The Hemoglobin of *Urechis caupo*

THE cDNA- DERIVED AMINO ACID SEQUENCE*

James R. Garey and Austen F. Riggs

From the Department of Zoology, University of Texas, Austin, Texas 78712

The nucleotide sequence of a cDNA transcript containing part of the 5' noncoding region, the entire coding region, and the entire 3' noncoding region has been determined. The protein sequence predicted from the coding region matches almost exactly the amino-terminal sequence and the sequence of several peptides from *Urechis caupo* F-I globin. Only 11-20% of the amino acid positions are identical with those of other known globins.

*Urechis caupo* belongs to the Echiura, a minor protostome phylum closely allied to the annelids. The animals live in U-shaped burrows in intertidal mudflats along the Pacific coast. Their respiratory physiology has been thoroughly studied (4, 5), as have the hemoglobins found within the coelomic red cells (6-9). The hemoglobin is tetrameric and binds oxygen noncooperatively and almost independently of pH. The binding is not significantly affected by mono- or divalent cations, organic phosphates, or changes in concentration. Although the tetramers are essentially homomeric, the hemoglobin is electrophoretically heterogeneous due to the very small differences between the chains, which may be either a small number of amino acid changes or post-translational modifications (9). The hemoglobin appears to function in storage, rather than transport of oxygen. We have previously characterized the hemoglobin from this animal and have shown that the nucleated red cells contain globin messenger RNA (2, 9).

We report here the cloning and sequencing of the cDNA corresponding to the globin and the amino acid sequence of the F-I globin of *U. caupo* derived from the cDNA.

EXPERIMENTAL PROCEDURES

Oligonucleotide Synthesis and Extension—The amino acid sequence, Gin-Asp-His-Trp-Phe, corresponding to residues 11-15 of *U. caupo* F-I globin (9) was chosen for the synthesis (Bachem Fine Chemicals, Torrance, CA) of a mixed oligonucleotide probe,  

$$5' - A C C A _ { G G G } ^ { G G G } ,$$

complementary to the F-I globin message. The probe was extended by using it as a primer for double-stranded cDNA synthesis (10) from poly(A)* RNA isolated from red cells of *U. caupo* (see below) and subsequently cloned directly into MI3mp8 with EcoRI linkers. The clone of the extended probe, UCG-1, was identified by the Sanger dideoxy sequencing method (11); the sequence corresponded exactly to the first 14 amino-terminal residues of F-I globin.

RNA Preparation and Analysis—Specimens of *U. caupo* were collected at Bodega Bay, CA, and maintained in a marine aquarium. RNA, isolated from 5 ml of packed red cells as described by Cox (12) was obtained twice over a column of oligo(dT)-cellulose in order to isolate the poly(A)* RNA. Northern analysis (13) was carried out on 2 µg of total and 2 µg of poly(A)* RNA following electrophoresis on 1% agarose gels in 50% formamide, 10% formic acid, and 0.1 M sodium phosphate buffer, pH 6.8.

Preparation and Analysis of *cDNA*—A cDNA library was constructed (10) in Agt10 and screened by hybridizing nick-translated UCG-1 to nitrocellulose filters containing replicas of Agt10 plaques (15). Insert DNA from a positive plaque, UCG-2, was subcloned into MI3mp8 and sequenced from both ends by the Sanger dideoxy method (11).

Preparation Analysis and Sequencing—*U. caupo* F-I globin was prepared by limited proteolysis with CNBr (15) and chromatographed on a column of Sepacryl S-200 superfine (2.5 x 165 cm) in 6 M guanidine HCl, 0.2 M sodium acetate, pH 6.0 (Fig. 2a). Peak F was rechromatographed by high-performance liquid chromatography (Fig. 2b) as described (16). F-I globin (100 mg) was also digested with BNPS-skatole* reagent (15) and chromatographed on Sepacryl S-200 (Fig. 3a) as described above. Peaks B and E were rechromatographed with high performance liquid chromatography (Fig. 3, b and c). Ammonium acid compositions were determined as described (9) on a Beckman 121MB amino acid analyzer at the Protein Sequencing Center at the University of Texas. Sequence determinations with a Beckman model 890 Sequencer equipped with a cold trap made use of the 0.1 M Quadrol program (No. 121178) with the addition of polybrene (17, 18). The phenylthiohydantoin derivatives of the amino acids were identified by high performance liquid chromatography as described (16).

RESULTS AND DISCUSSION

The cDNA library contained about 1200 recombinants, of which seven hybridized to UCG-1. The largest of these, UCG-2, was completely sequenced (Fig. 4). When sequenced, the cDNA region of this clone is only 11 bp long. This is probably shorter than that of the original mRNA because of clipping of the 5' loop by S1 nuclease during cDNA preparation. A 423-bp open reading frame that codes for *U. caupo* F-I globin follows the start.

*The abbreviations used are: BNPS-skatole, the product formed by reaction of N-bromosuccinimide with 2-(2-nitrophenylsulfenyl)-3-methyl indole, used as name for peptides cleaved with this reagent; bp, base pairs; SDS, sodium dodecyl sulfate.

*This work was supported by Grants PCM8202760 and DMB-8502887 from the National Science Foundation, Welch Foundation Grant F-213, and National Institutes of Health Grant GM28410. Preliminary accounts of some of this work have been presented (1-3). The work is based in part on a Ph.D. dissertation by J. R. G., University of Texas, Austin, 1985. 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.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBankTM/EMBL Data Bank with accession number(s) J02624.

† Recipient of a predoctoral fellowship from the Robert A. Welch Foundation. Present address: Dept. of Hematology, University of Utah School of Medicine, Salt Lake City, UT 84132.

‡ To whom reprint requests should be addressed.
The amino-terminal sequence (34 residues, Ref. 9) and the partial sequence of a large CNBr fragment and two BNPS-skatole fragments (Fig. 4, Table I) were compared with the amino acid sequence predicted from UCG-2. CNBr fragment 4B corresponds to residues 98–119, BNPS-skatole fragment 3B corresponds to residues 58–91, and BNPS-skatole fragment 3E corresponds to residues 130–141. Amino acid residue 4 is threonine in UCG-2, although it was found to be alanine in both the amino-terminal sequence and the extended probe, UCG-1. Because this is a neutral substitution, it may not account for an electrophoretic difference unless the substitution caused a change in interaction between other parts of the globin chain. However, differences in other undetected positions could account for at least some of the heterogeneity reported earlier in *U. caupo* F-I hemoglobin (9).

A long (295 bp) 3' noncoding region follows the termination codon and ends with nine nucleotides of the poly A tail. Mammalian and frog α and β globin genes have 88–130 nucleotides in the 3' noncoding region, legHb globin genes have up to 174 bp, whereas the genes of seal and human...
myoglobin have 548 and 531 bases, respectively (20, 21). A possible polyadenylation site, AATAAA, found in the cDNA of vertebrate messages (22) starts at base 697. The sequence GGTTTTA starts 11 bases upstream from the polyadenylation site. This sequence is also found upstream from the initiation site. This sequence must await the sequences of other invertebrate genes. No other correspondence could suggest sufficiently closely to suggest that their sequences of the globins from man, lamprey, Molpadiana, Chironomus, Glyceria, Lumbricus, Anadara, Aplysia, and soybean (23–26). The number of identical positions ranges from 11 (Lumbricus) to 20% (Glyceria) with a mean of 14%, although 51% of the amino acid sequence of Urechis globin corresponds to positions in at least one of the globins compared. These low levels of identity preclude conclusions as to the phylogenetic position of the Echiura.

Fig. 5 shows that, with few exceptions, a close correspondence exists between the hydrophobicity pattern of the Urechis F-I globin chain and that of the human β chain. We have used for the hydrophobicity the mean area of the amino acid side chain buried upon transfer from the standard state to the folded protein as described by Rose et al. (27). The detailed, residue-by-residue correspondence strongly suggests that the overall conformation of the two chains is very similar, and that we are justified in the following discussion in assuming that the Urechis hemoglobin has the same helical segments that are present in the human β chain. It is impossible to recognize the close correspondence if one uses the running average employed by Kyte and Doolittle (28) which blurs these details.

The Urechis globin A helix appears to start with the highly conserved threonine in position 3. All the charged residues of the A helix appear to be external and available to solvent. The highly conserved tryptophan occurs at position 14 in the Urechis chain. This residue (215) acts as an AE helix spacer in the human chain; we presume a similar function here. Three residues, -Ile-Lys-Gly-, join the B helix appears to be too far away to form such a linkage. Although position B12 is internal in the human chain, the corresponding lysyl residue of the Urechis chain might be external if the intersubunit contacts are different than in human hemoglobin.

The short Urechis C helix appears in Fig. 5 to have a very different hydrophobicity pattern from that of the human chain, but this depends largely on the glycine at C3. The tabulation of Rose et al. (27) which we are using gives Gly as the least buried, but then it has the least to bury so this discrepancy may be misleading. The glycine could easily occupy a hydrophobic pocket without serious energetic penalty. For this reason we have ignored the glycines in parts of Fig. 5. The Urechis chain has the highly conserved Phe at position CD1; in human hemoglobin this residue is packed close to the heme. The patterns for the E, F, FG, and G segments shown in Fig. 5 correspond sufficiently closely to suggest that their lengths are probably very similar in the two chains. The EF interhelical segment appears to have 9 residues in the Urechis chain—3 residues fewer than in the human β chain. Glutamine replaces the normal distal histidine in the Urechis chain. A striking feature of the F helix in the Urechis chain is the presence of a proline at F6, just two residues away from the proximal histidine. This proline seems certain to cause a shift in the F helix, perhaps similar to that found with the proline.

### Table I

| Edman cycle | BNPS-skatole fragment 3B | BNPS-skatole fragment 3E | CNBr peptide 4B |
|-------------|--------------------------|--------------------------|------------------|
| Amino acid  | Yield (nmol)             | Amino acid             | Yield (nmol)     |
| 1 Asn       | 4.6 Gly                  | 35.4 Gly               | 6.4             |
| 2 Pro       | 2.7 Asp                  | 34.0 Ile               | 10.6            |
| 3 Ala       | 6.5 Ala                  | 28.2 Thr               | 0.8             |
| 4 Tyr       | 4.2 Ala                  | 45.0 Pro               | 0.6             |
| 5 Lys       | 1.6 Gly                  | 22.4 Lys               | 3.0             |
| 6 Ala       | 2.9 Val                  | 51.0 His               | 1.2             |
| 7 Gln       | 1.8 Leu                  | 24.0 Phe               | 2.0             |
| 8 Thr       | 3.3 Val                  | 9.2 Gly               | 0.3             |
| 9 Leu       | 2.6 Leu                  | 2.6 Gln                | 0.7             |
| 10 Thr      | 0.5 Ala                  | 8.6 Leu                | 4.5             |
| 11 Val      | 2.6 Met                  | 13.0 Leu               | 2.7             |
| 12 Ile      | 2.6 Lys                  | 1.3 Lys                | 2.4             |
| 13 Asn      | 1.6 Leu                  | 1.9                   |
| 14 Tyr      | 1.1 Val                  | 2.3                   |
| 15 Leu      | 2.5 Gly                  | 1.6                   |
| 16 Asp      | 1.9 Gly                  | 1.4                   |
| 17 Lys      | 0.5 Val                  | 2.0                   |
| 18 Val      | 2.0 Phe                  | 0.7                   |
| 19 Val      | 1.5 Gln                  | 0.2                   |
| 20 Asp      | 3.0 Gln                  | 0.8                   |
| 21 Ala      | 1.3 Gln                  | 1.4                   |
| 22 Leu      | 1.5 Phe                  | 1.3                   |
| 23 Gly      | 0.6                      |
| 24 Gly      | 0.9                      |
| 25 Asn      | 0.3                      |
| 26 Ala      | 0.5                      |
| 27 Gly      | 0.6                      |
| 28 Ala      | 0.5                      |
| 29 Leu      | 0.4                      |
| 30 Met      | 0.2                      |
| 31 Lys      | 0.2                      |
| 32 Ala      | 0.4                      |
| 33 Lys      | 0.2                      |
| 34 Val      | 0.3                      |

Fig. 5. Progressive hydrophobicity of F-I globin (solid line) compared with human β globin (dashed line). Letters designate the helical regions of human hemoglobin. The glycine residues of the FG and G segments were ignored for greater clarity.

![Hydrophobicity graph](image-url)
in the middle of the G helix of Glycera hemoglobin where it causes a 25° bend (29). The alignment of the G and H helices shown in Fig. 5 suggests that the GH corner and the H helix together are shorter by 3 residues. If the H helix is the same length as that in the human β chain, then the GH segment would be only 2 residues in length. The 180° turn generated by the GH corner would be energetically easier if the GH be ultimately necessary to explain these observations. However, we suggest that the F helix were shorter.

Comparison of the 17 residues which form the αβ contacts in human hemoglobin with those in Urechis indicates that only 2 are identical. A similar comparison for the αβ2 interface shows that only 2 of 13 residues are identical. However, 5 of 11 residues known to form the heme contacts in human β globin (23) are identical with those in Urechis globin. Thus, although the heme pocket residues appear to be conserved, the data indicate that the intersubunit contacts are not.

Thus the tetrameric Urechis hemoglobin appears to lack the αβ contacts which, in human hemoglobin, shift during oxygenation. Analysis of the CO and O2 association-dissociation kinetics at 20° (30) indicates that Urechis hemoglobin can be best classified as a low affinity “T-state” hemoglobin (31). Furthermore, O2 and CO recombination kinetics studied in the nanosecond time regime show very little geminate recombination, a characteristic of T-state hemoglobin. Recent extended x-ray absorption fine structure measurements (32) on the HbCO form are consistent with a very large out-of-plane position for the iron atom when CO-ligated, which is also consistent with assignment of Urechis hemoglobin to the T-state. X-ray diffraction determination of the structure will be ultimately necessary to explain these observations. However, we suggest that the F6 proline close to the proximal histidine may in part be responsible for the unusual functional properties.

The subunits of the tetrameric hemoglobin of the clam, Scapharca inequivalvis, are assembled very differently from those which form the tetramers of vertebrate hemoglobins (30). The major difference is that the E and F helices form major intersubunit contacts in the clam hemoglobin, whereas the G and H helices form the corresponding contacts in tetramers of vertebrate hemoglobin. If the assembly of the tetramers of Urechis hemoglobin were similar to that in Scapharca we might expect to see some indication of this in the hydrophobicity plot. However, the hydrophobicity profiles for the putative E and F helices in Urechis and Anadara trapezii (related to S. inequivalvis) are both very similar to those in the human β chain so that no conclusion can be drawn. Another feature of Scapharca globin and presumably that of Anadara is the additional pre-A helix (33). This helix is absent in all other known globins including those of Urechis and man. The “pre-A” sequence of lamprey globin exists as an extended chain rather than a helix (34).

Codon usage does not appear to differ significantly from that of other known globin genes (human α and β, mouse β, and chicken β from Refs. 35; seal myoglobin, legHb, and Chironomus from Refs. 20, 36, and 37, respectively) except that Urechis globin appears to be the only one to utilize the CGA codon for arginine, the UCG codon for serine, and the UUA codon for leucine. Vertebrate globins use mostly CAG to code for glutamine, whereas the invertebrates appear to use only CAA.

We have preliminary evidence (3) that multiple globin genes occur in Urechis. A Southern analysis of genomic DNA cut with EcoRI and probed with nick-translated UCG-2 DNA showed the presence of four bands of 4, 9, 15, and 17 kilobases in size. If the multiple genes are all expressed, then they are probably very similar in size because Northern analysis (Fig. 6) of Urechis red cell RNA using the same probe shows only a single band. The previously discussed difference of one nucleotide between UCG-1 and UCG-2 indicates that at least two of the multiple genes are almost identical. This could explain the electrophoretic heterogeneity seen in Urechis hemoglobin (9). This is a situation similar to that seen in the globin genes of Chironomus (37), where gene A codes for threonine at position 57 and gene B codes for isoleucine.

Acknowledgments—We thank Karen Haschke for expert laboratory assistance and Dr. David Wolstenholme for support of J. R. G. during part of this project. We thank Dr. Larry J. Parkhurst for helpful comments.

REFERENCES
1. Garey, J. R., and Riggs, A. F. (1983) Fed. Proc. 42, 1971
2. Garey, J. R., Brodeur, R. D., and Riggs, A. F. (1983) Life Chem. Rep. Suppl. 1, 387–391
3. Garey, J. R., and Riggs, A. F. (1986) in Invertebrate Oxygen Carriers (Linzen, B., ed), Springer-Verlag, New York, in press
4. Hall, V. E. (1931) Biol. Bull. 61, 406–416
5. Pritchard, A., and White, F. N. (1931) Physiol. Zool. 54, 44–54
6. Redfield, A. C., and Florkin, M. (1931) Biol. Bull. 61, 185–210
7. Hall, R. E., Terverwilliger, R. C., and Terverwilliger, N. B. (1981) Comp. Biochem. Physiol. 70B, 363–367
8. Mangum, C. P., Terverwilliger, R. C., Terverwilliger, N. B., and Hall, R. E. (1983) Comp. Biochem. Physiol. 76A, 253–257
9. Garey, J. R., and Riggs, A. F. (1984) Arch. Biochem. Biophys. 228, 320–331
10. Huynh, T. V., Young, R. A., and Davis, R. W. (1984) in DNA Cloning: A Practical Approach (Glover, D., ed), Vol. I, pp. 49–78, IRL Press, Oxford
11. Sanger, F., Nicklen, S., and Coulson, A. R. (1977) Proc. Natl. Acad. Sci. U.S.A. 74, 5463–5467
12. Cox, R. H. (1966) Methods Enzymol. 12B, 120–129
13. Ménatin, T., Fräsch, E. F., and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual, pp. 109–112, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
14. Rigby, P. W. J., Dieckmann, M., Rhodes, C., and Berg, P. (1977) J. Mol. Biol. 113, 257–251
15. Allen, G. (1961) in Sequencing of Proteins and Peptides (Work, T. S., and Burdon, R. H., eds), p. 68, Elsevier/North-Holland, New York
25. of the region were sequenced completely in both directions; about used, 14,3029-3035
22. The coding region and about noncoding region were sequenced in only one direction.

FIG. 1. Map of clone UCG-2 showing sequencing strategy used. The coding region and about 200 bases of the 3' noncoding region were sequenced completely in both directions; about 90 bases of the 3' noncoding region were sequenced in only one direction.

APPENDIX

FIG. 4. The complete sequence of UCG-2. Noncoding regions are lower case. The numbers indicate the amino acid positions. The polyadenylation site is underlined, and the sequence GGTTTTA confirmed by protein sequencing are marked with arrows. The symbol * marks the position different in UCG-2 and the amino acid sequence as described in the text.

5' cagtttga ATG

1. Gly Leu Thr Thr Ala Gln Ile Lys Ile Gin Gin Asp His Trp Phe
   GGT CTT ACA ACA GCT CAA ATC AAA GCC ATC GAT CAT TGG TTT
   1

2. Leu Asn Ile Lys Gly Cty Gya Leu Gin Ala Ala Asp Ser Ile Phe
   CTT AAT CAC AAT AAG GAA TGT TGT CAG GCG GAA GGT TCC ATC TTT
   1

3. Thr Tyr Leu Thr Ala Tyr Pro Gly Asp Leu Ala Phe His
   TTC AAG TAC TCT ACT GCT TCT GGG GAT TTA GCG TTT CAC
   1

4. Lys Phe Ser Ser Val Pro Leu Tyr Gly Leu Arg Ser Asn Pro Ala
   AAG TTT TGG TCC TTC TCT CTC TCT GGC CTG CTA GGC TGG AAA CAA GCA
   1

5. Tyr Ala Gin Thr Leu Thr Val Ile Asn Tyr Leu Asp Lys Val
   TAT AAA GCC CAG ACT CTA ACA GTC ATG ATA AAC TAC TGG GAT AAA CGT
   1

6. Val Asp Ala Leu Gly Gyl Asn Ala Asp Met Lys Ala Lys
   ATG GAC GCT CTG GCC GGC CAA AAT GCA GAT GGT ATT AAG GCG AAG
   1

7. Val Pro Ser His Asp Ala Met Gly Ile Pro Leu His Phe Gly
   GTC CCA AGT GAC GGC ATG GGC AAG ACC CAG CAT TGG TGG
   1

8. Gin Leu Leu Leu Val Gly Cty Gla Phe Gin Gin Gin Ser
   CCA CTT TGG AAG TGG GAT GGA GAT TGC CAA GAA GAG TCG
   1

9. Ala Asp Pro Thr Thr Val Ala Ala Trp Gly Asp Ala Gly
   GCA GAT GGC AGC AAT GTC GCT GGC TGG GGC GAC GCT GCT GTC
   1

10. Leu Val Ala Ala Met Lys
   CTG GTC GCC GGC ATG AAG TAA accganaagcgctgctactcaccacagca
   1

11. gagctcgctgttattgtagctggcagccagagccagctggcttttcggccatgtgctgttattagc
   actactathtagctgctacatcctcggcagctgttattgtagctggcagccagagccagctggcttttcggccatgtgctgttattagc
   1

12. aactcggagctcctcgtggactctcctcggcagctgttattgtagctggcagccagagccagctggcttttcggccatgtgctgttattagc
   1

13. aactcggagctcctcgtggactctcctcggcagctgttattgtagctggcagccagagccagctggcttttcggccatgtgctgttattagc
   1

14. aactcggagctcctcgtggactctcctcggcagctgttattgtagctggcagccagagccagctggcttttcggccatgtgctgttattagc
   1

15. aactcggagctcctcgtggactctcctcggcagctgttattgtagctggcagccagagccagctggcttttcggccatgtgctgttattagc
   1

16. aactcggagctcctcgtggactctcctcggcagctgttattgtagctggcagccagagccagctggcttttcggccatgtgctgttattagc
   1

17. aactcggagctcctcgtggactctcctcggcagctgttattgtagctggcagccagagccagctggcttttcggccatgtgctgttattagc
   1

18. aactcggagctcctcgtggactctcctcggcagctgttattgtagctggcagccagagccagctggcttttcggccatgtgctgttattagc
   1

19. aactcggagctcctcgtggactctcctcggcagctgttattgtagctggcagccagagccagctggcttttcggccatgtgctgttattagc
   1

20. aactcggagctcctcgtggactctcctcggcagctgttattgtagctggcagccagagccagctggcttttcggccatgtgctgttattagc
   1

21. aactcggagctcctcgtggactctcctcggcagctgttattgtagctggcagccagagccagctggcttttcggccatgtgctgttattagc
   1

22. aactcggagctcctcgtggactctcctcggcagctgttattgtagctggcagccagagccagctggcttttcggccatgtgctgttattagc
   1

23. aactcggagctcctcgtggactctcctcggcagctgttattgtagctggcagccagagccagctggcttttcggccatgtgctgttattagc
   1

24. aactcggagctcctcgtggactctcctcggcagctgttattgtagctggcagccagagccagctggcttttcggccatgtgctgttattagc
   1

25. aactcggagctcctcgtggactctcctcggcagctgttattgtagctggcagccagagccagctggcttttcggccatgtgctgttattagc
   1

26. Fisher, W. K., Gilbert, A. T., and Thompson, E. O. P. (1984) J. Biol. Chem. 259, 4739-4749
27. Brauer, A. W., Margolies, M. N., and Haber, E. (1975) Biochemistry 14, 3029-3035
28. Klapfer, D. G., Wilde, E. E., and Capra, J. D. (1978) Anal. Biochem. 85, 126-131
29. Bonaventura, J., and Kitto, G. (1973) in Comparative Physiology (Bolis, L., Schmidt-Nielsen, K., and Madrell, S. H. P., eds) pp. 493-507, Elsevier/North-Holland, Amsterdam
30. Blanchetot, A., Wilson, V., Wood, D., and Jeffreys, A. J. (1983) Nature 301, 732-734
31. Weller, P., Jeffeys, A. J., Wilson, V., and Blanchetot, A. (1984) EMBO J. 3, 439-446
32. Gil, A., and Proudfoot, N. J. (1984) Nature 312, 473-474
33. Dickerson, R. E., and Geis, I. (1983) Hemoglobin: Structure, Function, and Pathology, Benjamin-Cummings, Menlo Park, CA
34. Garlick, R. L., and Riggs, A. F. (1982) J. Biol. Chem. 257, 9005-9015
35. Mauri, F. C. (1985) Ph.D. dissertation, University of Texas, Austin, TX