but have been domesticated in a common ancestor of chicken and human

Sabine E Hammer [1], Sabine Strehl, Sylvia Hagemann

Homologue de l’élément P de la drosophile
retrouvé chez le zebrafish (*Pdre2*) et chez l’Homme (Phsa)

Domaine THAP similaire en position N-terminale

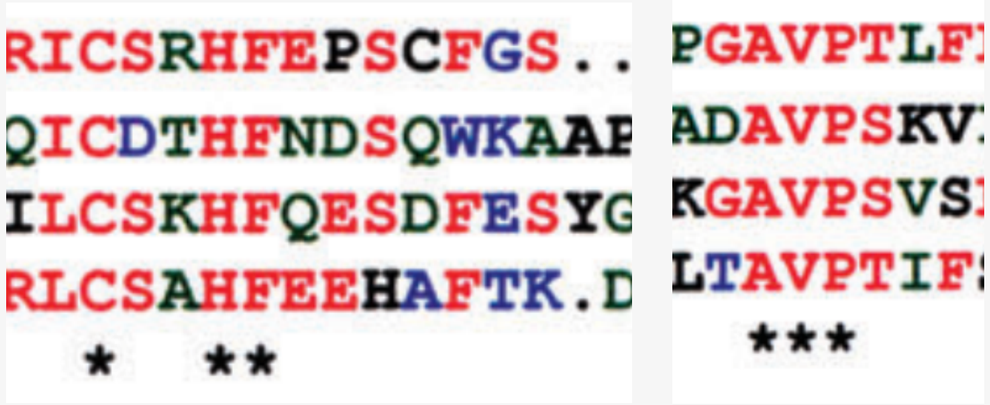
Les protéines THAP?

Famille des protéines THAP: différents paralogues chez chaque espèce (THAP 1 à THAP 11)
En commun: domaine THAP situé en **position N-terminale**
Très **bonne conservation** entre orthologues:
(exemple : THAP 1 chez l'homme, souris, zebrafish)
Domaine de fixation à l'ADN

Enables sequence-specific DNA binding activity and transposase activity. Involved in DNA integration and transposition, DNA-mediated. [provided by Alliance of Genome Resources, Apr 2022]

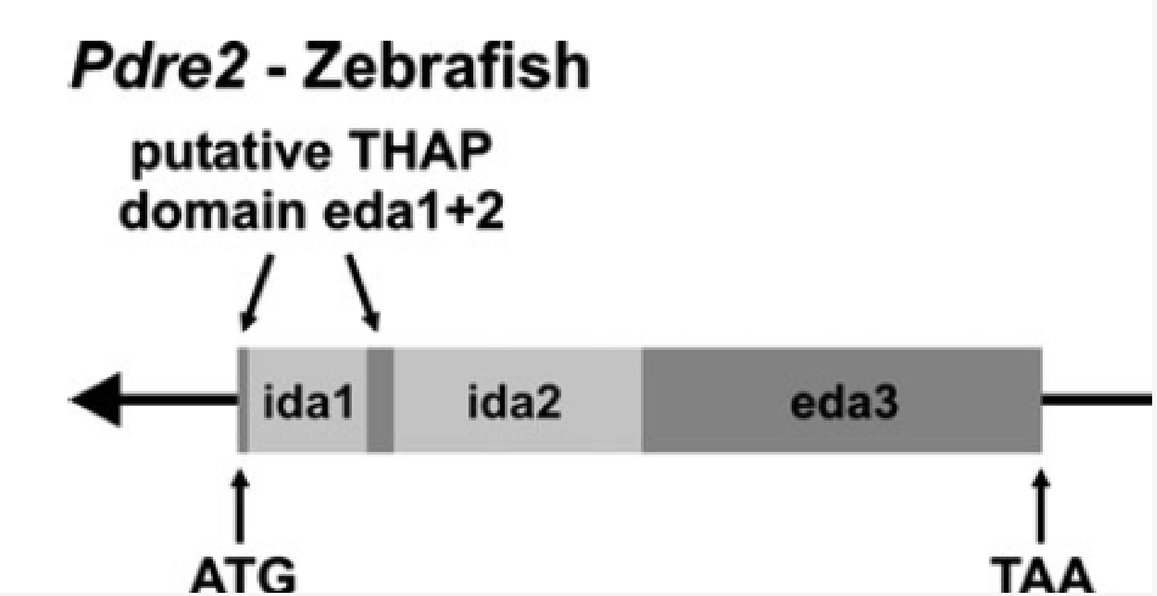
Domaine THAP

Extrait de la publication



Les objectifs du projet

Confirmation/vérification des résultats publiés



1. **Positions de l'orthologue de THAP9 chez le Zebrafish**

2. **Présence du motif caractéristique des protéines THAP**

Recherche d’une séquence THAP9 chez le Zebrafish

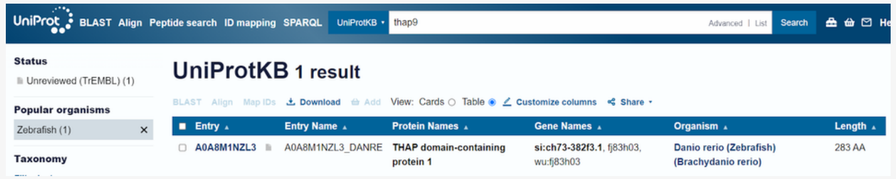
Sur NCBI

((thap[Title] AND (zebrafish[Organism])))

Name/Gene ID	Description	Location	Aliases
thap11 ID: 406390	THAP domain containing 11 [*Danio rerio* (zebrafish)]	Chromosome 7, NC_007118.7 (29041470..29042887, complement)	wu:fa14e04, wu:fc33b05, zgc:65871, zgc:76876
e2f6 ID: 560495	E2F transcription factor 6 [*Danio rerio* (zebrafish)]	Chromosome 20, NC_007131.7 (29676870..29683806, complement)	fb34f06, si:ch211-195d17.2, wu:fb34f06, zgc:136607
thap4 ID: 567037	THAP domain containing 4 [*Danio rerio* (zebrafish)]	Chromosome 2, NC_007113.7 (22653123..22659453, complement)	nitrobindin, si:ch211-105d4.5
thap12b ID: 436953	THAP domain containing 12b [*Danio rerio* (zebrafish)]	Chromosome 15, NC_007126.7 (19900304..19908909)	prkrir, prkrirb, zgc:86870
thap12a ID: 406354	THAP domain containing 12a [*Danio rerio* (zebrafish)]	Chromosome 10, NC_007121.7 (32058817..32064405)	fb55g11, prkrir, prkrira, prkrirl, wu:fb55g11, zgc:55697
thap7 ID: 100006965	THAP domain containing 7 [*Danio rerio* (zebrafish)]	Chromosome 19, NC_007130.7 (24129414..24136233, complement)	im:7136959, si:ch211-282e16.1
thap3 ID: 541481	THAP domain containing, apoptosis associated protein 3 [*Danio rerio* (zebrafish)]	Chromosome 23, NC_007134.7 (30726006..30727595, complement)	zgc:110849
thap1 ID: 692315	THAP domain containing, apoptosis associated protein 1 [*Danio rerio* (zebrafish)]	Chromosome 5, NC_007116.7 (431958..441310)	zgc:136597
LOC100535121 ID: 100535121	THAP domain-containing protein 6-like [*Danio rerio* (zebrafish)]	Chromosome 15, NC_007126.7 (44106184..44116954, complement)	
LOC110438581 ID: 110438581	THAP domain-containing protein 11-like [*Danio rerio* (zebrafish)]		

Avec ou sans RefSeq: **Pas de THAP9**

SUR UNIPROT



Pas de THAP9 mais une THAP1

Récupération des séquences sur Uniprot

BX511023 et **Q9H5L6** au format FASTA

P-homologous **zebrafish** sequence (Pdre2) from GenBank

Séquence protéique THAP9 **humaine**

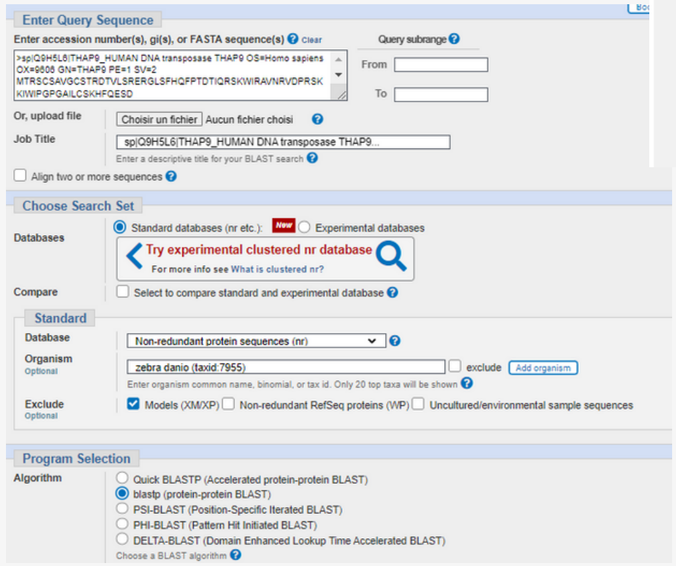
Alignement Séquence THAP Humaine contre Organisme Zebrafish

BX511023 et Q9H5L6

P-homologous zebrafish sequence

Séquence protéique THAP9 humaine

BlastP



Contre Swissprot

select all 2 sequences selected GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
RecName: Full=THAP domain-containing protein 1 [Danio rerio]	Danio rerio	50.4	50.4	9%	2e-07	33.71%	225	Q1JPT7.2
RecName: Full=THAP domain-containing protein 11 [Danio rerio]	Danio rerio	36.6	36.6	8%	0.006	30.67%	257	Q6TGZ4.2

Contre NR

select all 20 sequences selected GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
uncharacterized protein LOC664763 [Danio rerio]	Danio rerio	71.2	71.2	20%	2e-13	25.73%	201	NP_001231992.2
LOC553397 protein [Danio rerio]	Danio rerio	71.2	71.2	21%	3e-13	25.35%	216	AAH93414.1
LOC553397 protein [Danio rerio]	Danio rerio	70.9	70.9	20%	5e-13	25.73%	225	AAI21761.1
THAP domain-containing protein 7 [Danio rerio]	Danio rerio	71.2	71.2	9%	4e-12	42.39%	627	NP_001096668.1
Si:ch211-282e16.1 protein [Danio rerio]	Danio rerio	71.2	71.2	9%	5e-12	42.39%	641	AAH92176.1
LOC100005466 protein [Danio rerio]	Danio rerio	65.9	65.9	9%	2e-10	35.71%	619	AAI39695.1
uncharacterized protein LOC100003014 [Danio rerio]	Danio rerio	56.6	56.6	11%	1e-07	32.17%	415	NP_001070137.1
uncharacterized protein LOC100151273 [Danio rerio]	Danio rerio	55.5	55.5	14%	1e-07	30.08%	283	NP_001243612.1
Si:bY184L24.4 (novel protein) [Danio rerio]	Danio rerio	53.1	53.1	9%	2e-07	35.96%	178	CAD58737.1
ARL14 effector protein [Danio rerio]	Danio rerio	53.5	53.5	9%	4e-07	33.71%	224	NP_001074222.1
Zgc:153292 protein [Danio rerio]	Danio rerio	54.3	54.3	11%	7e-07	31.30%	415	AAI71353.1
Zgc:153292 protein [Danio rerio]	Danio rerio	54.3	54.3	11%	7e-07	31.78%	431	AAI55218.1
Zgc:136597 [Danio rerio]	Danio rerio	50.8	50.8	9%	1e-06	33.71%	158	AAI16603.1
THAP domain-containing protein [Danio rerio]	Danio rerio	51.2	51.2	9%	3e-06	36.56%	292	NP_001373405.1
THAP domain-containing protein 1 [Danio rerio]	Danio rerio	50.4	50.4	9%	4e-06	33.71%	225	NP_001410789.1
uncharacterized protein LOC795946 [Danio rerio]	Danio rerio	48.1	48.1	9%	6e-05	33.71%	413	NP_001077038.1
L(3)mbt-like 2 (Drosophila) [Danio rerio]	Danio rerio	46.6	46.6	12%	2e-04	29.41%	805	AAI71358.1
lethal(3)malignant brain tumor-like protein 2 [Danio rerio]	Danio rerio	46.2	46.2	12%	3e-04	29.41%	805	NP_956326.1
THAP domain containing, apoptosis associated protein 3 [Danio rerio]	Danio rerio	43.1	43.1	19%	9e-04	27.23%	213	AAH91927.1
THAP domain-containing protein 3 [Danio rerio]	Danio rerio	43.1	43.1	11%	0.001	28.04%	213	NP_001014316.2

BX511023 absente

Pas de THAP9 dans les 2 Blast

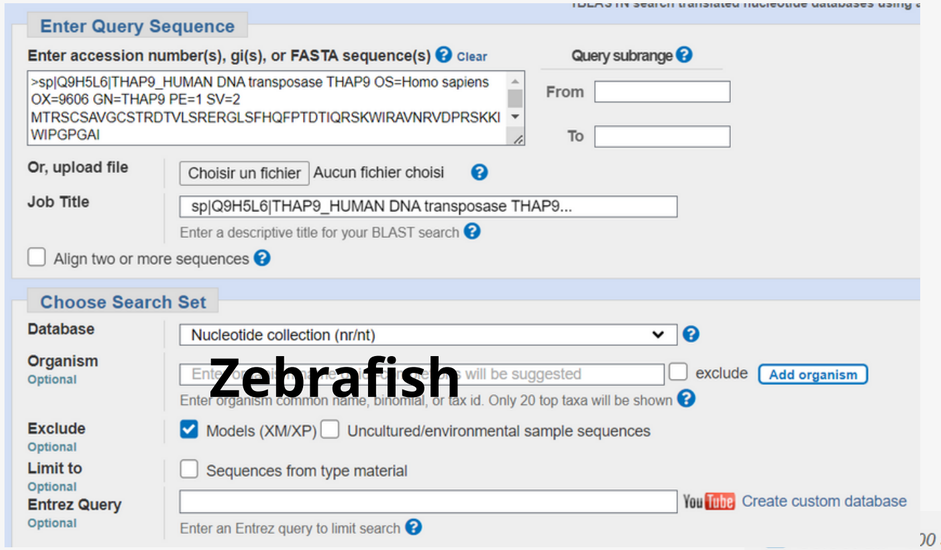
Alignement Séquence THAP Humaine contre Organisme Zebrafish

BX511023 et Q9H5L6

P-homologous zebrafish sequence

Séquence protéique THAP9 humaine

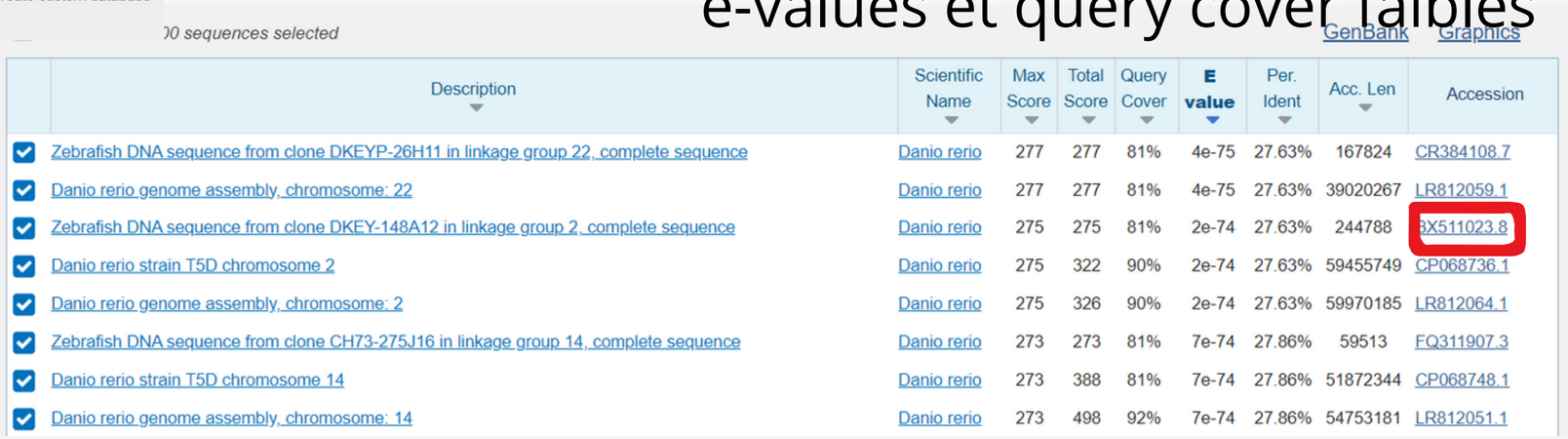
tBlastN



Zebrafish

e-values et query cover faibles

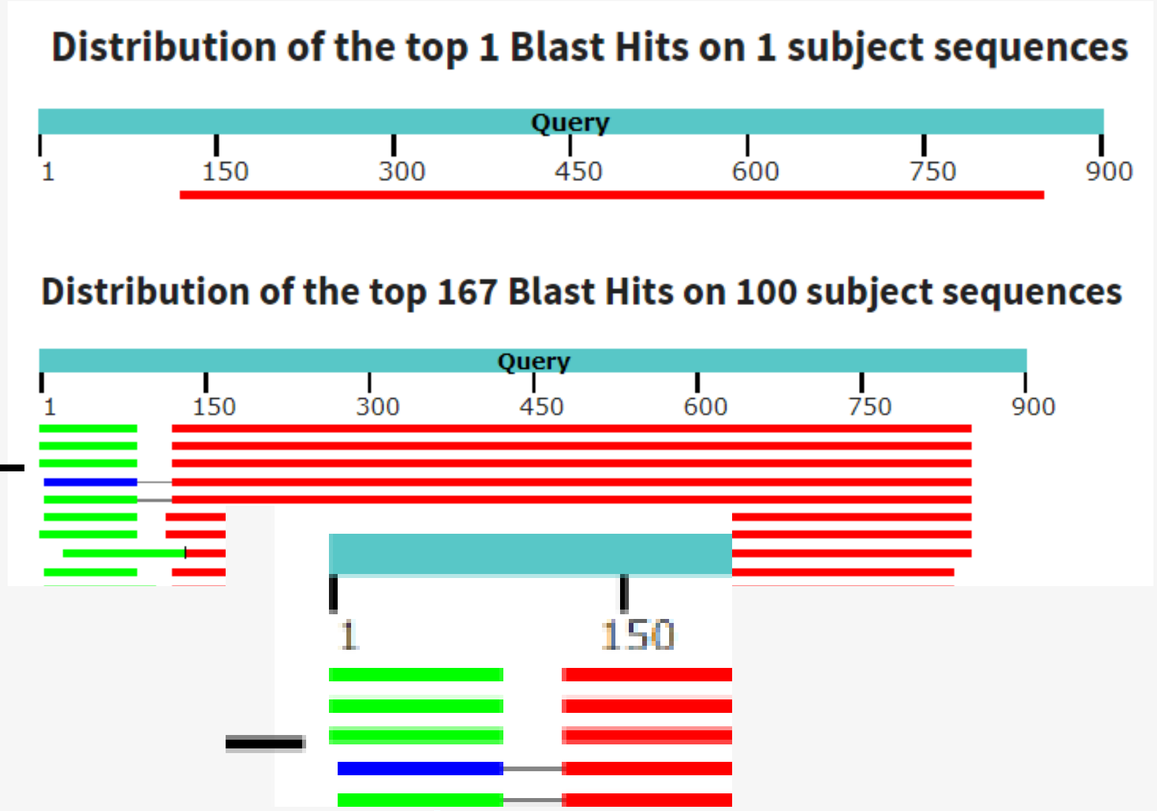
Séquence BX511023



	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Zebrafish DNA sequence from clone DKEYP-26H11 in linkage group 22, complete sequence	Danio rerio	277	277	81%	4e-75	27.63%	167824	CR384108.7
☑	Danio rerio genome assembly, chromosome: 22	Danio rerio	277	277	81%	4e-75	27.63%	39020267	LR812059.1
☑	Zebrafish DNA sequence from clone DKEY-148A12 in linkage group 2, complete sequence	Danio rerio	275	275	81%	2e-74	27.63%	244788	BX511023.8
☑	Danio rerio strain T5D chromosome 2	Danio rerio	275	322	90%	2e-74	27.63%	59455749	CP068736.1
☑	Danio rerio genome assembly, chromosome: 2	Danio rerio	275	326	90%	2e-74	27.63%	59970185	LR812064.1
☑	Zebrafish DNA sequence from clone CH73-275J16 in linkage group 14, complete sequence	Danio rerio	273	273	81%	7e-74	27.86%	59513	FQ311907.3
☑	Danio rerio strain T5D chromosome 14	Danio rerio	273	388	81%	7e-74	27.86%	51872344	CP068748.1
☑	Danio rerio genome assembly, chromosome: 14	Danio rerio	273	498	92%	7e-74	27.86%	54753181	LR812051.1

Alignement Séquence THAP Humaine contre Organisme Zebrafish

tBlastN



Sequence BX511023

Partie N-term manquante

Zebrafish DNA sequence from clone DKEY-148A12 in linkage group 2, complete sequence

Sequence ID: BX511023.8 Length: 244788 Number of Matches: 1

Range 1: 202236 to 204425 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
275 bits(704)	2e-74	Compositional matrix adjust.	214/760(28%)	372/760(48%)	58/760(7%)	+3

```
Query  122     EDHNYSLKTPLTIGAEKLAEVQQMLQVSKKRLI--SVKNYRMIKKRKGLRLIDALVEEKL  179
               EDH+Y+L +P ++  +  A   Q+LQ  +K+L   S K+ R+  K   L+ +   +++KL
Sbjct  202236  EDHSYNLASPRSLKRKNQA-TDQILQKYRKKLKIESQKSRRLKNKISTLKDVVTELKKKL  202412

Query  180     L-SEETECLLRAQFSDFK---WELYNWRETDEYSAEMKQFACTLYLCSSKVYDYVRKILK  235
               L S E   LL +     K    ++   ++T +YS E+KQFA TL+  S K YDYVR   +
Sbjct  202413  LISNECASLLESIDEVPKHILLKIQQGKKTAKYSEELKQFATTLHFYSPKAYDYVRDNFQ  202592

Query  236     --LPHSSILRTWLSKCQPSPGFNSNIFSFLQRRVENGDQLYQ--YCSLLIKSMPLKQQLQ  291
                 LPHS  +R W S      PGF    F+ L+  VE   +  +   C+L++  M + +  +
Sbjct  202593  KALPHSHTIRNWYSSVSADPGFTVASFTALKCHVEENKERGKETVCALMMDEMYIHKMTE  202772

Query  292     WDPSSHSLQGFMDFGLGKLDADETPLASETVLLMAVGIFGHWRTPLGYFFVNRASGYLQA  351
               +  +     G++D G G++D      LA++ ++LM V +   W+ P+ YF +    G  +A
Sbjct  202773  F--AGDQFHGYVDIGTGEID---NTLATQALVLMVVAVNESWKIPIAYFLITGMDGSEKA  202937

Query  352     QLLRLTIGKLSDIGITVLAVTSDATAHSVQMAKALGIHIDGDDMKCTFQHPSSSSQQIAY  411
                ++R ++ +L ++G+ V+++T DA    ++ M + LG  ++ ++MK  F HP   +Q++
Sbjct  202938  NVIRESLSRLHEVGVKVISLTCDAPTTNLAMIRELGADLNINNMKPYFMHPEDPTQKVHV  203117

Query  412     FFDSCHLLRLIRNAFQNFQSIQFINGIA-HWQHlvelvaleeqelSNM-ERIPSTLANLK  469
                  D+CH+L+L+RNAF +    +  +G    W+++  L  L+E+E   +  ++       +
Sbjct  203118  ILDACHMLKLLRNAFASSLEFETEDGNKIKWKYIEALNELQEKEGLRLGNKLKMAHLQWR  203297

Query  470     NHVLKVNSATQLFSESVASALEYL-LSLDLPPFQNCIGTIHFLRLINNLFDIFNSRNCYG  528
                  +KV+ A QLFS SVA A+E+    L +  F+ C  T+ FLR ++  FD+ NSRN  G
Sbjct  203298  KQKMKVHLAAQLFSSSVADAIEFCEQGLKMEEFKGCAATVQFLRTVDAAFDVLNSRNPLG  203477

Query  529     KGLKGPLLPETYSKINHVLIEAKTIFVTLSDTSNNQIIkgkqklgflgf---llNAESLK  585
               KG K P+   T  ++  +L +A+++   L    N+++              + N  S
Sbjct  203478  KGFKAPIKTTTKDRVETILKQAESMLRGLKVQQYNKMVPLHTTKKKTAIYGFIANGRSAL  203657

Query  586     WLYQNYVFPKVMPFPYLLTYKFSHDHLELFLKMLRQVLVTSSSPTCMAFQKAYYNLETRY  645
                +Y + V     P  YLLTYK S DHLELF   +R     + +P    F+ AY  L  R+
Sbjct  203658  NIYHDLVERPNAPCRYLLTYKLSQDHLELFFAAIRARGGHNDNPNARQFRGAYKRLLVRH  203837

Query  646     KFQ---------DEVFLSKVSIFDISIARRKDLALWTVQRQYGVSVTKTVFHEEGICQDW  696
               + +         D  ++   +++ARR ++ L  V             E     D
Sbjct  203838  QVKTGTGNCLLLDNTYMLNSTPASVNVARRLEVQLVEVD----------VPENDAVPDL  203984

Query  697     SH-CSLSEALLDLSDHRRNLICY-AGYVANKLSALLTCEDCITALYASDLKASKIGSLLF  754
                H CSLSE       ++   I Y  G+V  K+   +TC  C  AL  +D  A +    +
Sbjct  203985  PHVCSLSE-------YKEAAIHYITGFVVKKMKEKITCLPCSQAL-TTDGAAHE---FIH  204131

Query  755     VKKKNGLHFPSESLCRVINICERVVRTHSRMAIFELV-SKQRELYLQQKILCELSGHINL  813
               +K + GL  PS  +  V    E+ ++     +  +L   +   L  +L  ++L   +  +L
Sbjct  204132  LKNRGGLQKPSPGMVSVCLTTEKCIQRKITTSGGQLPRGRGITLAISNEVLANCAER-DL  204308

Query  814     FVDVNKHLFDGEVCAINHFVKLLKDIIICFLNIRAKNVAQ  853
               F  ++ H+F   V  +NH   L+K   I +  +R  + A+
Sbjct  204309  FPQLHSHMFATSV-EMNHIHLLVKMASIWYSKVRFNHFAR  204425
```

Conclusion sur Blast

BLASTP	tBlastN
Séquences de protéines THAP1 et THAP11 Mauvais % d'identité Mauvais query cover	On retrouve **BX511023** Query cover plutôt bon 48% de similarité **Partie N-terminale absente** N-term: motif des protéines THAP

Recherche du domaine THAP sur la séquence BX511023

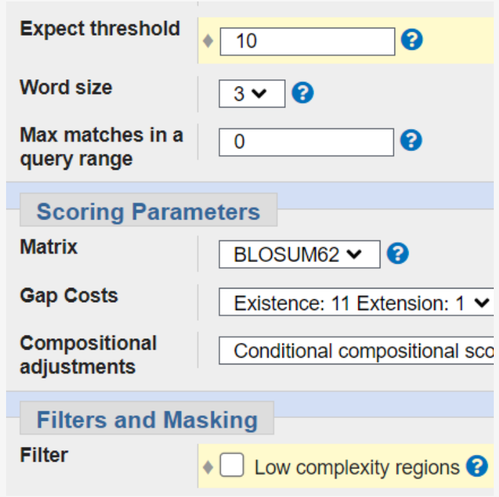
BX511023 et Q9H5L6

Séquence chez le Zebrafish

Séquence protéique THAP9 humaine

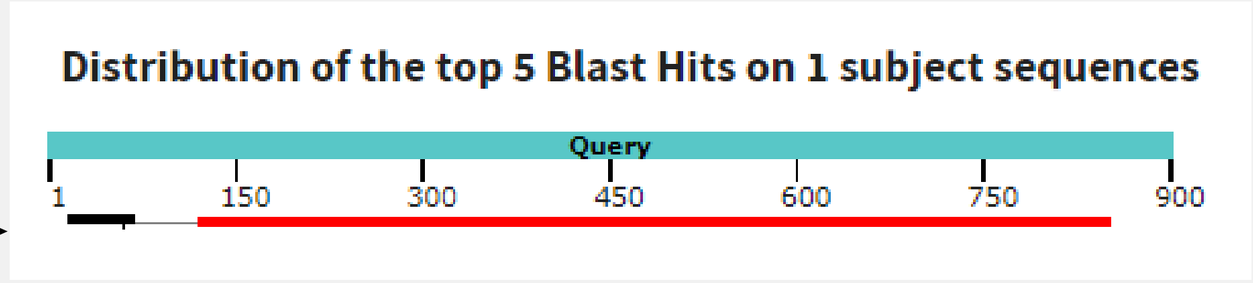
tblastN

Paramètres modifiés :
expect threshold et word size



	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	BX511023.8 Zebrafish DNA sequence from clone DKEY-148A12 in linkage group 2, complete se…		298	369	87%	1e-86	28.82%	244788	Query_404113

l sequences selected



1 fragment supplémentaire en partie N-term

Range 1: 202236 to 204425 Graphics ▼ Next Match ▲ Previous

Score	Expect	Method	Identities	Positives	Gaps	Frame
298 bits(762)	1e-86	Compositional matrix adjust.	219/760(29%)	377/760(49%)	58/760(7%)	+3

```
Query  122     EDHNYSLKTPLTIGAEKLAEVQQMLQVSKKRLI--SVKNYRMIKKRKGLRLIDALVEEKL  179
Sbjct  202236
Query  180
Sbjct  202413
Query  236
Sbjct  202593
Query  292
Sbjct  202773
Query  352
Sbjct  202938  NVIRESLSRLHEVGVKVISLTCDAPTTNLAMIRELGADLNINNMKPYFMHPEDPTQKVHV  203117
Query  412     FFDSCHLLRLIRNAFQNFQSIQFINGIA-HWQHLVELVALEEQE-LSNMERIPSTLANLK  469
               D+CH+L+L+RNAF +    +  +G    W+++  L  L+E+E L    ++       +
Sbjct  203118  ILDACHMLKLLRNAFASSLEFETEDGNKIKWKYIEALNELQEKEGLRLGNKLKMAHLQWR  203297
Query  470     NHVLKVNSATQLFSESVASALEYL-LSLDLPPFQNCIGTIHFLRLINNLFDIFNSRNCYG  528
                 +KV+ A QLFS SVA A+E+    L +  F+ C  T+ FLR ++  FD+ NSRN  G
Sbjct  203298  KQKMKVHLAAQLFSSSVADAIEFCEQGLKMEEFKGCAATVQFLRTVDAAFDVLNSRNPLG  203477
Query  529     KGLKGPLLPETYSKINHVLIEAKTIFVTLSDTSNNQII---KGKQKLGFLGFLLNAESLK  585
               KG K P+    T   ++  +L +A+++   L     N+++     K+K    GF+ N  S
Sbjct  203478  KGFKAPIKTTTKDRVETILKQAESMLRGLKVQQYNKMVPLHTTKKKTAIYGFIANGRSAL  203657
Query  586     WLYQNYVFPKVMPFPYLLTYKFSHDHLELFLKMLRQVLVTSSSPTCMAFQKAYYNLETRY  645
                +Y + V     P  YLLTYK S DHLELF   +R     + +P    F+ AY  L  R+
Sbjct  203658  NIYHDLVERPNAPCRYLLTYKLSQDHLELFFAAIRARGGHNDNPNARQFRGAYKRLLVRH  203837
Query  646     KFQ---------DEVFLSKVSIFDISIARRKDLALWTVQRQYGVSVTKTVFHEEGICQDW  696
               + +         D  ++   +   +++ARR ++ L  V                E    D
Sbjct  203838  QVKTGTGNCLLLDNTYMLNSTPASVNVARRLEVQLVEVD-----------VPENDAVPDL  203984
Query  697     SH-CSLSEALLDLSDHRRNLICY-AGYVANKLSALLTCEDCITALYASDLKASKIGSLLF  754
                H CSLSE       ++   I Y  G+V  K+   +TC  C  AL  +D  A +    +
Sbjct  203985  PHVCSLSE-------YKEAAIHYITGFVVKKMKEKITCLPCSQAL-TTDGAAHE---FIH  204131
Query  755     VKKKNGLHFPSESLCRVINICERVVRTHSRMAIFELV-SKQRELYLQQKILCELSGHINL  813
               +K + GL  PS  +  V    E+ ++     +  +L   +   L +  ++L   +   +L
Sbjct  204132  LKNRGGLQKPSPGMVSVCLTTEKCIQRKITTSGGQLPRGRGITLAISNEVLANCAER-DL  204308
Query  814     FVDVNKHLFDGEVCAINHFVKLLKDIIICFLNIRAKNVAQ  853
               F  ++ H+F   V  +NH   L+K   I +  +R  + A+
Sbjct  204309  FPQLHSHMFATSV-EMNHIHLLVKMASIWYSKVRFNHFAR  204425
```

Range 2: 200145 to 200201 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
24.3 bits(51)	4.1	Compositional matrix adjust.	9/19(47%)	11/19(57%)	0/19(0%)	+3

```
Query  53      WIPGPGAILCSKHFQESDF  71
               W P    + LCS HF+E  F
Sbjct  200145  WKPTTSSRLCSAHFEEHAF  200201
```

Match ▲ Previous Match

Gaps	Frame
%) 0/19(0%)	+3

Range 3: 152773 to 152904 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
23.5 bits(49)	5.9	Compositional matrix adjust.	11/44(25%)	23/44(52%)	0/44(0%)	-1

```
Query  18      SRERGLSFHQFPTDTIQRSKWIRAVNRVDPRSKKIWIPGPGAIL  61
               ++++ +  H     TIQ ++ +  +N +DPRS  +      A+L
Sbjct  152904  TKQK*IIVHNSKNKTIQ*NQVLHILNNIDPRSVNVACVDTTALL  152773
```

Range 4: 101879 to 101941 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
23.5 bits(49)	6.4	Compositional matrix adjust.	8/21(38%)	13/21(61%)	0/21(0%)	-3

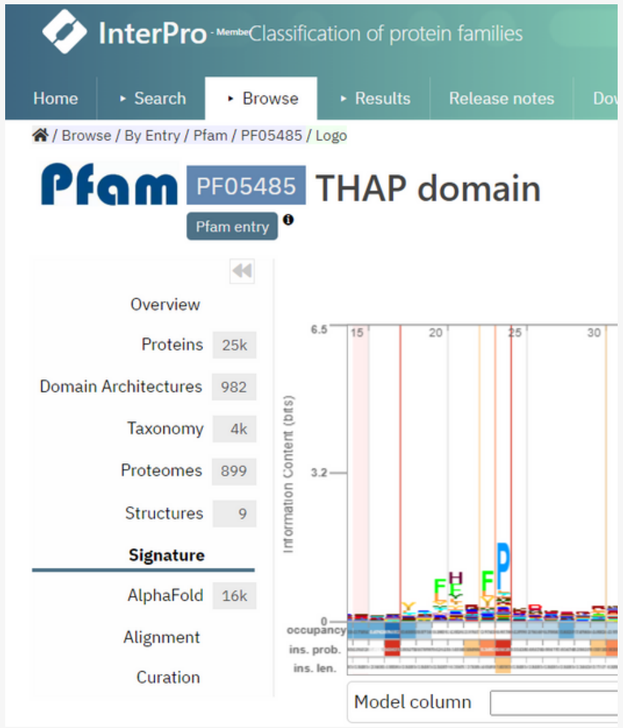
```
Query  395     MKCTFQHPSSSSQQIAYFFDS  415
               ++ T +HP S  QQ+ YF  +
Sbjct  101941  IEKTIRHPESIKQQVGYFVGA  101879
```

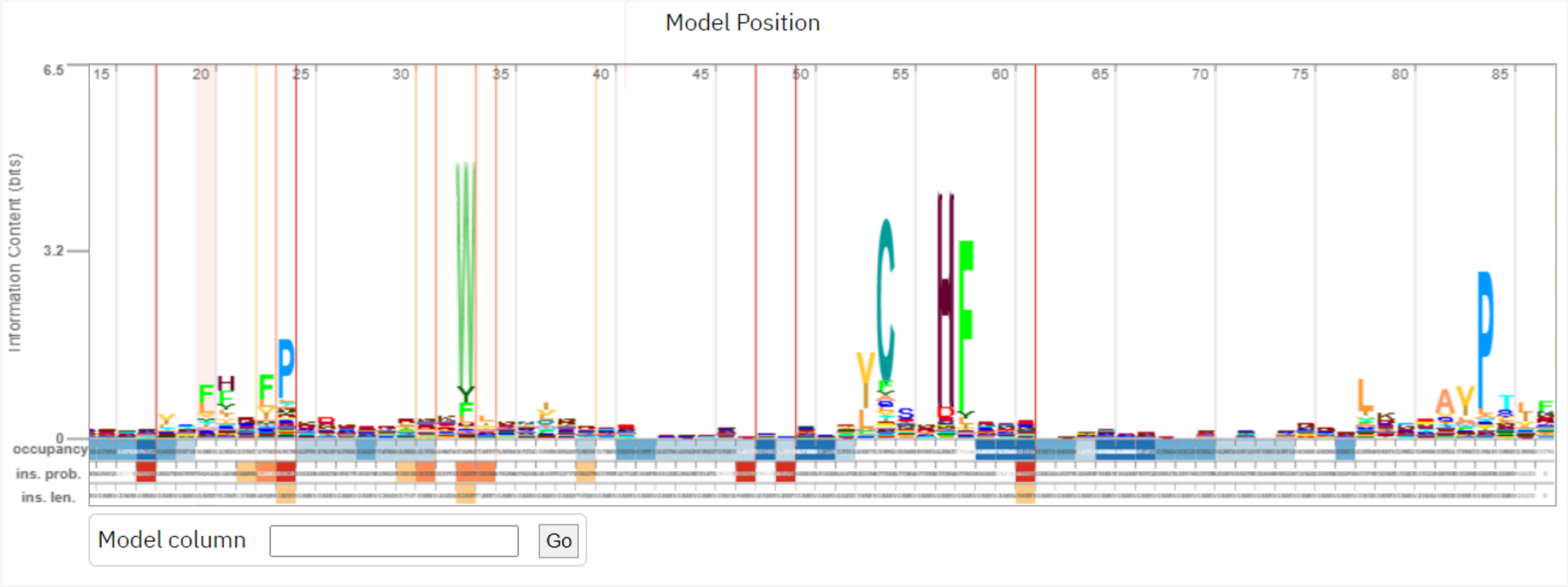
Segment AVP absent sûrement plus loin

Recherche du logo du domaine THAP

Pour connaître le motif à chercher

Sur Interpro: Logo du domaine THAP



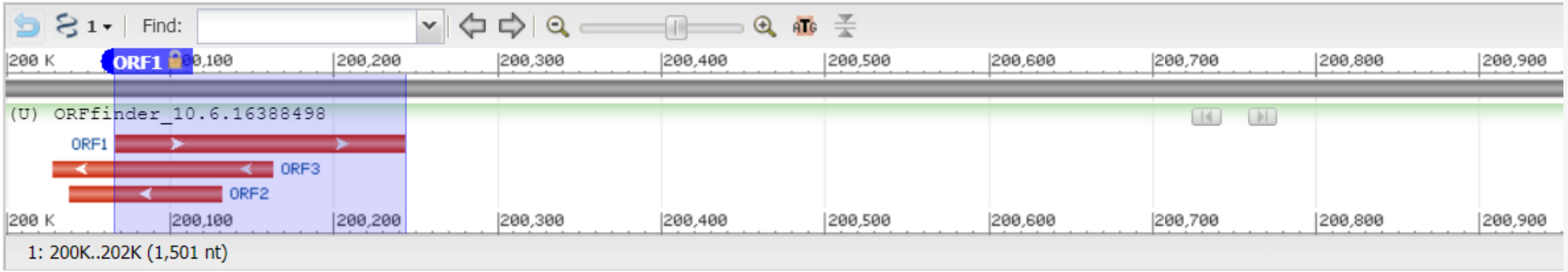


Recherche d’ORF avec ORF Finder

Open Reading Frame Viewer

Sequence

ORFs found: 3 **Genetic code: 1** **Start codon: 'ATG' only**
ORFs were calculated on the interval from 200000 to 200500 nt



ORF1 (58 aa) Display ORF as... **Mark**

```
>lcl|ORF1
MFCSFPLGRHDAARLKQWVVNMRRNGWKPTTSSRLCSAHFEEHAFTKDLK
VKKNLQNN
```

Mark subset... Marked: 0 Download marked set as Protein FASTA

Label	Strand	Frame	Start	Stop	Length (nt | aa)
ORF1	+	3	200067	200243	177 | 58
ORF3	-	1	200163	200029	135 | 44
ORF2	-	3	200131	200039	93 | 30

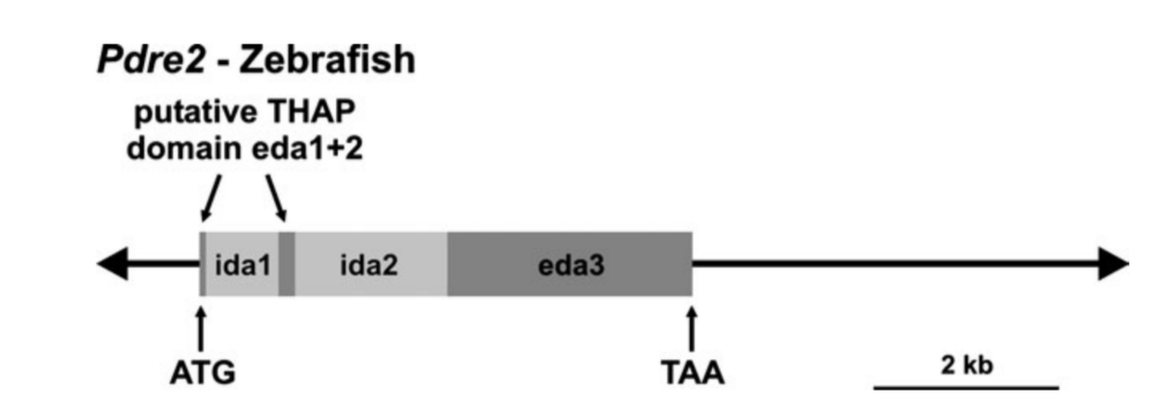
Une partie du motif THAP visible

La partie N-terminale du domaine THAP est manquante dans la séquence BX511023

Recherche du domaine THAP sur la séquence BX511023

Domaine haché sur plusieurs exons

Impossible de le récuperer en entier



Positions Recherchées à partir de l'article	Positions Trouvées
199271 - 199302	Aucune correspondance
200078 - 200230	200145 - 200201
201890 - 204500	202236 - 204425

Reconstitution de la séquence protéique utilisation des positions de la publications

FASTA ▾ Send to: ▾

Zebrafish DNA sequence from clone DKEY-148A12 in linkage group 2, complete sequence

GenBank: BX511023.8

GenBank Graphics

Change region shown

- ○ Whole sequence
- ◉ Selected region

from: 199000 to: 203000

Update View

199271 - 199302

>BX511023.8:199271-199302 Zebrafish DNA sequence from clone DKEY-148A12 in linkage group 2, complete sequence

ATGAGTTGTTCAGCAGTAAACTGTTCTAACCG

200078 - 200230

>BX511023.8:200078-200230 Zebrafish DNA sequence from clone DKEY-148A12 in linkage group 2, complete sequence

TTTTCCTTTAGGTAGGCATGATGCAGCAAGATTGAAACAATGGGTGGTGAACATGAGACGCAACGGCTGG
AAACCAACAACTTCATCCCGTCTTTGCAGTGCTCATTTTGAAGAGCATGCTTTCACCAAGGATTTGAAGG
TAAAAAAAAATCT

201890 - 204500

>BX511023.8:201890-204500 Zebrafish DNA sequence from clone DKEY-148A12 in linkage group 2, complete sequence

TACAGCAGTTCCAACTATATTCTCATTTCCTTGTCATCTGGTGAAATCCTCTTGTGTAGGGGTAAGACAC
AAGACTTTTGCGAATAAGAATTCTGAGGTAAGTGGCTCCTCTCTGGTTGTTCAGCAAAGATGCAGGACA
GTAAAAAGGATTCTCCAGTTGACAAGTCTCTTAGCGTGCACTCAGCTGCTGATTACAGCAATGCCACCTT
GACTGAACATGCTATTGGCAACTTCTCAGTTAGTAATGACCTATGTCAAGTTGACACTGCTGTAACGGTG
GATATTAATGAGGATCATAGTTATGCAACTTTAAGTTCCACACATGAAACTCCTAAGTTTATTAAAGAGG
ATCACAGCTACAACCTTGCGTCTCCGAGATCTTTAAAACGGAAAAATCAGGCCACTGACCAAATTCTACA
AAAATACAGGAAAAAGCTTAAAATTGAATCCCAAAAGTCTAGGCGTTTAAAAAACAAAATCTCCACCTTA
AAGGATGTCGTCACAGAACTCAAGAAGAAGCTGTTGATATCCAATGAATGTGCTTCTTTGCTTGAGTCAA
TTGATGAAGTGCCAAAGCACATCTTGTTGAAAATACAGCAGGGGAAAAAAACAGCAAAGTACTCAGAGGA
GTTGAAGCAGTTTGCCACAACCTTGCATTTTTATTCCCCAAAGGCATACGACTACGTACGTGACAATTTT
CAGAAAGCGCTGCCACACAGTCACACCATTCGAAATTGGTACAGCAGCGTATCAGCTGATCCTGGATTTA
CTGTGGCATCTTTCACAGCGCTGAAATGTCATGTAGAGGAGAATAAAGAGAGAGGAAAAGAAACAGTATG
TGCCCTAATGATGGATGAAATGTATATTCACAAGATGACTGAATTTGCAGGGGACCAGTTTCATGGCTAT
GTGGACATTGGTACTGGAGAAATTGACAACACGTTGGCTACCCAGGCTTTGGTCCTCATGGTTGTGGCTG
TAAATGAGTCATGGAAAATTCCCATAGCATACTTTCTCATCACAGGTATGGATGGATCAGAGAAGGCCAA
TGTCATTAGGGAGAGCCTTAGCAGACTTCATGAGGTTGGTGTCAAAGTCATCTCTCTTACATGTGATGCC
CCCACGACCAACTTGGCTATGATAAGGGAACTTGGCGCTGACCTAAATATTAACAACATGAAGCCTTACT
TCATGCATCCTGAGGACCCTACTCAGAAAGTACATGTGATCCTGGATGCATGCCACATGCTAAAGCTTCT
TCGCAATGCCTTTGCATCTTCCCTTGAATTTGAGACAGAAGATGGAAATAAGATTAAGTGGAAGTACATC
GAAGCCCTCAATGAACTCCAAGAAAAAGAAGGTCTGCGACTGGGAAACAAGCTGAAGATGGCTCATTTAC
AATGGCGAAAACAAAAAATGAAAGTCCATCTGGCAGCACAGCTCTTCAGCAGCAGTGTGGCAGATGCTAT
AGAGTTCTGTGAGCAAGGCTTGAAAATGGAGGAATTCAAAGGTTGTGCTGCAACTGTGCAGTTTTTGCGC
ACTGTAGATGCAGCTTTTGATGTACTGAATAGCAGAAATCCTCTTGGAAAAGGTTTCAAAGCCCCAATCA
AAACAACAACTAAAGATCGGGTTGAGACCATTTTAAAGCAGGCAGAATCCATGCTTCGTGGGTTGAAGGT
CCAGCAGTACAACAAGATGGTGCCACTTCACACAACCAAAAAGAAGACTGCCATATATGGCTTCATTGCT
AATGGCAGAAGTGCACTTAACATCTATCATGATCTTGTTGAAAGGCCCAATGCACCATGCAGATATCTTC
TTACCTACAAATTAAGTCAGGACCACTTGGAGCTTTTCTTTGCTGCCATAAGAGCAAGAGGTGGTCACAA
TGATAATCCTAATGCCAGGCAGTTTAGAGGAGCATACAAACGCCTACTTGTAAGGCACCAAGTGAAAACA
GGCACAGGGAACTGTTTGCTTCTAGACAACACTTACATGCTCAACTCAACACCTGCATCTGTAAATGTTG
CCCGAAGATTGGAAGTACAGTTGGTGGAAGTGGACGTGCCCGAGAATGATGCTGTTCCTGACCTCCCTCA
TGTATGCAGTCTTTCTGAATACAAAGAGGCAGCAATTCATTATATCACAGGATTTGTGGTGAAGAAGATG
AAAGAGAAAATCACCTGCTTGCCATGCTCACAGGCATTAACAACTGATGGTGCTGCACATGAATTCATCC
ATCTAAAAAACAGAGGTGGTCTGCAGAAACCATCACCAGGCATGGTGTCTGTGTCTGACCACAGAGAA
ATGCATTCAGAGAAAGATCACAACAAGTGGAGGTCAGCTTCCCCGTGGACGTGGAATCACACTTGCTATT
TCCAACGAGGTGTTAGCAAATTGTGCAGAGAGAGATCTTTTTCCACAACTCCACAGCCACATGTTTGCTA
CAAGCGTTGAAATGAATCATATTCATTTATTGGTTAAAATGGCATCCATTTGGTATAGCAAAGTCAGATT
TAACCACTTTGCAAGAAGGGGAGCTGAAATAGCAAAAGATGGCAAAATGGTTCGCAGACAACTGACTAAA
TTGATTCATTTTTATGGTGAA

Séquence assemblée

ATGAGTTGTTCAGCAGTAAACTGTTCTAACCG
GTGGTGAACATGAGACGCAACGGCTGGAAACC
GCTTTCACCAAGGATTTGAAGGTAAAAAAAAATCTTACAGCAGTTCCAACTATATTCTCATTTCCTTGTCATCTG
GTGAAATCCTCTTGTGTAGGGGTAAGACACAAGACTTTTGCGAATAAGAATTCTGAGGTAAGTGGCTCCTCTCTG
TTGTTCAGCAAAGATGCAGGACACAGTAAAAAGGATTCTCCAGTTGACAAGTCTCTTAGCGTGCACTCAGCTGCT
GATTACAGCAATGCCACCTTGACTGAACATGCTATTGGCAACTTCTCAGTTAGTAATGACCTATGTCAAGTTGAC
ACTGCTGTAACGGTGGATATTAATGAGGATCATAGTTATGCAACTTTAAGTTCCACACATGAAACTCCTAAGTTT
ATTAAAGAGGATCACAGCTACAACCTTGCGTCTCCGAGATCTTTAAAACGGAAAAATCAGGCCACTGACCAAATT
CTACAAAAATACAGGAAAAAGCTTAAAATTGAATCCCAAAAGTCTAGGCGTTTAAAAAACAAAATCTCCACCTTA
AAGGATGTCGTCACAGAACTCAAGAAGAAGCTGTTGATATCCAATGAATGTGCTTCTTTGCTTGAGTCAATTGAT
GAAGTGCCAAAGCACATCTTGTTGAAAATACAGCAGGGGAAAAAAACAGCAAAGTACTCAGAGGAGTTGAAGCAG
TTTGCCACAACCTTGCATTTTTATTCCCCAAAGGCATACGACTACGTACGTGACAATTTTCAGAAAGCGCTGCCA
CACAGTCACACCATTCGAAATTGGTACAGCAGCGTATCAGCTGATCCTGGATTTACTGTGGCATCTTTCACAGCG
CTGAAATGTCATGTAGAGGAGAATAAAGAGAGAGGAAAAGAAACAGTATGTGCCCTAATGATGGATGAAATGTAT
ATTCACAAGATGACTGAATTTGCAGGGGACCAGTTTCATGGCTATGTGGACATTGGTACTGGAGAAATTGACAAC
ACGTTGGCTACCCAGGCTTTGGTCCTCATGGTTGTGGCTGTAAATGAGTCATGGAAAATTCCCATAGCATACTTC
TCATCACAGGTATGGATGGATCAGAGAAGGCCAATGTCATTAGGGAGAGCCTTAGCAGACTTCATGAGGTTGGTG
TCAAAGTCATCTCTCTTACATGTGATGCCCCCACGACCAACTTGGCTATGATAAGGGAACTTGGCGCTGACCTAA
ATATTAACAACATGAAGCCTTACTTCATGCATCCTGAGGACCCTACTCAGAAAGTACATGTGATCCTGGATGCAT
GCCACATGCTAAAGCTTCTTCGCAATGCCTTTGCATCTTCCCTTGAATTTGAGACAGAAGATGGAAATAAGATTA
AGTGGAAGTACATCGAAGCCCTCAATGAACTCCAAGAAAAAGAAGGTCTGCGACTGGGAAACAAGCTGAAGATGG
CTCATTTACAATGGCGAAAACAAAAAATGAAAGTCCATCTGGCAGCACAGCTCTTCAGCAGCAGTGTGGCAGATG
CTATAGAGTTCTGTGAGCAAGGCTTGAAAATGGAGGAATTCAAAGGTTGTGCTGCAACTGTGCAGTTTTTGCGCA
CTGTAGATGCAGCTTTTGATGTACTGAATAGCAGAAATCCTCTTGGAAAAGGTTTCAAAGCCCCAATCAAAACAA
CAACTAAAGATCGGGTTGAGACCATTTTAAAGCAGGCAGAATCCATGCTTCGTGGGTTGAAGGTCCAGCAGTACA
ACAAGATGGTGCCACTTCACACAACCAAAAAGAAGACTGCCATATATGGCTTCATTGCTAATGGCAGAAGTGCAC
TTAACATCTATCATGATCTTGTTGAAAGGCCCAATGCACCATGCAGATATCTTCTTACCTACAAATTAAGTCAGG
ACCACTTGGAGCTTTTCTTTGCTGCCATAAGAGCAAGAGGTGGTCACAATGATAATCCTAATGCCAGGCAGTTTA
GAGGAGCATACAAACGCCTACTTGTAAGGCACCAAGTGAAAACAGGCACAGGGAACTGTTTGCTTCTAGACAACA
CTTACATGCTCAACTCAACACCTGCATCTGTAAATGTTGCCCGAAGATTGGAAGTACAGTTGGTGGAAGTGGACG
TGCCCGAGAATGATGCTGTTCCTGACCTCCCTCATGTATGCAGTCTTTCTGAATACAAAGAGGCAGCAATTCATT
ATATCACAGGATTTGTGGTGAAGAAGATGAAAGAGAAAATCACCTGCTTGCCATGCTCACAGGCATTAACAACTG
ATGGTGCTGCACATGAATTCATCCATCTAAAAAACAGAGGTGGTCTGCAGAAACCATCACCAGGCATGGTGTCTG
TGTGTCTGACCACAGAGAAATGCATTCAGAGAAAGATCACAACAAGTGGAGGTCAGCTTCCCCGTGGACGTGGAA
TCACACTTGCTATTTCCAACGAGGTGTTAGCAAATTGTGCAGAGAGAGATCTTTTTCCACAACTCCACAGCCACA
TGTTTGCTACAAGCGTTGAAATGAATCATATTCATTTATTGGTTAAAATGGCATCCATTTGGTATAGCAAAGTCA
GATTTAACCACTTTGCAAGAAGGGGAGCTGAAATAGCAAAAGATGGCAAAATGGTTCGCAGACAACTGACTAAA
TTGATTCATTTTTATGGTGAA

quence traduite (avec transeq)

1

SCSAVNCSNRFPLGRHDAARLKQWVVNMRRNGWKPTTSSRLCSAHFEEHAFTKDLKVKK
TAVPTIFSFPCHLVKSSCVGVRHKTFANKNSEVSGSSLLFSKDAGHSKKDSPVDKSLSVHSAA
DYSNATLTEHAIGNFSVSNDLCQVDTAVTVDINEDHSYATLSSTHETPKFIKEDHSYNLASPRSLK
RKNQATDQILQKYRKKLKIESQKSRRLKNKISTLKDVVTELKKKLLISNECASLLESIDEVPKHILLKI
QQGKKTAKYSEELKQFATTLHFYSPKAYDYVRDNFQKALPHSHTIRNWYSSVSADPGFTVASFT
ALKCHVEENKERGKETVCALMMDEMYIHKMTEFAGDQFHGYVDIGTGEIDNTLATQALVLMVVA
VNESWKIPIAYFLITGMDGSEKANVIRESLSRLHEVGVKVISLTCDAPTTNLAMIRELGADLNINNM
KPYFMHPEDPTQKVHVILDACHMLKLLRNAFASSLEFETEDGNKIKWKYIEALNELQEKEGLRLG
NKLKMAHLQWRKQKMKVHLAAQLFSSSVADAIEFCEQGLKMEEFKGCAATVQFLRTVDAAFDV
LNSRNPLGKGFKAPIKTTTKDRVETILKQAESMLRGLKVQQYNKMVPLHTTKKKTAIYGFIANGRS
ALNIYHDLVERPNAPCRYLLTYKLSQDHLELFFAAIRARGGHNDNPNARQFRGAYKRLLVRHQVK
TGTGNCLLLDNTYMLNSTPASVNVARRLEVQLVEVDVPENDAVPDLPHVCSLSEYKEAAIHYITG
FVVKKMKEKITCLPCSQALTTDGAAHEFIHLKNRGGLQKPSPGMVSVCLTTEKCIQRKITTSGGQ
LPRGRGITLAISNEVLANCAERDLFPQLHSHMFATSVEMNHIHLLVKMASIWYSKVRFNHFARRG
AEIAKDGKMVRRQLTKLIHFYGE

>_2

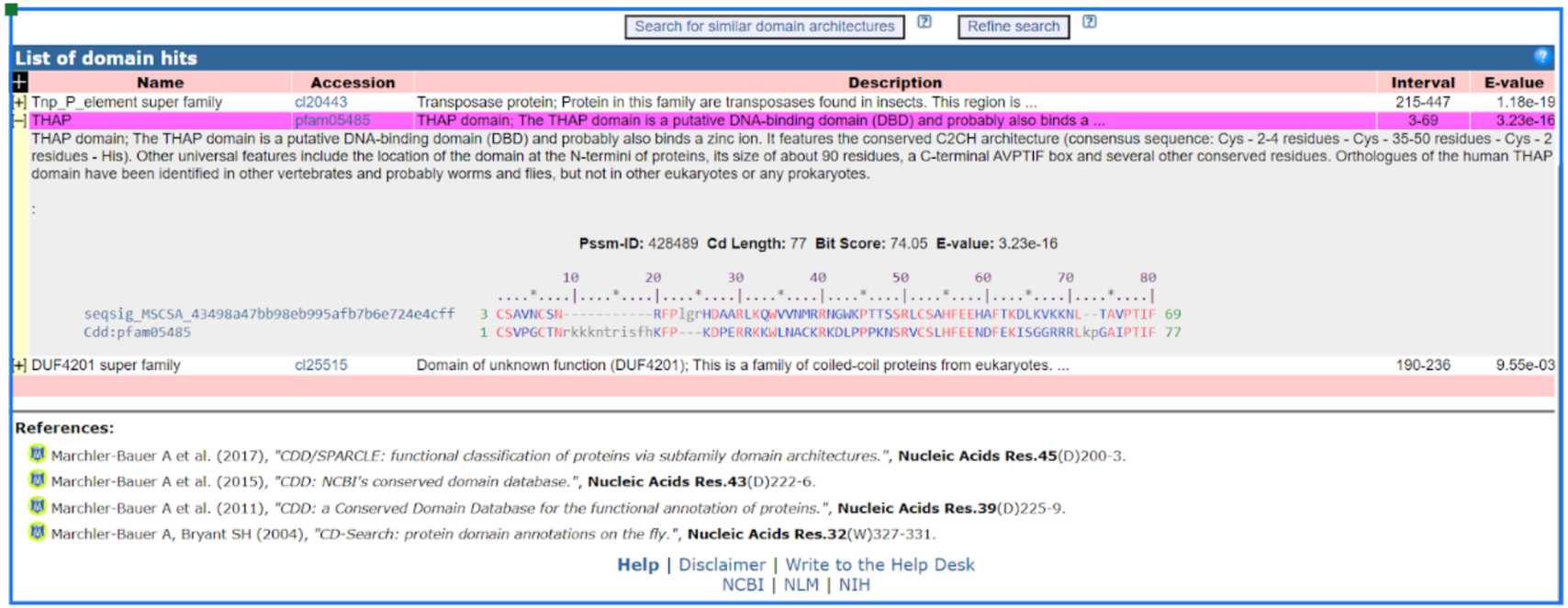
*VVQQ*TVLTVFL*VGMMQQD*NNGW*T*DATAGNQQLHPVFAVLILKSMLSPRI*R*KKILQQFQLY
SHFLVIW*NPLV*G*DTRLLRIRILR*VAPLCCSAKMQDTVKRILQLTSLLACTQLLITAMPP*LNMLLA
TSQLVMTYVKLTLL*RWILMRIIVMQL*VPHMKLLSLLKRITATTLRLRDL*NGKIRPLTKFYKNTGKS
LKLNPKSLGV*KTKSPP*RMSSQNSRRSC*YPMNVLLCLSQLMKCQSTSC*KYSRGKKQQSTQR
S*SSLPQPCIFIPQRHTTTYVTIFRKRCHTVTPFEIGTAAYQLILDLLWHLSQR*NVM*RRIKREEKK
QYVP**WMKCIFTR*LNLQGTSFMAMWTLVLEKLTTRWLPRLWSSWLWL*MSHGKFP*HTFSSQV
WMDQRRPMSLGRALADFMRLVSKSSLLHVMPPRPTWL**GNLALT*ILTT*SLTSCILRTLLRKYM*
SWMHATC*SFFAMPLHLPLNLRQKMEIRLSGSTSKPSMNSKKKKVCDWETS*RWLIYNGENKKK
SIWQHSSSAAVWQML*SSVSKA*KWRNSKVVLQLCSFCAL*MQLLMY*IAEILLEKVSKPQSKQQ
LKIGLRPF*SRQNPCFVG*RSSSTTRWCHFTQPKRRLPYMASLLMAEVHLTSIMILLKGPMHHADI
FLPTN*VRTTWSFSLLP*EQEVVTMIILMPGSLEEHTNAYL*GTK*KQAQGTVCF*TTLTCSTQHLH
L*MLPEDWKYSWWKWTCPRMMLFLTSLMYAVFLNTKRQQFIISQDLW*RR*KRKSPACHAHRH*
QLMVLHMNSSI*KTEVVCRNHHQAWCLCV*PQRNAFRERSQQVEVSFPVDVESHLLFPTRC*QI
VQREIFFHNSTATCLLQALK*IIFIYWLKWHPFGIAKSDLTTLQEGELK*QKMAKWFADN*LN*FIFM
VX

>_3

ELFSSKLF*PFSFR*A*CSKIETMGGEHETQRLETNNFIPSLQCSF*RACFHQGFEGKKKSYSSSN
YILISLSSGEILLCRGKTQDFCE*EF*GKWLLSVVQQRCRTQ*KGFSS*QVS*RALSC*LQQCHLD*
TCYWQLLS***PMSS*HCCNGGY**GS*LCNFKFHT*NS*VY*RGSQLQPCVSEIFKTEKSGH*PNS
TKIQEKA*N*IPKV*AFKKQNLHLKGCRHRTQEEAVDIQ*MCFFA*VN**SAKAHLVENTAGEKNSKV
LRGVEAVCHNLAFLFPKGIRLRT*QFSESAATQSHHSKLVQQRIS*SWIYCGIFHSAEMSCRGE*R
ERKRNSMCPNDG*NVYSQDD*ICRGPVSWLCGHWYYWRN*QHVGYPGFGPHGCGCK*VMENS
HSILSHHRYGWIREGQCH*GEP*QTS*GWCQSHLSYM*CPHDQLGYDKGTWR*PKY*QHEALLH
AS*GPYSESTCDPGCMPHAKASSQCLCIFP*I*DRRWK*D*VEVHRSPQ*TPRKRRSATGKQAED
GSFTMAKTKNESPSGSTALQQQCGRCYRVL*ARLENGGIQRLCCNCAVFAHCRCSF*CTE*QKS
SWKRFQSPNQNNN*RSG*DHFKAGRIHASWVEGPAVQQDGATSHNQKEDCHIWLHC*WQKCT*
HLS*SC*KAQCTMQISSYLQIKSGPLGAFLCCHKSKRWSQ**S*CQAV*RSIQTPTCKAPSENRHR
ELFASRQHLHAQLNTCICKCCPKIGSTVGGSGRARE*CCS*PPSCMQSF*IQRGSNSLYHRICGE
EDERENHLLAMLTGINN*WCCT*IHPSKKQRWSAETITRHGVCVSDHREMHSEKDHNKWRSASP
WTWNHTCYFQRGVSKLCRERSFSTTPQPHVCYKR*NESYSFIG*NGIHLV*QSQI*PLCKKGS*NS
KRWQNGSQTTD*IDSFLW*X

Recherche des domaines conservées

Alignement CD Search séquence 1 :



Search for similar domain architectures | Refine search

List of domain hits

	Name	Accession	Description	Interval	E-value
[+]	Tnp_P_element super family	cl20443	Transposase protein; Protein in this family are transposases found in insects. This region is ...	215-447	1.18e-19
[-]	THAP	pfam05485	THAP domain; The THAP domain is a putative DNA-binding domain (DBD) and probably also binds a ...	3-69	3.23e-16

THAP domain; The THAP domain is a putative DNA-binding domain (DBD) and probably also binds a zinc ion. It features the conserved C2CH architecture (consensus sequence: Cys - 2-4 residues - Cys - 35-50 residues - Cys - 2 residues - His). Other universal features include the location of the domain at the N-termini of proteins, its size of about 90 residues, a C-terminal AVPTIF box and several other conserved residues. Orthologues of the human THAP domain have been identified in other vertebrates and probably worms and flies, but not in other eukaryotes or any prokaryotes.

Pssm-ID: 428489 **Cd Length:** 77 **Bit Score:** 74.05 **E-value:** 3.23e-16

```
                                  10        20        30        40        50        60        70        80
                                  ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
seqsig_MSCSA_43498a47bb98eb995afb7b6e724e4cff   3 CSAVNCSN-----------RFPlgrHDAARLKQWVVNMRRNGWKPTTSSRLCSAHFEEHAFTKDLKVKKNL--TAVPTIF 69
Cdd:pfam05485                                   1 CSVPGCTNrkkkntrisfhKFP---KDPERRKKWLNACKRKDLPPPKNSRVCSLHFEENDFEKISGGRRRLkpGAIPTIF 77
```

	Name	Accession	Description	Interval	E-value
[+]	DUF4201 super family	cl25515	Domain of unknown function (DUF4201); This is a family of coiled-coil proteins from eukaryotes. ...	190-236	9.55e-03

References:

Marchler-Bauer A et al. (2017), *"CDD/SPARCLE: functional classification of proteins via subfamily domain architectures."*, **Nucleic Acids Res.45**(D)200-3.

Marchler-Bauer A et al. (2015), *"CDD: NCBI's conserved domain database."*, **Nucleic Acids Res.43**(D)222-6.

Marchler-Bauer A et al. (2011), *"CDD: a Conserved Domain Database for the functional annotation of proteins."*, **Nucleic Acids Res.39**(D)225-9.

Marchler-Bauer A, Bryant SH (2004), *"CD-Search: protein domain annotations on the fly."*, **Nucleic Acids Res.32**(W)327-331.

Pssm-ID: 428489 **Cd Length:** 77 **Bit Score:** 74.05 **E-value:** 3.23e-16

```
          10        20        30        40        50        60        70        80
....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
3 CSAVNCSN-----------RFPlgrHDAARLKQWVVNMRRNGWKPTTSSRLCSAHFEEHAFTKDLKVKKNL--TAVPTIF 69
1 CSVPGCTNrkkkntrisfhKFP---KDPERRKKWLNACKRKDLPPPKNSRVCSLHFEENDFEKISGGRRRLkpGAIPTIF 77
```

Alignement CD Search séquence 2 :

Pas de résultat

Alignement CD Search séquence 3 :

Pas de résultat

Protéine reconstituée domaine THAP complet

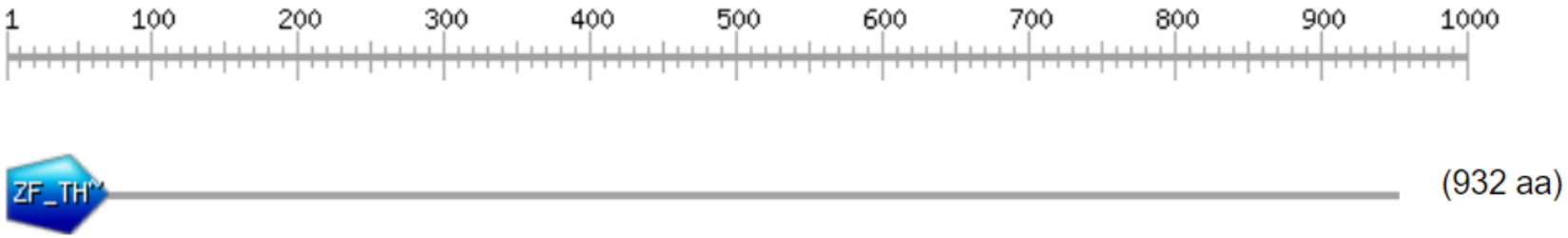
← Protéine reconstituée

← Logo PFAM

Scan Prosite

hits by profiles: [1 hit (by 1 profile) on 1 sequence]

Upper case represents match positions, lower case insert positions, and the '-' symbol represents deletions relative to the matching profile.



PS50950 **ZF_THAP** *Zinc finger THAP-type profile* :

l - 69: **score =** 12.707

----------MscSAVNCSNRFPlgrHDAARLKQWVVNMRRNG--WKPTTSSRLC
SAHFEEHAF-TKDLKVKKNLTAVPTIF

Predicted feature:

ZN_FING	-	46	THAP-type ; degenerate	[condition: not(<5=C> and <10=C> and <55=C> and <58=H> or <5=[CH]> and <10=[CH]> and <55=[CH]> and <58=[CH]>)]

Domaine THAP complet

Confirmation présence du motif
Comparaison protéine reconstituée / protéine THAP9 humain

par BlastP

Search Parameters	
Program	blastp
Word size	3
Expect value	0.05
Hitlist size	100
Gapcosts	11,1
Matrix	BLOSUM62
Filter string	F
Genetic Code	1
Window Size	40
Threshold	11
Composition-based stats	2

Download Graphics Sort by: E value

_1

Sequence ID: Query_27705 Length: 932 Number of Matches: 2

Range 1: 178 to 907 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
298 bits(762)	3e-91	Compositional matrix adjust.	219/760(29%)	377/760(49%)	58/760(7%)

```
Query  122  EDHNYSLKTPLTIGAEKLAEVQQMLQVSKKRLI--SVKNYRMIKKRKGLRLIDALVEEKL  179
            EDH+Y+L +P ++  +  A   Q+LQ  +K+L   S K+ R+  K   L+ +   +++KL
Sbjct  178  EDHSYNLASPRSLKRKNQA-TDQILQKYRKKLKIESQKSRRLKNKISTLKDVVTELKKKL  236

Query  180  L-SEETECLLRAQFSDFK---WELYNWRETDEYSAEMKQFACTLYLCSSKVYDYVRKILK  235
            L S E   LL +     K     ++    ++T +YS E+KQFA TL+   S K YDYVR    +
Sbjct  237  LISNECASLLESIDEVPKHILLKIQQGKKTAKYSEELKQFATTLHFYSPKAYDYVRDNFQ  296

Query  236  --LPHSSILRTWLSKCQPSPGFNSNIFSFLQRRVENGDQLYQ--YCSLLIKSMPLKQQLQ  291
              LPHS  +R W S      PGF     F+ L+  VE    +  +   C+L++   M + +  +
Sbjct  297  KALPHSHTIRNWYSSVSADPGFTVASFTALKCHVEENKERGKETVCALMMDEMYIHKMTE  356

Query  292  WDPSSHSLQGFMDFGLGKLDADETPLASETVLLMAVGIFGHWRTPLGYFFVNRASGYLQA  351
            +  +     G++D G G++D      LA++ ++LM V +   W+ P+ YF +     G  +A
Sbjct  357  F--AGDQFHGYVDIGTGEID---NTLATQALVLMVVAVNESWKIPIAYFLITGMDGSEKA  411

Query  352  QLLRLTIGKLSDIGITVLAVTSDATAHSVQMAKALGIHIDGDDMKCTFQHPSSSSQQIAY  411
             ++R ++ +L ++G+ V+++T DA    ++ M + LG  ++ ++MK  F HP   +Q++
Sbjct  412  NVIRESLSRLHEVGVKVISLTCDAPTTNLAMIRELGADLNINNMKPYFMHPEDPTQKVHV  471

Query  412  FFDSCHLLRLIRNAFQNFQSIQFINGIA-HWQHLVELVALEEQE-LSNMERIPSTLANLK  469
             CH+L+L+RNAF +    +  +G    W+++  L  L+E+E L     ++        +
     HMLKLLRNAFASSLEFETEDGNKIKWKYIEALNELQEKEGLRLGNKLKMAHLQWR  531

     VNSATQLFSESVASALEYL-LSLDLPPFQNCIGTIHFLRLINNLFDIFNSRNCYG  528
     V+ A QLFS SVA A+E+    L +  F+ C  T+ FLR ++  FD+ NSRN  G
     VHLAAQLFSSSVADAIEFCEQGLKMEEFKGCAATVQFLRTVDAAFDVLNSRNPLG  591
```

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	_1								

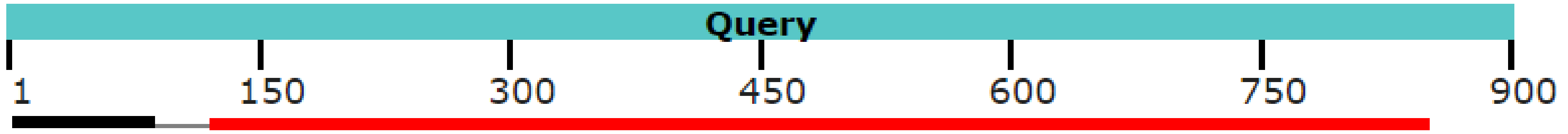
Range 2: 2 to 68 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
34.3 bits(77)	3e-05	Compositional matrix adjust.	26/88(30%)	37/88(42%)	24/88(27%)

```
Query  4   SCSAVGCSTRDTVLSRERGLSFHQFPT---DTIQRSKWIRAVNRVDPRSKKIWIPGPGAI  60
           SCSAV CS R                   FP    D  +  +W+  + R           W P   +
Sbjct  2   SCSAVNCSNR-------------FPLGRHDAARLKQWVVNMRR------NGWKPTTSSR  41

Query  61  LCSKHFQESDFESYGIRRKLKKGAVPSV  88
           LCS HF+E   F  +  ++ K     AVP++
Sbjct  42  LCSAHFEEHAF-TKDLKVKKNLTAVPTI  68
```

Distribution of the top 2 Blast Hits on 1 subject sequences



Fragment N-term récupéré plus long

Motifs CS, HF et AVP conservés

Conclusions (1/2)...

Ce que nous avons réussi à faire...

Confirmation de la présence du domaine THAP chez le Zebrafish
Séquence BX511023) aux positions de l'article

Séquence protéique BX511023 inexacte
domaine THAP incomplet: manque motif APTV

Après ajustement de la prédiction de la protéine
récupération du domaine THAP complet
>séquence reconstituée et traduction
TSSRLCSAHFEEHAFTKDLKVKKNLTAVPTIFSFPCHLVKSSC

Conclusions (2/2) ...

Par rapport à la publication :

- Cette protéine a probablement été renommée
- Sur NCBI: pas d’update sur le nom

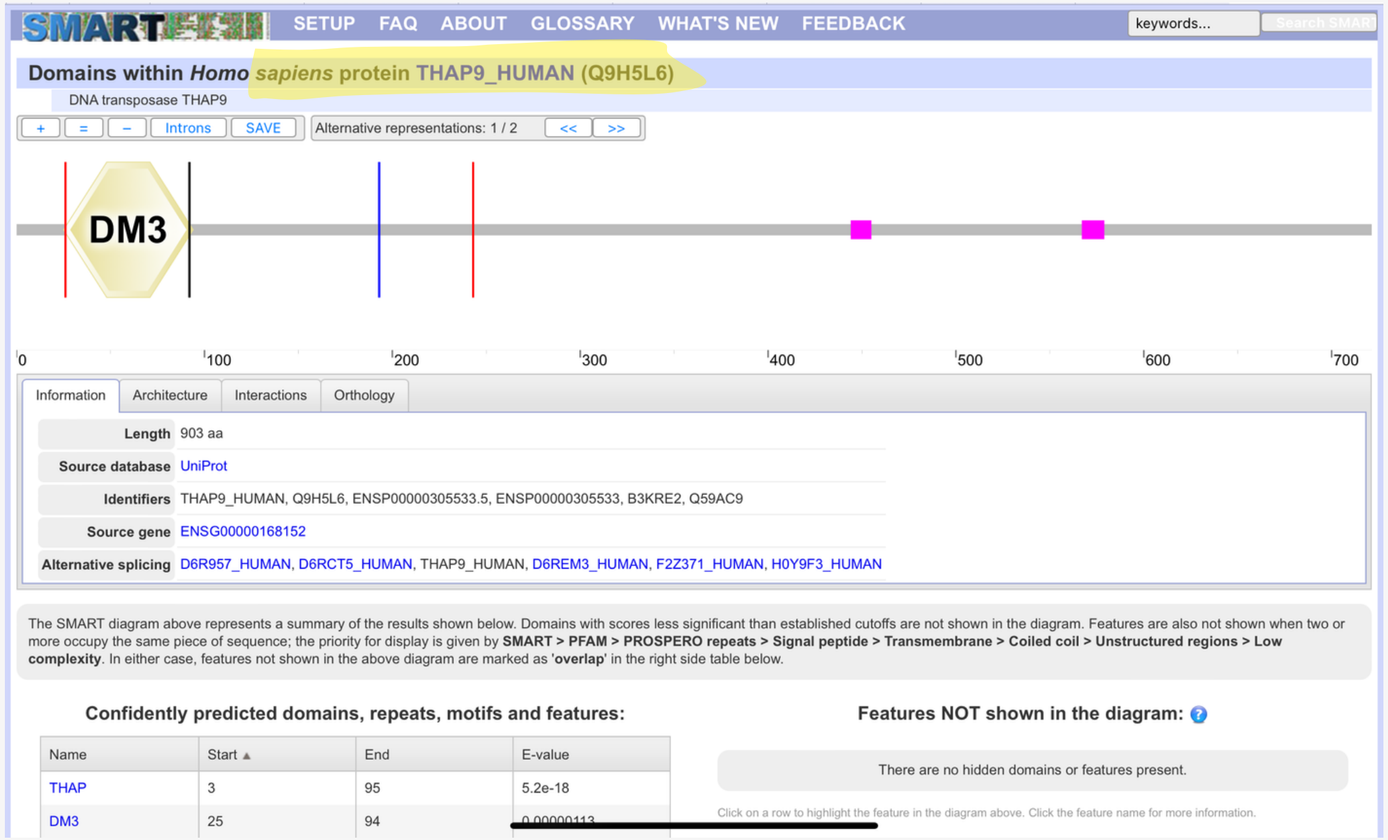
Les limites :

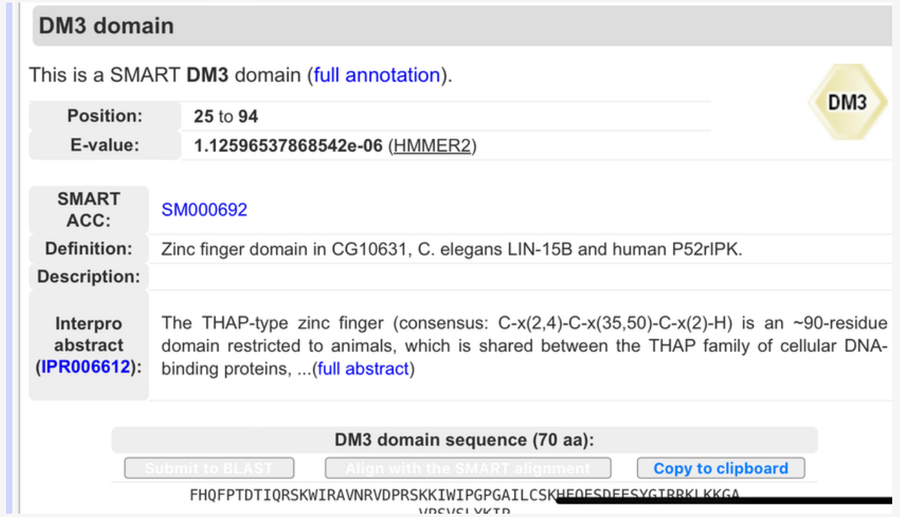
Nous n’avons jamais considéré les introns.

Recherche du logo du domaine THAP

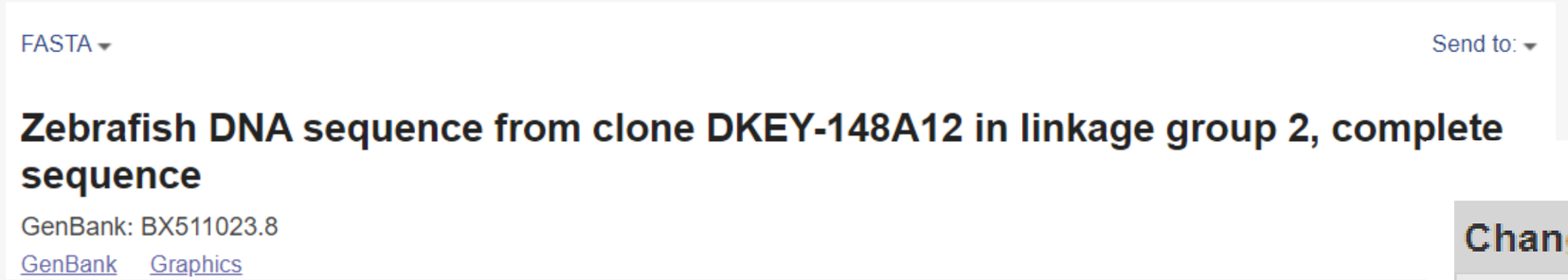
Pour connaître le motif à chercher

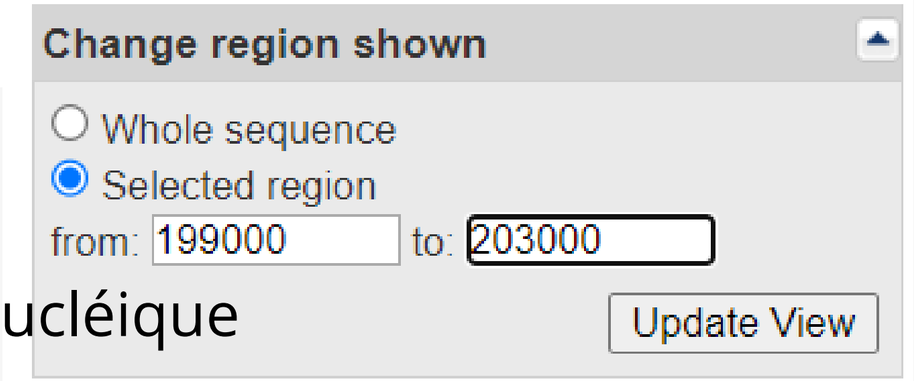
Sur SMART





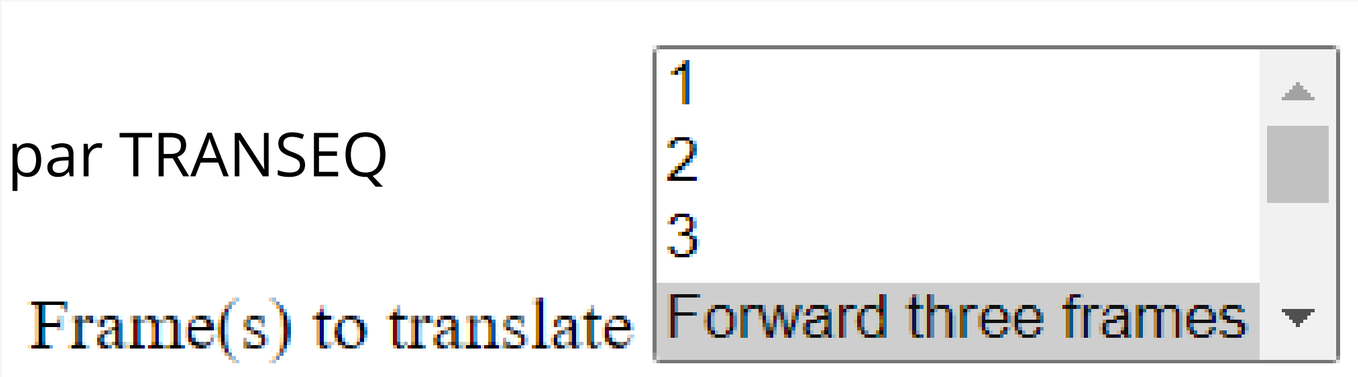
Reconstitution de la proteine à partir de l'ADN utilisation positions publiées



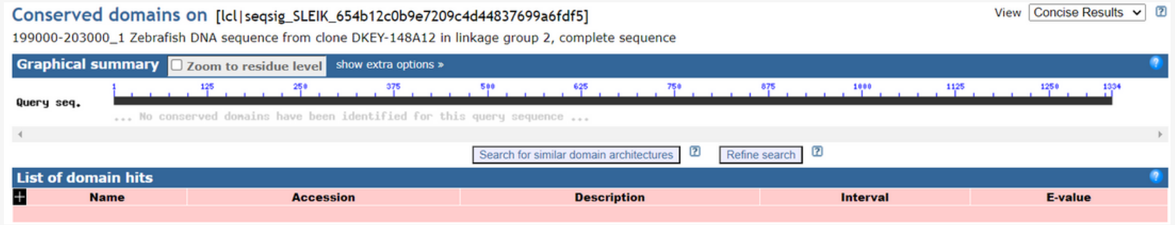


Sélection portion séquence nucléique

Transcription de la séquence par TRANSEQ



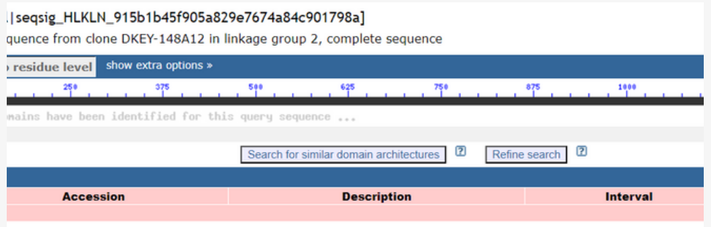
CD SEARCH

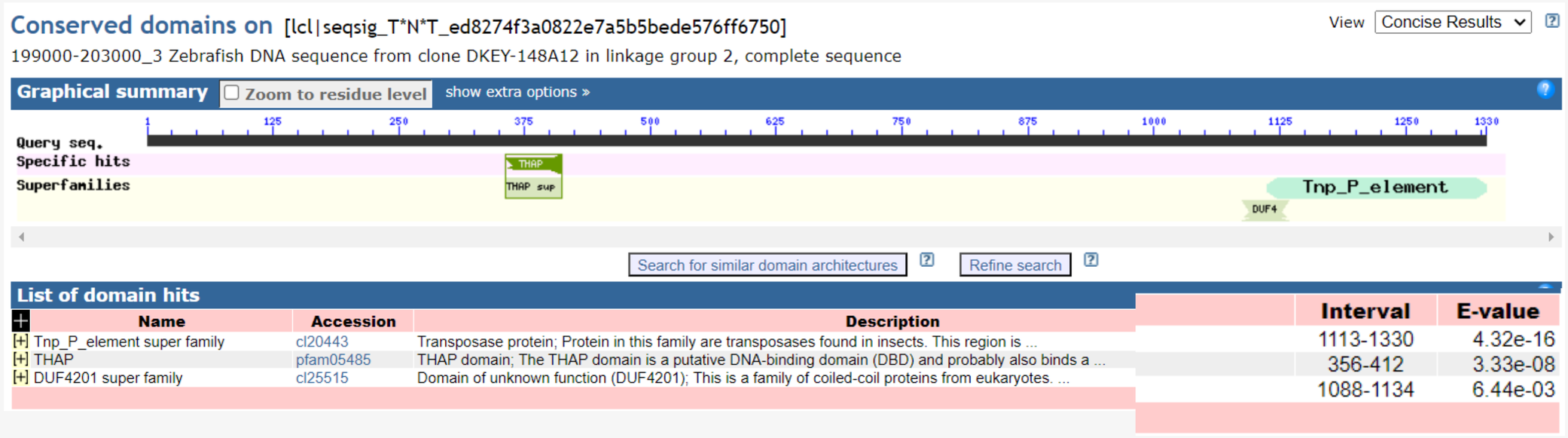


--> Localiser le domaine THAP:

on retrouve la même zone du domaine:

il manque toujours la partie N-term.





La 3ième séquence nous donne une correspondance avec le domaine THAP.

Position de l'alignement : 200065 - 200235

Pas de motif THAP dans la séquence BX511023: probablement en amont

PFAM

Deuxieme alignement obtenu lors du tBlastN

Range 2: 200145 to 200201 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
24.3 bits(51)	4.1	Compositional matrix adjust.	9/19(47%)	11/19(57%)	0/19(0%)	+3

```
Query   53       WIPGPGAILCSKHFQESDF   71
                 W P   + LCS HF+E  F
Sbjct   200145   WKPTTSSRLCSAHFEEHAF   200201
```

```
            10         20         30         40         50
     ....*....|....*....|....*....|....*....|....*....|....*..
356  FCSFPlgrHDAARLKQWVVNMRRNGWKPTTSSRLCSAHFEEHAFTKDLKVKKNLQNN 412
 19  HKFP---KDPERRKKWLNACKRKDLPPPKNSRVCSLHFEENDFEKISGGRRRLKPGa 72
```

Même zone