THE JOURNAL OF BIOLOGICAL CHEMISTRY
© 1990 by The American Society for Biochemistry and Molecular Biology, Inc.

Vol. 265, No. 24, Issue of August 25, pp. 14285-14291, 1990
Printed in U.S.A.

Structural Interpretation of the Amino Acid Sequence of a Second Domain from the *Artemia* Covalent Polymer Globin*

(Received for publication, December 13, 1989)

Luc Moens‡§, Marie-Louise Van Hauwaert‡, Koen De Smet‡, Kris Ver Donck‡, Yves Van De Peer‡, Jozef Van Beeumen¶, Shoshana Wodak‖, Philippe Alard‖, and Clive Trotman**

*From the ‡Department of Biochemistry, University of Antwerp (UIA), Wilrijk B 2610, Belgium, the ¶Laboratory of Microbiology, State University of Ghent, Ghent B 9000, Belgium, the ‖Biological Macromolecular Conformation Unit, Free University of Brussels (ULB), and Plant Genetic Systems, Brussels B 1050, Belgium, and the **Department of Biochemistry, University of Otago, Dunedin, New Zealand*

***Artemia* has a complex extracellular hemoglobin of M_r 260,000 comprising two globin chains (M_r 130,000) each of which is a polymer of eight covalently linked domains of M_r 16,000. The primary structure of this polymeric globin was studied to understand how globin folded domains are ordered within a globin chain and, in turn, how the latter associate into a functional hemoglobin molecule.**

Here we report the amino acid sequence of a second domain, E7 (M_r 16,081, excluding the heme), and interpretations of sequence data by computer-assisted alignment and modeling. This clearly shows that, as with domain E1 (Moens, L., Van Hauwaert, M.-L., De Smet, K., Geelen, D., Verpooten, G., Van Beeumen, J., Wodak, S., Alard, P., & Trotman, C. (1988) *J. Biol. Chem.* 263, 4679–4685), domain E7 is compatible with a globin folded structure of the β-type chain. Several specific differences of domains E7 and E1 from the classic globins are identified. They possibly can be interpreted in terms of specific requirements for a double octameric functional molecule.

The characterization of a bacterial (1) and protozoan (2) globin and the observation that hemoglobins occur more frequently in plants than originally expected (3) have recently attracted attention to the study of the structure and evolution of globins in primitive organisms. Indeed, hemoglobins and myoglobins not only occur in vertebrates, but they are also widely but disparately distributed in all other phyla (4). The episodic distribution in invertebrates is more likely a matter of gene expression rather than gene possession (5, 6). In contrast with higher eukaryotes, where they are always intracellular and tetrameric, the extracellular hemoglobins of lower organisms show a wide variety in molecular architecture both in the native molecules and in their constituent globin chains (4, 7). Extracellular hemoglobins always show high M_r values (10^5 to 1.2×10^7) necessary to minimize loss by excretion. This high M_r is obtained by aggregation (Annelida) of many myoglobin-like low M_r chains into a functional molecule or by the covalent linking of such chains, as structural units or domains, into polymeric globins (M_r 32,000–300,000) (*Mollusca* and *Arthropoda*) (4, 5, 7–9). Based on protein and gene structure, it is generally accepted that all globins evolved from an ancestral chain (5, 6, 10, 11). This ancestral globin itself probably arises from an even older heme-binding protein of the cytochrome b_5 type (12). Comparison of the primary structures of all known globins confirms this hypothesis (13, 14) and shows a vast set of motifs (15) determining the globin fold (16). It is not known if a polymeric globin structure can be reconciled with the classic globin model.

The arthropod *Artemia*, an anostracan branchiopod crustacean, has hemoglobins of M_r 260,000. Each molecule is a dimer of two similar sized subunits (M_r 130,000). Each subunit represents a globin chain which exists in two forms (α and β), thus making possible three different phenotypes (Hb1, α_2, Hb2, $\alpha\beta$, and Hb3, β_2). The physical and physiological properties of these hemoglobins are well documented (9, 17–20). The M_r 130,000 globin chain itself comprises eight heme-binding domains of M_r ~16,000 which are sequentially linked through peptide bonds (8, 14, 21, 22).

To elucidate the structural and evolutionary relationships of these polymeric globins to the classic globin family, we determined the primary structure of another domain and analyzed sequence data by computer-assisted alignment and modeling.

MATERIALS AND METHODS[1]

RESULTS

The amino acid sequence of domain E7 from the *Artemia* globin was determined by automated Edman degradation of the amino-terminal segment and by manual sequencing of peptides obtained by cleavage with trypsin and chymotrypsin. The data relevant in reconstructing the sequence are summarized in Fig. 4. Due to the fact that only two sets of peptides were generated and that not all of them could be purified to homogeneity, a few overlaps are missing. However, alignment with *Artemia* domain E1 (8) and several key globins allows unambiguous reconstruction of the total sequence (Figs. 6 and 12). The proposed sequence (152 residues; M_r 16,081, excluding the heme) is shown in Fig. 5 along with a globin tertiary structure diagram.

DISCUSSION

Cleavage of Domains from Intact Globin Chains

Analysis of the fragment mixture, obtained after limited digestion of hemoglobin with subtilisin, strongly suggests that

* This work was supported by Fund for Joint Basic Research (Belgium) Programs 2.0012.82 and 2.0042.85. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ To whom reprint requests should be addressed.

[1] Portions of this paper (including "Materials and Methods," Figs. 1–4 and 11–14, and Table 1) are presented in miniprint at the end of this paper. Miniprint is easily read with the aid of a standard magnifying glass. Full size photocopies are included in the microfilm edition of the Journal that is available from Waverly Press.

Tables 2–6 are available as supplementary material from the authors.

the hydrolytic enzyme has preferential access to the more exposed regions in between the structural units or domains (21, 22). This is confirmed by comparison of the amino termini of several fragments, clearly showing the presence of structures similar to the A, B, and C motifs of classic globins (36, 37). The purified fragments of M_r 16,000 (E1–E8) can thus be considered as equivalent to domains. However, when sequencing the amino termini of domains E1 and E7, in both cases, a minor protein fraction having the same sequence but starting a few residues earlier was observed (Fig. 6 and Table 2). This illustrates that the hydrolytic enzyme, due to its nonspecificity, cleaves within the interdomain regions rather arbitrarily. The domain population (fraction E) (21, 22) results from proteolytic cleavage of a total hemoglobin preparation containing both globin chains. Therefore, it is impossible to attribute a given domain to the α or β chain.

Interpretation of Structure of Domain E7

A search in the National Biomedical Research Foundation Data Bank with FASTP (29) shows that the domain E7 sequence scores the highest similarity to the members of the globin family. It also displays the invariant globin landmarks (15, 38) as CD1, Phe and F8, His, and some of the more highly conserved residues, including Trp(A12), Pro(C2), Phe(CD4), and distal His(E7), confirming its globin nature (Figs. 5 and 6). Within the globin family, the highest similarity is observed with the β-globins and some invertebrate globins (*Busycon canaliculatum* and *Chironomus* chain 10) in contrast with *Artemia* domain E1, which shows the highest similarity to the myoglobins (8, 14, 15). Therefore, the human β chain was adapted as a structural template, and the standard numbering system was used, based on sperm whale myoglobin. The sequence of domain E7 will be discussed now in relation to the recognized regions of the globin structure (15, 38) (Figs. 5 and 6).

Region NA (Residues 1–2)—The conserved Leu at position NA2 of the majority of globins (15) can also be recognized in domain E7. No pre-A helix region, similar to the consensus linker sequence (-Val-Asp-Pro-Ile-Thr-Gly-) of domain E1, is present (8, 36, 37, 39). This confirms the arbitrary nature of the cleavage between domains.

Helix A (Residues 3–18)—Between residues 3 and 18, the A motif is well recognizable (Figs. 5 and 6). Indeed, Glu(A4), Ile(A8), Ile(A11), and Trp(A12) are able to form ridges and grooves with Glu(H7), Tyr(H8), Lys(H10), and Gly(H11), realizing A/H helix packing. Also, the identical Trp(A12) allows a hydrogen bond with the E helix, whereas Ile(A8) and Leu(A15) are able to cluster at the bottom of the heme pocket (38). This strongly suggests not only the presence of an A helix, but also that the spatial relationship of the A helix to the rest of the molecule is similar to that of the globin fold generally.

Substitution at the surface of Lys to Ile at position A14 seems to be in contrast with the functional conservation of the A motif (15). Despite the fact that a hydrophobic residue at the surface is less plausible, Val and Leu are noted in this position in several vertebrate and invertebrate species (15, 40) (Fig. 12). In addition, as the *Artemia* hemoglobin quaternary structure is probably quite different from that of the classical hemoglobins, it is not known which parts of the domain E7 are really buried or at the surface and whether an external salt bridge at position A14 is necessary.

Region AB (Residues 19–20)—Alignment in the A, B, and C helices places two (possibly 3) residues, -Ala-Val-(Gly), between the A and B helices. Although a single AB residue (Ala or Gly) is typical in globins, longer AB turns are recognized in *Tylorrhynchus, Lumbricus, Cerithidea, Busycon,* and *Glycera* (14, 15, 40). The insertion of 3 residues in the AB region of domain E7 does not cause a distortion of the AB corner as shown by BRUGEL modeling using a myoglobin fragment (GH5–H7) anchored at the A and B helix boundaries of the human β chain template.

Helix B (Residues 21–36)—Compared to the templates of Bashford *et al.* (15), the only discrepancy is the unique observation of a Met at position B9 (template 1). However, the nature of its side chain is fully compatible with an interior position (template 2). The B/E helix packing in globins is normally formed by ridges into grooves of the ± 4 n type whereby the volume of the sterically adjacent residues B6/B10 and E8/E4 is important (38). At positions B10 and E4, a rather big side chain is expected, which is also the case in domain E7 (Phe and Leu, respectively). At positions B6 and E8, the rather bulky side chains of Gln and Leu are observed, whereas small side chains were expected. However, domain E7 is not the only exception. A Gln or Glu is frequently observed at position B6 in the *Chironomus* globins, whereas a Ser or Thr occurs in the Leg hemoglobins.

Several invertebrates also show a somewhat bigger side chain at position E8 as Met (*Tylorrhynchus* chains 1 and 2A) and Leu/Ile (*Aplysia limacina, Aplysia kurodai*). This suggests

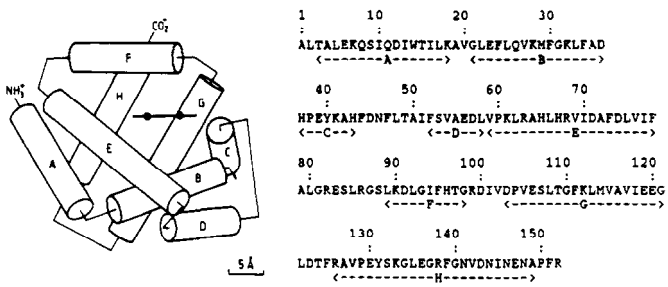

FIG. 5. **Amino acid sequence of *Artemia* domain E7 shown with diagram of three-dimensional structure of *Chironomus* erythrocruorin.**

FIG. 6. **Alignment of amino acid sequence of *Artemia* domains E1 and E7 with sequences of human β chain and *Chironomus* globin.** Residues between *brackets* were derived from an additional peptide present during amino-terminal sequencing.

```
                    NA1 5      A1  5   10   15   AB1  B1  5   10   15   C1  5    CD1 5    D1  5
                      :  :       :   :    :    :    :   :    :    :   :   :    :   :    :   :
Hb BETA     :    -- ------V HL  TPEEKSAVTALWGKV-  ---  NVDEVGGEALGRLLVV  YPWTQRF  FESFGDLS  TPDAVMG
ARTEMIA E1: [SL]ERVDPIT GL  SGLEKNAILDTWGKVR  G--  NLQEVGKATFGKLFAA  HPEYQQM  FRFFQGVQ  LAF-LVQ
ARTEMIA E7:   [GLYAP-K]AL   TALEKQSIQDIWTILK  AV-  GLEFLQVKMFGKLFAD  HPEYKAH  FDNFLTAI  FSVAEDL
CHIRON III: -- ------- -L   SADQISTVQASFDKVK  G--  ----DPVGILYAVFKA  DPSIMAK  FTQFAG-K  DLESIKG

                    E1  5   10   15   20   EF1 5   10    F1  5   10   FG1 5   G1  5   10   15     GH1 5
                      :   :    :    :    :   :    :     :   :    :    :   :   :    :    :      :   :
Hb BETA     :    NPKVKAHGKKVLGAFSDGLA  HL-D---NLKGT  FATLSELHCD  KL--HV  DPENFRLLGNVLVCVLAHH  FGKEF
ARTEMIA E1:    SPKFAAHTQRVVSALDQTLL  ALNRPSDQFVYM  IKELGLDHIN  RGT---  DRSFVEYLKESLGDSVDEF  TVQSF
ARTEMIA E7:    VPKLRAHLHRVIDAFDLVIF  AL--GRESLRGS  LKDLGIFHTG  RD--IV  DPVESLTGFKLMVAVIEEG  LDT-F
CHIRON III:    TAPFETHANRIVGFFSKIIG  ELP----NIEAD  VNTFVASHKP  RG---V  THDQLNNFRAGFVSYMKAH  T--DF

                    H1  5   10   15   20   25   HC1
                      :   :    :    :    :    :   :
Hb BETA     :    TPPVQAAYQKVVAGVANALAHKYH--  ---
ARTEMIA E1:    GEVIVNPLNEGLRQA----------  ---
ARTEMIA E7:    --RAVPEYSKGLEGRFGNVDNINENA  PFR
CHIRON III:    A-GAEAAWGATLDTFFGMIFSKM---  ---
```

that in these cases the B/E helix packing is less compact. If, in domain E7, the angle between the B and E helices is to conform to the globin fold, a class 2 rather than a class 3 crossing is likely (41). BRUGEL modeling shows that a sterically acceptable solution can be found by rotating the Gln side chain at position B6 in between the Leu(E8) and Ile(E12) positions in the mutated β chain (Fig. 7). All residues involved in the B/G helix packing (B5, B9, B13; G15, G11, G7; B8, B9, B16) are fully compatible, including the Met at position B9, as the volume of the side chains is less important for this contact (15, 38).

Helix C and Region CD (Residues 37–51)—Alignment in the critical heme environment of the CD-D region of domain E7 with other globins is unmistakable with the presence of the invariant residue Phe(CD1) and the quasi invariants Pro(C2) and Phe(CD4) (15, 38). Consequently, it reinforces the alignment of the B helix. As in myoglobins, domain E7 has a His at position C1 which is considered to represent the ancestral condition that existed before the cytochromes and the globins diverged (12). The very conservative Thr at position C4 is substituted with a much bulkier Tyr, resulting in the only penalty against the templates of Bashford *et al.* (15) for the CD-D region. This substitution not only occurs in domain E7, but also in domain E1 and in the homologous position of the amino terminus of the *Artemia* globin α and β chains (36, 37). No other globins have a Tyr at position C4; it therefore seems to be a specific feature of the *Artemia* globins with unknown functional significance. BRUGEL modeling showed that the bulky Tyr could be accommodated between the C and G helices with its -OH at the surface in a model based on β-globins (Fig. 8), but was subject to steric hindrance in a myoglobin-based model as used for domain E1 (8).

Helix D (Residues 52–58)—This helix-predictive sequence (secondary structure prediction) is compatible with the retention in domain E7 of a small D helix similar to the β-globins. In domain E1, a helix shorter by 1 residue was assigned to this region.

Helix E (Residues 59–78)—Placing His^{65} at the E7 location (distal His) fulfills the major requirements for the E helix motif (15). (see Ref. 15) Indeed, with the exception of the already discussed Leu at position E8, the side chain character of the residues involved in the B/E helix packing is mainly conserved. The highly conserved requirement for a Val at position E11 in contact with the heme is met (16). The most significant discrepancies with Bashford template 2 occur at the end of the helix (Leu(E17), Phe(E20)). The replacement

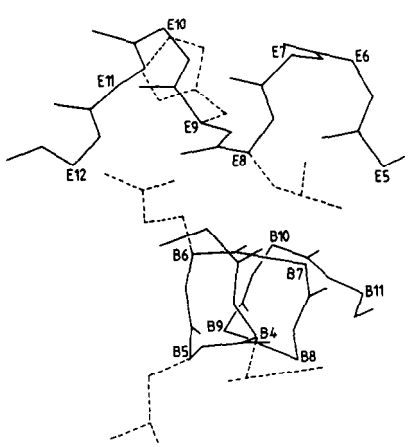

FIG. 7. **B-E contact region.** The B-E contact region of *Artemia* domain E7 was modeled with a BRUGEL software package as described under "Materials and Methods." The major side chains (- - -) are shown on the fragments of the B and E helices.

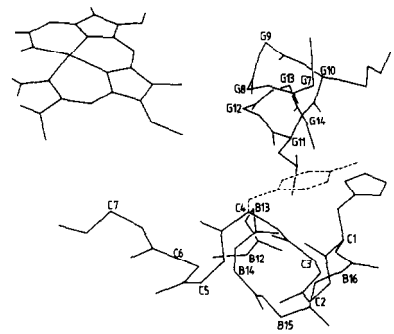

FIG. 8. **BC-G region.** The BC corner of *Artemia* domain E7 was modeled with a BRUGEL software package as described under "Materials and Methods." The side chains at positions C4 (- - -), C1, G10, and G11 (——) are shown.

of the 2 adjacent surface residues E17 and E20 by Leu and Phe, respectively, suggests the possibility of a hydrophobic surface region adapted to the special quaternary requirements of the *Artemia* globin. Domain E1 similarly has a unique Leu at position E20.

Region EF and Helix F (Residues 79–98)—Alignment of the F helix sequence is influenced by the obligatory proximal His at position F8. This brings about a Leu at position F4 and a Phe at position F7, allowing not only normal heme contact (F4, F7, F8), but also the expected F/H helix packing (± 4 *n*; F1, F4, F8) (38). In contrast, domains E1 and E7 are exceptional in having large hydrophobic residues (Leu and Ile, respectively) at surface position F6 where hydrophilic residues are expected. However, this is not unique as in several invertebrates (*Chironomus* globin chain 10, *Glycera, Anadara broughtioni*) together with some vertebrates, a Val is observed at this position (40).

Alignment of the E and F helices brings residues 79–88 into the EF corner. Due to the variability of this region in length and side chain character, the similarity to domain E1 as well as to the other globins is low, resulting in poor modeling from fragments available in the National Biomedical Research Foundation Data Bank.

Region FG and Helix G (Residues 99–121)—The FG-G region permits alternative alignments. A first alternative (Fig. 6) places the sequence -Val-Asp-Pro-Val- at positions FG4–F3 as found in human α-globin. This brings a Ser and a Gly at positions G5 and G8, respectively, where a Phe and a Leu, important in heme contact and in the formation of the F/G helix packing, are expected. However, BRUGEL modeling using the β chain as template, suggests that the resulting decrease in heme contact is compensated for by substitution at position G9 of a small side chain by the aromatic ring of Phe. The expected crossover point of the G and H helices (positions G9 and H16, respectively) is now formed by 2 large Phe residues, whereas a small and a large side chain are expected. However, residue size and conformation can be combined in different ways to give the same type of packing as shown for sperm whale myoglobin, *Chironomus* globin 3A, and a Leg hemoglobin (38).

A second, less attractive alternative places Pro^{104} at position G1. This brings about the hydrophobic Leu and Phe at positions G5 and G8, allowing heme contact, but also the hydrophilic Glu at position G16, whereas this side chain is conventionally buried and involved in the G/H helix packing (38).

Region GH and Helix H (Residues 122–152)—A search was made in the Brookhaven National Laboratory Data Bank for a structure of the same length as the domain E7 sequence in this region and with matching flanking coordinates over G17–GH1 and H5–H8 for substitution into the β-globin template.

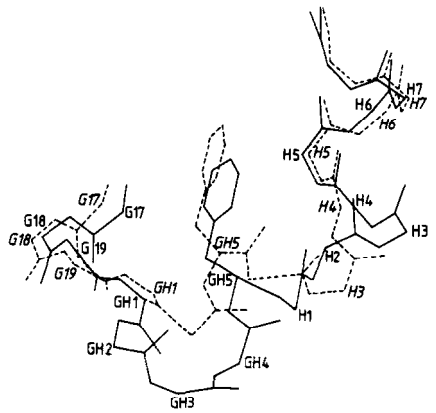

FIG. 9. **GH corner.** The GH corner of *Artemia* domain E7 was modeled with BRUGEL software package as described under "Materials and Methods." Human β chain; - - -, extracted fragment from *Chironomus* 3.——Side chains at position GH5 are shown.

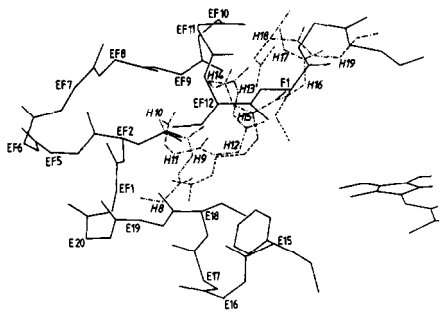

FIG. 10. **EF region.** The EF region of *Artemia* domain E7 was modeled with a BRUGEL software package as described under "Materials and Methods."——, EF corner with some hydrophobic side chains; - - -, H helix fragment; - - -, side chains at positions F1, H15, and H18.

The corresponding fragment from another globin, *Chironomus* globin 3A, was the best match. This substitution is illustrated in Fig. 9. The well conserved Phe at position GH5 is retained in the same position and orientation despite the 3-residue deletion. The G (from position G7) and the H helices, corresponding in higher globins to the third exon, are always the least conserved (14, 15, 42).

The alignment in Fig. 6 achieves the lowest total penalty against the Bashford templates (see Ref. 15). To explore the major discrepancies, Val(H5), Pro(H6), Arg(H15), and Asn(H18), the EF corner, and the H helix were modeled with a BRUGEL software package. The hydrophobic Val at surface position H5 may form a van der Waals contact with Phe at position GH5, whereas Pro at position H6 is, due to the deletions between positions GH3 and H3, in relative terms, the second residue, and thus can be considered as the H helix starter. Asn at position H18 is not only sterically in too close contact with Leu at position F1, but, like Arg at position H15, is in a hydrophobic environment. Both problems can be solved by rotating the total H helix a few degrees counterclockwise along the amino/carboxyl-terminal axis. This brings Asn(H18) to the surface and reduces sterical contact with Leu(F1), whereas the Arg side chain at position H15 can reach the surface in the space left by Gly(H11) (Fig. 10).

The H helix sequence is considerably more predictive of nonhelical and turn structure than that in other globins (secondary structure prediction); but in view of the expected adaptation of the carboxyl-terminal region to an interdomain linkage, this is not unexpected. The latter is confirmed by the recognition of a region similar to the linker sequence -Val-Asp-Pro-Val-Ile-Thr-Gly- in the H helix (H19–H25) (36, 37, 39).

Overall Structure of Artemia Globin Domains and Intact Hemoglobin Molecule

In domain E7, as in domain E1, the observed length discrepancies occur mainly in the interhelical turns (AB, EF, FG, and GH), where they would least disturb the general folding. At a few surface positions (A14, A17, E20, and F6), unexpectedly, hydrophobic residues are observed. In addition, comparison of the hydropathic profiles (43) of domains E7 and E1 with that of the human β chain (Fig. 11) shows that the CD-D region of both domains is more hydrophobic, whereas the end of the H helix is more hydrophilic. These differences, which do not disturb the overall folding, may reflect specific adaptations to the association of the covalently linked domains within a globin chain and of the association of both globins into the native dimer.

In conclusion, the fact that the domain E7 sequence can be interpreted in terms of the globin fold and that discrepant regions can be successfully modeled using the human β chain as a template strongly suggests that the E7 domain is compatible with a globin folded structure.

A maximum parsimony tree was constructed with some selected globins, including *Artemia* domains E1 and E7 (30–32). It is clear that this tree (Figs. 12–14) is inferior to those of Goodman *et al.* (14). Howeverr as expected, both domains E1 and E7 cluster with higher globins and close to the Insecta.

The reconstructed ancestral globin sequence (Fig. 14) is in full agreement with Bashford template 1 (see Ref. 15) and is 42% identical to domain E7, illustrating its homology (14, 44) to the globin family. Assuming that the domains are globin folded structures (4.5 × 3.5 × 2.5 nm), a speculative model for *Artemia* hemoglobin can be proposed. As no extensive linker sequences are present, neighboring domains must be oriented in such a way that the adapted HC and NA regions are close to each other. This may be achieved by a circular assembly of the eight covalently linked domains having their E and F helices coplanar as shown in Fig. 15. Heme is oriented toward the hollow center, and the EF turn and the interdomain linker sequence toward the periphery. The latter would agree with the preferential access of subtilisin to it (21, 22). Supposition of two such octomers with the order of domains in opposite directions and with the domains contacting each other at points of the E and F helices will provide a stable dimer. The resulting dimensions, 12–14 nm diameter and 6–7 nm thick, are in agreement with electromicroscopic observations (18). In this way, the unusual hydrophobic character of residues A14, E17, E20, and F6 may support the proposed model.

REFERENCES

1. Wakabayashi, S., Matsubara, M. & Webster, D. A. (1986) *Nature* **322,** 481–483
2. Iwaasa, H., Takagi, T. & Shikama, K. (1989) *J. Mol. Biol.* **208,** 355–358
3. Landsmann, J., Dennis, E. S., Higgins, T. J. V., Applebey, C. A., Kortt, A. A. & Peacock, W. J. (1986) *Nature* **324,** 166–168
4. Vinogradov, S. N. (1985), *Comp. Biochem. Physiol. B Comp. Biochem.* **82,** 1–15
5. Dickerson, R. E. & Geis, I. (1983) *Hemoglobin: Structure, Function, Evolution and Pathology* The Benjamin Cummings Publishing Co., Inc., Menlo Park, CA
6. Lewin, R. (1985), *Science* **226,** 328
7. Wood, E. J. (1980) *Essays Biochem.* **1,** 1–47
8. Moens, L., Van Hauwaert, M.-L., De Smet, K., Geelen, D., Verpooten, G., Van Beeumen, J., Wodak, S., Alard, P. & Trotman, C. (1988) *J. Biol. Chem.* **263,** 4679–4685
9. Moens, L., De Smet, K., Van Hauwaert, M.-L., Geelen, D.,

Verpooten, G. & Van Beeumen, J. (1990) in *Biochemistry and Cell Biology of Artemia* (MacRae, T., Bagshaw, J. & Warner, A., eds) CRC Press, Inc., Boca Raton, FL, in press
10. Collins, F. S. & Weismann, S. M. (1984) *Prog. Nucleic Acid Res. Mol. Biol.* **31,** 315–465
11. Jhiang, S. M., Garey, J. R. & Riggs, A. F. (1988), *Science* 240, 334–336
12. Runnegar, B. (1984) *J. Mol. Evol.* **21,** 33–41
13. Goodman, M. (1981) *Prog. Biophys. Mol. Biol.* **37,** 105–164.
14. Goodman, M., Pedwaydon, J., Czelusniak, J., Suzuki, T., Gotoh, T., Moens, L., Shishikura, ., Walz, D. & Vinogradov, S. N. (1988), *J. Mol. Evol.* **27,** 236–249
15. Bashford, D., Chothia, C. & Lesk, A. M. (1987) *J. Mol. Biol.* **196,** 199–216
16. Perutz, M. F. (1979) *Annu. Rev. Biochem.* **48,** 327–386
17. Moens, L. & Kondo, M. (1978), *Eur. J. Biochem.* **82,** 65–72
18. Wood, E. J. F., Barker, C., Moens, L., Jacob, W., Heip, J. & Kondo, M. (1981) *Biochem. J.* **193,** 353–359
19. D'Hondt, J., Moens, L., Heip, J., D'Hondt, A. & Kondo, M. (1978) *Biochem. J.* **171,** 705–710
20. Heip, J., Moens, L. & Kondo, M. (1978) *Dev. Biol.* **63,** 247–251
21. Moens, L., Geelen, D., Van Hauwaert, M.-L., Wolf, G., Blust, R., Witters, R. & Lontie, R. (1984) *Biochem. J.* **223,** 861–869
22. Moens, L., Van Hauwaert, M.-L. & Wolf, G. (1985) *Biochem J.* **227,** 917–924
23. Cohen, S. A., Bidlingmeyer, B. A. & Tarvin, T. L. (1986) *Nature* **320,** 769–770
24. Hewick, R. M., Hunkapillar, M. W., Hood, L. E. & Dreyer, W. J. (1981) *J. Biol. Chem.* **256,** 7990–7997
25. Hunkapillar, M. W., Hewick, R. M., Dreyer, W. J. & Hood, L. E. (1983) *Methods Enzymol.* **91,** 399–413
26. Chang, J. Y., Brauer, D. & Wittmann-Liebold, B. (1978), *FEBS Lett.* **93,** 205–214
27. Chang, J. Y. (1981), *Biochem. J.* **199,** 537–545
28. Chen, R. (1976), *Hoppe-Seylers's Z. Physiol. Chem.* **357,** 873–886
29. Stockwell, P. A. & Petersen, G. B. (1987), *CABIOS* **3,** 37–43
30. Felsenstein, J. (1985) *Evolution* **39,** 783–791
31. Eck, R. V., & Dayhoff, M. D. (1966) *Science* **152,** 363
32. Fitch, W. M. (1971) *Syst. Zoo.* **20,** 406–416
33. Delhaize, P., Van Belle, D., Bardiaux, M., Alard, P., Hamers, P., Van Cutsem, E. & Wodak, S. (1985) *J. Mol. Graphics* **3,** 116–119
34. Jones, A. T. & Thirup, S. (1986) *EMBO J.* **5,** 819–822
35. McLachlan, A. D. (1979) *J. Mol. Biol.* **128,** 49–79
36. Moens, L., Ver Donck, K., De Smet, K., Van Hauwaert, M.-L., Van Beeumen, J., Alard, P., Wodak, S. & Trotman, C. N. A. (1989) *NATO Adv. Study Inst. Ser.* **174,** 429–438
37. Moens, L., Wolf, G., Van Hauwaert, M.-L., De Baere, I., Van Beeumen, J., Wodak, S. & Trotman, C. N. A. (1990) in *The Biology of Artemia* (Brown, R., Sorgeloos, P. & Trotman, C. N. A., eds) CRC Press, Inc., Boca Raton, FL, in press
38. Lesk, A. M. & Chothia, C. (1980), *J. Mol. Biol.* **136,** 225–270
39. Manning, A. M., Marshall, C. J., Powell, R. J., Trotman, C. N. A. & Tate, W. (1989) *NATO Adv. Study Inst. Ser.* **174,** 413–425
40. Barker, W. C., Hunt, L. T., George, D. G., Yeh, L. S., Chen, H. R., Blomquist, M. C., Seibel-Ross, E. I., Elzanowski, A., Hong, M. K., Ferrich, D. A., Bair, J. K., Chen, S. L. & Ledley, R. S. (1986) *Protein Sequence Data Base of the Protein Information Resource,* National Biomedical Research Foundation, Washington, D. C.
41. Chothia, C., Levitt, M. & Richardson, D. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74,** 4130–4134
42. Eaton, W. A. (1980) *Nature* **284,** 183–185
43. Kyte, J. & Doolittle, R. F. (1982) *J. Mol. Biol.* **157,** 105–132
44. Lewin, R. (1985) *Science* **237,** 1570

Supplemental Material to :

Structural interpretation of the amino acid sequence of a second domain from the Artemia covalent Polymer globin.

Luc Moens (1), Marie-Louise Van Hauwaert (1), Koen De Smet (1), Kris Ver Donck (1), Yves Van de Peer (1), Jozef Van Beeumen (2), Shoshana Wodak (3), Philippe Alard (3) and Clive Trotman (4).

MATERIALS AND METHODS

1. Purification of Artemia hemoglobin domain E7

The Artemia hemoglobin was prepared according to previously published methods (9,17).
The E7 domain was purified to homogeneity from the subtilisin digested hemoglobin by gel filtration and chromatofocusing on a Mono P HR 5/20 column equilibrated in 25 mM Bis-Tris iminodiacetic acid pH 7,1 and developed with 9% Polybuffer 7-4, iminodiacetic acid pH 4.0 (Fig. 1). Final purity was judged by two dimentional electrophoresis.

2. Amino acid analysis

An appropriate amount of the apoprotein or tryptic peptides was hydrolysed in evacuated sealed tubes in 6N HCl for 24 hours at 110°C. The amino acid composition was determined using a Jeol 6 AH amino acid analyser. Protection of tryotophan was performed by the addition of 4% thioglycollic acid during hydrolysis. The amino acid composition of the chymotryptic peptides was determined using the Pico-Tag method (23).

3. Amino acid sequence determination

Automated amino acid sequence determination was performed using a vapor phase sequencer (Applied Biosystems model 470 A) (24,25).
PTH amino acids were identified and quantified by reverse phase HPLC on a 4,6 mm x 250 mm C18 column (IBM). Manual sequencing was performed as described previously (26,27).

4. Enzymatic cleavage

Cleavage was performed with trypsin and chymotrypsin in 0,2 M ethylmorpholine-acetate pH 8,3 at an enzyme concentration of 2% for 3 hours at 37°C. The reaction was terminated by lyophilisation.

5. Peptide separation and purification

The acid soluble peptides were separated by reverse phase HPLC on a 4,6 x 250 mm micro-Bondapack column using a gradient of 0,1% TFA in water to 0,1% TFA in acetonitrile at a flow rate of 1 ml over 120 min. Peptides were detected at 220 nm.
Additional purification was performed by preparative cellulose TCL using the solvent system of Chen (28).

6. Nomenclature of peptides

Tryptic and chymotryptic peptides were designated T and C respectively.
Peptides were numbered for each cleavage according to their position in the amino acid sequence of the chain starting from the amino terminus. Only relevant peptides were numbered.

7. Software package

The NBRF-databank was used in combination with software written by P.A. Stockwell (29).
Phylogeny tree construction is based on maximum parsimony analysis of 30 globin sequences. A program called PROTPARS (phylogeny interference package : version 2.9) (30) based on a new algorithm, intermediate between the approaches of Eck and Dayhoff (31) and Fitch (32) which infers an unrooted phylogeny from protein sequences, is used.

8. Model building of the E7 domain

A 3D model of the E7 domain was built using the crystal structures of the human β chain as a template and following sequence alignments shown in Fig. 5 and 12. The BRUGEL software package was used throughout this work (33).
Whenever necessary, β chain sidechains were replaced by those corresponding to the E7 sequence.
The conformation of the new sidechains were then adjusted to interact optimally with surrounding protein atoms in an automatic procedure which consists in a systematic search for minimum energy conformations as a function of the sidechain dihedral angles. In regions of length discrepancy, the deletions were modeled using fragments from the crystal structure of other proteins. These fragments were chosen as a result of a systematic search through a database containing atomic coordinates and other relevant information on known protein structures, using a procedure similar to that described by Jones (34,35). The criteria for the choice were based on structural similarity to the regions in the 3 chain immediately preceding and following the deletions and on fragment length. Information on amino acid sequence was not used. The fragments were then fitted into the β chain using coordinate superposition and their sequence was altered to match that of the E7 domain using the procedure for amino acid substitutions described above.

DETAILS OF SEQUENCE DETERMINATION

1. Amino acid composition of chain E7

In table 1 the experimentally determined amino acid composition for E7 is compared with the number of residues derived from the proposed sequence. For comparison the amino acid composition of E1 is included.

2. Sequence determination

2.1. Automated sequence determination

The amino terminal sequence of chain E7 was determined up to residue 42 starting from 5 nMol apoprotein as described in materials and methods.
Although at each cycle several amino acid residues were identified, the sequence could be deduced unambiguously from their relative variations in concentration taking into account the previous and next cycle. (Table 2)

2.2. Manual sequence determination

An appropriate amount of E7 was digested respectively with trypsin and chymotrypsin.
The resulting peptide mixtures were separated by R.P. chromatography. (Fig. 2, 3)
Peptides T1, T2, T13 and T14 were pure enough, as judged by TLC, to determine their amino acid composition and sequence (Table 3, 4) without additional purification.
The other tryptic peptides as well as the chymotryptic peptides were further purified by preparative TLC before determining their amino acid composition and sequence (Table 5, 6).
The manual sequenced peptides covering residues 16-41 completely confirm the result obtained by automated sequence determination. (Table 2, Fig. 4)

The data from automated and manual sequencing and the amino acid compositions of the relevant peptides (Fig. 4) summarized in Table 2 to 6 are deposited with the Journal of Biological Chemistry who will reproduce them on individual demand.

3. Reconstruction of the sequence of chain E7

The amino acid sequence of chain E7 was reconstructed from the above summarized data. (Fig. 4).

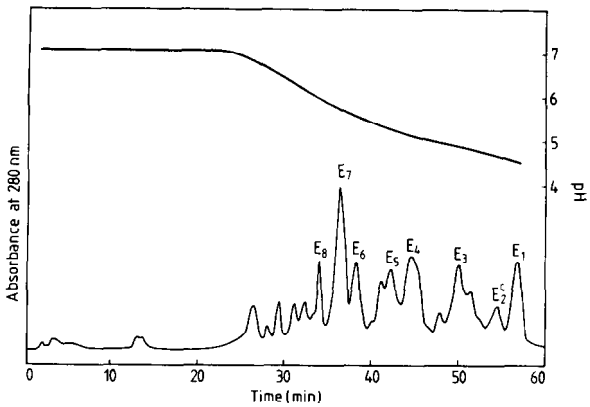

Fig. 1. Chromatofocusing of Artemia hemoglobin domains. A domain population selected by gelfiltration was loaded onto a Mono P HR 5/20 column equilibrated in 25 mM Bis-Tris imino-diacetic acid pH 7.1 and developed with 9% polybuffer 7-4, imino-diacetic acid pH 4.0 at a flow rate of 1 ml/min. Fractions were collected manually.

Table I : Comparison of the amino acid composition of chain E_7 with the number of residues derived from the proposed sequence. For comparison the amino acid composition of chain E_1 (9) is included.

Amino Acid	E_7 Determined Residues/mol	SD	E_7 Derived from the proposed sequence	E_1 Derived from the proposed sequence
Asp	16,38	0,44	16	14
Thr	8,86	0,28	9	7
Ser	7,84	0,31	6	8
Glu	13,29	0,27	14	21
Pro	3,96	0,22	5	4
Gly	10,51	0,48	11	12
Ala	14,43	0,30	13	10
1/2 Cys	-	-	-	-
Val	8,75	0,19	12	13
Met	2,18	0,18	2	2
Ile	7,68	1,90	10	5
Leu	16,78	0,55	19	17
Tyr	1,77	0,10	2	3
Phe	12,30	0,30	13	12
Trp	1,00	nd	1	1
Lys	7,34	0,68	9	7
His	4,76	0,61	5	3
Arg	9,16	0,97	8	8
	147,21		152	147

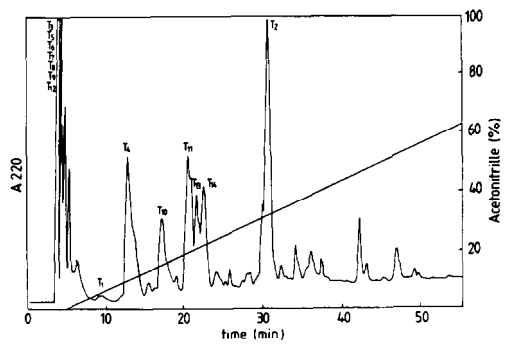

Fig. 2. Reverse phase chromatography of the tryptic peptides from Artemia domain E7. Buffers were 0,1% TFA in water (A) and 0,1% TFA in acetonitrile (B). A linear gradient was developed from 0% A to 100% B in 120 min at a flow rate of 1 ml/min. Peptides were numbered according to their order in the final sequence. Detection was performed at 220 nm.

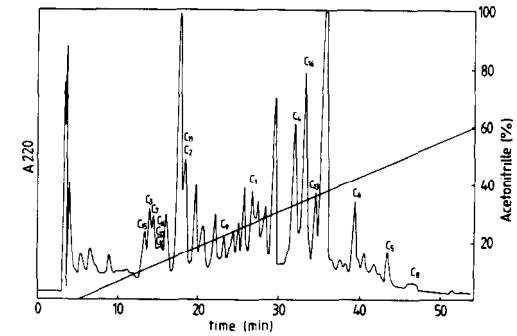

Fig. 3. Reverse phase chromatography of the chymotryptic peptides from Artemia domain E7. Conditions as in Fig. 2.

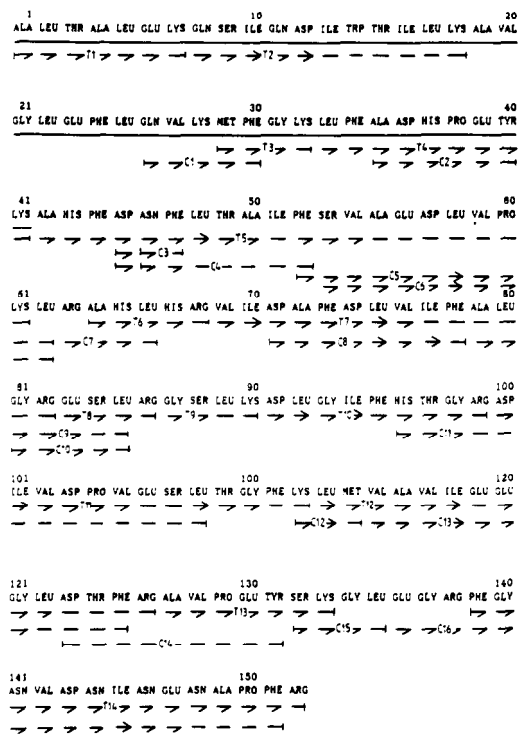

Fig. 4. Summary of the data used to establish the complete amino acid sequence of Artemia domain E7. T and C, tryptic and chymotryptic peptides respectively. Automated sequence determination : ———. Manual sequence determination : . . . . Identification of the DABTH amino acid residue by : TLC ⇝, HPLC ⇢, TLC + HPLC ⇛ . Substrated from amino acid composition of peptides : ——.

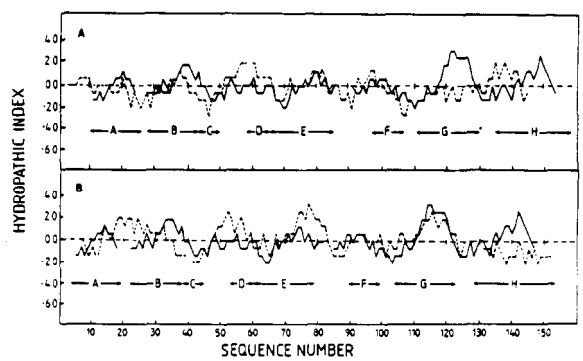

Fig. 11. Comparison of the hydrophobicity profiles of Artemia domain E1 and E7 with the human β chain. The hydrophobicity profiles were calculated according to Kyte and Doolittle (1982) using a window length of 9. A : Artemia domain E1 (----) versus human β chain (——). B : Artemia domain E7 (----) versus human β chain (——). The standard numbering and helix notification based on myoglobin is used.

```
                NA1      10   A1       10   15   AB1   B1       10   15   C1    7   CD1   8   D1    7   E1       10   15   20
                :        :    :        :    :    :     :        :    :    :     :   :     :   :     :   :        :    :    :
Hb ALPHA    : ---------VL  SPADKTNVKAAWGKVG  A---  HAGEYGAEALERMFLS  FPTTKTY  FPHF-DLS  -----HG  SAQVKGHGKKVADALTNAVA
Hb BETA     : ---------VHL TPEEKSAVTALWGKV-  ----  NVDEVGGEALGRLLVV  YPWTQRF  FESFGDLS  TPDAVMG  NPKVKAHGKKVLGAFSDGLA
MYO WHALE   : ---------VL  SEGEWQLVLHVWAKVE  A---  DVAGHGQDILIRLFKS  HPETLEK  FDRFKHLK  TEAEMKA  SEDLKKHGVTVLTALGAILK
CHIRON III: ----------L  SADQISTVQASFDKVK  G---  ----DPVGILYAVFKA  DPSIMAK  FTQFAG-K  DLESIKG  TAPFETHANRIVGFFSKIIG
CHIRON IX   : --------DPV  SSDEAAIRASWAGVK  H---  ----NEVDILAAVFSD  HPDIQAR  FPQFAG-K  DLASIKD  TGAFATHAGRIVGFISEIVA
GLYCERA     : ---------GL  SAAQRQVIAATWKDIA  GND-  NGAGVGKDCLIKHLSA  HPQMAAV  FG-FSG--  --AS---  DPAVADLGAKVLAZIGVAVS
APLYSIA     : ---------SL  SAAEADLAGKSWAPVF  A---  NKNANGLDFLVSLFEK  FPDVSAF  FADFADLL  --SKISA  DTKLSPVIRDVFSRILTFLN
SCAPHARCA   : PSVQGAA-AQL  TADVKKDLRDSWKVIG  S---  DKKGNGVALMTTLFAD  NQETIGY  FKRL----  GVSQGMA  NDKLRGHSITLMYALQNFID
TYLO 1      : ---------TDC  GILQRIKVKQQWAQVY  SVGE  SRTDFAIDVFNNFFRT  NPD-RSL  FNRV----  -NGDNVY  SPEFKAHMVRVFAGFDILIS
ARTEMIA E1: -ERVDPIT-GL  SGLEKNAILDTWGKVR  G---  NLQEVGKATFGKLFAA  HPEYQQM  FRFFQGVQ  LAF-LVQ  SPKFAAHTQRVVSALDQTLL
ARTEMIA E7: ----------AL  TALEKQSIQDIWTILK  AV--  GLEFLQVKMFGKLFAD  HPEYKAH  FDNFLTAI  FSVAEDL  VPKLRAHLHRVIDAFDLVIF
LUMBRICUS   : --------KKQC  GVLEGLEKVKSEWGRAY  GSGH  DREAFSQAIWRATFAQ  VPESRSL  FKRVHGDH  TS-----  DPAFIAHAERVLGGLDIAIS
PETROMYZON: PIVDTGSVAPL  SAAEKTKIRSAWAPVY  S---  NYETSGVDILVKFFTS  TPAAQEF  FPKFKGLT  TADQLKK  SADVRWHAERIINAVNDAVA
LEG-Hb      : --------GAL  TESQAALVKSSWEEFN  A---  NIPKHTHRFFILVLEI  APAAKDL  FS-FLK-G  TSEVPQN  NPELQAHAGKVFKLVYEAAI

                EF1      10   F1       10   FG1   G1       10   15   GH1 5   H1       10   15   20   25   HC1
                :        :    :        :    :     :        :    :    :   :   :        :    :    :    :    :
Hb ALPHA    : HV-------DDMPNA  LSALSDLHAH  KL--RV  DPVNFKLLSHCLLVTLAAH  LPAEF  TPAV-HASLDKFLASVSTVLTSKYR--  ------
Hb BETA     : HL-------DNLKGT  FATLSELHCD  KL--HV  DPENFRLLGNVLVCVLAHH  FGKEF  TPPV-QAAYQKVVAGVANALAHKYH--  ------
MYO WHALE   : K-------KGHHEAE  LKPLAQSHAT  KH--KI  PIKYLEFISEAIIHVLHSR  HPGDF  GADA-QGAMNKALELFRKDIAAKYKEL  GYQG--
CHIRON III: EL-------PNIEAD  VNTFVASHKP  RG---V  THDQLNNFRAGFVSYMKAH  T--DF  A-GA-EAAWGATLDIFFGMIFSKM---  ------
CHIRON IX   : LVGNESNAPAMATL  INELSTSHHN  RG---I  TRGQFNEFRSSLVSYLSSH  A--SW  N-DATADAWTHGLDNIFGMIFAHL---  ------
GLYCERA     : HLGDZGKMVAQ---  MKAVGVRHKG  YGNKHI  KGQYFEPLGASLLSAMEHR  IGGKM  NAAA-KDAWAAAYADISGALISGLQS-  ------
APLYSIA     : EAAF---AGKMSAM  LKQFAKEHVG  FG---V  GSAQFENVRSMFPGFVASI  SAPP-  ---AGDAAWTLLFGLIIDALKAAGK--  ------
SCAPHARCA   : Q---LDNTDDLVCV  VEKFAVNHIT  RK---I  SAAEFGKINGPIKKVLASK  NF---  -----GDKYANAWARLVAVVQAAL---  ------
TYLO 1      : V---LDDKPVLDQA  LAHYAAFHKQ  FGT--I  PFKAFGQTMEQTIAEHIHG  -----  --AD-IGAWRACYAEQIVTGITA----  ------
ARTEMIA E1: AL--NRPSDQFVYH  IKELGLDHIN  RGT---  DRSFVEYLKESLGDSVDEF  TVQSF  GAVI-VNFLNEGLRQA-----------  ------
ARTEMIA E7: AL-----GRESLRGS  LKDLGIFHTG  RDI--V  DPVESLTGFKLMVAVIEEG  LDT-F  --RA-VPEVSKGLEGRPGNVDNINENA  PFR---
LUMBRICUS   : TLD---QPATLKEE  LDHLQVQMEG  R---KI  PDNYFDAFKTAILHVVAAQ  LGDAI  ACDG-FARVLPQVLREGKIGHH-----  ------
PETROMYZON: SM----DDTEKMSMK  LRDLSGKHAK  SF--QV  DPQYFKVLAAVIADTVAAG  -----  ----DAGFEKLMSMICILLRSAY----  ------
LEG-Hb      : QLEV-TGVVVSDAT  LKNLGSVHVS  KG---V  ADAHFPVVKEAILKTIKEV  VGAKW  SEELNSAWTIAYDELAIVIKKEMDDAA  ------
```

Fig. 12. **Alignment of some selected globin sequences with Artemia domain E1 and E7.** The sequences as given in the NBRF databank were used. The standard numbering system based on myoglobin is used.

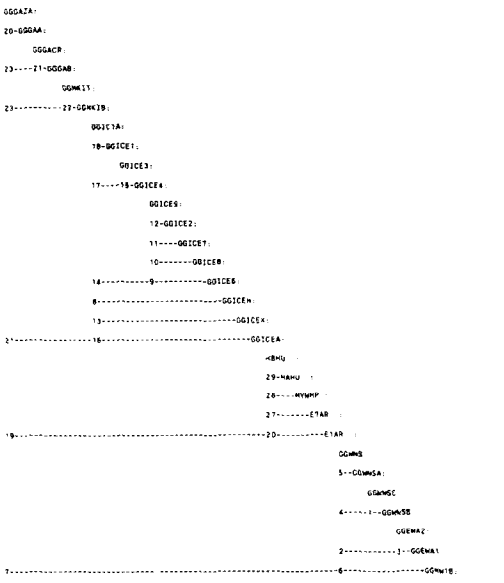

Fig. 13. **Maximum parsimony tree of selected globin chains.** Maximum parsimony tree was calculated as described in "Materials and Methods". The NBRF codes are :

```
GGGAZA   Aplysia kurodai                    HBHU    Homo B chain
GGGAA    Aplysia limacina                   HAHU    Homo a chain
GGGACR   Caristhidia rhizophorarum          MYWHP   Physeter catadon myoglobin
GGGAB    Busycon canaliculatum              E1AR    Artemia domain E1
GGNKIT   Anadara trapezia                   E7AR    Artemia domain E7
GGNKIB   Anadara broughtonii                GGNHS   Tylorhynchus heterochaetus (globin) small chain
GGICE1   Chironomus thummi thummi globin 1  GGNHSA  Tylorhynchus heterochaetus globin A
GGICIA   Chironomus thummi thummi globin 1A GGNHSB  Tylorhynchus heterochaetus globin B
GGICE2   Chironomus thummi thummi globin 2  GGNHSC  Tylorhynchus heterochaetus globin C
GGICE3   Chironomus thummi thummi globin 3  GGEWAI  Lumbricus terrestris globin I
GGICE4   Chironomus thummi thummi globin IV GGEWAZ  Lumbricus terrestris globin IIIA
GGICE6   Chironomus thummi thummi globin VI GGNWIB  Glycera dibranchiata
GGICE7   Chironomus thummi thummi globin VII
GGICE8   Chironomus thummi thummi globin VIIB
GGICEM   Chironomus thummi thummi globin VIII
GGICE9   Chironomus thummi thummi globin IX
GGICEX   Chironomus thummi thummi globin X
```

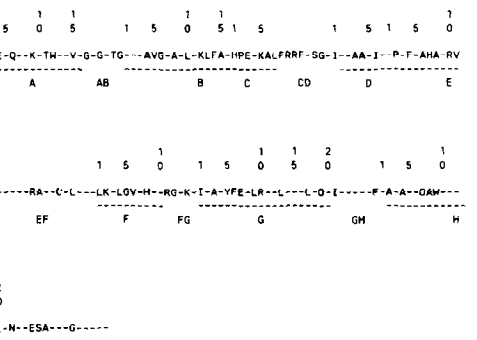

Fig. 14. **Ancestral globin sequence reconstructed by maximum parsimony method.**