Rabbit Globin mRNA

ANALYSIS OF T1 RNASE DIGESTION FRAGMENTS*

(Received for publication, May 7, 1976, and in revised form, December 14, 1976)

GARY V. PADDOCK,† RAYMOND POON, HOWARD C. HEINDELL, JUDITH ISAACSON,§ AND WINSTON SALSER¶

From the Department of Biology and Molecular Biology Institute, University of California at Los Angeles, Los Angeles, California 90024

Rabbit globin complementary DNA made with RNA-dependent DNA polymerase (reverse transcriptase) was used as a template for in vitro synthesis of 32P-labeled RNA and deoxysubstituted RNA. The sequences of the nucleotides in most of the fragments resulting from combined ribonuclease T1 and alkaline phosphatase digestion have been determined. In addition, the 3' nearest neighbor was determined for several fragments resulting from digestion with T1 ribonuclease. The utility of the deoxysubstitution technique was demonstrated by the ease with which the sequences of pyrimidine-rich fragments could be determined. Many sequences thus determined were long enough to fit uniquely with the α- or β-globin amino acid sequences. The positions of these fits were found to be clustered, leading us to believe that only certain regions of the complementary DNA are transcribed by Escherichia coli RNA polymerase. Other unique characteristics of RNA synthesis from a complementary DNA template include a high yield of free poly(A) and the fact that one must use low rather than high salt buffers to obtain transcripts of high molecular weight.

While rapid progress has been made in determining the primary structures of prokaryotic nucleic acids, nucleotide sequencing of mammalian mRNAs has met with less success. There has been, however, considerable progress in obtaining and characterizing pure mammalian mRNA species either from specialized systems synthesizing one or a few proteins in large amounts (e.g. reticulocytes) or, where such advantages do not present themselves, by more advanced methods such as immunoprecipitation of polyribosomes synthesizing a specific protein. We have for some time been working on the sequence determination of the rabbit α- and β-globin mRNAs which can easily be obtained in a pure state from reticulocytes. As a first step toward this goal we have determined a catalogue of T1 ribonuclease digestion fragment sequences.

In the past it has been exceedingly difficult to determine accurately the sequences of such a large catalogue of T1 fragments. For example, Woese and co-workers (1) have pointed out many difficulties in the determination of the catalogue of T1 fragments from the 16 S rRNA of Escherichia coli and had to develop special techniques to obtain the correct sequences (2). Not only do the globin mRNA sequences yield a large number of T1 fragments but many of these fragments proved difficult to sequence with the more commonly used techniques. One of our major aids in solving such problems has been the use of deoxysubstitution techniques (3, 4). This paper represents the first extensive use of such techniques and provides extensive characterizations of the deoxysubstitution products which should be useful to others employing this technique. This approach has proved especially effective for solving runs of pyrimidines within a T1 fragment and for solving the sequences of mixtures of T1 fragments.

Progress in sequence analysis of mammalian mRNAs has been hindered by the fact that ribosomal RNA or ribosomal RNA fragments are frequently a major contaminant of the mRNA preparations thus purified and by the great difficulty in labeling with 32P efficiently in vivo in order to obtain the very high specific activities needed for such work. The discovery of the RNA-dependent DNA polymerase from avian myeloblastosis virus (5, 6) and the subsequent demonstration that this enzyme can synthesize a complementary DNA (cDNA) copy of the globin mRNA when oligo(dT) is used as a primer (7, 8) suggested to us a way of overcoming these problems using a novel approach which has been presented in preliminary papers on this work (9, 10). The strategy has been to synthesize globin cDNA and use it as a template for in vitro RNA synthesis with E. coli RNA polymerase and highly 32P-labeled nucleoside triphosphate precursors. By using oligo(dT) primer for reverse transcriptase one can ensure that only poly(A)-containing mRNA serves as template for the synthesis of cDNA and since the subsequent syntheses are carried out in vitro very high 32P specific activities can be utilized. The radioactive RNA transcript can then be fragmented with appropriate nucleases and subjected to sequence analysis by the method of Sanger and collaborators (11–13). This approach has also been used by others (14).

The 9 S fraction of rabbit reticulocyte RNA has been shown to code for the α- and β-globin polypeptides (15), which contain 141 and 146 amino acids, respectively. Their amino acid se-
quences have been determined (16). The $\alpha$- and $\beta$-globin mRNAs were estimated to be approximately 670 nucleotides from measurements by formamide acrylamide electrophoresis (17). Each of these mRNAs has also been found to contain a poly(A) tract 50 to 70 nucleotides long (18).

In the presence of actinomycin D, reverse transcriptase synthesizes single strand cDNA with a length comparable to that of the RNA template as measured by alkaline sucrose gradient centrifugation. This cDNA can specifically reanneal to the RNA template (7, 8, 19, 20), and it has provided a useful assay for globin sequences in numerous biological systems. For example, it has been used as a probe to test for the presence of globin sequences in heterogeneous nuclear RNA (HnRNA) (21), to measure the relative amount of hemoglobin mRNA in thalassemic reticulocytes (22, 23), to follow the presence of globin sequences in heterogeneous nuclear RNA (HnRNA) (21), to measure the relative amount of hemoglobin mRNA in thalassemic reticulocytes (22, 23), to follow the control of transcription of various tissue-specific chromatin in vitro (24, 25), and to assay for the number of genes coding for globins (26, 27).

The available evidence suggests that the globin cDNA synthesized in vitro is a faithful and perhaps even a complete copy of the reticulocyte 9 S RNA. Nevertheless, comparison of the nucleotide sequence with the amino acid sequence would provide a more stringent test of the fidelity of the reaction since a relatively high degree of mispairing could have gone undetected in the earlier reannealing experiments. In fact, preliminary results using the methods outlined above (3, 14) indicate that cDNA is a faithful copy of mRNA. More importantly, direct sequence analysis of mammalian mtRNAs should permit us to answer a number of fundamental questions about the detailed nature of the signals for initiation and termination of protein synthesis, the signals for addition of poly(A) sequences and for cleavage of the mRNA from any HnRNA precursors, and the detailed nature and extent of base pairing within the primary gene product itself as well as the large untranslated sequences.

The in vitro RNA synthesis reaction using synthetic, single-stranded cDNA as a template, and also report on the sequence determinations of most of the ribonuclease T1 fragments of rabbit hemoglobin messenger, several of which are long enough to fit uniquely into the rabbit globin sequences. Our results demonstrate that reverse transcriptase synthesizes faithful cDNA copies of mRNA sequences and that deoxysubstitution techniques provide an useful tool for the accurate determination of the sequences obtained by transcribing these cDNAs.

**EXPERIMENTAL PROCEDURES**

**Materials**

Rabbit globin cDNA, synthesized in vitro in the presence of actinomycin D by the method of Kacian et al. (8), was the kind gift of F. Ramirez, A. Gambino, A. Bank, and D. Kacian of Columbia University. The cDNA was synthesized in the presence of actinomycin D and at deoxynucleoside triphosphate concentrations that would promote synthesis of cDNA which was a complete copy of the mRNA template (28). The avian myeloblastosis virus reverse transcriptase was provided by the Office of Program Resources and Logistical, Viral Cancer Program, Viral Oncology, Division of Cancer Cause and Prevention, National Cancer Institute, Bethesda, Md. 20014.

_Escherichia coli_ RNA polymerase, isolated by the method of Burgess (29), was either obtained as gifts from C. Clibenson and D. Eisenberg, U.C.L.A., or from R. Burgess, University of Wisconsin, or was made in our laboratory by a newer Burgess procedure (30). u1 RNase was the kind gift of C. A. Decker (31, 32), University of California, Berkeley. U2 RNase and T1 RNase were purchased from Calbiochem, La Jolla, Calif.

**Methods**

**Chromatography and Electrophoresis**

Sepharose chromatography was used in some cases to obtain rapid approximate estimates of the lengths of synthesized RNAs (Sepharose 4B, Pharmacia). The column (0.6 x 12 cm) was made and chromatography performed by the same procedures as the Bio-Gel P-60 column used for desalting the RNA product as discussed in a later section.

Oligo(dT)-cellulose chromatography was carried out using grade T2 oligo(dT)-cellulose obtained from Collaborative Research. The chromatography procedure was the same as that described by Aviv and Leder (36) except that the intermediate elution step with 0.1 M KCI was omitted. The column (0.3 x 2 cm), equilibrated with 0.01 M Tris, pH 7.6, 0.5 M KCI, was packed tightly (flow rate less than 1 drop per min). The sample was made 0.5 M KCI, loaded onto the column, and washed with approximately 1 ml of the equilibrating buffer. Sequences containing poly(A) were then eluted with 0.01 M Tris, pH 7.6.

Bacteriophage T4 glicy1-tRNA used as a size marker in polyacrylamide gel electrophoresis was prepared according to Stahl et al. (37). It is a circular molecule long and the nucleotide sequence has been determined (37).

Polyethyleneimine (PEI)-cellulose plates were prepared as described by Whitcombe et al. (39). DEAE-cellulose plates were made according to Brownlee and Santer (40). Radioactivity of RNA fragments on DEAE-paper was determined in a Beckman LS-233 liquid scintillation counter, usually by Cerenkov counting.

**In Vitro RNA Synthesis Procedures**

Synthesis of $^{32}$P-labeled RNA—Most of the RNA syntheses were carried out in a low ionic strength buffer containing the following in $100 \mu$l final volume: 4 $\mu$mol of Tris, pH 7.9, 1 $\mu$mol of MgCl$_2$, 0.01 $\mu$mol of EDTA, 0.01 $\mu$mol of dithiothreitol, 0.04 $\mu$mol of potassium phosphate, pH 7.5, and 20 $\mu$mol of each nucleoside triphosphate with one labeled with $^{32}$P in the $\alpha$ position (New England Nuclear, specific activity 100 Ci/mmol), in some cases another labeled with $^{35}$S (Schwarz/Mann, specific activity around 20 Ci/mmol). The complete reaction mixture minus ATP, CTP, and UTP was preincubated with 10 $\mu$g of RNA polymerase, 1 $\mu$g of globin cDNA, 1 $\mu$g of DNA, 20 $\mu$mol of each ribonucleoside triphosphate, and 60 $\mu$mol of the selected deoxynucleoside triphosphate except in the case of the nucleoside triphosphate carrying the radioactive label. One to two milliliters of $^{32}$P-labeled nucleoside triphosphate with a specific activity of approximately 100 Ci/mmol was used per synthesis. Incubation was for 60 min at 37°C. In initial experiments, the nucleoside triphosphate precursors were not repurified prior to synthesis of deoxysubstituted RNA. In subsequent experiments, however, the nucleoside triphosphates were repurified in order to obtain completely deoxysubstituted RNA consistently (e.g. if d$\beta$-substituted RNA was desired, all remaining $\alpha$CTP had to be removed from the four TNP precursors, dCTP, dGTP, dATP, and dUTP).

Desalting of RNA Product—The labeled RNA product was heat-denatured at 100°C for 2 min and quick-cooled in ice. The mixture was centrifuged at 10,000 rpm for 100 min, and the supernatant was used as a probe.
Sequence of mRNA from Globin cDNA: T1 RNase Oligonucleotides

[...]

Sequence Determination Procedures

32P-labeled RNA was prepared for digestion with T1 ribonuclease or for combined digestion with T1 ribonuclease and bacterial alkaline phosphatase. Because of its freedom from DNase contamination, T1 ribonuclease was often used in place of T1 ribonuclease, especially when analysis of deoxysubstituted RNA was carried out. Carrier RNA (E. coli RNA) (20 µg) was added to the pooled fractions, usually 300 to 400 µl in volume. The material was evaporated to a smaller volume (10 to 30 µl) in a silarized tube by a combination of vacuum, vortex mixing, and heat. The sample was then transferred to a silarized polyethylene sheet and dried in a vacuum desiccator.

Enzymatic digests, electrophoresis, and other subsequent analysis and sequence determination procedures were carried out as described by Sanger and colleagues (11-13) with some additions and modifications. Alkaline hydrolysis of RNA fragments was carried out as described by Paddock and Abelson (41). Analysis of deoxysubstituted RNA fragments was carried out as in Paddock et al. (3). Alkaline hydrolysis of RNA fragments was carried out and analyzed by the following procedure: oligonucleotide samples were re-suspended in 10 µl of 0.2 M LiOH in capillaries and held overnight at 37° in a humiﬁer to minimize evaporation. The samples were then neutralized with 5 µl of 0.4 N HCl containing 10 µg each of GMP, AMP, UMP, and CMP. The neutralized samples containing unla-beled mononucleotides were then spotted on PEI-cellulose plates for chromatography as in Whitcombe et al. (39). The PEI-cellulose plates were washed in methanol tanks for 15 min to remove LiCl from the samples and the plates were then dried and chromatographed in a solvent system of 1 M acetic acid:1 M LiCl (25:1, v/v). Ultraviolet quenching was used to locate the four mononucleotide absorbance markers, and autoradiography was performed to locate radioactive alkaline hydrolysis products. Complete spleen nuclease digestion of deoxysubstituted RNA fragments was carried out as follows: the oligonucleotide sample was re-suspended in 10 µl of 0.1 M LiOH, 0.002 M EDTA, 0.0 M ammonium tartrate, and 0.05% Tween 80. Digestion was overnight at 37° in a humiﬁer. Ten micrograms of each nucleoside monophosphate (3 rNMPs plus the appropriate dNMP) were added and the mixture spotted on PEI-cellulose plates for analysis as above.

RESULTS

In Vitro RNA Synthesis Characteristics

The in vitro RNA synthesis reaction using globin cDNA as template has some unique features: (a) poly(A) synthesis exceeds the synthesis from the heteropolymeric region of the template, (b) the size of the RNA synthesized is relatively small (roughly 70 nucleotides) when the synthesis is performed in high salt, and larger (approaching template size) when synthesis is carried out in low ionic strength buffer.

Poly(A) Synthesis – The single-stranded cDNA template with its stretch of deoxythymidylic acid at the 5’ end poses unique problems for in vitro RNA synthesis. The incorporation of AMP is 5- to 10-fold more than that observed for other nucleotides (Fig. 1a), suggesting that considerable poly(A) is synthesized. This is substantiated by the presence of large pancreatic ribonuclease-resistant radioactive fragments (average size about 90 nucleotides) in digests of material synthesized with [α-32P]ATP as the radioactive precursor (Fig. 2). Synthesis of poly(A) can be reduced 2-fold by preincubation with GTP for a short period before adding the other nucleoside triphosphates (Fig. 1b), suggesting that a substantial fraction of the poly(A) is synthesized by polymerase molecules which initiate directly on the oligo(dT) sequence. This has been confirmed by experiments involving digestion with T1 RNase after the RNA synthesized in vitro was purified by oligo(dT)-cellulose chromatography (data not shown). These results indicate that even when preincubation with rGTP is used, 80% of the heteropolymeric product is not covalently attached to the poly(A) sequences.

In an attempt to decrease the poly(A) synthesis, the antibiotic rifampicin was added to the reaction mixture after preincubation with GTP but before the addition of the other nucleoside triphosphates. Since rifampicin is known to inhibit RNA chain initiations but not chain elongations with double-stranded DNA as template (42), it was thought that the addi-
tion of the antibiotic would limit initiations to those regions primed with GTP. But in the case of the *in vitro* RNA synthesis with globin cDNA template, rifampicin stopped RNA synthesis immediately, apparently inhibiting RNA chain elongation as well as chain initiation (Fig. 3). The discrepancy between this and previous results may lie with the difference in templates: globin cDNA is a single-stranded DNA, whereas previous work with rifampicin was done with double-stranded templates. A similar result has been observed with a different inhibitor by Walter et al. (43) who showed that heparin inhibited RNA chain initiations when double-stranded DNA was used as a template and all RNA synthesis when single-stranded DNA was used.

Ohasa and Tsugita (44) observed poly(A) synthesis in the absence of DNA template under certain conditions, due to the inherent activity of the α subunit of the *Escherichia coli* RNA polymerase. However, in our reaction conditions, no poly(A) synthesis is detected without template or when bacteriophage M13 DNA is used as a template. The poly(A) synthesis phenomenon we observe resembles that observed by Chamberlain and Berg (45), who reported a template-dependent synthesis of poly(A) from heat-denatured λ, T2, or calf thymus DNA when ATP is the only nucleoside triphosphate present. Unlike our results, they observed that the synthesis of poly(A) was strongly suppressed when other nucleoside triphosphates were added to the reaction mixture. They postulated that oligo(dT) tracts within the single-stranded DNAs were priming the synthesis of poly(A) longer than the oligo(dT) tracts due to slippage which would be limited when other nucleoside triphosphates were present. In our case, since the oligo(dT) tracts were located at the 5’ ends rather than the interior of the template molecules, the addition of other nucleoside triphosphates would not be expected to inhibit the synthesis of poly(A) according to the model of Chamberlin and Berg.

The presence of large amounts of poly(A) does not interfere with the fingerprint analysis unless [α-32P]ATP is the radioactive precursor. Fig. 4 shows fingerprints of combined ribonuclease T1 and bacterial alkaline phosphatase digests of *in vitro* synthesized globin RNA in which the radioactive precursors are [α-32P]GTP (Fig. 4a), [α-32P]UTP (Fig. 4b), and [α-32P]CTP (Fig. 4c). The fingerprints contain the graticule patterns expected of a combined ribonuclease T1 and alkaline phosphatase digest, and the unlabeled poly(A) does not pose any problems. When [α-32P]ATP is the precursor, we have found that further purification steps are necessary to remove poly(A) sequences since commercial bacterial alkaline phosphatase and bovine pancreatic ribonuclease preparations contain nuclease activities which cleave poly(A) sequences. Fig. 5, a and b, shows that such cleavages yield polydisperse [32P]-oligoadenylate which can mask the graticule pattern of standard fingerprints. Note that the ribonuclease T1 finger-
Sequence of mRNA from Globin cDNA: T1 RNase Oligonucleotides

Fig. 5. Two-dimensional electrophoresis of globin RNA labeled with \([\alpha-^{32}P]\)ATP. The \textit{in vitro} RNA synthesis from globin cDNA template was performed in high salt (0.15 M KCl). The RNA was purified as described under "Experimental Procedures." The following enzymatic digests were carried out in the absence of NaCl: a, combined RNase T1 and alkaline phosphatase; b, pancreatic ribonuclease; c, ribonuclease T1.

print (Fig. 5c) has faint spots resembling a regular RNase T1 graticule leaving the undigested \([^{32}P]\)poly(A) at the origin, indicating that this enzyme is highly pure and specific. By pretreating alkaline phosphatase with diethylpyrocarbonate (46) and by using pancreatic ribonuclease under high salt conditions (47), one can eliminate such degradations of poly(A) tracts. Alternatively, we have found it convenient to remove the poly(A) sequences by passing the enzymatic digests through an oligo(dT)-cellulose column (36). Fig. 6 shows a fingerprint of a combined RNase T1 and alkaline phosphatase digestion of \textit{in vitro} synthesized RNA labeled with \([\alpha-^{32}P]\)ATP and purified on an oligo(dT)-cellulose column. This step, of course, is unnecessary if the labeled precursor is any nucleoside triphosphate other than ATP.

\textbf{Influence of Salt—}The influence of KCl on the \textit{in vitro} synthesis of RNA with \textit{E. coli} RNA polymerase has been extensively studied. Most of the templates used in these studies are double-stranded, such as bacteriophage T4 DNA and \textit{E. coli} DNA, and the results of these studies indicate that KCl allows more extensive synthesis and reinitiation of synthesis (48). However, in the \textit{in vitro} RNA synthesis with single-stranded globin cDNA as template, we have found that smaller transcripts are produced at higher salt concentrations. For example, in the presence of 0.15 M KCl ("high salt"), the globin RNA synthesized \textit{in vitro} has an average length of 70 nucleotides (Fig. 7a), whereas when KCl is omitted a much larger product is obtained (Fig. 7b). Electrophoresis of \textit{in vitro} synthesized globin RNA with KCl omitted in a 10% acrylamide gel (data not shown) indicates that most of the RNA produced is in the size range of 70 to 300 nucleotides. Maitra \textit{et al.} (49) observed a similar salt effect on the product length of \textit{in vitro} synthesized RNA using denatured T4 DNA as a template.

It was hoped that the smaller transcripts synthesized in high salt might originate from a specific region of the cDNA template; however, fingerprints obtained from material synthesized in high and low ionic strength buffers were similar (data not shown).

\textbf{Two-Dimensional Fingerprints}

Figs. 4 and 6 show fingerprints of RNA synthesized \textit{in vitro} from globin cDNA with each of the four labeled precursors. The fingerprints are similar with expected differences due to the labeling (for example, due to the phosphatase treatment, the graticule containing no uridine residues should not be visible when \([\alpha-^{32}P]\)UTP is the radioactive precursor). Subsequent analysis of the fragments by various enzymatic cleavages shows that the fingerprint patterns are highly reproducible. This consistency extends over different syntheses from the same batch of cDNA and from cDNA preparations made at different times. This is a necessary prerequisite for any subsequent sequence determinations, since procedures involving labeling with one radioactive precursor at a time require that the experimenter be able to recognize corresponding fragments from different fingerprints with a high degree of confidence.

To ensure proper correlation of data between different experiments, we have also carried out experiments in which three of the four nucleoside triphosphates (rGTP, rUTP, and rCTP) were labeled. Fig. 8 shows the fingerprint of a U1 and bacterial alkaline phosphatase digest of such \(^{32}P\)-labeled RNA. The fingerprint has the characteristic pattern expected for
globin RNA and subsequent analysis of the oligonucleotides obtained from this digest confirmed sequences deduced from digests of RNA labeled with a single radioactive precursor as depicted in Figs. 4 and 6.

**Sequence Determinations of Fragments**

In the primary digestions of the globin RNA synthesized *in vitro* for sequence analysis, bacterial alkaline phosphatase was used in combination with ribonuclease T1 to remove terminal phosphates, thereby reducing the negative charge on each fragment. This provides better separation of the large fragments rich in uridine by increasing their electrophoretic mobilities in the DEAE-cellulose dimension.

The basic methods for determining the sequences of fragments in a ribonuclease T1 fingerprint have been well described (11-13). Generally, digestions with pancreatic ribonuclease, ribonuclease U2 and/or alkaline hydrolysis, and subsequent analysis sufficed to elucidate their primary structures. However, when runs of pyrimidines occurred partial digestions with spleen phosphodiesterase were frequently used. Partial digestion with carboxymethyllysine-41-pancreatic RNase (41) as well as deoxysubstitution techniques described (11-13) were used in combination with ribonuclease T1 to remove terminal phosphates, thereby reducing the negative charge on each fragment. This provides better separation of the large fragments rich in uridine by increasing their electrophoretic mobilities in the DEAE-cellulose dimension.

The basic methods for determining the sequences of fragments in a ribonuclease T1 fingerprint have been well described (11-13). Generally, digests with pancreatic ribonuclease, ribonuclease U2 and/or alkaline hydrolysis, and subsequent analysis sufficed to elucidate their primary structures. However, when runs of pyrimidines occurred partial digestions with spleen phosphodiesterase were frequently used. Partial digestion with carboxymethyllysine-41-pancreatic RNase (41) and deoxysubstitution techniques described in Paddock *et al.* (3) were used to determine the more difficult sequences. Using these techniques, nearly all of the fragments resulting from combined ribonuclease T1 and alkaline phosphatase digestions have been sequenced.

Fig. 9 is a line drawing of a fingerprint labeled with [α-32P]CTP, [β-32P]UTP, and [γ-32P]CTP (see Fig. 8), with numbers assigned to each fragment spot. Several fragments overlap on the fingerprint and were repurified by one-dimensional homochromatography on DEAE-cellulose thin layer plates (40) before secondary digestions were performed. With this procedure, Fragments 26 and 29 can be resolved, for example, and Spot 58 separates into three distinct fragments.

Table II gives the data obtained by the techniques described. Although most fragment sequences have been determined, the sequence data for Fragment 58b remain ambiguous, and a few other fragments are of such low yield that their sequences probably will not be determined (see Table I). For many of the longer sequences, especially those obtained with partial spleen digestions, confirmation has been obtained from deoxysubstitution data. Because of many sequence similarities among the fragments, we have been able in many instances (as noted in Table II) to use small fragments as markers for identifying the products of digestion of larger fragments. For example, Fragment 4, C-C-U-G, was used as an electrophoretic mobility marker to allow an unambiguous interpretation of the results of a U2 digest of Fragment 14, C-C-A-C-C-U-G; Fragment 17, U-C-C-U-G, was used as a marker for a partial spleen digest of Fragment 42, U-C-C-U-C-U-G; and pancreatic digests of deoxy-C-substituted Fragments 4, C-C-U-G, and 7, C-C-C-U-G, were used as markers for a pancreatic digest of deoxy-C-substituted Fragment 26, C-C-U-C-C-C-U-G. For most of the fragments, we also electrophoresed the U2 digests and partial spleen digest of fragments from the different labeling experiments side by side in at least one experiment.

Most of the sequence determinations using the data in Table II were quite straightforward. However, the derivation of the sequences for the oligonucleotides contained in Spot 34 was an exception. The derivations are displayed in Fig. 10, which shows the sequence overlaps that can be obtained from the larger digestion products for Spot 34. From its position in the fingerprint we can deduce that all of the oligonucleotides of fragments 26 and 29 can be resolved, for example, and Spot 58 separates into three distinct fragments.
Sequence of mRNA from Globin cDNA: T1 RNase Oligonucleotides

Fig. 8 (left). Two-dimensional electrophoresis of a U1 RNase plus bacterial alkaline phosphatase digest of globin RNA labeled with $\alpha$-32PdGTP, $\alpha$-32PdUTP, and $\alpha$-32PdCTP as described under "Experimental Procedures." Fig. 9 (right). Line drawing of fingerprint. Line drawing of Fig. 8. Numbers identify the fragments, most of which have been sequenced and are shown in Table I. Nonphosphatased T1 digestion fragments appear in our fingerprints in low yield and are not included in the diagram.

Table I

<table>
<thead>
<tr>
<th>Spot No.</th>
<th>Sequences</th>
<th>Spot No.</th>
<th>Sequences</th>
<th>Spot No.</th>
<th>Sequences</th>
<th>Spot No.</th>
<th>Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GUG</td>
<td>20</td>
<td>GUAUG</td>
<td>34b</td>
<td>GCAUCAUAG</td>
<td>54</td>
<td>GCUUACUCUCAG</td>
</tr>
<tr>
<td>2</td>
<td>GCUG</td>
<td>21a</td>
<td>GCUCUCUG</td>
<td>34c</td>
<td>GAAUCAUCAG</td>
<td>55</td>
<td>GCAAAAAUUAUG</td>
</tr>
<tr>
<td>3</td>
<td>GAUG</td>
<td>21b</td>
<td>GUCUCUG</td>
<td>34d</td>
<td>GAUCAUCUG</td>
<td>57</td>
<td>GAAUCAUCUCUCAG(c)</td>
</tr>
<tr>
<td>4</td>
<td>GCCUG</td>
<td>22</td>
<td>GCUCUCUG</td>
<td>35</td>
<td>GCUCUCUCUCAG</td>
<td>58a</td>
<td>GAAUUCUUUCUUAUG(3)</td>
</tr>
<tr>
<td>5a</td>
<td>GACUG</td>
<td>23</td>
<td>GAUUG</td>
<td>36</td>
<td>GUAUUCUCUG</td>
<td>58b</td>
<td>(Py-)GCAUCAUCUCUCAG(t)</td>
</tr>
<tr>
<td>5b</td>
<td>GCAUG</td>
<td>24a</td>
<td>GCAUG</td>
<td>37</td>
<td>GUUUG</td>
<td>58c</td>
<td>GCUUACUCUCUCUCAG(s)</td>
</tr>
<tr>
<td>5c</td>
<td>GAUCUG</td>
<td>24b</td>
<td>GAUCUG</td>
<td>38</td>
<td>GUAUUG</td>
<td>61</td>
<td>GCG</td>
</tr>
<tr>
<td>6</td>
<td>GAUGUAG</td>
<td>25</td>
<td>GAUCUCUG</td>
<td>39</td>
<td>GUCUCUCU</td>
<td>62</td>
<td>GAAG</td>
</tr>
<tr>
<td>7</td>
<td>GCCUCUG</td>
<td>26</td>
<td>GCCUCUCUG</td>
<td>40</td>
<td>GUCUCUC</td>
<td>63</td>
<td>GCCG</td>
</tr>
<tr>
<td>8</td>
<td>GCCACUG</td>
<td>27a</td>
<td>GCCACUG</td>
<td>42</td>
<td>GCUCUCU</td>
<td>64</td>
<td>GCAG</td>
</tr>
<tr>
<td>9</td>
<td>GACUGU</td>
<td>27b</td>
<td>GCUCUGU</td>
<td>43</td>
<td>GCACUCUG</td>
<td>65</td>
<td>GACG</td>
</tr>
<tr>
<td>10</td>
<td>GAUCUGU</td>
<td>27c</td>
<td>GAUCUGU</td>
<td>44</td>
<td>GCUCUCUGU</td>
<td>66</td>
<td>GAGG</td>
</tr>
<tr>
<td>11</td>
<td>GCCACUGU</td>
<td>28</td>
<td>GCCACUCUGU</td>
<td>45</td>
<td>GCAUCAUCUGU</td>
<td>67</td>
<td>GCCGG</td>
</tr>
<tr>
<td>12</td>
<td>GCAUCUGU</td>
<td>29</td>
<td>GCAUCUGU</td>
<td>47</td>
<td>GACUCAUCUGU</td>
<td>68a</td>
<td>GCGG</td>
</tr>
<tr>
<td>13</td>
<td>GCCUCUCCG</td>
<td>30a</td>
<td>GCCUCUCCG</td>
<td>49</td>
<td>GUCUCAUUG</td>
<td>68b</td>
<td>GCCG</td>
</tr>
<tr>
<td>14</td>
<td>GCCACUCUG</td>
<td>30b</td>
<td>GCCACUCUG</td>
<td>50</td>
<td>GUAUCAUUG</td>
<td>68c</td>
<td>GCCG</td>
</tr>
<tr>
<td>15</td>
<td>GCAAAUGU</td>
<td>31</td>
<td>GCCAUCUCUG</td>
<td>51</td>
<td>GUCUCUCUG</td>
<td>69</td>
<td>GCAAG</td>
</tr>
<tr>
<td>16</td>
<td>GACUGU</td>
<td>33</td>
<td>GCCUCUCUG</td>
<td>52</td>
<td>GUCUCUCUG</td>
<td>70</td>
<td>GAAGA</td>
</tr>
<tr>
<td>17</td>
<td>GCCUGU</td>
<td>34a</td>
<td>GCAUCAUAG</td>
<td>53</td>
<td>GUCUCUCUG</td>
<td>71</td>
<td>GCAAG</td>
</tr>
</tbody>
</table>

Spot 34 should have the base composition 3Ap, 2Up, 2Cp, and 1 G-OH. Therefore, modified pancreatic RNase product 19, A-A-C-U-U-Cp(Ap), cannot be derived from the same oligonucleotide as either product 14, (Py-)A-C-C-U-G, or product 18, (Py-)A-U-C-A-U-G; combining 19 with either 14 or 18 would yield a fragment with 3 or more uridine residues which could not run in this position on the fingerprint. In the same way we may conclude that product 18, A-A-U-Cp(Ap), and products 18 and 19 are derived from different oligonucleotides. Therefore, we may conclude that the number of oligonucleotides present in Spot 34 is greater than three unless one such oligonucleotide has the sequence A-A-U-C-A-C-C-U-G. This latter sequence may be eliminated because it has the wrong number of cytidine residues and, more importantly, because it is the wrong size: all of the components of Spot 34 are octanucleotides as shown by their co-migration upon homochromatography on a
DEAE-cellulose thin layer plate, and the overlap data summarized in Fig. 10 and discussed below clearly establish the sequences.

The most significant overlaps for Spot 34 are as follows. The four and one-half nucleotide overlap between products 17 and 18, establishes the sequence for oligonucleotide 34a. The three nucleotide overlap between products 6 and 19 establishes the sequence for oligonucleotide 34c. Oligonucleotide 34d is defined by the three and one-half nucleotide overlap between products 15 and 7 and the two base overlap between products 15 and 2. Admittedly we are relying on the nucleotide composition constraints to place product 2 in Fragment 34b instead of into Fragment 34d. That the nucleotide composition we had deduced by fingerprint position is correct however, is shown by products 9 and 12 which demand that Gp is the 5' neighbor of the sequence we have deduced for Fragment 34b which is defined by the three nucleotide overlap between spots 13 and 14. This leaves only the sequence determination for Fragment 34b which is defined by the three nucleotide overlap between products 15 and 7 and the two base overlap between products 15 and 2. Admittedly we are relying on the nucleotide composition constraints to place product 2 in Fragment 34b instead of into Fragment 34d. That the nucleotide composition we had deduced by fingerprint position is correct however, is shown by products 9 and 12 which demand that Gp is the 5' neighbor of the sequence we have deduced for Fragment 34b. The remaining data displayed in Fig. 10 and shown in Table II are also supportive of the four sequences that were determined.

An excellent example of the power of the deoxysubstitution technique is provided by the ease with which the method solves the sequence of Fragment 26, C-C-U-C-C-R-G. It proved to be very difficult to obtain the correct partial digestion conditions for analysis with an exonuclease such as spleen phosphodiesterase due to the high number of cytidine residues. However with deoxysubstitution, the sequence could be determined in a single experiment. When the 32P label was introduced in [α-32P]dCTP, pancreatic RNase digestion of this fragment (digestion only after Up is now possible) gives products C-C-U and C-C-U. Subsequent analysis with spleen phosphodiesterase shows that only C-C-U has label in Up. Thus C-C-U must be the 3' neighbor of C-C-U. Confirmation of the sequence is provided in a dC-substitution experiment with 32P label introduced in [α-32P]dCTP. Pancreatic RNase digestion results in only C-C-C-U appearing with radioactive label.

Several T1 RNase fragments separated in the second dimension on DEAE-cellulose thin layer plates by homochromatography have also been analyzed. The fingerprint of T1-digested globin RNA separated in this manner is shown in Fig. 11 and the analysis is presented in Table III. The analyses of these fragments are consistent with the data for phosphatase-treated fragments already discussed and, in addition, have allowed us to determine the nucleotides adjacent to the 3' ends for some of the fragments with "nearest neighbor" analysis. These nearest neighbor assignments are shown in Table I.

**Amino Acid Correspondence of Nucleotide Fragments**

We have devised a computer program which converts a given nucleotide sequence into all possible amino acid sequences and searches for possible fits within the α- and β-globin chains (each of the three possible reading frames are tried and partially specified codons at each end of the fragment are filled in with all possible sequences so that all possible fits will be detected). As expected, nucleotide sequences up to six nucleotides in length usually have possible fits at more than one position, while sequences eight nucleotides or longer usually fit the amino acid sequences at a unique location or not at all (10). When any fragment is found to have unique fits, it can be stated that either it comes from the corresponding position in the mRNA, or if the single observed fit is fortuitous, then the fragment must be from the untranslated regions of the mRNAs. As the fragment sequences become longer and longer, the likelihood of "chance" fit becomes less. As has been shown previously (10), most of the sequences of eight nucleotides or longer fit into single locations in the globin amino acid chains or can be assigned to the untranslated regions. Among them are the sequences for Spots 13, 26, 31, 47, 48, and 54 which fit into the α chain, and Spots 30b, 33, 34c, 35, 45, 52, 53, 57, and 58c which fit into the β chain (Table IV). We were able to assign unambiguously eight fragments of eight nucleotides or longer to the untranslated regions on this basis (Table V).

Among the assignments thus made, that for Spot 48 (A-A-U-U-U-C-A-A-G) deserves special comment. It fits the amino acid sequence valine-asparginine-phenylalanine-lysine beginning at position 96 of the α chain, but it must be noted that experiments using isoacceptor lysyl-tRNAs have led to some.
controversy as to whether the lysine (position 99) is coded by AAA or AAG (50, 51). If the lysine is indeed coded by AAA as suggested by Woodward and Herbert (50), then this fragment may actually have come from the untranslated region of the α chain. On the other hand, the discrepancy could also have arisen from possible differences in strains of rabbits used in the previous experiments.

As can be seen from Table IV, the fragment fits are clustered in a region starting at amino acid position 80 in the (γ chain and at amino acid 40 in the β chain. Through the termination codon for each chain, the gaps between fits average seven amino acids for the α chain and eight amino acids for the β chain, with a range of 0 to 15 for each chain. Unless the nucleotide composition is grossly different in the regions where fits have been found and in those regions where there are not fits (i.e., the proportion of guanosine is much higher in the no-fit region resulting in T1 fragments too small for unique fits), we must assume that our in vitro transcription yields in the no-fit region must be too low to be seen in a fingerprint of a T1 digest. Fig. 7b provides support for the hypothesis that we are not transcribing the entire cDNA in yields sufficient to sequence. While some of RNA synthesized under low-ionic strength conditions appears as long as the cDNA, a majority of the RNA fragments appear much shorter. As mentioned earlier, we have also electrophoresed globin RNA synthesized in vitro on acrylamide gels and found a size range of 70 to 300 nucleotides. That much fragments are not of full length suggests that the "no-fit" regions may be under-transcribed in our system. We have also searched the possible nucleotide sequences for these "no-fit" regions using codons for the known amino acids. Several large T1 fragments are obligatory, yet their sequences do not exist in our catalogue of fragments in Table I. This is in contrast to the data communicated to us by Weissman3 for the human globin mRNAs, for which sequences appearing to represent the entire translated portions of the α chain have been detected. At the present time, we have no good explanation for the difference in results; a difference in the conformation for the cDNAs synthesized from the human and rabbit α-mRNAs could account for the varying results. Transcription from human β-cDNA appears to be more similar to that of rabbit with most RNA synthesis commencing near amino acid position 40. However, weak transcription is also observed from human cDNA for regions corresponding to earlier amino acid positions.4

As shown in Table V, we have also assigned eight fragments to the untranslated regions of the mRNAs. In consideration of the clustering evidence displayed in Table IV, these have been assigned to the untranslated regions at the 3' ends of the 2 molecules. Comparison of data with that obtained by Proudfoot5 for the 3' ends of the mRNAs (52, 53) shows that Fragments 29, 34A, 36, 53, and 58A fit into the β chain 3' untranslated region and none of the fragments in Table V fit into the α chain 3' untranslated region. A comparison with the Proudfoot sequences for the β chain 3' untranslated region show that we have found all the predicted T1 fragments (Table I), including 58A which is at the 3' end just prior to the poly(A) region (52). The remaining Fragments 34B, 34D, and 358B in Table V not yet assigned probably belong in the β' untranslated regions not yet sequenced by Proudfoot. Thus, it would appear that we

---

3 S. Weissman, personal communication of unpublished results.
4 N. Proudfoot, personal communication of unpublished results.
Sequence of mRNA from Globin cDNA: T1 RNase Oligonucleotides

The nucleotide sequences we have determined for the T1 RNase digestion fragments represent completion of the first stage in our effort to sequence the rabbit globin mRNAs. We have had to employ a variety of novel techniques to determine this catalogue of sequences. The most useful of these approaches was deoxysubstitution (3, 4), in which a deoxynucleotide is substituted for the corresponding ribonucleotide in our in vitro synthesized RNA, thus specifically blocking certain cleavages and gaining more specificity for nucleases such as pancreatic RNase.

With the aid of a computer programmed with the genetic code and the globin amino acid sequences we have been able to position uniquely fragments of eight or more nucleotides. This has allowed us to test for the fidelity of transcription of avian myeloblastosis virus reverse transcriptase and Escherichia coli RNA polymerase. As discussed in our preliminary paper (10) and strongly confirmed by the additional detailed data presented here, we were able to determine from our T1 digestion data that there were neither systematic errors nor random errors through the two transcription steps with these enzymes. Moreover, this high degree of fidelity is retained even when deoxynucleotides are substituted for transcription by the RNA polymerase in the second step.

We have been able to position 15 fragments defining a total of 138 nucleotides into the translated regions of the two globin mRNAs. While this represents significant progress, much of the globin mRNA's sequence is still not represented either because of gaps between the fragment fits or because of regions not transcribed from the cDNA into 32P-labeled RNA. However, even with only these partial data, through comparison of our sequences for rabbit sequences with those obtained by Marotta et al. (14) for the human globin mRNAs, we have determined that mutations that do not affect changes in amino acids in the translated regions of the mRNAs are accepted at a 9-fold higher rate than are mutations that do affect the amino acid sequence. (See Ref. 54 for a more detailed discussion of this point. As discussed in Ref. 55, these results suggest that only a small portion of the single copy DNA in the cell can be accurately maintained as functional genes.)

We have determined that eight sequences clearly do not fit in the translated regions of the mRNA. Five of these sequences coincide precisely with data obtained by Proudfoot for a 76-nucleotide sequence adjacent to the poly(A) at the 3' end of the β chain. From our Fragment 58a it is clear that E. coli RNA polymerase transcribes the entire untranslated region of the β-globin mRNA sequence. On the other hand, we appear to be missing the sequences predicted by Proudfoot's data for the untranslated portion of the α chain. The simplest interpretation of those results is that E. coli RNA polymerase does not transcribe the untranslated region at the 3' end of the α chain and that Fragments 34b, 34d, and 58b must fall into that portion of the untranslated region of the β chain for which Proudfoot has no data. Comparison of the two sets of data indicates that the untranslated sequence adjacent to the poly(A) at the 3' end of the β chain must be approximately 160 nucleotides in length. From current estimates of the sizes of the β-globin mRNA and its poly(A) sequences and from the known size of the structural gene and the above estimate of the 3' untranslated sequence we may tentatively conclude that there is an untranslated sequence of substantial size, perhaps more than 50 nucleotides, at the 5' end of the mRNA.

If one is ultimately to sequence the globin mRNA molecules in their entirety, techniques must be developed to isolate specific smaller regions to provide overlap data needed to order the ribonuclease T1 fragments. Unfortunately, partial RNase T1 digests of the in vitro synthesized globin RNA have been unsuccessful, perhaps due to the lack of a specific secondary structure of the RNA transcript. Partial digestion with carboxymethyllysine-41-pancreatic RNase yielded products in a useful size range but the task of separating the many fragments is difficult. However, the relatively small gap size between fits for the T1 fragments suggested that if we could obtain primary cleavages at nucleotides other than guanosine, those gaps could be filled in. In order to fill in the gaps between fits, we and others developed the deoxysubstitution technique in order to obtain these other specific cleavages (3, 4), and we are currently carrying out such analyses.

The only apparent drawback to the use of reverse transcriptase for sequencing eukaryotic mRNAs is the possibility of incomplete copying, either during the synthesis of the cDNA or during the subsequent transcription of the cDNA with RNA polymerase. In either case, it is the sequences corresponding to the 5' end of the mRNA which are most likely to be omitted. It has been found (29) that reverse transcriptase does not always make cDNA as long as the mRNA template. However, our rabbit globin cDNA is a full length or nearly full length copy (56) and so cannot explain the clustering of T1 fragment fits which must instead indicate that the RNA polymerase does not transcribe a complete copy of the cDNA. This is in contrast to the success observed by Weissman and collaborators for fairly complete RNA transcription from human cDNA.

We have been able to overcome the problem by the insertion of the cDNA copy of mRNA into bacterial plasmids (57). Such inserts which have also been made by others (58-60) can be sequenced using the powerful new DNA sequencing techniques now available. It appears therefore that the complete sequence determination of eukaryotic mRNAs is becoming a realistic possibility.

Acknowledgments We are indebted to our many colleagues who supplied us with the enzymes needed for this work, and are especially thankful to Drs. D. Kacian, R. Gambino, F. Ramirez, and A. Bank who supplied us with the rabbit globin cDNA during the early stages of this work. We are very appreciative of the helpful discussions with Drs. N. Proudfoot and S. Weissman regarding their nucleotide sequence determinations. We also would like to thank undergraduates D. Brown, R. Main, R. Saltzman, R. Bronte, R. Friedman, and laboratory technicians M. Childs, S. Chu, and V. Jones for helping us in our work. Finally, we thank Kay Ayres and Georgia Rahn for preparing the manuscript. We thank Bruce Wallace for preparing the final photographs.

Note Added in Proof—We have recently published the sequence of a 169 nucleotide region of the rabbit β-globin mRNA obtained by analysis of cloned cDNA sequences (62). This brings to 228 the number of β chain mRNA nucleotides we have unambiguously assigned exclusive of the 76 nucleotides assigned by Proudfoot (53).

A. Maxam and W. Gilbert, personal communication (cf. Ref. 61).
Rabbit globin mRNA: analysis of T1 RNAse digestion fragments.
G V Paddock, R Poon, H C Heindell, J Isaacson and W Salser


Access the most updated version of this article at http://www.jbc.org/content/252/10/3446

Alerts:
• When this article is cited
• When a correction for this article is posted

Click here to choose from all of JBC's e-mail alerts

This article cites 0 references, 0 of which can be accessed free at http://www.jbc.org/content/252/10/3446.full.html#ref-list-1