

SHORT COMMUNICATIONS

S. Pietrokovski · S. Henikoff

A helix-turn-helix DNA-binding motif predicted for transposases of DNA transposons

Received 29 October 1996 / Accepted: 28 January 1997

Abstract A helix-turn-helix (HTH) DNA-binding motif is identified in transposase sequences in Tc1, mariner and pogo DNA transposon. The findings are supported by results of various sequence analysis methods. Tc1 transposases are also predicted to contain another DNA-binding region. These findings are in accord with experimental evidence obtained from Tc1A, Tc3A and pogo transposases. The pogo family transposases, but not the pogo-type transcription factors, contain the HTH motif, suggesting that HTH structures are essential for Tc1/mariner/pogo transposition. Analysis of multiple sequence alignments enabled the identification of the HTH motif in distantly related protein sequences.

Key words Helix-turn-helix · DNA-binding motif · pogo · Tc1/mariner · Protein structure/function prediction

Introduction

Traces of ancient genetic invasions were recently detected in humans. Inactive and fragmentary remains of mariner and pogo DNA transposable elements were found to be abundant in human and other mammalian genomes (Auge-Gouillou et al. 1995; Morgan 1995; Oosumi et al. 1995; Robertson 1996; Robertson et al. 1996; Smit and Riggs 1996). The degenerate state of

these elements and their weak similarity to DNA transposons in other organisms made them difficult to detect, and thus sensitive database search tools were required. As a result DNA mobile elements are now known to occur in bacteria (Henikoff 1992; Selbitschka et al. 1995), protozoa (Doak et al. 1994), plants (Flavell et al. 1994) and most, if not all, animal phyla (Lam et al. 1996b; Plasterk 1996; Robertson 1995).

The most widespread animal transposons are from the Tc1 and mariner families. In vitro transposition systems have been developed for members of both (Lampe et al. 1996; Vos et al. 1996) and a Tc1-like transposon has been used as a vector for germ-line transformation between genera (Loukeris et al. 1995). Tc1/mariner transposons are small (1000–2000 bp), have short inverted repeats and encode a single protein. This protein is a transposase required and apparently sufficient for transposition in the absence of host factors (Lampe et al. 1996; Vos et al. 1996). Four conserved acidic residues in the C-terminal half of the transposase were identified as its catalytic domain (van Luenen et al. 1994; Vos and Plasterk 1994). This domain is related to that of retroviral integrases, bacterial IS3-like and Mu bacteriophage transposases (Doak et al. 1994; Grindley and Leschziner, 1995). However, there are non-transposase proteins similar to sequences from the pogo family (Smit and Riggs 1996; Toth et al. 1995; Tudor et al. 1992).

The protein region that binds the transposon inverted repeats was localized at the N-terminal domains of the Tc1A and Tc3A transposases (Colloms et al. 1994; Vos and Plasterk 1994). A few transposases from the Tc1 family were also found to have in their N-terminal regions sequence segments similar to protein regions with helix-turn-helix (HTH) DNA-binding motifs (Franz et al. 1994; Selbitschka et al. 1995; Vos et al. 1993). Aided by a new method for motif classification we show that all transposases of the Tc1, mariner and pogo families have a HTH DNA-binding motif in their N-terminal domains.

Communicated by D. J. Finnegan

S. Pietrokovski (✉)
Fred Hutchinson Cancer Research Center,
1100 Fairview ave. N, Seattle, WA 98109, USA

S. Henikoff
Howard Hughes Medical Institute, Fred Hutchinson
Cancer Research Center, Seattle, WA 98104, USA

Materials and methods

The sequences used to construct the Tc1 and mariner alignments are from Robertson (1995) Izsvak et al. (1995) Robertson and Asplund (1996), Robertson et al. (1996) and Lam et al. (1996a, b). The Tigger1 transposase sequence was retrieved from the repbase database of human repetitive elements (Jurka et al. 1992). Other sequences are found in public sequence databases. Database searches were performed with the BLAST programs (Altschul et al. 1990) on the SwissProt protein database (Bairoch and Boeckmann 1991) and Genbank nucleotide database (Benson et al. 1993).

Local multiple alignments were first found by the Block Maker (Henikoff et al. 1995) and MEME (Bailey and Elkan 1994) systems. Other regions in the sequences were then aligned by the Macaw program (Schuler et al. 1991) and a few gaps were manually inserted to increase the sequence similarity across the alignments.

Block searches of sequence databases were done with the Blimps and Multimot programs (Henikoff et al. 1995). The LAMA program (Petrokovski 1996) was used for block vs. block comparisons and database searches. The program is available on the WWW at URL "<http://blocks.jhrc.org/blocks-bin/LAMA-search>" Program hth (Conrad Halling, unpublished) was used to implement the Dodd and Egan HTH motif prediction method (Dodd and Egan 1990).

Results and discussion

Identifying conserved regions in the Tc1 and mariner transposases

Sequences belonging to the Tc1 family and to the mariner family were multiply aligned from end to end. In each family the conserved sequence regions (blocks) span most of the sequences (Fig. 1). The central and C-terminal regions are very similar to the multiple alignments of Henikoff and Robertson (Henikoff 1992; Henikoff et al. 1995; Robertson 1995; Robertson and Asplund 1996) and include the probable catalytic domain (Doak et al. 1994). Although the N-terminal region of the Tc1 transposases is less conserved than the catalytic region (Colloms et al. 1994), it can be confidently aligned using advanced methods for local multiple alignments. The multiple alignments were objectively constructed for each family, using only sequence similarity and assuming the same block order in each sequence.

Querying databases with the Tc1 and mariner blocks

Searching sequence databases with the transposase blocks as queries identified many Tc1/mariner-related sequences (not shown). Block A from the Tc1 family and block C from the mariner family were also significantly similar to regions in various DNA-binding proteins. These proteins include eukaryotic paired domain proteins, bacterial insertion sequence transposases, transcription regulators, RNA polymerase sigma factors and resolvases (not shown). Most of these regions are known or predicted to contain HTH DNA-binding structures. HTH structures are found in diverse proteins that bind specific DNA sequences (Pabo and Sauer 1992) including transposases (Petrokovski 1996; Selbitschka et al.

1995). The A block from Tc1 and C block from mariner families are both in the N-terminal DNA-binding domains of these protein families (Fig. 1B).

To confirm the presence of HTH motifs in the N-terminal domains of Tc1 and mariner transposases, a search of block databases (Petrokovski 1996) was performed using the transposase blocks as queries. This "block-vs-block" search can be more diagnostic of a relationship than sequence-vs-sequence and block-vs-sequence searches (Petrokovski 1996). Searching the Blocks (Henikoff and Henikoff 1991) database identified significant similarities between different HTH entries and both Tc1 block A and mariner block C (Table 1). The Blocks database HTH entries are of HTH motifs in families of bacterial DNA-binding proteins: eight families of transcription regulators, two families of sigma transcription factors and a family of insertion sequence transposases. No other entries in the databases searched were similar to any of the Tc1/mariner N-terminal blocks, except for a similarity between the LacI family HTH entry and block Tc1 C. This pattern of similarity to known HTH blocks was previously shown to be highly specific for HTH motifs (Petrokovski 1996).

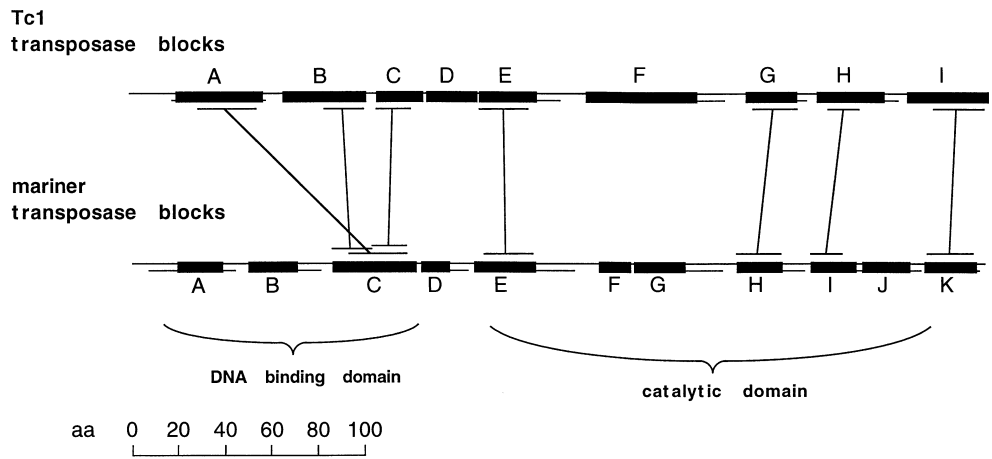
An additional analysis for HTH motifs was performed with the Dodd and Egan method. This method examines single sequences using a HTH-specific scoring matrix (Dodd and Egan 1990). The analysis identified probable HTH motifs in most, but not all, of the transposase sequences inside Tc1 block A and mariner block C (Fig. 1B). The identification failures might be due to the mostly prokaryotic sequences used to construct the Dodd and Egan HTH matrix and the relatively limited sensitivity of methods that analyze single sequences at a time.

Predicted HTH motif corresponds to DNA binding region

The above findings, based on three different methods, strongly indicate that the N-terminal regions of both Tc1 and mariner transposases contains a HTH motif. Experimental data support this conclusion. Deletion analyses by Plasterk and coworkers have shown that the N-terminal regions of the Tc1A and Tc3A transposases specifically bind to their respective inverted repeats (Colloms et al. 1994; Vos et al. 1993). The proposed Tc1 HTH block contains segments from both these regions.

Sporadic reports of HTH motifs have been based on analysis of some individual Tc1-type transposases, including Tc1A, Minos and the bacterial *ISRm2011-2* (Franz et al. 1994; Selbitschka et al. 1995; Vos et al. 1993). The Tc1A and Minos segments are found in the proposed Tc1 HTH block. However, another report claimed to identify a leucine zipper motif in the corresponding region (Ivics et al. 1996). This putative leucine zipper would be specific for fish Tc1 transposases, would overlap a paired-related domain and would require the insertion of gaps in the multiple alignment. We could

A



B

Tc1

Block:	A	B	C
	HHHHHHHhthtHHHHHHHH	HHHHHHHH	HHHHHHHH
Tc1A	18 IVAGFEQGIPTKMLALQIORSFSPSTIWKVIKKYQTEKS 54	62 GRPRVTTTHRMDRNILRSAREDPHRTATDIQMI I SSP 97	102 PSKRTVRRRLQQAGLHGRKP 121
Tc3A	15 LDVVKLLNLSLHMSRKRISRSRHCIREYKLDPVSYGT 51	56 PRRKALSVRDERNVIRAASNSCK-TARDIRNELQLS 90	91 ASKRTILNVIKRSGVIVRQK 110
Tss1	15 IVDLHKSGSSLGAI SKHLKVPRSSVQIVIRKYKHGT 51	59 GRRRVLSPRDERTLVRKVQINPRTTAKDLVKMLEET 94	98 VSI STVKRVLYRHNKGRSA 117
Tdrla	15 IVDLHKSGSSLVTISRCLKVPHSSVQTNLNRNYKQHG 51	59 GRRRILCPRDERAMVRTVCINPRTKAKDLVKMLAEA 94	97 RVSLSTVKRVLYHGLKGHSA 116
Tzf	21 IVSRHKAGEGYRNI SAALKVPMSTATSIRKWKMPGT 57	65 GRPSKLSVRGRRSLVREVINNPMVTLSELQRSSVER 100	104 YRRTTICAAIHQSGLYGRVA 123
Txr	20 LSHLSSISKVYKAIKSKALGLORTIVRAIHKWKHGT 56	64 GRPTKITPRAQQLIREATKDPRTTSKELQASLASI 99	103 VHDSTIRKRLGKNGLHGRFP 122
Txz	14 LLKLRKQKQKPIREMATILGEAKSTVYVYILKKEITGE 50	58 GRPQKTTVVDHRI I SMGKRNPFTRAKQVNNTLQEV 93	97 ISKSTIKRRLYESKYRGGTA 116
Bari	15 IVARFKAGTFAAKIAEYIORSRTVYVYILKKEITGE 51	59 GRKPVLDRQCRQILGVVAKNPSASPVKIALESKNT 94	99 VSSSTIRRRLEADFKTYVV 118
S	14 TYYNHQLGKSIPELVEIFSVSRKTVYNILNRAEKEGR 50	58 GCKTKINKRVDRLIMRKAIANPRI SVRSLAQDIREE 93	98 VSHETVRQVILRHRYSSRVA 117
HB1	3 ILKLRKEGKTYKDIQKTLKCSAKMVSNAIKYKWKPEN 39	41 GTKHKTDDIEDRRIVSYSKVYRFASFRIKSELNLG 76	77 ISDVTIRRRLLNQNFARS 96
Minos	15 IRDYFKSGKTLTEISKOLNLPKSSVHGVIQIFKKNGN 51	60 GRTSAITPRDKRQIAKIVKADRRQSLRN LASKWSQT 95	100 VKREWTRHELKSI GYGFYKA 119
Uhu	13 ILEHFKIGYSYROI AKMVNLSLTTTFVNI IRRFVDENR 49	57 APNKIFTEQEERRI IRKIRENPKLSAPKLTQQVQDE 92	97 CSVQTVRRVVLHNDHFNARVP 116
Paris	14 VYFHYHKGKCAKELAE MFSIKLRTTYNIINRAEKENR 50	58 GRPAKLSRRDH SKILKQINENPQTS LRQLALDLKND 93	98 VSHETVRKVLKMHKYSQJIA 117
TCL4.7			8 VSLSTVRRHLHAADIHNYKP 27

mariner

Block:	A	B	C
			HHHHHHHhthtHHHHHHHH
mar1 Hs	17 FKMGRKAAETTRNINAF 35	47 WFKFKCGDESLEDEERSGRP 67	79 IEADPLTTTREVAEELNVDHSTVVRHLKQIGKVKKL 114
mar2 Hs	15 VRLGWKNGEII DALRKYV 33	45 WITRFKKGQDNVEDEAHSGRP 65	81 IEEDQQLTAETIANFIDISIGSAYTILTEKLLKLSKL 116
mar1 Gp	16 FVKGENESQA AKIVNGVY 34	45 WFRFRFSGIFDVKDAPRTYGR 65	77 IEVDRHVSSRNIA--HKT DHTVNLNHLKAGFKKKL 110
mar1 De	16 FVKGKSARETFREINGVY 34	46 WFRFRFAGENDTMDKPAGGRP 66	78 IELDRHVASRDIAOEMGVSHQTIINHLQKAGYKKL 113
mar1 Am	16 FRKGNASQA HKLCAVY 34	46 WFDKFRSGDFSLKDEKRSGRP 66	78 IDSDRHSTTRETAEKLVHSHTCIENHLKQIGVYQKL 113
mar1 Hc	14 FYRGTSAEATARRINNVY 32	44 WFQFRFSGDFDLQNGPR-GRP 63	76 V*ADPQSOSTSEI AAGFGVSDKTVLTYLQIGKVKKL 111
mar1 Dt	14 FHRGATTRQAVGNINSVYP 32	44 WFKRFRSGDFDMSNQPRS-RP 63	76 VEADSSQSALELASKFGVAKSIIILHLKQINKVKKL 111
mar1 Csp	14 FKLGRKAAETARDINDAF 32	44 WFKKFRGDESLEDDERSGRP 64	76 VNANTRVTLRELAELDVTPTI SNHLKEIGTKTKL 111
mar1 Dm	19 FHLKKTAAESHRLMVEAF 37	49 WFQRFKSGDFDVKDEHGKPP 69	81 LDEDDAQOTOKOLAEOLVSSQAVSNRLREMGKI QKV 116
mar1 Mo	19 FHLKKSASEAHLLEVEAF 37	49 WYENFSKGNFDLENEPRGRP 69	81 LDQNNRQTOSELAEOLKVTREAI STRLKMKG#ISKL 116
mar1 Md	21 FPAKKTAAESHRLMVEVY 39	51 WFQRFKSGDFDTEDEKERPGP 71	83 LDEDCQTOEELAKSLGVTOQAI SKRLKAAGYIQQK 118
mar1 Cp	14 FLKGNVTVEAKTWDNEFP 32	44 WYAKFKRGMSTEDGERSGRP 64	80 ILNDRKMKLIEIAEALKISKERVGHIIHQYLDMRKL 115
mar1 Mp	14 FLKGGK#TAEAKTWDLE*FP 32	44 WYAKFKRGMSTEDGKHTGRP 64	80 ILNDRKMKLIEIADIVKISKERVHHTI IREYLYGRKL 115
mar1 Bm		41 RAINRCNETSSVCDKRSGRP 61	77 IRRNPVKOKILSREMKIAPRTMSRI LKDDGLAAY 112
mar1 Ce	20 FPQICNCNEARRNMCVVG 38	50 WFEKFTKKNYLDLDDKPR*DRS 70	82 LEDDPRATNRELSATLKHPOKTI INHLHETGRVEKF 117
mar2 Ce	19 FRLGHSAMEAERNICGAMG 37	49 WFQKFKNGDFSL EEIERSGRP 69	81 VEEEPRLSREMEEKLECCHSTIARHLGRIGPTSKL 116
mar5 Ce	13 FHRNGVAAKSIARRLKVSE 31	37 TIARFKELGNFSDRSRGRP 57	72 FRHNSGRSVRAMARELKI SOSSLCRMVKNLKLKAY 107

Fig. 1A, B Conserved regions in the Tc1 and mariner transposases. **A** Block diagram. Blocks are depicted as filled rectangles. The longest and shortest distances between blocks in each family are represented by the lines between the blocks. The locations of the DNA binding and catalytic domains are the same for both families. Similarities between blocks are shown as lines connecting the Tc1 and mariner blocks. **B** Block alignment of the N-terminal regions. Blocks are labeled with their names. The predicted structure is shown above the blocks with “H” corresponding to alpha helix and “t” to turn. Each sequence segment is flanked by the coordinates of its first and last amino acid. The marks beneath each block indicate every fifth (“:”) and tenth (“I”) position. Segments identified as probable [> 3 SD above the mean (Dodd and Egan 1990)] HTH motifs by the hth program (C. Halling, unpublished) are underlined. No other probable HTH motifs were identified by the hth program in the Tc1/mariner sequences. Unknown amino acids due to stop codons and frame

shifts are marked by “*” and “#”, respectively. The species listed are nematode *C. elegans* (Tc1A, Tc3A, mar1 Ce, mar2 Ce and mar5 Ce); human (mar1 Hs and mar2 Hs); Atlantic salmon *S. salar* (Tss1) (zebrafish *D. rerio* (Tdrla and Tzf); Pacific hagfish *E. stouti* (Tes1) (the sequence was aligned only in the central and C-terminal blocks); clawed frog *X. laevis* (Txr and Txz); fruit fly *D. melanogaster* (Bari, S and HB1); fruit fly *D. hydei* (Minos); fruit fly *D. heteroneura* (Uhu); fruit fly *D. virilis* (Paris); false codling moth *C. leucotreta* (TCL4.7); tsetse fly *G. palpalis* (mar1 Gp); fruit fly *D. mar1* De; honey bee *A. mellifera* (mar1 Am); giant silkworm *H. cecropia* (mar1 Hc); flatworm *D. tigrina* (mar1 Dt); staphylinid beetle *Carpelimus*. sp (mar1 Csp); fruit fly *D. mauritiana* (mar1 Dm); predatory mite *M. Occidentalis* (mar1 Mo); Hessian fly *M. destructor* (mar1 Md); green lacewing *C. plorabunda* (mar1 Cp); mantispid *M. pulchella* (mar1 Mp); silkworm moth *B. mori* (mar1 Bm)

Table 1 Blocks Database entries similar to Tc1 A and mariner C query blocks

Database block ^a	Tc1 A block			Mariner C block		
	Positions ^b	z-score	Expected ^c	Positions	z-score	Expected
crp BRP ^d BL00042B	1–30/7–36	5.7	$2.1 \cdot 10^{-1}$	2–28/6–32	6.2	$8.9 \cdot 10^{-2}$
gntR BRP BL00043	1–32/5–36	8.1	$\leq 1.0 \cdot 10^{-6}$	1–28/3–30	8.8	$\leq 1.0 \cdot 10^{-6}$
lysR BRP BL00044	–	–	–	2–19/13–30	6.4	$5.3 \cdot 10^{-2}$
lacI BRP BL00356	1–25/10–34	6.4	$4.6 \cdot 10^{-2}$	1–28/8–35	7.9	$\leq 1.0 \cdot 10^{-6}$
luxR BRP BL00622	8–44/1–37	9.0	$\leq 1.0 \cdot 10^{-6}$	15–45/6–36	9.8	$\leq 1.0 \cdot 10^{-6}$
deoR BRP BL00894A	–	–	–	5–40/1–36	8.0	$\leq 1.0 \cdot 10^{-6}$
iclR BRP BL01051A	1–26/11–36	5.8	$1.8 \cdot 10^{-1}$	–	–	–
tetR BRP BL01081	–	–	–	16–43/6–33	8.4	$\leq 1.0 \cdot 10^{-6}$
Sigma-54 BL00717F	–	–	–	1–23/3–25	6.6	$3.5 \cdot 10^{-2}$
Sigma-70 BL01063B	12–46/1–35	7.8	$3.3 \cdot 10^{-3}$	14–46/1–33	6.8	$2.1 \cdot 10^{-2}$
IS30 BL01043A	10–40/1–31	9.0	$\leq 1.0 \cdot 10^{-6}$	18–41/7–30	7.0	$1.3 \cdot 10^{-2}$

^a Block entries are from the Blocks Database v9.1.

^b The positions are of the aligned regions for each pair of blocks in the order database block/query block

^c The score cutoff is 5.6 Z-score units and corresponds to the top 7.7×10^{-5} percentile of chance scores (Petrokovski 1996).

The probability that the similarity occurs by chance is calculated for a search of the same number (3300) of blocks.

^d BRP, bacterial regulatory proteins; sigma-54 and sigma-70 are bacterial RNA polymerase sigma factors; IS30 refers to bacterial IS30 transposases.

not confirm the presence of a leucine zipper in Tc1/mariner sequences. The lack of agreement between conventional single-sequence search methods emphasizes the need for the more diagnostic multiple-alignment-based tools used in this work.

Similarity between the Tc1 and mariner blocks predicts a bipartite DNA binding domain in Tc1 transposases

We compared the Tc1 and mariner transposase blocks with each other to examine the relation between the conserved regions of the two groups (Fig. 1A and Table 2). The blocks containing the putative catalytic

residues (Doak et al. 1994) and the C-terminal ends of the Tc1 and mariner transposases are similar to each other, confirming earlier analyses (Doak et al. 1994; Henikoff and Henikoff 1992; Robertson 1995). The mariner and Tc1 HTH regions are similar to each other, with the HTH motifs predicted by the Dodd and Egan method (Fig. 1B) aligned with each other. The two blocks following the Tc1 HTH block are also similar to the mariner HTH block. The first block is similar to the N-terminal half of the mariner HTH motif and the second is similar to the C-terminal half (Fig. 1A). This indicates that the structure adopted by the sequences in these two blocks is two DNA-binding helices separated by three to seven amino acids. Nonspecific DNA-binding activity was observed for these regions in the Tc1A

Table 2 Similarity between Tc1 and Mariner blocks

Blocks	Positions	z-score	Expected ^a	
Tc1 A	Mariner C	10–34/8–32	8.6	$\leq 1.0 \cdot 10^{-6}$
Tc1 B	Mariner C	18–34/1–17	4.6	$6.3 \cdot 10^{-2}$
Tc1 C	Mariner C	1–15/18–32	8.5	$\leq 1.0 \cdot 10^{-6}$
Tc1 E	Mariner E	1–21/5–25	4.6	$6.6 \cdot 10^{-2}$
Tc1 G	Mariner H	4–22/1–19	12.7	$\leq 1.0 \cdot 10^{-6}$
Tc1 H	Mariner I	6–16/1–11	6.3	$1.9 \cdot 10^{-3}$
Tc1 I	Mariner K	12–33/2–23	7.5	$1.0 \cdot 10^{-4}$

^a The similarity was calculated as in Table 1 except that the expected probability that a given similarity occurs by chance is calculated for searches of 100 blocks and the Z-score cutoff is 4.5 (corresponding to the top 7.3×10^{-4} percentile of chance scores).

and Tc3A transposases (Colloms et al. 1994; Vos and Plasterk 1994). We propose that the N-terminal regions of both Tc1 and mariner transposases have a specific DNA-binding HTH in Tc1 block A and mariner block C, and that Tc1 transposases also have a DNA-binding domain with little or no sequence specificity in Tc1 blocks B and C.

The bipartite DNA-binding domain predicted for Tc1 transposases bears similarity to the paired domain found in the Pax transcription factors (Franz et al. 1994; Ivics et al. 1996; Vos et al. 1993). The domain consists of two DNA-binding motifs that each recognize distinct DNA sequences (Czerny et al. 1993; Epstein et al. 1994; Jun and Desplan 1996; Xu et al. 1995). DNA binding of the C-terminal region depends on the N-terminal HTH motif and is dispensable in some Pax proteins (Bertuccioli et al. 1996; Cai et al. 1994; Czerny et al. 1993; Epstein et al. 1994). The C-terminal region is less conserved than the N-terminal region and has a HTH-like structure composed of three α -helices with loops of 7–10 amino acids between them in the *Drosophila* Prd protein (Xu et al. 1995). Franz et al. (1994) found sequence similarity between the N-terminal region of Tc1 transposases and the paired domain of Prd, aligning a segment in the Tc1 block B with the Prd C-terminal α 5 helix. We confirm this by finding this region of block B to be similar to the first mariner HTH helix (Table 2).

Bipartite DNA-binding domains containing a HTH structure are known for other proteins as well. Yeast RAP1 protein apparently binds to telomeric DNA via two domains with HTH structures. Both domains are structurally related to homeodomains, but the C-terminal domain has an additional helix and a large loop. This domain was suggested to bind less specifically than the N-terminal domain and to contribute mainly to the general binding affinity (König et al. 1996). PurR and lac repressor DNA-binding domains consist of a HTH structure next to a minor groove-binding hinge helix (Lewis et al. 1996; Schumacher et al. 1994). The POU domain of the Oct-1 transcription factor binds DNA using two consecutive HTH structures (Klemm et al. 1994; Sturm and Her 1988). Myb oncoprotein has a similar organization with two tandem repeats having HTH-like structures. One of these repeats is essential for

the specificity of the DNA-binding and the other stabilizes it (Ording et al. 1994; Tanikawa et al. 1993).

Presence of a HTH motif correlates with transposase activity in the pogo family

To demonstrate the predictive value of the Tc1/mariner blocks we compared them with very distantly related protein sequences from the pogo transposon family. Sequence segments similar to the HTH blocks were found in the N-terminus of the proteins, and the presence of HTH motifs was also indicated by the Dodd and Egan (1990) method (Fig. 2). The DNA-binding domain of the human centromere protein CENP-B protein is known to lie in its N-terminal 125 residues and our predicted HTH motif is exactly the one found to be similar to the HTH of homeodomains (Yoda et al. 1992). The similarity of pogo sequences to Tc1/mariner transposases had been recognized only in their central and C-terminal parts (Robertson 1996; Smit and Riggs 1996). We extend the alignment between the families to the N-terminal region and predict a HTH DNA-binding structure in pogo transposases. Recent experiments have confirmed that the DNA-binding domain of pogo transposase is contained within the N-terminal 75 amino acid residues and that mutations affecting the putative HTH structure of the protein greatly reduce this activity (H. Wang and D. J. Finnegan, personal communication).

Three non-transposase proteins of the pogo family were not predicted to have HTH structures using either the Tc1/mariner HTH blocks or the Dodd and Egan method. These proteins are the PDC2 (Hohmann 1993) and RAG3 (Prior et al. 1996) fungal transcription factors and the murine Jerky protein, inactivation of which causes epileptic seizures in mice (Toth et al. 1995). Comparison of these sequences to the conserved regions of the other pogo transposases confirmed the sequence similarity between all the sequences outside the HTH structure predicted for the transposases (not shown). Sequence database searches using BLAST readily detected similarity between pogo family members but failed to align N-terminal segments predicted to be HTH structures with these three non-transposase proteins (not

	HHHHHHHcttHHHHHHH	HTH
		Z-score
Tigger1	MASKCSSERKSXTSLTLNQLKLEMIKLSEEG MSKAEIGQKLGLLRQTVSQVVNAKEKFLKE	5.8
Pogo	MGKTKRVVGLTLKEKLIIELVTNK VDKKEICAKFKCDRSTVNRILOK TNEIHEA	3.0
Aot1	MIKTSaipPKIPKSKSRVEQEGRILLAI SAIKKQETSSFRKAAEIFNIPIATLRYRLNGG SFRNDT	4.2
Pot2	MKQYTEKQLISAINDVNNG NPIAKTSRKWGI PRSTLQ SRLKGSQPYKKA	3.2
Fcc1	MPQQRSIQTSCEGRISLATAS YRNNPKQSVRALAVAYDVPKSTLQ RRLHGTHARSEI	2.7
CENP-B Hs	MGPKRRQLTFREKSRIIQE VEENPDLRKGEIARRFNI PPSTL STILK NKRAILLAS	5.6

Fig. 2 HTH motifs in pogo sequences. Individual sequences in the box are similar to the mariner C block, whose secondary structure prediction is shown as in Fig. 1 *Bold* segments mark probable HTH motifs identified by the hth program (Dodd and Egan, 1990). The hth program Z-scores (standard deviations from the mean) are shown for each sequence. All sequences begin at their N-terminus. Similar results were obtained with Tc1 block A. The sequence descriptions and

database accessions are Tigger1, human Tigger1 transposase (rebase: Tigger 1); Pogo, *D. melanogaster* pogo transposase (GenBank:X59837); Aot1, *A. oryzae* Aot1 transposase (GenBank:D45179; Robertson et al. 1996); Pot2, *M. grisea* Pot2 transposase (PIR: 1078658); Fcc1, *C. carbonum* Fcc1 transposase (GenBank:U40479); CENP-B Hs, human centromere protein B (SwissProt:P07199)

shown). The apparent lack of HTH motifs implies that the HTH structure is essential for pogo, and probably Tc1/mariner, transposition.

Conclusion

Sensitive methods for comparing multiple protein sequences allowed us to reliably identify sequence similarities beyond the reach of single-sequence search methods. The block-vs-block comparison method (Petrokovski 1996) provided functional and structural identifications of conserved protein regions. Thus, multiple alignments can be useful for tasks other than phylogenetic reconstruction and searches for additional family members. The detection of similarity between different protein families and their corresponding motifs enables us to transfer the knowledge gained for one family to the other.

Acknowledgements We thank H. Robertson and C. Desplan for useful comments and providing manuscripts prior to publication. Part of this work was done at the 1996 "Identifying Features in Biological Sequences" workshop of the Aspen Center for Physics. S. P. is a Howard Hughes Medical Institute Fellow of the Life Sciences Research Foundation.

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410
- Auge-Gouillou C, Bigot Y, Pollet N, Hamelin MH, Meunier-Rotival M, Periquet G (1995) Human and other mammalian genomes contain transposons of the mariner family. *FEBS Lett* 368:541–546
- Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of Second International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, Calif, pp 28–36
- Bairoch A, Boeckmann B (1991) The SWISS-PROT protein sequence data bank. *Nucleic Acids Res* 19:2247–2249
- Benson D, Lipman DJ, Ostell J (1993) GenBank. *Nucleic Acids Res* 21:2963–2965
- Bertuccioli C, Fasano L, Jun S, Wang S, Sheng G, Desplan C (1996) In vivo requirement for the paired domain and homeodomain of the paired segmentation gene product. *Development* 122:2673–2685
- Cai J, Lan Y, Appel LF, Weir M (1994). Dissection of the *Drosophila* paired protein: functional requirements for conserved motifs. *Mech Dev* 47:139–150
- Colloms SD, van Luenen HG, Plasterk RH (1994) DNA binding activities of the *Caenorhabditis elegans* Tc3 transposase. *Nucleic Acids Res* 22:5548–5554
- Czerny T, Schaffner G, Busslinger M (1993) DNA sequence recognition by Pax proteins: bipartite structure of the paired domain and its binding site. *Genes Dev* 7:2048–2061
- Doak TG, Doerder FP, Jahn CL, Herrick G (1994) A proposed superfamily of transposase genes: transposon-like elements in ciliated protozoa and a common "D35E" motif. *Proc Natl Acad Sci USA* 91:942–946
- Dodd IB, Egan JB (1990) Improved detection of helix-turn-helix DNA-binding motifs in protein sequences. *Nucleic Acids Res* 18:5019–5026
- Epstein JA, Glaser T, Cai J, Jepeal L, Walton DS, Maas RL (1994) Two independent and interactive DNA-binding subdomains of the Pax6 paired domain are regulated by alternative splicing. *Genes Dev* 8:2022–2034
- Flavell AJ, Pearce SR, Kumar A (1994) Plant transposable elements and the genome. *Curr Opin Genet Dev* 4:838–844
- Franz G, Loukeris TG, Dialektaki G, Thompson CR, Savakis C (1994). Mobile Minos elements from *Drosophila hydei* encode a two-exon transposase with similarity to the paired DNA-binding domain. *Proc Natl Acad Sci USA* 91:4746–4750
- Grindley ND, Leschziner AE (1995) DNA transposition: from a black box to a color monitor. *Cell* 83:1063–1066
- Henikoff S (1992) Detection of *Caenorhabditis* transposon homologs in diverse organisms. *New Biol* 4:382–388
- Henikoff S, Henikoff JG (1991) Automated assembly of protein blocks for database searching. *Nucleic Acids Res* 19:6565–6572
- Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* 89:10915–10919
- Henikoff S, Henikoff JG, Alford WJ, Petrokovski S (1995) Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene* 163:GC 17–26
- Hohmann S (1993) Characterization of *PDC2*, a gene necessary for high level expression of pyruvate decarboxylase structural genes in *Saccharomyces cerevisiae*. *Mol Gen Genet* 241:657–666
- Ivics Z, Izsvák Z, Minter A, Hackett PB (1996) Identification of functional domains and evolution of Tc1-like transposable elements. *Proc Nat Acad Sci USA* 93:5008–5013
- Izsvak Z, Ivics Z, Hackett PB (1995) Characterization of a Tc1-like transposable element in zebrafish (*Danio rerio*). *Mol Gen Genet* 247:312–322
- Jun S, Desplan C (1996) Cooperative interactions between paired domain and homeodomain. *Development* 122:2639–2650
- Jurka J, Walichewicz J, Milosavljevic A (1992) Prototypic sequences for human repetitive DNA. *J Mol Evol* 35:286–291
- Klemm JD, Rould MA, Aurora R, Herr W, Pabo CO (1994) Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules. *Cell* 77:21–32
- König K, Giraldo R, Chapman L, Rhodes D (1996) The crystal structure of the DNA-binding domain of yeast RAP1 in complex with telomeric DNA. *Cell* 85:125–136
- Lam WL, Lee T-S, Gilbert W (1996a). Active transposition in zebrafish. *Proc Nat Acad Sci USA* 93:10870–10875
- Lam WL, Seo P, Robison K, Virk S, Gilbert W (1996b) Discovery of amphibian Tc1-like transposon families. *J Mol Biol* 257:359–366
- Lampe DJ, Churhill MEA, Robertson HM (1996) Purified marine transposase is sufficient to mediate transposition in vitro. *EMBO J* 15:5470–5479
- Lewis M, Chang G, Horton NC, Kercher MA, Pace HC, Schumacher MA, Brenan RG, Lu P (1996) Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science* 271:1247–1254
- Loukeris TG, Livadaras I, Arca B, Zabalou S, Savakis C (1995) Gene transfer into the medfly, *Ceratitis capitata*, with a *Drosophila hydei* transposable element. *Science* 270:2002–2005
- Morgan GT (1995) Identification in the human genome of mobile elements spread by DNA-mediated transposition. *J Mol Biol* 254:1–5
- Oosumi T, Belknap WR, Garlick B (1995) Mariner transposons in humans. *Nature* 378:672
- Ording E, Kvavik W, Bostad A, Gabrielsen OS (1994) Two functionally distinct half-sites in the DNA recognition sequence of the Myb oncoprotein. *Eur J Biochem* 222:113–120
- Pabo CO, Sauer RT (1992) Transcription factors: structural families and principles of DNA recognition. *Annu Rev Biochem* 61:1053–1095
- Petrokovski S (1996) Searching database of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res* 24:3836–3845
- Plasterk RH (1996) The Tc1/mariner transposon family. *Curr Top Microbiol Immunol* 204:125–143

- Prior C, Fukuhara H, Wesolowski-Louvel M (1996) *RAG3* gene and transcriptional regulation of the pyruvate decarboxylase gene in *Kluyveromyces lactis*. *Mol Microbiol* 20:765–772
- Robertson HM (1995) The Tc1-mariner superfamily of transposons in animals. *J Insect Physiol* 41:99–105
- Robertson HM (1996) Members of the pogo superfamily of DNA-mediated transposons in the human genome. *Mol Gen Genet* 252:761–766
- Robertson HM, Asplund ML (1996) *Bmmar1*: a basal lineage of the mariner family of transposable elements in the silkworm moth, *Bombyx mori*. *Insect Biochem Mol Biol*, 26:945–954
- Robertson HM, Zumpano, KL, Lohe AR, Hartl DL (1996). Reconstructing the ancient mariners of humans. *Nature Genet* 12:360–361
- Schuler GD, Altschul SF, Lipman DJ (1991) A workbench for multiple alignment construction and analysis. *Proteins* 9:180–190
- Schumacher MA, Choi KY, Zalkin H, Brennan RG (1994) Crystal structure of LacI member, PurR, bound to DNA: minor groove binding by alpha helices. *Science* 266:763–770
- Selbitschka W, Arnold W, Jording D, Kosier B, Toro N, Pühler A (1995) The insertion sequence element *ISRm2011–2* belongs to the *IS630*-Tc1 family of transposable elements and is abundant in *Rhizobium meliloti*. *Gene* 163:59–64
- Smit AF, Riggs, AD (1996) Tiggers and other DNA transposon fossils in the human genome. *Proc Nat Acad Sci USA* 93:1443–1448
- Sturm RA, Herr W (1988). The POU domain is a bipartite DNA-binding structure. *Nature* 336:601–660
- Tanikawa J, Yasukawa T, Enari M, Ogata K, Nishimura Y, Ishii, S, Sarai A (1993) Recognition of specific DNA sequences by the c-myc protooncogene product: role of three repeat units in the DNA-binding domain. *Proc Natl Acad Sci USA* 90:9320–9324
- Toth M, Grimsby J, Buzsaki G, Donovan GP (1995) Epileptic seizures caused by inactivation of a novel gene, *jerky*, related to centromere binding protein-B in transgenic mice. *Nature Genet* 11:71–75
- Tudor M, Lobočka M, Goodell M, Pettitt J, O'Hare K (1992). The pogo transposable element family of *Drosophila melanogaster*. *Mol Gen Genet* 232:126–134
- Van Luenen HG, Colloms SD, Plasterk RH (1994) The mechanism of transposition of Tc3 in *C. elegans*. *Cell* 79:293–301
- Vos JC, Plasterk RH (1994) Tc1 transposase of *Caenorhabditis elegans* is an endonuclease with a bipartite DNA binding domain. *EMBO J* 13:6125–6132
- Vos JC, van Leunen HG, Plasterk RH (1993) Characterization of the *Caenorhabditis elegans* Tc1 transposase in vivo and in vitro. *Genes Dev* 7:1244–1253
- Vos JC, De Baere I, Plasterk RH (1996) Transposase is the only nematode protein required for in vitro transposition of Tc1. *Genes Dev* 10:755–761
- Xu W, Rould MA, Jun S, Desplan C, Pabo CO (1995) Crystal structure of a paired domain-DNA complex at 2.5 Å resolution reveals structural basis for Pax developmental mutations. *Cell* 80:639–650
- Yoda K, Kitagawa K, Masumoto H, Muro Y, Okazaki T (1992) A human centromere protein, CENP-B, has a DNA binding domain containing four potential alpha helices at the NH₂ terminus, which is separable from dimerizing activity. *J Cell Biol* 119:1413–1427