Amino acid sequence of a putative transposase protein of the medaka fish transposable element *Tol2* deduced from mRNA nucleotide sequences

Akihiko Koga[1], Miho Suzuki[1,2], Yuki Maruyama, Makiko Tsutsumi, Hiroshi Hori*

Division of Biological Sciences, Graduate School of Science, Nagoya University, Nagoya 464-8602, Japan

Received 2 September 1999; received in revised form 26 October 1999

Abstract ***Tol2*** **is a terminal-inverted repeat transposable element of the medaka fish** ***Oryzias latipes*****. It is one of a few elements of this class so far demonstrated to be active in vertebrates, thus providing a unique tool for establishing a gene tagging system. For the purpose of identifying its transposase, we analyzed the structures of mRNAs originating from the** ***Tol2*** **element. The results indicated that transcription of** ***Tol2*** **is initiated at several sites, the four open reading frames in** ***Tol2*** **roughly corresponding to exons, and that two main forms of mRNAs, covering exons 1–4 and exons 2–4, are present in medaka fish cells. One or both of these mRNAs are likely to encode a transposase, the amino acid sequence of which was deduced.**

Key words: Transposable element; Transposase; *Tol2*; Rapid amplification of cDNA ends; Transcription initiation site; Medaka fish

1. Introduction

Transposable elements of the terminal-inverted repeat class are thought to move in a cut-and-paste fashion. The enzyme key to their transposition is the transposase, which is, in many of the elements so far examined, encoded by the gene residing in the elements themselves. In the case of the *Activator* (*Ac*) element of maize [1,2], five exons in a 'full-length' *Ac* copy contribute to the production of a mRNA, which is translated into a transposase protein consisting of 807 amino acids [3,4].

Tol2 is a terminal-inverted repeat transposable element in the medaka fish *Oryzias latipes*, present as 10–30 copies in its genome [5]. Because its in vivo excision has already been demonstrated, it is a unique tool for establishing a gene tagging system in fish and possibly also in other vertebrate species. *Tol2* contains four open reading frames (ORFs), for 117, 352, 102 and 118 amino acids, having amino acid sequence similarity with members of the *hAT* family [6], which is a group of transposable elements including *hobo* of *Drosophila*, *Ac* of maize and *Tam3* of snapdragon. An autonomous copy of *Tol2* has already been identified by introducing its clone into zebrafish *Danio rerio*, which does not harbor *Tol2*, and observing its excision [7]. It was also shown that removal of an internal region of the *Tol2* clone leads to loss of the excision reaction [7]. These results, taken together with the observation of amino acid sequence similarity of *Tol2* ORFs to the *Ac* transposase, suggest that *Tol2* encodes its transposase just like *Ac* and other *hAT* family elements.

For the purpose of identifying the transposase of *Tol2*, with the ultimate purpose of applying *Tol2* to establishing a gene tagging system, we analyzed the structure of mRNAs representing the *Tol2* DNA sequence by cloning and sequencing cDNAs prepared from medaka fish cells. The cloning procedure included the biotinylated cap trap method [8] by which 5′ ends of mRNAs can be obtained. We here report the results, together with amino acid sequences deduced from mRNA nucleotide sequences.

2. Materials and methods

2.1. Fish samples

Fish of the orange-colored variant [9] were used. These fish contain about 20 *Tol2* copies per diploid genome [10].

2.2. Preparation and analysis of RNA

Total RNA was extracted from eight adult fish by the acid guanidinium thiocyanate-phenol-chloroform method [11]. Poly(A)-tailed RNA was isolated from the total RNA using an Oligotex-dT30 column (W9021A, Takara Shuzo, Kyoto, Japan). First-strand cDNA was synthesized using a Ready-To-Go T-Primed First-Strand kit (27-9263-01, Pharmacia Biotech, Uppsala, Sweden). To isolate *Tol2*-specific cDNAs, rapid amplification of cDNA ends (RACE) was performed with the polymerase chain reaction (PCR) primers listed in Table 1. 3′ RACE was conducted by PCR using a *Tol2*-specific primer and an adapter primer, as described previously [12], and 5′ RACE by the biotinylated cap trap method [8], which provides accurate positions of transcription initiation.

2.3. Analysis of DNA

Cloning, sequencing and hybridization experiments were conducted as described previously [13]. Nucleotide positions on the 4682 bp *Tol2* sequence (DDBJ/EMBL/GenBank accession no. D84375) are denoted NP throughout this report. For example, NP25 indicates the 25th nucleotide position of the *Tol2* sequence. The sequences of two mRNAs obtained in the present study (mRNA:1–4 and mRNA:2–4) have been deposited in the DDBJ/EMBL/GenBank with accession numbers AB031079 and AB031080.

3. Results

3.1. Comparison of amino acid sequences between Tol2 and Ac

To infer which of the four ORFs of *Tol2* could contribute to its transposase protein, we compared their amino acid sequences with that of the *Ac* ORFa protein (SWISS-PROT, P08770), which has already been demonstrated to be a transposase [3,4]. The comparison revealed four amino acid blocks having sequence similarities (Fig. 1). These blocks are included in segments 'a', 'b' and 'c' of *Ac* [14] that exhibit amino

*Corresponding author. Fax: (81) (52) 789 2974.
E-mail: hori@bio.nagoya-u.ac.jp

[1] These authors contributed equally to this work.
[2] Present address: Department of Zoology, Graduate School of Science, Kyoto University, Kyoto 606-8502, Japan.

Abbreviations: ORF, open reading frame; PCR, polymerase chain reaction; RACE, rapid amplification of cDNA ends

PII: S0014-5793(99)01479-9

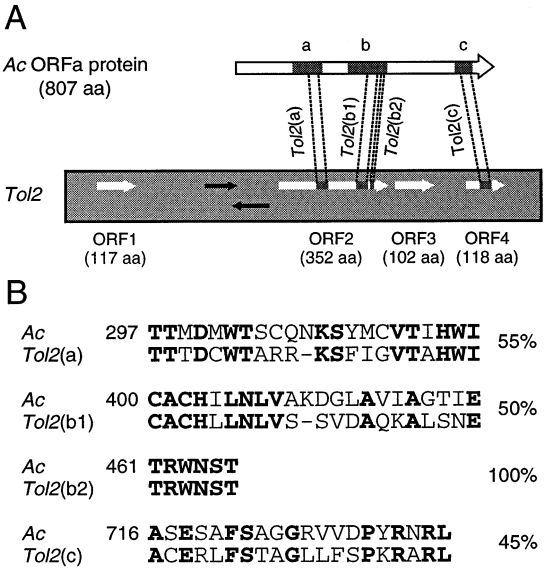

Fig. 1. Amino acid sequence similarity between the *Ac* ORFa protein and the *Tol2* ORFs. (A) Amino acid blocks exhibiting sequence similarities are schematically illustrated. The white arrows inside *Tol2* indicate the spans and direction of its ORFs. The black arrows indicate internal inverted repeats. Segments a, b and c are segments in the *Ac* ORFa protein for which amino acid sequence similarities with *hobo* and *Tam3* have been established [14]. *Tol2*(a), *Tol2*(b1), *Tol2*(b2) and *Tol2*(c) are the amino acid blocks in *Tol2*, found by comparing the *Tol2* ORFs and the *Ac* ORFa protein. (B) The sequences of the four blocks are shown together with those of the corresponding regions in *Ac*. The numbers on the left are the positions of the first amino acid residues in the *Ac* ORFa protein. Boldface letters are identical amino acids. The percentages of identical amino acids are noted on the right.

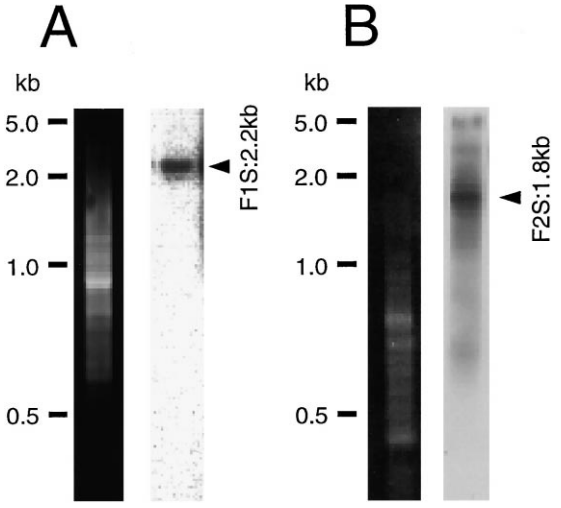

Fig. 2. 3′ RACE of medaka fish mRNA for *Tol2*-specific transcripts. First-strand cDNA synthesis from mRNA was conducted by RT-PCR with primer ST, the 3′ half of which consists of consecutive T residues that are expected to anneal to poly(A) tails of mRNA. To isolate cDNAs to which primer F1 or primer F2 can anneal, PCR was performed with each of these primers and primer S, which represents the 5′ half of primer ST. (A) PCR products with primers F1 and S. PCR was performed for 25 cycles of 94°C for 25 s, 68°C for 30 s, 72°C for 3 min. An aliquot of the product was electrophoresed on a 2.0% agarose gel (left panel) and, after transfer to a nylon membrane, hybridized with a probe for ORF2 (NP2512–NP2968) (right panel). (B) PCR products with primers F2 and S. The PCR conditions were 35 cycles of 94°C for 25 s, 53°C for 30 s, 72°C for 3 min. The product was electrophoresed (left panel) and hybridized with the same probe (right panel).

acid sequence identities of 40–65% with *hobo* and *Tam3*. The first three blocks are included in ORF2 and the last one is in ORF4. Therefore, the putative transposase protein of *Tol2* is likely to contain amino acids encoded by ORF2 and ORF4. Based on this inference, we chose two regions as *Tol2*-specific primers to be used for 3′ RACE: the 5′-terminus of ORF1 (primer F1) and the 5′-terminus of ORF2 (primer F2).

3.2. Isolation of Tol2-specific cDNAs by 3′ RACE

3′ RACE with primers F1 and S (with a sequence artificially added to the poly(A) tail) and subsequent Southern hybridization analysis resulted in a single band of 2.2 kb (Fig. 2A). The same analysis with primers F2 and S revealed a single band of 1.8 kb (Fig. 2B). The DNA fragments responsible for these two bands were cloned into the pBluescript vector and designated F1S:2.2kb and F2S:1.8kb, respectively. Because there is a possibility of misincorporation of nucleotides with reverse transcription (RT-) PCR and its subsequent nested PCR and DNA sequencing, five clones from independent RT-PCR reactions were prepared for both F1S:2.2kb and F2S:1.8kb.

Table 1
PCR primers used in the present study

Primer	Sequence	Target[a]
Primers for 3′ RACE		
F1	GAAGTGACGTCATGTCACATCTATTACCAC	338–367 (ORF1)
F2	AATGCACCCAAATTACCTCAAAAACTACTCT	2167–2196 (ORF2)
ST	AACTGGAAGAATTCGCGGCCGCAGGAA(T_{18})	Poly(A)
S	GGAAGAATTCGCGGCCGCAGG	Primer ST
Primers for 5′ RACE		
R1x	CAGGTCAAGGTGCTGTGC	388–371 (ORF1)
R1y	GGTAATAGATGTGACATGACG	365–345 (ORF1)
R1z	TCACTTCCAAAGGACCAATGAAC	344–321 (ORF1)
R2x	AACTGAGTCAACTTTCAGTTG	2275–2255 (ORF2)
R2y	GCTTACTGCTGGAAGCATGG	2255–2236 (ORF2)
R2z	CGATCTTTCTCTTCTGTGCTGTC	2225–2202 (ORF2)
ABC	GTTACAGCTGGAGGGATGTTGGGCTTAAGGATG(C_{18})	Poly(G)
A	GTTACAGCTGGAGGGATGTTG	Primer ABC
B	GAGGGATGTTGGGCTTAAGGATG	Primer ABC

[a]Numbers indicate nucleotide positions in the 4682 bp *Tol2* DNA sequence.

3.3. Sequences of the 3′ RACE products

Every clone contained 1–4 nucleotides different from the corresponding nucleotides in the *Tol2* DNA sequence. However, the consensus sequences of the respective five clones did not contain mismatches from the *Tol2* DNA sequence. We use the consensus sequences as those of F1S:2.2kb and F2S:1.8kb.

The following results were obtained for the F1S:2.2kb sequence. (1) It contains a reading frame that covers most of the entire sequence. (2) It is composed of four exons which roughly correspond to the four ORFs. (3) Three introns are spliced out essentially according to the consensus sequence of the splicing site [15]. (4) A poly(A) tail is added to NP4493, 123 bp after the stop codon 'TGA'. (5) A six nucleotide block (ATTAAA) which resembles the consensus poly(A) addition signal [16] is present 18 bp before the poly(A) tail.

The sequence of clone F2S:1.8kb proved to be identical to the last 1.8 kb portion of clone F1S:2.2kb.

3.4. Isolation of 5′ RACE products

3′ RACE does not provide information on the sequence of upstream regions beyond the target-specific primers (F1 or F2 in the present study). We went on to clone the 5′-terminus regions of *Tol2*-specific cDNAs by the cap trap method [8] and subsequent nested PCRs. In the analysis of the upstream region of ORF1, 10 fragments of different lengths were observed (Fig. 3A), indicating that there are at least 10 different transcripts for which the transcription initiation sites are upstream of ORF1. However, with eight of the 10 products, the lengths were found to be apparently larger than the distance between primer R1z and the 5′ end of *Tol2* (0.4 kb), suggesting that they are transcribed from outside *Tol2*. We cloned and sequenced the two shortest products, with lengths of about 0.4 and 0.2 kb, and designated them as CR1:0.4kb and CR1:0.2kb, respectively.

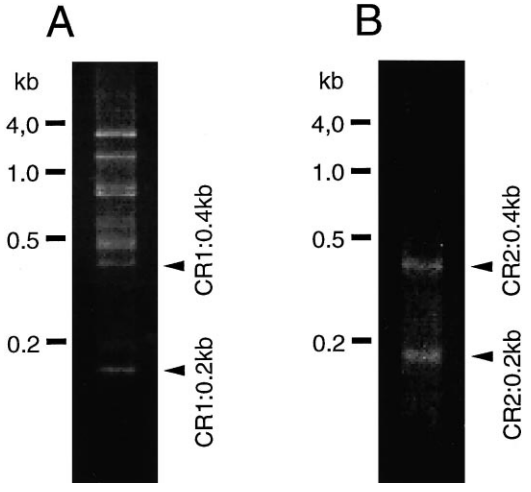

Fig. 3. 5′ RACE of medaka fish mRNA for *Tol2*-specific transcripts. (A) PCR products with the primers for ORF1. The first step of the 5′ RACE was RT-PCR to synthesize first-strand cDNA with the *Tol2*-specific primer R1x. Next, mRNA-cDNA hybrid molecules which possess a cap structure were isolated by the method described [8]. After removing the RNA portion, poly(G) was added to the 3′ ends of the first-strand cDNAs. The second strand was then synthesized using primer ABC that contains a poly(C) segment to anneal to poly(G). Finally, two rounds of nested PCRs were conducted: the first round with primers A and R1y (35 cycles of 94°C for 25 s, 50°C for 30 s, 72°C for 30 s), the second round with primers B and R1z (30 cycles of 94°C for 25 s, 57°C for 30 s, 72°C for 30 s). An aliquot of the product was electrophoresed on a 2.0% agarose gel. (B) PCR products with the primers for ORF2. The same analysis was conducted by substituting primers R1x, R1y and R1z with R2x, R2y and R2z, respectively.

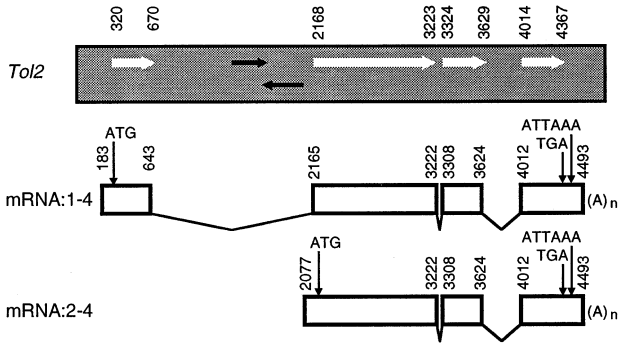

Fig. 4. Schematic illustration of the structures of the putative mRNAs. The numbers indicate nucleotide positions (NP) of the first and the last nucleotides of the ORFs in *Tol2* (white arrows) and the exons in the mRNAs (open boxes).

In the analysis of the upstream region of ORF2, two fragments of 0.4 and 0.2 kb were observed (Fig. 3B). We cloned and sequenced these two fragments and designated them as CR2:0.4kb and CR2:0.2kb, respectively.

3.5. Sequences of the 5′ RACE products

It is often the case, for transposable elements, that their sequences are represented in read-through transcripts of adjacently or closely located genes. Of the four 5′ RACE products, two (CR1:0.4kb and CR2:0.4kb) proved to have such features. The other two products (CR1:0.2kb and CR2:0.2kb) were revealed to have transcription initiation sites inside the *Tol2* element.

The sequence of CR1:0.2kb was identical to the nucleotides at NP183–NP344, except for a single nucleotide. The mismatched nucleotide is residue T at the first position of this clone, being A (NP183) in the *Tol2* sequence. This may have arisen from an error in PCRs, as observed in the 3′ RACE. The sequence of CR2:0.2kb was identical to the nucleotides at NP2077–NP2225 with no mismatched nucleotides.

3.6. Possible sequences of mRNAs and proteins

CR1:0.2kb was identified as a 5′-terminus of a mRNA transcribed from upstream of ORF1 and F1S:2.2kb was isolated as the 3′ portion of a mRNA containing ORFs 1–4. We combined these two sequences and designated the combination as mRNA:1–4. Similarly, we combined the sequences of CR2:0.2kb and F2S:1.8kb and designated it as mRNA:2–4. The sequences of these two hypothetical mRNAs, together with deduced amino acid sequences, have been deposited in the databases (see Section 2). Their overall structures, including nucleotide positions of exon-intron boundaries, are shown in Fig. 4.

4. Discussion

4.1. mRNA transcribed from Tol2

The fact that we could obtain cDNAs originating from *Tol2* indicates that the *Tol2* ORFs are transcribed into mRNAs in medaka fish cells. The four ORFs were revealed to roughly

correspond to the exons. Our discussion hereafter deals with the exons rather than the ORFs.

4.2. Transcription initiation sites

The transcription initiation sites for mRNA:1–4 and mRNA:2–4 were determined to be NP183 and NP2077, respectively. It is obvious that there were more mRNAs carrying longer sequences at their 5′ ends (see Fig. 3). A plausible explanation for the presence of such mRNAs is that their transcription is initiated outside *Tol2*, probably due to promoters located adjacent or close to *Tol2* at different chromosomal locations. Evidence for the presence of mRNA:2–4 is relatively weak because the possibility of F2S:1.8kb being produced from mRNA:1–4 cannot be ruled out.

TATA box-like sequences were not found in the vicinity of NP183 and NP2077. It remains to be determined what factors work as promoters for these mRNAs.

4.3. Possible amino acid sequence of the Tol2 transposase

mRNA:1–4 is 2319 bp long and contains a reading frame of 685 amino acids. mRNA:2–4 is 1946 bp in length and encodes 576 amino acids in a single reading frame. These two putative proteins include all of the four amino acid blocks exhibiting sequence similarities with the *Ac* transposase protein (see Fig. 1). Thus, one or both of them are likely to be a transposase for *Tol2*. It is also possible that the two proteins are competitive in function, one being a transposase and the other being a repressor.

4.4. Variation in Tol2

It was suggested by the 5′ RACE that several transcripts are produced from different *Tol2* copies. However, the 3′ RACE yielded only a single product with primers F1 and S (F1S:2.2kb) and also a single product with primers F2 and S (F2S:1.8kb). In addition, these 3′ RACE products were 'complete' sequences composed of exons 1–4 and exons 2–4, respectively. This is consistent with our previous findings for the structure of *Tol2* copies [10]: no restriction map variation was evident among more than 200 *Tol2* copies and no sequence variation among five randomly chosen *Tol2* clones. We have proposed, therefore, that all or most *Tol2* copies are 'full-length' and potentially autonomous [10]. The contrast in the number of products between the 5′ RACE and the 3′ RACE may be due to the presence of several mRNAs with differences in their 5′ regions but with identity in their bodies.

4.5. Possibilities for further analysis

We have already constructed clones which represent the mRNA:1–4 and mRNA:2–4 sequences by combining the 5′ RACE and 3′ RACE products. In vitro transcription and in vitro translation from these clones were also successful. Another tool required, which is near completion in our lab, is an experimental system to accurately determine the excision frequency. This system should be useful for confirming the transposase function of the in vitro mRNA and protein products and, in addition, for their analysis. We are ready to deliver our clones on request.

Acknowledgements: We are grateful to H. Inagaki, K. Kawakami and H. Ohtsubo for helpful discussions. This work was partly supported by Grant no. 10216025 to A.K. and no. 08640786 and 09554053 to H.H. from the Ministry of Education, Science, Sports and Culture of Japan and also by the Takeda Science Foundation to A.K.

References

[1] McClintock, B. (1948) Carnegie Inst. Wash Year Book 47, 155–169.
[2] Fedoroff, N.V., Wessler, S. and Shure, M. (1983) Cell 35, 235–242.
[3] Kunze, R., Stochaj, U., Laufs, J. and Starlinger, P. (1987) EMBO J. 6, 1555–1563.
[4] Coupland, G., Baker, B., Schell, J. and Starlinger, P. (1988) EMBO J. 7, 3653–3659.
[5] Koga, A., Suzuki, M., Inagaki, H., Bessho, Y. and Hori, H. (1996) Nature 385, 30.
[6] Atkinson, P.W., Warren, W.D. and O'Brochta, D.A. (1993) Proc. Natl. Acad. Sci. USA 90, 9693–9697.
[7] Kawakami, K., Koga, A., Hori, H. and Shima, A. (1998) Gene 225, 17–22.
[8] Carninci, P., Kvam, C., Kitamura, A., Ohsumi, T., Okazaki, Y., Itoh, M., Kamiya, M., Shibata, K., Sasaki, N., Izawa, M., Muramatsu, M., Hayashizaki, Y. and Schneider, C. (1996) Genomics 37, 327–336.
[9] Hirose, E. and Matsumoto, J. (1993) Pigment Cell Res. 6, 45–51.
[10] Koga, A. and Hori, H. (1999) Genet. Res. Camb. 73, 7–14.
[11] Chomczynski, P. and Sachi, N. (1987) Anal. Biochem. 162, 156–159.
[12] Inagaki, H., Bessho, Y., Koga, A. and Hori, H. (1994) Gene 150, 319–324.
[13] Koga, A., Inagaki, H., Bessho, Y. and Hori, H. (1995) Mol. Gen. Genet. 249, 400–405.
[14] Feldmar, S. and Kunze, R. (1991) EMBO J. 10, 4003–4010.
[15] Mount, S.M. (1982) Nucleic Acids Res. 10, 459–472.
[16] Sheets, M.D., Ogg, S.C. and Wickens, M.P. (1990) Nucleic Acids Res. 18, 5799–5805.