Volume 3 no.7 July 1976 Nucleic Acids Research

Nucleotide sequence from the coding region of rabbit β-globin messenger RNA

N.J. Proudfoot

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England, UK.

Received 7 June 1976

ABSTRACT

A sequence of 89 nucleotides from rabbit β-globin mRNA has been determined and is shown to code for residues 107 to 137 of the β-globin protein. In addition, a sequence heterogeneity has been identified within this 89 nucleotide long sequence which corresponds to a known polymorphic variant of rabbit β-globin.

INTRODUCTION

The sequence analysis of globin messenger RNA (mRNA) as with several other eukaryotic mRNAs has rapidly progressed in the last few years. Most of this progress can be ascribed to the discovery of reverse transcriptase (1) and this enzyme's subsequent use in the *in vitro* synthesis of complementary DNA from a mRNA template. cDNA synthesis is initiated by the hybridisation of an oligo(dT) primer to the 3' terminal poly(A) sequence present in most mRNAs (2). Thus Poon *et al.* (1974) (3) and Marotta *et al.* (1974) (4) described the sequence analysis of ^{32}P-labelled complementary RNA obtained by transcription of globin cDNA. These and more recent studies (5,6) established the sequences of various T_1 ribonuclease products, many of which have been shown to code for small sections of either the α- or β-globin proteins. Proudfoot and Brownlee (1974) (7) used a somewhat different approach to sequence the 3' terminal region of rabbit β-globin mRNA. They obtained short, ^{32}P-labelled cDNA using the reverse transcriptase activity of DNA polymerase I (*E. coli*) (8,9) and directly sequenced this using various DNA sequencing procedures (endonuclease IV digestion (10-12) and partial exonuclease digestion (13,14) in particular). This approach has been successfully applied to other mRNAs (15-17) as well as to both human (18) and rabbit globin mRNAs (12), so that substantial parts of the 3' non-coding regions of several different eukaryotic mRNAs are now sequenced.

Although the direct sequence analysis of cDNA is a particularly

suitable means of determining the 3' non-coding sequence of a mRNA, I will describe in these studies the sequence analysis of rabbit globin cDNA complementary to the coding region of the mRNA. These studies have allowed the sequence determination of 89 nucleotides that code for rabbit β-globin, amino acids 107-137. This result clearly illustrates that oligo(dT) primed cDNA may be directly used to determine the sequence of a mRNA at an internal position in the molecule and not just in the 3' terminal region.

MATERIALS AND METHODS

(a) Transcription methodology

(i) Using DNA polymerase I (E. coli). Rabbit globin mRNA [purified from rabbit reticulocytes as previously described (2)] was copied into ^{32}P-labelled cDNA using subtilisin-treated DNA polymerase I (Boehringer Mannheim Corp., West Germany) in the presence of Mn^{2+} (8,12) with oligo $(dT)_{10}$ (P.L. Biochemicals Inc., Wis., U.S.A.) as primer. The transcription reaction was carried out for 1 hr at 37°C in a sealed tube [for details see (12)].

(ii) Using reverse transcriptase (Avian Myloblastosis virus). Rabbit globin mRNA was copied into ^{32}P-labelled cDNA using reverse transcriptase (a gift from Dr. J.W. Beard) with oligo$(dT)_{10}$ as primer using the conditions previously described (12). However, each of the four deoxyribonucleotide triphosphates were α-^{32}P-labelled in turn (NEN, Boston, Mass., U.S.A.).

The cDNA obtained with either DNA polymerase or reverse transcriptase was purified from the reaction mixture by phenol extraction followed by ethanol precipitation and then finally acid precipitation [see (12)].

(b) Sequencing techniques

The cDNA obtained with both reverse transcriptase and DNA polymerase was digested with endonuclease IV (10-12) and the products of digestion were then fractionated in two dimensions using electrophoresis on Cellogel at pH 3.5 in the first dimension and homochromatography in the second dimension. (Cellogel was obtained from Reeve-Angel Ltd., Ashford, Kent, U.K.). The separated oligonucleotides were then eluted and subjected to sequence analysis procedures as previously described in detail (12). The oligonucleotides were isolated from several endonuclease IV fingerprints of cDNA labelled with different input labels. They were then subjected to partial venom exonuclease digestion technique (11,14), depurination (14) and nearest neighbour analysis (13). Sufficient information was thus obtained to determine their sequences.

RESULTS AND DISCUSSION

(a) Sequence analysis of endonuclease IV digestion products

Rabbit globin cDNA was synthesised using both DNA polymerase I (12) and reverse transcriptase. However, the DNA polymerase I derived cDNA was made using high concentrations of nucleoside triphosphate precursors (1 mM) as under such conditions longer cDNA molecules of 200 or more nucleotides may be produced (12). Figure 1 shows the endonuclease IV (a deoxyribonuclease specific for cytidine residues) fingerprints (11,12) of these two preparations. As indicated the two fingerprints are very similar.

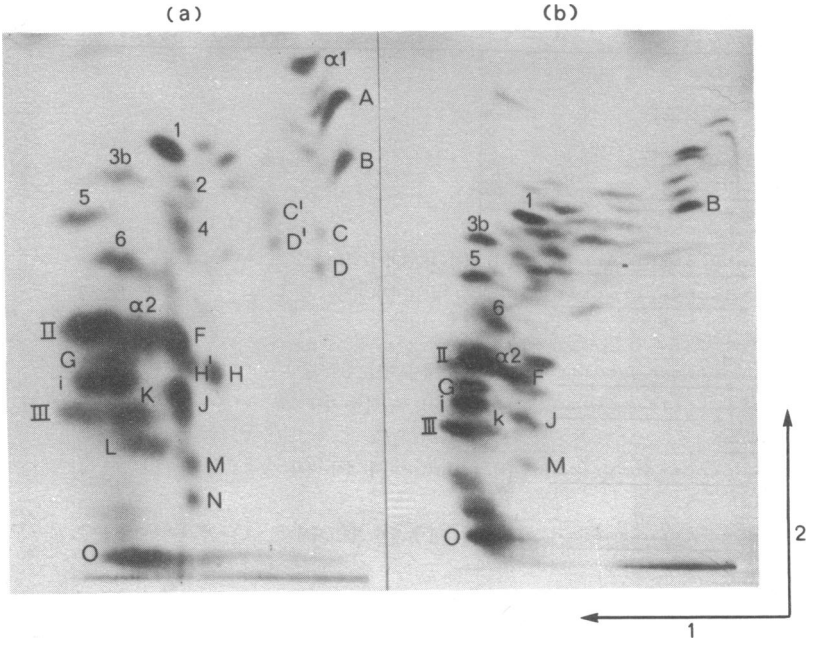

Fig. 1. Radioautographs of two endonuclease IV fingerprints (11,12) of rabbit globin cDNA obtained with DNA polymerase I (E. coli) (a) and reverse transcriptase (b). 1 denotes electrophoresis at pH 3.5 on Cellogel, 2 denotes homochromatography. Sequence data on the oligonucleotides labelled with capital letters is contained within Table 1. The sequences of the other labelled oligonucleotides have been previously reported (12). Both cDNAs were synthesised using [α-^{32}P] dCTP.

However, the reverse transcriptase cDNA fingerprint contains many more low yield spots, reflecting the DNA's greater length (19,20). In a previous communication (12), the sequence and location of many of the predominant oligonucleotides (labelled with numbers and Roman numerals) were described.

However several oligonucleotides (labelled with capital letters) were not included in this 3' non-coding region sequence as no overlaps could be found.

Table I contains the sequence data obtained for these endonuclease IV products. As indicated, a combination of partial venom exonuclease digestion (11,14) depurination (14) and nearest neighbour analysis (13) data allows the complete sequence determination of many of the oligonucleotides.

TABLE I. Sequence data on endonuclease IV derived oligonucleotides

Spot No.	Partial exonuclease digestion data	Depurination data	Nearest neighbour analysis data: Input label dCTP C A G X	dATP C A G X	dGTP C A G X	dTTP C A G X
A	C-A-C-C-A-G-C ←		-	-	-	-
B	C-A-C-C-A-G-C-C-A-C ←		-	-	-	-
C	T† C-A-A-C-A-A-C-C-A-G ←		-	-	-	-
D	T† C-A-A-C-A-A-C-C-A-G-C-A ←		-	-	-	-
E	C-A-C-C-A-G-C-C-A-C-C-A ←		3 2† 1 1	-	-	-
F	C-A-C-C-A-C-C-T-T-C-T-G-A-T-A-G-G ←	C[A], C_2[A], (C_3,T_3)[G], T[A]	1 1 0 1	-	0 1 1 1	-
G	C-T-T-T-G-C-C-A-A-A-A-T-G-A-T-G-A- ← G-A	(C,T_3)[G], C_2[A], T[G], T[G]	-	1 3 3 0	0 1 0 3†	-
H	T† C-A-A-C-A-A-C-C-A-G-C-A-C-G-T-T-G ←	C[A], C[A], C_2[A], C[A], C[G], T_2[G]	1 3 1† 1	3 2† 0 1	1 1† 0 1	-
I	C-C-T-G-C-A-C-C-T-G-A-G-G-A-G-T-G- ← A-A-T-T	(C_2,T)[G], C[A], (C_2,T)[G], T[G]	1 1 3§ 0	1† 1 3 0	0 2 1 3	-
J	C-A-C-C-A-C-C-T-T-C-T-G-A-T-A-G-G- ← C-A-G	C[A], C_2[A], (C_3,T_3)[G], T[A], C[A]	2 2† 1 2	2 0 1 2	0 2 1 1	-
K	C-T-T-T-G-C-C-A-A-A-A-T-G-A-T-G-A- ← G-A-C-A-G	(C,T_3)[G], C_2[A], T[G], T[G], C[A]	1 1 1 1	2 3 3 0	0 2 0 3	-
L	C-T-T-T-G-C-C-A-A-A-A-T-G-A-T-G-A- ← G-A-C-A-G-C-A	(C,T_3)[G], C_2[A], T[G], T[G], C[A], C[A]	1 1 2 1	1 1† 1 0	0 2 0 3	-
M	C-A-C-C-A-G-C-C-A-C-C-A-C-C-T-T-C- ← T-G-A-T-A-G-G	C[A], C_2[A], C_2[A], C_2[A], (C_3,T_3)[G], T[A]	-	H 0 1 1	0 2† 1 1	-
N	C-A-C-C-A-G-C-C-A-C-C-A-C-C-T-T-C- ← T-G-A-T-A-G-G-C-A-G	C[A], C_2[A], C_2[A], C_2[A], (C_3,T_3)[G], T[A], C[A]	-	H 0 1 1	0 3 1 1	-
O		(C_3,T_3), C_2, C_2, C_2, C, C	-	-	-	-

Partial exonuclease digestion data: ← denotes sequence deduced by partial venom exonuclease digestion technique. Arrow covers deduced sequence. † denotes sequence heterogeneity.

Depurination data: 5' and 3' terminal phosphates of depurination products are omitted. [ ] denotes that the base within the brackets was deduced by nearest neighbour considerations. Thus depurination products obtained from [α-^{32}P]dGTP or [α-^{32}P] dATP input label oligonucleotides contain G or A residues (respectively) on their 3' sides.

Nearest neighbour analysis data: The section is divided into four parts for oligonucleotides labelled by each of the different input labels as indicated ([α-^{32}P]labelled). Each input label is divided into a further four columns indicating the products of analysis: Cp, Ap, Gp, and Tp or pCp (not separated in the fractionation system). These are denoted by C, A, G and X respectively. Ratios of product yields are listed. O denotes zero yield, whilst H denotes high yield. It should be noted that pCp derives from the 5' terminus of all endonuclease IV derived oligonucleotides. † denotes uncertainty of numerical value. § denotes pCpCp is resistant to spleen phosphodiesterase [the principal enzyme used in nearest neighbour analysis (12)]. This product, which derives from the 5' end of spot i, co-migrated with Gp in the fractionation system.

The partial venom exonuclease digestion fingerprints of spots E, J, i, L and H' are shown in Fig. 2, in combination covering most of the sequence data. It should be noted that spots C', D' and H' are variants of spots C, D and H (related oligonucleotides - see Table I) and contain a dC to dT base change.

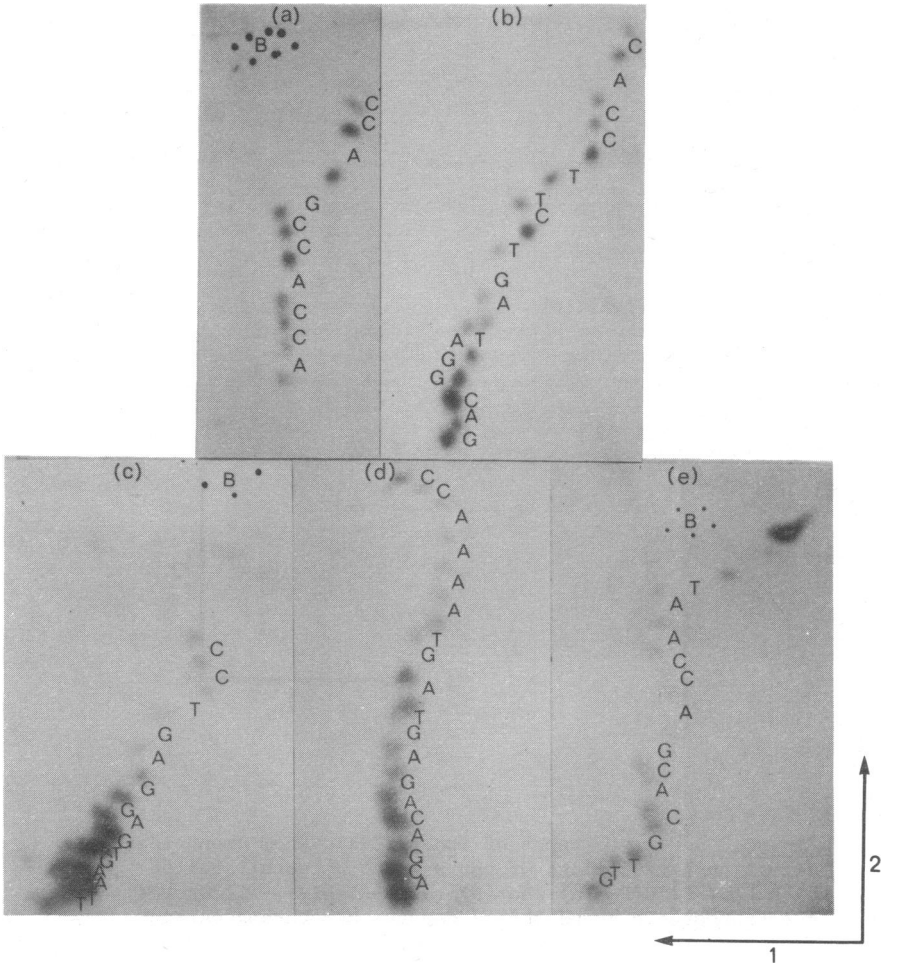

Fig. 2. Radioautographs of partial venom exonuclease digestion fingerprints E, J, i, L and H' (labelled with [α-^{32}P]dCTP, [α-^{32}P]dCTP, [α-^{32}P]dATP, [α-^{32}P]dGTP and [α-^{32}P]dGTP respectively) from the endonuclease IV fingerprints as in Fig. 1 [(a), (b), (c), (d) and (e) respectively], fractionated in two dimensions: first dimension pH 3.5 electrophoresis (1); second dimension (2) homochromatography. The prefix d for deoxy is omitted in the figures and table.

This result is illustrated by Fig. 3 which compares the partial venom exonuclease digestion fingerprints of spots D and D'. The significance of this sequence heterogeneity will be discussed below.

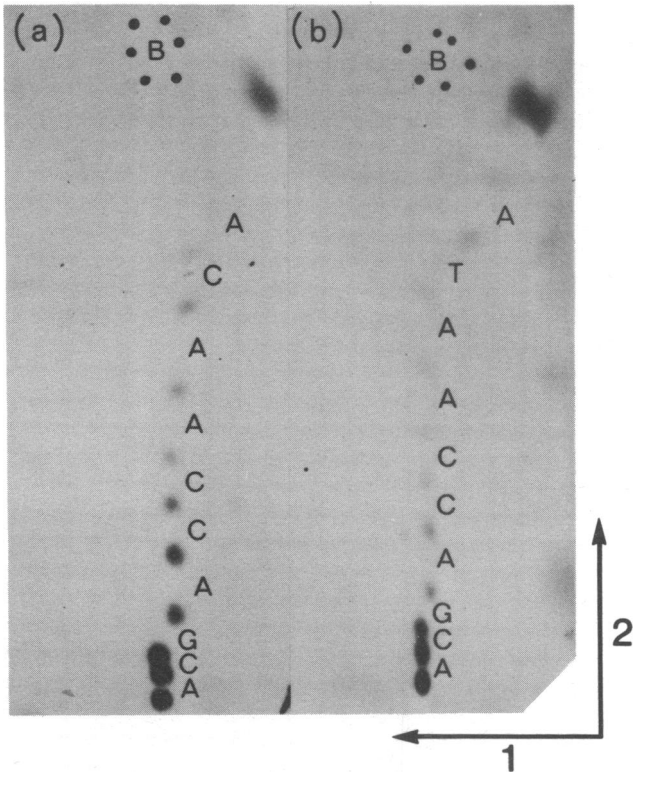

Fig. 3. Radioautographs of two partial venom exonuclease digestion fingerprints of spots D and D' both labelled with [α-^{32}P]dCTP [(a) and (b) respectively]. 1 denotes electrophoresis at pH 3.5; 2 denotes homochromatography.

Figure 4 summarises the sequence data obtained for the different oligonucleotides and shows that all fourteen of them (described in Table I) derive from one of the four larger oligonucleotides N, I, L and H [numbered (1), (2), (3) and (4) respectively in the figure]. In particular, spots A, B, E, F, J and M derive from N, G and K from L, and C and D from H. This result reflects the partial nature of the endonuclease IV activity, as described

by others (10,11). These smaller oligonucleotides provided a very important cross-check on regions of the larger oligonucleotides that were not unequivocally sequenced. However, no oligonucleotides have been identified that overlap these four sequences. Their positions within the whole mRNA were established by other considerations as described below.

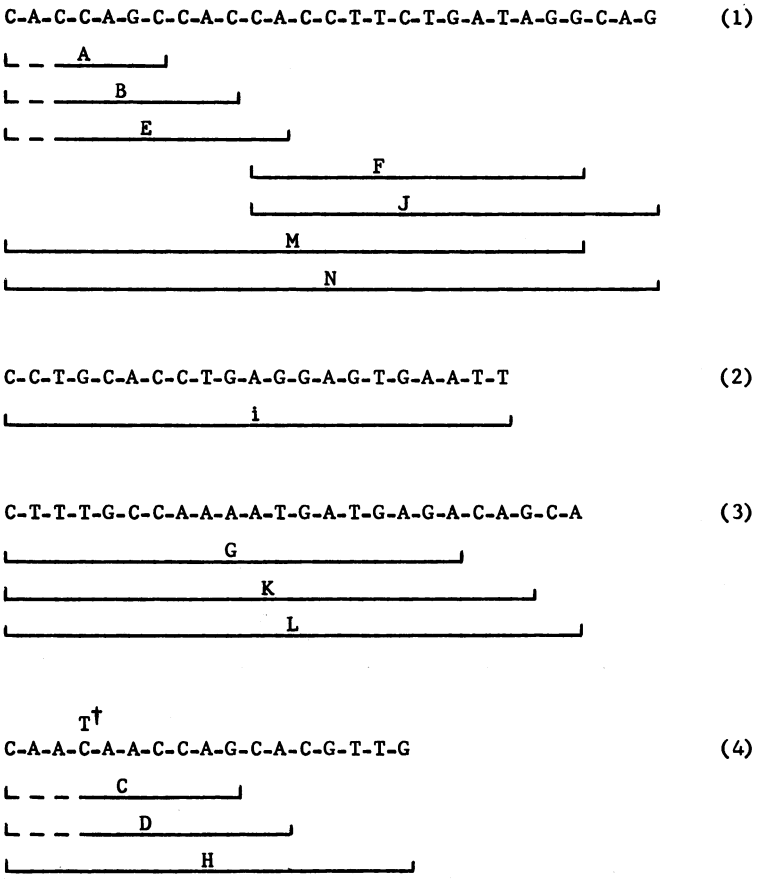

Fig. 4. Nucleotide sequences of spots N, i, L and H [(1), (2), (3) and (4)]. Positions of related oligonucleotides - see Table I - are drawn in. Regions of different oligonucleotides that have not been independently sequenced are denoted by dashed line. † denotes sequence heterogeneity.

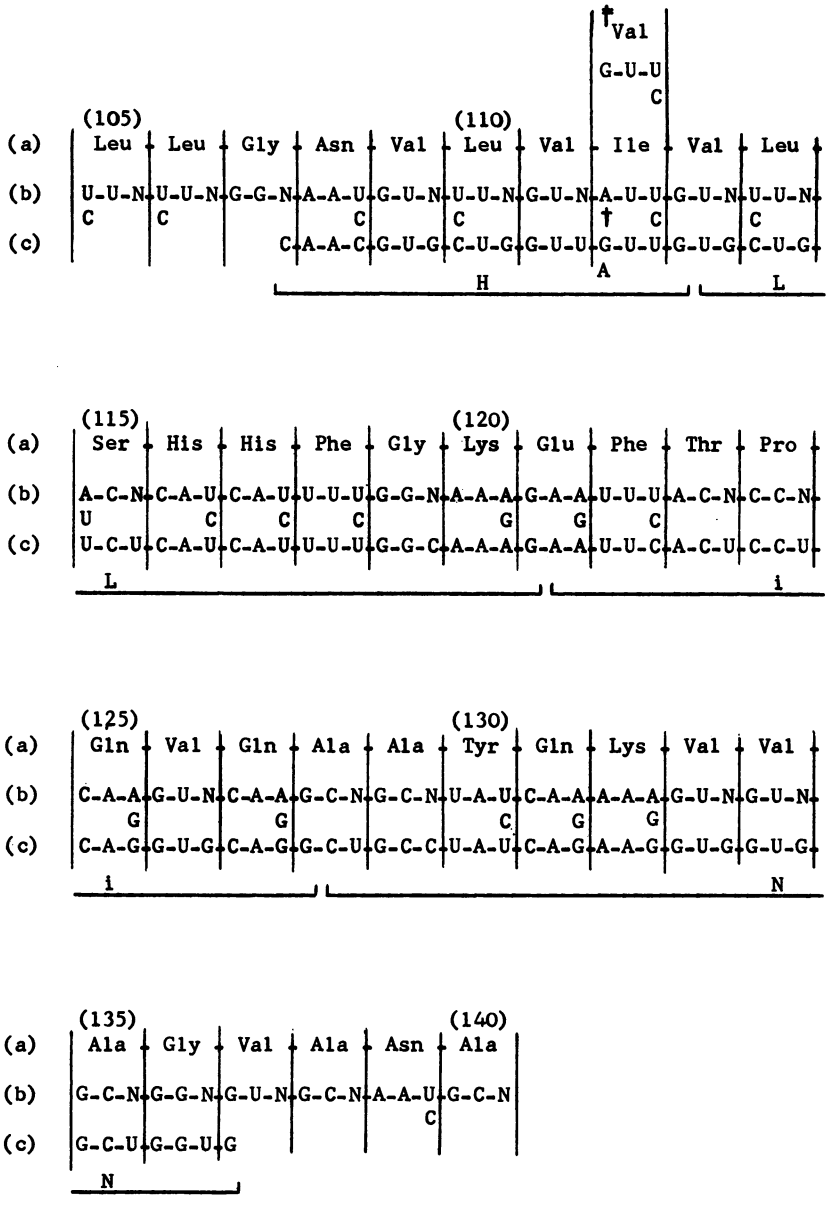

Fig. 5. Comparison of the RNA sequence deduced from the cDNA nucleotides N, i, L and H (indicated by horizontal brackets) to the RNA sequence predicted by rabbit β-globin (residues 105-140). (a) denotes amino acid sequence, (b) denotes protein predicted RNA sequence, (c) denotes RNA sequence complementary to cDNA sequences. † denotes nucleotide sequence heterogeneity. ‡ denotes amino acid sequence heterogeneity (19).

(b) cDNA oligonucleotides correspond to β-globin (residues 107-137)

Figure 5 demonstrates that the four cDNA oligonucleotide sequences (drawn in their RNA sense) code for a section of rabbit β-globin (21). Indeed, these four oligonucleotides may be arranged in order and are shown to be contiguous with each other in the mRNA sequence. A sequence of 89 nucleotides is therefore established for the β-globin mRNA towards the carboxy terminal end of the coding region. The interesting sequence heterogeneity demonstrated to exist in the cDNA sequence requires an A to G base change in the RNA sequence. Therefore a valine to isoleucine amino acid variation in the protein primary structure is predicted. Indeed, such an amino acid variation is well established. Thus rabbit β-globin amino acid residue 112 is known to be polymorphic and may be either isoleucine or valine (22). It appears that the rabbits from which the globin mRNA used in these experiments was isolated contain approximately equal amounts of valine and isoleucine at residue 112 in the β-globin sequence (i.e. spots C, D and H are in nearly equivalent yield to C', D' and H' - see Fig. 1a).

Consideration of Fig. 1(a) reveals that all of the significantly high yield oligonucleotides (labelled) have been accounted for in either the 3' terminal sequences of the α- and β-globin mRNA described elsewhere (12) or in the β-globin mRNA coding region presented here. However, these coding and non-coding region sequences are likely to be at least 100 nucleotides apart. This is because the β-globin mRNA coding sequence extends for 30 nucleotides to the 3' side of the sequence described here, and Paddock et al. (23) have recently described several oligonucleotide sequences in the 3' non-coding region of rabbit β-globin mRNA, in addition to the 75 3' terminal nucleotides already sequenced (12). It would therefore seem likely that this 100 nucleotide sequence connecting the coding region sequence described here to the 3' terminal sequence previously reported (12), is very resistant to endonuclease IV [possibly due to secondary structure, which is known to inhibit the enzyme (11)]. Undigested DNA would not be visible on the two-dimensional fingerprint as it remains at the origin of the first dimension (not shown in the figure).

CONCLUSIONS

The sequence of 89 nucleotides coding for residues 107-137 of rabbit β-globin described here contains few points of particular interest other than to further substantiate the validity of the genetic code. It does not appear to have any significant secondary structure, unlike the RNA sequences determined for the 3' non-coding regions of various mammalian mRNAs (12,16).

However, there does appear to be some degree of triplet codon selection. In particular valine occurs seven times in the sequence and is coded for by either GUU or GUG but not GUA or GUC. Similar levels of triplet codon selection have been suggested by Paddock et al. (23) and Forget et al. (24). Paddock et al. (23) have reported the sequence of three T_1-oligonucleotides that correspond to large sections of the sequence described here, so that there is strong supporting evidence for the data presented in this paper. Also, the fact that the sequence technology used in these experiments is sufficiently sensitive to pick up a known β-globin (residue 112) polymorphic variant is a good vindication of accuracy.

ACKNOWLEDGEMENTS

I thank Dr. G.G. Brownlee for invaluable help and discussion, both during the experiments and in the preparation of the manuscript. I would also like to acknowledge helpful discussions with Drs. C. Milstein, T. Hunt, G.V. Paddock and W. Salser. Dr. N.J. Proudfoot is a Junior Beit Memorial Fellow.

REFERENCES

1. Temin, H.M. and Baltimore, D. (1972) Advan. Virus Res. 17, 129-186.
2. Brawerman, G. (1974) Ann. Rev. Biochem. 42, 621-642.
3. Poon, R., Paddock, G.V., Heindell, H., Whitcome, P., Salser, W., Kacian, D., Bank, A., Gambino, R. and Ramirez, F. (1974) Proc. Nat. Acad. Sci. USA 71, 3502-3506.
4. Marotta, C.A., Forget, B.G., Weissman, S.M., Verma, I.M., McCaffrey, R.P. and Baltimore, D. (1974). Proc. Nat. Acad. Sci. USA 71, 2300-2304.
5. Salser, W., Bowen, S., Browne, D., El Adli, F., Fedoroff, N., Fry, K., Heindell, H., Paddock, G.V., Poon, R., Wallace, B. and Whitcome, P. (1976) Fed. Proc. 35, 23-35.
6. Forget, B.G., Marotta, C.A., Weissman, S.M. and Cohen-Solal, M. (1975) Proc. Nat. Acad. Sci. USA 72, 3614-3618.
7. Proudfoot, N.J. and Brownlee, G.G. (1974) Nature (London) 252, 359-362.
8. Proudfoot, N.J. and Brownlee, G.G. (1974) FEBS Letters 38, 179-183.
9. Modak, M.J., Marcus, S.L. and Cavalieri, L.F. (1973) Biochem. Biophys. Res. Commun. 55, 1-7.
10. Sadowski, P.D. and Bakyta, I. (1972) J. Biol. Chem. 247, 405-412.
11. Galibert, F., Sedat, J.W. and Ziff, E.B. (1974) J. Mol. Biol. 87, 377-407.
12. Proudfoot, N.J. (1976) J. Mol. Biol. in the press.
13. Sanger, F., Donelson, J.E., Coulson, A.R., Kössel, H. and Fischer, D. (1974) J. Mol. Biol. 90, 315-333.
14. Ling, V. (1972) J. Mol. Biol. 64, 87-102.
15. Milstein, C., Brownlee, G.G., Cartwright, E.M., Jarvis, J.M. and Proudfoot, N.J. (1974) Nature (London) 252, 354-359.
16. Cheng, C.C., Brownlee, G.G., Carey, N.H., Doel, M.T., Gillam, S. and Smith, M. (1976) J. Mol. Biol. in the press.
17. Dixon, G.H., Davies, P.L., Ferrier, L.N., Gedamu, L. and Iatrou, K. (1977) Progress in Nucleic Acid Research and Molecular Biology, ed. W.E. Cohn (Academic Press, New York) 17, in the press.

18. Proudfoot, N.J. and Longley, J. (1976) manuscript in preparation.
19. Ross, J., Aviv, H., Scholnick, E. and Leder, P. (1972) Proc. Nat. Acad. Sci. USA 69, 264-268.
20. Rabbitts, T.H. and Milstein, C. (1975) Eur. J. Biochem. 52, 125-133.
21. Dayhoff, M.O. (ed.) (1969) Atlas of Protein Sequence and Structure, Vol. 4, National Biomedical Research Foundation, Silver Spring.
22. Bricker, J. and Garrick, M.D. (1974) Biochim. Biophys. Acta 351, 437-441.
23. Paddock, G.V., Heindell, H.C., Isaacson, J., Poon, R., Ramirez, F., Bank, A., Kacian, D. and Salser, W. (1976) submitted to Proc. Nat. Acad. Sci. USA.
24. Forget, B.G., Marotta, C.A., Weissman, S.M., Verma, I.M., McCaffrey, R.P. and Baltimore, D. (1974) Ann. N.Y. Acad. Sci. 241, 290-309.