Volume 7 Number 5 1979 Nucleic Acids Research

Molecular cloning and sequence analysis of adult chicken β globin cDNA

Robert I.Richards*†, John Shine[1†], Axel Ullrich[2†], Julian R.E.Wells* and Howard M.Goodman†

*Department of Biochemistry, University of Adelaide, Box 498, G.P.O., Adelaide, South Australia 5001, and †The Howard Hughes Medical Institute, Department of Biochemistry and Biophysics, University of California, San Francisco, CA 94143, USA

Received 4 September 1979

ABSTRACT

The molecular cloning and nucleotide sequence analysis of adult chicken β globin mRNA is reported. DNA sequences derived from in vitro transcription of globin mRNA were purified and amplified as recombinant DNA using the plasmid pBR322. Sequence analysis of several clones coding for β globin strongly suggests that transcription errors may be generated near the 5' end of transcripts in vitro by reverse transcription. The complete sequence of the longest β globin insert containing 51 bases of the 5' untranslated region as well as the complete coding and 3' untranslated regions has been determined.

INTRODUCTION

The globin gene system, in a variety of species, has become well established as a model for extensive analysis of eukaryote gene expression (3,4). This has been mainly due to the accessibility of the structural sequences by mRNA isolation, and the availability of in vivo systems for study such as cultured erythroblasts (5) and Friend erythroleukaemic cells (6).

In view of the available data on the mammalian systems, similar studies of the chicken globin genes will provide information concerning the divergence of structural and putative control regions. Analysis of globin amino acid sequences suggests that the chicken genes are relatively primative and lie close on the genealogical tree to the time of separation of globin into α and β chain types (7).

MATERIALS AND METHODS

Double stranded cDNA was synthesised from chicken globin mRNA (prepared as previously described (8)), by sequential reverse transcriptase reactions (9). Following S_1 nuclease treatment to open the hairpin loop, the DNA was made blunt-ended with E. coli DNA polymerase I (Klenow fragment) and ligated

to synthetic linker DNA encoding the HindIII recognition site (9). This material was then digested with HsuI (an isoschizomer of HindIII) and electrophoresed on a 6% polyacrylamide gel. DNA from the 500-700 bp region of the gel was electroeluted and ligated to HindIII digested, dephosphorylated pBR322 (9). The recombinant molecules were transformed into E. coli χ1776 as described (10).

Cells carrying recombinant plasmids were selected on the basis of their ampicillin resistance, tetracycline sensitivity and absence of the HaeIII digestion fragment of pBR322 containing the HindIII restriction site.

Plasmid DNA was prepared from those clones which gave HaeIII fragment patterns corresponding to HaeIII digested double-stranded globin cDNA. The inserted DNA was isolated by HindIII cleavage and electrophoresis on polyacrylamide gels. DNA fragments isolated for sequencing were end-labelled by either incubation with T4 polynucleotide kinase and γ-^{32}P-ATP or Klenow fragment catalyzed exchange of the 3' residue (11).

Sequencing reactions and gel electrophoresis were as described by Maxam and Gilbert (12).

All manipulations involving recombinant DNA were in accordance with NIH (USA) or ASCORD (Australia) guidelines.

RESULTS

Restriction endonuclease digestion of double-stranded cDNA transcribed from reticulocyte mRNA yielded information on the existence of cleavage sites in the major species present. The HaeIII digest gave two bands of approximately 260 base pairs each. Miniscreening of the recombinant DNA molecules with HaeIII indicated the presence of either one or the other of these two bands and suggested that these two fragments (and the clones containing them) were derived from two different major cDNA species. The cleavage patterns determined by digestion of the cDNA with HpaII, HhaI and AluI (Fig. 1) were used in establishing the sequencing strategy shown in Fig. 2. HpaI, XmaI and KpnI did not cleave the cDNA, whereas the cleavage site for PstI was subsequently found to lie in the β mRNA 3' untranslated region. SstI cleaved some of the cDNA sequences but no site has yet been determined in the β or α (13) cloned sequences.

Insert DNA isolated from several clones by HindIII digestion was labelled with T4 polynucleotide kinase and γ-^{32}P-ATP, digested with HaeIII restriction endonuclease and the labelled fragments separated on 6% polyacrylamide gels and subjected to sequence analysis (12). Sequences coding for adult β globin

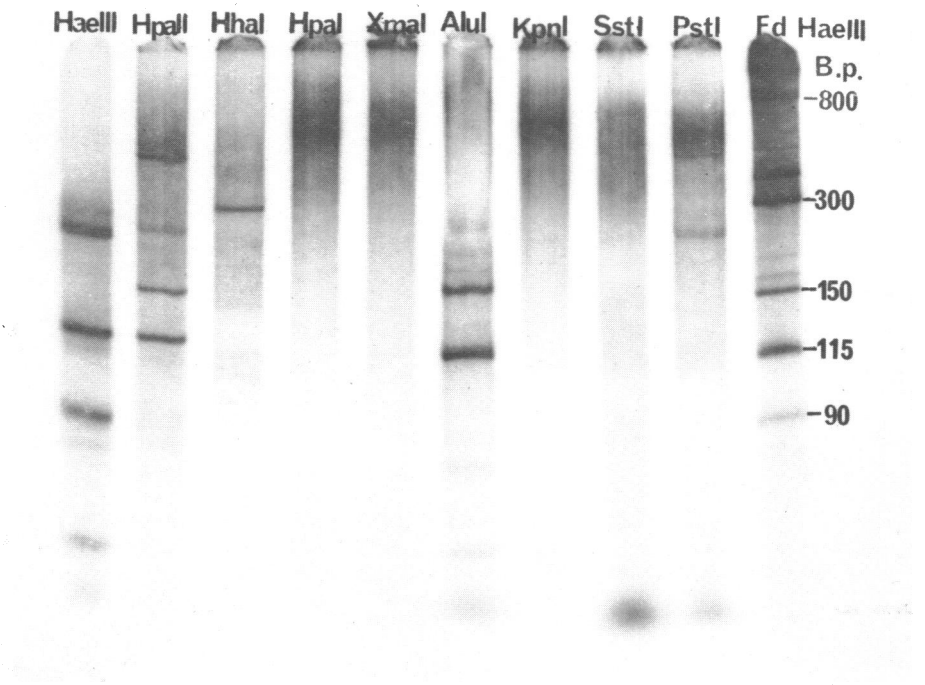

Figure 1. Restriction endonuclease digestion of chicken globin double-stranded cDNA. 3 x 10^4 c.p.m. (Cerenkov) of ^{32}P-labelled cDNA were incubated with each enzyme for 1 hour, electrophoresed on a 6% polyacrylamide slab gel and autoradiographed. Molecular weight markers are Fd phage DNA HaeIII fragments.

(14) and the partially characterized α globin (15) were found. The longest β coding insert, pCGβ-3, was completely sequenced (Fig. 3). In this case there was complete agreement with the amino acid sequence established by Matsuda et al. (14). Where possible, both strands were sequenced and where this proved difficult one strand was sequenced several times.

In addition to pCGβ-3, five other β coding inserts have been partially sequenced. All of these agree with corresponding sequences in pCGβ-3, except for a few bases confined to the 5' end of each insert (with respect to the mRNA sequence). This suggests that incorrect bases are inserted by reverse transcriptase during the "loop" formation or more likely arose during the repair process with E. coli DNA polymerase I in the blunt-ending reaction. It follows that several bases at the 5' end of pCGβ-3 may be incorrect.

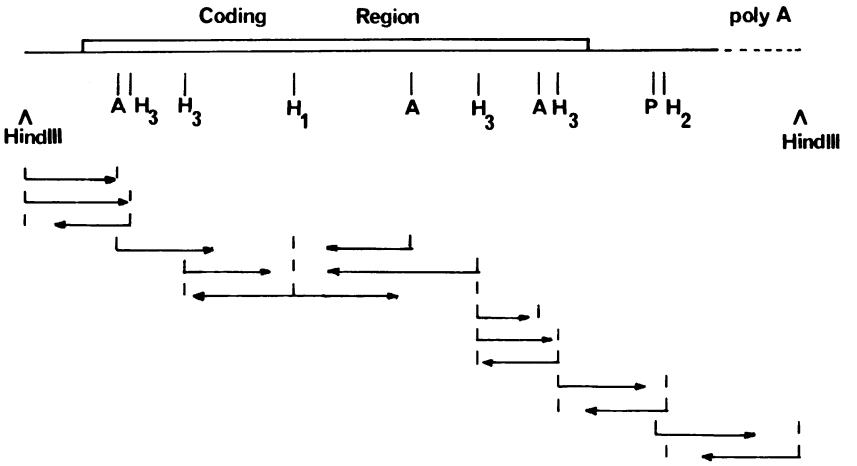

Figure 2. Sequencing strategy of pCGβ-3. Only those restriction sites used in the analysis are shown (A-AluI, H_1-HhaI, H_2-HpaII, H_3-HaeIII, P-PstI). Arrows indicate the direction and extent of sequencing.

Confirmation that these errors are due to in vitro reactions is obtained by inspection of the sequence of the β coding clone pCGβ-4 (Figs. 3 and 4) which terminates within the coding sequence and has an incorrect 5'-terminal sequence (Fig. 3 and 4). A similar result has been obtained for α coding clones (13). A potential model for the generation of errors is diagrammed in Figure 5.

DISCUSSION

The nucleotide sequence of chicken β-globin mRNA, as deduced from the sequence of the cloned cDNA, has a relatively high GC content (57% compared with 41% for the genome (16)). This is particularly evident in the redundant bases within the coding region in which 49% of codons are NNC and 30% NNG (Table 1). This selection in the mRNA sequence for a high GC content (also found in growth hormone (17) and chorionic somatomammotropin (18) mRNA) may result in a more stable overall secondary structure with a high degree of nuclease resistance. In addition, the stable secondary structure may be an essential feature of processing of precursor mRNA. In one possible conformation of β mRNA, many of the G and C residues in the third position of codons are involved in hydrogen bonding (13). Despite the high GC content there is a relatively low frequency of the

```
                  5'                                                    start
pCGβ-3                GCUCAGACCUCCUCCGUACCGACAGCCACACGCUACCC UCCAACCGCCGCC AUG
pCGβ-2                                  ...GGGAUAACACGCUACCC UCCAACCGCCGCC AUG
pCGβ-1                                     ...UAGCACGCUACCCCUCCAACCGCCGCC AUG

                                       10                                  20
val his trp thr ala glu glu lys gln leu ile thr gly leu trp gly lys val asn val
GUG CAC UGG ACU GCU GAG GAG AAG CAG CUC AUC ACC GGC CUC UGG GGC AAG GUC AAU GUG
pCGβ-4                                                                     GUG
                                       30                                  40
ala glu cys gly ala glu ala leu ala arg leu leu ile val tyr pro trp thr gln arg
GCC GAA UGU GGG GCC GAA GCC CUG GCC AGG CUG CUG AUC GUC UAC CCC UGG ACC CAG AGG
GCU GUC GGU GGG GCC GAA GCC CUG GCC AGG
                                       50                                  60
phe phe ala ser phe gly asn leu ser ser pro thr ala ile leu gly asn pro met val
UUC UUU GCG UCC UUU GGG AAC CUC UCC AGC CCC ACU GCC AUC CUU GGC AAC CCC AUG GUC

                                       70                                  80
arg ala his gly lys lys val leu thr ser phe gly asp ala val lys asn leu asp asn
CGC GCC CAC GGC AAG AAA GUG CUC ACC UCC UUU GGG GAU GCU GUG AAG AAC CUG GAC AAC

                                       90                                 100
ile lys asn thr phe ser gln leu ser glu leu his cys asp lys leu his val asp pro
AUC AAG AAC ACC UUC UCC CAA CUG UCC GAA CUG CAU UGU GAC AAG CUG CAU GUG GAC CCC

                                      110                                 120
glu asn phe arg leu leu gly asp ile leu ile ile val leu ala ala his phe ser lys
GAG AAC UUC AGG CUC CUG GGU GAC AUC CUC AUC AUU GUC CUG GCC GCC CAC UUC AGC AAG

                                      130                                 140
asp phe thr pro glu cys gln ala ala trp gln lys leu val arg val val ala his ala
GAC UUC ACU CCU GAA UGC CAG GCU GCC UGG CAG AAG CUG GUC CGC GUG GUG GCC CAU GCC

                          stop
leu ala arg lys tyr his
CUG GCU CGC AAG UAC CAC  UAA   GCACCAGCACCAAAGAUCACGGAGCACCUACAACCAUUGCAUGCACCU

                                                                          3'
GCAGAAAUGCUCCGGAGCUGACAGCUUGUGACAAAUAAAGUUCAUUCAGUGACACUC poly(A)
```

Figure 3. Complete nucleotide sequence of the mRNA corresponding to pCGβ-3. Sequences of pCGβ-1, pCGβ-2 and pCGβ-4 are included to show 5' terminal heterogeneity.

C - G doublet, normally characteristic of eukaryotic DNA (19) (19 C - G compared with 43 G - C).

Kafatos et al. (20) have carried out an extensive analysis of the

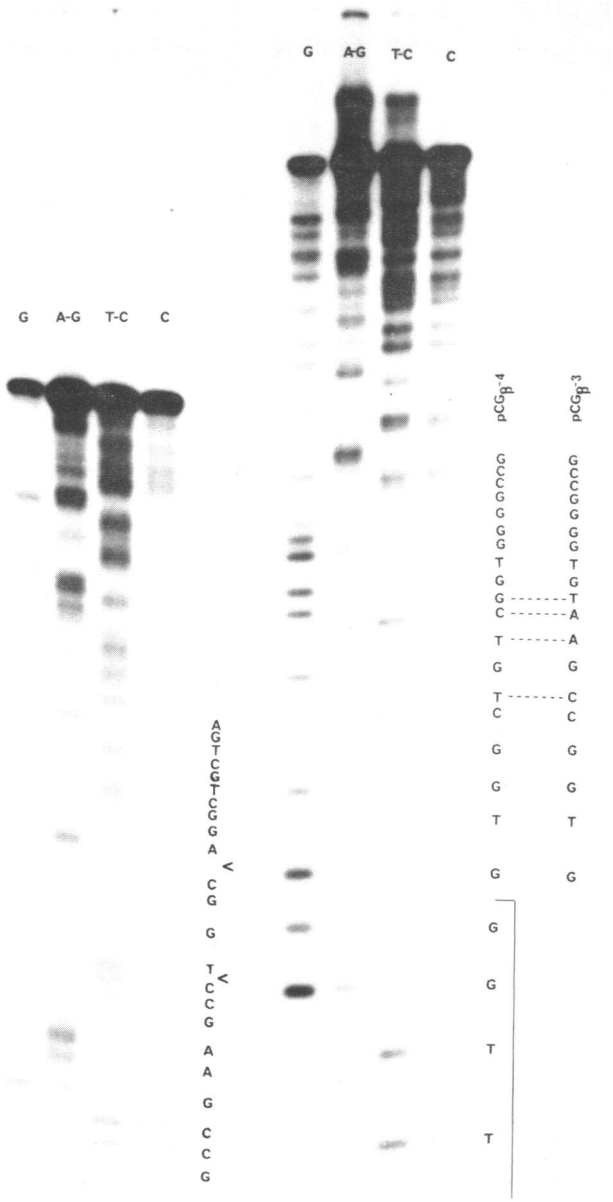

Figure 4. Autoradiograph of sequencing gel showing the 5' terminal sequence of pCGβ-4. The relevant sequence of pCGβ-3 is included to show heterogeneity. The bracket shows the base sequence from the synthetic HindIII linker DNA. The arrows mark sites where C residues are absent from the sequence due to methylation of EcoRII sites by E. coli.

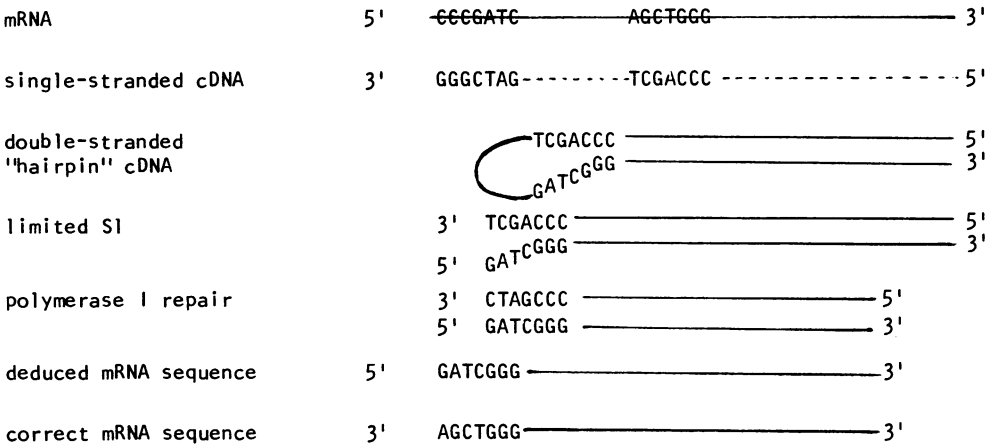

Figure 5. Potential mechanism for generation of "errors" at the 5'-end of cDNA. Limited S1 cleavage of the hairpin-loop generated during synthesis of dS cDNA results in a duplex molecule terminating in two non-paired strands. The unpaired 3'-end of the strand complementary to the 5'-terminus of mRNA is then removed by the 3'-5' exonuclease activity of DNA polymerase I and then re-synthesised using the other unpaired strand as template. This results in the incorporation of "incorrect" sequences since this portion of the cDNA is derived from sequences further towards the 5' end of the mRNA.

homology between rabbit and human β globin. Comparison of the chicken sequence with those derived from rabbit (21,22,23) and human (22,24,25) reveals some interesting features. In pCGβ-3, of the 51 bases in the 5' untranslated region (AUG not included) there are at most 25 bases homologous with rabbit and 30 bases homologous with human (Fig. 6). No more than four contiguous bases are found to be homologous in any part of this region suggesting that if a ribosome binding site exists, analogous to that in prokaryotes, then there is little selective pressure to maintain the specificity of this sequence. A similar conclusion has been previously arrived at by a comparison of 5'- untranslated regions from a variety of eukaryote mRNAs (26). In the case of both human and rabbit β mRNA there are six more bases (not shown in Fig. 6) before the 7meG cap which suggests that pCGβ-3 does not contain the complete 5'-untranslated region of chicken β globin.

Comparison of respective 3'-untranslated regions shows a similar degree of homology as described for the 5' ends (Fig. 7). Of the 108 bases in the chicken sequence, 57 are homologous with rabbit and 52 with human. For rabbit and human β globin mRNAs there is a region immediately after

	U		C		A		G		
U	Phe	3	Ser	-	Tyr	-	Cys	2	U
	Phe	5	Ser	5	Tyr	2	Cys	1	C
	Leu	-	Ser	-	Term	1	Term	-	A
	Leu	-	Ser	-	Term	-	Trp	4	G
C	Leu	1	Pro	1	His	3	Arg	-	U
	Leu	6	Pro	4	His	4	Arg	3	C
	Leu	-	Pro	-	GluN	1	Arg	-	A
	Leu	11	Pro	-	GluN	4	Arg	-	G
A	Ile	1	Thr	3	AspN	1	Ser	-	U
	Ile	6	Thr	4	AspN	6	Ser	2	C
	Ile	-	Thr	-	Lys	1	Arg	-	A
	Met	1	Thr	-	Lys	9	Arg	3	G
G	Val	-	Ala	4	Asp	1	Gly	1	U
	Val	5	Ala	11	Asp	5	Gly	4	C
	Val	-	Ala	-	Glu	4	Gly	-	A
	Val	7	Ala	1	Glu	3	Gly	3	G

Table 1. Codon utilization of chicken β globin mRNA. The Table shows the preference for codons ending in G or C, and the discrimination against those containing C-G.

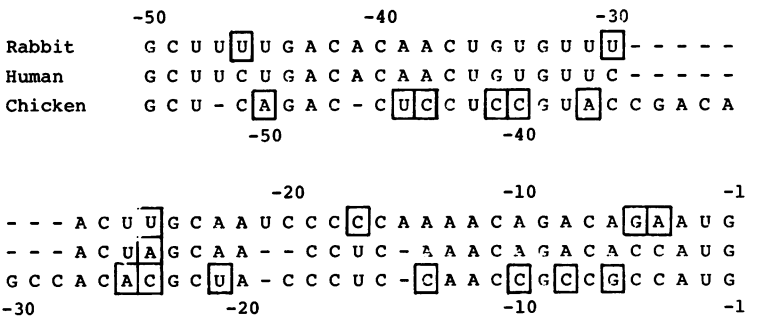

Figure 6. Comparison of the 5' untranslated regions of chicken, human (19, 20) and rabbit (21) β globin mRNA. Boxes show base changes between the three sequences. Sequences have been aligned to show maximum homology.

the termination codon of complete divergence and the same is true for the chicken sequence. This is followed by a region of homology, which contains 24 deletions (or insertions) and 10 base changes (chicken compared to rabbit), up to the highly conserved AAUAAA sequence. From this sequence to the poly(A) tract there is only one deletion but 8 base changes (both in rabbit and human) suggesting that the length is more critical than the sequence.

Comparison of the chicken and rabbit β chain coding regions shows 120 base changes. Of these, 66 are involved in an amino acid change at that position, while the remaining 54 conserve the protein sequence. The 72% nucleotide sequence homology between the chicken and rabbit coding regions is significantly higher than the homology between the non-coding regions (49% 5' end; 54% 3' end); a result consistent with selection at the amino acid level being a significant factor in the maintenance of nucleotide sequence.

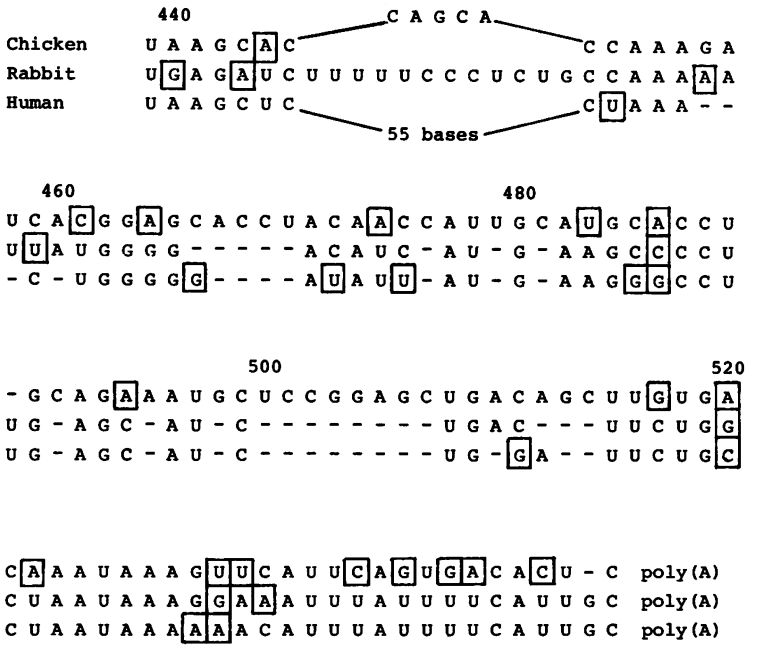

Figure 7. Homology in the 3' untranslated regions of chicken, human (19, 20) and rabbit (21) β globin mRNA. Boxes show base changes between the three sequences. Sequences have been aligned to show maximum homology.

ACKNOWLEDGEMENTS

We thank the Office of Program Resources and Logistics Viral Cancer Program, Viral Oncology, Division of Cancer Cause and Prevention, National Cancer Institute, Bethesda, Maryland for Reverse Transcriptase. R.I.R. is the recipient of a Commonwealth Postgraduate Research Award and a grant from the University of Adelaide Research and Publications Committee. R.I.R. is grateful to Professor W.J. Rutter for encouragement and permission to work in his department. This work was supported in part by grants from A.R.G.C. (D2-76/16788) to J.R.E.W. and NIH (CA14026) to H.M.G.

REFERENCES

1. Present address: Molecular Biology Unit, Research School of Biological sciences, Australian National University, Canberra, Australia.
2. Present address: Genentech, Inc., 460 Point San Bruno Blvd., So. San Francisco, California 94080, USA.
3. Weatherall, D.J. and Clegg, J.B. (1979) Cell, 16, 467-479.
4. Van den Berg, J., Van Ooyen, A., Mantei, N., Schambock, A., Grosveld, G., Flavell, R.A. and Weissmann, C. (1978) Nature, 276, 37-44.
5. Crawford, R.J. and Wells, J.R.E. (1978) Biochemistry, 17, 1591-1596.
6. Lo, S.-C., Aft, R., Ross, J. and Mueller, G.C. (1978) Cell, 15, 447-453.
7. Goodman, M., Moore, G.W. and Matsuda, G. (1975) Nature, 253, 603-608.
8. Crawford, R.J., Scott, A.C. and Wells, J.R.E. (1977) Eur. J. Biochem., 72, 291-299.
9. Seeburg, P.H., Shine, J., Martial, J.A., Baxter, J.D. and Goodman, H.M. (1977) Nature, 270, 486-494.
10. Goodman, H.M. and MacDonald, R.J. (1979) Methods in Enzymology, in press.
11. Donelson, J.E. and Wu, R. (1972) J.Biol.Chem., 247, 4654-4660.
12. Maxam, A. and Gilbert, W. (1977) Proc. Natl. Acad. Sci. USA, 74, 560-564.
13. Richards, R.I., manuscript in preparation.
14. Matsuda, G., Maita, T., Mizuno, K. and Ota, H. (1973) Nature New Biol., 244, 244.
15. Cummings, I.W., Liu, A.Y. and Salser, W.A. (1978) Nature, 276, 418-419.
16. Sinclair, J.H. and Brown, D.D. (1971) Biochemistry, 10, 2761-2769.
17. Martial, J.A., Hallewell, R.A., Baxter, J.D. and Goodman, H.M. (1979) Science, 205, 602-606.
18. Shine, J., Seeburg, P.H., Martial, J.A., Baxter, J.D. and Goodman, H.M. Nature, 270, 494-499.
19. Russell, G.J., Walker, P.M.B., Elton, R.A. and Subak-Sharpe, J.H. (1976) J.Mol.Biol., 108, 1-15.
20. Kafatos, F., Efstratiadis, A., Forget, B.G. and Weissman, S.M. (1977) Proc. Natl. Acad. Sci. USA, 74, 5618-5622.
21. Baralle, F. (1977) Cell, 10, 549-558.
22. Proudfoot, N.J. (1977) Cell, 10, 559-570.
23. Efstratiadis, A., Kafatos, F.C. and Maniatis, T. (1977) Cell, 10, 571-585.
24. Baralle, F. (1977) Cell, 12, 1085-1095.
25. Marotta, C.A., Wilson, J.T., Forget, B.G. and Weissman, S.M. (1977) J.Biol.Chem., 252, 5040-5053.
26. Baralle, F. and Brownlee, G.G. (1978) Nature, 274, 84-87.