THE JOURNAL OF BIOLOGICAL CHEMISTRY
© 1991 by The American Society for Biochemistry and Molecular Biology, Inc.

Vol. 266, No. 21, Issue of July 25, pp. 13789-13795, 1991
Printed in U.S.A.

The Polymeric Hemoglobin Molecule of *Artemia*

INTERPRETATION OF TRANSLATED cDNA SEQUENCE OF NINE DOMAINS*

(Received for publication, November 30, 1990)

Clive N. A. Trotman, Anthony M. Manning, Luc Moens‡, and Warren P. Tate

From the Department of Biochemistry, University of Otago, Dunedin, New Zealand and the ‡Department of Biochemistry, Universitaire Instelling Antwerpen, B-2610 Antwerp, Belgium

Translated cDNA for *Artemia* hemoglobin provided sequence data for almost nine domains, from the fourth residue of the A helix of one domain through 1405 residues to a stop codon after the ninth domain. The domain sequences were all different (homology between pairs 17–38%) but aligned well with each other and with conventional globins, satisfying the requirements for Phe at CD1, His at F8 and most other highly conserved features of globins including His at E7. Features found to be characteristic of *Artemia* globin and present in all nine domains were Phe at B10, Tyr at C4, Gly at F5, Phe at G5 and Gly at H22. Approximately 14 residues including a consensus -Val-Asp-Pro-Val-Thr-Gly-Leu- were available to form the linker between each pair of domains.

The *Artemia* sequence data were compared with the crystal structures of *Chironomus thummi thummi* erythrocruorin III and sperm whale myoglobin in order to identify features of structural similarity and to examine the consequences of the differences. The *Artemia* sequences were compatible with the main helices and critical features of the globin fold. Possible modifications to the C helix, FG turn, and GH turn were studied in terms of molecular coordinates.

Artemia has large hemoglobin molecules of estimated molecular weight 260,000, sufficient for 16 conventional globin domains (for review, see Ref. 1). Previous work on the isolated molecules showed the presence of two different M_r 130,000 subunits (α and β) (2), distinguishable biophysically (3), physiologically (4, 5), and immunologically (6), each comprising a number of concatenated domains resulting from the translation of a single mRNA (7). The two different subunits pair to form three alternative functional hemoglobins: HbI ($\alpha\alpha$), HbII ($\alpha\beta$) and HbIII ($\beta\beta$). By limited proteolysis of HbII, a number of polypeptides have been isolated (8), two of which have been sequenced and shown to be compatible with a globin-like structure (9, 10).

There are many unanswered questions about the structure of this unusual hemoglobin. What is the number of heme domains in each subunit, and is it the same in α and β? How uniform are the domains within a subunit? Are they all functional? What length is the interdomain linker? How are the domains organized within a subunit? How do the two subunits interact at the quaternary level? How similar are the domains to other invertebrate and vertebrate globins?

A cDNA made from an *Artemia* hemoglobin mRNA has been cloned (11) and sequenced (12) to answer some of these questions and for comparison with the genomic DNA in studies of the expression and regulation of the genes. From the sequence we have inferred that the multiple-domain structure could have arisen by gene duplication and fusion (12).

In the present paper we interpret the translated cDNA sequence in structural terms, in the absence of any crystallographic information on the protein. Almost nine domains are present in a continuous translation. Amino acid sequence data (13) complete the missing N terminus region and show it to be the α subunit. An alignment of nine sequences of similar domains from the same molecule provides a unique resource for the identification of characteristic features of the *Artemia* hemoglobin and provides a strong consensus for comparison with other species.

As well as having a convincing homology with invertebrate and vertebrate globin sequences, the *Artemia* domains displayed significant deviations from convention including an unprecedented Tyr at position C4. These deviations, together with a novel and well conserved interdomain linker of about 14 residues, broaden our knowledge of the constraints governing structure within the globin family. An understanding of how the specialized features of the *Artemia* hemoglobin are related to the specialized high salt, potentially low oxygen and sometimes high temperature niche of the organisms, will increase our understanding of structure-function relationships in oxygen carriers.

MATERIALS AND METHODS[1]

RESULTS AND DISCUSSION

Translation of the cDNA Sequence Data

Translation of the cDNA revealed a continuum of nine globin-like sequences with conserved residues and recognizable motifs in key locations (Table I), although only eight domains had been expected (1, 27). Since our cDNA contains a stop codon but not a start codon, the N-terminal 9 residues have been inferred from an overlapping polypeptide sequence.

Prominent globin features included the invariant CD1 Phe and F8 His, and the almost invariant E7 His (standard myoglobin numbering). Other typical globin features found in most domains included -Gly-Leu-Ser- at the start, a Trp at A12, -His-Pro-Glu- in the C-helix, and the common E11 Val. At five other sites the *Artemia* alignment was unanimous, namely a B10 Phe, C4 Tyr, F5 Gly, G5 Phe, and H22 Gly. A

* This work was supported by grants from the Lottery Grants Board and the University Grants Committee. 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.

[1] Portions of this paper (including "Materials and Methods" and Fig. 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.

TABLE I

Artemia globin domains T1–T9 aligned with sperm whale myoglobin, C. thummi thummi erythrocruorin III and N terminus of Petromyzon marinus globin V

Part of the *Artemia* linker and the A helix are repeated at both ends for convenience. The *Artemia* sequence from A4 to the C-terminus is the translation of the cDNA sequence (12). From the N terminus to position A3 was taken from the directly determined amino acid sequence (13), the first 40 residues of which had the following differences from T1: A13 (S), A14 (I), A16 (N), B13 (N). Deviations from the template of Bashford *et al.* (21) are emphasized.

	Linker	NA	A 1	AB	B 1	C
		12	1234567890123456	12	1234567890123456	1234567
Artemia T1	AEVS	GI	LVSDKATIKRTWATVT		DLPSFGRNVFLSVFAA	KPE**Y**KNL
Artemia T2	LRRQIDLEVT	GL	SCVDVANIQESWSKVS	G	DLKTTGSVVFQRM**I**NG	HPE**Y**QQL
Artemia T3	SSLKRVDPIT	GL	SGLEKNAILSTWGKVR	G	NLQEVGKATFGKLFTA	HPE**Y**QQM
Artemia T4	RQADIVDPVT	HL	TGRQKEMIKASWSKAR	T	DLRSLGQELFMRMFKA	HPE**Y**QTL
Artemia T5	ATSEEADPVT	GL	YGKEI**V**ALRQAFAAVT	P	RNVEIGKRVFAKLFAA	HPE**Y**KNL
Artemia T6	FQLGQVDSNT	L	TALEKQSIQDIWSNLR	ST	GLQDLAVKIFTRLFSA	HPE**Y**K**L**L
Artemia T7	LQLERINPIT	GL	SAREVAVVKQTWN**L**VK	P	DLMGVGMRIFKSLFEA	FPA**Y**QAV
Artemia T8	QQSYKQDPVT	GI	TDAEKALVQESWD**L**LK	P	DLLGLGRKIFTKVFTK	HPD**Y**Q**I**L
Artemia T9	IGLKEVNPQN	AF	SAYDIQAVQRTWA**L**AK	P	DLMGKGAMVFKQLFTD	H**G** **Y**QPL
Consensus	LQL VDPVT	GL	S EKAAIQ TW V	P	DL GLG VF LFTA	HPEYQ L
Myoglobin		VL	SEGEWQLVLHVWAKVE	A	DVAGHGQDILIRLFKS	HPETLEK
Chironomus		L	SADQISTVQASFDKVK	G	DPVGILYAVFKA	DPSIMAK

	CD	D	E 1 2	EF 1
	123 45678	1234567	12345678901234567890	12345678901
Artemia T1	F**V**E FRNI	PASELAS	SERL**LY**HGGRVLSSIDEAIA	GIATPDRAVKT
Artemia T2	FRQ FRDV	DLDKLGE	SNSF**V**AHV**F**RVVAAFDGIIH	ELDNNQFIVST
Artemia T3	FRFS QGM	PLASLVE	SPKFAAHTQRVVSALDQTL**L**	ALNRPSDFVYM
Artemia T4	F**V**NKGFADV	PLVSLRE	DERF**I**SH**M**ANVLGGFDTLLQ	NLDESSYFIYS
Artemia T5	FKK FEQY	SVEELPS	TDAFH**Y**H**I**S**L**VMN**R**FSSI**G**K	VIDDNVSFVYL
Artemia T6	FTGR FGNV	DNINE	NAPFKAH**L**HRVLSAFD**I**VIS	TLDDSEHLIRQ
Artemia T7	FPK FSDV	PLDKLED	TPAVGKHS**I**SVTT**K**LDELIQ	TLDEPANLALL
Artemia T8	FTRTGFGDT	PLTKLDD	NPAFGTH**II**KVMRAFDHVIQ	ILGKPKTLMAY
Artemia T9	FSN LAQY	EITGLEG	SPELNTHARNVMA**Q**LDTLVG	SLQNSIELGQS
Consensus	F GF DV	PLDKLEE	SPAF AHIIRV SAFD LIQ	TLDBP LVY
Myoglobin	FDR FKHLK	TEAEMKA	SEDLKKHGVTVLTALGAILK	KGHHEAE
Chironomus	FTQ FAG K	DLESIKG	TAPFETHANRIVGFFSKIIG	ELPN IEAD

	F 1	FG	G 1	GH
	1234567890	1234	1234567890123456789	12345
Artemia T1	L**L**ALGERHIS	RG**T**	**V**RRHFEAFS**Y**AFID**E**LKQR	G V
Artemia T2	LKKLGEQHIA	RG**T**	DISHFQNFR**V**TLLEYLKEN	G M
Artemia T3	IKELG**L**DHIN	RG**T**	DRSHFEN**Y**Q**V**VFIEYLKET	LGDSL
Artemia T4	LRNLGDAHIQ	RK**A**	GTQHFRSFEAILIPILQES	Q G L
Artemia T5	LKKLGREHIK	RGL	SRKQFDQF**V**EL**Y**IA**E**ISSE	L S
Artemia T6	LKDLG**L**FHTR	LG**M**	TRSHFDNFATAFLSV**A**QDI	APNQL
Artemia T7	ARQLGEDHIV	LRV	NKPMFKSFGKVLVRLLEND	LGQRF
Artemia T8	LRS**V**GADHIA	TNV	ERRHFQAFSNALIPVMQHD	LKAQL
Artemia T9	LAQLGKDHVP	RKV	NRVHFKDFAEHFIPLMKAD	LGDEF
Consensus	LK LGEDHIA	RG	RSHFZNF A IP LKED	LGDQL
Myoglobin	LKPLAQSHAT	KHKI	PIKYLEFISEAIIHVLHSR	HPGDF
Chironomus	VNTFVASHKP	RGV	THDQLNNFRAGFVSYMKAH	T D F

	H 1	HA 1	A 1
	12345678901234567890	1234567890123456	123456789012
Artemia T1	ESADLAAWRRGWDNIVNVLEAGL	LRRQIDLEVTGL	SCVDVANIQESW
Artemia T2	NGAQKASWNKAFDAF**E**KYISMGL	SSLKRVDPITGL	SGLEKNAILSTW
Artemia T3	DEFT**V**KSFNHVFEVIISFLNEGL	RQADIVDPVTHL	TGRQKEMIKASW
Artemia T4	DAAS**V**EAWKKFFDV**S**IGVIAQGL	KVATSEEADPVTGL	YGKEI**V**ALRQAF
Artemia T5	DT GRNGLEKVLT**F**ATGVIEQGL	FQLGQVDSNT L	TALEKQSIQDIW
Artemia T6	TVLGRESLNKGFKLMHGVIEEGL	LQLERINPITGL	SAREVAVVKQTW
Artemia T7	SSFASRSWHKAYDVIVEYIEEGL	QQSYKQDPVTGI	TDAEKALVQESW
Artemia T8	RPDA**V**AAWRKGLDRIIGIIDQGL	IGLKEVNPQNAF	SAYDIQAVQRTW
Artemia T9	TPLAESAWKRAFDVMIATIEQGQ	EGSSHALSSFLT	NPVA*
Consensus	D AAVA WNK FDVIIGVIEQGL	LQL VDPVTGL	S LEKAAIQ W
Myoglobin	GADAQGAMNKALELFRKDIAAKYKELGYQG*		
Chironomus	A GAEAAWGATLDTFFGMIFSKM*		
Petromyzon		PIVDTGSVAPL	SAAEKTKIRSAW

consensus -Val-Asp-Pro-Val-Thr-Gly-Leu- characterized the interdomain linker.

The nine domains (translates T1–T9, Table I) had cross-homology of between 17% (T5/T9) and 38% (T2/T3). Domain T3 was recognized as the polypeptide E1 sequenced directly by Moens *et al.* (9, 27); small differences between T3 and E1 were considered to be a consequence of polymorphism since doubt surrounds the taxonomy of commercial *Artemia* cysts. The E7 polypeptide sequence (10, 28) was not found as such, although regions of similarity to it were present in T6, and it may belong to the other subunit.

When the sequences were compared with the database of globin sequences (16) by FASTP, high homology scores were achieved against invertebrate globins, notably those of *Chironomus* (of which the database contains 11), *Lumbricus*, *Tylorrhynchus*, and molluscs. If submitted as fragments corresponding to vertebrate globin exons, the central fragments showed the most consistent recognition of invertebrate species, particularly *Chironomus*. Vertebrate globins, especially myoglobins, were widely recognized by *Artemia* domains and fragments with FASTP, in some cases scoring higher than invertebrates. The composite hydropathicity profile of the *Artemia* domains (Fig. 1) tends, however, to resemble *Chironomus* globin III where that differs from myoglobin, for example in the C helix, D-E boundary, and late H helix. Regions of markedly greater hydrophobicity in *Artemia* than in either *Chironomus* or myoglobin are observed in the C, D, E, and G helices, and EF turn; these could manifest the multiple interdomain and intersubunit associations implicit in a multimeric structure.

Homology and Structural Consequences at Individual Residues

The alignment was refined with the aid of the Bashford *et al.* (21) theoretical template II, which incorporates the more significant residues. Discrepancies between the aligned sequences and the template were relatively few (Table II), some being sufficiently consistent to imply that they should not be penalized. After assignment of the conserved sites and the major helices A to H (Table I), disparities in length fell mostly into the interhelical turns and an additional interdomain linker (HA).

In the following discussion we examine possible structural consequences where a sequence deviates from the knowledge base defining the globin family. The advantage of having nine *Artemia* sequences is that they can be cross-correlated, as well as being compared with conventional globins. Some deviations were modeled by altering the structural coordinates of either *Chironomus* globin III or sperm whale myoglobin. Sequence homology, however good, is not proof of structural similarity, but there would be no realistic prospect of an unrelated sequence fulfilling the prerequisites of the globin fold since the contribution of each successive conforming residue is factorial.

A Helix—Following the NA region we expect an initial α helix and regard a Trp at A12 as almost diagnostic, Bashford *et al.* (21) finding it in 198 globins out of 226. Eight *Artemia* domains had a Trp in the right place while the exception, T5, had a Phe which is the best substitute in the database (21 of the remaining 28 sequences). The Val at A6 in T5 is sterically comparable with the Thr often observed at A6 and is not unknown at this site (21), but it introduces a recurring theme in the tendency for substitutes to be more hydrophobic in *Artemia*. The next such example is the Leu appearing at A14 in T7, T8, and T9, which incurs a template penalty of 1.0 and has been observed in only three other globins in the database. The A14 Leu appears to be reliable since the three *Artemia* domains in which it appears show a high degree of consistency from A8 to B6.

TABLE II

Against each domain is shown any position and residue incurring a penalty (between 0 and 1) against the Bashford et al. *(21) theoretical template II, listed under the penalty score*

Italics indicate buried sites (21), but E14 is listed as a surface site by Perutz *et al.* (29).

Domain	Penalty				
	0.2	0.5	0.7	1.0	Total
T1	*FG4* Thr	CD2 Val G1 Val	*C4* Tyr	E5 Leu E6 Tyr F2 Leu G10 Tyr *G15* Glu	6.9
T2	*B14* Ile *FG4* Thr	E5 Val G10 Val	*C4* Tyr	E9 Phe H16 Glu	4.1
T3	*FG4* Thr *G8* Tyr	G10 Val H5 Val	*C4* Tyr	E20 Leu F6 Leu	4.1
T4		CD2 Val H5 Val	*C4* Tyr *E8* Met *FG4* Ala *H15* Ser	E5 Ile	4.8
T5	*E8* Ile	A6 Val G9 Val	*C4* Tyr *E19* Gly *G12* Tyr	E6 Tyr E10 Leu *E14* Arg *G15* Glu H014 Phe	8.3
T6	*FG4* Met		*C4* Tyr *E8* Leu G16 Ala	C6 Leu E17 Ile F6 Leu	5.3
T7			*C4* Tyr	A14 Leu E9 Ile *E14* Lys	3.7
T8	*E8* Ile	H5 Val	*C4* Tyr *F4* Val	A14 Leu C6 Ile E9 Ile	5.1
T9			*C2* Gly *C4* Tyr	A14 Leu *E14* Gln	3.4

AB Turn—In domain T6 a strong alignment in the A, B, and C helices dictates the placement of 2 residues, -Ser-Thr- (followed by Gly), between the A and B helices. AB turns sometimes differ from the more usual single Ala or Gly and the T6 sequence is not unlike the -Ser-Val-Gly- of *Tylorrhynchus* (16) in having a branched side-chain at AB2. Chou and Fasman predictions (19) on the AB region for domain T6 gave a turn product (×10^5) of 17.2, comparable with the other *Artemia* domains (9.5–23.3). *Chironomus* globin III has a similar value (18.8) while in whale myoglobin the low value (4.0) does not predict the turn.

B Helix—The fit against template II (21) is excellent, the only minor deviation being from Phe to Ile at B14 in T2, while even the more restrictive template I finds only a Gln in T9 at B12 where Arg is common. In all *Artemia* domains B10 is a distinctive Phe, which is found in only five instances in the database against 204 examples of a B10 Leu. Although it does not feature in the templates, a B6 Gly, as found in eight domains, is well conserved in globins and will be considered with the E helix.

C Helix—The high degree of conservation in the critical heme environment is the key to the alignment of the adjacent regions. The Tyr at C4 is not known in other species but is universal in the *Artemia* domains (Table III). The -His-Pro-Glu- found at C1–3 in domains T2–T6 is identical to many myoglobins. The variants on this theme in T1, T7, and T8 are conservative except for the C1 Phe in T7, which is unexceptionable since Phe is almost as common as His at C1 in the database.

A Pro at C2 is regarded as practically invariant (222/226 in the database) but a striking exception was found in T9, in which the C helix sequence was also shorter than expected. The compelling alignment of T9 through B13–14, C1, C4–5,

TABLE III

Alignment around the C helix

Residues predicted to be in turns by the algorithm of Chou and Fasman are underlined. The last column gives the highest turn product found within the partial sequence.

	B13	C1	C7	CD1	Turn product × 10^5
T1	V F A A	K P E Y K N	L	F	15.9
T2	M I N G	H P E Y Q Q	L	F	40.6
T3	L F T A	H P E Y Q Q	M	F	40.6
T4	M F K A	H P E Y Q T	L	F	40.6
T5	L F A A	H P E Y K N	L	F	40.6
T6	L F S A	H P E Y K L	L	F	40.6
T7	L F E A	F P A Y Q A	V	F	8.7
T8	V F T K	H P D Y Q I	L	F	94.3
T9	L F T D	H G Y Q P	L	F	16.4
Consensus	L F T A	H P E Y Q Q	L	F	
Chironomus	V F K A	D P S I M A	K	F	
Myoglobin	L F K S	H P E T L E	K	F	

C7, and CD1 leaves few options for adjustments and it is inferred that a single Gly occupies the place of the conventional C2 Pro and its following residue.

The sequence variants found in this region were all turn-predictive (Table III). One interpretation of the generally high scores is that tight constraints on the structure of this part of the heme environment are reflected in a highly specified turn, as distinct from turns such as the AB where a hinge is required during folding.

A reconstruction of the C4 region to incorporate a Tyr was based on the structure of *Chironomus* Hb III, which has a large hydrophobic Ile at C4. Alignment of the C-α–C-β bond of the replacement Tyr with the equivalent bond of the C4 Ile *in situ* allowed space for the Tyr ring in the vicinity of the heme (Fig. 2*A*). Van der Waals conflicts with the B14 and CD1 Phe rings, which are conserved in *Artemia*, were resolvable by rotation of the bonds of the C-β atom of the mutant sidechain, placing the C4 ring in an existing hydrophobic environment.

The unprecedented deletion of C2 or C3 in domain T9 was explored by modeling. In the *Chironomus* Hb III structure, deletion of C2 and C3 left the C terminus of C1 and the N terminus of C4 suitably distanced for the gap to be filled by a single Gly (Fig. 2*B*). The amide nitrogen of the *Chironomus* Hb III C4 Ile and the carboxyl oxygen of the C1 Asp were nominally 2.99 Å apart, or 2.94 Å if the latter oxygen were exchanged with the C2 amide nitrogen. This distance is easily bridged by 1 Gly. Main-chain reconstruction is speculative, although clues to the quaternary arrangement (13) suggest that this region in *Artemia* is exposed and free from the constraints of interdomain switching imposed on mammalian hemoglobins.

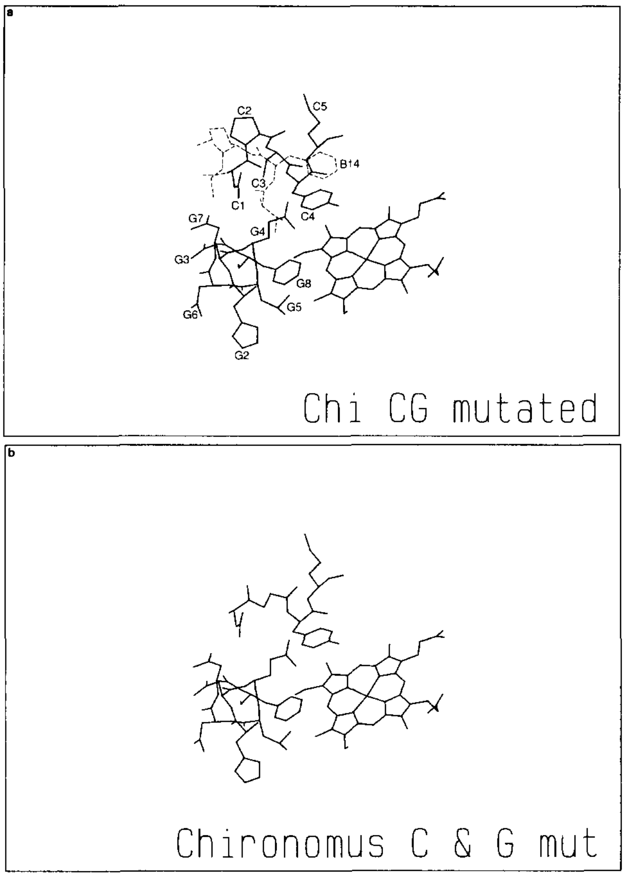

FIG. 2. *A*, a region of the *Chironomus* Hb III heme environment after replacement of the C4 Ile with a Tyr to simulate *Artemia* globins. The upper fragment is B14–B16 (*dashed line*, right to left) then C1–C5, from left to right; lower left fragment is the G helix receding from the viewer. *B*, as above, with C2 Pro and C3 Ser replaced by a single Gly (B helix fragment omitted for clarity).

CD Turn and D Helix—The pivotal locations of CD1 Phe and E7 His leave little scope for adjustment in the short region between them. A convincing alignment with the buried D5 Ile and D2 Leu of *Chironomus* Hb III, with a Pro or hydrophilic residue at D1, places most length variations in the CD turn. Variations in loop structures can be inconsequential but the globin CD turn is special in forming part of the heme environment. It is instructive to identify the hydrophobic environment usually associated with CD4, which is Phe in 206/226 sequences in the database, and with CD7 (although the latter is anomalously hydrophilic in *Chironomus* Hb III). The alignment suggests that both of these sites are represented in *Artemia* provided the first half of the CD turn accommodates 1 extra residue in T6 and 2 in T4 and T8. In T6 the late CD turn or early D helix appears to be truncated.

Domain T3 requires mention as the only one without an ideal CD4 candidate. T3 is equivalent to the polypeptide E1 where the amino acid sequence has a perfectly acceptable Phe at CD4 (9, 27) in place of the Ser in T3. The polypeptide and cDNA sequences in this part of the structure have been carefully re-examined, and the difference is attributed to polymorphism.

E Helix—Exceptions are rare to the rule of a distal His at E7 accompanied by a Val at E11. *Artemia* fitted this rule, strengthened by an identical spacing of 32 residues between E7 and the F8 proximal His in every domain. The requirements for buried sidechains at E4, E12, and E15 were also consistently satisfied.

This strong alignment is not conclusive evidence that E7 His in *Artemia* fulfils the conventional distal His role, however, since there are exceptions and *Chironomus* Hb III happens to be an example. The E7 His side-chain in *Chironomus* Hb III is displaced toward the start of the E helix and the E11 Ile side-chain assumes its function (30), with one of its γ-carbon atoms positioned closest to the Fe (4.48 Å). In *Artemia* the E7 His, E11 Val motif suggests a conventional heme environment but a Val is a plausible alternative to a distal Ile since it has analogous γ-carbon atoms.

The E helix was the region with the most points of difference from template II (21), but they follow a pattern. Of the 19 discrepancies, 15 are hydrophobic and of these, 11 appear at the equivalent of relatively exposed sites in higher globins, namely E5 (three instances), E6 (2), E9 (3), E10, E17, and E20. None of these sites is hydrophobic in *Chironomus* Hb III, except for an interesting polymorphism at E6 where either Thr or Ile is present in a 2:1 ratio (30). A net increase in hydrophobicity associated with the surface of the E helix may have important implications for the association of domains either within or between the polymeric subunits. In three domains a large hydrophilic residue appears at E14, which the template treats as buried (21) but is listed by Perutz *et al.* (28) as surface.

The appearance of Gly at both B6 and E8 in higher globins is rationalized as permitting a class 1 crossing (31) of the B and E helices with the main-chains in close proximity. This combination is not totally conserved among vertebrate globins and is not characteristic of invertebrates, the *Chironomus* Hb III crossing involving Pro at B6 and Val at B7, fitting between Ala at E8 and Val at E12. The effect is that the helix axes cross a little further apart, 8.3 Å at an angle of 85° in *Chironomus* Hb III and 8.1 Å and 80° in myoglobin. In common with many invertebrates *Artemia* has retained a B6

Gly in eight domains, but only T1 has a Gly at E8.

EF Turn and F Helix—The proximal His at F8 confirms the overall F helix alignment including the highly conserved F4 Leu in the heme environment, with an acceptable F4 Val in T8. A Gly was common to all *Artemia* domains at F5, as was a branched side-chain at the F9 position.

As in the E helix, the main discrepancies from the template (21) were hydrophobic replacements at surface sites, namely one instance of a Leu at F2 and two of a Leu at F6. The E and F helices are approximately coplanar and it is not unknown for this face to be involved in interdomain contact (32).

The strong alignment in the E and F helices, particularly at E7, E11, and F8, indicates a consistent EF turn of 10 residues. The EF turn is conspicuous for being the most uniform of the *Artemia* turns in both length and composition (apart from the CD turn) but its excess of 3 or 4 residues over *Chironomus* Hb III or myoglobin respectively would make structure prediction speculative.

FG Turn—The alignment is influenced by the "severe" size constraint (21) at FG4 which is satisfied with a penalty of 0 or 0.2 in all domains except T4, where it is 0.7. Furthermore, domains T1–T5 and also T9 have an Arg aligned with the Arg in *Chironomus* Hb III at FG2 (FG1 being deleted) which hydrogen bonds to the heme porpionate and assumes the role of the myoglobin FG2 His (9). Three domains (T6–T8) do not have an equivalent hydrogen bond donor. Domain T6 has a nearby Arg at F10 which, if its side-chain were reoriented, would remain 1 Å too distant for re-establishing the Hbond, as can be appreciated from Fig. 3*A*. The T7 and T8 hydrogen bond donors at FG3 would be too far from the propionate. Changing the alignment of T7 or T8 to fit myoglobin (Fig. 3*B*) so that FG1 is occupied, does move hydrogen donors into FG2, but creates new problems: Arg is too large, the constraints on FG4 are violated and the distance between FG2 and G1 (9.36 Å) cannot be spanned by the single Val available.

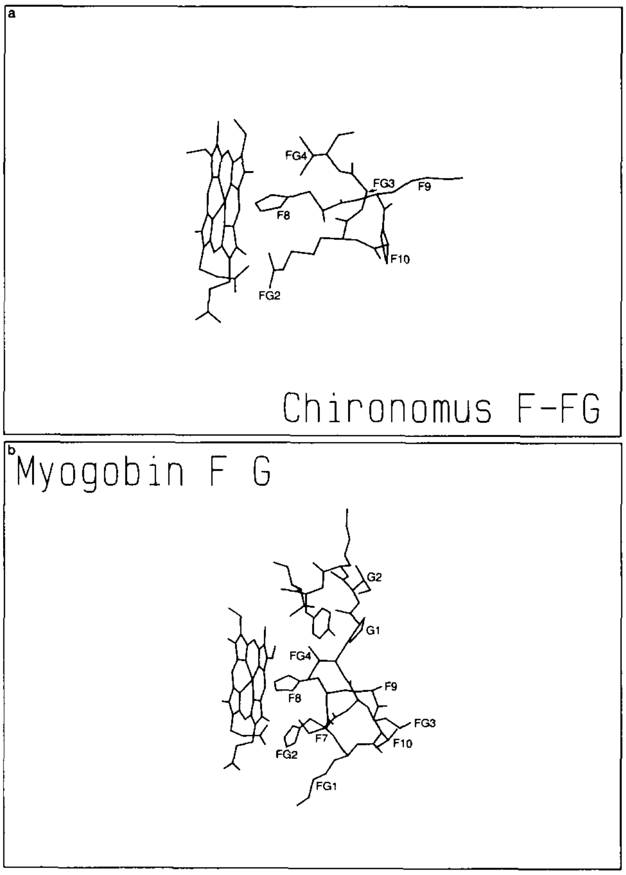

FIG. 3. *A*, *Chironomus* Hb III from F8 His (nearest viewer) to FG4 (distant) showing interaction with heme. Six *Artemia* domains have an equivalent Arg to FG2. The condition in the remaining three domains is discussed in the text. *B*, myoglobin in similar orientation to *A*.

While T1–T5 and T9 can conform to the *Chironomus* Hb III model, where the heme pocket design permits a degree of heme reversal about the α-γ meso axis (30), it appears that T6, T7, and T8 may have a different mode of heme-packing.

G Helix—The G helix buried sites having severe and "medium" size constraints (21) at G5 and G8, respectively, are well satisfied in the *Artemia* alignment, in which all are occupied by Phe except for a Tyr at G8 in T3. This alignment is reinforced with few exceptions at the "low" severity interior positions G12, G15, and G16, and in seven domains by a conspicuous His at G4 which is rare in the database (4/226). The G helix alignment further strengthens that of the FG turn.

GH Turn and H Helix—A strong alignment toward the latter part of the H helix (and the linker following) indicated the GH turn to be variable in length. The appearance of Trp at H8 in six domains was a key to the alignment since Trp is scarce in globins, and H8 is one of the few positions (A5, A12, C3, H8) where its side-chain is accommodated with any regularity. The Leu or Phe found in three domains are common substitutes for Trp at H8.

Comparison of myoglobin and *Chironomus* Hb III structures revealed practically identical distances between the GH1 and H8 α-carbons (12.93 Å) or between GH1 and H5 (10.85 Å), although *Chironomus* Hb III uses 3 fewer residues (Fig. 4). Variations in length appear to be tolerated around the GH turn and a skeleton structure comprises GH1, H5, and H8 with a hydrophobic GH5 positioned between the G and H helices.

Five of the *Artemia* sequences (T3, T6–T9) are the same length as myoglobin through this region and a Leu or Phe appears at GH5 appropriately. The four shorter ones have variously 4, 5, or 6 residues available to insert between GH1 and H5, thus the structure is not constant. Tolerable buried GH5 residues are suggested in Table I, and it is observed that T1, T3, T4, and T8 have an H5 side-chain that alternatively could rotate to occupy the GH5 space.

Interdomain Linker—Where does the interdomain linker begin to deviate from the conventional globin fold? The indications are that the H helix is fully retained since between H5 and the end of the template at H19 there were only three discrepancies. This, with the spare residues after each H helix, implies that the linker utilizes additional sequence as distinct from an adaptation of the H or A helices.

Between 13 and 16 extra residues are available following H21 and preceding A1 for interdomain linkage, which we designate HA (absorbing any positions equivalent to HC and NA). No structural precedent is available, but the sequences provide some significant clues. The well conserved character of the HA alignment and its neighbors suggests that the linkers are of uniform structure. This in turn is compatible with a regular arrangement of domains such as the cylindrical morphology seen in electron micrographs (3).

HA1 and HA2 are unanimously Gly-Leu except in T9 where it is not a true linker. The next 5 residues (7 in T4; HA3 to HA9) contain a high proportion with large, hydrophilic side-chains. Then follows a consensus -Val-Asp-Pro-Val-Thr-Gly-Leu- over HA10–HA16 which is one of the longest and clearest in the *Artemia* alignment and has consistent turn forming

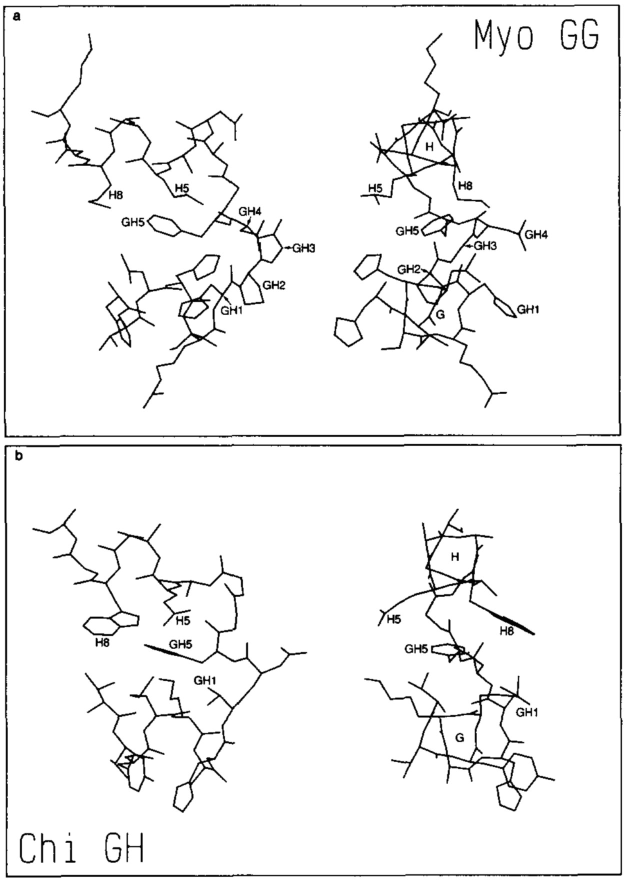

FIG. 4. **Comparison of the region GGHH in whale myoglobin (*A*) and *Chironomus* Hb III (*B*).** Side views (*left*) and external end views (*right*). Distances between GH1 and H5 or H8 α-carbon atoms are indistinguishable and the position of the GH5 Phe is similar. Modeling of the GH turns of the *Artemia* domains is discussed in the text.

potential. Finally, conservation of the NA2 (HA16) Leu, Ile, or Phe, classically assigned to contact with the H helix, suggests that the globin fold is restored by this point. Domain E1 isolated after limited proteolysis had been cut between HA7 and HA8 at each end (27), and domain E-7 ran from HA15 to HA8 (10), compatible with the linker containing the potentially flexible bridge structure that favors proteolysis.

Few leads toward the possible structure of this linker sequence are to be found in the database or in other interprotein linkers (33). A fragment of human interleukin-1-β has the sequence -Leu-Gln-Leu-Glu-Ser-Val-Asp-Pro-, which bears comparison with HA5–HA12 of *Artemia*. This part of interleukin-1-β is described (34) as a bent β sheet leading to an extended turn (we do not have coordinates). A more attractive prospect, however, may exist in *Petromyzon* globin V, which has an unusually long sequence of 11 residues preceding the A helix (26, 35). This extension has a remarkably similar composition to the *Artemia* linkers, although the sequences do not correspond, and demonstrates the acceptability of an additional structure on this part of the globin molecule. Studies in progress indicate that an extended *Petromyzon* globin V leader could provide the structural template for the *Artemia* linker and would be compatible with an annular arrangement of nine domains.

In conclusion, the interplay of data from an analysis of nine related domains has provided a higher level of resolution than could be derived from one or two sequences. The C4 Tyr, for example, is unique among species studied. In the absence of definitive structural data we nevertheless have an image of a multimeric molecule in which the domains are conventional yet characterstic, constructed from the usual A–H helices with variable turns and a lengthy EF loop. The relatively uniform linkers and the hint of hydrophobicity associated with the E helix are suggestive of an orderly arrangement of domains within the subunit and, by implication, its partner.

Acknowledgments—We thank the Trustees of the E. G. Johnstone Trust, of the J. D. S. Roberts Trust, and of the Todd Foundation, New Zealand. Expert advice on computing problems was generously given by Drs. Peter Stockwell, Geoff Wyvill, and Brendan Murray.

REFERENCES

1. Moens, L., Wolf, G., Van Hauwaert, M-L., De Baere, I., Van Beeumen, J., Wodak, S. & Trotman, C. N. A. (1990) in *Artemia Biology* (Browne, R. A., Sorgeloos, P. & Trotman, C. N. A., eds) pp. 187–219, CRC Press, Inc., Boca Raton
2. Moens, L. & Kondo, M. (1978) *Eur. J. Biochem.* **82,** 65–72
3. Wood, E. J., Barker, C., Moens, L., Jacob, W., Heip, J. & Kondo, M. (1981) *Biochem. J.* **193,** 353–359
4. D'Hondt, J., Moens, L., Heip, J., D'Hondt, A. & Kondo, M. (1978) *Biochem. J.* **171,** 705–710
5. Wolf, G., Van Pachtenbeke, M., Moens, L. & Van Hauwaert, M-L. (1983) *Comp. Biochem. Physiol.* **76B,** 731–736
6. Marshall, C. J., Cutfield, J. F., Trotman, C. N. A. & Tate, W. P. (1987) *Biochem. Int.* **15,** 925–933
7. Manning, A. M., Ting, G. S., Mansfield, B. C., Trotman, C. N. A. & Tate, W. P. (1986) *Biochem. Int.* **12,** 715–724
8. Moens, L., Geelen, D., Van Hauwaert, M-L., Wolf, G., Blust, R., Witters, R. & Lontie, R. (1984) *Biochem. J.* **223,** 861–869
9. 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
10. Moens, L., Van Hauwaert, M-L., De Smet, K., Ver Donck, K., Van De Peer, Y., Van Beeumen, J., Wodak, S., Alard, P. & Trotman, C. (1990) *J. Biol. Chem.* **265,** 14285–14291
11. Manning, A. M., Powell, R. J., Trotman, C. N. A. & Tate, W. P. (1990) *New Biologist* **2,** 77–83
12. Manning, A. M., Trotman, C. N. A. & Tate, W. P. (1990) *Nature* **348,** 653–656
13. Moens, L., Ver Donck, K., De Smet, K., Van Hauwaert, M. L., Van Beeumen, J., Allard, P., Wodak, S. & Trotman, C. N. A. (1989) *NATO Adv. Study Inst. Ser. Ser. A Life Sci.* **174,** 429–438
14. Stockwell, P. A. (1987) in *Nucleic Acid and Protein Sequence Analysis: A Practical Approach* (Bishop, M. J. & Rawlings, C. J., eds) pp. 19–45, IRL Press, Oxford
15. Lipman, D. J. & Pearson, W. R. (1985) *Science* **227,** 1435–1441
16. 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.
17. Stockwell, P. A. & Petersen, G. B. (1987) *CABIOS* **3,** 37–43
18. Stockwell, P. A. (1983) *Trends Biochem. Sci.* **13,** 322–323
19. Rawlings, N., Ashman, K. & Wittmann-Liebold, B. (1983) *Int. J. Pept. Protein Res.* **22,** 515–524
20. Hopp, T. P. & Woods, K. R. (1981) *Proc. Natl. Acad. Sci. U. S. A.* **78,** 3824–3838
21. Bashford, D., Chothia, C. & Lesk, A. M. (1987) *J. Mol. Biol.* **196,** 199–216
22. Crabbe, M. J. C. & Appleyard, J. R. (1989) *Desktop Molecular Modeller,* Oxford University Press, Oxford
23. Manning, A. M., Marshall, C. J., Powell, R. J., Trotman, C. N. A. & Tate, W. P. (1989) *NATO Adv. Study Inst. Ser. Ser. A Life Sci.* **174,** 413–425
24. Steigemann, W. & Weber, E. (1979) *J. Mol. Biol.* **127,** 309–338
25. Phillips, S. E. V. (1980) *J. Mol. Biol.* **142,** 531–554
26. Honzatco, R. B., Hendrickson, W. A. & Love, W. E. (1985) *J. Mol. Biol.* **184,** 147–164
27. Moens, L., Van Hauwaert, M-L., Geelen, D., Verpooten, G. & Van Beeumen, J. (1987) in *Artemia Research and Its Applications* (Decleir, W., Moens, L., Slegers, H., Jaspers, E. & Sor-

geloos, P., eds) Vol. 2, pp. 93–98, Universa Press, Wetteren, Belgium

28. De Smet, K., Van Hauwaert, M-L., Moens, L. & Van Beeumen, J. (1987) in *Artemia Research and its Applications* (Decleir, W., Moens, L., Slegers, H., Jaspers, E. & Sorgeloos, P., eds) Vol. 2, pp. 41–51, Universa Press, Wetteren, Belgium
29. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965) *J. Mol. Biol.* **13,** 669–678
30. Osmulski, P. A. & Leyko, W. (1986) *Comp. Biochem. Physiol.* **85B,** 701–722
31. Richmond, T. J. & Richards, F. M. (1978) *J. Mol. Biol.* **119,** 537–555
32. Royer, W. E., Love, W. E. & Fenderson, F. F. (1985) *Nature* **316,** 277–280
33. Argos, P. (1990) *J. Mol. Biol.* **211,** 943–958
34. Priestle, J. P., Schar, H-P. & Grutter, M. G. (1988) *EMBO J.* **7,** 339–343
35. Hombrados, I., Rodewald, K., Neuzil, E. & Braunitzer, G. (1983) *Biochimie* **65,** 247–257

Supplemental material to:

"The Polymeric Hemoglobin Molecule of *Artemia*
Interpretation of Translated cDNA Sequence of Nine Domains"

Clive N.A.Trotman, Anthony M.Manning, Luc Moens and Warren P.Tate

MATERIALS AND METHODS

1. Computer Programs

The following programs were kindly provided and maintained by Dr Peter Stockwell:
NLDNA translates and performs a comprehensive range of other manipulations on DNA sequences (14);
FASTP searches protein sequence databases for homologues (15); this was operated in conjunction with a database of 329 globin sequences extracted from the NIH Protein Information Resource, PIR (16);
HOMED is a homology editing facility (17,18);
CHUFAS performs secondary structure prediction on peptides by the method of Chou and Fasman with the parameters of Rawlings, Ashman and Wittmann-Liebold (19);
HYDROP plots relative hydrophobicity along a polypeptide sequence; parameters selected were those of Hopp and Woods (20) with a window length of 6 residues.

Other programs: SIMPLATE, a program for testing the fit between a sequence and a template (21), and accompanying template files were a gift from Dr Donald Bashford. It was accompanied by a database of 226 globin sequences, which is the database referred to unless otherwise stated. Diagrams of molecular structures were constructed from coordinates in the Brookhaven Protein Data Bank (Brookhaven National Laboratory, New York), modified using molecular graphics program DTMM (22) as discussed in the text, and drawn with a Hewlett-Packard ColorPro plotter.

2. Amino Acid Sequencing

The first nine residues from the N-terminus were obtained by sequencing the polypeptide as previously described (13).

3. Cloning and Sequencing of *Artemia* Hemoglobin cDNA

Methods for cloning, mapping and sequencing of *Artemia* hemoglobin cDNA have been published previously (11, 12, 23). Each cDNA fragment sequence was translated in three reading frames by program NLDNA. As a control, the inferred complementary sequence was also translated. Each of the six translations was compared with a database of 329 globin sequences by program FASTP to search for homology. The database was a subset comprising all members of the globin family extracted from the entries in the PIR database (16). The translated coding sequences normally showed marked homology to globins in one reading frame. Any changes of reading frame or anomalies found in the analysis of the translation (below), if at all equivocal on the original gels, were resolved by further DNA sequencing.

4. Analysis of Inferred Amino Acid Sequence

The entire sequence was divided into putative heme-binding domains (Table 1). Each domain was tested to measure its conformity with the Bashford *et al.* (21) globin template II by program Simplate. This program subdivides a domain into 'patterns', which broadly correspond to the major helices of the globin fold, by testing alternative possible starting points for each pattern and allocating penalty scores to discrepant residues. Low-penalty alternatives were further eliminated by intra-domain constraints such as the requirement for patterns not to overlap, by observations of consistency between the nine *Artemia* domains, and by further considerations discussed in the text.

5. Molecular Reconstruction

Three structural coordinate files from the Brookhaven database were selected on the basis of the FASTP results as structural templates for testing the compliance of different parts of the *Artemia* sequence with the prerequisites of the globin fold. These structures were: *Chironomus thummi thummi* deoxy-erythrocruorin (Hb) III (24), file 1ECD; sperm whale deoxy-myoglobin (25), file 1MBD; and *Petromyzon marinus* cyano-methemoglobin V (26), file 2LHB. The acceptability of certain residue placements necessary in fitting the *Artemia* domain sequence to the *Chironomus* Hb III or myoglobin mainchain was investigated by substituting the sidechain and then checking bond lengths, angles, and surrounding Van der Waals clearances, and by considerations discussed in the text. The resulting structures were output to the Hewlett-Packard plotter.

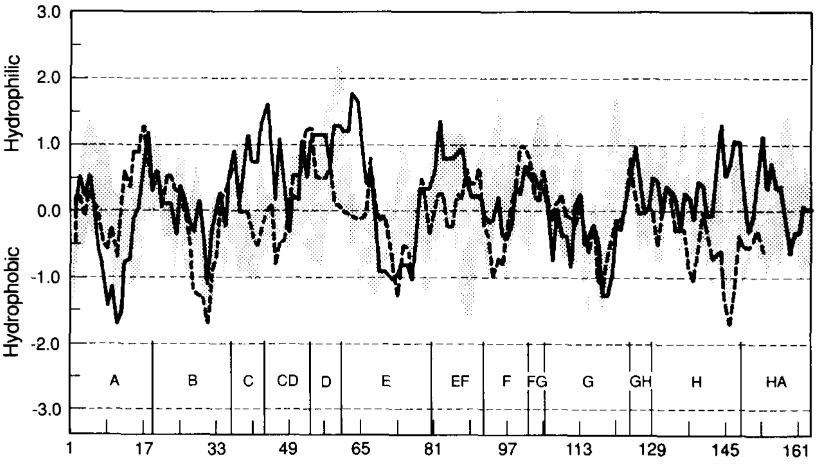

Fig. 1. Composite hydropathicity profile of *Artemia* globin domains (gray envelope). Heavy traces are those of *Chironomus* globin III (continuous line) and whale myoglobin (broken line). All sequences are in alignment as in Table 1. Hydrophobicity scale: Hopp and Woods (20); window length: 6.