Two-domain haemoglobin of the blood clam *Barbatia lima* resulted from the recent gene duplication of the single-domain δ chain

Tomohiko SUZUKI,* Yoshitada KAWASAKI, Tomikazu ARITA and Akio NAKAMURA

Department of Biology, Faculty of Science, Kochi University, Kochi 780, Japan

The blood clam *Barbatia lima* subsp. from Amami-Oshima, Japan, expresses three types of haemoglobins in erythrocytes: a tetramer (α₂β₂), a homodimer (δ.mock) and a polymer consisting of two kinds of chains, a 34 kDa two-domain (2D) globin and a δ chain. This is in sharp contrast to the congeneric clams *B. reeveana* (a North American species) and *B. lima* from Kochi, Japan, each containing a tetramer and a polymer consisting of the 2D globin, but not the δ chain. We have determined the cDNA-derived amino acid sequences of all four chains, α (163 residues), β (155 residues), δ (152 residues) and 2D (308 residues) of *B. lima* (Amami-Oshima). The α chain has an extremely long N-terminal extension of 20 residues that may form a ‘pre-A helix’, and this makes the α chain the longest known globin. *B. lima* α and β chains show about 50% sequence identity with the α and β chains, respectively, of tetrameric globin from a related clam, *Anadara trapezium*. The *B. lima* homodimeric δ chain shows 71–74% identity with each of the two domains of the 2D chain, but only 39% identity with the homodimeric γ chain of *Anadara*. In addition, the alignment of amino acid sequences of the δ chain and the two domains of the 2D chain revealed that the δ chain lacks one amino acid (Lys) at the C-terminus, suggesting that the C-terminal Lys (codon AAA or AAG) of the two domains of 2D chain could result from the stop codon TAA in the δ chain by nucleotide substitutions. These results, together with the fact that the δ and 2D chains form a polymeric haemoglobin, indicates that the δ chain is the ancestral single-domain globin for the 2D globin. The δ chain is expressed only in *B. lima* (Amami-Oshima), and appears to be a relic of molecular evolution.

INTRODUCTION

Several arcid clams, such as *Anadara* (*Scapharca*), *Barbatia* and *Calyptiogenia*, contain abundant haemoglobins in circulating erythrocytes. Among the invertebrate haemoglobins, the molluscan forms show a remarkable diversity in subunit structure; so far, homodimeric, heterodimeric, tetrameric and 2D polymeric haemoglobins have been found [1,2]. Anadara, a common blood clam around the seacoast, has two types of haemoglobin: a homodimer (γ₂δ₂) and a tetramer (α₂β₂, δ₂) [3], while the related clam *Barbatia reeveana* has a tetramer and a 400 kDa polymer consisting of an unusual two-domain (2D) chain [4]. Interestingly, the 2D globin gene of *B. reeveana* has two novel introns, a ’pre-coding’ intron and a ’bridge’ intron that separates the two domains [5], in addition to the two introns conserved in all of the vertebrate globin genes.

The blood clam tetrameric haemoglobin possesses a novel ‘back-to-front’ quaternary structure relative to vertebrate haemoglobin in that the G and H helices are on the outside of the molecule and the E and F helices form important intersubunit contacts [6]. This means that the tetrameric assembly has been acquired independently in vertebrates and molluscs. Thus clam haemoglobins appear to provide an interesting model to elucidate the complex molecular evolution of invertebrate haemoglobins.

The cDNA-derived or chemical amino acid sequences were determined for *Anadara* (*Scapharca*) tetrameric α and β and homodimeric γ chains [7,8], *B. virescens* heterodimeric chains I and II [9,10] and *B. reeveana* 2D chain [5]. Here we report the cDNA-derived amino acid sequences of all four chains, α, β, δ and 2D of the dimer, tetramer and polymer haemoglobins of *B.

*lima* subsp. from Amami-Oshima, Japan. These sequence data clearly show that the δ chain is the ancestral single-domain globin for the unusual 2D globin.

MATERIALS AND METHODS

Isolation of haemoglobins and their constituent chains

*Barbatia lima* subsp. were collected from Amami-Oshima, Japan. The red cells were washed three times with 3% NaCl. The haemolysate was centrifuged at 8500 g for 5 min, and the supernatant was applied to a gel filtration column of Ultrogel AcA44. The column was equilibrated and eluted with 50 mM phosphate buffer (pH 7.2) at a flow rate of 0.5 ml/min. To separate constituent polypeptide chains, three haemoglobin fractions were applied directly to a reverse-phase column (Cosmosil 5C₁₈-300, Nacalai). The column was equilibrated with 30% acetonitrile in 0.1% trifluoroacetic acid (TFA) and eluted with a linear gradient of 30–90% acetonitrile in 0.1% TFA over 30 min at a flow rate of 1 ml/min. The molecular mass of the isolated chains was estimated by SDS/PAGE in the presence or absence of a reducing agent.

Protein sequence analysis

The isolated polypeptide chains α, β and δ were pyridylethylated and digested with trypsin at an enzyme/substrate ratio of 1:100 in 0.1 M NH₄HCO₃ at 37°C for 2 h or with pepsin at an enzyme/substrate ratio of 1:100 in 5% formic acid at 37°C for...
**Figure 1** Nucleotide and derived amino acid sequences of two of the cDNAs of *B. lima* (Amami-Oshima) haemoglobins

Arrows a–d indicate the primers used in the PCR amplification. Polyadenylation signals (AATAAA) are boxed. The amino acids determined by protein sequencing are underlined. Left panel, 2D chain. The second domain begins with Met at position 154 (indicated by the arrow). Right panel, homodimeric δ chain.

1 h. The digested products were isolated by reverse-phase chromatography. The column (Cosmosil SC18-300) was equilibrated with 0.1% TFA and eluted with a linear gradient of 0–90% acetonitrile in 0.1% TFA over 120 min at a flow rate of 1 ml/min. Some peptides were purified further by rechromatography on the same column with a linear gradient of 0–90% acetonitrile in 10 mM ammonium acetate over 60 min at a flow rate of 1 ml/min. The amino acid sequence of peptides was determined with an automated protein sequencer (Model 476A, Applied BioSystems).

**Figure 2** Nucleotide and derived amino acid sequences of two of the cDNAs of *B. lima* (Amami-Oshima) haemoglobins

Arrows e–h indicate the primers used in the PCR amplification. Polyadenylation signals (AATAAA) are boxed. The amino acids determined by protein sequencing are underlined. Left panel, tetrameric α chain. Right panel, tetrameric β chain.
Figure 3 Gel filtration of *B. lima* (Amami-Oshima) haemoglobins on a Ultrogel AcA44 column

The column (2.8 cm x 98 cm) was equilibrated and eluted with 50 mM phosphate buffer (pH 7.2). Fraction size, 5 ml/tube.

cDNA sequence analysis

mRNA was prepared from the erythrocytes of *B. lima* (Amami-Oshima) with a FastTrack mRNA Isolation Kit (Invitrogen). The single-stranded cDNA was synthesized with avian reverse transcriptase using the oligo-dT adaptor 5'-GGATCGGAAT-TCCCGGGTT-3', as a primer.

The 3' half of the cDNA of the 2D chain was amplified for 30 cycles, each consisting of 0.5 min at 94 °C for denaturation, 1 min at 50 °C for annealing and 1 min at 72 °C for primer extension, by PCR [11]. The enzyme Extra Taq DNA polymerase (Takara) with high fidelity was used. The primers used were the oligo-dT adaptor and redundant oligomer (24-mer complementary to the sequence shown in Figure 1, left panel) and c (24-mer complementary to the sequence shown in Figure 1, left panel) were designed from the sequence of the 3' region. The 5' region of the cDNA was first amplified for 30 cycles, each consisting of 1 min at 94 °C for denaturation, 1 min at 50 °C for annealing and 2 min at 72 °C for primer extension, by PCR, using the oligo-dC adaptor (5'-GAATTCGGATCC-3') and primer b. Then the amplified products were re-amplified using the oligo-dC adaptor and primer c, to improve the specificity. The 280 bp products amplified were subcloned in the SmaI site of pUC18 and sequenced as described above.

Furthermore, the whole length of the cDNA of the *B. lima* 2D globin was amplified by PCR using the oligo-dT adaptor and non-redundant primer d (20-mer based on the first-20-nucleotide sequence of the 5' non-coding region; Figure 1, left panel). The amplified products were sequenced directly with a PRISM dye terminator cycle sequencing kit.

The whole length of the cDNA of δ chain was amplified with the same non-redundant primer d as used in the 2D chain (Figure 2, left and right panels) and the oligo-dT adaptor. The amplified products were subcloned in the plasmid pUC18 and sequenced. The 3' halves of cDNAs of α and β chains were amplified with the oligo-dT adaptor and redundant primers e and g, respectively, based on the amino acid sequences (see Figure 2, left and right panels respectively). The amplified products were subcloned in the plasmids pUC18 or pCRII (Invitrogen) and sequenced.

The 5' halves of cDNAs of α and β chains were amplified with the oligo-dC adaptor and non-redundant primers f and h respectively (Figure 2, left and right panels respectively). The amplified products were subcloned in the plasmid pCRII and sequenced.

RESULTS AND DISCUSSION

Figure 3 shows the typical elution profile of haemoglobins from *B. lima* subsp. (Amami-Oshima) on an Ultrogel AcA44 column.

Table 1 Subunit and chain compositions of *Barbatia* haemoglobins

<table>
<thead>
<tr>
<th>Source</th>
<th>Subunit</th>
<th>Chain composition (kDa)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>B. lima</em> (Amami-Oshima)</td>
<td>Dimer</td>
<td>δ (17.3)</td>
<td>This work</td>
</tr>
<tr>
<td></td>
<td></td>
<td>α (18.3)</td>
<td>This work</td>
</tr>
<tr>
<td></td>
<td></td>
<td>β (17.6)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tetramer</td>
<td>α (16)</td>
<td>Suzuki et al. [12]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>β (16)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Polymer</td>
<td>δ (17.3)</td>
<td>This work</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2D (35.2)</td>
<td></td>
</tr>
<tr>
<td><em>B. lima</em> (Kochi)</td>
<td>Tetramer</td>
<td>α (16)</td>
<td>Suzuki and Arita [19]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>β (16)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Polymer</td>
<td>2D (35.2)</td>
<td></td>
</tr>
<tr>
<td><em>B. reeveana</em></td>
<td>Tetramer</td>
<td>α (17)</td>
<td>Grinich and Terwilliger [4]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>β (16)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Polymer</td>
<td>2D (35.2)</td>
<td>Naito et al. [5]</td>
</tr>
<tr>
<td><em>B. virescens</em></td>
<td>Dimer (heteromeric)</td>
<td>I (17.8)</td>
<td>Suzuki et al. [9]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>II (17.8)</td>
<td>Suzuki et al. [10]</td>
</tr>
</tbody>
</table>
Three haemoglobin fractions corresponding to the polymer (~290 kDa), tetramer (70 kDa) and dimer (30 kDa) were separated by reverse-phase chromatography [12]. Table 1 summarizes the subunit and chain compositions of B. lima (Amami-Oshima) haemoglobins, with those of B. lima (Kochi), B. reeveana and B. virescens haemoglobin.

We have succeeded in amplifying cDNAs of all four chains, 2D, 2, α and β, of B. lima (Amami-Oshima) by PCR. The complete nucleotide and derived amino acid sequences of the cDNAs are shown in Figures 1 and 2.
The whole length of the cDNA (693 bp) of the δ chain was successfully amplified by PCR using the non-redundant primer d designed from the 5′ non-coding region of the 2D chain. The open reading frame is 459 nucleotides in length and encodes a protein with 155 amino acids, of which 44 residues were identified directly. The molecular mass for the protein was calculated to be 17622 Da.

The cDNA-derived amino acid sequences of four chains of B. lima (Amami-Oshima) are aligned with those of chains I and II of heterodimeric haemoglobin from B. virescens [9,10] and α, β and γ chains of tetrameric and homodimeric haemoglobins from the related blood clam Anadara trapezia [7] in Figure 4. The percentages of identity between the chains are summarized in Table 2. B. lima α and β chains show 51–52 % sequence identity with α and β chains of Anadara. The B. lima homodimeric δ chain shows strong 71–74 % identity with each of the two domains of B. lima 2D chain, but shows low similarity (39 % identity) with the homodimeric γ chain of Anadara. This observation is consistent with the fact that the δ and 2D chains form a polymeric haemoglobin. The B. lima (Amami-Oshima) 2D chain shows strong sequence identity (78–79 %) with those of B. reeveana and B. lima (Kochi) (alignment not shown). The sequence identity between B. reeveana and B. lima (Kochi) 2D chains is 89 %. On the other hand, the sequence identity between the two domains of B. lima (Amami-Oshima) was about 75 %.

The α chain of B. lima (Amami-Oshima) has an extremely long N-terminal extension of 20 residues, and this makes the α chain the longest globin known so far (Figure 4). The N-terminal extension of about 10 residues of Scapharca (Anadara) haemoglobin is known to form a ‘pre-A-helix’ [6]. Secondary structure analysis of this extension of B. lima α chain using a computer program in GeneWorks 2.4 (Teijin) does predict formation of an α-helix.

The alignment of amino acid sequences of the δ chain and two domains of the 2D chain of B. lima (Amami-Oshima) revealed that the δ chain lacks one amino acid (Lys) at the C-terminus (Figure 4). It is generally known that a basic residue (Lys, Arg or His) is structurally rather common at the C-terminus of vertebrate and invertebrate globins. In addition, the second domain of the 2D chain begins with the initiation Met. These observations suggest that the C-terminal Lys (codon AAA or AAG) of the two

The tree was constructed by the method of Feng and Doolittle [18]. The tree is an unrooted tree and does not suppose constancy for the evolutionary rate. The bar represents the relative evolutionary distance for the branch length [18].
domains of the 2D chain could result from the stop codon TAA in the δ chain by nucleotide substitutions (Figure 1). Thus it is possible that the δ chain is the ancestral single-domain globin for the 2D globin. This is supported by the phylogenetic analysis shown in Figure 5. The 2D globin would have resulted from gene duplication and loss of a stop codon of the ancestral δ chain. This δ chain is expressed only in B. lima (Amami-Oshima), and appears to be a relic of molecular evolution.

The 2D haemoglobin chain also occurs in Daphnia (water flea) [13,14] and the nematodes Ascaris and Pseudoterranova [15,16], of which the latter two have been sequenced completely. Each of the two domains of the nematode haemoglobins shows relatively low similarity (46–52%) to each other. This similarity is close to that (44%) between human α and β chains, which were separated by gene duplication about 450 million years ago [17]. On the other hand, the two domains of the Barbatia 2D chain are highly similar (80% identity). Goodman et al. [17a] estimated that the divergence time of the molluscan globins of Aplysia limacina and A. kurodai, with 84% sequence identity, is about 50 million years ago. Therefore the gene duplication time of the Barbatia 2D globin would also be estimated at around 50 million years ago.

A phylogenetic tree was constructed from the 10 amino acid sequences aligned in Figure 4 using a computer program by Feng and Doolittle [18]. The tree separated clam globins into three distinct clusters, a cluster for tetrameric (α and β) and homodimeric (γ) chains of Anadara and B. lima, a cluster for heterodimeric chains I and II of B. virens, and a cluster for the 2D and δ chains of B. lima (Figure 5). This tree clearly indicates that the 2D chain evolved from the δ chain recently, and also suggests that the Anadara γ chain might correspond to the ancestral globin for the tetrameric chains α and β of Anadara and B. lima. The heterodimeric chains I and II are unique gene products expressed only in B. virens [12]. This remarkable diversity in haemoglobin constituent chains and subunit structure of Barbatia and Anadara (Scapharca) haemoglobins could be a consequence of a physiologically less important role of blood clam haemoglobin relative to vertebrate haemoglobins [1].

We thank Dr Serge N. Vinogradov of Wayne State University for giving us invaluable suggestions.

REFERENCES