

# **Annotation Structurale: Caractérisation d'un domaine THAP chez le Zebrafish**

Présenté par Célia, Ema et Véronique

# Contexte...



> [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](#)<sup>1</sup>, [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

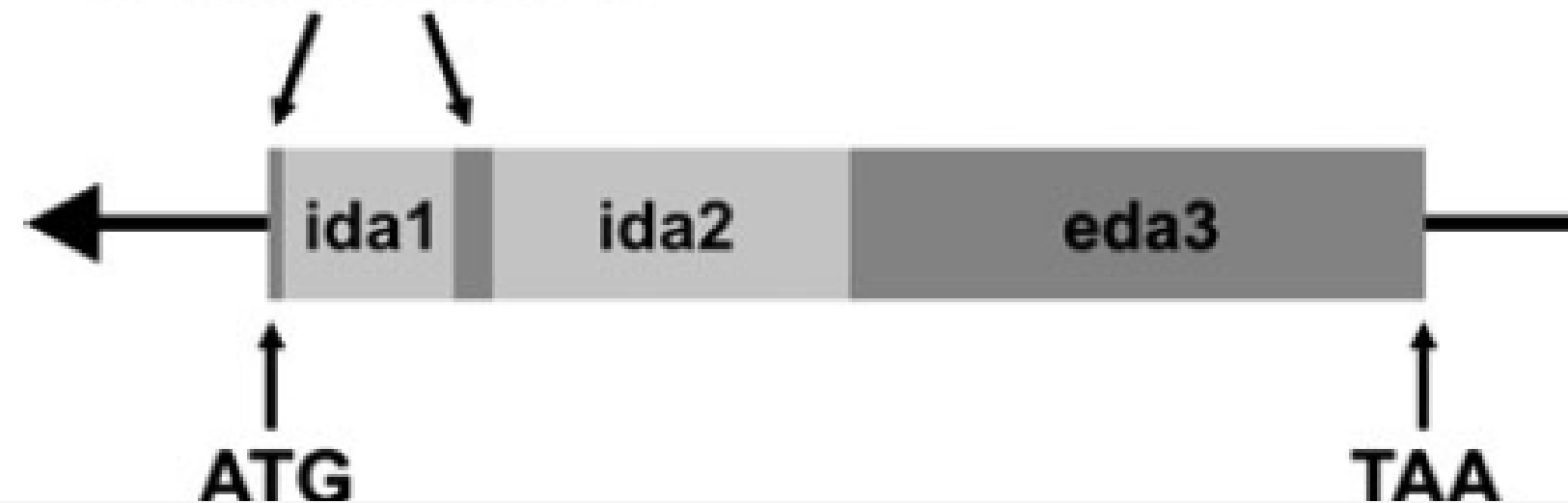


# Les objectifs du projet

## Confirmation/vérification des résultats publiés

### *Pdre2* - Zebrafish

putative THAP  
domain eda1+2



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

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

## SUR UNIPROT

UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB thap9 Advanced | List Search

Status  
Unreviewed (TrEMBL) (1)

UniProtKB 1 result

Popular organisms  
Zebrafish (1)

Entry Entry Name Protein Names Gene Names Organism Length  
 A0A8M1NZL3  A0A8M1NZL3\_DANRE THAP domain-containing protein 1 si:ch73-382f3.1, fj83h03, wu:fj83h03 Danio rerio (Zebrafish) (Brachydanio rerio) 283 AA

Taxonomy

**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

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

Query subrange From  To

Or, upload file  Aucun fichier choisi

Job Title  Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Databases  Standard databases (nr etc.) new  Experimental databases

[Try experimental clustered nr database](#) For more info see What is clustered nr?

Compare  Select to compare standard and experimental database

Standard

Database  ?

Organism   exclude  Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude  Models (XM/XP)  Non-redundant RefSeq proteins (WP)  Uncultured/environmental sample sequences

Program Selection

Algorithm  Quick BLASTP (Accelerated protein-protein BLAST)  blastp (protein-protein BLAST)  PSI-BLAST (Position-Specific Iterated BLAST)  PHI-BLAST (Pattern Hit Initiated BLAST)  DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST) Choose a BLAST algorithm

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
<input checked="" type="checkbox"/>	<a href="#">RecName: Full=THAP domain-containing protein 1 [Danio rerio]</a>	<a href="#">Danio rerio</a>	50.4	50.4	9%	2e-07	33.71%	225	<a href="#">Q1JPT7.2</a>
<input checked="" type="checkbox"/>	<a href="#">RecName: Full=THAP domain-containing protein 11 [Danio rerio]</a>	<a href="#">Danio rerio</a>	36.6	36.6	8%	0.006	30.67%	257	<a href="#">Q6TGZ4.2</a>

## Contre NR

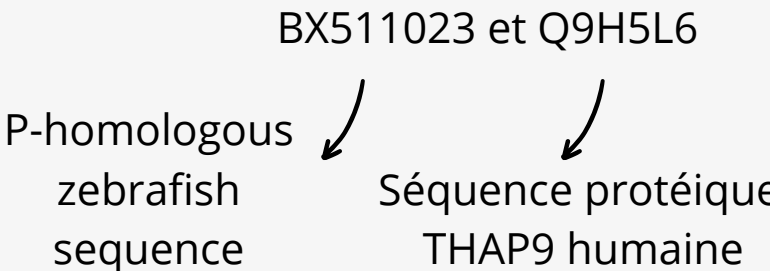
select all 20 sequences selected [GenPept](#) [Graphics](#) [Distance tree of results](#) [Multiple alignment](#) [MSA Viewer](#)

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

**BX511023 absente**  
**Pas de THAP9 dans les 2 Blast**

# Alignement Séquence THAP Humaine contre Organisme Zebrafish

## tBlastN



**Enter Query Sequence**

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)

>sp|Q9H5L6|THAP9\_HUMAN DNA transposase THAP9 OS=Homo sapiens  
OX=9606 GN=THAP9 PE=1 SV=2  
MTRSCSAVGCSTRDVTLSRERGLSFHQFPTDIQRSKWIRAVNRVDPRSKKI  
WIPGPGAI

Query subrange [?](#)

From

To

Or, upload file  Aucun fichier choisi [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

---

**Choose Search Set**

Database  [?](#)

Organism  will be suggested  exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Exclude  Models (XM/XP)  Uncultured/environmental sample sequences

Limit to  Sequences from type material

Entrez Query  [YouTube](#) [Create custom database](#)

Enter an Entrez query to limit search [?](#)

e-values et query cover faibles

**Séquence  
BX511023**



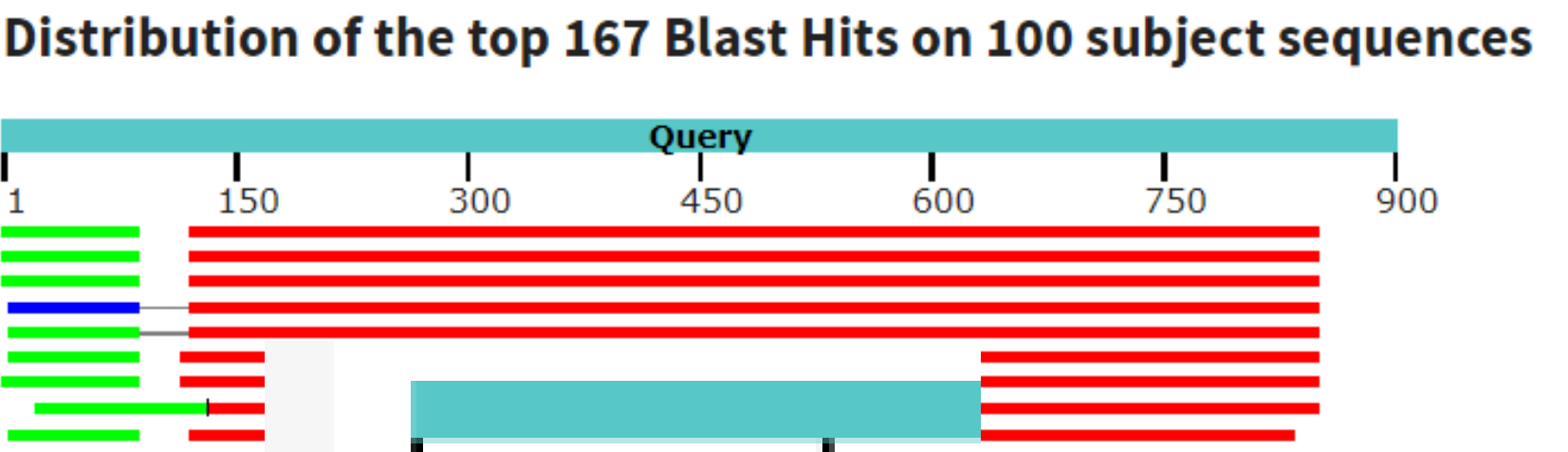
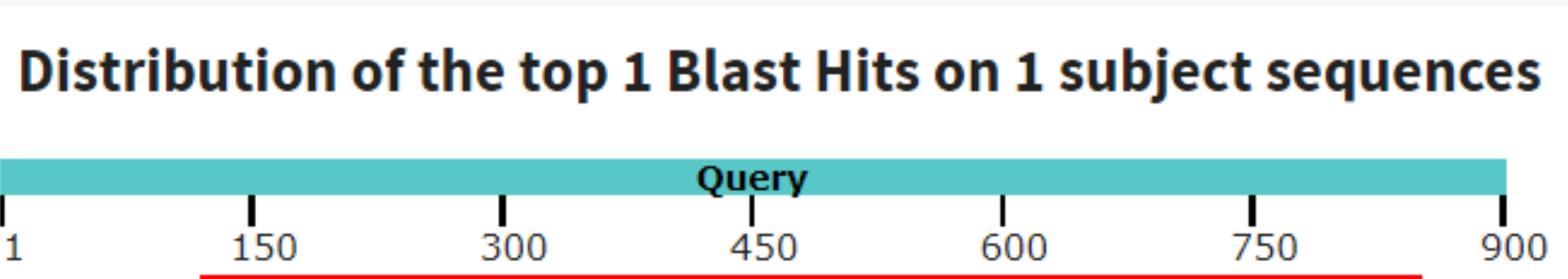
10 sequences selected

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



# 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

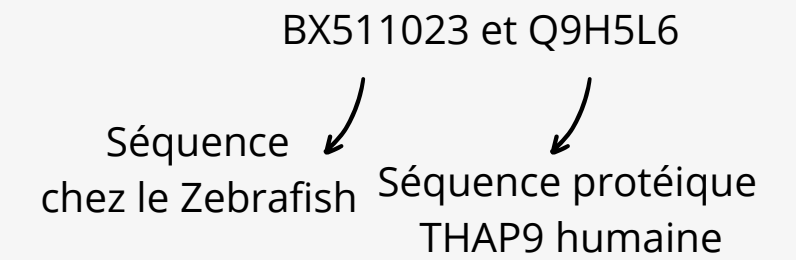
Range 1: 202236 to 204425 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#)

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		EDHNYSLKPTLTIGAEKLAEVQQLQVSKKRLI--SVKNYRMIKKRRKGLRLIDALVEEKL				179
Sbjct 202236		EDH+Y+L +P ++ + A Q+LQ +K+L S K+ R+ K L+ + +++KL				202412
Query 180		L-SEETECLLRAQFSDFK---WELYNWRETDEYSAEMKQFACTLYLCSKVVYDVRKILK				235
Sbjct 202413		LISE LL + K ++ ++T +YS E+KQFA TL+ S K YDYVR +				202592
Query 236		--LPHSSILRTWLSKQPSPGFNSNIFSLQRRVENGDLQYQ--YCSLLIKSMPLKQQLQ				291
Sbjct 202593		LPHS +R W S PGF F+ L+ VE + + C+L++ M + + +				202772
Query 292		WDPSSHSLQGFMDFGKGLDADETPLASETVLLMAVGIFGHWRTPLGYYFVNRASGYLQA				351
Sbjct 202773		+ + G++D G G++D LA++ ++LM V + W+ P+ YF + G +A				202937
Query 352		QLLRLTIGKLSDIGITVLAVTSDATAHSVQMAKALGIHIDGDDMKCTFQHPSSSSQOIAY				411
Sbjct 202938		++R ++ +L ++G+ V+++T DA ++ M + LG ++ ++MK F HP +Q++				203117
Query 412		FFDSCHLLRLIRNAFQNFQSIQFINGIA-HWQHlvelvaleeqelsNM-ERIPSTLANLK				469
Sbjct 203118		D+CH+L+L+RNAF + + +G W+++ L L+E+E + ++ +				203297
Query 470		NHVLKVNATQLFSESVASALEYL-LSLDLPFPQNCIGTIHFLRLINLFDIFNSRNCYG				528
Sbjct 203298		+KV+ A QLFS SVA A+E+ L + F+ C T+ FLR ++ FD+ NSRN G				203477
Query 529		KGLKGPLLPETYSKINHVLEIAKTIKIFVTLSDTSNNQIIkgkqkfgf---llNAESLK				585
Sbjct 203478		KG K P+ T ++ +L +A+++ L N+++ + N S				203657
Query 586		WLYQNYVFPKVMPPYLLTYKFSHDHLELFLKMLRQVLVTSSTPCMAFQKAYNLETRY				645
Sbjct 203658		+Y + V P YLLTYK S DHLELF +R + +P F+ AY L R+				203837
Query 646		KFQ-----DEVFLSKYSIFDISIARRKDLALWTVQRQYGVSVTKTVFHEEGICQDN				696
Sbjct 203838		+ + D ++ + + +ARR ++ L V E D				203984
Query 697		SH-CSLSEALLDLSHRRNLICY-AGYVANKLSALLTCEDCITALYASDLKASKIGSLLF				754
Sbjct 203985		H CSLSE ++ I Y G+V K+ +TC C AL +D A + +				204131
Query 755		VKKKNGLHFPSESLCRVINICERVVTRHSRMAIFELV-SKQRELYLQKILCELSEGHINL				813
Sbjct 204132		+K + GL PS + V E+ ++ + +L + L + ++L + +L				204308
Query 814		FVDVNKHLFDGEVCAINHFVLLKDIICFLNIRAKNVAQ				853
Sbjct 204309		F ++ HF V +NH L+K I + +R + A+				204425

# Conclusion sur Blast

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

# Recherche du domaine THAP sur la séquence BX511023



**Paramètres modifiés :  
expect threshold et word size**

## tblastN

Expect threshold

Word size

Max matches in a query range

**Scoring Parameters**

Matrix

Gap Costs Existence: 11 Extension: 1

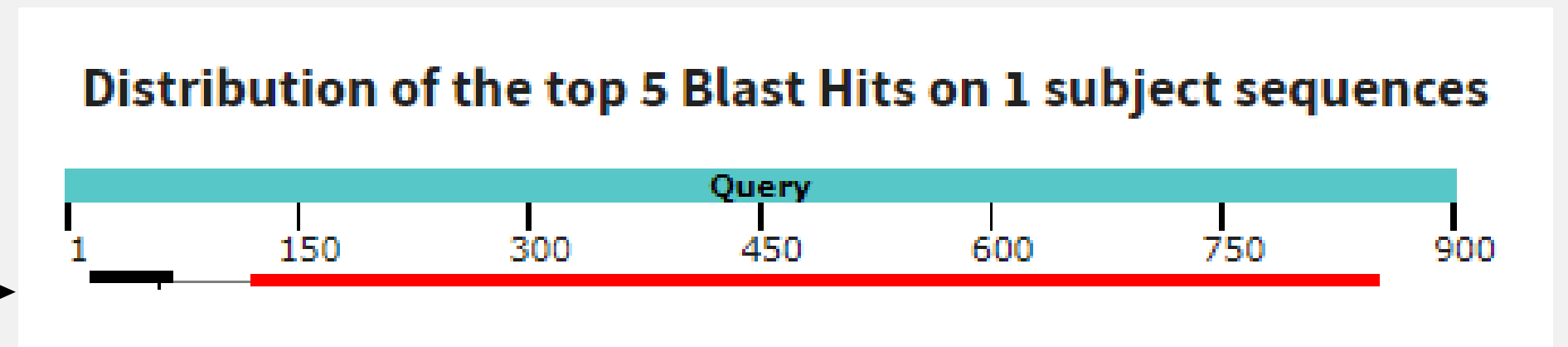
Compositional adjustments Conditional compositional sc

**Filters and Masking**

Filter  Low complexity regions

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/> 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

**1 fragment supplémentaire  
en partie N-term**



Range 1: 202236 to 204425 [Graphics](#)

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

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 EDHNSLKTPLTIGAEKLAEVQQLQVSKKRLI--SVKNYRMIKKRGLRLIDALVEEKL 179

Sbjct 202236

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

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

Gaps	Frame
0/19(0%)	+3

Query 53 WIPGPGAILCSKHFQESDF 71

Sbjct 202773

W P + LCS HF+E F

Sbjct 200145 WKPTTSSRLCSAHFEEHAF 200201

Sbjct 202938 NVIRELSRLCHEVGKVISLTCDAPTINLAMPRELGADLNINNPYPFMPLEDPQKVRV 203117

Query 412 FFDSCHLLRLIRNAFQNFQSIQFINGIA-HWQHLVELVALEEQE-LSNMERIPSTLANLK 469

Sbjct 203118 ILDACHMLKLLRNAFASSLEFETEDGNKIKWKYIEALNELQEKEGLRLGNKLMMAHLQWR 203297

Query 470 NHVLKVN SATQLFSESVASALEYL-LSLDLPFPQNCIGTIHFLRLINLFDIFNSRNCYG 528

Sbjct 203298 KQKMKVHLAAQLFSSSVADAIEFCEQGLKMEEFKGAATVQFLRTVDAADFVLSNRNPLG 203477

Query 529 KGLKGPLLPETYSKINHVLIEAKTIFVTLSDTSNNQII---KGKQKLGFLGFLNLAESLK 585

Sbjct 203478 KGFKAPIKTTTKDRVETILKQAESMLRGLKVQYQYKMPVPLHTTKKTAIYGFIANGRSAL 203657

Query 586 WLYQNYVFPKVMPPYLLTYKFSHDHLELF LKMLRQVLVTSSSPTCMFQKAYYNLETRY 645

Sbjct 203658 NIYHDLVERPNAPCRYLLTYKLSQDHLELFFAAIRARGGHNDNPNAQFRGAYKRLLVHR 203837

Query 646 KFQ-----DEVFLSKVSIFDISIARRKDLALWTVQRQYGVSVTKTVFHEEGICQDW 696

Sbjct 203838 QVKTGTGNCLLLDNTYMLNSTPASVNVARRLEVQLVEVD-----VPENDAVPDL 203984

Query 697 SH-CSLSEALLDLSDHRRNLICY-AGYVANKLSALLTCEDCITALYASDLKASKIGSLLF 754

Sbjct 203985 PHVCSLSE-----YKEAAIHVITGFVVKMKKEKITCLPCSQAL-TTDGAAHE---FIH 204131

Query 755 VKKKNGLHFPSESLCRVINICERVVTRTHSRMAIFELV-SKQRELYLQQKILCELSGHINL 813

Sbjct 204132 LKNRGGGLQKPSGPMVSVCLTTEKCIQRKITTSGGQLPRGRGITLAISNEVLNCAER-DL 204308

Query 814 FVDVNKHLFDGEVCAINHFKLLKDIIICFLNIRAKNVAQ 853

Sbjct 204309 FPQLHSHMFATSV-EMNHIHLLVKMASIWYSKVRFNHFAR 204425

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 SRERGLSFHQFPTDTIQRSKWIRAVNRVDPKSKKIWIWIPGPGAIL 61

Sbjct 152904 TKQK\*IIVHNSKNKTIQ\*NQVLHILNNDPRSVNVACVDTTALL 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 MKCTFQHPSSSSQIAYFFDS 415

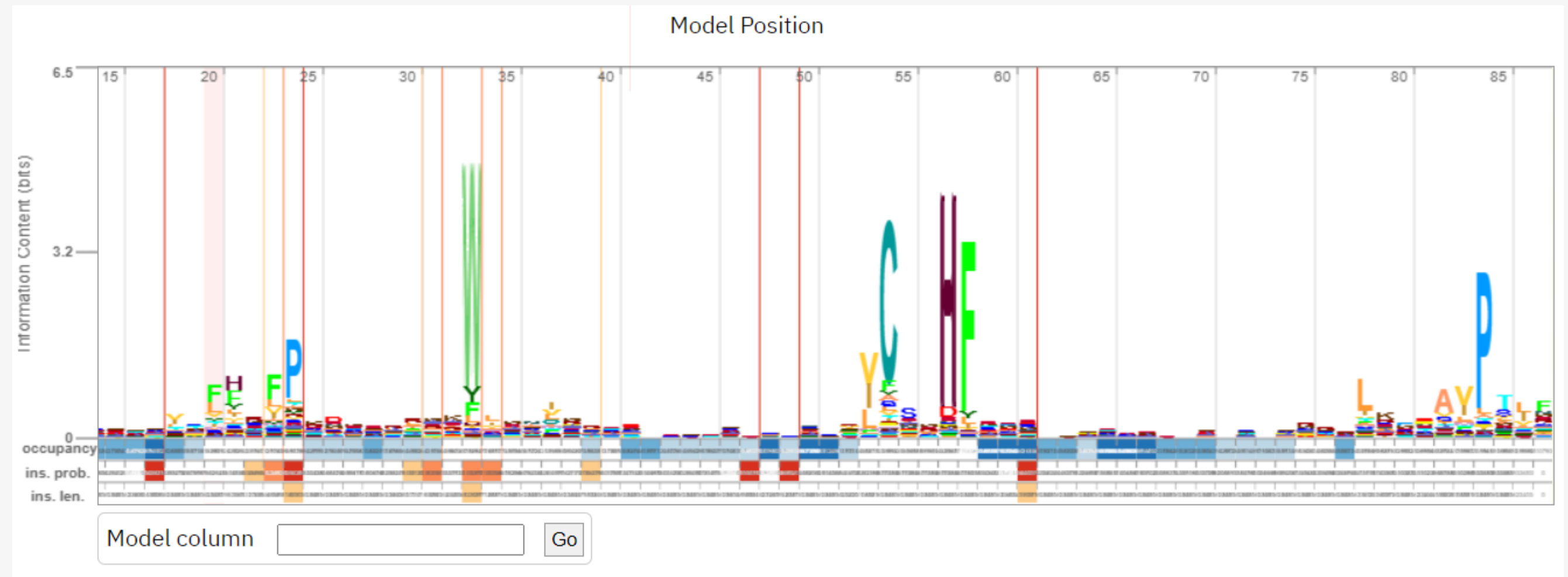
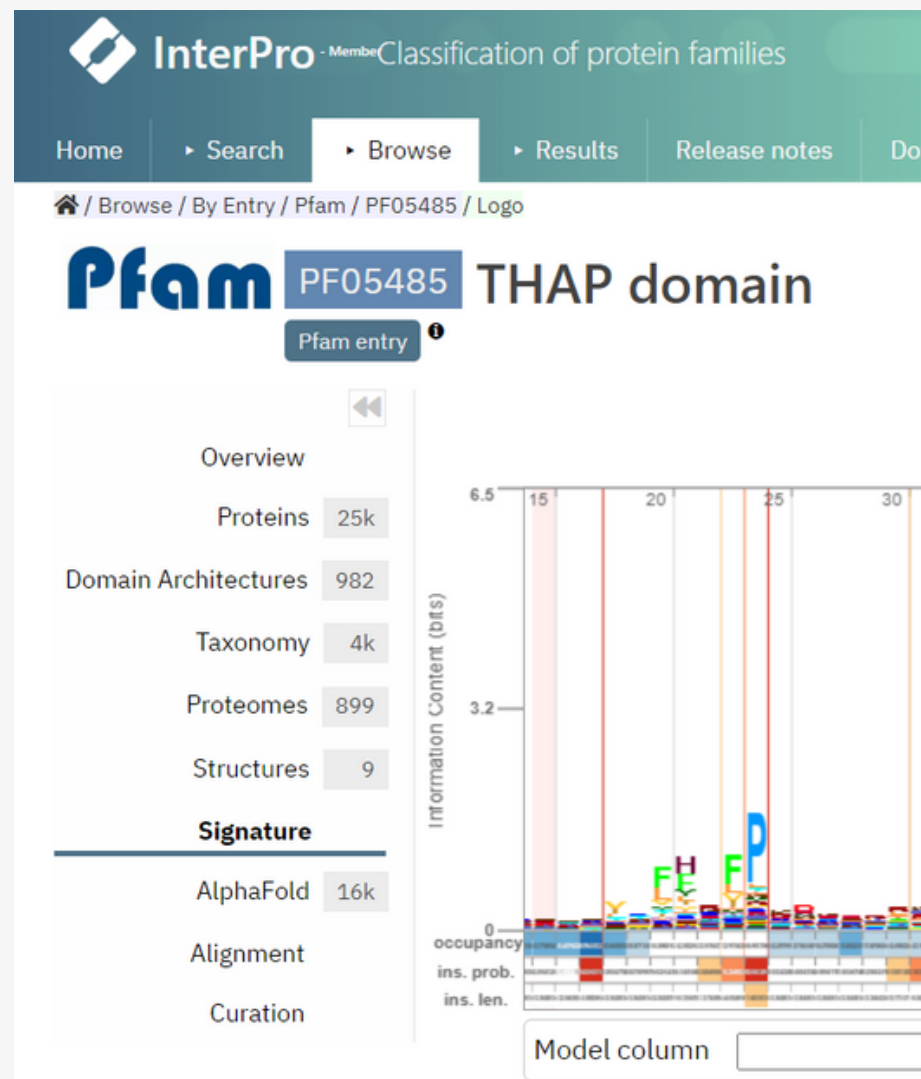
Sbjct 101941 IEKTIRHPESIKQQVGYFVGA 101879

**Segment AVP absent  
sûrement plus loin**

# Recherche du logo du domaine THAP

Pour connaître le motif à chercher

Sur Interpro: Logo du domaine THAP

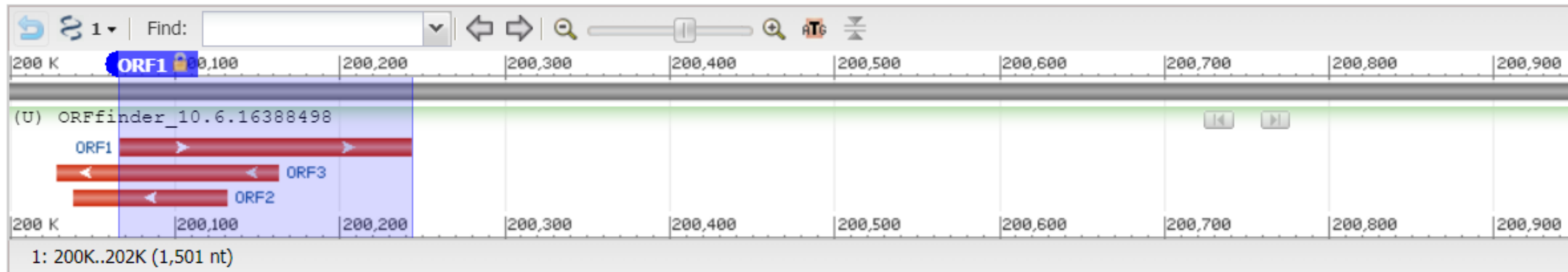


# Recherche d'ORF avec ORF Finder

## 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

Mark subset...

Marked: 0

Download marked set

as

Protein FASTA

```
>lcl|ORF1
MFCSPFLGRHDAARLKQWVVMRRNGWKPTTSSRLCSAHFEEHAFTKDLK
VKKNLQNN
```

Une partie du motif THAP visible

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

**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 récupérer en entier

## *Pdre2* - Zebrafish



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:  to:

[Update View](#)

199271 - 199302  
>BX511023.8:199271-199302 Zebrafish DNA sequence from clone DKEY-148A12 in linkage group 2, complete sequence  
ATGAGTGTTCAGCAGTAAACTGTTCTAACCGT  
200078 - 200230  
>BX511023.8:200078-200230 Zebrafish DNA sequence from clone DKEY-148A12 in linkage group 2, complete sequence  
TTTTCTTTAGGTTAGGATGATGCAAGATTGAAACAATGGTGGTGAACATGAGACGCCAGGCGTGGAAACCAACACCTTCATCCCGCTTTGAGATGCTCATTTTGAAGAGATGCTTTCCACCAAGGATTTGAAGGTA  
201890 - 204500  
>BX511023.8:201890-204500 Zebrafish DNA sequence from clone DKEY-148A12 in linkage group 2, complete sequence  
TACAGCAGTTCCAACCTTATCTCAATTTCCCTGTCATCTGGTGAATCCCTTTGTTAGGGGTAAGCACAAAGACTTTTGGCAATAAAGACTTTCAGCAAAAGATGCAGGACACAGATAAARAGGATTCTCCAGTGTGCAAGTCTCTTAGCGTGCACCTAGCTGCTGATTACAGCAATGCCACTGACTGAACATGCTATGGCACTTCTCAGTTAGTAAATGACCTATGTCAGTTGACACTGCTGAACGGTGTATTAATGAGGATCATAGTTATGCAACTTTAAGTCCACACATGAACTCCTAAGTTTATTAAGAGGATCACAGCTACAACTTGGCTCTCCGAGATCTTTAAAACGGAAAAATCAGGCCACTGACCAAAATCTACA  
AAAATACAGGAAAAAGCTTAAAAATGAATCCCAAAAGTCTAGGCGTTTTAAAAACAAAATCTCCACTTAAGGATGTCGCACAGAACTCAAGAAAGGACTGTTGATATCCAAATGAATGTGCTTCTTTGCTTGAAGTCAA  
TTGATGAAGTGCACAAGACATCTTTGTTGAAAATACAGCAGGGGAAAAAAACAGCAAAAGTACTCAGAGGAGTTGAAGCAGTTTGCCCAACCTTGCATTTTTATTCCCAAAGGCATACGACTACGTGACAAATTT  
CAGAAGCGCTGCCACACAGTCCACCACTTCGAAAATGGTACAGCAGCTATCAGCTGATCCTGGATTTA  
CTGTGGCATCTTTACAGCGCTGAAAATGTATGTAGAGGAGAAATAAAGAGAGAGAAAAAGAAAACAGATG  
TGCCCTAATGATGGATGAAATGATAATTCACAAGATGACTGAAATTTGCAAGGACCGATTTTCATGGCTAT  
GTGGACATTTGATGAGAAATGACAACAGCTTGGCTACCCAGGCTTTGGCTCATGGTGTGGGCTG  
TAAATGAGTGTGAAAAATCCCAATAGCATACTTTCTCATCACAGTATGGATGGATCAGAAAGGCCAA  
TGTCATTAGGGAGAGCCCTTAGCAGACTTCATGAGGTTGGTGTCAAAGTATCTCTCTTACATGTGATGCC  
CCCAAGCAACTTTGGCTATGATAAGGGAAGTTGCGCTGACCTAAATTAACAACATGAAGCCCTTACT  
TCATGATCCTGAGGACCCTACTCAGAAAGTACATGTGATCCTGGATGCACTGCCACATGCTAAAAGCTTCT  
TGCAATGCCCTTTCATCTTCCCTTGAATTTGAGACAGAAGATGAAAAAAGATTAAGATTAAGTGGAAAGTACATC  
GAAAGCCCTCAATGAACCCAGAAAAAGAGGCTGCGACTGGAAAAAGCTGAAGATGGCTCATTAC  
AATGGGCAAAAAAATAAGATGAAAGTCCATCTGCGAGCAGACTCTTACAGCAGTGTGGCAGATGCTAT  
AGAGTCTGTGAGCAAGGCTTGAATAAGGGAATCAAGGTTGTGCTGCAACTGTGAGATTTTGGG  
ACTGTAGTGCAGCTTTGATGTACTGAATAGCAGAAAAATCCTCTTGGAAAAGGTTTTCAAAAGCCCCAATCA  
AAACAACACTAAAGATCGGGTGGAGCCTTTAAAAGCAGGCAGAAATCCATGCTTTCGTTGGGTTGAAGG  
CCAGCATCAACAAAGATGGTGCACCTTCACACAAACAAAAAAGAGACTGGCATATGAGCTTCAATGCT  
AATGGCAGAAGTGCATTAACATCTATCATGATCTTTGTTGAAAGGCCAAATGCACCATGCGATATCTTC  
TTACCTACAAATTAAGTCCAGGACACTTTGGAGCTTTCTTTGCTGCCATTAAGAGCAGAGGTTGGTCAAAA  
TGATAATCCTAATGCCAGGCACTTTAGAGGAGCATAACACGCTACTTGTAAAGCACAAGTGAAGAAAC  
GGACACAGGGAAGCTTCTCTTAGCAGCAACTTACATGCTCACTCAACACTGCATCTGTAAATGTTG  
CCGAAGATTTGGAAGTACAGTTGGTGGAGTGAAGTGGCAGCAAGTGAAGTGTCTTCTGACCTCCTC  
TGATGACAGTCTTCTGAATAACAAGAGGCAGCAATCATTATACAGGAATTTGGTGGTGAAGAGATG  
AAGAGAAAAATCAACTGCTTTCAGAGCACTTCAACAGGCATTAACAAGTGAAGTGGTGTGACATGAAATCACTC  
ATCTAAAAACAGAGGTTGGTGTGCAAGAACCAATCAACAGGCAATGCTGCTGCTGCTGACCAAGAGAA  
ATCATTTCAGAGAAAGATCACAAACAGTGGAGTTCAGCTTCCCGTGGAGCTGGAATCAGTATGCTATT  
TCCAAAGAGGTTGATGAAATGTCAGAGAGAGACTTTTTCCAAACTCCACAGCACAATGTTTGTCTA  
CAAGCGTGAAGTGAATCATATTTCATTTTGGTAAATGGCATCCTATTTGGTATAGCAAAGTCA  
TAACCACTTTCAGAAAGGAGGAGCTGAAATAGCAAAAGATGGCAAAATGGTTCGAGCAGCACTGACTAAA  
TTGATTCATTTTTATGGTGA

### Séquence assemblée

ATGAGTGTTCAGCAGTAAACTGTTCTAACCGT  
GTGGTGAACATGAGACGCAACGGCTGGAAACCT  
GCTTTCACCAAGGATTTGAAGTAAAAAAAATCTTACAGCAGTTCCAACCTATATTCTCATTTCCTTGTCTATCTG  
GTGAAATCCTCTGTGTAGGGTAAGACACAAGACTTTGCGAATAAGAATCTGAGGTAAGTGGCTCCCTCTCTG  
TTGTTACAGCAAAGATGCAGGACACAGTAAAAGGATTTCCAGTTGACAAGTCTCTTAGCGTGCACCTCAGCTGCT  
GATTACAGCAATGCCACCTTACTGAACTATGCTATTGGCACTTCTCAGTTAGTAAATGACCTATGTCAGTTGAC  
ACTGCTGTAACGGTGGATATTAATGAGGATCATAGTTTGAACCTTAAAGTCCACACATGAACTCCTAAGTCTT  
ATTAAGAGGATCACAGCTACAACTTGGCTCTCCGAGATCTTTAAAACGGAAAAATCAGGCCACTGACCAAAAT  
CTACAAAAATACAGGAAAAAGCTTAAAATGAATCCCAAAAGTCTAGGCGTTTTAAAAACAAAATCTCCACCTTA  
AAGGATGTCGTACAGAACTCAAGAAGAAGTGTGATATCCAATGAATGTGCTTCTTTGCTTGAAGTCAATGAT  
GAAGTGCCAAAGCACATCTTGTGAAAATACAGCAGGGGAAAAAAACAGCAAAAGTACTCAGAGGAGTTGAAGCAG  
TTTGCCACAACCTTGCATTTTTATTCCCAAAGGCATACGACTACGTACGTGACAAATTTTCAGAAAGCGCTGCCA  
CAGACTCACACCATTCGAAATTTGGTACAGCAGCGTATCAGCTGATCCTGGATTTA  
CTGAAATGTCTAGAGGAGAAATAAAGAGAGAGAAAAAAGAAAAAGAGTACTCAGAGGAGTTGAAGCAG  
ATCAAGATGCTAGAGGAGAAATAAAGAGAGAGAAAAAAGAAAAAAGTACTGTGCCTAATGATGGATGAAATGTAT  
ATTCACAAGATGACTGAATTTGCAAGGAGCCAGTTTCATGGCTATGTGGACTGGTACTGGAGAAATGACAAC  
ACGTTGGCTACCCAGGCTTTGGTCTCATGGTGTGGCTGTAAATGAGTCAAGAAATCCCATAGCATACTTC  
TCATCACAGGATGGATGGATCAGAGAAGGCCAATGTCATTAAGGAGAGCCCTTAGCAGACTTCATGAGTTGGTG  
TCAAAGTATCTCTTACATGTATGACCCCAAGCACTTGGCTATGATAAGGGAAGTGGCGCTGACCTAA  
ATATTAACAACATGAAGCCTTACTTTCATGCTCCTGAGGACCCTACTCAGAAAGTACATGTATGATCCTGGATGCAT  
GCCACATGCTAAAGCTTCTTCGCAATGCCTTTCATCTTCCCTTGAATTTGAGACAGAAGATGAAATAAAGATTA  
AGTGGAAAGTACATCGAAGCCCTCAATGAATCCAAAGAAAAAGAGTCTGCGACTGGGAAACAGCTGAAGATGG  
CTCATTACAATGGCGAAAACAAAAATGAAAGTCCATCTGGCAGCACAGCTCTCAGCAGCAGTGTGGCAGATG  
CTATAGAGTCTGTGAGCAAGGCTTGAATAAGGAGAAATCAAAGGTTGTGCTGCAACTGTGCAAGTTTTTGGCA  
CTGTAGATGCAGCTTTTGATGTACTGAATAGCAGAAATCCTCTTGGAAAAGGTTTCAAAGCCCAATCAAACAA  
CAACTAAAGATCGGGTTGAGACCATTTAAAGCAGGCAGAAATCCATGCTTGGTGGGTTGAAGTCCAGCAGTACA  
ACAAGATGGTGCACCTTCCACACAACCAAAAAGAAAGTCCATATATGGCTTCAATGCTAATGGCAGAAGTGCAC  
TTAACATCTATCATGATCTTGTGAAAGGCCAATGCAACATGCAGATATCTTCTTACCTACAAATTAAGTCAAGG  
ACCCTTGGAGCTTTTCTTTGCTGCCATAAAGAGCAAGAGGTTGGTACAATGATAATCCTAATGCCAGGCAAGTTA  
GAGGAGCATACAAACGCTTACTTGTAAAGGCACCAAGTAAAACAGGCACAGGGAAGCTTTGGTCTTAGACAACA  
CTTACATGCTCAACTCAACCTGATCTGTAAATGTTGCCCAAGATTTGGAAGTACAGTTGGTGGAAAGTGGAGC  
TGCCCGAGAATGATGCTGTTCTGACCTCCCTCATGATGATGAGTCTTCTGAATACAAAGAGGCAGCAATTCATT  
ATATCACAGGATTTGTTGGTGAAGAAAGTAAAGAGAAAAATCACTGCTTGCATGCTCACAGGCATTAACAAGT  
ATGGTGCTGCACATGAATTCATCCATCTAAAAAACAGAGGTGGTCTGCAAAACCATCACAGGCATGGTGTCTG  
TGTGTCTGACCACAGAGAAATGCATTCAGAGAAAGATCAACAAGTGGAGGTCAGCTTCCCGTGGAGCTGGAA  
TCACACTTGTATTTCAACAGAGGTTAGCAAAATGTCAGAGAGAGATCTTTTCCCAACTCCACAGCCACA  
TGTTTGCTACAAGCGTTGAAATGAATCATATTCATTTATTGGTAAAATGGCATCCATTTGGTATAGCAAAGTCA  
GATTTAACCCTTTGCAAGAAGGGGAGCTGAAATAGCAAAAGATGGCAAAATGGTTCGAGCAGCAACTGACTAAA  
TTGATTTCATTTTTATGGTGA

### Séquence traduite (avec transeq)

1  
MCSAVNCSNRFPLGRHDAARLQWVWVNMRRNGWKPTTSSRLCSAHFEEHFTKDLKLVKK  
TAVPTIFSFCHLVKSSCVGVRHKTFAKNSEVSGSSLLFSKIDAGHSHKDDSPVDKLSVSHSAA  
DYSNATLTHEAIGNFVSNLDCQVDTVAVTVDINEDHSYATLSSSTHETPKFIKEDHSYNLASPRSLK  
RKNQATDQILQKYRKKLIESQKSRRLLKNKISTLKDVTVELKKLLISNECASLLESIDEVPKHILLKI  
QQGKTKAKYSEELKQFATLHFYSPKAYDYVRDNFQKALPHSHTIRNWSVSSADPGFVTSFT  
ALKCHVEENKERGKETVFCALMMDMDEMYIHKMTEFAGDQFHGYVDIGTEIDNLTALQALVLMVVA  
VNESWKIPIAYFLITGMDGSEKANVIRELSRLHEVGVKVISLTCDAPTNLAMIRELGADLNINNM  
KPYFMHPEDPTQKVHVILDACHMLKLLRNFASLSLEFETEDGNKIKWKYIEALNELQEKEGLRLG  
NKLKMAHLQWRKQKMKVHLAQLFSSVADAEIEFCEQLKMEEFKGCATVQFLRTVDAADFV  
LNSRNPLGKGFAPKIPIKTTTDRVETILKQAESMLRGLKVQYQNKMVLPHHTKKTAIYGFANGRS  
ALNIYHDLVERPNAPCRYLTLTYKLSQDHLEFFAARARGGHNDNPNARQFRGAYKRLLVRHQVK  
TGTGNLILLNTYMLNSTPASNVARLLEVQLVEVDVPENDAVPDLPHVCSLEYKEAAIHIYITG  
FVVKMKEKITCLPSCQALTDGAAHEFIHLKNRGGLQKPSGMVSVLCTTEKCIQRKITTSGGQ  
LPRGRGITLAISNEVLANCAERDLFPQLHSHMFATSVEMNHILLVKNMASIYYSKVRFNFHARRG  
AEIAKDGMVRRQLTKLIHIFYGE  
>\_2  
"VVQQ"VTLVFL"VGMMQQD"NINGW"DATAGNQQLHPVFAVILKLSMLSPRI"R"KKILQQFQLY  
SHFLVIW"NPLV"G"DTRLLRIRLR"VAPLCCSAKMQDQTVKRLQLTLLACTQLLITAMP"LNMLLA  
TSQLVMTYKLTLL"RWILMRIIWMQL"VPHMKLLSLLKRITATLRLRDL"NGKIRPLTKFYKNTGKS  
LKLNPKSLGV"KTCSPP"RMSQNSRRSC"YPMNVLLCLSQLMKCQSTSC"KYSRGKKQQSTQR  
S"SSLPQCFIPQRTTITTYVTFIRKRVHTVTFPEIGTAAYQLILDLLVHLSQR"NVM"RRIKREKK  
QYV"WMKCFIR"LNQGTSMAMVTLVLEKLTTRVPRVWSSWLWL"MSHGKFP"HTFSSQV  
WMDQRRPMSLGRALADFMRVLSKSSLLHVMPRPVTL"GNLALT"ILTT"SLTSCILRTLRLKYM"  
SWMHATC"SFAMPLHLNLQRKMEIRLSGSTSKPMSNKKKKVCWDWETS"RWLIYNGENKKK  
SIWQHSSSAVWQML"SSVSKA"KWRNSKVLQLCSFCAL"MQLLMY"IAELLEKVSFQPSKQQ  
LKLGLRPF"SRQNPFCV"RSSSTTRWCHFTQPKRRLPYMASLLMAEVHLSIMILLKGMHMHADI  
FLPTN"VRTTWFSLLP"EQEVVTLMIIMPGLSLEEHTNAYL"GTK"KQAQGTVCF"TTITCSTQHLH  
L"MLPEDWKYSWWWKTCPRMMLFLTSLMAYFLNTRKQFIIISQDLW"RR"KRKSPACHAHRH"  
QLMVLHMNSSJ"KTEVCRNHHQAWCLCV"QORNFRERSQVQVSVFPVDVESHLLFPTRC"QI  
VQREIFFHNSTATCLLQALK"IIFIYWLKWHPFGIKASDLTLLEQEGELK"QKMAKWFADN"LN"FIFM  
X  
>\_3  
ELFSSKLF"PFSEF"A\*CSKIETMGGEHETQRLTNNFIPSLQCSF\*RACFHQGFEGKKKSYSSSN  
YLILSSGSEILLCRGKTDQCFE\*EF\*GKWLLSVVQQRCTQ\*KGESS\*QVS\*RALSC\*LQQCHLDS\*  
TCYWQLL\*\*PMSS\*HCCNGGY\*\*GS\*LCNFKFHT\*NS\*Y\*YRGSQDQPCVSEFKTKSGH\*P  
TKIQEKA"N\*IPKY\*AFKQNLHLKGRHRTQEEAVDIQ\*MCFFA\*VN\*\*SAKAHLVENTAGEKNSKV  
LRGVEAVCHNLAFLPKGIRLT\*QFSESAATQSHSKLVQQRIS\*SWIYCGIFHSAEMSCRGE\*R  
ERKRNSHPCNDG\*NVYSQDD\*ICRGPVSVLWLGHWYWRN\*QHVGYPFGFPHGGCCK\*VMENS  
HSILSHRYGWIREDGQCH\*GEP\*QTS\*GWCQSHLSYM\*CPHDQLGYDKGTWR\*PKY\*QHEALLH  
AS\*GPYSESTCDPDCMPHAKASSQCLCIPF\*IDRRVK\*D\*VEVHRSQP\*TPRKRRSATGKQAE  
GSFTMAKTNPSPGSTALLQQCGRCYRVL\*ARLENGGIQRLCCNCVFAHRCRSF\*CTE\*QKS  
SWKRFQSPNQNN\*RSG\*DHFAGRIHASWVEGPAVQQDGTSHNQKEDCHIWLHC"WQKCT\*  
HL\*SC\*KAQCTMQISSYLQIKSGPLGAFLCHCKSKRWSQ\*\*S\*QVAV\*RSIQTPCKAPSENHRH  
ELFASRQHLHAQLNTICKCCPKIGSTVGGSGRARE\*CCS\*PESCMLQSF\*IQRGSNSLYHRICGE  
EDERENHLLAMLTGINN\*WCCT\*IHPKQRWSAETITRHGVCSVDRHEMSEKDHNKWRSASP  
WTWNHCTYFORGVSKLRCRETSFSTTPPHVCYKR\*NEYSYFIQ\*NGIHLV\*QSQI\*PLCKKGS\*NS  
KRWQNGSQDQTT\*IDSLVWX



# Recherche des domaines conservés

## 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
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.

Help | Disclaimer | Write to the Help Desk  
NCBI | NLM | NIH

## Alignement CD Search séquence 2 :

Pas de résultat

## Alignement CD Search séquence 3 :

Pas de résultat

**Protéine reconstituée  
domaine THAP complet**

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-----RFP lgrHDAARLKQVVNMRRNGWKPTTSSRLCSAHFEEHAFTKDLKVKKNL--TAVPTIF 69
1  CSVPGCTNrkkkntrisfhKFP--KDPERRKWLNACKRKDLPPPKNRSVCSLHFEENDFEKISGRRRLkpGAIPTIF 77
  
```

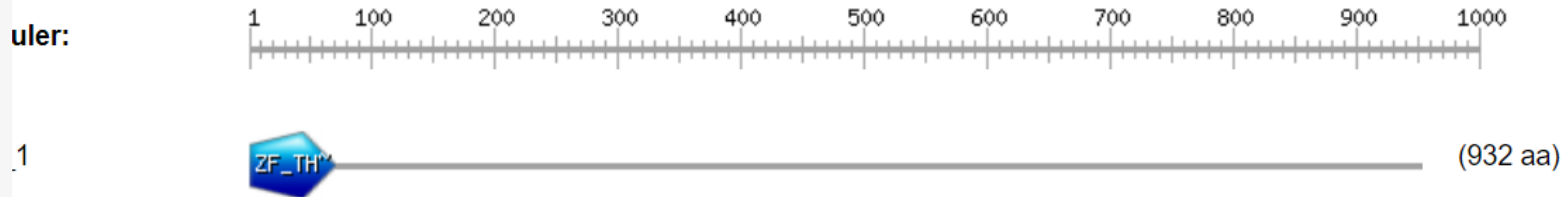
← Protéine reconstituée

← 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 :

1 - 69: score = 12.707

-----MscSAVNCSNRFP1grHDAARLKQwVVMRRNG--WKPTTSSRLC  
SAHFEEHAF-TKDLKVKKNLTAVPTIF

**Domaine THAP complet**

**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]>)]

# 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](#)

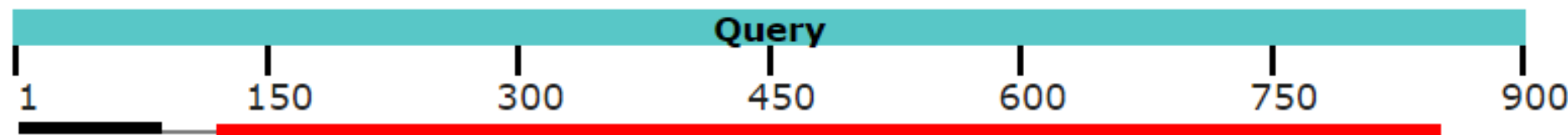
Score	Expect	Method	Identities	Positives	Gaps
298 bits(762)	3e-91	Compositional matrix adjust.	219/760(29%)	377/760(49%)	58/760(7%)
Query 122	EDHNYSLKPTLTIGAEKLAEVQQLQVSKKRLI--SVKNYRMKIKRKRGLRIDALVEEKL				179
Sbjct 178	EDH+Y+L +P ++ + A Q+LQ +K+L S K+ R+ K L+ + +++KL				236
Query 180	L-SEETECLLRAQFSDFK---WELYNWREDEYSAEKQFACCTLYLCSKSSKVVYVRKILK				235
Sbjct 237	L S E LL + K ++ ++T +YS E+KQFA TL+ S K YDYVR +				296
Query 236	--LPHSSILRTHLSKCPSPGFNSNIFSLQRRVENGQDLYQ--YCSLLIKSMPKQQLQ				291
Sbjct 297	LPHS +R W S PGF F+ L+ VE + + C+L++ M + + +				356
Query 292	MDPSSHSLQGFMDFLGKLDADETPLASETVLLMAVGIFGHWRTPLGYFFVNRASGYLQA				351
Sbjct 357	+ + G++D G G++D LA++ ++LM V + W+ P+ YF + G +A				411
Query 352	QLRLTIGKLSDIGITVLAVTSDATAHSVQMAKALGIHIDGDMKCTFQHPSSSSQQIAY				411
Sbjct 412	++R ++ +L ++G+ V+++T DA ++ M + LG ++ ++MK F HP +Q++				471
Query 412	FFDSCHELLRLIRNAFQNFQSIQFINGIA-HWQHLVELVALEEQE-LSNMRIPSTLANLK				469
Sbjct 531	+ + + +G W+++ L L+E+E L ++ +				531
Query 528	/NSATQLFSESVASALEYL-LSLDLPPFQNCIGTIHFLRLINLFDIFNSRNCYG				528
Sbjct 591	/+ A QLFS SVA A+E+ L + F+ C T+ FLR ++ FD+ NSRN G				591

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/> _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	SCSAVGCSTRDVTLSRERGLSFHQFPT---DTIQRSKWIRAVNVRVDRSCKKIWIPIGPGAI				60
Sbjct 2	SCSAV CS R FP D + +W+ + R W P +				41
Query 61	LCSKHFQESDFESYGIRRKLKKGAVPSV				88
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

**TSSRLCSAHFEEHAFTKDLKVKNLTA**VPTIFSFCHLVKSSC

# 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

SMART SETUP FAQ ABOUT GLOSSARY WHAT'S NEW FEEDBACK keywords... Search SMART

Domains within *Homo sapiens* protein THAP9\_HUMAN (Q9H5L6)  
DNA transposase THAP9

+ = - Introns SAVE Alternative representations: 1 / 2 << >>

DM3

0 100 200 300 400 500 600 700

Information Architecture Interactions Orthology

Length 903 aa  
Source database UniProt  
Identifiers THAP9\_HUMAN, Q9H5L6, ENSP00000305533.5, ENSP00000305533, B3KRE2, Q59AC9  
Source gene ENSG00000168152  
Alternative splicing D6R957\_HUMAN, D6RCT5\_HUMAN, THAP9\_HUMAN, D6REM3\_HUMAN, F2Z371\_HUMAN, H0Y9F3\_HUMAN

The SMART diagram above represents a summary of the results shown below. Domains with scores less significant than established cutoffs are not shown in the diagram. Features are also not shown when two or more occupy the same piece of sequence; the priority for display is given by SMART > PFAM > PROSPERO repeats > Signal peptide > Transmembrane > Coiled coil > Unstructured regions > Low complexity. In either case, features not shown in the above diagram are marked as 'overlap' in the right side table below.

Confidently predicted domains, repeats, motifs and features:

Name	Start ▲	End	E-value
THAP	3	95	5.2e-18
DM3	25	94	0.00000113

Features NOT shown in the diagram: ?

There are no hidden domains or features present.

Click on a row to highlight the feature in the diagram above. Click the feature name for more information.

**DM3 domain**

This is a SMART **DM3** domain ([full annotation](#)).

Position: 25 to 94  
E-value: 1.12596537868542e-06 (HMMER2)

SMART ACC: SM000692  
Definition: Zinc finger domain in CG10631, C. elegans LIN-15B and human P52rIPK.  
Description:

Interpro abstract (IPR006612): The THAP-type zinc finger (consensus: C-x(2,4)-C-x(35,50)-C-x(2)-H) is an ~90-residue domain restricted to animals, which is shared between the THAP family of cellular DNA-binding proteins, ...([full abstract](#))

DM3 domain sequence (70 aa):

Submit to BLAST Align with the SMART alignment Copy to clipboard

FHQFPTDTIQRSKWIRAVNRVDPKSKKIWIIPGPGAILCSKHEQESDFESYGTBRKIKKGA  
VPEVGLVYKTR

# Reconstitution de la proteine à partir de l'ADN

## utilisation positions publiées

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:  to:

Sélection portion séquence nucléique

Transcription de la séquence par TRANSEQ

Frame(s) to translate

1

2

3

Forward three frames

# CD SEARCH

--> Localiser le domaine THAP:  
on retrouve la même zone du domaine:  
il manque toujours la partie N-term.

Conserved domains on [lcl|seqsig\_SLEIK\_654b12c0b9e7209c4d44837699a6fdf5] View Concise Results ?  
199000-203000\_1 Zebrafish DNA sequence from clone DKEY-148A12 in linkage group 2, complete sequence

Graphical summary  Zoom to residue level show extra options >

Query seq. ... No conserved domains have been identified for this query sequence ...

Search for similar domain architectures ? Refine search ?

List of domain hits

Name	Accession	Description	Interval	E-value
... No conserved domains have been identified for this query sequence ...				

[seqsig\_HLKLN\_915b1b45f905a829e7674a84c901798a]  
quence from clone DKEY-148A12 in linkage group 2, complete sequence

Zoom to residue level show extra options >

... No conserved domains have been identified for this query sequence ...

Search for similar domain architectures ? Refine search ?

Accession	Description	Interval
... No conserved domains have been identified for this query sequence ...		

Conserved domains on [lcl|seqsig\_T\*N\*T\_ed8274f3a0822e7a5b5bede576ff6750] View Concise Results ?  
199000-203000\_3 Zebrafish DNA sequence from clone DKEY-148A12 in linkage group 2, complete sequence

Graphical summary  Zoom to residue level show extra options >

Query seq. Specific hits Superfamilies

THAP THAP sup Tnp\_P\_element DUF4

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 ...	1113-1330	4.32e-16
[+] THAP	pfam05485	THAP domain; The THAP domain is a putative DNA-binding domain (DBD) and probably also binds a ...	356-412	3.33e-08
[+] DUF4201 super family	cl25515	Domain of unknown function (DUF4201); This is a family of coiled-coil proteins from eukaryotes. ...	1088-1134	6.44e-03

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 FCSFP1grHDAARLKQWVVMRRNGWKPTTSSRLCSAHFEEHAF TKDLKVKKNLQNN 412

19 HKFP---KDPERRKKWLNACKRKDLPPPKNSRVCSLHFEENDFEKISGGRRRLKPGa 72

Même zone