Rabbits, T. H., Forster, A., Smith, M., & Gellam, S. (1977) *Eur. J. Immunol. 7*, 43–48.
Rabbits, T. H., Forster, A., Dunnick, W., & Bently, D. L. (1980) *Nature (London) 283*, 351–356.
Rigby, P. W. J., Dieckmann, M., Rhodes, C., & Berg, P. (1977) *J. Mol. Biol. 113*, 237–251.
Sakano, H., Hüppi, K., Heinrich, G., & Tonegawa, S. (1979) *Nature (London) 280*, 288–294.
Seidman, L. G., & Leder, P. (1978) *Nature (London) 276*, 790–795.
Shepherd, J., Mulvihill, E., & Palmiter, R. (1977) *J. Cell Biol. 75*, 353a.
Southern, E. M. (1975) *J. Mol. Biol. 98*, 503–517.
Stalder, J., Groudine, M., Dogson, J. B., Engel, J. D., & Weintraub, H. (1980) *Cell 19*, 973–980.
Stavnezer, J., Huang, R. C., Stavnezer, E., & Bishop, J. M. (1974) *J. Mol. Biol. 88*, 43–63.
Steinmetz, M., & Zachau, H. G. (1980) *Nucleic Acids Res. 8*, 1693–1706.
Storb, U., Hager, L., Wilson, R., & Putnam, D. (1977) *Biochemistry 16*, 5432–5438.
Storb, U., Arp, B., & Wilson, R. (1980) *Nucleic Acids Res. 8*, 4681–4687.
Walfield, A., Storb, U., Selsing, E., & Zentgraft, H. (1980) *Nucleic Acids Res. 8*, 4689–4707.
Warner, N., Harris, A., & Gutman, G. (1975) in *Membrane Receptors of Lymphocytes* (Seligman, M., Prud'homme, J. L., & Kourilsky, I. M., Eds.) pp 203–220, Elsevier, New York.
Weintraub, H., & Groudine, M. (1976) *Science (Washington, D.C.) 193*, 848–856.
Wilson, R., Miller, J., & Storb, U. (1979) *Biochemistry 18*, 5013–5021.
Wilson, R., Storb, U., & Arp, B. (1980) *J. Immunol. 124*, 2071–2076.
Wu, C., Bingham, P. M., Livak, K. J., Holmgren, R., & Elgin, S. C. R. (1979) *Cell 16*, 797–806.

Sequence Determination and Analysis of the 3′ Region of Chicken Pro-α1(I) and Pro-α2(I) Collagen Messenger Ribonucleic Acids Including the Carboxy-Terminal Propeptide Sequences†

Forrest Fuller‡ and Helga Boedtker*

ABSTRACT: Three pro-α1 collagen cDNA clones, pCg1, pCg26, and pCg54, and two pro-α2 collagen cDNA clones, pCg13 and pCg45, were subjected to extensive DNA sequence determination. The combined sequences specified the amino acid sequences for chicken pro-α1 and pro-α2 type I collagens starting at residue 814 in the collagen triple-helical region and continuing to the procollagen C-termini as determined by the first in-phase termination codon. Thus, the sequences of 272 pro-α1 C-terminal, 260 pro-α2 C-terminal, 201 pro-α1 helical, and 201 pro-α2 helical amino acids were established. In addition, the sequences of several hundred nucleotides corresponding to noncoding regions of both procollagen mRNAs were determined. In total, 1589 pro-α1 base pairs and 1691 pro-α2 base pairs were sequenced, corresponding to approximately one-third of the total length of each mRNA. Both procollagen mRNA sequences have a high G+C content. The pro-α1 mRNA is 75% G+C in the helical coding region sequenced and 61% G+C in the C-terminal coding region while the pro-α2 mRNA is 60% and 48% G+C, respectively, in these regions. The dinucleotide sequence pCG occurs at a higher frequence in both sequences than is normally found in vertebrate DNAs and is approximately 5 times more frequent in the pro-α1 sequence than in the pro-α2 sequence. Nucleotide homology in the helical coding regions is very limited given that these sequences code for the repeating Gly-X-Y tripeptide in a region where X and Y residues are 50% conserved. These differences are clearly reflected in the preferred codon usages of the two mRNAs.

Collagen is a fibrillar, structural protein responsible for the physical integrity of organs, tissues, and gross skeletal structures of vertebrates and of at least some invertebrates as well (Adams, 1978). The structures and functions of collagens have been recently reviewed (Fessler & Fessler, 1978; Prockop, et al., 1979a; Bornstein & Byers, 1980; Eyre, 1980; Bornstein & Sage, 1980; Olsen & Berg, 1979). Analysis of polypeptide chains implies that higher animals produce at least five different collagens from at least seven genes. Various cells synthesize different collagens and modify them to produce structural matrices specific to the cell type. Consequently, collagen expression is carefully regulated during differentiation and embryogenesis (von der Mark et al., 1976; Linsenmayer & Toole, 1977; Bornstein & Sage, 1980). Failure to correctly express or modify specific collagens has been implicated as the causative factor in several diseases of man and other animals (Prockop et al., 1979b; Bornstein & Byers, 1980). In order to understand how the cell regulates expression of these different collagens, it is necessary to determine the levels and rates of synthesis of collagen mRNA[1] and pre-mRNAs at various stages during development and in both normal and abnormal cells. It is also important to define the organization of the collagen sequences in the genome to ascertain how this

† From the Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, Massachusetts 02138. *Received June 27, 1980.* Supported by Grants from the National Institutes of Health (HD-01229) and the Muscular Dystrophy Association, Inc.

* To whom reprint requests should be addressed.
‡ To whom requests for sequence data and computer programs should be addressed.

[1] Abbreviations used: cDNA, complementary DNA; mRNA, messenger RNA; poly(A), poly(adenylic acid); mC, 5-methyldeoxycytidine; I, inosine; helical, a peptide (or nucleotides coding for peptide) characterized by the repeating sequence $(Gly\text{-}X\text{-}Y)_n$; C-terminal, amino acids (or nucleotides coding for them) which occur in the carboxyl ends of procollagens and do not contain the repeating sequence, $(Gly\text{-}X\text{-}Y)_n$.

0006-2960/81/0420-0996$01.00/0

family of genes evolved and how the gene structure may relate to expression.

To achieve these goals, cDNAs complementary to chicken cranial (type I) procollagen mRNAs have been synthesized and cloned (Lehrach et al., 1978, 1979; Sobel et al., 1978; Yamamoto et al., 1980). Here we report DNA sequences from five such clones previously isolated and characterized in this laboratory (Lehrach et al., 1978, 1979). The nucleotide sequences specify the complete pro-α1 and pro-α2 carboxy-terminal propeptide and telopeptide sequences. The C-terminal propeptides have proven refractory to conventional protein sequencing due to their rapid cleavage and proteolysis during collagen maturation prior to fibrinogenesis (Fietzek & Kühn, 1976; Fessler & Fessler, 1978). In addition, the residues in the pro-α2 cyanogen bromide peptide CB5 which were not previously determined (Dixit et al., 1979) have now been specified. The combined sequences reported here correspond to about one-fifth of the triple-helical coding region, all of the C-terminal coding region, and probably all of the 3′ non-translated sequences of two chicken collagen mRNAs.

Materials and Methods

Preparation of DNA for Sequencing. Recombinant plasmids, pCg1, pCg26, pCg54, pCg13 and pCg45, were amplified in *E. coli* K12 strain χ1776 or HB101 and the DNAs were isolated as previously described (Lehrach et al., 1979).

Enzymes. Restriction enzymes *Alu*I, *Ava*I, *Bam*HI, *Hae*III, *Hin*fI, *Hpa*II, *Kpn*I, *Pst*I, and *Sau*3A and DNA polymerase I were obtained from Bethesda Research Laboratories. *Eco*RI was obtained from New England Biolabs. *Hin*dIII was a gift from John Wozney and T4 polynucleotide kinase was a gift from H. Lehrach.

DNA Sequence Determination. DNA sequencing was carried out essentially as described by Maxam & Gilbert (1980). Restriction fragments were labeled at 5′ ends with γ-[^{32}P]ATP and T4 polynucleotide kinase or at 3′ ends with α-[^{32}P]dCTP and α-[^{32}P]dTTP and *E. coli* DNA polymerase I. Radionucleotides were purchased from Amersham Corp. Labeled fragments were either restricted with a secondary restriction enzyme or subjected to strand separation. Polyacrylamide gels, 0.45 mm thick by 40 cm long, were typically employed and were run at voltages between 1200 and 1800 V. Usually two 10% gels and one or two 6% gels were used for each determination. This allowed reading to commence at base 3 to base 14 and to continue, in the best case, to base 325. Where determination of the first few nucleotides was desired, a 20%, 1.5-mm-thick gel was employed in order to minimize salt effects. Sequences were compiled separately by hand and finally transcribed and concatenated on two long graph papers. The final sequence was independently checked against the original autoradiographs and compiled on an Apple II Plus microcomputer programmed for sequence analysis.

Containment. All bacteria containing recombinant plasmids were grown in an approved P2 physical containment laboratory in compliance with the National Institutes of Health Guidelines for Recombinant DNA Research (NIH Guidelines, 1976), as revised (NIH Guidelines, 1978).

Results and Discussion

DNA Sequence Determination. A primary concern in reporting a nucleotide sequence is the accuracy of the determination, particularly when a protein sequence is to be derived from the nucleotide sequence alone. Sources of error in such determinations are twofold: gel artifacts or poorly carried out reactions which result in an incorrect interpretation of the sequence autoradiograph and the unfaithful reproduction of mRNA sequences during either cDNA synthesis or subsequent amplification in *E. coli*. To eliminate errors of the former type, we have determined sequences multiple times, on both strands, and have required that autoradiographs be interpreted by two readers. Errors of the latter type have been circumvented by determining sequences from overlapping clones.

Restriction maps of the five cDNA clones have been previously published (Lehrach et al., 1978, 1979). Figures 1 and 2 depict restriction maps which correspond to the pro-α1 and pro-α2 mRNA sequences, respectively. The sequence contained in each cloned cDNA insert which corresponds to the mRNA sequences is indicated by the heavy lines above the restriction map. For reasons discussed below (see Sequence Rearrangements in cDNA Clones) we have depicted pCg26 and pCg54 in each of two possible orientations with respect to the mRNA sequence. It may be helpful initially to consider each of these as a distinct clone (i.e., a total of five pro-α1 cDNA clones). Individual sequence determinations[2] on each clone are depicted by the arrows. The only sequences which have not been determined on both DNA strands are pro-α1 sequences between −393 and −179 and between −142 and −20. As indicated in Figure 1, all but about 130 nucleotides in the first region have been sequenced more than once. Moreover, autoradiographs corresponding to these regions are very clear and have been identically interpreted by two readers. The protein sequence dictated by these sequences is identical with that determined for chicken skin α1(I) collagen (Fietzek & Kühn, 1976; Dixit et al., 1978; J. H. Highberger et al., unpublished results).

It is also notable that although we have been able to determine the sequence of a DNA strand to the first base when desired, we have relied on such a sequence only for the first few nucleotides of the *Hin*dIII ends of the cloned DNAs. The seven 5′-terminal nucleotides in such sequences are defined by the synthetic DNA (pCCA↑pAGCTTGG) which has been artificially attached to the end of the cDNAs. In order to ensure that small restriction fragments have not been accidentally omitted from the sequence and to eliminate misinterpretations of autoradiographs often caused by "salt" effects on the mobilities of small fragments, we defined all the sequences around internal restriction sites by sequence determinations extending across those sites.

The precaution of sequencing both strands of pro-α2 helical coding regions was necessitated by the frequent occurrence of *Eco*RII sites in this region. The second C in such sequences (pCCA/TGG) is methylated in many *E. coli* strains and does not appear on the sequence autoradiograph (Ohmori et al., 1978). Although mC's may be defined by the lack of a band in all lanes of the autoradiograph, we find bands of variable intensities in the pyrimidine lanes at these positions. Therefore it appears likely that the large number of sites saturates the methylase activity, resulting in partial methylation. Variation in the effect of adjacent sequences on methylase affinity could then explain the apparent variability in the extent of methylation.

Finally, it is important to point out that S1 protection experiments (Lehrach et al., 1979) have shown that the majority of cloned sequences are totally complementary to calvaria

[2] Sequence coordinates: coordinate numbers refer to bonds between nucleotides on the coding strand. Coordinates of restriction sites refer to the bond cleaved by the restriction endonuclease. Specific nucleotides are referred to by the coordinate 3′ to the nucleotide. The bond between the last helical coding nucleotide and the first C-terminal coding nucleotide is assigned the zero coordinate. Thus, numbers less than one correspond to helical coding regions and numbers greater than zero correspond to C-terminal coding and noncoding regions.

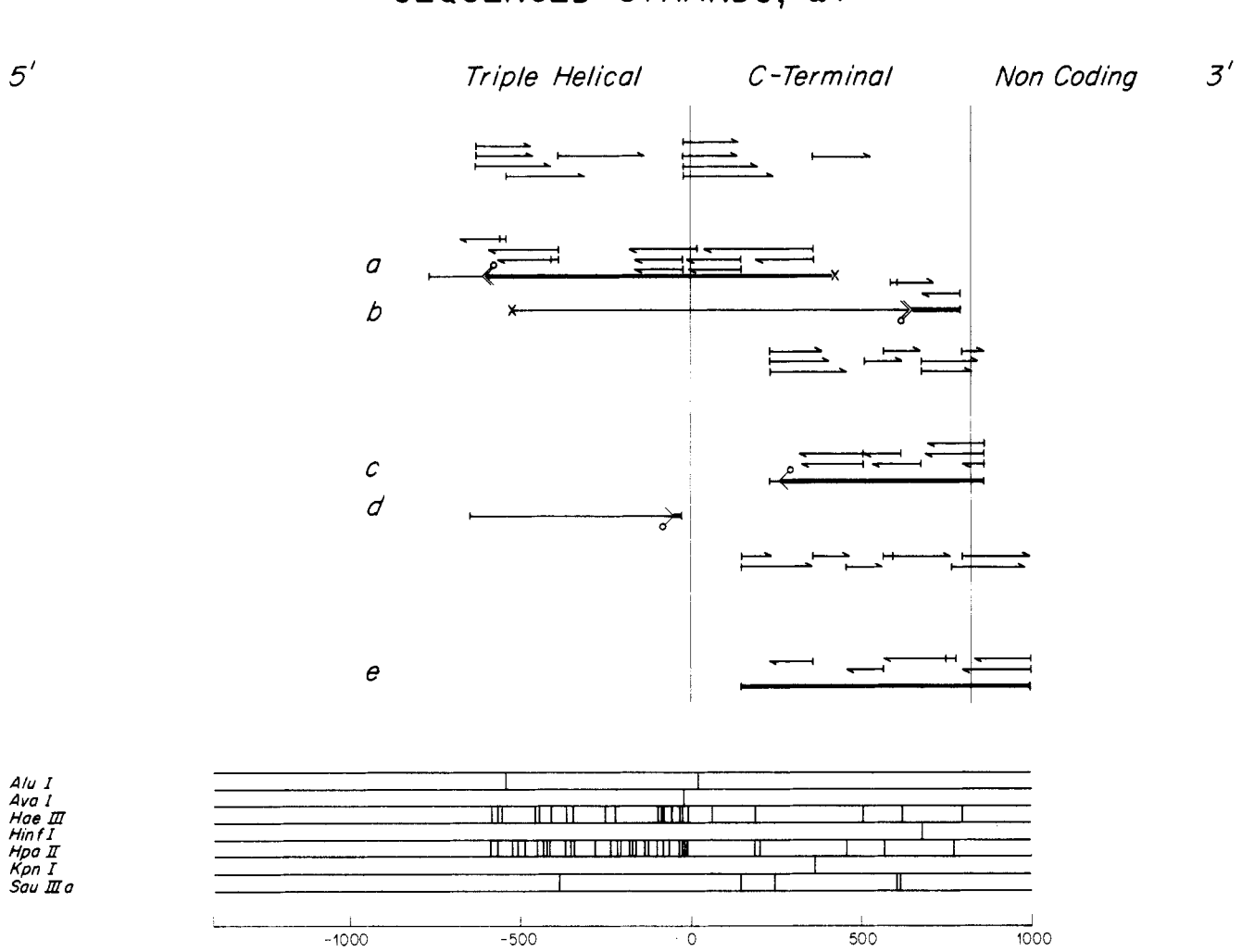

FIGURE 1: Restriction map of the pro-α1 mRNA sequence depicting the sequenced strands of cloned cDNA inserts: (a) pCg54; (b) pCg54 drawn in inverted orientation; (c) pCg26; (d) pCg26 drawn in inverted orientation; (e) pCg1. Left to right corresponds to 5′ to 3′ mRNA. For (a–d), only the part of the cloned DNA which corresponds to the mRNA restriction map below is shown in heavy line. (<°) indicates the junctions between cDNA segments in pCg54 and pCg26 with the circle always on the coding strand side of the larger segment. This is intended to help the reader visualize the difference in orientation between (a) and (b) and between (c) and (d). Arrows indicate individual sequence determinations. (|—) indicates the labeled restriction cut and that reading began within the first 14 nucleotides; (|–|—) indicates the labeled restriction cut and that reading began at the second vertical line. The arrowhead indicates the position at which reading ended and the type of end labeling: (—), 5′-end label of sense strand; (—), 5′-end label of antisense strand; (—), 3′-end label of sense strand; (—), 3′-end label of antisense strand. (×) indicates a lost *Hin*dIII site due to a deletion.

mRNAs. All coding sequences reported here have either been determined on more than one clone or have been shown to be complementary to mRNA sequences. This strongly argues that we have determined sequences which accurately correspond to the type I procollagen mRNA sequences. Differences, if they exist, must be limited to a very few single nucleotide mutations in regions which were not sequenced on overlapping clones.

Coding Sequences. The concatenated sequences and corresponding translation products are presented in Figure 3 and 4 for pro-α1 and pro-α2, respectively. For brevity, we present only the coding strand. The end of the helical coding region is depicted by a vertical line at coordinate 0. (▲) indicates the collagen carboxy-terminal peptidase site which defines the end of the telopeptide. Propeptides begin at this site and end at the termination codon (***). To optimize alignment between pro-α1 and pro-α2 amino acid sequences, 36 spaces (Δ) have been inserted into the pro-α2 nucleotide sequence. The amino acid sequences dictated by the two mRNA sequences are so similar (65% identical in the helical regions and 60% identical in the propeptides) that we have determined sequences from pCg13 only where significant differences between pro-α1 and pro-α2 peptides exist or where information is not available in pCg45.

Amino acid sequences for the α1(I) chicken collagen have been reported (Fietzek & Kühn, 1976; Dixit et al., 1978; J. H. Highberger et al., unpublished data). The corresponding sequences in Figure 3 are in perfect agreement with these except for the terminal leucine residue within the telopeptide where we find the sequence Phe-Ser-Phe-Leu. Since this residue was not degraded in the original amino acid sequence determination (Dixit et al., 1978), it is possible that the tripeptide Phe-Ser-Phe was overlooked. Calf α2 amino acid sequences have also been reported for this region (Fietzek & Kühn, 1976). Of 52 residues which correspond to part of the sequence in Figure 4, 44 or 83% are identical. This agrees well with the 80% homology previously observed between calf and chicken α2 residues (Dixit et al., 1977a). Most of the differences are conservative amino acid substitutions. In addition, we have determined sequences (unpublished data) corresponding to 52 residues previously reported (Dixit et al., 1979) for the chicken α2 cyanogen bromide peptide CB5 and

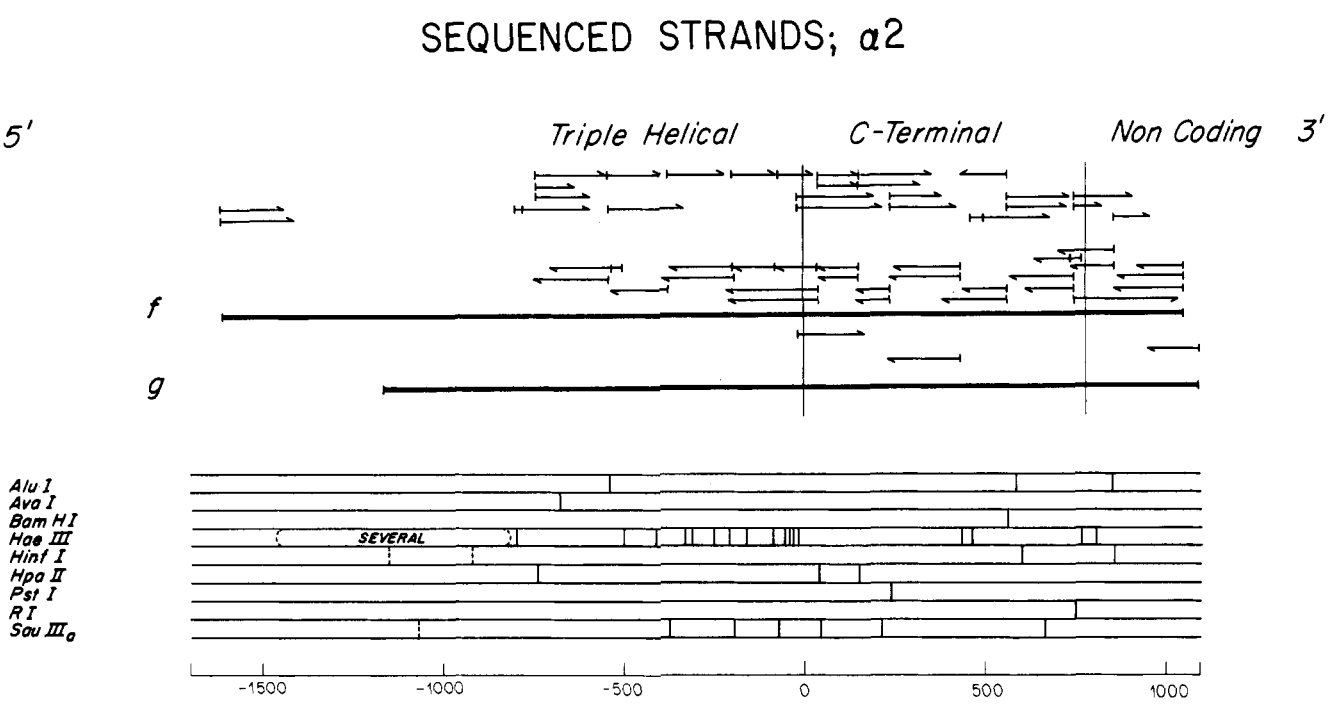

FIGURE 2: Restriction map of the pro-α2 mRNA sequence depicting the sequenced strands of cloned cDNA inserts: (f) pCg45; (g) pCg13. Restriction sites indicated by dashed lines have been mapped but not sequenced. Other symbols are the same as in Figure 1.

```
-602
 GGT GAA CGC GGT CCT CCC GGC CCC ATG GGA CCC CCC GGC CTT GCT GGC CCC CCT GGT GAA GCT GGA CGT GAG GGT GCT CCC GGT GCC GAA
-GLY-GLU-ARG-GLY-PRO-PRO-GLY-PRO-MET-GLY-PRO-PRO-GLY-LEU-ALA-GLY-PRO-PRO-GLY-GLU-ALA-GLY-ARG-GLU-GLY-ALA-PRO-GLY-ALA-GLU
-512
 GGT GCC CCC GGT CGC GAC GGT GCT GCC GGT CCC AAG GGT GAC CGT GGT GAG ACG GGC CCT GCC GGC CCC CCT GGT GCT CCC GGT GCC CCC
-GLY-ALA-PRO-GLY-ARG-ASP-GLY-ALA-ALA-GLY-PRO-LYS-GLY-ASP-ARG-GLY-GLU-THR-GLY-PRO-ALA-GLY-PRO-PRO-GLY-ALA-PRO-GLY-ALA-PRO
-422
 GGT GCC CCC GGC CCC GTC GGT CCT GCT GGC AAA AAT GGA GAT CGC GGT GAG ACC GGC CCC GCT GGT CCC GCC GGC CCC CCC GGT CCC GCT
-GLY-ALA-PRO-GLY-PRO-VAL-GLY-PRO-ALA-GLY-LYS-ASN-GLY-ASP-ARG-GLY-GLU-THR-GLY-PRO-ALA-GLY-PRO-ALA-GLY-PRO-PRO-GLY-PRO-ALA
-332
 GGT GCT CGT GGT CCT GCT GGT CCC CAA GGT CCT CGC GGC GAC AAA GGC GAG ACC GGT GAA CAG GGA GAC AGA GGC ATG AAG GGC CAC AGA
-GLY-ALA-ARG-GLY-PRO-ALA-GLY-PRO-GLN-GLY-PRO-ARG-GLY-ASP-LYS-GLY-GLU-THR-GLY-GLU-GLN-GLY-ASP-ARG-GLY-MET-LYS-GLY-HIS-ARG
-242
 GGC TTC TCC GGT CTC CAG GGC CCA CCC GGT CCT CCC GGC GCT CCT GGT GAA CAA GGT CCC TCC GGT GCT TCC GGT CCC GCC GGT CCA CGC
-GLY-PHE-SER-GLY-LEU-GLN-GLY-PRO-PRO-GLY-PRO-PRO-GLY-ALA-PRO-GLY-GLU-GLN-GLY-PRO-SER-GLY-ALA-SER-GLY-PRO-ALA-GLY-PRO-ARG
-152
 GGT CCT CCT GGT TCC GCC GGT GCC GCC GGC AAA GAC GGT CTC AAT GGG CTG CCC GGC CCC ATC GGC CCC CCC GGC CCC CGC GGT CGC ACC
-GLY-PRO-PRO-GLY-SER-ALA-GLY-ALA-ALA-GLY-LYS-ASP-GLY-LEU-ASN-GLY-LEU-PRO-GLY-PRO-ILE-GLY-PRO-PRO-GLY-PRO-ARG-GLY-ARG-THR
-62                                          Ava I
 GGT GAA GTC GGC CCC GTT GGT CCC CCC GGC CCT CCC GGC CCC CCG GGT CCT CCC GGC CCC CCC|AGC GGC GGC TTC GAC TTC AGC TTC CTG
-GLY-GLU-VAL-GLY-PRO-VAL-GLY-PRO-PRO-GLY-PRO-PRO-CLY-PRO-PRO-GLY-PRO-PRO-GLY-PRO-PRO|SER-GLY-GLY-PHE-ASP-PHE-SER-PHE-LEU
28
 CCT CAA CCG CCC CAA GAG AAG GCG CAC GAC GGC GGC CGC TAC TAC CGA GCC GAC GAC GCC AAC GTG ATG CGC GAC CGC GAC CTG GAG GTC
-PRO-GLN-PRO-PRO-GLN-GLU-LYS-ALA-HIS-ASP-GLY-GLY-ARG-TYR-TYR-ARG-ALA-ASP▲-ASP-ALA-ASN-VAL-MET-ARG-ASP-ARG-ASP-LEU-GLU-VAL
118
 GAC ACC ACC CTG AAG AGC CTC AGC CAA CAG ATC GAG AAC ATC CGC AGC CCC GAA GGC ACC CGC AAG AAC CCG GCC CGC ACC TGC CGG GAC
-ASP-THR-THR-LEU-LYS-SER-LEU-SER-GLN-GLN-ILE-GLU-ASN-ILE-ARG-SER-PRO-GLU-GLY-THR-ARG-LYS-ASN-PRO-ALA-ARG-THR-CYS-ARG-ASP
208
 CTG AAG ATG TGC CAC GGC GAC TGG AAG AGC GGC GAG TAC TGG ATC GAC CCC AAC CAG GGC TGC AAC CTG GAT GCC ATC AAG GTC TAC TGC
-LEU-LYS-MET-CYS-HIS-GLY-ASP-TRP-LYS-SER-GLY-GLU-TYR-TRP-ILE-ASP-PRO-ASN-GLN-GLY-CYS-ASN-LEU-ASP-ALA-ILE-LYS-VAL-TYR-CYS
298                                                    Kpn I
 AAC ATG GAG ACG GGT GAG ACG TGC GTC TAC CCC ACG CAA GCC ACC ATT GCC CAG AAG AAC TGG TAC GTC AGC AAG AAC CCC AAG GAG AAG
-ASN-MET-GLU-THR-GLY-GLU-THR-CYS-VAL-TYR-PRO-THR-GLN-ALA-THR-ILE-ALA-GLN-LYS-ASN-TRP-TYR-LEU-SER-LYS-ASN-PRO-LYS-GLU-LYS
388
 AAG CAC GTC TGG TTC GGC GAG ACG ATG AGC GAC GGC TTC CAG TTT GAG TAC GGC GGT GAG GGC TCC AAC CCG GCT GAT GTC GCC ATC CAA
-LYS-HIS-VAL-TRP-PHE-GLY-GLU-THR-MET-SER-ASP-GLY-PHE-GLN-PHE-GLU-TYR-GLY-GLY-GLU-GLY-SER-ASN-PRO-ALA-ASP-VAL-ALA-ILE-GLN
478
 CTG ACC TTC CTG CGC CTG ATG TCC ACC GAG GCC ACC CAG AAC GTC ACC TAC CAC TGC AAG AAC AGC GTC GCC TAC ATG GAC CAC GAC ACC
-LEU-THR-PHE-LEU-ARG-LEU-MET-SER-THR-GLU-ALA-THR-GLN-ASN-VAL-THR-TYR-HIS-CYS-LYS-ASN-SER-VAL-ALA-TYR-MET-ASP-HIS-ASP-THR
568
 GGC AAC CTG AAG AAG GCT CTG CTG CTC CAG GGA GCC AAC GAG ATC GAG ATC AGG GCC GAA GGA AAC AGC CGC TTC ACC TAT GGG GTC ACC
-GLY-ASN-LEU-LYS-LYS-ALA-LEU-LEU-LEU-GLN-GLY-ALA-ASN-GLU-ILE-GLU-ILE-ARG-ALA-GLU-GLY-ASN-SER-ARG-PHE-THR-TYR-GLY-VAL-THR
658          Hinf I
 GAG GAT GGC TGC ACG AGT CAC ACT GGA GCA TGG GGC AAA ACA GTG ATT GAG TAC AAG ACG ACG AAG ACC TCG CGC CTG CCC ATC ATT GAC
-GLU-ASP-GLY-CYS-THR-SER-HIS-THR-GLY-ALA-TRP-GLY-LYS-THR-VAL-ILE-GLU-TYR-LYS-THR-THR-LYS-THR-SER-ARG-LEU-PRO-ILE-ILE-ASP
748
 TTG GCT CCT ATG GAC GTT GGC GCT CCG GAC CAG GAA TTT GGC ATT GAC ATC GGC CCC GTC TGC TTT TTG TAA ACA GGA AAA AAA AAG AAA
-LEU-ALA-PRO-MET-ASP-VAL-GLY-ALA-PRO-ASP-GLN-GLU-PHE-GLY-ILE-ASP-ILE-GLY-PRO-VAL-CYS-PHE-LEU-***
838
 AAG AAA AGA AAA AAA AAA AAA AAA AAG CCC CCC CAA CGC GTG ACA GGA GAG AGT AAT AAT TAT AAT AAT TAA TAA AAA AAA AAA AAA AAA
928
 AAA AAA AAC TGC CAA AAA ATG GAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AA
```

FIGURE 3: Sense strand sequences for pro-α1 cDNA cloned inserts. Numbers on the left indicate the coordinate of the first nucleotide in the row (i.e., the bond 3′ to the first nucleotide in the row). Corresponding amino acids are indicated. Important amino acids are underlined. Recognition sequences for restriction enzymes which cut uniquely are indicated. Noncoding sequences which are overlined are discussed under Results and Discussion. Other symbols: (|) end of the helical-coding region and beginning of the C-terminal coding region (coordinate 0); (▲) carboxyprotease cleavage site (end of the telopeptide); (***) first in-phase termination codon; (△) spaces which must be inserted into the pro-α2 sequence for best amino acid alignment (not included in the coordinates).

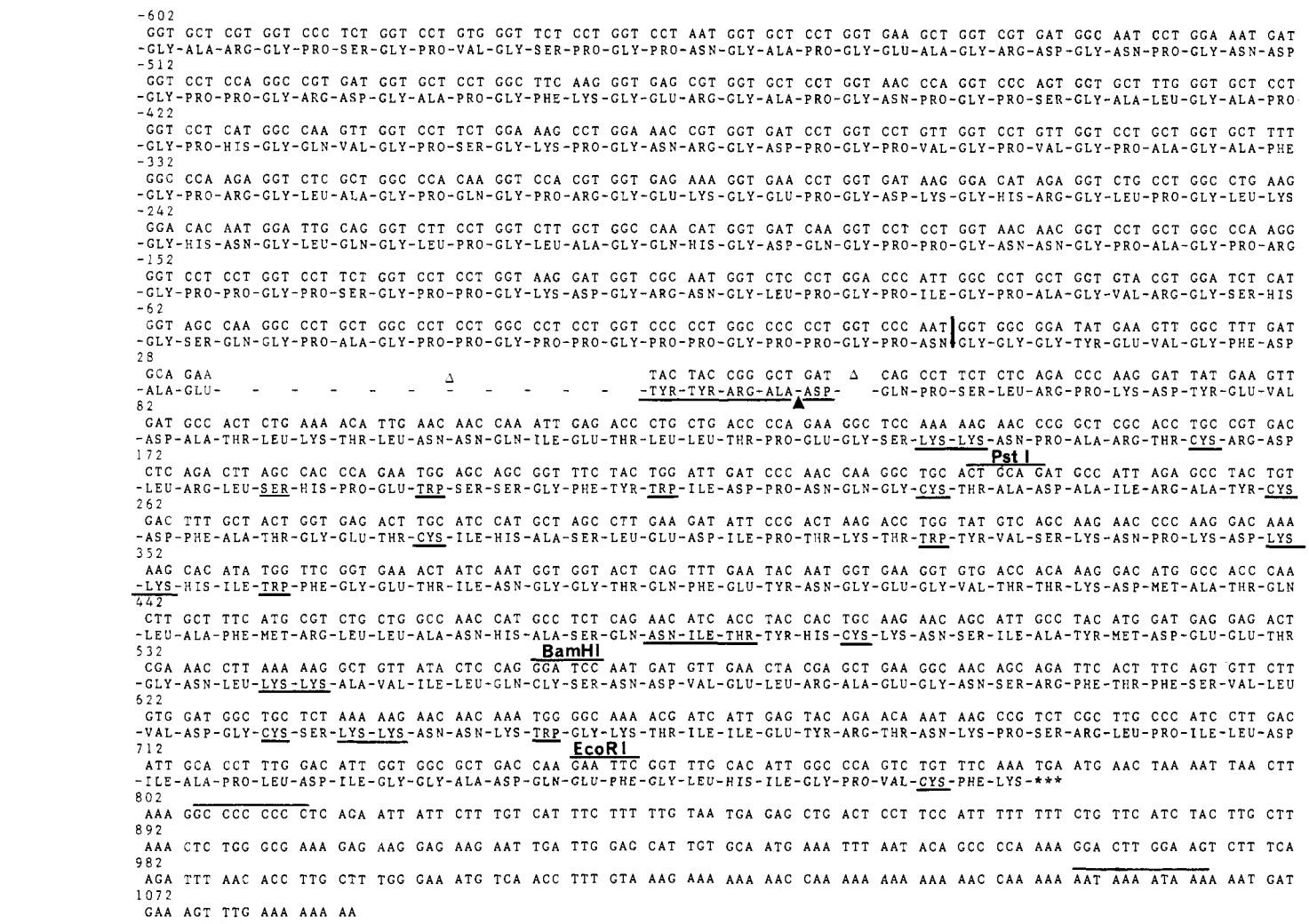

FIGURE 4: Sense strand sequence for pro-α2 cDNA cloned inserts. See Figure 3 for details.

Table I: Comparison of Nucleotide Sequences Coding for Pro-α1 and Pro-α2 Collagen Chains

nucleotides sequenced	frequency							
	triple helical		C-terminal		noncoding		poly(A)	
	α1, 603[a]	α2, 603[a]	α1, 816[a]	α2, 780[a]	α1, 133[a]	α2, 300[a]	α1, 37[a]	α2, 8[a]
G+C	0.75	0.60	0.61	0.48	0.23	0.29	0	0
pCpG	0.103	0.018	0.081	0.019	0.015	0.003		
pCpGR[b]	0.139	0.091	0.093	0.057	0.013	0.020		
pCpCpGpG[c]	0.046	0.006[d]	0.006	0.003	0	0		
homology	65%		63%		43%			

[a] Number of nucleotides sequenced. [b] Calculated for the random distribution of C and G. [c] *Hpa*II site. [d] In 1599 helical coding nucleotides.

find them to be in complete agreement. Residues specified by the pro-α2 sequences in Figure 4 combined with previous determinations of the amino-terminal sequence of CB5 define all but six residues of this peptide. The nucleotide sequence immediately preceding that in Figure 4 corresponds to the hexapeptide sequence Gly-Val-Ser-Gly-Pro-Pro (unpublished data). Thus, only the sequence of CB4 is needed to complete the amino acid sequence of mature chicken α2(I) collagen (Kang & Gross, 1970; Highberger et al., 1971; Dixit et al., 1977a,b, 1979).

The complete amino acid sequences of the type I collagen telopeptides and C-terminal propeptides are also depicted in Figure 3 and 4. Amino acids which are conserved and may be important to function are underlined. None of the pro-α2(I) C-terminal propeptide residues and only a very small fraction of the pro-α1(I) C-terminal propeptide residues (Showalter et al., 1980) have been determined by protein sequencing. The comparative analysis of these amino acid sequences is discussed elsewhere (Fuller and Boedtker, unpublished experiments).

3′ Nontranslated Sequences. The A+T rich, nontranslated regions of both mRNAs are similar in composition to those of other eukaryotic mRNA species (O'Hare et al., 1979; Konkel et al., 1979; Roskam & Rougeon, 1979; Nakanishi et al., 1979). One difference is notable, however. Both mRNAs contain an oligo(dC) tract 7 (α1) or 8 (α2) nucleotides long, preceded by a G, and 45 (α1) or 23 (α2) nucleotides downstream from the stop codon. The significance of this sequence is unknown.

The lengths of the 3′ nontranslated regions are 133 and 300 nucleotides for pro-α1 and pro-α2 mRNAs, respectively (Table I and Figures 3 and 4). Both contain the canonical pAAUAAA sequence preceding the poly(A) addition site

(Proudfoot & Brownlee, 1976). Two such sequences occur in the $\alpha 2$ sequence 16 and 21 nucleotides preceding the poly(A) addition site. R loop and partial sequence analysis of a pro-$\alpha 2$ genomic clone indicates a nontranslated length of approximately 250 to 350 nucleotides (unpublished data). We conclude, therefore, that our sequences cover the entire pro-$\alpha 2$ 3′ noncoding sequence.

The pro-$\alpha 1$ noncoding sequence is somewhat odd in that the pAAUAAA sequence is followed by $p(A)_{22}CTGCC-(A)_6TGG$-poly(A). The pro-$\alpha 1$ noncoding region contains another unusually long oligo(dA) tract of 18 nucleotides. It is possible then that the 37 A's at the end of pCg1 are in fact encoded in the genomic DNA and that our sequence does not contain the entire noncoding portion of the pro-$\alpha 1$ mRNA.

Four Point Mutations in pCg26. The number of nucleotides sequenced in each coding and noncoding region is shown in Table I along with other statistical data. Of the total, 354 pro-$\alpha 1$ C-terminal coding, 42 pro-$\alpha 1$ noncoding, 345 pro-$\alpha 2$ C-terminal coding, and 87 pro-$\alpha 2$ noncoding nucleotides have been determined on two overlapping clones; 306 pro-$\alpha 1$ C-terminal coding nucleotides have been determined on three overlapping clones. The total redundancy of multiply determined nucleotides is thus 2574. Four differences in nucleotide sequence are evident. This implies that the process of cDNA synthesis and amplification in *E. coli* is only moderately faithful. The probable error rate in our hands is $^4/_{2574}$ or about 1.5×10^{-3}, which compares favorably with error rates found for reverse transcription of ϕX DNA (Gopinathan et al., 1979).

It is of interest to note that these point mutations are clustered on pCg26 at positions 628, 834, 840, and 845 in the pro-$\alpha 1$ sequence. They are all G's in the pCg1 sequence and A's in the pCg26 sequence. The pCg1 sequence probably reflects the correct mRNA sequence since the G in the coding region defines a glycine codon (as opposed to an arginine codon in pCg26) and a glycine codon is present in the same position in the pro-$\alpha 2$ sequence. The sequence in Figure 3 corresponds to the pCg1 sequence at these positions. The correctness of the pCg1 sequence is also confirmed by the sequence of an independently isolated pro-$\alpha 1$ clone (Showalter et al., 1980).

Lomedico et al. (1979) have found four discrepancies in the sequences of proinsulin cDNA clones and have observed that they are *not* likely due to polymorphism. These changes are also G to A transitions with respect to the coding strand and occur at about the same frequency as we find. Since these changes apparently occur on only one strand, it is likely that the most frequent error in cloning cDNA sequences is the misincorporation of T for C during cDNA synthesis. This is supported by the finding that reverse transcriptase tends to overincorporate T (Falvey et al., 1976) and by the low discrepancy rate between sequences of genomic DNA clones (Konkel et al., 1979). Furthermore, since the errors in our clones are clustered, it is possible that only a subpopulation of reverse transcriptase molecules is responsible. In fact, Gopinathan et al. (1979) have observed variation in the misincorporation rate between different preparations of reverse transcriptase.

Sequence Rearrangements in cDNA Clones. We have previously reported (Lehrach et al., 1979) that sequences at one end of pCg26 and at one end of pCg54 do not correspond to contiguous mRNA sequences. Because we have determined sequences from three overlapping clones, an unambiguous consensus sequence can be derived. Comparing the sequences of the aberrant ends of these clones with the consensus mRNA sequence indicates that these sequences in fact correspond to distal mRNA sequences. Thus, each clone contains two segments of double-stranded cDNA. The cDNA segments in pCg26 correspond to nucleotides −49 to −24 and to nucleotides 264–862 in Figure 3. pCg54 contains segments corresponding to nucleotides −602 to 425 and to nucleotides 658–805. There are only four ways to combine two segments of DNA and preserve the antiparallel arrangement of strands. Two of these will result in coding strands being connected to coding strands, and two will result in coding strands being connected to noncoding strands. The sequences cloned in both pCg26 and pCg54 fall into only one of these four possible classes. Specifically, the junction between the cDNA segments contains the 5′ ends of coding strands fused to the 3′ ends of noncoding strands. The cDNA segments are thus inverted relative to each other and are connected in a head-to-head fashion as opposed to tail-to-tail.

In addition to these two clones, a pro-$\alpha 2$ clone, pCg75, has this distinctive structure (unpublished data). An independently isolated pro-$\alpha 1$ clone, pCOL3, has been reported (Yamamoto et al., 1980). The restriction map of this clone suggests that it too may be a rearranged clone consisting essentially of the large end of pCg26 and the small end of pCg54 (personal observation). Although this constitutes a limited number of data, it is possible to conclude that the events which have given rise to these rearrangements (1) occur at a high frequency (at least in pro-$\alpha 1$ clones), (2) are not recurrent (none of our clones has been observed to rearrange subsequent to initial cloning), (3) may be limited to a few sites of rearrangement, (4) result in sequence inversions where the junction contains the 5′ ends of coding strands, and (5) do not appear to enjoin heterologous cDNA sequences. A trivial explanation for these structures is that they are due to dimerization of cDNAs during the linker addition ligation step. However, cDNA dimers formed in this manner would be expected to contain random ends, randomly associated. Since this is *not* the case, a more likely explanation is that these relatively precise rearrangements occur either during cDNA synthesis or upon transfection into *E. coli*.

To schematically represent the ends of pCg26 and pCg54, we have drawn these clones in each of two possible orientations in Figure 1. In each orientation only the part of the cloned DNA which corresponds to the mRNA sequence is positioned in accordance with the restriction map below and drawn in heavy line. The portion of the cloned sequence which is drawn in light line does not correspond to the mRNA restriction map in that region (e.g., the *Hin*fI site at one end of pCg54 is indicated only in Figure 1b and not in 1a). The symbol <° indicates the junction between cDNA segments with the circle always on the coding strand side of the larger segment. pCg54 also contains a third section of cDNA in the junction region. This collagen-like decanucleotide sequence, pCGGTTCCCCC, immediately precedes the sequence in Figure 3. It cannot be the remnant of a defective *Hin*dIII linker caught in the junction and probably corresponds to mRNA sequences 5′ to those shown in Figure 3. Thus, a deletion of a segment of helical coding DNA may also have occurred in pCg54.

Comparison with Previously Reported Sequences. Part of the pro-$\alpha 1$ sequence shown in Figure 3 overlaps the 600 nucleotide sequence reported for pCOL3 by Showalter et al. (1980). While the agreement between the two sequences is substantial, there are four discrepancies in the C-terminal coding region and six in the noncoding region. It is possible that some of these differences may be due to polymorphism of the pro-$\alpha 1$ sequence since large numbers of chicken embryos were used in mRNA preparations. However, we note the

following. Five positions in the pCOL3 sequence were given ambiguous assignments. Two of the differently assigned nucleotides (450 and 461) and four ambiguously assigned nucleotides (434, 452, 462, and 464) occur very near a *Hpa*II site (457). We have obtained agreeing sequences from two clones for all these positions. Two other discrepancies in the coding region (726 and 780) were identical on all three of our pro-α1 clones. Four differences in the lengths of oligo(dA) tracts in the noncoding region are possible miscounts since clearly interpretable autoradiographs have been obtained from independent sequence determinations of both strands of pCg1. However, we cannot rule out the possibility of heterogeneous incorporation of T due to slippage during reverse transcription. The 3′-terminal pCCA where we indicate only poly(A) is probably the *Hin*dIII linker sequence. Finally, the sequence pCCGGCC has been identified in the noncoding region where we find pCTGCC. The former sequence contains both a *Hae*III site and a *Hpa*II site. Furthermore, nucleotide 780 is the complement G of a mC in our sequence and thus defines an *Eco*RII site. No such site is indicated at this position in the pCOL3 sequence. Restriction mapping of pCOL3 by the same authors (Yamamoto et al., 1980) does *not* indicate *Hpa*II or *Hae*III sites in the corresponding region and *does* indicate an *Apy*I site at about 780 (*Apy*I is an isoschizomer of *Eco*RII). These observations lead us to conclude that the 15 differences between our assignments and those of Showalter et al. (1980) cannot be the result of errors in the determination of the sequences reported here and that explanations involving possible polymorphism of the pro-α1 sequence should await further confirmation of the pCOL3 sequence.

Sequence Anomalies. One region in the pro-α1 sequence (−30 to −10) has proven difficult to sequence. This region contains a partially repeated sequence which is about 90% G+C. On all autoradiographs depicting this region, 3–6 bands run together. Moreover, the positions of these compressed bands also vary. The sequence in this region has been unambiguously determined by comparing autoradiographs obtained from different sequence determinations. Brown & Smith (1979) first observed such anomalies and found that resolution could be increased by increasing temperature during electrophoresis. We have found compressions in a number of sequences when lower voltages or lower concentrations of urea were used during electrophoresis; however, all but the ones between −30 and −10 could be totally resolved at higher (45–55 °C) temperatures.

Another sequencing anomaly is apparently associated with the repetitive and G+C-rich sequence typical of the helical coding region. When reading the antisense strand, tracts of G's corresponding to proline anticodons are interrupted by two C's corresponding to glycine anticodons. The C doublet often migrates in an anomalous position, as shown in Figure 5b. The reason for this is unknown; however, the true position of the C doublet can be determined from the sequence alone because, in Maxam–Gilbert sequencing, the bands corresponding to G's in a tract are progressively lighter from 5′ to 3′ (Figure 5; Maxam & Gilbert, 1977). Thus, a dark band in the G lane represents the first G in a string of G's and the last light band represents the 3′ terminal G in the tract. The two C's must lie between the two G tracts in Figure 5b and the sequence is thus pAGGGCCGGGGGGACC. The protein sequence which requires glycine in every third position (i.e., C doublets separated by seven nucleotides) also specifies this sequence.

G+C Content and pCG Frequency. The average G+C content determined for separate regions of both mRNA sequences are shown in Table I. The G+C content of the helical coding regions are unusually high compared to the average G+C content of chicken DNA, which is 44% (Turner et al., 1963), and they are also high relative to the G+C content of the corresponding C-terminal coding regions. This is understandable in light of the fact that approximately 67% of the amino acids in the helical region of collagen chains is glycine, proline, or alanine (Eastoe, 1967). The variation in G+C content is more finely detailed in Figure 6 where we have plotted it as a function of position in the nucleotide sequence. Each point in the diagram depicts the fraction of G+C in a tract of nucleotides of specified length and is plotted at the position of the central nucleotide in that tract. Sliding the position of the tract by one nucleotide at a time results in a curve which is discontinuous (on the vertical axis) if short tracts are specified and loses detail if long tracts are specified. To smooth out the curve and still depict details, we have superimposed curves obtained for various tract lengths (Figure 6b). From this it is readily apparent that the pro-α1 sequence is consistently higher in G+C content (over 90% in four places) than is the pro-α2 sequence, except in the noncoding region. It is also apparent that the curves mimic each other most at the end of the helical coding region and in the C-terminal coding region. The curves are nearly coincident at the end of the coding region. The sequences in these regions are in fact much more homologous than those in the surrounding regions, especially at the end of the C-terminal coding region (see Homology).

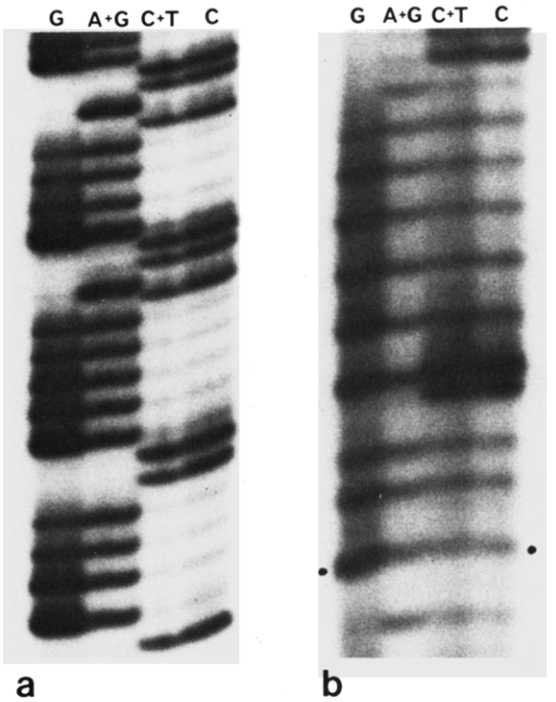

FIGURE 5: Sequences of antisense strands in the helical coding region of pCg54: (a) typical sequence; (b) sequence showing anomalous migration of C doublet nested between two tracts of G's. Note the decrease in intensity of each G in a tract of G's. The sequence in (b) is pAGGGCCGGGGGGAC.

The higher G+C content of the pro-α1 mRNA in the regions sequenced is consistent with the observation that genomic pro-α1 DNA fragments contain a higher G+C content than genomic pro-α2 DNA fragments as determined by banding position in actinomycin CsCl gradients (Wozney et al., 1979). The helical coding regions are some 3000 nucleotides long and comprise about three-fifths of the entire length of each mRNA. Based on total mRNA lengths of 4900 (pro-α1) and 5100 (pro-α2) nucleotides (Rave et al., 1979) and assuming that

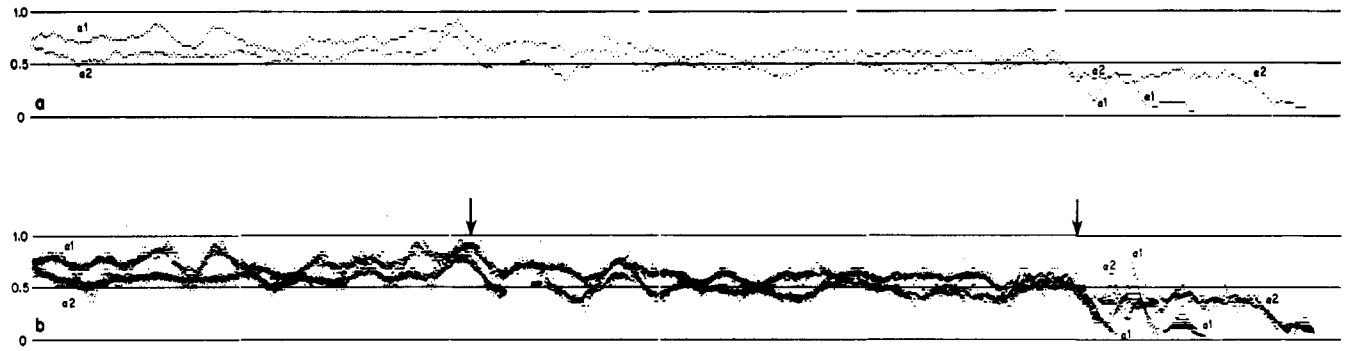

FIGURE 6: Plot of fractional G+C content as a function of position in the sequence. (a) Averaged over 49 contiguous nucleotides and plotted at the center. (b) Averages over 25, 31, 37, 43, 49, and 55 nucleotides are superimposed. For alignment of pro-α1 and pro-α2 sequences, 36 false insertions are made into the pro-α2 sequence (see Figures 3 and 4). These positions are not plotted but are assigned an arbitrary value of 0.5 (G+C content) for averaging purposes. The arrows locate the 3′ end of the helical coding region (left) and the C-terminal coding region (right).

Table II: Codon Usages in Helical Coding Regions

	U	α1	α2	C	α1	α2	A	α1	α2	G	α1	α2	
U	Phe	0	1	Ser	0	5	Tyr	0	0	Cys	0	0	U
	Phe	1	1	Ser	4	0	Tyr	0	0	Cys	0	0	C
	Leu	0	0	Ser	0	0	term	0	0	term	0	0	A
	Leu	0	2	Ser	0	0	term	0	0	Trp	0	0	G
C	Leu	1	2	Pro	13	36	His	0	4	Arg	3	7	U
	Leu	2	2	Pro	36	6	His	1	1	Arg	7	1	C
	Leu	0	0	Pro	2	6	Gln	2	5	Arg	0	0	A
	Leu	1	2	Pro	1	0	Gln	2	1	Arg	0	0	G
A	Ile	0	1	Thr	0	0	Asn	2	6	Ser	0	1	U
	Ile	1	0	Thr	3	0	Asn	0	4	Ser	0	1	C
	Ile	0	0	Thr	0	0	Lys	3	1	Arg	2	2	A
	Met	2	0	Thr	1	0	Lys	2	5	Arg	0	1	G
G	Val	1	3	Ala	12	14	Asp	1	7	Gly	38	45	U
	Val	2	0	Ala	11	0	Asp	5	0	Gly	24	14	C
	Val	0	1	Ala	0	0	Glu	6	2	Gly	4	8	A
	Val	0	1	Ala	0	0	Glu	4	2	Gly	1	0	G

the 5′ N-terminal coding and noncoding sequences are approximately as G+C rich as the C-terminal coding regions, we would estimate the G+C content of the pro-α1 and pro-α2 mRNAs to be 69% and 54%, respectively.

An interesting feature of the collagen mRNA sequences is the relatively frequent occurrence of the dinucleotide pCG. This dinucleotide sequence is unusually rare in the DNAs of most vertebrates, occurring about 5 times less frequently than expected for random distribution of G and C (Russel et al., 1976). The frequency of pCG occurrence in each region of the collagen sequences that we have determined as well as that expected for the region if G's and C's were randomly distributed is also shown in Table I. Although both pro-α1 and pro-α2 sequences have a higher pCG frequency than that expected for total chicken DNA (about 0.01), the frequency in the pro-α1 mRNA is consistently 5-fold higher than in the pro-α2 mRNA. This does not simply reflect the differences in G+C content. As can be seen in Table I, the pro-α1 sequence contains pCG at a frequency almost equal to that expected for random distribution of G and C (in an unusually G+C-rich mRNA) while the frequency in the pro-α2 sequence is 3 to 7 times lower than predicted for a random distribution. The difference in pCG frequency is also reflected in the number of *Hpa*II sites (pCCGG) present in the helical coding regions (Table I). In contrast, *Hae*III sites (pGGCC) are abundant in triple-helical coding regions of both mRNAs (Figures 1 and 2).

Homology. We have employed computer-generated homology plots (data not shown) as described by Konkel et al. (1979) to determine the lengths and frequencies of homologies in the sequences reported here. The method visually depicts contiguous homologies irrespective of the alignment of the sequences. Despite the overall nucleotide homology of the two mRNAs as aligned in Figures 3 and 4 (Table I), the plots indicate that cross-homology is *not* extensive. In the helical coding regions it is essentially limited to short ($\leq$10 nucleotide), scattered tracts. The two largest homologies are 14 nucleotides each. Self-homologies in the same region (excluding the collinear, perfect homology) seem to be more frequent and, in the pro-α2 sequence, longer. The longest pro-α2 self-homology is 19 nucleotides. Evidently, these are two very divergent nucleotide sequences (F. Fuller and H. Boedtker, manuscript in preparation). Surprisingly, the longest cross-homologies are found in the C-terminal coding regions. Here, two homologies of lengths 19 and 23 and one 93% homology 29 nucleotides long are evident. Homology in the noncoding region is essentially limited to the poly(dA) tracts.

Codon Usage. Collagen mRNAs, like other mRNAs (Grantham et al., 1980), are biased in their use of synonymous codons. This is demonstrated in Tables II and III. However, the collagen bias is further exaggerated by the large fraction of glycine, alanine, and proline codons. Most obvious is the small fraction of codons with purines, especially G, in the wobble position. A χ^2 test indicates that the probability of this degree of bias occurring at random is <0.001 for both helical-coding and C-terminal-coding regions of each mRNA. The bias is consistent with the finding that the most abundant $tRNA^{Gly}$ species in calvaria is cognate to GGC and GGU codons and that the species cognate to GGG is rare (Carpousis et al., 1977). Only one of 134 helical coding glycine codons

Table III: Codon Usages in C-Terminal Coding Regions

	U	α1	α2	C	α1	α2	A	α1	α2	G	α1	α2	
U	Phe	3	3	Ser	0	4	Tyr	1	3	Cys	0	2	U
	Phe	7	7	Ser	2	2	Tyr	10	8	Cys	8	5	C
	Leu	0	0	Ser	0	0	term	1	0	term	0	1	A
	Leu	2	4	Ser	1	0	term	0	0	Trp	5	5	G
C	Leu	0	6	Pro	2	2	His	0	2	Arg	0	2	U
	Leu	3	3	Pro	7	4	His	6	4	Arg	9	2	C
	Leu	0	1	Pro	0	3	Gln	5	4	Arg	1	1	A
	Leu	12	5	Pro	4	3	Gln	7	4	Arg	1	1	G
A	Ile	4	9	Thr	1	9	Asn	0	4	Ser	1	1	U
	Ile	9	5	Thr	13	7	Asn	14	12	Ser	10	7	C
	Ile	0	2	Thr	1	3	Lys	1	8	Arg	0	5	A
	Met	7	3	Thr	7	1	Lys	16	11	Arg	1	0	G
G	Val	1	5	Ala	4	8	Asp	3	10	Gly	2	10	U
	Val	8	2	Ala	11	7	Asp	17	7	Gly	18	9	C
	Val	0	0	Ala	1	3	Glu	3	12	Gly	3	3	A
	Val	3	2	Ala	1	0	Glu	15	5	Gly	1	0	G

sequenced is GGG. The high G+C content of collagen mRNAs may require that G be avoided wherever possible in order to limit secondary structure (Bachra, 1976). A large degree of secondary structure in these mRNAs is indicated by their resistance to reverse transcription (unpublished data), by compressions of bands on sequencing gels, and by their propensity to aggregate in solution (Boedtker et al., 1976; unpublished data). Furthermore, the higher G+C content of the pro-α1 mRNA should impart a more stable secondary structure on this mRNA as compared to the pro-α2 mRNA. Accordingly, Tolstoshev et al. (1979) have shown that, in 85% formamide (sucrose gradients), the pro-α1 mRNA sediments faster than the pro-α2 mRNA even though the pro-α2 mRNA is larger (Rave et al., 1979; Adams et al., 1979).

The distribution of isoaccepting tRNAs in a given tissue seems to mimic the codon bias of the major mRNA species (Garel, 1974; Carpousis et al., 1977; Osterman, 1979; Grantham & Gautier, 1980). Therefore, tissues expressing collagen sequences would be expected to contain very low concentrations of tRNAs with U and C in the anticodon wobble because both are cognate to G. Since tRNAs with U in the anticodon wobble are also cognate to A, selective pressure should discriminate against codons with A in the third position. Thus, a low frequency of A in the third position of collagen codons is a necessary consequence of a low frequency of G in the third position.

The absence of significantly long homology tracts and the differences in G+C content, in the frequency of pCG, and in the number of *Hpa*II and *Eco*RII sites can all be accounted for by the differing bias in codon usages of the two mRNAs. Tables II and III show that the pro-α1 mRNA prefers C in the third position while the pro-α2 mRNA prefers U. The χ^2 probabilities of this degree of bias occurring at random are 0.054, <0.001, <0.001, and 0.46 for α1 helical, α2 helical, α1 C-terminal and α2 C-terminal coding regions, respectively. Eliminating the most frequent codons (for proline, alanine, and glycine) from the calculation for the α1 helical coding region gives a probability of <0.005. Thus, dipeptides which are coded for frequently in the helical regions (Pro-Gly, Pro-Ala, Ala-Gly, Gly-Ala, and Ala-Ala) necessitate a higher frequency of pCG in the pro-α1 mRNA. The difference in the numbers of *Hpa*II and *Eco*RII sites is a consequence of this pyrimidine bias in proline codons in Pro-Gly dipeptides.

Conclusions

The chicken pro-α1(I) and pro-α2(I) collagen mRNAs code for two very similar, coexpressed, and interdependent polypeptides, yet they differ substantially in nucleotide sequence. These differences are due mostly to a U/C (α2/α1) bias in the choice of codon third positions. As a result, sequence homologies are frequently interrupted, and no large homology tracts are found in the sequences reported here; however, the lack of a U bias in the pro-α2 C-terminal coding region is consistent with the slightly longer cross-homology tracts found in the C-terminal coding regions as compared to the helical coding regions. These findings corroborate our earlier observations that cross-hybridization between α1 and α2 cDNA clones is insignificant under stringent filter hybridization conditions (Lehrach et al., 1978; Rave et al., 1979). The sequences are therefore capable of serving as distinct probes for their respective mRNA, pre-mRNA, and genomic sequences. However, under nonstringent conditions these sequences can hybridize to unique genomic fragments of an organism as divergent from chicken as the nematode *Caenorhabditis elegans* (D. Hirsh, personal communication). Therefore, these cDNA clones may be capable of serving as probes for collagen coding sequences in other organisms under appropriate hybridization conditions. Moreover, the α1 sequences of type I and type II collagens are more homologous than are the α1(I) and the α2(I) collagens (Dixit et al., 1977; Bornstein & Sage, 1980). Therefore, pCg54 may be a useful probe for type II collagen coding sequences as well.

Differences in the sequences of the two mRNAs result in significant differences in G+C content and in the number of pCG dinucleotides, *Eco*RII sites, and *Hpa*II sites. Although these particular properties may have no direct in vivo significance (e.g., the number of *Hpa*II sites), they serve to illustrate how such sequence differences can result in marked physical differences of possible biological consequence. Because either G or I in the anticodon wobble is cognate to both C and U, it is unlikely that the relative rates of translation of these two collagen mRNAs could be effected by the relative abundance of isoacceptor tRNAs; however, relative rates of translation and pre-mRNA processing could be effected by differences in the RNA secondary structure dictated by the U/C bias. The extent of methylation of Cs (usually at pCG sites) in genomic DNAs has been found to correlate with expression of certain genes in vivo (van der Ploeg & Flavell, 1980). Differential methylation of the two mRNAs due to the difference in frequency of pCG sequences could also effect the relative rates of expression. The U/C bias need not necessarily effect expression of the two sequences and may instead indicate differential treatment of these sequences at a more subtle level. For instance, it may be a consequence of processes involved

in the maintenance of these two sequences as distinct entities within the genome.

If the codon bias found in these collagen coding sequences is the result of processes which act on related sequences which are coexpressed, then the mouse β^{maj} and β^{min} globin (Konkel et al., 1979), chicken ovalbumin and X (Heilig et al., 1980), and rat insulin I and II (Lomedico et al., 1979) coding sequences might also be expected to show this bias. Unfortunately, the variety of amino acids and the lack of a strong bias against purines in the third position of noncollagen codons mean that a U/C bias should be difficult to detect in most closely related sequences. In fact these sequences do not demonstrate overall U/C biases; however, biases in individual codon usages are evident. Arginine codons in the rat insulin sequences do show a strong U/C (rat insulin I/rat insulin II) bias. The hemoglobin α and β genes also code for related polypeptides which are coexpressed. Here, a larger statistical sampling is available, and a codon bias similar to that found in collagen coding sequences is evident. If the codon usages of chicken (Richards et al., 1979; Salser et al., 1979), human (Forget et al., 1979; Lawn et al., unpublished results), mouse (Konkel et al., 1979; Nishioka & Leder, 1979), and rabbit (Efstratiadis et al., 1977; Heindell et al., 1978; Hardison et al., 1979; van Ooyen et al., 1979) globins are summed and compared, it is apparent that both α- and β-globin codons are generally biased against purines in the third position. When pyrimidine choices are compared, most α-globin codons are seen to prefer C in the third position while β-globin codons either prefer C to a lesser degree, show no preference at all, or prefer U in the third position. Moreover, the α and β globins are similar to $\alpha 1$ and $\alpha 2$ collagens in that the two polypeptides are not only coexpressed but are functionally interdependent. It remains to be determined, then, whether the differences observed between these two collagen coding sequences are unique to chicken collagens or to collagens in general or whether they represent common processes which act on other coding sequences and which are merely exaggerated in collagen coding sequences.

Acknowledgments

We are indebted to Hans Lehrach and John Wozney for gifts of enzymes, to D. Hirsh and B. Olsen for communicating results prior to publication, and to Alan Maxam for helpful advice. We thank Tom McDonnell for confirmational reading of autoradiographs. We are especially grateful to Paul Doty for direction and reading of the manuscript and for procuring the computer.

Supplementary Material Available

Additional sequence data, tables of restriction site coordinates, and computer programs can be obtained directly from the authors. Requests to view sequence autoradiographs and two-dimensional homology plots should also be directed to the authors.

References

Adams, E. (1978) *Science (Washington, D.C.) 202*, 591–598.
Adams, S. L., Alwine, J. C., deCrombrugghe, B., & Pastan, I. (1979) *J. Biol. Chem. 254*, 4935–4938.
Bachra, B. N. (1976) *J. Mol. Evol. 8*, 155–173.
Boedtker, H., Frischauf, A. M., & Lehrach, H. (1976) *Biochemistry 15*, 4765–4770.
Bornstein, P., & Byers, P. H. (1980) *Collagen Metabolism in Current Concepts*, The Upjohn Co., Kalamazoo, MI.
Bornstein, P., & Sage, H. (1980) *Annu. Rev. Biochem. 49*, 957–1003.
Brown, N. L., & Smith, M. (1977) *J. Mol. Biol. 116*, 1–30.
Carpousis, A., Christner, P., & Rosenbloom, J. (1977) *J. Biol. Chem. 252*, 8023–8026.
Dixit, S. N., Seyer, J. M., & Kang, A. H. (1977a) *Eur. J. Biochem. 73*, 213–221.
Dixit, S. N., Seyer, J. M., & Kang, A. H. (1977b) *Eur. J. Biochem. 81*, 599–607.
Dixit, S. N., Seyer, J. M., Kang, A. H., & Gross, J. (1978) *Biochemistry 17*, 5719–5722.
Dixit, S. N., Mainardi, C. L., Seyer, J. M., & Kang, A. H. (1979) *Biochemistry 18*, 5416–5422.
Eastoe, J. E. (1967) in *Treatise on Collagen* (Ramashandran, G. N., Ed.) Vol. I, pp 1–67, Academic Press, New York.
Efstratiadis, A., Kafatos, F. C., & Maniatis, T. (1977) *Cell 10*, 571–586.
Eyre, D. (1980) *Science (Washington, D.C.) 207*, 1315–1322.
Falvey, A. K., Weiss, B., Krueger, L. S., Kantor, J. A., & Anderson, W. F. (1976) *Nucleic Acids Res. 3*, 79–88.
Fessler, J. H., & Fessler, L. I. (1978) *Annu. Rev. Biochem. 47*, 129–162.
Fietzek, P. P., & Kühn, K. (1976) *Int. Rev. Connect. Tissue Res. 7*, 1–60.
Forget, B. G., Cavallesco, C., deRiel, J. K., Spritz, R. A., Choudary, R. V., Wilson, J. T., Wilso, L. B., Reddy, V. B., & Weissman, S. M. (1979) in *ICN-UCLA Symposia on Molecular and Cellular Biology* (Axel, R., Maniatis, T., & Fox, C. F., Eds.) Vol. 14, p 367, Academic Press, New York.
Garel, J. P. (1974) *J. Theor. Biol. 43*, 211–225.
Gopinathan, K. P., Weymouth, L. A., Kunkel, T. A., & Loeb, L. A. (1979) *Nature (London) 278*, 857–859.
Grantham, R., & Gautier, C. (1980) *Naturwissenschaften* (in press).
Grantham, R., Gautier, C., Gouy, M., Mercier, R., & Pave, A. (1980) *Nucleic Acids Res. 8*, 49–62.
Hardison, R. C., Butler, E. T., Lacy, E., Maniatis, T., Rosenthal, N., & Efstratiadis, A. (1979) *Cell 18*, 1285–1297.
Heilig, R., Perrin, F., Gannon, F., Mandel, J. L., & Chambon, P. (1980) *Cell 20*, 625–637.
Heindell, H. C., Liu, A., Paddock, G. V., Studnicka, G. M., & Salser, W. A. (1978) *Cell 15*, 43–54.
Highberger, J. H., Kang, A. H., & Gross, J. (1971) *Biochemistry 10*, 610–616.
Kang, A. H., & Gross, J. (1970) *Biochemistry 9*, 796–804.
Konkel, D. A., Maizel, J. V., Jr., & Leder, P. (1979) *Cell 18*, 865–873.
Lehrach, H., Frischauf, A. M., Hanahan, D., Wozney, J., Fuller, F., Crkvenjakov, R., Boedtker, H., & Doty, P. (1978) *Proc. Natl. Acad. Sci. U.S.A. 75*, 5417–5421.
Lehrach, H., Frischauf, A. M., Hanahan, D., Wozney, J., Fuller, F., & Boedtker, H. (1979) *Biochemistry 18*, 3146–3152.
Linsenmayer, T. F., & Toole, B. P. (1977) *Birth Defects: Orig. Art. Ser. 13*, 19–35.
Lomedico, P., Rosenthal, N., Efstratiadis, A., Gilbert, W., Kolodner, R., & Tizard, R. (1979) *Cell 18*, 545–558.
Maxam, A., & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 560–564.
Maxam, A., & Gilbert, W. (1980) *Methods Enzymol. 65*, 499–560.
Nakanishi, S., Inoue, A., Kita, T., Nakamura, M., Chang, A. C. Y., Cohen, S. N., & Numa, S. (1979) *Nature (London) 278*, 423–434.
NIH Guidelines for Recombinant DNA Research (1976) *Fed. Regist. 41*, 27902–27943.

NIH Guidelines for Recombinant DNA Research (1978) *Fed. Regist. 43*, 60080-60105.
Nishioka, Y., & Leder, P. (1979) *Cell 18*, 875-882.
O'Hare, K., Breathnach, R., Benoist, C., & Chambon, P. (1979) *Nucleic Acids Res. 7*, 321-334.
Ohmori, H., Tomizawa, J. I., & Maxam, A. M. (1978) *Nucleic Acids Res. 5*, 1479-1485.
Olsen, B. R., & Berg, R. A. (1979) in *Secretory Mechanisms* (Hopkins, C. R., & Duncan, C. J., Eds.) pp 57-78, Cambridge University Press, Cambridge, England.
Osterman, L. A. (1979) *Biochimie 61*, 323-342.
Prockop, D. J., Kivirikko, K. I., Tuderman, L., & Guzman, N. A. (1979a) *New Engl. J. Med. 301*, 13-23.
Prockop, D. J., Kivirikko, K. I., Tuderman, L., & Guzman, N. A. (1979b) *New Engl. J. Med. 301*, 77-85.
Proudfoot, N. J., & Brownlee, G. G. (1976) *Nature (London) 263*, 211-214.
Rave, N., Crkvenjakov, R., & Boedtker, H. (1979) *Nucleic Acids Res. 6*, 3559-3567.
Richards, R. I., Shine, J., Ullrich, A., Wells, J. R. E., & Goodman, H. M. (1979) *Nucleic Acids Res. 7*, 1137-1146.
Roskam, V. G., & Rougeon, F. (1979) *Nucleic Acids Res. 7*, 305-320.
Salser, W. A., Cummings, I., Liu, A., Strommer, J., Padayatty, J., & Clarke P. (1979) in *Cellular and Molecular Regulation of Hemoglobin Switching* (Stamatoyannopoulos, G., & Nienhuis, A. W., Eds.) p 621, Grune and Stratton, New York.
Showalter, A. M., Pesciotta, D. M., Eikenberry, E. F., Yamamoto, T., Pastan, I., deCrombrugghe, B., Fietzek, P. P., & Olsen, B. R. (1980) *FEBS Lett. 111*, 61-65.
Sobel, M. E., Yamamoto, T., Adams, S. L., Dilauro, R., Avvedimento, E. V., deCrombrugghe, B., & Pastan, I. (1978) *Proc. Natl. Acad. Sci. U.S.A. 75*, 5846-5850.
Tolstoshev, P., Haber, R., & Crystal, R. (1979) *Biochem. Biophys. Res. Commun. 87*, 818-826.
Turner, A. M., Bell, E., & Darnell, J. E. (1963) *Science (Washington, D.C.) 141*, 1187-1188.
van der Ploeg, L. H. T., & Flavell, R. A. (1980) *Cell 19*, 947-948.
van Ooyen, A., van den Berg, J., Mantel, N., & Weissmann, C. (1979) *Science (Washington, D.C.) 206*, 337-344.
von der Mark, H., von der Mark, K., & Gay, S. (1976) *Dev. Biol. 48*, 237-249.
Wozney, J., Hanahan, D., Fuller, F., & Boedtker, H. (1979) *Fed. Proc., Fed. Am. Soc. Exp. Biol. 38*, 617.
Yamamoto, T., Sobel, M. E., Adams, S. L., Avvedimento, V. E., DiLaure, R., Pastan, I., deCrombrugghe, B., Showalter, A., Pesciotta, D., Tietzek, P., & Olsen, B. R. (1980) *J. Biol. Chem. 255*, 2612-2615.

Partial Digestion of tRNA-Aminoacyl-tRNA Synthetase Complexes with Cobra Venom Ribonuclease†

Olga O. Favorova,‡ Franco Fasiolo,* Gérard Keith, Stanislas K. Vassilenko,§ and Jean-Pierre Ebel

ABSTRACT: Transfer RNA molecules have been labeled with ^{32}P at the 5′ or 3′ end and digested with cobra venom ribonuclease, which preferentially cuts double-stranded regions. The products of yeast $tRNA^{Phe}$ and $tRNA^{Val}$ were analyzed by high-resolution gel electrophoresis. In the free state, these tRNAs were cut predominantly in the acceptor and anticodon stems. Minor cuts occurred in the TΨ stem in $tRNA^{Val}$. The topography of zones interacting with their cognate synthetases was studied by determining the tRNA regions shielded by protein. Nearly 100% protection was found in the anticodon and acceptor stem of $tRNA^{Val}$, while in $tRNA^{Phe}$ only the stem of the anticodon was protected. Noncognate interactions between $tRNA^{Phe}$ and tryptophanyl-tRNA synthetase from beef pancreas were examined. The beef enzyme did not protect $tRNA^{Phe}$ despite the fact that efficient misaminoacylation occurred. The pattern of shielding obtained for each tRNA-synthetase complex was compared with the results of direct ultraviolet cross-linking experiments with these complexes.

The interaction between aminoacyl-tRNA synthetases and their cognate tRNAs has been the subject of considerable investigation over the past years [for recent general reviews, see Kisselev & Favorova (1974), Schimmel (1977), Ofengand (1977), and Goddard (1975)]. Among the approaches investigated, the use of partial digestion with nucleases as a method for the study of complex formation was applied to several systems by Hörz & Zachau (1973), Dube (1973), and Dickson & Schimmel (1975). The principle of this approach is that regions of the tRNA which are cut by the RNase become resistant when they are shielded by the protein. In all the systems studied so far, the ribonulceases employed (A, T_1, or T_2) generate cleavages in mild digestion conditions only in the nonhelical sections of the tRNA (Hörz & Zachau, 1973; Dickson & Schimmel, 1975).

The isolation from cobra Naja oxiana venom by one of the authors (S.V.) of an RNase without base specificity which preferentially cuts structured regions of an RNA (Vassilenko & Babkina, 1965; Vassilenko & Ryte, 1975) made it possible to determine the parts of the tRNA involved in the cloverleaf helical sections. We have digested yeast $tRNA^{Phe}$ and $tRNA^{Val}$, with these nucleases both in the free state and complexed with their synthetases. We compared these results

† From the Laboratoire de Biochimie, Institut de Biologie Moléculaire et Cellulaire du CNRS, 67084 Strasbourg Cedex, France. *Received May 9, 1980.*
‡ Permanent address: Institute of Molecular Biology of the Academy of Sciences of USSR, B334 Moscow, USSR.
§ Permanent address: Institute of Organic Chemistry, Siberian Branch of the Academy of Sciences, Novosibirsk 90, USSR.

0006-2960/81/0420-1006$01.00/0