Volume 11 Number 5 1983 Nucleic Acids Research

Complete nucleotide sequence of a cloned cDNA derived from the major adult α-globin mRNA of *X. laevis*

R.M.Kay*, R.Harris, R.K.Patient+ and J.G.Williams

Imperial Cancer Research Fund, Mill Hill Laboratories, Burtonhole Lane, London, NW7 1AD, UK

Received 1 February 1983; Accepted 14 February 1983

ABSTRACT

The complete sequence of a cloned cDNA derived from the major adult α-globin mRNA of Xenopus laevis (the South African Clawed Toad) is presented. The sequence contains the complete coding and 3' non-coding regions of the mRNA and part of the 5' non-coding region. The amino acid sequence of the encoded α-globin polypeptide has been deduced and is compared to other α-globin polypeptides. We find that the sequence is equally diverged from a bullfrog tadpole α-globin polypeptide and human α-globin polypeptide suggesting that these three sequences may have diverged from a common ancestral sequence several hundred million years ago.

INTRODUCTION

The blood of the adult Xenopus laevis, contains two major globin polypeptides and at least four minor species (1,2). We have previously reported the isolation, by cDNA cloning, of almost complete DNA copies of the mRNA sequences coding for the two major adult globins and have assigned one to the α-globin family and the other to the β-globin family on the basis of conserved C-terminal amino acid sequence (3). The complete sequence of the β-globin cDNA has been reported confirming its designation (4). This paper reports the complete sequence of the cloned cDNA encoding the major adult α-globin polypeptide of X.laevis. The sequence is shown to contain the complete coding and 3' non-coding regions of the mRNA and part of the 5' non-coding region. The complete α-globin amino acid sequence is thus deduced.

Since the amphibian ancestor of X.laevis diverged from the mammalian and avian lines several hundred million years ago, it is of interest to study the evolutionary conservation of the mRNA and protein sequences of functionally related genes. Thus we have compared the X.laevis sequence with the available α-globin mRNA sequences from human (5), rabbit (6), mouse (7), and chicken (8) sources. The availability of a much larger collection of α-globin polypeptide sequences (9) has made it possible to compare the X.laevis α-globin sequence with homologous sequences from mammals, birds,

reptiles and fish as well as other amphibia (9,10).

Although this paper reports the first complete X.laevis α-globin coding sequence, two previous reports by other workers describing partial mRNA (11) and gene (12) sequences have appeared. The various sequences differ in a number of places and these differences are assessed.

MATERIALS AND METHODS

The recombinant plasmid pXG 6C1 containing a cDNA copy of the major adult X.laevis α-globin mRNA has previously been described (3). Growth and purification of the plasmid DNA were performed as previously reported (3).

Preparation of DNA fragments for sequence determination. Restriction enzyme digests of the pXG 6C1 DNA were end labelled by one of the following methods. Fragments containing protruding 5'-ends or blunt ends were ^{32}P-labelled at their 5'-ends with T_4 polynucleotide kinase and γ^{32} P-ATP either by an exchange reaction (13) or following dephosphorylation with calf intestinal alkaline phosphatase (14). Fragments containing overlapping 5'-ends were labelled at their 3'-termini with reverse transcriptase using one α^{32} P-dNTP and three unlabelled dNTPs. Labelling of 3'-termini with blunt ends or over-lapping 3'-ends was accomplished using T_4 polymerase and α^{32} P-dNTP in the presence of 3 unlabelled dNTPs (15). Individual labelled restriction fragments were separated by agarose gel electrophoresis and purified by electroelution from gel slices. Following an appropriate second restriction enzyme digest, fragments to be sequenced were finally purified on preparative urea/acrylamide gels.

Sequence determination. The nucleotide sequence of appropriately labelled restriction fragments was determined by the Maxam and Gilbert sequencing reactions (16) followed by analysis on thin (0.4mm) urea/acrylamide gels of 5%, 10% and 20% acrylamide. Fig. 1 shows the restriction sites used in the sequencing and the extent of sequence determined from each site. The complete sequence was established by confirming each nucleotide in both DNA strands and sequencing across all restriction sites (with one exception, see Fig. 1).

RESULTS AND DISCUSSION

The coding sequence. The complete nucleotide sequence (556 residues) of the coding strand of the globin cDNA insert in pXG 6C1 plasmid is shown in Fig. 2. The orientation of the sequence had previously been established (3) and was reconfirmed by the identification of a poly (dA) sequence adjacent to the poly (dC) linker at one end of the insert. Examination of

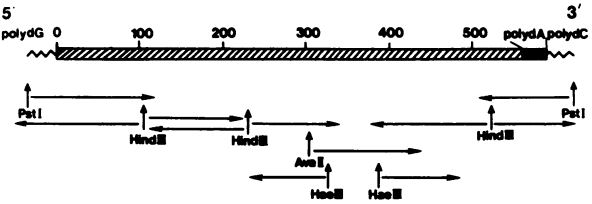

Figure 1 Restriction enzyme cleavage map for the X.laevis α-globin cDNA insert. Sites shown are those used to determine the complete DNA sequence of this region. Arrows indicate the direction and distance of sequence determination performed on both strands with the exception of the sequencing from the PstI restriction enzyme sites. This was performed only on 3' end-labelled DNA. The HindIII restriction site at nucleotide 228 was the only site which was not sequenced across.

the coding strand revealed only one open reading frame capable of directing the synthesis of a globin polypeptide. This frame starts at the ATG initiation codon nearest the 5' end and remains open to code for 141 amino acids, closing with the termination triplet TAA.

Comparison of this amino acid sequence with other known globin sequences confirms that the sequence is that of an α-globin polypeptide. The sequence can be aligned exactly with the N- and C-terminal amino acids of mammalian, avian and amphibian adult α-globin sequences (9). This alignment reveals that 77 amino acids (55%) are identical to those present in human α-globin, 70 amino acids (50%) are identical to those found in both newt and chicken α-globins and 66 amino acids (47%) are identical to those found in viper α-globin. Furthermore, of the 28 amino acids invariant among all α-globins, including known fish, 27 are also conserved in X.laevis. The amino acid at position α51 is changed from glycine to asparagine. By contrast, if the α-globin sequence is aligned for maximum homology with the known β-globins (taking account of the 3 insertion/deletion events assumed by Dayhoff (9), the X.laevis α-globin shows only 55 amino acids (39%) identical to those present in the human β-globin and 46 amino acids (33%) identical to X.laevis β-globin. Of the highly conserved amino acids in β-globin (not including fish) only 24 out of 41 (58%) are also conserved in the X.laevis α-globin. Thus we are confident in our assignment of this sequence as a member of the α-globin family.

The comparisons above show that the X.laevis α-globin chain is almost equally diverged from mammalian, for example the human, α-globin chain and from another amphibian, the newt α-globin chain. This suggests

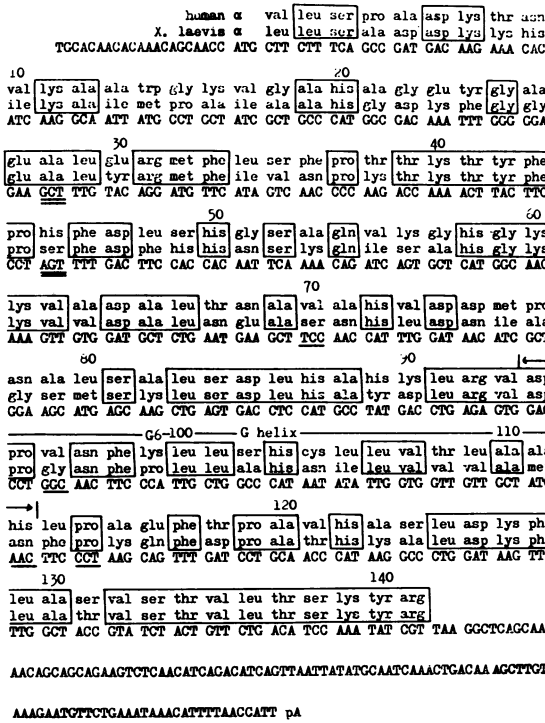

Figure 2 The complete nucleotide sequence of the coding strand of the cloned adult X.laevis major α-globin cDNA. Immediately above the sequence is the derived α-globin amino acid sequence and the human α-globin sequence is presented for comparison. Identical amino acids are boxed and the G helical region of the α-globin is indicated. The amino acid G6 (α99) is identified as the unusual residue unique to amphibia (see text). Several differences from previously published partial sequences of this mRNA are underlined. Those codons underlined once differ from Richardson et al. (11) and those underlined twice differ from Partington and Barelle(12). At codons 28/29 we have read an extra T residue which is absent in the published sequence but which must be present in order to assure the correct phase of the coding sequence. At codon 45 our sequence reads AGT as opposed to AAT in the published sequence. This difference may reflect polymorphism in the coding sequence. At codon 70, an A residue is reported between the two C residues in the published sequence. Again, this would cause an unacceptable phase change in the coding sequence. At codon 96, our sequence reads GGC as opposed to AGC in the published sequence. Polymorphism could account for this difference but it should be noted that this G residue in our sequence lies at an EcoRII methylation site and can only be confirmed by reading the sequence in both DNA strands. At codon 112 a T residue lies between the A and C residues in the published sequence and phase is restored by loss of a T residue in codon 114. Polymorphism is unlikely to be the cause of this two-fold difference. Partington and Barelle (12) have reported the same sequence as ours in this region. The 3' terminal sequence contains 9 residues more in our sequence than that reported by Richardon et al. and this last difference is discussed more fully in the text.

that these three α-chains have been diverging separately for several hundred million years from a common ancestral gene. The sequence of a bullfrog tadpole α-globin chain has been reported (10). When this sequence is compared to the X.laevis α-chain 78 amino acids (55%) are identical. A similar degree of homology (56%) was observed when comparing the tadpole α chain with the 71 amino acids of known sequence from the bullfrog adult α chain (10). Thus the gene duplications giving rise to the distinct tadpole and adult hemoglobins may have occured at or before the divergence of amphibian and mammalian lines from a common ancestor at least 350 million years ago (17).

Maruyama et al. (10) noted that, in the bullfrog tadpole α-chain, an unusual freature of the sequence was the presence of a proline (α99) at position 6 of the G helix (Fig. 2). This proline residue is also present in the X.laevis α-chain. The only other known α-globin polypeptide containing a proline at this position is the newt α-chain. In all other α-globin sequences including fish (9), this residue is a lysine.

The 5' non-coding region. The 5' non-coding region of the sequence presented here contains 21 nucleotides. During preparation of the double stranded cDNA, the 3' terminal sequence is used to self-prime second strand synthesis (19) resulting in a hairpin structure which is subsequently cleaved by S1 nuclease. Thus, inevitably, the extreme 5' terminal sequence is lost. Comparison of the available 5' non-coding sequence with other α-globin mRNA's reveals only one potentially significant conservation of sequence - trinucleotide ACC prior to the ATG initiation codon.

The 3' non-coding region. The presence of a complete 3' non-coding region in the cDNA is confirmed by the presence of a 30 nucleotide poly (dA) stretch adjacent to the poly (dC) linker. The region contains 110 nucleotides following the termination triplet and compares with 109 nucleotides in man (5) 101 in mouse (7) 96 in chicken (8) and 86 in rabbit (6). The only obvious region of conserved sequence is the hexanucleotide AATAAA close to the 3' poly (dA). This sequence is found in almost all eukaryotic mRNAs and is thought to direct the terminal addition of poly A. The distance of this sequence from the poly (dA) is somewhat shorter in X.laevis (13 nucleotides) than human and rabbit (15 nucleotides), and chicken (19 nucleotides). Richardson et al. (11) have reported the partial sequence of the X.laevis α-globin mRNA in which the 3' terminal sequence

is 9 nucleotides shorter than the sequence presented here. It is possible that the difference between the sequences has resulted from a deletion of this region during their cloning procedure which involved the insertion of a cDNA:mRNA hybrid into the plasmid vector. This method has been reported to be associated with 3' end deletions (20). Alternatively, it cannot be excluded that there is some microheterogeneity at the poly A addition site.

*Present address: Genetics Institute, 225 Langwood Avenue, Boston, MA, USA

+Present address: King's College, Department of Biophysics, 26-29 Drury Lane, London, WC2B 5RL, UK.

REFERENCES

1. Battaglia, P. and Melli, M. (1977) Dev. Biol. 60, 337-350
2. Hentschel, C., Kay, R.M. and Williams, J.G. (1979) Dev. Biol. 72, 350-363
3. Kay, R.M., Harris, R., Patient, R.K. and Williams, J.G. (1980) Nuc. Acid Res. 8, 2691-2707
4. Williams, J.G., Kay, R.M. and Patient, R.K. (1980) Nuc. Acids Res. 8, 4247-4258
5. Wilson, J.T., Wilson, L.B., Reddy, V.B., Cavallesco, C., Ghosh, P.K., de Riel, J.K., Forget, B.G. and Weissman, S.M. (1980) J. Biol. Chem. 255, 2807-2815
6. Heindell, H.C., Liu, A., Paddock, G.V., Studnicka, G.M. and Salser, W.A. (1978) Cell 15, 43-45
7. Nishioka, Y. and Leder, P. (1979) Cell 18, 875-882
8. Deacon, N.J., Shine, J. and Naora, H. (1980) Nuc. Acids Res. 8, 1187-1199
9. Dayhoff, M.O. (1975) Atlas of protein sequence and structure. National Biomedical Research Foundation, Silver Spring, Maryland.
10. Maruyama, T., Watt, K.W.K. and Riggs, A. (1980) J. Biol. Chem. 255, 3285-3293.
11. Richardson, C., Cappello, J., Cochran, M.D., Armentrout, R.W. and Brown, R.D. (1980) Dev. Biol. 78, 161-172
12. Partington, G.A. and Barelle, F.E. (1981) J. Mol. Biol. 145, 463-470
13. Berkener, K.L. and Folk, W.R. (1977) J. Biol. Chem. 252, 3176-3184
14. Weaver, R.F. and Weissman, C. (1979) Nuc. Acids Res. 7, 1175-1193
15. Soeda, E., Arrand, J.R., Smolar, N. and Griffin, B.E. (1979) Cell 17, 357-370
16. Maxam, A. and Gilbert, W. (1980) Meth. Enz. 65(1), 499-560
17. Goodman, M., Moore, G.W. and Matsuda, G. (1975) Nature 253, 603-608
18. Patient, R.K., Elkington, J.A., Kay, R.M. and Williams, J.G. (1980) Cell 21, 565-573
19. Maniatis, T., Kee, S.G., Efstradiadis, A. and Kafatos, F.C. (1976) Cell 8, 163-182
20. Zain, S., Sambrook, J., Roberts, R.J., Keller, W., Freid, M. and Dunn, A.R. (1979) Cell 16, 851-863