Proc. Natl. Acad. Sci. USA
Vol. 74, No. 11, pp. 5145–5149, November 1977
Medical Sciences

The nucleotide sequences of the untranslated 5′ regions of human α- and β-globin mRNAs

(cDNA/restriction endonuclease)

JUDY C. CHANG*, GARY F. TEMPLE*, RAYMOND POON*†, KURT H. NEUMANN*, AND YUET WAI KAN*†‡§

* Hematology Service and † Howard Hughes Medical Institute Laboratory, San Francisco General Hospital, San Francisco, California 94110; and Departments of * Medicine and ‡ Laboratory Medicine, University of California, San Francisco, California 94143

Communicated by Rudi Schmid, August 5, 1977

ABSTRACT **The complete sequences of the untranslated 5′ regions of human α- and β-globin mRNAs were determined by sequence analysis of full-length cDNAs. The single-stranded cDNAs were digested with the restriction endonuclease *Hae* III, and the two 3′-terminal fragments of 75 and 132 nucleotides, complementary to the 5′ termini of the α- and β-globin mRNAs, respectively, were isolated and sequenced. Including the initiation codon AUG, the untranslated 5′ regions of human α- and β-globin mRNAs contain 41 and 54 nucleotides, respectively, and exhibit striking homologies with the corresponding sequences in the rabbit. Human α- and β-globin mRNAs have five bases in the region of the initiation codon that may form base pairs with the 3′ terminus of 18S rRNA. Stable secondary structures with hairpin loops can be constructed in the untranslated 5′ regions.**

The nucleotide sequences of globin mRNAs have been the focus of intensive investigation. Partial sequences of rabbit and human globin mRNAs were initially obtained by use of RNA sequencing techniques (1, 2). The development of rapid DNA sequencing methods (3, 4) has made it possible to derive mRNA sequences by analysis of double-stranded cDNAs digested with restriction endonucleases (5–8) or of single-stranded cDNAs primed with oligodeoxyribonucleotides complementary to specific regions of the mRNAs (9–12). To date, the entire rabbit β-globin mRNA and part of the rabbit α-globin mRNA sequence have been determined (5–12). In the human, the β-globin mRNA from the initiation codon AUG to the 3′-poly(A) end has been sequenced, and portions of the translated region of the α-globin mRNA and the complete untranslated 3′ end are known (10, 13–16).

The primary structure of the untranslated 5′ region of mRNA is of great interest because it may play a role in ribosomal binding and the initiation of protein synthesis. Lockard and RajBhandary (17) determined the sequence of the 5′ termini of the separated rabbit α- and β-globin mRNAs by direct RNA sequence analysis. By analyzing the cDNAs synthesized with a primer complementary to the region of the initiation codon, Baralle (11, 12) sequenced the complete untranslated 5′ regions of both rabbit α- and β-globin mRNAs.

We report here the complete sequences of the untranslated 5′ region of the human α- and β-globin mRNAs. We utilized the ability of the restriction enzyme *Hae* III to cleave single-stranded DNA at G-G-C-C sequences (18, 19). Full-length single-stranded α- and β-globin cDNAs were digested with *Hae* III and the two 3′-terminal fragments corresponding to the 5′ termini of the α- and β-mRNAs were identified and isolated. The sequences of these two fragments were then determined according to the method of Maxam and Gilbert (4) by labeling either at their 3′ ends with ribonucleoside [^{32}P]diphosphate with terminal deoxyribonucleotidyl transferase, or at their 5′ ends with ^{32}P by using polynucleotide kinase.

MATERIALS AND METHODS

Source of Human Globin mRNA. Blood was obtained from a patient with sickle cell disease, one with hemoglobin H disease, and one with homozygous β thalassemia. RNA was prepared from these samples and the poly(A)-rich RNA was isolated by two passages over oligo(dT)-cellulose as described (20).

Preparation of Globin cDNAs. Single-stranded globin cDNAs were prepared according to the method of Verma *et al.* (21). Nonradioactive dATP, dGTP, and TTP at 400 μM each and radioactive dCTP at 100 μM were used to maximize the synthesis of full-length cDNAs as shown by Efstratiadis *et al.* (22). The radioactive precursors used for the synthesis of internally labeled cDNA in the different experiments were [^{3}H]dCTP (New England Nuclear, 23 Ci/mmol) diluted to a specific activity of 2.3 Ci/mmol and [α-^{32}P]dCTP (New England Nuclear, 110–115 Ci/mmol) diluted to a specific activity of 25 or 0.1 Ci/mmol to yield high- or low-specific activity ^{32}P products, respectively.

Digestion of cDNA with Restriction Enzyme. cDNA was digested with the restriction endonuclease *Hae* III (New England Biolabs), 200 units/μg of DNA, in 6 mM Tris·HCl, pH 7.4/6 mM $MgCl_2$/6 mM NaCl/6 mM 2-mercaptoethanol for 16 hr at 37°. To analyze the restriction fragments, digested high-specific activity [^{32}P]cDNA was separated by electrophoresis on a 5% polyacrylamide (acrylamide/bisacrylamide, 19:1) slab gel in a buffer containing 90 mM Tris·borate (pH 8.3) and 4 mM EDTA.

Identification of the 3′-Terminal *Hae* III Fragments. Ribonucleoside [^{32}P]monophosphate residues were added to the 3′ ends of the internally ^{3}H-labeled cDNA by a method modified from Higuchi *et al.* (23). The 100-μl reaction mixture contained 0.2M K cacodylate (pH 7.1), 1 mM $CoCl_2$, 2 mM 2-mercaptoethanol, 1 μg of [^{3}H]cDNA, 500 pmol of [α-^{32}P]GTP or [α-^{32}P]UTP (New England Nuclear, 100–150 Ci/mmol), and 25 units of terminal deoxyribonucleotidyl transferase (obtained from Winston Salser). After incubation for 1 hr at 37°, the reaction was terminated by ethanol precipitation. To remove all but one of the ribonucleotide residues added to the 3′ ends, the cDNA was incubated in 0.3 M NaOH at 37° for 16 hr, neutralized, and recovered by ethanol precipitation. The 3′-^{32}P-labeled cDNA was digested with *Hae* III and analyzed on a 5% polyacrylamide gel.

5′-^{32}P-Labeling of the 3′ Terminal *Hae* III Fragments. cDNA, labeled internally with low-specific activity [^{32}P]dCMP,

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. §1734 solely to indicate this fact.

§ To whom reprint requests should be addressed at: San Francisco General Hospital, San Francisco, CA 94110.

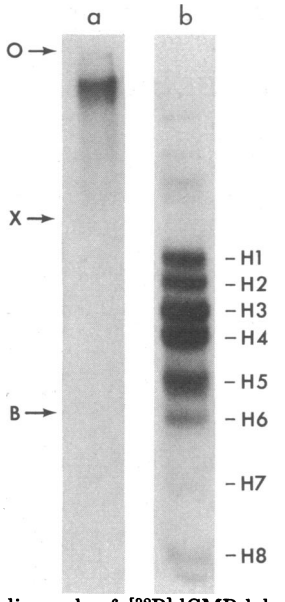

FIG. 1. Autoradiograph of [^{32}P]dCMP-labeled human globin cDNAs on 5% polyacrylamide gel. Lanes: a, single-stranded cDNA; b, *Hae* III digestion products of cDNA. Arrows indicate positions of origin (O), xylene cyanol blue (X), and bromphenol blue (B) markers. The faint bands above H1 were not found consistently and represented products of incomplete digestion.

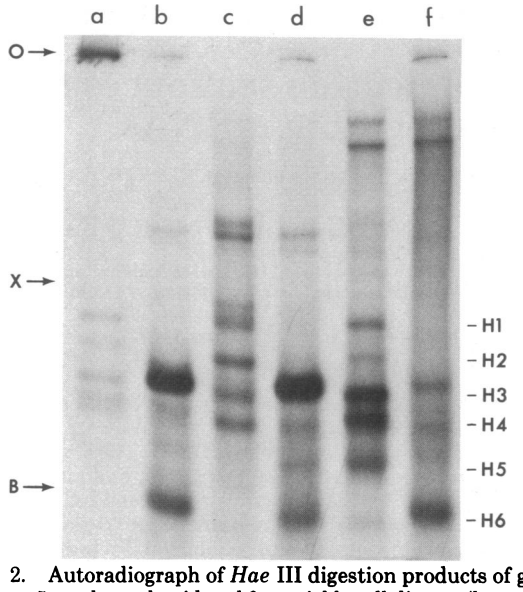

FIG. 2. Autoradiograph of *Hae* III digestion products of globin cDNAs on 5% polyacrylamide gel from sickle cell disease (lanes a and b), hemoglobin H disease (lanes c and d), and β-thalassemia (lanes e and f). Lanes a, c, and e, internally ^{32}P-labeled cDNAs; lanes b, d, and f, [^{3}H]cDNAs labeled at 3′ ends with [3′, 5′-^{32}P]UDP.

was digested with *Hae* III. The 3′-terminal α- and β-globin cDNA fragments were isolated on a 5% polyacrylamide gel and further purified on a 12% gel. These fragments were then dephosphorylated and labeled at their 5′ ends with polynucleotide kinase (New England Biolabs) and [γ-^{32}P]ATP (New England Nuclear, >5000 Ci/mmol) by the method of Weiss *et al.* (24).

Determination of the Size of the 3′-Terminal cDNA Fragments. The ^{32}P-labeled 3′-terminal cDNA fragments were separated by electrophoresis on a 5% polyacrylamide gel in 98% formamide and compared to ^{32}P-labeled *Hin*dII- and *Hin*dIII-digested bacteriophage λ DNA fragments of known sizes, as described by Maniatis *et al.* (25).

Sequence Analysis of the 3′-Terminal Fragments. The purified 3′-terminal fragments of α- and β-globin cDNA prepared from sickle cell mRNA and ^{32}P-labeled at either the 3′ or the 5′ end were repurified by electrophoresis on 20% polyacrylamide gel in 7 M urea/90 mM Tris·borate, pH 8.3/4 mM EDTA. The fragments in the slowest migrating bands were eluted from the gel and sequenced according to the method of Maxam and Gilbert (4).

RESULTS

Hae III digestion of single-stranded unseparated human α- and β-globin cDNAs yielded at least eight bands on 5% polyacrylamide gel electrophoresis (Fig. 1). The 3′-terminal cDNA fragments were identified by adding a ^{32}P-labeled ribonucleotide to the [^{3}H]cDNA before digestion with *Hae* III (Fig. 2). Two prominent bands corresponding to H3 and H6 were observed. To determine whether these bands were of α or β origin, cDNAs from a patient with hemoglobin H disease (β > α) and homozygous β-thalassemia (α > β) were similarly prepared and digested. H3 was prominently labeled in the former and H6 in the latter, indicating that they were derived from β- and α-globin cDNA, respectively.

The sizes of these 3′-terminal cDNA fragments were ascertained by comparison to bacteriophage λ DNA restriction fragments of known size on 5% polyacrylamide gel in 98% formamide. There were approximately 75 nucleotides in the α-globin 3′-cDNA fragment (H6) and 135 nucleotides in the β fragment (H3) (Fig. 3).

Figs. 4 and 5 show the results of DNA sequence analysis of the 3′-terminal fragments of the α- and β-globin cDNAs, labeled with ^{32}P at their 3′ ends. The DNA sequences obtained represented the untranslated 5′ region of the α- and β-globin mRNAs. The corresponding mRNA sequences of the untranslated regions are shown in Fig. 6, compared to their counterparts in the rabbit globin sequences as determined by Baralle (11, 12). Including the initiation codon AUG, the untranslated 5′ region of the α-globin mRNA contained 41 nucleotides and that of β-globin mRNA contained 54 nucleotides.

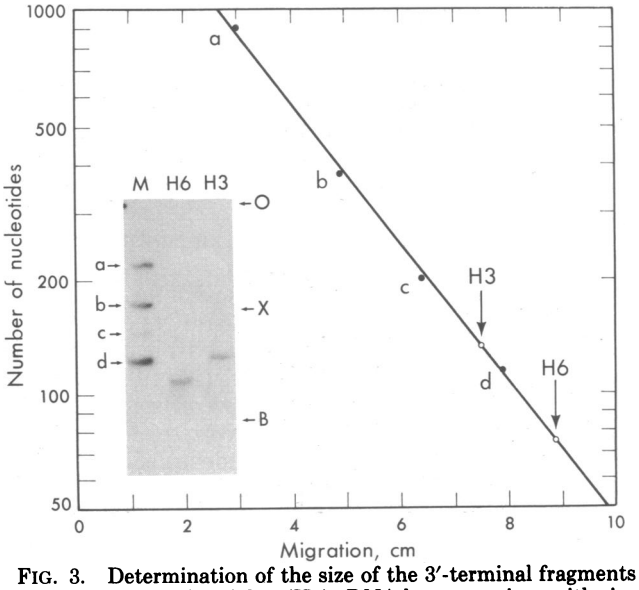

FIG. 3. Determination of the size of the 3′-terminal fragments of α-globin (H6) and β-globin (H3) cDNA by comparison with size markers (M) from λ DNA restriction fragments. (*Inset*) Electrophoresis of these fragments.

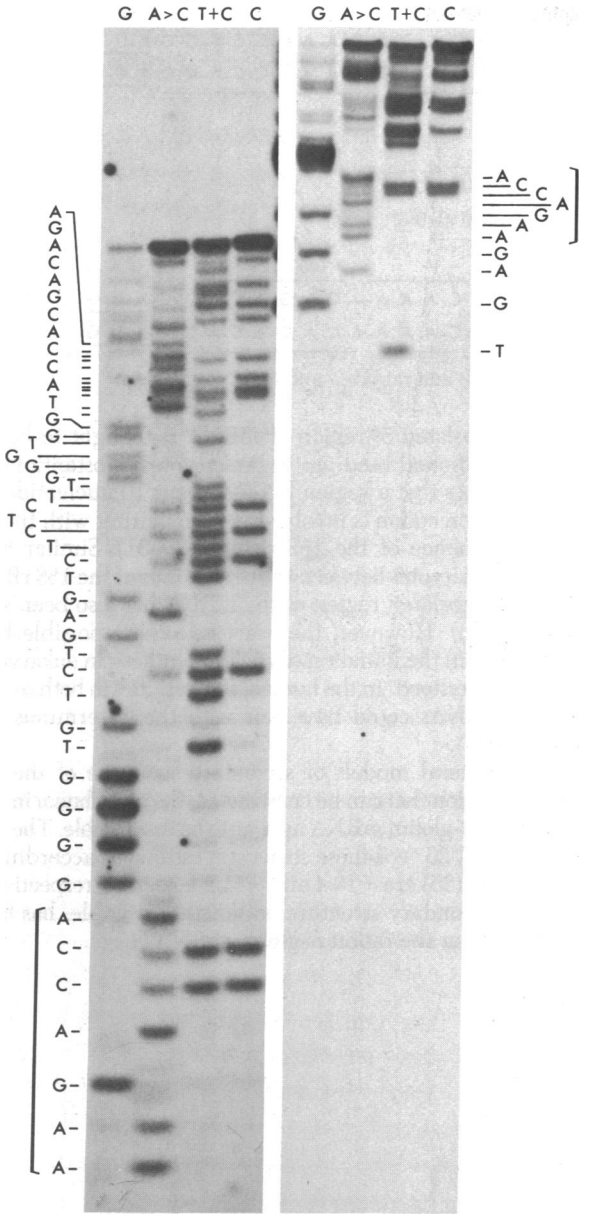

FIG. 4. Autoradiograph of the sequencing gel of fragment H6 from α-globin cDNA labeled at the 3′ end with [3′,5′-^{32}P]GDP. Four reactions were used such that cleavage occurred: only at guanosine (G), more frequently at adenosine than at cytidine (A > C), at both thymidine and cytidine (T + C), and only at cytidine (C). The left set was loaded on the gel 16 hr before the right. The bracket shows the region of overlap.

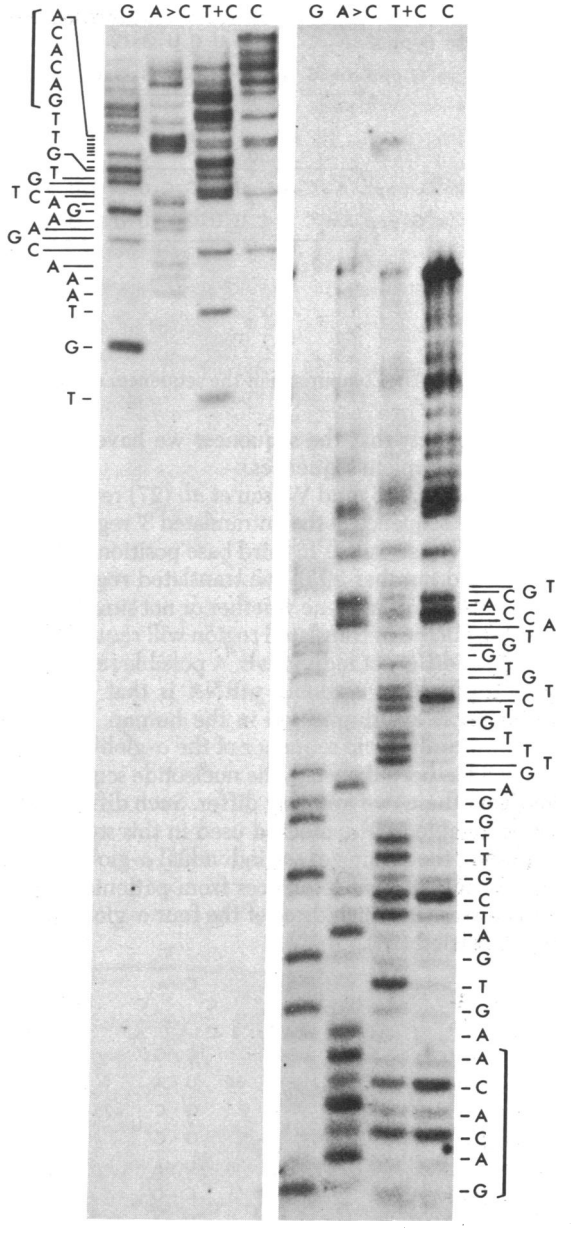

FIG. 5. Autoradiograph of the sequencing gel of 3′-^{32}P-labeled fragment (H3) from β-globin cDNA. The right set was loaded 40 hr before the left.

The methylation of some bases could not be determined from these experiments.

By sequencing the 3′-terminal cDNA fragments labeled with ^{32}P at their 5′ ends, we have determined the sequences corresponding to the translated 5′ region of the α- and β-globin mRNAs from the first *Hae* III site to the initiation codon and beyond, providing an overlap of at least 43 nucleotides for α and 12 for β (gels not shown). The sequences of the translated regions from the initiation codon to the *Hae* III site are shown in Table 1.

DISCUSSION

We have analyzed the 3′-terminal cDNA fragments produced by digestion of single-stranded globin cDNAs and determined the sequence of the 41 and 54 nucleotides in the untranslated 5′ region of the human α- and β-globin mRNAs, respectively. The sequences most likely represent the entire untranslated 5′ region of the globin mRNAs, except for the 5′ m^{7}Gppp "cap" structure. When Baralle (11, 12) compared the rabbit cDNA sequences he obtained to the 5′-mRNA sequences obtained by direct RNA analysis by Lockard and RajBhandary (17), he found that the cDNAs extended to the residue adjacent to the "cap". Only the first two nucleotides of human α- and β-globin mRNAs, 5′m^{7}GpppAm, have been directly determined (26). Our finding of T as the last nucleotide indicates that the cDNAs we sequenced could have extended to the penultimate nucleotide of the 5′ end of the mRNAs. The untranslated regions of human and rabbit α- and β-globin mRNAs show extensive homologies (Fig. 6): 76% of the α sequences and 78% of the β sequences are identical, including 12 of the first 13 nucleotides of the former and 18 of the first 20 of the latter. These simi-

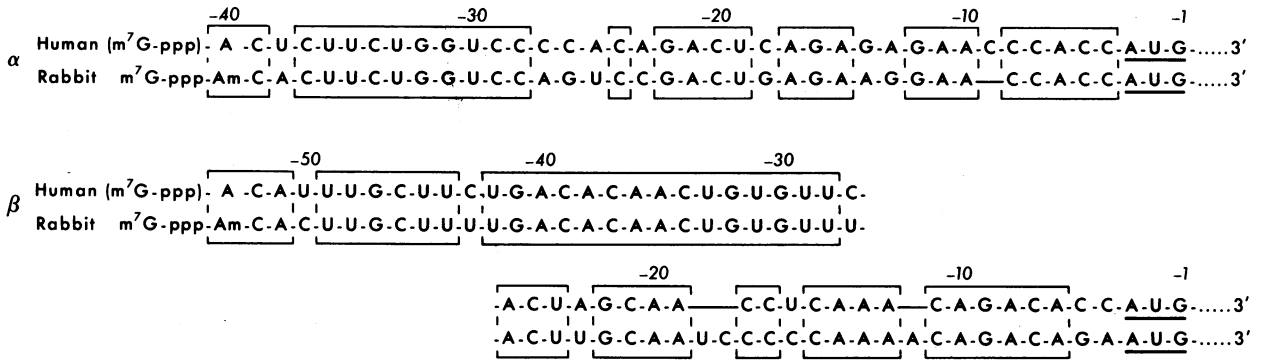

FIG. 6. Comparison of the sequences of the untranslated regions of human and rabbit α- and β-globin mRNAs.

larities also suggest that the sequences we have determined represent the complete sequences.

Weissman *et al.* (16) and Wilson *et al.* (27) reported heterogeneity in the sequence of the untranslated 3′ region of human α-globin mRNA, as well as in third base position of the codon for amino acid number 50 in the translated region of the β-globin mRNA. To determine whether or not similar polymorphism occurs in the untranslated region will require studies of mRNAs from different individuals. A possible reason for variation in sequence in α-globin mRNA is that the α-globin structural genes are duplicated in the human. Although no difference in amino acid sequence of the α-globin chain from the two loci has been detected, the nucleotide sequences of the mRNA from these two loci may differ. Such differences may not be detectable by the method used in this study, and will require either the cloning of the individual α-globin sequences or the determination of sequences from patients with hemoglobin H disease in which three of the four α-globin structural genes are deleted (28).

The untranslated 5′ region of mRNA is thought to be involved in ribosomal binding. In prokaryotes, substantial evidence indicates that a region located about 10 nucleotides 5′ to the initiation codon is involved in base pairing with the 3′-terminal sequence of the 16S rRNAs (29–31). Similar base pairing in eukaryotes between the 3′ terminus of the 18S rRNA and the untranslated region of the mRNA has also been suggested (32–34). However, the exact role these possible base pairings play in the initiation of protein synthesis in eukaryotes is not yet understood. In the human, five residues in both α- and β-globin mRNAs could base pair with the 3′-terminus 18S rRNA (Fig. 7).

Of the several models of secondary structure of the untranslated region that can be constructed, the ones shown in Fig. 7 for α- and β-globin mRNA appear to be most stable. The free energies (ΔG, 25°) of these structures estimated according to Tinoco *et al.* (35) are −10.4 and −11.2 kcal/mol, respectively. A similar secondary structure, although less stable, has been constructed for the rabbit α-globin mRNA (12).

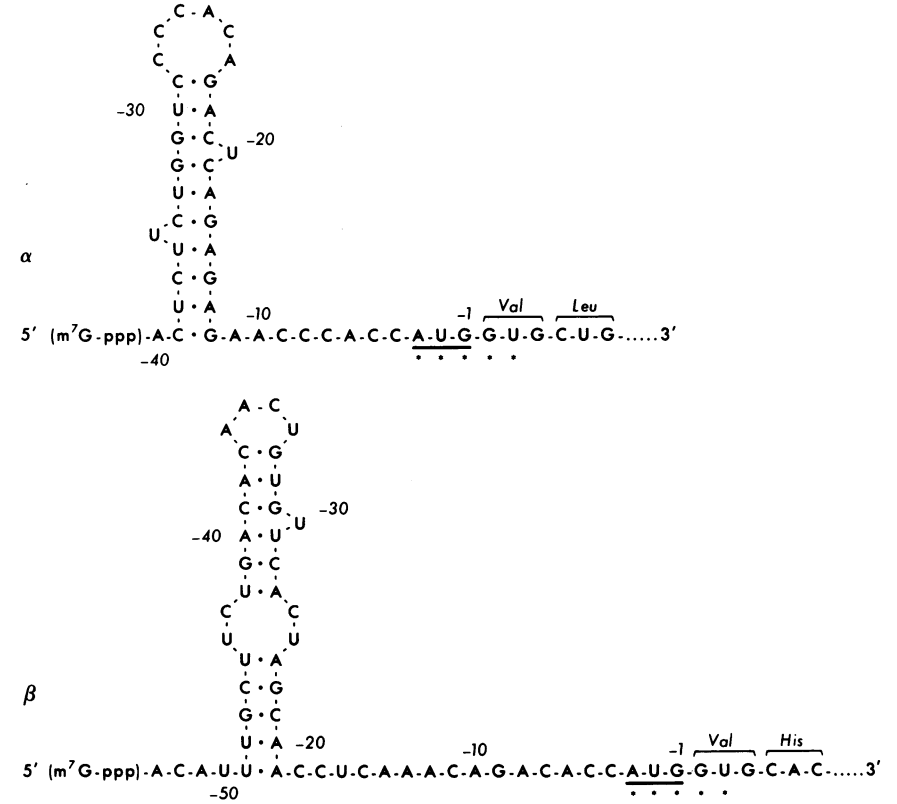

FIG. 7. Possible secondary structures in the untranslated regions of the human α- and β-globin mRNAs. Asterisks indicate possible base pairings with the 3′ terminus of 18S rRNA.

Table 1. Nucleotide and corresponding amino acid sequences of the translated regions of human α- and β-globin mRNAs derived from the 3′-terminal cDNA fragments

α-Globin mRNA

	1				5					10			
	Val	Leu	Ser	Pro	Ala	Asp	Lys	Thr	Asn	Val	Lys	Ala	
5′...	AUG	GUG	CUG	UCU	CCU	GCC	GAC	AAG	ACC	AAC	GUC	AAG	G(CC) ...3′

β-Globin mRNA

	1				5					10			
	Val	His	Leu	Thr	Pro	Val	Glu	Lys	Ser	Ala	Val	Thr	
5′...	AUG	GUG	CAC	CUG	ACU	CCU	GUG	GAG	AAG	UCU	GCC	GUU	ACU
			15					20					
	Ala	Leu	Trp	Gly	Lys	Val	Asn	Val	Asp	Glu	Val	Gly	
	GCC	CUG	UGG	GGC	AAG	GUG	AAC	GUG	GAU	GAA	GUU	GGU	
	25												
	Gly	Glu	Ala										
	GGU	GAG	G(CC)...3′										

The roles of the primary and secondary structures in the untranslated regions of the mRNAs in protein synthesis are not clear. In the human, a type of homozygous β° thalassemia in which β-globin mRNA is present but fails to function has been identified (36). Delineation of the primary structure of such defective mRNAs may provide valuable information on the relationship between mRNA structure and function.

We thank Mr. Allen Maxam and Dr. John Shine for their advice on the DNA sequencing technique, Ms. Andree Dozy for technical assistance, Ms. Jennifer Gampell for editorial assistance, and the Office of Program Resources and Logistics, Viral Cancer Program, Viral Oncology, Division of Cancer Cause and Prevention, National Cancer Institute, Bethesda, MD for the reverse transcriptase. This work was supported in part by grants from the National Institutes of Health (AM 16666), The National Foundation–March of Dimes, and a contract from Maternal and Child Health, Department of Health, State of California. Y.W.K. is an Investigator of the Howard Hughes Medical Institute.

1. Poon, R., Paddock, G. V., Heindell, H., Whitcome, P., Salser, W., Kacian, D., Bank, A., Gambino, R. & Ramirez, F. (1974) *Proc. Natl. Acad. Sci. USA* **71**, 3502–3506.
2. Forget, B. G., Marotta, C. A., Weissman, S. M., Verma, I. M., McCaffrey, R. P. & Baltimore, D. (1974) *Ann. N.Y. Acad. Sci.* **241**, 290–309.
3. Sanger, F. & Coulson, A. R. (1975) *J. Mol. Biol.* **94**, 441–448.
4. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 560–564.
5. Efstratiadis, A., Kafatos, F. C. & Maniatis, T. (1977) *Cell* **10**, 571–585.
6. Maniatis, T., Sim, G. K., Efstratiadis, A. & Kafatos, F. C. (1976) *Cell* **8**, 163–182.
7. Liu, A., Paddock, G. V., Heindell, H. C. & Salser, W. (1977) *Science* **196**, 192–196.
8. Browne, J. K., Paddock, G. V., Liu, A., Clarke, P., Heindell, H. C. & Salser, W. (1977) *Science* **195**, 389–391.
9. Proudfoot, N. J. (1976) *J. Mol. Biol.* **107**, 491–525.
10. Proudfoot, N. J. & Longley, J. I. (1976) *Cell* **9**, 733–746.
11. Baralle, F. (1977) *Cell* **10**, 549–558.
12. Baralle, F. (1977) *Nature* **267**, 279–281.
13. Marotta, C. A., Forget, B. G., Cohen-Solal, M., Wilson, J. T. & Weissman, S. M. (1977) *J. Biol. Chem.* **252**, 5019–5031.
14. Cohen-Solal, M., Forget, B. G., Prensky, W., Marotta, C. A. & Weissman, S. M. (1977). *J. Biol. Chem.* **252**, 5032–5039.
15. Marotta, C. A., Wilson, J. T., Forget, B. G. & Weissman, S. M. (1977) *J. Biol. Chem.* **252**, 5040–5051.
16. Weissman, S. M., Forget, B. G., Marotta, C. A. & Wilson, J. W. (1977) *Clin. Res.* **21**, 519A.
17. Lockard, R. E. & RajBhandary, U. L. (1976) *Cell* **9**, 747–760.
18. Horiuchi, K. & Zinder, N. D. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 2555–2558.
19. Blakesley, R. W. & Wells, R. D. (1975) *Nature* **257**, 421–422.
20. Temple, G. F., Chang, J. C. & Kan, Y. W. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 3047–3051.
21. Verma, I. M., Temple, G. F., Fan, H. & Baltimore, D. (1972) *Nature New Biol.* **235**, 163–167.
22. Efstratiadis, A., Maniatis, T., Kafatos, F. C., Jeffrey, A. & Vournakis, J. N. (1975) *Cell* **4**, 367–378.
23. Higuchi, R., Paddock, G. V., Wall, R. & Salser, W. (1976) *Proc. Natl. Acad. Sci. USA* **73**, 3146–3150.
24. Weiss, B., Live, T. R. & Richardson, C. C. (1968) *J. Biol. Chem.* **243**, 4530–4542.
25. Maniatis, T., Jeffrey, A. & Van de Sande, H. (1975) *Biochemistry* **14**, 3787–3794.
26. Cheng, T-c., Thompson, B. J. & Kazazian, H. H., Jr. (1976) *Blood* **48**, 998A.
27. Wilson, J. T., Forget, B. G., Wilson, L. B. & Weissman, S. M. (1977) *Science* **196**, 200–202.
28. Kan, Y. W., Dozy, A. M., Varmus, H. E., Taylor, J. M., Holland, J. P., Lie-Injo, L. E., Ganesan, J. & Todd, D. (1975) *Nature* **255**, 255–256.
29. Shine, J. & Dalgarno, L. (1974) *Proc. Natl. Acad. Sci. USA* **71**, 1342–1346.
30. Steitz, J. A. & Jakes, K. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 4734–4738.
31. Steitz, J. A., Wahba, A. J., Laughrea, M. & Moore, P. B. 1977 *Nucleic Acids Res.* **4**, 1–15.
32. Dasgupta, R., Shih, D. S., Saris, C. & Kaesberg, P. (1975) *Nature* **256**, 624–627.
33. Legon, S., Robertson, H. D. & Prensky, W. J. (1976) *J. Mol. Biol.* **106**, 23–37.
34. Legon, S. (1976) *J. Mol. Biol.* **106**, 37–53.
35. Tinoco, I., Borer, P. N., Bengler, B., Levine, M. D., Uhlenbeck, O. C., Crothers, D. M. & Gralla, J. (1973) *Nature New Biol.* **246**, 40–41.
36. Kan, Y. W., Holland, J. P., Dozy, A. M. & Varmus, H. E. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 5140–5144.