The Fish Biology Journal MEDAKA, 7: 11–14 (1995)

Characteristics of medaka genes and their promoter regions

Minoru Tanaka

Reproductive Biology, National Institute for Basic Biology, Myodaiji-cho, Okazaki 444, Japan.

(Received September 14, 1995)

Abstract Here I briefly describe the characterization of the medaka (*Oryzias latipes*) genome. The size of the medaka genome was estimated to be 680–850 megabases by comparing the medaka genome with that of the pufferfish. This value corresponds to one-fourth to one-fifth of the entire human genome, probably due to short introns and short intergenic regions. However the size distribution of medaka introns so far examined appears similar to that of human introns. The location of the introns is conservative between human and medaka. Moreover some important *cis*-elements and *trans*-acting factors functioning in the promoters of medaka genes also seem to be conservative between the two species. This allows medaka to be a suitable model for genomic analysis as a miniature vertebrate.

Introduction

Development of molecular biology techniques has enabled scientists to attempt elucidating nucleotide sequence of the entire genome of several species. Determination of the entire genomic nucleotide sequence in species having a relatively small genome such as *Escherichia coli, Saccharomyces cerevisiae, Caenorhabditis elegans* and *Arabidopsis thaliana* is in progress. Although a human genome project is also in progress, more time will be needed to complete the determination because of the length and complexity of the human genome.

Recently Brenner *et al.* (1993) proposed the pufferfish (*Fugu rubripes*) as a model for vertebrate genome analysis. They showed that Fugu has 0.4–0.5 pg DNA per haploid genome and is estimated to be 400 Mb (megabases) in genomic length. This is equivalent to 1/7.5 of the entire length of the human genome and is 29 times longer than the yeast genome. Unique sequences occupy more than 90% of the entire Fugu genome. These facts promise Fugu to be one of the ideal models for genome analysis which phylogenetically locates between human and other organisms.

What about medaka genomic sequences?

Several methods to produce transgenic medaka have been devised and the function of the introduced genes was assayed using exogenous promoters such as the chicken δ-crystallin promoter (Ozato *et al.*, 1986). ES-like cell lines have been developed in medaka (Ozato *et al.*, 1994). These situations demand more information on genomic structure of medaka, which is useful for cloning genes and constructing expression vectors in medaka. Here I briefly survey the characteristics of the medaka genomic structure.

Medaka genome size

Uwa and Iwata (1981) reported medaka (*Oryzias latipes*) karyotype shows 48 chromosomes (2n = 48) consisting of one pair of subtelocentrics and 23 pairs of acrocentrics. They further showed that the amount of DNA per nucleus is calculated to be 1.7 pg when carp erythrocyte nucleus was used as a standard. Based on this result, the entire length of the medaka genome is estimated to be 680–850 Mb, which corresponds to one-fourth to one-fifth of the entire human genome.

Medaka gene structure

The number of isolated genes from medaka is now increasing. As a typical medaka gene which is comparable to the human counterpart, I describe the structure of the medaka *P*-450 aromatase gene (human gene name; *CYP19*) (Tanaka *et al.*, 1995). The *P*-450 aromatase (abbreviated as *P*-450arom) gene encodes an enzyme which catalyzes conversion of androgen to estrogen and belongs to a large *P*-450 gene family. The deduced amino acid sequence of medaka *P*-450arom shows 51% homology with that of human (Means *et al.*, 1989; Harada *et al.*, 1990). Since homologies of other genes so far cloned are between 50 and 70%, *P*-450arom can be said to be a conventional and ordinary gene.

Comparison of *P*-450arom gene structures between human and medaka is shown in Fig. 1.

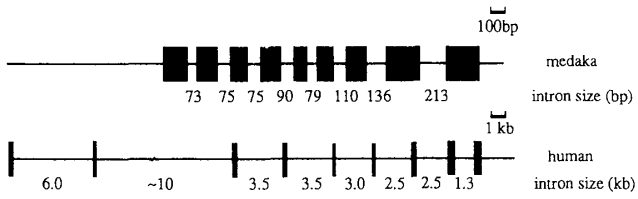

Fig. 1. Comparison of *P*-450arom gene organization between medaka and human. The black boxes indicate exons. The numbers below the lines show the size of introns. Note that the size of medaka introns is represented by base pair units whereas human introns by kilobase pairs (Means *et al.*, 1989; Harada *et al.*, 1990).

Sequencing of the whole medaka gene showed that the basic organizations of the *P*-450arom genes were identical to each other. The coding region of *P*-450arom gene in both medaka and human consists of nine exons. The medaka introns are inserted exactly at the same positions as those of human at the nucleotide level. The 5' and 3' ends of the medaka intron sequences conform to the GT/AG rule as described by Breathnach and Chambon (1981) (medaka consensus sequence deduced from *P*-450arom; 5' GTRRRY----Y-YYYYAG 3', where R and Y indicate purine and pyrimidine residues, respectively). However, there are also great differences between the two genes. Medaka *P*-450arom spans only 2.5 kb on the genome whereas the human counterpart spreads over in more than 70 kb. This indicates that medaka *P*-450arom consists of very small introns. The smallest one is 73 bp and the largest one is only 213 bp. Human *P*-450arom introns range from 1.3 kb to more than 10 kb. This is not an exceptional case for the medaka genome. Another steroidogenic *P*-450 gene, *P*-450c17, also contains extremely small introns when compared with human counterpart (Matsuyama, unpublished). These examples indicate that medaka genes are a miniature version of the human gene due to harboring smaller introns.

This discussion raises another question. Does only the small size of intronic sequences contribute to the smaller size of the entire medaka genome? I summarized the size distribution of medaka introns so far obtained and compared it with that of human introns (Ogata, unpublished) in Figs. 2 and 3. Surprisingly, the figures show that human genes with small introns appear as frequently as medaka genes and the distribution profile seems almost similar between the two species. The frequency of less than 200 bp introns is high in both species. As the intron sizes become larger, the number of the introns becomes less frequent. This suggests that intergenic sequences are much shorter in medaka genome than in human genome. However, the pool of medaka genes is not large enough to allow for conclusive analysis. Recently the genomic structure of the Huntington's disease gene was published by the Fugu genome project. The Fugu gene is 7.4 times shorter (23 kb) than human counterpart (170 kb) (Baxendale *et al.*, 1995). Assuming that Fugu and medaka have the same number and similar size of genes, 40–55% of medaka total genome is calculated to be occupied by unique sequences.

Considering the examples of medaka genes having smaller introns, the distribution profile implies that genes having extremely long introns occur more frequently in the human genome than in medaka genome, which might contribute to the differences of the genome size between these species.

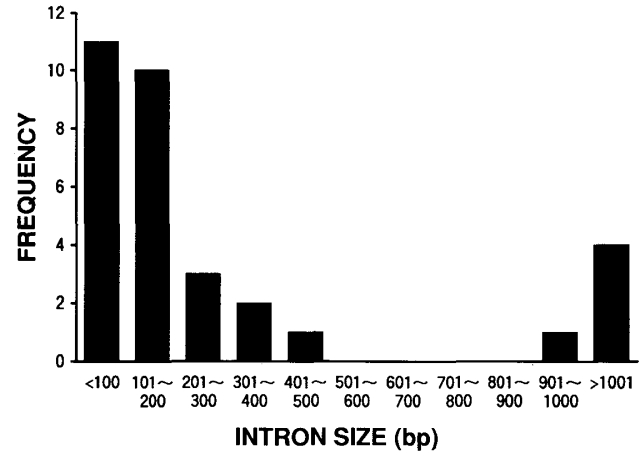

Fig. 2. Distribution of the size of medaka introns. The vertical axis indicates the frequency of each intron size. The total number of medaka introns surveyed is 33.

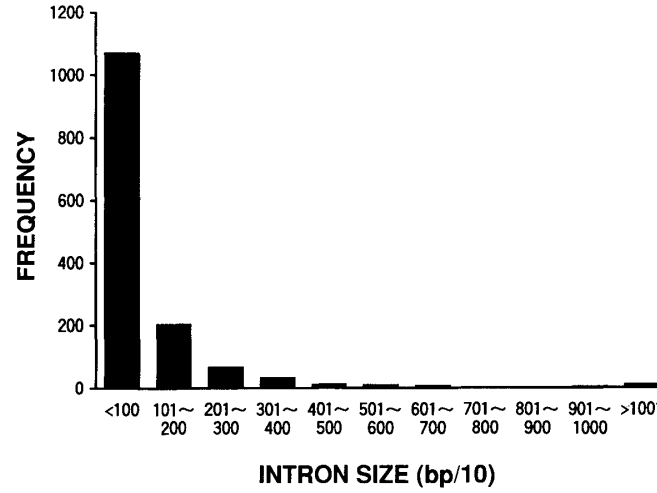

Fig. 3. Distribution of the size of human introns. The vertical axis indicates the frequency of each intron size. The unit of the intron sizes is 10 base pairs. The total number of human introns is 1425 which are compiled from Genbank Release 86.0.

Promoter structure

There are few examples showing the precise characterization of medaka gene promoters, although the number of examples is expected to increase. We characterized the promoter structure of medaka for the 5' upstream region of *P*-450arom gene by primer extension analysis and S1 nuclease mapping (Tanaka *et al.*, 1995). Fig. 4 illustrates the promoter sequence and exon 1 of medaka *P*-450arom.

Mammalian steroidogenic *P*-450 genes require a *trans*-acting factor, called SF-1/Ad4BP, to regulate their expression. This factor binds to the Ad4 motif (YCAAGGTYA) conferring the responsiveness to intracellular cAMP levels. Since medaka *P*-450arom activity in isolated ovarian follicles is enhanced by the increase of intracellular cAMP levels, we searched for cAMP responsive elements. Surprisingly, 107 and 135 bp upstream from the transcriptional initiation site we found motifs perfectly matching the mammalian consensus sequence (Fig. 4 and Table 1). Moreover, another two tandemly repeated hexameric sequences (TGACCT/A) are found 40 and 51 bp upstream from the transcriptional initiation site. This motif is complementary to the Ad4 motif and is the half-site of the estrogen responsive element (ERE-half). When the medaka promoter region is compared with the human *P*-450arom promoter region, a moderate homology (60–80%) is locally recognizable at the nucleotide level. This may suggest the occurrence of other possible *cis*-elements which are functionally conservative between human and medaka.

The result of the promoter sequence analysis encouraged us to clone SF-1/Ad4BP homologues from medaka ovarian follicular cells, where *P*-450arom is expressed. Conventional polymerase chain reaction technique successfully allowed us to clone putative SF-1/Ad4BP cDNA fragments which share more than 90% homology in DNA binding domain with mammalian counterpart (Tanaka, unpublished). It is likely that medaka SF-1/Ad4BP-like *trans*-acting factor functions to bind to conservative SF-1/Ad4BP motifs found in the *P*-450arom promoter. These results, together with the occurrence of the SF-1/Ad4BP consensus sequence, suggest that functionally important

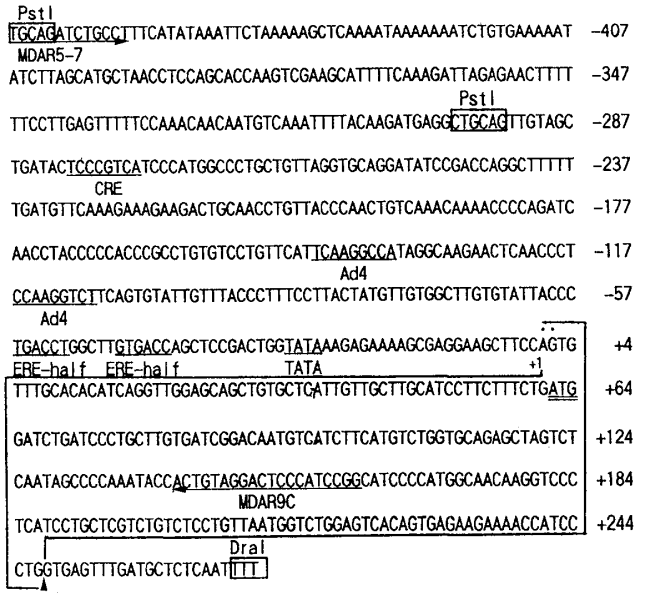

Fig. 4. Nucleotide sequence of the promoter region and exon 1 of medaka ovarian *P*-450arom (Tanaka *et al.*, 1995). Nucleotide sequence enclosed by a black line is exon 1. Dots above the nucleotides show major transcriptional initiation sites, the first nucleotide of which is numbered at +1. Putative TATA box (TATA), SF-1/Ad4BP binding motif (Ad4), half-sites of estrogen responsive element (ERE-half) and cAMP-responsive element (CRE) are underlined. Horizontal arrows with the names of MDAR5-7 and MDAR9C indicate the location of primers used for determination of transcriptional initiation sites. A putative translational initiation codon is doubly underlined. The boundary of intron and exon is indicated by a reverse triangle.

Table 1. Consensus sequence of SF-1/Ad4BP motif. The names of genes and their coding proteins are as follows. *P*-450c17; cytochrome *P*-450 17α-hydroxylase/lyase, *P*-45011β; cytochrome *P*-450 11β-hydroxylase, *P*-450scc; cytochrome *P*-450 cholesterol side chain cleavage enzyme, *P*-450arom; cytochrome *P*-450 aromatase. + indicates that SF-1/Ad4BP experimentally binds to the sequence. The other sequences were surveyed and compiled from the putative steroidogentic *P*-450 promoters. Modified from Morohashi *et al.* (1992).

Gene	Source	Consensus Sequence (C/T) CAAGGT (C/T)A
P-450c17	human	TCAAGGTGA +
	bovine	AGAAGGTCA +
		ACAAGGTGA
p-45011β	human	GAGAGGTCA
		CAGAGGTCA
		CCAAGGCTC
	bovine	CCAAGGACC +
		CCAAGGTCT +
		GGAAGGGCA +
	mouse	CCAAGGCTC
P-450scc	human	TCAAGGCCA +
		CCAAGGTGA +
		CCAAGGTCT +
		TCAAGGTCA +
P-450arom	human	GGGAGGTCA
		AGGAGGTCA +
	rat	CCAAGGTCA
	medaka	CCAAGGTGT
		TCAAGGCCA

regions of the promoter are conserved between medaka and mammals and that the identification of basic regulatory regions between *trans*-acting factors and *cis*-elements can be assessed by direct sequence comparison between medaka and other vertebrates.

Actually Brenner's group found the functionally regulatory elements conserved between Fugu and mouse by identification of possible *cis*-elements by the direct sequence comparison first, followed by producing transgenic mice with Fugu promoter constructs (Aparicio *et al.*, 1995, Popperl *et al.*, 1995). The deletion of putative Fugu *cis*-elements resulted in a tissue-specific loss of expression in transgenic mice.

It is not surprising to find a TATA box 25 bp upstream from the transcriptional initiation site of medaka *P*-450arom gene since TATA box binding protein (TBP) is known to exit from yeast to human.

As described above, the size of the medaka genome is small probably due to short introns and short intergenic sequences. The functionally important *cis*-elements in medaka promoters are possibly identifiable by sequence comparison with mammalian corresponding regions and are expected to be functionally compatible to those in mammalian promoters. This possibility suggests that medaka is an outstanding model suitable for genomic analysis.

Acknowledgments

I would like to thank Prof. Nagahama for giving me a chance to write this manuscript. I also acknowledge Dr. Ogata in Kyoto University and Dr. Nakai in Osaka University in preparation of Fig. 3. Finally I thank Dr. M.S. Grober for critical reading of this article. This study is supported in part by Grant-in-Aid (07740638) from the Ministry of Education, Science, Sport and Culture of Japan.

References

Aparicio, S., A. Morrison, A. Gould, J. Gilthorpe, C. Chaudhuri, P. Rigby, R. Krumlauf and S. Brenner (1995) *Proc. Natl. Acad. Sci. USA,* **92**: 1684–1688.

Baxendale, S., S. Abdulla, G. Elgar, D. Buck, M. Berks, G. Micklem, R. Durbin, G. Bates, S. Brenner and H. Lehrach (1995) *Nature Genetics*, **10**: 67–76.

Breathnach, R. and P. Chambon (1981) *Ann. Rev. Biochem.,* **50**: 349–383.

Benner, S., G. Elgar, R. Sandford, A. Macrae, B. Venkatesh and S. Aparicio (1993) *Nature,* **366**: 265–268.

Harada, N., K. Yamada, K. Saito, N. Kibe, S. Dohmae and Y. Takagi (1990) *Biochem. Biophys. Res. Commun.,* **166**: 365–372.

Means, G.D., M.S. Mahendroo, C.J. Corbin, J.M. Mathis, F.E. Powell, C.R. Mendelson and E.R. Simpson (1989) *J. Biol. Chem.,* **264**: 19385–19391.

Morohashi, K., S. Honda, Y. Inomata, H. Handa and T. Omura (1992) *J. Biol. Chem.*, **267**: 17913–17919.

Ozato, K., H. Kondoh, H. Inohara, T. Iwamatsu, Y. Wakamatsu and T.S. Okada (1986) *Cell. Differ.,* **19**: 237–244.

Ozato, K. and Y. Wakamatsu (1994) *Dev. Growth Differ.*, **36**: 437–443.

Popperl, H., M. Bienz, M. Studer, S.-K. Chan, S. Aparicio, S. Brenner, R.S. Mann and R. Krumlauf (1995) *Cell,* **81**: 1031–1042.

Tanaka, M., S. Fukada, M. Matsuyama and N. Yoshitaka (1995) *J. Biochem.,* **117**: 719–725.

Uwa, H. and A. Iwata (1981) *Chrom. Inf. Serv.,* **31**: 24–26.