Biochem. J. (1984) **222**, 561–570
Printed in Great Britain

Nucleotide sequence determination of guinea-pig casein B mRNA reveals homology with bovine and rat α_{s1} caseins and conservation of the non-coding regions of the mRNA

Len HALL,* Jane E. LAIRD and Roger K. CRAIG
Courtauld Institute of Biochemistry, The Middlesex Hospital Medical School, London W1P 7PN, U.K.

(*Received 12 January 1984/Accepted 30 May 1984*)

Nucleotide sequence analysis of cloned guinea-pig casein B cDNA sequences has identified two casein B variants related to the bovine and rat α_{s1} caseins. Amino acid homology was largely confined to the known bovine or predicted rat phosphorylation sites and within the 'signal' precursor sequence. Comparison of the deduced nucleotide sequence of the guinea-pig and rat α_{s1} casein mRNA species showed greater sequence conservation in the non-coding than in the coding regions, suggesting a functional and possibly regulatory role for the non-coding regions of casein mRNA. The results provide insight into the evolution of the casein genes, and raise questions as to the role of conserved nucleotide sequences within the non-coding regions of mRNA species.

Until the advent of recombinant DNA technology, detailed structural analysis of the major protein components of milk, the acid-precipitable caseins, was confined predominantly to those isolated from bovine milk. Such studies resulted in the determination of the amino acid sequence of the four major caseins [α_{s1} (Mercier *et al.*, 1971; Grosclaude *et al.*, 1983), α_{s2} (Brignon *et al.*, 1977), β (Grosclaude *et al.*, 1983; Ribadeau-Dumas *et al.*, 1972) and κ (Mercier *et al.*, 1973)], demonstrated that a number of minor casein components result from the post-translational proteolytic cleavage of the β-casein (Gordon *et al.*, 1972), and revealed the existence of genetic variants, presumed to reflect gene deletions or point mutations leading to altered gene products (Thompson & Farrell, 1974). Available amino acid sequence data also suggested that the caseins from different species exhibited a high rate of sequence divergence, although short regions containing multiple phosphoserine residues showed homology. The latter observations indicate that part at least of the different casein polypeptides may have evolved from a common ancestral gene (Ribadeau-Dumas *et al.*, 1975).

The cloning and nucleotide sequence determination of cDNA representative of rat α_{s1} (Hobbs & Rosen, 1982), bovine (Willis *et al.*, 1982) and guinea-pig (Hall *et al.*, 1984) casein mRNA species, has provided new insight into the manner in which the casein-gene family has evolved, in addition to providing hybridization probes necessary to initiate studies on the genomic organization of the casein-gene family and the manner in which hormones modulate their expression. We have already described the complete nucleotide sequence of two guinea-pig milk protein mRNA species, those encoding α-lactalbumin (Hall *et al.*, 1982) and casein A (Hall *et al.*, 1984). Here we present the nucleotide sequence of an abundant guinea-pig mRNA encoding casein B, and provide preliminary evidence for the existence of a less-abundant mRNA encoding a casein B variant. Nucleotide sequence analysis of cloned cDNA has revealed that the two casein B mRNA species differ by a single nucleotide within the coding region (G or T), giving rise to a premature termination codon (TAA) in the minor species.

Comparison of the deduced amino acid sequence of the major guinea-pig casein B mRNA with published amino acid sequences of caseins from other species reveals regions of homology with bovine and rat α_{s1} caseins. Sequence conservation was particularly striking within the signal peptide and at potential phosphorylation sites, suggesting that selective pressures have maintained these functionally important domains. Similar comparisons of guinea-pig casein B and rat α_{s1} casein mRNA species at the nucleotide level

Abbreviation used: Hepes, 4-(2-hydroxyethyl)-1-piperazine-ethanesulphonic acid.

* Present address and address for correspondence and reprint requests: Department of Biochemistry, University of Bristol, University Walk, Bristol BS8 1TD, U.K.

revealed unusually high conservation of sequence within the 5′ and 3′ non-coding regions. Further striking homology was observed within the 5′ non-coding region and that encoding the signal peptide of a number of mRNA species encoding different caseins from several species. The possible implications of these findings are discussed.

Materials and methods

Materials

Restriction endonucleases *Bst*NI, *Fnu*4HI and *Rsa*I were purchased from Biolabs through CP Laboratories, Bishop's Stortford, Herts., U.K.; all other restriction enzymes, phage-T4 DNA polymerase and phage-T4 polynucleotide kinase were from Bethesda Research Labs., Cambridge, U.K. Avian-myeloblastosis-virus reverse transcriptase (batch no. G-91180) was provided by Dr. J. W. Beard, Life Sciences, St. Petersburg, FL 33707, U.S.A.

Recombinant plasmids

The construction and characterization of pgpK27, a recombinant plasmid containing a cDNA copy of guinea-pig casein B mRNA inserted into the *Eco*RI site of pAT153 using homopolymeric dA–dT tails, has been described previously (Craig *et al.*, 1981). A second recombinant, pgpK48, containing a cDNA insert with a restriction map identical with that of pgpK27, but inserted into the pAT153 vector in the opposite orientation, was subsequently isolated from the same cDNA library.

Chemical DNA sequencing of recombinant plasmids

Procedures for the preparation of restriction fragments, 5′- and 3′-end labelling of DNA and Maxam–Gilbert chemical sequencing of the labelled products were exactly as described previously (Hall *et al.*, 1982, 1984).

Chemical DNA sequencing after primer extension

The 382-base-pair *Sal*I–*Pst*I fragment (see Fig. 1*a*) was excised and gel purified from 100 μg of pgpK27 plasmid DNA (Maxam & Gilbert, 1980). A 5 μg portion of this fragment was restricted with *Fnu*4HI, then denatured by incubation at 100°C for 3 min in water. The DNA fragments were then labelled at their 5′-termini with phage-T4 kinase and [γ-^{32}P]ATP, and the products purified by electrophoresis on a denaturing 8% (w/v)-polyacrylamide/urea sequencing gel, and two separate primers, the 21-base *Sal*I–*Fnu*4HI (primer A) and the adjacent 91-base *Fnu*4HI–*Fnu*4HI (primer B) fragments, eluted (Maxam & Gilbert, 1980). Post-nuclear poly(A)-containing RNA (15 μg), isolated from lactating-guinea-pig mammary gland (Craig *et al.*, 1979), was heated at 95°C for 2 min in water to minimize secondary structure and then incubated with either 40 ng of primer A or 140 ng of primer B in 150 μl of 50 mM-Hepes, pH 7.6, containing 0.45 M-NaCl, 1 mM-EDTA, 0.1% (w/v) sodium dodecyl sulphate, 25% (v/v) formamide, at 43°C for 4 h. The hybrid was then purified by ethanol precipitation, extended with avian-myeloblastosis-virus reverse transcriptase, and the products either analysed on denaturing 8% polyacrylamide/urea sequencing gels or subjected to Maxam–Gilbert chemical sequencing as described previously (Hall *et al.*, 1984). Primer extension using a 60-base *Pst*I–*Hin*dIII fragment (see Fig. 1*a*) labelled with phage-T4 polynucleotide kinase and [γ^{32}-P]ATP at the *Hin*dIII site, was performed in an identical manner.

Alignment of sequences for homology analyses

Guinea-pig, rat and bovine casein protein sequences could be readily aligned by eye. Nucleotide sequence alignment of guinea-pig casein B and rat α_{s1} casein cDNA species was initially performed by eye, inserting gaps where necessary to maximize alignment. This was then confirmed by applying DIAGON, an interactive computer-graphics program (Staden, 1982) designed for comparing and aligning nucleic acid and amino acid sequences. In particular, this program is ideal for determining the points of insertions or deletions, which appear on the two-dimensional plot as parallel diagonals.

Results and discussion

Nucleotide sequence analysis and deduction of amino acid sequence

By using the sequencing strategy shown in Fig. 1(*a*) the complete nucleotide sequence of the casein B cDNA insert contained within recombinant plasmid pgpK27 was determined by the method of Maxam & Gilbert (1980) (Fig. 2). Although this contained the entire 3′ non-coding region of the mRNA [as indicated by the presence of a 3′ poly(A) tail], its size of 788 base-pairs suggested that a significant proportion of the 5′-end was absent from this clone, when compared with previous estimates of the size of guinea-pig casein B mRNA (1050 $\pm$ 100 nucleotides) determinated by Northern blot analysis (Craig *et al.*, 1981). In an attempt to isolate clones containing casein B mRNA sequence absent from pgpK27, a *Hinf*I–*Hha*I fragment was isolated from the 5′-end of the pgpK27 sequence (see Fig. 1*a*), radiolabelled with [γ-^{32}P]ATP and phage-T4 polynucleotide kinase, and used to rescreen the original cDNA library (Craig *et al.*, 1981). A number of casein B clones

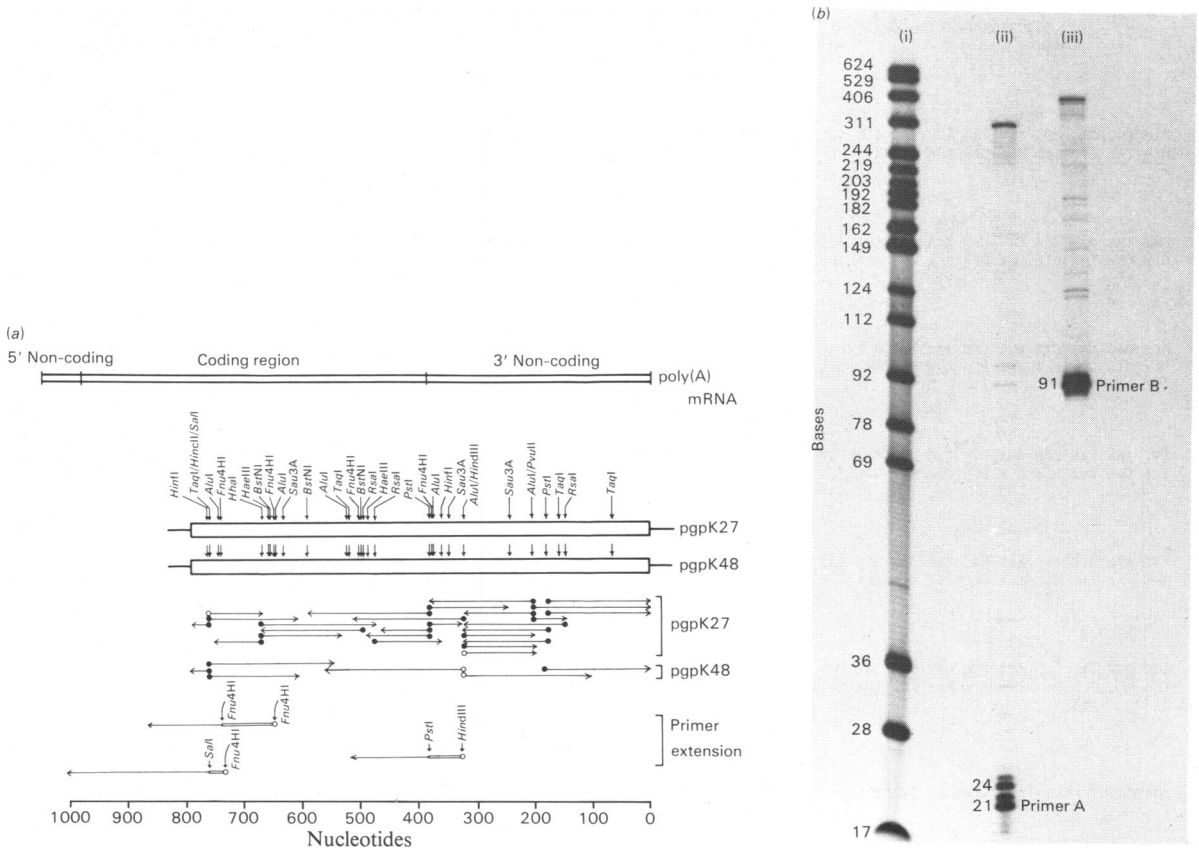

Fig. 1. *Guinea-pig casein B nucleotide-sequencing strategy*

(*a*) Restriction maps and sequencing strategy for guinea-pig casein B cDNA recombinant plasmids. Arrows indicate the direction and extent of sequencing of individual fragments; closed and open circles represent 3'- and 5'-end labelling respectively. (*b*) Determination of the 5'-sequence of guinea-pig casein B mRNA by primer extension. Extended primers (see the Materials and methods section) were analysed on a denaturing 8%-polyacrylamide/urea sequencing gel. Track (i), ^{32}P end-labelled, *Hpa*II-restricted pBR322 plasmid DNA marker fragments; track (ii), products of primer A extension; track (iii), products of primer B extension.

were identified and further characterized, but none was found to contain additional sequence to that present in pgpK27. However, one recombinant, pgpK48, contained a cDNA fragment identical in length with that cloned in pgpK27 (Figs. 1*a* and 2), but inserted in the opposite orientation within the pAT153 vector. DNA sequence analysis of the entire inserted cDNA sequence showed that, with the exception of a single base (see below), the cDNA species cloned into pgpK48 and pgpK27 were identical. Each contained a 3' poly(A) tract and terminated with the same nucleotide at the 5'-end.

To obtain additional casein B mRNA sequence information, two separate DNA primers were isolated from the 5'-end of the pgpK27 inserted sequence (see Fig. 1*a* and the Materials and methods section), hybridized to post-nuclear poly(A)-containing RNA from lactating guinea-pig mammary gland, and the resulting primer-templates extended with avian-myeloblastosis-virus reverse transcriptase. Analysis of the extended primers by electrophoresis on denaturing polyacrylamide/urea sequencing gels revealed a single major band in each case (Fig. 1*b*), as well as a number of minor bands created by premature termination of reverse transcription, which is common with this method and has been found to occur predominantly at C residues in the mRNA (Hall *et al.*, 1984). The major band in each case is consistent with an mRNA size of 1037 $\pm$ 10 nucleotides [excluding the poly(A) tail], very similar to that obtained by Northern blot analysis. Subsequent Maxam–Gilbert sequence analysis of these extended primers yielded 212 nucleotides of casein B mRNA sequence additional to that

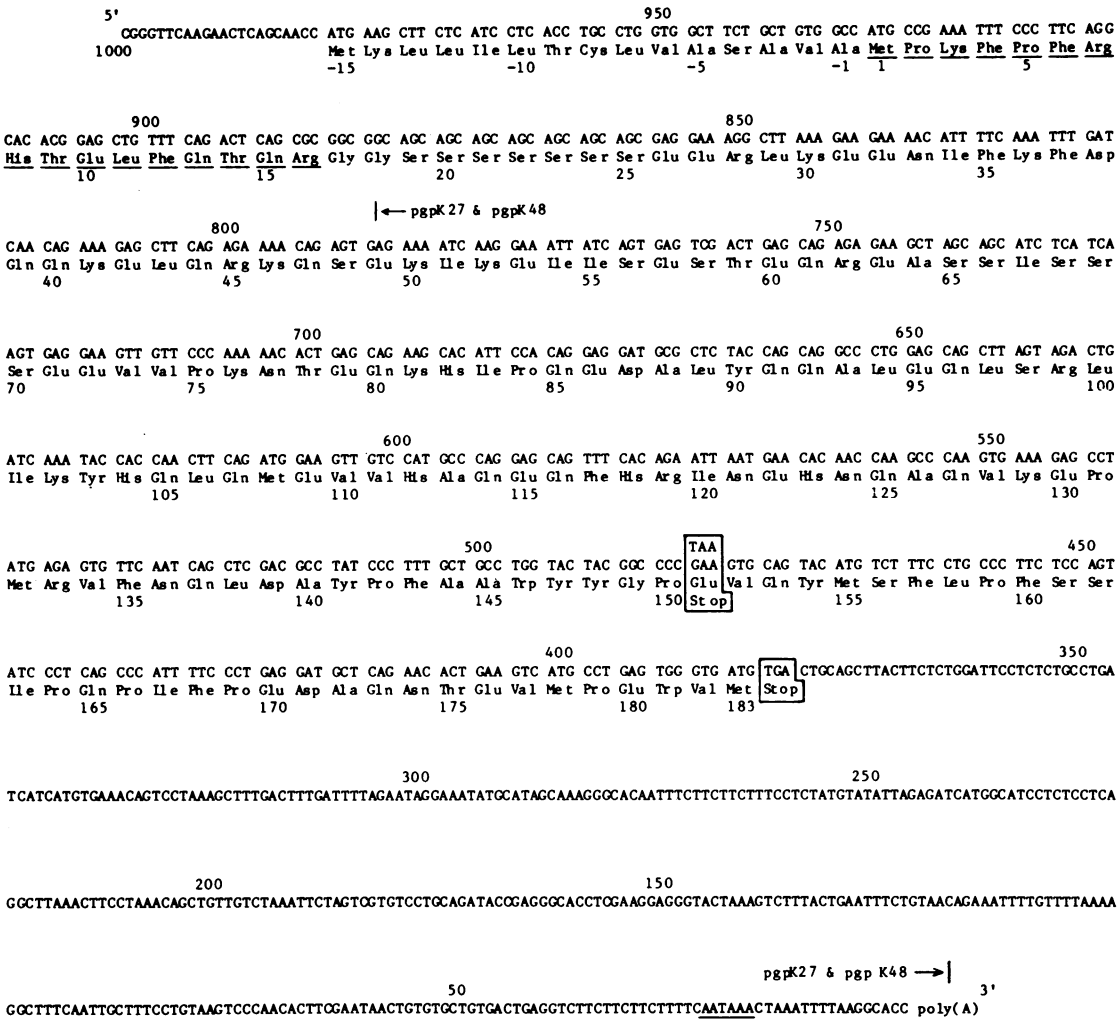

Fig. 2. *Nucleotide sequence of guinea-pig casein B cDNA*

The amino acid sequence (residues −15 to −1 refer to the pre-sequence; residues 1 to 183 represent the secreted protein) was deduced from the nucleotide sequence. Underlined residues were confirmed by limited protein-sequence analysis. Arrows indicate the 5'- and 3'-ends of cDNA species inserted in recombinant plasmids pgpK27 and pgpK48. The remaining 5'-sequence was determined by primer extension. The supposed polyadenylation enzyme recognition sequence is shown underlined. Boxed regions indicate the alternative termination codons found in plasmids pgpK27 and pgpK48 (see the text).

present in pgpK27 and pgpK48 (Fig. 2). However, with this technique it was not possible to determine with confidence the last 40 nucleotides at the extreme 5'-end of the casein B mRNA.

One reservation of this technique is that the 5'-end of the casein sequence has been obtained by primer extension of a population of mRNA species. It is therefore possible that the sequence derived by this means represents an mRNA species different from that used to generate pgpK27 and pgpK48. However, we would argue that this is extremely unlikely for several reasons: (i) hydridization conditions were suitably stringent during primer annealing (see the Materials and methods section); (ii) primer extension yielded a single major band with each primer (Fig. 1*b*), in agreement with the expected length for casein B mRNA obtained by Northern analysis; (iii) there is perfect agreement between the 50 nucleotides located at the extreme 5'-end of the pgpK27 inserted sequence obtained by sequencing the cloned cDNA, with that determined independent-

ly by primer extension using primer B; (iv) pgpK27 DNA is able to specifically bind only casein B mRNA, as indicated by hybrid-released translation assays *in vitro* (Craig *et al.*, 1981), and subsequent product analysis by antibody precipitation; (v) the sequence obtained by primer extension contains a region encoding the known *N*-terminal protein sequence (16 amino acid residues) of guinea-pig casein B (see Fig. 2). All available evidence, therefore, supports the belief that the sequence data obtained by primer extension and from analysis of the cloned cDNA sequence are derived from the same RNA species.

In view of the ease with which reverse transcription proceeded to the 5'-end of the mRNA during primer extension, it is difficult to explain why the two inserted cDNA sequences within the casein B clones terminate at the same point 788 nucleotides from the poly(A) tail. One possibility is that oligo(dT) may hybridize to the A-rich region located 836–846 nucleotides from the poly(A) tail (see Fig. 2), thereby prematurely terminating reverse transcription from the authentic poly(A) tail. In this respect it is interesting to note that the region immediately adjacent to the internal poly(A) tract (ATTTTC; nucleotides 829–834) is complementary to the 5'-end of the inserts within pgpK27 and pgpK48 (TAAAAG; nucleotides 782–787), thereby permitting a 'hairpin' configuration which, after S_1 cleavage during clone construction, would produce inserts of the length found in pgpK27.

Nucleotide sequence analysis of the two cloned cDNA sequences in combination with primer extension and subsequent nucleotide sequence analysis resulted in the determination of 1000 nucleotides of the casein B mRNA. This contained a single open reading frame encoding, in pgpK48, 198 amino acids, and, in pgpK27, 165 amino acids, before the first in-phase stop codon, the difference reflecting a single base change, G to T (nucleotide 483), converting the glutamic acid residue encoded within pgpK48 into a premature termination codon in pgpK27. In all other respects the nucleotide sequence within the coding and non-coding regions of the two cloned cDNAs was identical. The deduced amino acid sequences comprised a 15 amino acid hydrophobic 'signal' peptide, typical of most secretory proteins (Von Heijne, 1983), followed by 150 or 165 amino acids of the mature secreted proteins (see Fig. 2). Examination of the latter revealed an *N*-terminal methionine residue, the known *N*-terminal amino acid of mature guinea-pig casein B (Craig *et al.*, 1978), followed by a sequence of 15 amino acids, which were in agreement with the known *N*-terminal sequence of guinea-pig casein B (R. K. Craig, unpublished work). Thus the cDNA sequences cloned within pgpK27 and pgpK48 encode two casein B variants, an observation consistent with previous data, which indicated the existence of several forms, as judged by the presence of multiple casein B bands after translation of the purified casein B mRNA *in vitro* (see Pascall *et al.*, 1981).

In order to establish which of these cloned casein B mRNA species represented the predominant form in the lactating mammary gland, a further primer extension was performed. This utilized a denatured 60-base-pair *Pst*I–*Hin*dIII fragment, radiolabelled with [γ^{32}P]ATP and polynucleotide kinase, then hybridized to lactating-mammary-gland poly(A)-containing RNA and extended with reverse transcriptase. DNA sequence analysis of the extended primer (results not shown) identified a guanosine residue at nucleotide-483, demonstrating that the predominant casein B mRNA was the form cloned in pgpK48, and that the higher-M_r (21588) variant of casein B rather than the lower-M_r (17722) form of casein B predominated. The physiological significance of the smaller casein B variant is unknown, but may reflect the expression of two separate casein B genes, though we cannot exclude DNA polymorphism, as the mRNA used to prepare the original cDNA library was obtained from more than one animal. However the result demonstrates that lower-M_r casein variants may be generated by a single base change in the genomic DNA, and may not always reflect proteolytic cleavage of a higher-M_r parental protein (see Gordon *et al.*, 1972). Finally, at this stage we cannot rule out the possibility that the minor variant is simply a cloning artefact, although this seems unlikely since, as stated above, translation of purified casein B mRNA *in vitro* results in multiple casein bands. Definitive proof of multiple casein B genes, however, must await analysis of the genomic sequences.

Alignment of the deduced guinea-pig casein B protein sequence (excluding the signal sequence) with the various published bovine and rat caseins revealed homology only with the α_{s1} species (Fig. 3). This was largely confined to the positions and sequences of the known bovine or predicted rat casein kinase phosphorylation sites, and the signal pre-sequences, although other short regions of homology were also apparent. Although it is not known which serine residues are phosphorylated in the guinea-pig and rat caseins, data on the phosphorylation patterns of a number of caseins (Mercier, 1981) suggest that the sequence Ser-Xaa-A (where Xaa represents any amino acid and A is an acidic residue, usually glutamic acid, or a previously phosphorylated serine residue) is required for serine phosphorylation *in vivo*. If this is the case, then both of the major bovine phosphorylation sites are conserved in the guinea-pig

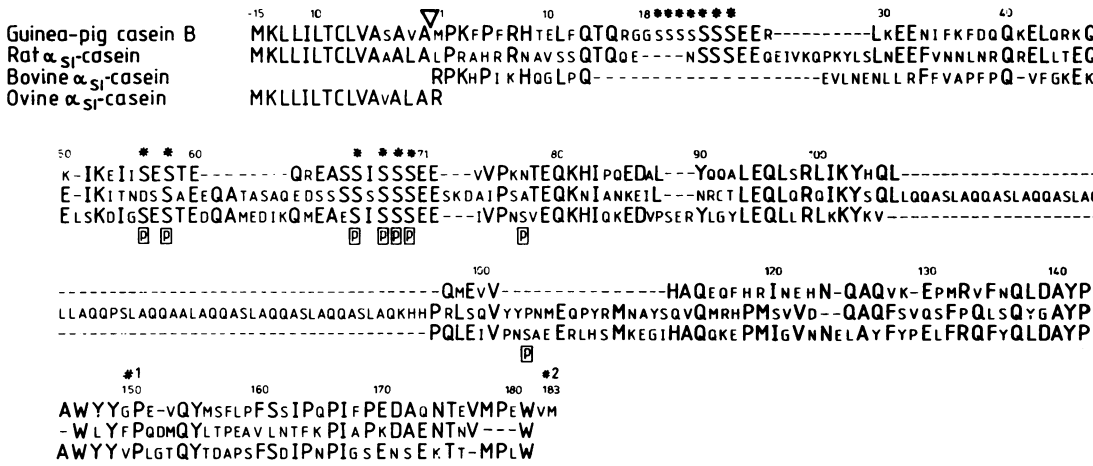

Fig. 3. *Comparison of the deduced amino acid sequence of guinea-pig pre-(casein B) with rat pre-(α_{s1}-casein), mature bovine α_{s1}-casein and the ovine α_{s1}-casein pre-sequence*

Conserved amino acids are shown in large capital letters. #1 and #2 represent the *C*-terminal residues of the minor and major casein B variants. Known phosphoserine residues in the bovine sequence (Mercier *et al.*, 1971) are indicated. Potential phosphoserine residues in the guinea-pig sequence are indicated by asterisks. Broken lines represent gaps introduced to maximize homology between the sequences. Rat and bovine sequences were obtained from Hobbs & Rosen (1982) and Mercier *et al.* (1978) respectively.

sequence (residues 66–72 and 56–60). However, guinea-pig casein B contains an additional potential phosphorylation site (residues 19–27), partially deleted in the rat sequence and totally absent from the bovine sequence. Also, in common with bovine α_{s1} casein, guinea-pig casein B does not contain the six-amino-acid repeat sequence identified in the rat α_{s1} casein (Hobbs & Rosen, 1982; see Fig. 3). Moreover, guinea-pig casein B, as demonstrated by analysis of the purified mature protein (Craig *et al.*, 1978), and in common with most other caseins, contains unusually high amounts of glutamic acid and proline residues, but no cysteine.

Comparison of guinea-pig and rat α_{s1} casein cDNA sequences

Comparison of the nucleotide sequence of the coding regions of guinea-pig and rat α_{s1} caseins (Hobbs & Rosen, 1982), after the introduction of gaps to maximize the alignment, revealed greater nucleotide (60%) than amino acid (43%) sequence homology (Table 1, Fig. 4). Almost a third (57/183) of the codons were identical and a further 52 differed by only a single base. Of the latter, 21 (40%) represented silent substitutions, a 'silent' mutation frequency significantly in excess of the value of 25% expected on a purely random basis (Jukes & King, 1979). High ratios of silent/replacement substitutions have been reported between homologous nucleotide sequences from different organisms (Hall *et al.*, 1982; Miyata *et al.*, 1980) and have been suggested to reflect evolutionary pressure to conserve important structural features.

Table 1. *Nucleotide and deduced amino acid sequence comparisons between guinea-pig and rat pre-(α_{s1}-casein) mRNA species*

Sequences were aligned with gaps introduced to maximize homologies as shown in Fig. 4. Only those regions common to both species were included in the above analysis. Silent and replacement substitutions were calculated only for codons differing by a single base.

	Fraction	(%)
Amino acid sequence		
Identical amino acids		
Total	79/183	43
Signal peptide	13/15	87
Mature protein	66/168	39
Nucleotide sequence		
Identical nucleotides		
Total	565/936	60
5′-Non-coding region	19/21	90
Signal peptide coding region	38/45	84
Mature protein coding region	298/513	58
Overall protein coding region	336/558	60
3′-Non-coding region	240/362	66
Identical codons	57/183	31
Two identical nucleotides/codon	52/183	28
Silent substitutions	21/52	40
Replacement substitutions	31/52	60

In this respect it can be seen from Fig. 4 that nucleotide differences were not distributed evenly throughout the coding regions of the two proteins, but instead there was a tendency for regions of homology to be clustered. In particular,

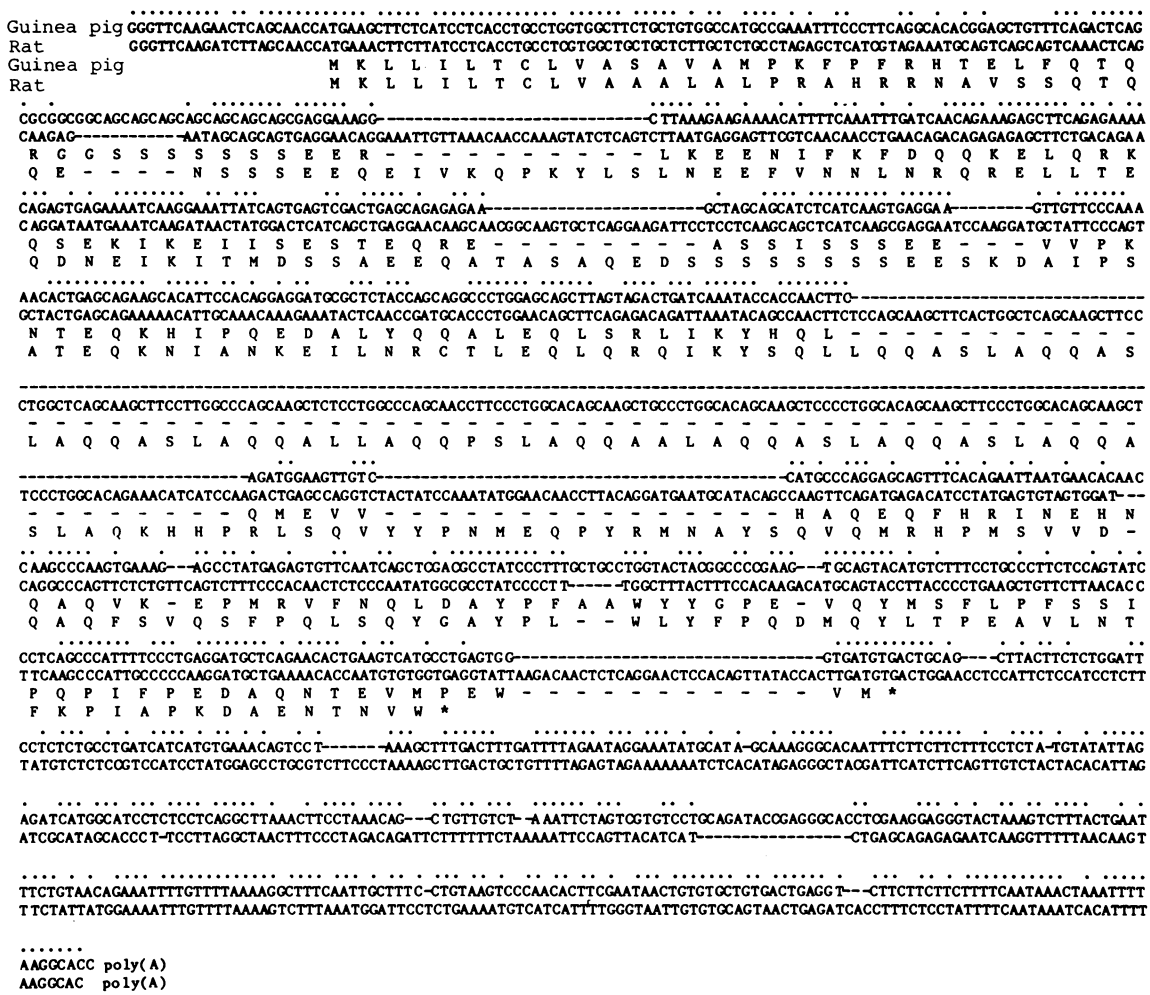

Fig. 4. *Comparison of the nucleotide sequence of guinea-pig casein B and rat α_{s1}-casein cDNA species*
Sequences were aligned with gaps introduced to maximize homology. Dots indicate nucleotides common to both sequences. Deduced protein coding regions (the larger casein B variant in the case of the guinea-pig sequence) are shown using the single-letter amino acid code. The rat sequence is from Hobbs & Rosen (1982) with corrections to the 3′ non-coding sequence (Rosen, 1983).

the two major phosphorylation sites and the signal presequence are highly conserved at the nucleotide level as well as at the amino acid level. In addition there are clearly other highly conserved regions, possibly important for protein secondary structure.

Comparison of the nucleotide sequences of the major potential phosphorylation sites in guinea-pig, rat and bovine α_{s1} casein mRNA species (Fig. 5), the latter obtained from the partial sequence reported by Willis *et al.* (1982), shows that, for a given phosphorylation site, all three species share a similar pattern of codon usage with respect to the serine residues, although the patterns differ between different phosphorylation sites. Thus TCN-TCN-AGN-GAG-GAA would represent the consensus sequence encoding the amino acid residues 68–72 of one guinea-pig phosphorylation site, whereas AGN-AGN-AGN-GAG-GAA represents the consensus sequence for the second phosphorylation site at amino acid residues 23–27. This observation would tend to argue against the multiple phosphorylation sites present in the α_{s1} casein being derived by intragenic duplication of a single ancestral sequence, a theory previously suggested on the basis of protein sequence data alone. In contrast, recent sequence analysis (Hall *et al.*, 1984) of guinea-pig casein A mRNA (an α_{s2}-type casein) has indicated a similar codon usage for two of the three major potential phosphorylation sites (TCN-TCN-AGN-GAG-GAA), in this case suggesting conservation of one class of phosphorylation site by duplication of a common sequence. Finally, the sequence encoding the two glutamic acid residues (GAGGAA) is highly con-

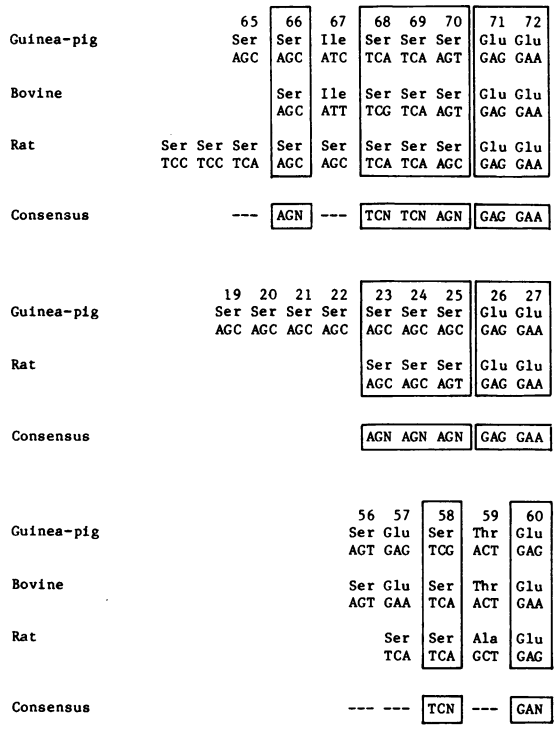

Fig. 5. *Comparison of the major potential phosphorylation sites in guinea-pig, rat and bovine α_{s1}-caseins*
Numbers refer to positions in the guinea-pig casein B sequence. Boxed regions indicate homology both in coding sequence and codon usage. Bovine and rat sequences are from Willis *et al.* (1982) and Hobbs & Rosen (1982) respectively.

served in the two major phosphorylation sites of this type (Fig. 5) present in the α_{s1} caseins, a property also common to all other classes of caseins sequenced to date (Hobbs & Rosen, 1982). Such conservation may in part be explained by the recent observation that the rat β-casein genomic sequence contains an exon/intron boundary between the glutamic acid codons (Hobbs & Rosen, 1982), thereby defining the first glutamic acid triplet (GAG).

Examination of the 5′ and 3′ non-coding regions of the guinea-pig and rat α_{s1} casein cDNA sequences (Fig. 4, Table 1) produced a somewhat unexpected result, in that, unlike results described within the globin-gene family (Miyata *et al.*, 1980; Nishioka & Leder, 1979), where considerable divergence is observed in the non-coding regions, the casein mRNA non-coding regions exhibited a higher degree of homology than did the coding region, excluding the signal peptides. In the case of the 3′ non-coding region this was 66% overall, although the distribution of differences was not uniform, with the highest degree of homology (76%) in the last third of the 3′ non-coding region, adjacent to the poly(A) tail. This observation is similar to recent studies on the α-actin genes (Ordahl & Cooper, 1983), which have shown strong homology within the A–T-rich 3′ untranslated regions adjacent to the putative poly(A)-addition signal (Proudfoot & Brownlee, 1974). It may be of significance that the guinea-pig casein A and B mRNA species and the rat α_{s1} casein mRNA contain a similar A–T-rich region. Thus, since nucleotide sequence divergence is slower in structurally important regions, but faster in areas where there are no RNA or protein sequence constraints (Miyata *et al.*, 1980; Nihioka & Leder, 1979), it must be concluded that the nucleotide sequence of at least a portion of the 3′ non-coding region has functional importance. One possibility is that this region is involved in casein mRNA stability, owing to interaction directly or indirectly with the peptide-hormone prolactin, the presence of which has been shown greatly to increase the half-life of rat casein mRNA (Guyette *et al.*, 1979). However, if this was so, one might expect other casein mRNA species to show significant homology with the α_{s1} casein in the 3′ non-coding sequence, and this does not appear to be the case. Consequently, since this homology is peculiar to the guinea-pig and rat α_{s1} casein mRNA species, it is more likely that these conserved sequences are related specifically to α_{s1} casein gene expression, possibly termination of transcription or post-transcriptional processing, though the A–T-rich region may be of more general significance.

Comparison of the 5′ non-coding regions of rat and guinea-pig α_{s1} casein cDNA sequence was limited to the 21 nucleotides obtained by primer extension in the latter case. However, within this limited region, homology was extremely high (90%), only two nucleotides differing between the two species. The possible implications of this are discussed below.

Comparison of the guinea-pig α_{s1} casein cDNA sequence with those of other casein classes

Alignment of the guinea-pig casein B (α_{s1}) cDNA sequence with other published casein sequences including guinea-pig casein A (an α_{s2} casein) (Hall *et al.*, 1984), rat α_{s1} and γ casein (Hobbs & Rosen, 1982), rat β-casein (Blackburn *et al.*, 1982), and mouse ε casein (Hennighausen *et al.*, 1982) showed that, in all cases, a reasonable alignment could be obtained in the 5′-non-coding and signal-peptide-coding regions (Fig. 6), but thereafter the different casein types lacked any significant homology (except within the phosphorylation sites, although the positions of these within the protein varied according to the class of casein). In particular, the signal-peptide sequence was highly conserved, with over half (26/45) of the nucleotides being

5' Non-coding region

```
                 ... ..... . ...  ..   .  ....  . . . .   ..    ...        . .
Rat αs1          AUCUUCUUGAUCAUCUC-CCAGCUUCUCUCACCCUACUCUUGGG-UUCAAGAUCUUAGCAACC
Guinea-pig αs1                                                CGGG-UUCAAGAACUCAGCAACC
Guinea-pig αs2             ACC---UGGCCUUCUUGUCU-----UCCACUUGGGAAAUAAGGACUCAGUAACC
Rat γ                        CAUCUA-CCUAUUCCUCUCGUCUUCCACUUGGGAAGCAAGGAUCAAGUAACC
Rat β            AUCCUCU-GAGCUU-----CAUCUUCUCUCUU-GUCCUCC---GCU-AAAGGACUUGACAGCC
Mouse ε                                                       AGGAAGCAAGGACUCAUUAAUC
```

Signal-peptide nucleotide sequence

```
                    ... ..  .  .  ..  .  .  ... ..  .. ..     ...  .  ..
Rat αs1             AUG AAA CUU CUU AUC CUC ACC UGC CUC GUG GCU GCU GCU CUU GCU
Guinea-pig αs1      AUG AAG CUU CUC AUC CUC ACC UGC CUG GUG GCU UCU GCU GUG GCC
Guinea-pig αs2      AUG AAG CUC UUC AUC UUC ACC UGC CUC UUG GCU GUG GCU CUC GCA
Rat γ               AUG AAG UUC UUC AUC UUC ACC UGC CUG GUG GCU GCU GCU CUG GCU
Rat β               AUG AAG GUC UUC AUC CUU GCC UGC CUU GUG GCA CUU GCU CUU GCA
Mouse ε             AUG AAG UUC AUC AUU CUG ACU UGC CUU UUG GCC GUU GCU CUU GCA
```

Signal-peptide amino acid sequence

	-15				-10					-5				-1
Consensus	Met	Lys	Val Leu Phe	Ile Leu Phe	Ile Met	Lys Leu Phe	Thr Ala	Cys	Leu	Leu Val	Ala	Leu Val Ser Ala	Ala	Leu Val	Ala

Fig. 6. *Comparison of the 5'-non-coding and signal-peptide-coding regions of casein mRNA species*
5'-Non-coding sequences were aligned at the initiating codon with gaps introduced to maximize homologies. Only the rat α_{s1}- and β-casein 5' non-coding sequences are complete, the rest are partial. Dots indicate nucleotides common to all six sequences. Sequences shown are rat α_{s1} and γ (Hobbs & Rosen, 1982), rat β (Blackburn *et al.*, 1982), mouse ε (Hennighausen *et al.*, 1982) and guinea-pig α_{s2} (Hall *et al.*, 1984) casein mRNA species. The casein signal-peptide amino acid consensus sequence was derived from the above nucleotide sequences in addition to pre-protein sequencing data for bovine α_{s1}- and α_{s2}-caseins (Mercier *et al.*, 1978) and porcine (Mercier & Gaye, 1980), rabbit (Mercier & Gaye, 1980) and bovine (Mercier *et al.*, 1978) β-caseins obtained by limited *N*-terminal analysis of the products synthesized *in vitro*.

common to all six caseins examined. As a result, comparison of the deduced amino acid sequences of the signal peptides obtained from nucleotide sequence analysis, together with those of bovine α_{s1} and α_{s2} caseins (Mercier *et al.*, 1978), and porcine (Mercier & Gaye, 1980), rabbit (Mercier & Gaye, 1980) and bovine (Mercier *et al.*, 1978) β-caseins (obtained by limited *N*-terminal analysis of the pre-proteins synthesized *in vitro*) indicate that 7 of the 15 residues are common to all 11 signal peptides (Fig. 6), variation within the remaining eight amino acids being restricted to only two or three different residues.

In view of the marked conservation within the signal peptides of various classes (α_{s1}, α_{s2}, β) of caseins, it would seem reasonable to assume that the casein signal sequence has some particular functional importance, possibly not common to all signal sequences. This observation is somewhat contrary to the increasing wealth of information on the structures of various signal peptides, which suggest that little amino acid sequence homology exists in the signal-peptide region among diverse groups of pre-proteins (Von Heijne, 1983; Steiner *et al.*, 1980). In general, the requirements for a signal peptide appear to be a central region rich in hydrophobic amino acids with large side chains and a region next to the cleavage site containing residues with small neutral side chains (Von Heijne, 1983). In addition, a requirement for a cysteine residue, either in the signal sequence or near the *N*-terminus of the mature protein, has recently been implicated in the membrane transport of secretory proteins (Walter & Blobel, 1980, 1981*a*), as well as in the interaction between the signal-recognition protein and the ribosomes (Walter *et al.*, 1981; Walter & Blobel, 1981*a*,*b*). Clearly, within the above constraints, a great deal of variation is possible within signal peptides, and one would therefore have expected the presequences of the α_{s1}-, α_{s2}- and β-caseins to have diverged much more.

The observed homology within the signal sequence is restricted to the so-called 'calcium-sensitive' caseins (α_{s1}, α_{s2} and β), the other major milk proteins (α-lactalbumin, whey acidic protein, and κ-casein) having quite different presequences both in length and amino acid sequence. It is therefore tempting to suggest that the casein presequence has evolved to fulfil a particular need,

possibly related to the exceptionally high levels of synthesis of secreted proteins during milk production by the lactating mammary gland.

Comparison of the sequenced 5′ non-coding regions of casein mRNA species (Fig. 6) also identified a surprisingly high degree of homology (after introduction of gaps to maximize the alignment). Of the published sequences, only the rat α_{s1}- and β-casein mRNA species represent the complete 5′ non-coding regions. In general the 5′ non-coding region is very rich in pyrimidine residues (almost 70% in the case of the rat α_{s1}-casein sequence), an observation of some interest, since both the 18S rRNA binding site (consensus sequence: AUGGUUCCGGA; see Hagenbuchle *et al.*, 1978; Lewin, 1980) and the sequence CUUCUCUC found in the 5′ non-translated regions of a number of steroid-sensitive mRNA species (Hobbs & Rosen, 1982), are both rich in U and C residues. Recently we have shown (Hall *et al.*, 1984) that guinea-pig casein A mRNA contains two potential ribosome-binding sites in its 5′ non-coding region, whereas multiple ribosome-binding sites can also be assigned to the rat α_{s1}- and β-casein sequences. It is therefore possible that conservation of the 5′ non-coding region (reflected also within the signal sequence; see above) represents a particularly efficient sequence in terms of ribosome interaction and initiation of protein synthesis, a conclusion consistent with the high translational activity of milk protein mRNA species *in vitro* (Craig *et al.*, 1976, 1979) and the high levels of casein synthesis *in vivo*. Alternatively, the conserved 5′ non-coding region, in common with the 3′ non-coding region, may reflect a role in mRNA stability, possibly through the formation of a stable secondary structure via interaction with other factors as has recently been suggested for the rat β-casein mRNA (Blackburn *et al.*, 1982).

The use of eukaryotic expression systems, in combination with mutagenesis *in vitro*, should now permit the design of experiments that will enable us to evaluate the possible role of the 3′ non-coding region in mRNA stability.

This work was supported by the Medical Research Council. We also thank Professor P. N. Campbell for his advice and encouragement.

References

Blackburn, D. E., Hobbs, A. A. & Rosen, J. M. (1982) *Nucleic Acids Res.* **10**, 2295–2307

Brignon, G., Ribadeau-Dumas, B., Mercier, J.-C. & Pelissier, J.-P. (1977) *FEBS Lett.* **76**, 274–279

Craig, R. K., Brown, P. A., Harrison, O. S., McIlreavy, D. & Campbell, P. N. (1976) *Biochem. J.* **160**, 57–74

Craig, R. K., McIlreavy, D. & Hall, R. L. (1978) *Biochem. J.* **173**, 633–641

Craig, R. K., Boulton, A. P., Harrison, O. S., Parker, D. & Campbell, P. N. (1979) *Biochem. J.* **181**, 737–756

Craig, R. K., Hall, L., Parker, D. & Campbell, P. N. (1981) *Biochem. J.* **194**, 989–998

Gordon, W. G., Groves, M. L., Greenberg, R., Jones, S. B., Kalan, E. B., Peterson, R. F. & Townsend, R. E. (1972) *J. Dairy Sci.* **55**, 261–263

Grosclaude, F., Mahe, M. F. & Ribadeau-Dumas, B. (1983) *Eur. J. Biochem.* **40**, 323–324

Guyette, W. A., Matusik, R. J. & Rosen, J. M. (1979) *Cell* **17**, 1013–1023

Hagenbuchle, O., Santer, M., Steitz, J. A. & Mans, R. J. (1978) *Cell* **13**, 551–563

Hall, L., Craig, R. K., Edbrooke, M. R. & Campbell, P. N. (1982) *Nucleic Acids Res.* **10**, 3503–3515

Hall, L., Laird, J. E., Pascall, J. C. & Craig, R. K. (1984) *Eur. J. Biochem.* **138**, 585–589

Hennighausen, L. G., Steudle, A. & Sippel, A. E. (1982) *Eur. J. Biochem.* **126**, 569–572

Hobbs, A. A. & Rosen, J. M. (1982) *Nucleic Acids Res.* **10**, 8079–8090

Jukes, T. H. & King, J. L. (1979) *Nature (London)* **281**, 605–606

Lewin, B. (1980) *Gene Expression*, vol. 2, 2nd edn., pp. 680–683, John Wiley and Sons, New York

Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560

Mercier, J.-C. (1981) *Biochimie* **63**, 1–17

Mercier, J.-C., Grosclaude, F. & Ribadeau-Dumas, B. (1971) *Eur. J. Biochem.* **23**, 41–51

Mercier, J.-C., Brignon, G. & Ribadeau-Dumas, B. (1973) *Eur. J. Biochem.* **35**, 222–235

Mercier, J.-C., Haze, G., Gaye, P. & Hue, D. (1978) *Biochem. Biophys. Res. Commun.* **82**, 1236–1245

Mercier, J.-C. & Gaye, P. (1980) *Ann. N.Y. Acad. Sci.* **343**, 232–251

Miyata, T., Yasunaga, T. & Nishida, T. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 7328–7332

Nishioka, N. & Leder, P. (1979) *Cell* **18**, 875–882

Ordahl, C. P. & Cooper, T. A. (1983) *Nature (London)* **303**, 348–349

Pascall, J. C., Boulton, A. P., Parker, D., Hall, L. & Craig, R. K. (1981) *Biochem. J.* **196**, 567–574

Proudfoot, N. J. & Brownlee, G. G. (1974) *Nature (London)* **252**, 359–362

Ribadeau-Dumas, B., Brignon, G., Grosclaude, F. & Mercier, J.-C. (1972) *Eur. J. Biochem.* **25**, 505–514

Ribadeau-Dumas, B., Grosclaude, F. & Mercier, J.-C. (1975) *Mod. Probl. Paediat.* **15**, 46–62

Rosen, J. M. (1983) *Nucleic Acids Res.* **11**, 6588

Staden, R. (1982) *Nucleic Acids Res.* **10**, 2951–2961

Steiner, D. F., Quinn, P. S., Chan, S. J., Marsh, J. & Tager, H. S. (1980) *Ann. N.Y. Acad. Sci.* **343**, 1–16

Thompson, M. P. & Farrell, H. M. (1974) in *Lactation* (Larson, B. L. & Smith, V. R., eds.), vol. 3, pp. 109–134, Academic Press, New York

Von Heijne, G. (1983) *Eur. J. Biochem.* **133**, 17–21

Walter, P. & Blobel, G. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 7112–7116

Walter, P. & Blobel, G. (1981*a*) *J. Cell Biol.* **91**, 551–556

Walter, P. & Blobel, G. (1981*b*) *J. Cell Biol.* **91**, 557–561

Walter, P., Ibrahimi, I. & Blobel, G. (1981) *J. Cell Biol.* **91**, 545–550

Willis, I. M., Stewart, A. F., Caputo, A., Thompson, A. R. & Mackinlay, A. G. (1982) *DNA* **1**, 375–386