Synapomorphic variations in the THAP domains of human THAP proteins and their homologs

Abstract

The THAP (Thanatos-associated protein) domain is a DNA-binding domain which binds DNA via a zinc coordinating C2CH motif. Although THAP domains share a conserved structural fold, they bind different DNA sequences in different THAP proteins, which in turn perform distinct cellular functions. In this study, we investigate (using multiple sequence alignment, *in silico* motif and secondary structure prediction) THAP domain conservation within the homologs of the human THAP (hTHAP) protein family. We report that there is significant variation in sequence and predicted secondary structure elements across hTHAP homologs. Interestingly, we report that the THAP domain can be either longer or shorter than the conventional 90 residues and the amino terminal C2CH motif within the THAP domain serves as a hotspot for insertion or deletion. Our results lay the foundation for future studies which will further our understanding of the evolution of THAP domain and regulation of its function.

Introduction

The THAP (Thanatos-associated protein) domain is a DNA-binding domain which is reported to be 80-90 amino acid residues long and mostly located at the amino terminal end of the corresponding protein. THAP domain-containing proteins have been recently reported in diverse groups of animals such as humans, chicken, zebrafish, *C. elegans* and *Drosophila melanogaster* (1–4). No known or predicted proteins containing THAP domains have been found in plants, yeast, fungi or bacteria, suggesting that the THAP domain is a novel protein domain restricted to animals.

THAP domain-containing proteins are involved in diverse cellular functions. For example, *Drosophila* P element transposase (DmTNP), the cell-cycle transcription factor E2F6 (1) and the transcriptional corepressor CtBP-1 in *C. elegans* (5). The human THAP protein family is a group of twelve proteins (hTHAP0-hTHAP11) which are all characterised by amino-terminal THAP domains (6) and are implicated in cell-cycle regulation, apoptosis, angiogenesis, pluripotency of stem cells (7–11). THAP family members have also been implicated in a variety of human diseases including heart disease (12), torsional dystonia (13) angiogenesis and cancer (14, 15).

Although the THAP domains of THAP proteins share low primary sequence identity (~10% sequence identity in human THAP proteins) as is typical of large DNA-binding protein families (16), there is strong conservation of the overall protein fold (Suppl. Fig. 1A), as well as secondary structure elements namely the characteristic β-α-β fold (1, 3, 6, 17, 18), with four loops (L1-L4) flanking and interconnecting the β sheets and the α helix. Recent structural studies illustrate how THAP proteins recognize specific DNA sites through bipartite recognition of adjacent major and minor grooves by specific residues (18): the β sheet interacts with the DNA major groove (GC rich sequence in DmTNP) while the carboxy terminal loop 4 (L4) interacts by π-stacking interactions with the DNA minor groove (AT rich sequence in DmTNP) via basic amino acid residues (3, 18, 19).

THAP domains, with more than 300 identified members, are the second most prevalent zinc-coordinating DNA-binding domains after the classical C_2H_2 class of zinc fingers (1, 6, 16, 20). The conserved sequence signatures in the THAP domain are: (i) C_2CH (consensus: Cys-$X_{2–4}$-Cys-$X_{35–50}$-Cys-X_2-His) zinc-coordinating motif, which is significantly different from the classical C2H2 zinc finger motifs (20) (ii) four invariant residues, Pro,Trp, Phr and Pro (iii) a consensus carboxy terminal AVPTIF box that marks the end of the THAP domain (6, 19). The invariant residues (named in ii) are a part of the β-α-β secondary structural fold. For example, the first conserved Pro is generally found at the beginning of Loop 2, Trp is found in the centre of helix1, Phe is generally found at the beginning of loop 3 (in 2JTG, it is at the beginning of beta 3) and the second conserved Pro is a part of the AVPTIF motif in L4 (Suppl. Fig. 1A). It has been experimentally demonstrated that the conserved sequence signatures and the consensus $\beta\alpha\beta$ structural fold (3, 18, 19) are indispensable for DNA binding by the THAP domain.

Not much is known about the importance of the four loops (L1-L4) in the β-α-β secondary structural fold. L4, which forms the carboxy terminal end of the THAP domain, contains basic residues (Arg65, Arg66 and Arg67 in DmTNP, Arg65 in hTHAP1) involved in binding the minor groove of DNA (17, 18) and the consensus AVPTIF motif. L4 is reported to be flexible unlike the rigid central core of the THAP domain (3, 8, 18, 19) and has been observed to undergo structural changes after binding to DNA in hTHAP1 (17). Interestingly, L4 in different THAP proteins is characterised by variability in length and primary sequence (3, 8, 18, 19).

The THAP domain is an example of a domain shared between DNA binding proteins and active DNA transposases. Some other examples of such shared domains include the DNA-binding domain of the BED zinc finger, which is shared by both chromatin-boundary element-binding proteins BEAF and DREF (Feschotte and Pritham 2007). It is interesting to note that each THAP protein [e.g., DmTNP (18), hTHAP1 (1), hTHAP5(21) and hTHAP11 (19, 22)] appears to bind distinct DNA sites (Suppl. Fig. 1B) despite overall similar structure (Suppl. Fig. 1C). This is similar to the basic leucine zipper motif (bZIP) containing proteins wherein different bZIP proteins bind different DNA sites (23). The difference in DNA binding specificities of THAP proteins is speculated to be due to the variation in the amino acid residues that form β sheets (which directly interacts with DNA), the number and sequence of amino acids before the first C of the C2CH motif and the length and composition of loop 4 (18, 19).

Till date, there is no comprehensive analysis of the possible diversity in sequence and structural elements in the THAP domains of THAP proteins and their homologs. In this study, we identify possible synapomorphic (sequence and structure) variations in the hTHAP proteins and their homologs using multiple sequence alignment, *in silico* secondary structure and motif prediction. We report conserved amino acid residues in the THAP domain in addition to the ones that are already reported. We identify interesting THAP protein homologs with THAP domains that are significantly longer and shorter than the conventional ~90 residue long THAP domain observed in available structures. Identification of a few hotspots for insertions and deletions within the THAP domain challenges the existing paradigms about this domain. This study opens avenues to investigate the evolutionary adaptations of a domain restricted to kingdom animalia.

Methods

Curation of THAP protein sequences

There are multiple databases which curate and document protein sequences based on either the families that the protein belongs to (Pfam) (24) or the common patterns in the protein sequence (PROSITE) (25). PROSITE was chosen for this study because the analysis is based on the sequence patterns of various THAP domains. The search string "THAP domain" identified PS50950 PROSITE documentation with 45 true positive protein sequences containing THAP type sequence patterns (Additional file 1). Of the 45 proteins, *C.elegans* CDC14, CTBP1, Lin36, Lin15B and *Drosophila* P element transposase (DmTNP) were THAP domain containing proteins which are not homologs of any human THAP protein. Thus, they are referred to as "Other THAP proteins" in this study.

Identification of the THAP domain in the curated protein sequences

Each human THAP protein was aligned with its homologs using Clustal Omega (26). Briefly, Clustal Omega uses HMM models to align multiple sequences and identify identical residues or residues with similar chemical properties, at a position, in the aligned set of sequences. The THAP domain of each human THAP protein homolog was manually identified by using the conserved AVPTIF motif as the carboxy terminal boundary of the domain. These THAP domain sequences for each homolog were stored in a separate word file.

Identification of conserved residues in the THAP domain

GLAM2 (27) was used to identify a gapped motif in the THAP domain. Briefly, the THAP domain protein sequence from each human THAP protein and its homologs were submitted to GLAM2. It uses an extension of gapless Gibbs sampling algorithm which examines the sequences provided by the user and gives an alignment of different segments of these sequences. This alignment is scored based on position- specific insertion and deletion possibilities. The conserved residues revealed by GLAM2 motifs were validated by multiple sequence alignment (MSA) of the THAP domains of each human THAP protein and its homologs generated using Clustal Omega (26). The GLAM2 and MSA results were carefully analyzed to record the conservation of residues within the THAP domain (C2CH, P, W, F, P, F). The homologs that did not have even one of the above-mentioned conserved residues or had replaced the conserved residues with some other amino acid were highlighted in the results.

THAP domain secondary structure prediction

The secondary structure elements of each THAP domain was predicted using JPRED (28), PSIPRED (29), SPIDER3 (30). Briefly, JPRED constructs a MSA using PSI-BLAST (31) for individual input sequences and uses it to predict local secondary structure using Jnet (32). Additionally, JPRED does a PDB search to identify possible structural homologs of the submitted

protein sequences. PSIPRED uses two feed forward neural networks to analyse the PSI-BLAST output. SPIDER3 uses bidirectional recurrent neural networks which capture non- local interactions to accurately predict the secondary structures of the given protein sequences.

Sequence Curation of human THAP protein homologs

PROTEIN database within the NCBI databank was used to extract the protein sequences of human THAP protein homologs. For each human THAP protein, a keyword search was performed using the protein name (for example, THAP1). Multiple protein sequences of the same protein were available for each organism. Thus, to avoid redundancy among the protein sequences, only the longest sequence was chosen as a THAP protein homolog. The entries which had partial or [PREDICTED] or hypothetical or uncharacterized protein in their names were excluded from the study.

Results

Conservation of invariant residues in the THAP domain of hTHAP family proteins and other THAP domain containing proteins

THAP domain-containing proteins (45 proteins identified by PROSITE) were divided into two groups (a) members of human THAP family (hTHAP0-hTHAP11) (b) Other THAP proteins (*C.elegans* CDC14, CTBP1, Lin36, Lin15B, DmTNP). Zebrafish E2F6 and *C. elegans* Him-17, which were earlier reported to contain THAP domains (1, 3) were also added to this group.

The THAP domain sequence in all the group a and b proteins was determined by the presence of the C2CH motif at the amino terminus and an AVPTIF box at the carboxy terminus, as described in the methods. Interestingly, more than one putative THAP domain was reported for *C. elegans* CDC14 (two THAP domains), Lin15B (two THAP domains) and Him-17 (six THAP domains) (1, 3). The two CDC14 THAP domains and six Him-17 THAP domains are respectively very different from each other except for the consensus invariant residues. Thus, only putative THAP domains, which retained the conserved Pro (required for DNA binding ; (3, 18, 19) of the AVPTIF box, was included in this study [one domain each for *C. elegans* CDC14 and Lin15B and four domains for Him-17 (1, 2, 3, 4)].

MSA (ClustalW) was independently performed for THAP domain sequences within each group (a and b), after which conserved residues were identified; these included the five residues namely 3 Cys and His of the C2CH motif and Pro of the AVPTIF motif, which have been earlier reported to be functionally indispensable for DNA binding by the THAP domain (3, 18, 19) as seen in Fig. 1A and 1B.

The THAP domain sequences for group a and b were independently submitted to GLAM2 which identifies underlying gapped motifs (as PWMs) in the input sequences after aligning them. (Highlighted by boxes in Fig. 1C and 1D). The C2CH motif and the AVPTIF box are two functionally distinct motifs in the THAP domain, wherein the C2CH motif coordinates the zinc ion and the AVPTIF box directly interacts with DNA. These two motifs also form different

structural folds, i.e., the C2CH motif folds into the $\beta\alpha\beta$ fold whereas the AVPTIF box forms a loop. Different THAP proteins have different inter motif spacers. For example, the intermotif spacer between the His of the C2CH motif and the Ala of the AVPTIF motif is 18 residues long in hTHAP1 and 24 residues in DmTNP. These GLAM2 identified gapped motifs for group a and b proteins were then analysed visually to record the conservation of specific residues within the THAP domain. The five invariant residues (Cys5, Cys10, Cys54, His57, Pro78; residue numbers correspond to hTHAP1, green circles in Fig. 1C and 1D) as well as some residues of unknown functional significance (Pro26, Trp36, Phe58; residue numbers correspond to hTHAP1; red circles, Fig. 1C and 1D) which were earlier reported to be conserved in human THAP proteins (6, 19) were found to be conserved. Interestingly, these residues were also conserved in group b proteins (Pro29, Trp39 and Phe61 in *C.elegans* CtBP -1; red circles in Fig. 1D).

In addition to these, three other residues (Phe22, Arg42, Leu72; residue numbers correspond to hTHAP1; blue circles, Fig. 1C) were found to be conserved in group a proteins. The Phe and Arg are a part of the C2CH zinc coordinating motif and Leu is within L4. However, the group b proteins only had a conserved Arg (Arg45 in *C.elegans* CtBP-1) as seen in Fig. 1D (blue circle). Interestingly, in the both group a and b proteins, the Phe of the AVPTIF motif was not seen to be strictly conserved (purple circle in Fig. 1C and D) and was replaced by Ser (hTHAP9), His (hTHAP10), Glu (CtBP-1), Val (DmTNP) or Pro (zE2F6).

Consensus secondary structural elements amongst many THAP proteins.

The THAP domain has been experimentally demonstrated to fold into a $\beta\alpha\beta$ (L1-β1- L2- α1- L3-β2-L4) secondary structural fold in hTHAP1 (2JTG, 2KO0), hTHAP2 (2D8R), hTHAP11 (2LAU), DmTNP (3KDE) and C. elegans CtBP-1 (2JM3). The secondary structure predictions of the THAP domains of each of the twelve human THAP proteins using JPRED, PSIPRED and SPIDER3 agree with the experimentally identified $\beta\alpha\beta$ secondary structural fold. Fig. 2A and 2B displays results from JPRED. Surprisingly, an additional β sheet (β1, green) of about five amino acid residues was predicted at the amino terminal region and another β sheet (β4, green) of length more than or equal to three amino acids residues was predicted within the L4 regions in hTHAP1, hTHAP4, hTHAP9 (Fig. 2A) albeit with a very low confidence score. The predicted β4 in hTHAP2, hTHAP5, hTHAP6, hTHAP8, hTHAP10 was not considered as it was less than 3 residues long and a typical β sheet is made of 3 -10 residues (33).

Surprisingly, in the "other THAP protein" group, only C.elegans CtBP-1, DmTNP and zE2F6 had a predicted $\beta\alpha\beta$ secondary structural fold in their THAP domains (Fig. 2B). Cdc14, Lin-15B and Him17 (3) had an additional short helix between the helix and β sheet whereas Lin-36, Him17(1), Him17(2), Him17(4) were predicted to have distinct structural folds with extra helices and sheets (Fig. 2C).

Interestingly, the length of β1(5 residues) and α1(10 residues) was conserved in the human THAP family (Fig. 2A) as well as CtBP-1, DmTNP and zE2F6 (Fig. 2B). On the other hand, the length of β2 and β3 varied slightly (Fig. 2). For example, β2 consists of five (hTHAP3, hTHAP5, DmTNP) or four (other hTHAP proteins, CtBP-1, zE2F6) residues while β3 consists of two

(DmTNP), three (hTHAP2, hTHAP8, hTHAP9), four (zE2F6, hTHAP1, hTHAP3, hTHAP6), five (hTHAP7, hTHAP0, hTHAP 5, hTHAP10) or six (hTHAP11, hTHAP4, CtBP-1) residues.

Loop 4, which interacts with DNA minor groove via basic residues, can be of diverse length

The length and sequence of L4 has been speculated to be important for different DNA binding specificities in different human THAP proteins (3, 18, 19). This is more significant in the light of structural studies that demonstrate direct interactions between the basic residues in L4 and the pyrimidine ring of a thymine base in their respective DNA binding sites (Arg65, Arg66 and Arg67 in DmTNP, Arg65 in hTHAP1) (17, 18).

DmTNP has a stretch of four consecutive basic amino acids (Lys64, Arg65, Arg66 and Arg67; Fig. 2B, highlighted in blue) in L4. In hTHAP9, three consecutive basic residues (Arg77, Arg78, Lys79) were predicted in L4-β4 (Fig. 2A, highlighted in blue). However, the L4 may also contain two consecutive basic residues as observed in hTHAP1 (Lys64, Arg65), hTHAP2 (Lys69, Lys70), hTHAP3 (Arg70, Lys71), hTHAP4 (Lys67, Arg68), hTHAP6 (Lys67, Lys68), hTHAP11 (Arg69, Lys70), CtBP-1 (Lys66, Lys67) or one basic residue as seen in hTHAP0, hTHAP5, hTHAP7, hTHAP8, hTHAP10 (highlighted in blue, Fig. 2). zE2F6 does not have a single basic residue in the predicted L4 (Fig. 2B). This raises the possibility that one basic residue in L4 might suffice for minor groove DNA interaction if complemented with another basic residue from another structural fold to form a positively charged surface around the DNA as has been previously suggested (3).

It has been speculated that the length of L4 determines DNA binding affinity: the longer the L4, the tighter the binding (18, 19). However, there are no experimental reports establishing this claim. Thus, it was interesting to observe significantly different lengths of L4 within the human THAP family of proteins (Table 1). It is tempting to hypothesize that hTHAP10 may have the strongest while hTHAP4 may have the weakest interaction with DNA.

Diversity in the THAP domain features in the homologs of hTHAP proteins

The THAP domain is a novel protein domain restricted to animals. The THAP domains of human THAP proteins appear to share structural similarity (Fig.2) and conserved invariant residues (Fig.1). Studying the THAP domain features in the homologs of human THAP proteins may provide insights into the evolution of THAP domains in humans. Thus, protein sequences of the homologs of each human THAP protein were extracted from the NCBI PROTEIN database. Several interesting observations were made:

a. The THAP domain can be longer or shorter than 90 residues

The THAP domain has been reported to be between 80-90 residues long, as demonstrated by structural analysis of the THAP domains of hTHAP1 (2JTG, 2KO0), hTHAP2 (2D8R), hTHAP11 (2LAU), DmTNP (3KDE), CtBP (2JMR. However, analysis of individual hTHAP protein homologs in different organisms revealed proteins with variable THAP domain length. For the purpose of this study, THAP domains of hTHAP protein homologs which are longer than 100

residues have been termed “Long THAP domains” (Table 2) and those with length less than or equal to 50 residues have been termed “Short THAP domains” (Table 3). It was interesting to observe that these homologs with long THAP domains often had high sequence similarity, as noted below:

1) hTHAP1: 7 (6 mammalian, 1 aves) out of 16 homologs had 152 residue long THAP domain (Fig. 3A). All 6 mammalian homologs had identical sequences except for two amino acid residues (Suppl. Fig. 2A). However, the aves homolog (Lonchura) varied considerably from the mammalian homologs (Suppl Fig. 3).
2) hTHAP4: 2 (mammalian) out of 11 homologs had 179 residue long identical THAP domains (Fig. 3B. Suppl. Fig. 2A).
3) hTHAP5: 2 (avian) out of 11 homologs had 128 residue long THAP domain (Fig. 3C) which are identical except for two residues (Suppl. Fig. 2A).
4) hTHAP7: (a) 2 (1 Actinopterygii, 1 mammalian) out of 22 homologs had 104 residues long THAP domain. Both these homologs had identical THAP domain sequences (b) 2 (mammalian) out of 22 homologs had 125 residue long identical THAP domains (Fig. 3D, Suppl. Fig. 2A).
5) hTHAP8: 2 (mammalian) out of 6 homologs had 116 residue long identical THAP domains (Fig. 3E, Suppl. Fig. 2A).
6) hTHAP9: 2 (1 insecta, 1 mammalian) out of 9 homologs had 127 residue long THAP domains (Fig. 3F), both of which were significantly different from each other (Suppl. Fig. 2A, highlighted in grey).

Some interesting observations about hTHAP protein homologs with short THAP domains include:

1. hTHAP3: Chinese alligator (Alligator sinensis) has 39 residue long THAP domain and black flying fox (Pteropus alecto) has 42 residue long THAP domain with deletion before the conserved F59.
2. hTHAP4: 2 [mammalian: white-tailed deer (Odocoileus virginianus texanus) and water buffalo (Bubalus bubalis)] out of 3 homologs has 50 residue long THAP domains (Fig. 3G, Suppl. Fig. 2B) with identical sequences except for two residues (Suppl. Fig. 2B), deletion before the conserved W38; thirteen-lined ground squirrel (Ictidomys tridecemlineatus) have a 31 residue long THAP domain with deletion before conserved H59.
3. hTHAP5: 10 [7 aves {medium ground finch (Geospiza fortis), barn owl (Tyto alba alba), , rock dove (Columba livia), peregrine falcon (Falco peregrinus), saker falcon (Falco cherrug), chuck-will's-widow (**Antrostomus carolinensis**) and wild turkey (Meleagris gallopavo)}, 1 reptilia (chinese softshell turtle; Pelodiscus sinensis), 1 actinopterygii (nile tilapia; Oreochromis niloticus), 1 chondrichthyes (thorny skate; Amblyraja radiata), out of 11 homologs had 42 residue long THAP domain (Fig. 3H, Suppl. Fig. 2B) which had 70% identity and deletion before conserved C57 and common pill bug/potato bug (Armadillidium vulgare) have a 46 residue long THAP domain .

4. hTHAP6: Both mammalian homologs (little brown bat (Myotis Lucifugus) and cheetah (Acinonyx jubatus) had 48 residue long THAP domain (Fig. 3I) which were identical except for two residues (Suppl. Fig. 2B) with deletion before the conserved C62
5. hTHAP8: 3 (mammalian, reptilia, aves) out of 4 homologs had 42 residue long THAP domain (Fig. 3J) which had 74% identity (Suppl. Fig. 2B) with deletion before the conserved C58.The prairie vole (**Microtus ochrogaster**) was also found to have 45 residue long THAP domain with deletion before conserved C58.

These interesting similarities in length and sequence of the THAP domains amongst the human THAP protein homologs led us to ask if the homologs with identical or similar THAP domains had sequence similarity beyond the THAP domain. That is, if the homologs with identical THAP domains were identical across the entire length of the protein. We could identify examples of three different possibilities. These are as follows:

(i) *Sequence similarity across the length of the entire protein*. e.g., 6 mammalian 152 residue long THAP domain containing THAP1 homologs, 2 mammalian 179 residue long THAP domain containing THAP4 homologs, 2 mammalian 128 residue long THAP5 THAP domain containing homologs, 2 mammalian 125 residue long THAP domain containing THAP7 homologs, both (1 mammalian, 1 actinopterygii) THAP7 homologs with 104 residues long THAP domain, 2 mammalian THAP8 homologs with 116 residue long THAP domain, 2 mammalian THAP4 homologs with 50 residue long THAP domain, 7 avian 42 residues long THAP domain containing THAP5 homologs and 2 mammalian 48 residue long THAP domain containing THAP6 homologs.

(ii) *Sequence similarity only within the THAP domain*. e.g., 3 THAP8 homologs (1 mammalian, 1 avian and 1 reptilian) with 42 residue long THAP domain.

(iii) *No sequence similarity across the length of the protein, i.e., within THAP domain and beyond THAP domain*. E.g., THAP9 homologs with 127 residue (1 mammalian, 1 insecta) long THAP domain.

b. Length of the THAP domain has no correlation with total protein length

We then asked if the length of a particular THAP domain depended on the length of the corresponding full-length protein. However, no correlation was observed between the THAP domain length and the total protein length. For example, the full length THAP proteins with long THAP domains were not longer than the proteins with either short or canonical ~90 residue longTHAP domains (THAP3, THAP4, THAP5, THAP6, THAP7, THAP8 in Fig. 4).

c. The amino terminal end of the THAP domain is a probable hot spot for insertions or deletions within the THAP domain

The diverse lengths of THAP domains in hTHAP homologs led us to look for a region within the domain which was conserved across the THAP domains of different lengths. GLAM2 was used to separately align the set of long or canonical (90 residue) THAP domains belonging to the homologs of each hTHAP protein. Comparison of gapped motifs of all the long THAP domains (Fig. 5A) and their canonical ~90 residue THAP domains for each hTHAP protein homolog (Fig. 5B) revealed that the carboxy terminal region was found to be more or less similar between the long THAP domains and the ~90 residue long THAP domains.

On the other hand, it was interesting to observe that the amino terminal region of the THAP domain, i.e the region from 1st to the conserved Trp residue within the C2CH motif, served as a probable hotspot for insertions within the longer THAP domain (Fig. 5A). These long THAP domain containing homologs had insertions immediately after the 1st Met or at specific positions such as before (in hTHAP1, hTHAP2, hTHAP3, hTHAP4, hTHAP6, hTHAP7) or immediately after (hTHAP5) the conserved Loop2 Pro, immediately before conserved Trp in $\alpha 1$ (hTHAP8, hTHAP9, hTHAP11) or before conserved Cys in $\beta 3$ (hTHAP10).

The amino terminal region, in addition to being a hot spot for insertions, also served as a hotspot for deletions. Most short THAP domain-containing homologs of hTHAP3, hTHAP4, hTHAP5, hTHAP6 and hTHAP8 proteins were found to have a deletion before the conserved CH of the C2CH motif (loop 4). These short THAP domain-containing homologs thus resemble the earlier described DM3 domain. It is a truncated THAP domain which has a deletion of the first 20 residues at the amino terminal region (6).

It was also interesting to see that even the homologs of hTHAP proteins with longTHAP domains (Fig. 5A) and short THAP domains had at least one basic amino acid in the predicted L4 region.

d. Insertions in the long THAP domain containing hTHAP protein homologs do not have a consensus sequence and secondary structure fold

The gapped motifs generated by GLAM2 for the long THAP domain sequences of 15 hTHAP1 homologs shows a lot of variation at the amino terminus (place where insertions are seen within the THAP domain) (Fig. 5A) as the insertion sequences differ (10% similarity) amongst themselves (Suppl. Fig. 4). Based on the variations in the sequences of insertions in the long THAP domains, it was logical to check if there was a consensus structural fold that was formed by these insertions as that might suggest a common functional role of the long THAP domain containing proteins. However, insertions in each long THAP domain had different secondary structure, despite being the homologs of the same hTHAP protein. For example, the THAP domains of Lonchura and Monodon homologs of hTHAP1 had different predicted secondary structural folds despite being of the same length (Suppl. Fig. 4).

However, the long THAP domains had some similar features. Specific residues were conserved (Fig. 5A) within the insertion sequences. For instance, in homologs of hTHAP0 (One Phe, one Gly, one Arg, one Pro one Asn), hTHAP1 (Three Arg, one Asn, one Gly, one His, one Pro), hTHAP2 (One Gly, one Val), hTHAP3 (One Ala), hTHAP4 (Three Arg, three Gly, two Pro,

one Ser, one Asn), hTHAP5 (One Ser, one Leu), hTHAP6 (One Phe), hTHAP8 (One Asp), hTHAP9 (One Arg). As only two homologs of hTHAP10 and hTHAP11 had longTHAP domains, we cannot comment on any conserved residues in the insertion sequences of these THAP domains.

e. Variations in the C2CH motif. It is expected that the C2CH motif of THAP domains would be completely conserved in all THAP domains, despite other variations in secondary structure elements and domain length. This is because of the role of this zinc-coordinating motif in DNA binding. However, it was surprising to observe that this motif was not conserved in many THAP domains.

Long THAP domains: Most long THAP domains did not show conservation of the amino-terminal C2CH zinc coordinating motif. Most either lacked two Cys residues (except hTHAP2) or have at least one Cys residue (in hTHAP3 and hTHAP7) within the first 40 residues of the C2CH motif (Fig. 5A).

90 residue THAP domains: It was also observed that the ~90 residue long THAP domain in certain homologs did not align with their hTHAP counterpart. Thus, we aligned these homologs separately (Suppl. Fig. 2c) and found that some of these were completely identical to each other. We could identify examples of four types of variations in the C2CH motif in these hTHAP protein homologs as mentioned below.

(i) *2 Cys in amino-terminal region, which do not align with C5 and C10 in hTHAP homolog*: hTHAP3 homologs in (1) desert tortoise (*Gopherus evgoodei*), painted turtle (*Chrysemys picta bellii*), pinta island tortoise (*Chelonoidis abingdonii*), three-toed box turtle (*Terrapene carolina triunguis*) are completely identical except for eight residues in desert tortoise [THAP3(1) Suppl. Fig. 2C] (2) hawaiian monk seal (*Neomonachus schauinslandi*), weddell seal (*Leptonychotes weddellis*) are completely identical to each other [THAP3(2) Suppl. Fig. 2C]
(ii) *Lacks 1st Cys in the first 6 residues*: hTHAP6 homologs in desert tortoise (*Gopherus evgoodei*), pinta island tortoise (*Chelonoidis abingdonii*) had completely identical THAP domains except for 3 residues. [THAP6(2) Suppl. Fig. 2C]
(iii) Lack 2nd Cys between 6th to 12th residues: hTHAP3 homolog in *Terraprene* (three-toed box turtle), hTHAP1 homologs of narrow-ridged finless porpoise (*Neophocaena asiaeorientalis*), Pacific white-sided dolphin (*Lagenorhynchus obliquidens),* Beluga whale (*Delphinapterus leucas*), Vaquita (*Phocoena sinus*), narwhal (*Monodon monoceros*) and long-finned pilot whale (*Globicephala melas*).
(iv) Lack both Cys in the first 12 residues: hTHAP6 homologs in wild Bactrian camel (*Camelus ferus*) and alpaca (*Vicugna pacos*) [THAP6(1) Suppl. Fig. 2C] and hTHAP4 homologs in cheetah (*Acinonyx jubatus*) and cat (*Felis catus*) [THAP4(1) Suppl. Fig. 2C]; hTHAP9 homologs in rohu (Labeo rohita), treeshrew (Tupaia chinensis), Okarito kiwi (*Apteryx rowi*), japanese quail

(Coturnix japonica), Kanglang fish (*Anabarilius grahami*), springtail (Folsomia candida), nine-banded armadillo (Dasypus novemcinctus) and bagworm moths (Eumeta japonica).

These observed variations within the C2CH motif were quite unexpected. It suggests the possibility of altered or modified functions of this motif, other than DNA binding, in the corresponding THAP protein homologs. It is interesting to note that closely related organisms that occupy similar habitats (e.g., hTHAP1 homologs in different whales), sometimes had similar variations in the C2CH motifs.

Overall, we demonstrate that there is considerable diversity amongst the THAP domains in different organisms. There are variations in domain length, secondary structure elements as well as the C2CH motif. These variant THAP domains need to be studied experimentally to understand how and why they have diverged. Did the long and short THAP domains evolve by respectively gaining insertions or undergoing deletions in a 90-residue THAP domain? Or were the long or short domains the ancestral versions of the THAP domain?

Discussion

Most proteins contain more than one domain which are usually functionally independent. Typically, a protein domain folds into a core structural motif independent of the rest of the protein and is reported to be more conserved at the tertiary structure level than at the amino acid sequence level (34, 35). Evolutionary divergence or conservation of a protein domain is an outcome of the combination of random mutations and selection restrictions imposed on the function of the domain (34, 35). The THAP domain is a C2CH zinc coordinating DNA binding domain. It is an example of a protein domain shared between DNA binding proteins and a DNA transposase (DmTNP) which is active in the host organism. There are several examples of domains which are shared by different proteins. For example, BED zinc finger DNA binding domain which is shared by chromatin boundary element binding proteins (BEAF, DREF) and AC1 and Hobo- like transposases in fungi, plants and animals.

The GLAM2 predicted gapped motif for the THAP domains of all twelve human THAP proteins illustrates the conservation of Cys, Pro, Trp and His residues of the C2CH motif, a Phe immediately flanking the C2CH motif and a Pro in the AVPTIF motif. We report three additional conserved residues (Phe, Arg within the C2CH motif and a Leu in the L4 region). We also report that the conserved residues vary between the THAP domains of human THAP proteins (Fig. 1A) and other THAP domain-containing proteins (Fig. 1B).

The THAP domain is reported to fold into a consensus $\beta\alpha\beta$ fold in spite of differences in the corresponding amino acid sequence. Here, we report the possible presence of a hitherto unreported third and fourth beta sheet in the predicted secondary structures of hTHAP1, hTHAP4, hTHAP9 proteins, albeit with low confidence.

Variations in the length and sequence of Loop 4 have been earlier speculated to impose functional diversity in the binding sites of different human THAP proteins. Our sequence and

secondary structure analysis provide more confidence to these speculations. We also report that except for hTHAP0, all the human THAP proteins had at least one basic amino acid residue in their respective L4 regions. This is important because the DmTNP structure demonstrates that these basic amino acid residues directly contact the nitrogenous base in the minor groove of the DNA. However, further experimental studies are required to establish if only one basic residue in L4 is sufficient to bind DNA.

We report, for the first time, that there are THAP domains of lengths different from the conventional 90 residues. Although there is no correlation between the length of the THAP domain and the length of the full protein, it was interesting to see that THAP homologs from different taxonomic classes have THAP domains of similar lengths. For example, THAP1 homologs in class aves [Bengalese finch (Lonchura striata domestica)] and in class mammalia [long-finned pilot whale (Globicephala melas)] had 152 residue long THAP domain (Fig. 3). Moreover, even in the same organism, the THAP domain sometimes appears to have evolved differently. For example, in two-lined caecilian (Rhinatrema bivittatum), the THAP domain is either 102 residues (THAP1), 131 residues (THAP5) or 109 residues (THAP7) (Fig. 3). Similarly, the sea anemone (Exaiptasia pallid) THAP domain is either 136 residues (THAP1), 245 residues (THAP6), 226 residues (THAP9) or 108 residues (THAP11). Interestingly, the long-finned pilot whale (Globicephala melas) has both a 152-residue long THAP domain (THAP1) and a 42 residue short THAP domain (THAP8).

A conserved carboxy terminal region of the long THAP domains suggests an important functional role for this region. This is interesting because the direct DNA binding residues that are important for both specificity and affinity, are also located in the C-terminal end of the domain (18). On the other hand, the variations in the amino terminal region (C2CH zinc coordinating motif) of the long THAP domains suggests possible modifications or loss of the DNA binding functions of the zinc finger motif.

A truncated THAP domain called the DM3 domain, which lacks the first 20 residues including the two Cys of the C2CH motif has been earlier reported by the SMART database (6). However, in this study we report the presence of short THAP domains (in THAP3, THAP4, THAP5, THAP6 and THAP8 homologs) which lack ~40 residues at the extreme N terminus. Since the short THAP domains lack the first two Cys of the C2CH motif as well as the conserved Pro and Trp residues within the C2CH motif, it is tempting to speculate that they may not bind DNA. However, these short THAP domains retain the CH of the C2CH motif, the conserved Phe residue adjacent to the H of C2CH motif and the AVPTIF motif, thus leading us to ask questions like do these proteins bind DNA with less affinity or bind to non-specific DNA regions? It would be interesting to study if these proteins play any role in the regulation of DNA binding.

The difference in the predicted secondary structural elements of “other THAP domain-containing proteins” as well as the insertions in the long THAP domains, leads us to speculate about possible diversities in the THAP domain structural fold. Future investigations of the structures as well as DNA binding capabilities of these THAP domain outliers, with diverse length

as well as sequence including variations in the conserved residues, may shed light on the evolutionary basis of the observed THAP domain diversity.

References

1. Clouaire T, Roussigne M, Ecochard V, Mathe C, Amalric F, Girard J-P. 2005. The THAP domain of THAP1 is a large C2CH module with zinc-dependent sequence-specific DNA-binding activity. Proc Natl Acad Sci U S A 102:6907–6912.
2. Hammer SE, Strehl S, Hagemann S. 2005. Homologs of Drosophila P Transposons Were Mobile in Zebrafish but Have Been Domesticated in a Common Ancestor of Chicken and Human. Molecular Biology and Evolution 22:833–844.
3. Liew CK, Crossley M, Mackay JP, Nicholas HR. 2007. Solution structure of the THAP domain from Caenorhabditis elegans C-terminal binding protein (CtBP). J Mol Biol 366:382–390.
4. Majumdar S, Rio DC. 2015. P Transposable Elements in Drosophila and other Eukaryotic Organisms. Microbiol Spectr 3: MDNA3-2014.
5. Reid A, Sherry TJ, Yücel D, Llamosas E, Nicholas HR. 2015. The C-terminal binding protein (CTBP-1) regulates dorsal SMD axonal morphology in Caenorhabditis elegans. Neuroscience 311:216–230.
6. Roussigne M, Kossida S, Lavigne A, Clouaire T, Ecochard V, Glories A, Girard J. 2003. The THAP domain : a novel protein motif with similarity to the DNA-binding domain of P element transposase Trends Biochem Sci. 28:2001–2004.
7. Macfarlan T, Kutney S, Altman B, Montross R, Yu J, Chakravarti D. 2005. Human THAP7 Is a Chromatin-associated , Histone Tail-binding Protein That Represses Transcription via Recruitment of HDAC3 and Nuclear Hormone Receptor Corepressor J Biol Chem. 280:7346–7358.
8. Bessière D, Lacroix C, Campagne S, Ecochard V, Guillet V, Mourey L, Lopez F, Czaplicki J, Demange P, Milon A, Girard J-P, Gervais V. 2008. Structure-Function Analysis of the THAP Zinc Finger of THAP1, a Large C2CH DNA-binding Module Linked to Rb/E2F Pathways. J Biol Chem. 283:4352–4363.
9. Dejosez M, Krumenacker JS, Zitur LJ, Passeri M, Chu L, Songyang Z, Thomson JA, Zwaka TP. 2008. Ronin Is Essential for Embryogenesis and the Pluripotency of Mouse Embryonic Stem Cells Cell. 2:1162–1174.
10. Parker JB, Palchaudhuri S, Yin H, Wei J, Chakravarti D. 2012. A Transcriptional Regulatory Role of the THAP11 – HCF-1 Complex in colon cancer cell function Mol Cell Biol. 1654–1670.
11. Aguilo F, Zakirova Z, Nolan K, Wagner R, Sharma R, Hogan M, Wei C, Sun Y, Walsh MJ, Kelley K, Zhang W, Ozelius LJ, Gonzalez-Alegre P, Zwaka TP, Ehrlich ME. 2017. THAP1: Role in Mouse Embryonic Stem Cell Survival and Differentiation. Stem Cell Reports 9:92–107.
12. Balakrishnan MP, Cilenti L, Mashak Z, Popat P, Alnemri ES, Zervos AS. 2009. THAP5 is a human cardiac-specific inhibitor of cell cycle that is cleaved by the proapoptotic Omi/HtrA2 protease during cell death. Am J Physiol Heart Circ Physiol. 297:H643–H653.
13. Fuchs T, Gavarini S, Saunders-Pullman R, Raymond D, Ehrlich ME, Bressman SB, Ozelius LJ. 2009. Mutations in the THAP1 gene are responsible for DYT6 primary torsion dystonia. Nature Genetics 41:286–288.
14. Leite K, Morais D, Reis S, Viana N, Moura C, Florez M, Silva I, Dip N, Srougi M. 2013. MicroRNA 100: a context dependent miRNA in prostate cancer. Clinics 68:797–802.
15. Morais DR, Reis ST, Viana N, Piantino CB, Massoco C, Moura C, Dip N, Silva IA, Srougi M, Leite KR. 2014. The involvement of miR-100 in bladder urothelial carcinogenesis changing the expression levels of mRNA and proteins of genes related to cell proliferation, survival, apoptosis and chromosomal stability. Cancer Cell International 14:119.
16. Luscombe NM, Austin SE, Berman HM, Thornton JM. 2000. An overview of the structures of protein-DNA complexes. Genome Biol 1: reviews001.1.
17. Campagne S, Saurel O, Gervais V, Milon A. 2010. Structural determinants of specific DNA-recognition by the THAP zinc finger. Nucleic Acids Res. 38:3466–3476.
18. Sabogal A, Lyubimov AY, Corn JE, Berger JM, Rio DC. 2010. THAP proteins target specific DNA sites through bipartite recognition of adjacent major and minor grooves. Nat Struct Mol Biol. 17:117–124.
19. Gervais V, Campagne S, Durand J, Muller I, Milon A. 2013. NMR studies of a new family of DNA binding proteins: The THAP proteins. J Biomol NMR 56:3–15.
20. Wolfe SA, Nekludova L, Pabo CO. 2000. DNA Recognition by Cys2His2 Zinc Finger Proteins. Annu Rev Biophys Biomol Struct 29:183–212.
21. Balakrishnan MP, Cilenti L, Ambivero C, Goto Y, Takata M, Turkson J, Li XS, Zervos AS. 2011. THAP5 is a

DNA-binding transcriptional repressor that is regulated in melanoma cells during DNA damage-induced cell death. Biochem Biophys Res Commun. 404:195–200.
22. Trung NT, Kremmer E, Mittler G. 2016. Biochemical and cellular characterization of transcription factors binding to the hyperconserved core promoter-associated M4 motif. BMC Genomics 17:693.
23. Miller M. 2009. The importance of being flexible: the case of basic region leucine zipper transcriptional regulators. Curr Protein Pept Sci. 10:244–269.
24. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44:D279–D285.
25. de Castro E, Sigrist CJA, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, Bairoch A, Hulo N. 2006. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res. 34: W362–W365.
26. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 7:539.
27. Frith MC, Saunders NFW, Kobe B, Bailey TL. 2008. Discovering Sequence Motifs with Arbitrary Insertions and Deletions. PLOS Comput Biol. 4:e1000071.
28. Drozdetskiy A, Cole C, Procter J, Barton GJ. 2015. JPred4: A protein secondary structure prediction server. Nucleic Acids Res. 43: W389–W394.
29. McGuffin LJ, Bryson K, Jones DT. 2000. The PSIPRED protein structure prediction server. Bioinformatics 16:404–405.
30. Heffernan R, Yang Y, Paliwal K, Zhou Y. 2017. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. Bioinformatics 33:2842–2849.
31. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.
32. Cuff JA, Barton GJ. 2000. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins. 40:502–511.
33. Reeb J, Rost B. 2019. Secondary Structure Prediction, p. 488–496. *In* Ranganathan, S, Gribskov, M, Nakai, K, Schönbach, C (eds.), Encyclopedia of Bioinformatics and Computational Biology. Academic Press, Oxford, pp. 488–496.
34. Ponting CP, Russell RR. 2002. The Natural History of Protein Domains. Annu Rev Biophys Biomol Struct 31:45–71.
35. Bagowski CP, Bruins W, Te Velthuis AJW. 2010. The nature of protein domain evolution: shaping the interaction network. Curr Genomics 11:368–376.
36. Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: a sequence logo generator. Genome Res 14:1188–1190.

Table 1

Human THAP protein	**L4 length (residues)**
hTHAP0	26
hTHAP1	16
hTHAP2	26
hTHAP3	26
hTHAP4	14
hTHAP5	24
hTHAP6	27

hTHAP7	23
hTHAP8	26
hTHAP9	15
hTHAP10	35
hTHAP11	17

Table 2

THAP protein	Organism name	THAP domain length
	Phyllostomus discolor	101
	Corvus cornix cornix	102
	Rhinatrema bivittatum	102
	Denticeps clupeoides	104
	Bos taurus	109
	Cyanistes caeruleus	119
	Exaiptasia pallid	136
	Sparus aurata	137
	Delphinapterus leucas	152
	Globicephala melas	152
	Lagenorhynchus obliquidens	152
	Lonchura striata domestica	152
	Monodon monoceros	152
	Neophocaena asiaeorientalis	152
THAP1	Phocoena sinus	152
	Chrysemys picta bellii	108
	Eumeta japonica	122
	Dasypus novemcinctus	124
	Trichonephila clavipes	131
	Gopherus evgoodei	132
	Equus caballus	166
THAP2	Liparis tanakae	201
THAP3	Lynx canadensis	104

	Pelodiscus sinensis	119
	Balaenoptera acutorostrata scammoni	120
	Chelonia mydas	131
	Myotis lucifugus	132
	Vicugna pacos	134
	Microtus ochrogaster	159
	Rhinolophus ferrumequinum	163
	Heterocephalus glaber	207
	Stylophora pistillata	276
	Pteropus alecto	100
	Catharus ustulatus	119
	Neomonachus schauinslandi	131
	Lonchura striata domestica	138
	Gallus gallus	160
	Lagenorhynchus obliquidens	179
	Orcinus orca	179
	Rhinolophus ferrumequinum	180
	Mustela erminea	185
	Trachypithecus francoisi	192
THAP4	Gorilla gorilla gorilla	196
	Equus caballus	101
	Aquila chrysaetos chrysaetos	104
	Astatotilapia calliptera	107
	Protobothrops mucrosquamatus	110
	Chrysemys picta bellii	113
	Tupaia chinensis	118
	Corvus cornix cornix	128
THAP5	Corvus moneduloides	128

	Rhinatrema bivittatum	131
	Fundulus heteroclitus	176
	Betta splendens	216
	Mus Pahari	102
	Nannospalax galili	122
	Alligator sinensis	136
THAP6	Exaiptasia pallid	245
	Astatotilapia calliptera	104
	Cricetulus griseus	104
	Maylandia zebra	104
	Thamnophis elegans	105
	Pan paniscus	108
	Rhinatrema bivittatum	109
	Oreochromis niloticus	110
	Cynoglossus semilaevis	112
	Phasianus colchicus	113
	Amblyraja radiata	117
	Canis lupus dingo	125
	Canis lupus familiaris	125
	Rhincodon typus	129
	Falco cherrug	141
	Chelonia mydas	146
	Denticeps clupeoides	147
	Microcebus murinus	151
	Salmo salar	163
	Sus scrofa	164
	Peromyscus leucopus	240
	Nannospalax galili	242
THAP7	Liparis tanakae	309

	Bos taurus	101
	Canis lupus dingo	116
	Canis lupus familirais	116
	Neomonachus schauinslandi	136
	Odocoileus virginianus texanus	142
THAP8	Stylophora pistillata	413
	Strigops habroptila	101
	Chelonoidis abingdonii	118
	Aquila chrysaetos chrysaetos	125
	Eumeta japonica	127
	Folsomia candida	127
	Coturnix japonica	138
	Tupaia chinensis	139
	Exaiptasia pallida	226
THAP9	Dasypus novemcinctus	285
	Bubalus bubalis	144
THAP10	Suricata suricatta	158
	Exaiptasia pallida	108
THAP11	Balaenoptera acutorostrata scammoni	175

Table 3

THAP protein	**Organism name**	**THAP domain length**
THAP3	Alligator sinensis	39
	Pteropus alecto	42
THAP6	Acinonyx jubatus	48
	Myotis lucifugus	48
THAP4	Ictidomys tridecemlineatus	31

	Bubalus bubalis	50
	Odocoileus virginianus texanus	50
THAP5	Amblyraja radiata	42
	Antrostomus carolinensis	42
	Columba livia	42
	Falco cherrug	42
	Falco peregrinus	42
	Geospiza fortis	42
	Meleagris gallopavo	42
	Oreochromis niloticus	42
	Pelodiscus sinensis	42
	Tyto alba alba	42
	Armadillidium vulgare	46
THAP8	Globicephala melas	42
	Terrapene carolina triunguis	42
	Tyto alba alba	42
	Microtus ochrogaster	45

Figure legends

Figure 1. Conserved residues within various THAP domains

MSA of (A) twelve human THAP proteins (B) other THAP proteins, highlight conserved residues when viewed in jalview. The conserved residues [functionally important residues (green), function

uncharacterised (red), novel conserved residues identified in this study (blue)] within the (C)THAP domain of all twelve hTHAP proteins and (D) other THAP proteins are represented as a position weighted matrix (PWM). Functionally important non-conserved residues (purple) are also highlighted.

Figure 2. Secondary structural elements of various THAP domains

Predicted $\beta\alpha\beta$ fold for THAP domains of (A) twelve hTHAP proteins (B) other THAP proteins (DmTNP, zE2F6, CtBP) (C) predicted structural fold different than consensus $\beta\alpha\beta$ fold for the THAP domains of CDC14, Lin15B, Lin36 and Him-17. Alpha helix (red), beta sheet (green), basic residues in L4 (blue)

Figure 3. Comparison of THAP domain length among hTHAP homologs

Graphical representation of THAP domains of long THAP domain containing homologs of (A) THAP1 (B) THAP4 (C) THAP5 (D) THAP7 (E) THAP8 and Short THAP domain containing homologs of (F) THAP4 (G) THAP5 (H) THAP6 (I) THAP8. Y axis represents THAP domain length in residues and X axis lists the scientific name of the organism. The vertical bars are colored according to the taxonomic class of organisms (color key in bottom panel).

Figure 4. Correlation between THAP domain length and total protein length

Graphical representation of comparison of THAP domain length with total protein length of long and short THAP domains. X axis represents the length of the THAP domain (orange) and length of the total protein (blue) and Y axis represents the scientific name of the organism.

Figure 5. Comparison of GLAM2 predicted motifs of long and ~ 90 residue THAP domains in hTHAP homologs

Different gapped motifs were predicted for (A) long THAP domains and (B) ~90 residue long THAP domains for each of the twelve hTHAP homologs. The GLAM2 predicted gapped motifs are represented as PWMs.

Supplementary figure 1. Conserved sequence and structure signatures of THAP domain

(A) The solution structure of THAP domain of human THAP1 (PDB ID: 2JTG) as viewed in Pymol. The conserved P in the C2CH motif is in loop2 (highlighted in red), the conserved W within the C2CH motif is in the first alpha helix (highlighted in green), the conserved F immediately after the H of C2CH motif is at the beginning of loop3 (highlighted in magenta) and the conserved P of AVPTIF motif is in loop4 (highlighted in blue). (B) DNA binding sites of different THAP domain containing proteins as created by Weblogo (36) or downloaded from http://jaspar.genereg.net/. (C) Pymol representations of structural alignments of THAP domains of hTHAP1(2JTG; gray) with hTHAP2 (2D8R; blue), hTHAP11 (2LAU; magenta), DmTNP (3KDE; red), CtBP-1 (2JM3; peach).

Supplementary figure 2. THAP domains of same length sometimes have high sequence identity

(A) Identical long THAP domains in hTHAP1, hTHAP4, hTHAP5, hTHAP7 and hTHAP8 homologs (B) Identical short THAP domains in hTHAP4, hTHAP5, hTHAP6 and hTHAP8 homologs (C) ~90 residue THAP domains of hTHAP3 [THAP3 (1), THAP3 (2)], hTHAP4 [THAP4(1)] and hTHAP6 homologs [THAP6(1), THAP6(2)] which do not align with the corresponding hTHAP protein

Supplementary figure 3. MSA of 152 residue long THAP domains in avian and mammalian homologs of hTHAP1.

Supplementary figure 4. Predicted secondary structural folds of long THAP domains found in different hTHAP1 homologs

Different secondary structural folds (predicted by JPRED) for the THAP domains in each of the fifteen long hTHAP1 domain homologs, compared to the 90-residue THAP domain in hTHAP1. The secondary structure in the insertion regions (before W of C2CH motif) differs in each of the long hTHAP1 domain homologs.

All Sequences used in this study: Additional file 1 (THAP Prosite.xls)

Declarations

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

Availability of data and material: All data generated or analysed during this study are included in this article (and its supplementary files).

Competing interests: The authors declare that they have no competing interests