Volume 10 Number 12 1982 Nucleic Acids Research

Comparative sequence analysis of the mRNAs coding for mouse and rat whey protein

Lothar G.Hennighausen and Albrecht E.Sippel
Institut für Genetik, University of Cologne, D-5000 Köln 41, FRG, and

Andrew A.Hobbs and Jeffrey M.Rosen
Department of Cell Biology, Baylor College of Medicine, Houston, TX 77030, USA

Received 28 April 1982; Accepted 19 May 1982

ABSTRACT

Whey acidic protein (WAP) is a major milk protein found in mouse and rat. Cloned WAP cDNAs from both species have been sequenced and the respective protein sequences have been deduced. Mouse and rat WAP (134 and 137 amino acids respectively) are acidic, cysteine rich proteins which contain a N-terminal signal peptide of 19 amino acids. Most of the cysteines are located in two clusters containing six cysteine residues each, arranged in an identical pattern. Comparison of the mouse and rat WAPs show that the signal peptide and the first cysteine domain are conserved to a greater extent than the rest of the protein. This result is reflected in the nucleotide sequence homology, where the regions coding for the signal peptide and cysteine domain I are the only regions where the rate of replacement substitution is lower than the rate of silent substitution. The 3' non-coding regions show a 91% conservation which is half the substitution rate for the coding region. This low rate of sequence divergence in the 3' non-translated region of the mRNA may indicate a functional importance for this region.

INTRODUCTION

During lactation the mammary gland secretes large amounts of milk protein which in murine milk principally consists of the acid precipitable caseins, and several whey proteins (1-4). The latter include α-lactalbumin and a novel cysteine-rich protein designated whey acidic protein (WAP, 2, 4-6). Murine WAP has a molecular weight of about 14,ooo d, an acidic isoelectric point, and may contain lipid (6). In addition, while mouse WAP is not phosphorylated (6), rat WAP contains up to three phosphate groups per molecule (2). Mouse WAP was recently identified as a member of the family of the 'four-disulfide core' proteins (7). A homologous whey acidic protein has not yet been described from the much better characterized ruminant milks.

Cloned cDNAs specific for the respective rat (8) and mouse (4) whey acidic proteins have been isolated and characterized. The rat cDNA clone was originally thought to encode the sequence for α-lactalbumin (pαLA32) on the basis of mRNA size (9,1o), hybrid arrested translation, and a comparison of

0305-1048/82/1012-3733$2.00/0

the restriction map with potential sites predicted from the α-lactalbumin amino acid sequence (8). Subsequently, the nucleotide sequence described in this paper revealed it to have no homology with α-lactalbumin, and therefore, this clone was designated pX32. In both species cloned WAP cDNAs hybridized to a single mRNA band approximately 620 nucleotides in length (4,8). In vitro translation studies have demonstrated that mouse WAP is synthesized as a 15,200 d preprotein which is processed by dog pancreatic microsomal membranes to an apparent molecular weight of 13,700 d (4). In mouse milk WAP is several fold more abundant than α-lactalbumin (5). In mid-lactating rat mammary gland WAP mRNA accounts for about 15% of the total poly $(A)^+$ mRNA fraction (8,11). The rate of accumulation of WAP mRNA during early pregnancy in the rat is significantly slower than that of the casein mRNAs, but then parallels casein mRNA accumulation during lactation (11). Hormonal induction studies using rat mammary explant cultures have shown not only that both prolactin and hydrocortisone are essential for maximal induction of WAP mRNA, but that the induction of this mRNA is regulated primarily by the steroid hormone (11). In addition, in the mouse, the WAP gene has been localized to a different chromosome from the casein gene family, which is on mouse chromosome 5 (12).

In this paper we report the nucleotide and derived protein sequences for rat WAP, and their comparison with the homologous mouse sequences. The most significant feature of these proteins is the presence of a high content of cysteine residues, which are arranged into two groups or domains each containing six cysteine residues. The positions of these cysteine residues are conserved both between the rat and mouse whey acidic protein and between the two domains within each protein. A comparative analysis of the cDNA sequences indicates that the amino acid sequences of one cysteine domain are conserved to an extent similar to that of the signal peptides. The remainder of these proteins, including the second cysteine domain, is diverging at a rate which is several-fold higher. The 3' non-coding regions of the molecules are more highly conserverd than the coding regions. This suggests that the 3' non-coding regions of the whey acidic protein mRNAs may have a functional significance.

MATERIALS AND METHODS

The isolation and characterization of plasmids containing cDNA inserts for rat and mouse whey acidic protein have been described previously (4,8). For sequence analysis of pX32, DNA fragments were labelled at either the

3' ends using the appropriate (α-^{32}P)dNTP and E. coli DNA polymerase (large fragment) (13) or at the 5' ends by the method of Maxam and Gilbert (14). Singly labelled fragments were obtained either after redigestion with a second restriction enzyme or after strand separation, and sequenced according to Maxam and Gilbert (14). The nucleotide sequence corresponding to amino acids -19 to +19 was determined by "primer extension" sequencing. The primer was a Sau 96I-Hinf I fragment 5' labelled at the Hinf I end. The fragment was used to prime cDNA synthesis as described previously (8) except that the reverse transcriptase reaction was carried out in the absence of added KCl. In the presence of 5o mM KCl, cDNA synthesis stopped abruptly at a position which correspond to the 5' end of the cDNA clone. The newly synthesized cDNA was denatured by boiling, and electrophoresed on an 8% sequencing gel. The fragment with the size of 24o nucleotides was eluted and sequenced. The sequence determination of the cDNA clones encoding mouse whey acidic protein (pWAP1 and pWAP2) has been described recently (7).

To compare the percentages or relative rates of silent and replacement substitutions, a modification of the procedure of Kafatos et al. (15) was used. In the present calculations a potential silent substitution site is defined as a position in a given sequence at which the nucleotide can be changed without changing the encoded amino acid (i.e., the 3rd positions of all codons except Trp and Met, as well as the 1st positions of some Leu and Arg codons, eg. CUG → UUG). Similarly, a potential replacement substitution site is defined as a position in a given sequence at which changing the nucleotide can change the encoded amino acid (i.e., the 1st and 2nd positions of all codons and the 3rd position of many codons, eg. AGT → AGG, changes Ser to Arg).

For these calculations each nucleotide change was determined as either silent, replacement or indeterminate. These designations were easily made in codons with single base changes. In codons with two or more base changes it was still often possible to make this distinction since, in whichever order the changes occurred, the change in the third position would always have been silent. In the indeterminate cases, the designation of the 3rd position depended upon the sequence of the changes, which gives rise to the range of values shown in the results. The percentage substitution was then calculated as follows: The number of silent substitutions was divided by the number of potential silent substitution sites. Previously all changes in the silent substitution sites were used in this calculation, disregarding whether or not they were silent (15). Since some changes in the silent substitution

sites will lead to amino acid replacement, the above ratio was multiplied by a correction factor. This factor is the sum of all possible (silent + replacement) substitutions at the potential silent substitution sites, divided by the number of possible silent substitutions. The replacement substitution rate was calculated in a similar manner. The number of replacement substitutions was divided by the number of potential replacement substitution sites. This result was then multiplied by a factor obtained by dividing the sum of all possible (silent + replacement) substitutions at the replacement sites by the sum of all possible replacement substitutions. In both calculations the results were expressed as percentages. Identical results were obtained whichever sequence was used as the parental "given" sequence.

RESULTS AND DISCUSSION

mRNA Sequence Analysis

The restriction map of pX32 and the strategy for sequence analysis are shown in Fig. 1.

The insert of pX32 covers 481 nucleotides which includes 13 Gs and 13 Cs, produced during the cloning procedure, and a 2o nucleotide poly(A) tail. Since WAP mRNA is 62o nucleotides in length, about 165 nucleotides of the 5' end of the mRNA are missing from the pX32 clone, depending upon the length of the poly(A) tail of the mRNA. Several attempts were made to isolate clones containing longer inserts, but were unsuccessful. A region of secondary structure in rat WAP mRNA, approximately 47o nucleotides from the 3' end of the mRNA, may have resulted in premature termination during cDNA synthesis and necessitated the use of partially denaturing conditions during

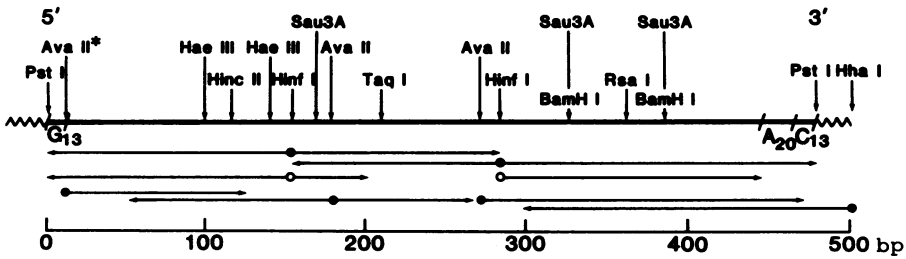

Figure 1

Restriction Endonuclease Map and Sequencing Strategy of pX32. Restriction fragments were labelled at the 3' ends (open circles) or 5' ends (closed circles).

primer extension (see Materials and Methods). Sufficient sequence information was obtained by "primer extension" to enable the complete amino acid sequence of the preprotein to be determined (Fig. 2). The restriction maps of the mouse WAP cDNA clones pWAP1 and pWAP2, the sequencing strategy, and the cDNA sequence has been described recently (7). The two mouse cDNAs cover together 6o2 nucleotides of the mRNA which includes 15 nucleotides of the 5' non-coding region as well as the complete coding and 3' non-coding regions, and a 53 nucleotide poly(A) tail.

The sequences of the rat and mouse WAP cDNAs are compared in Fig. 2. The initiation codon was assumed to be the first AUG followed by an open reading frame resulting in proteins of 134 and 137 amino acid residues in

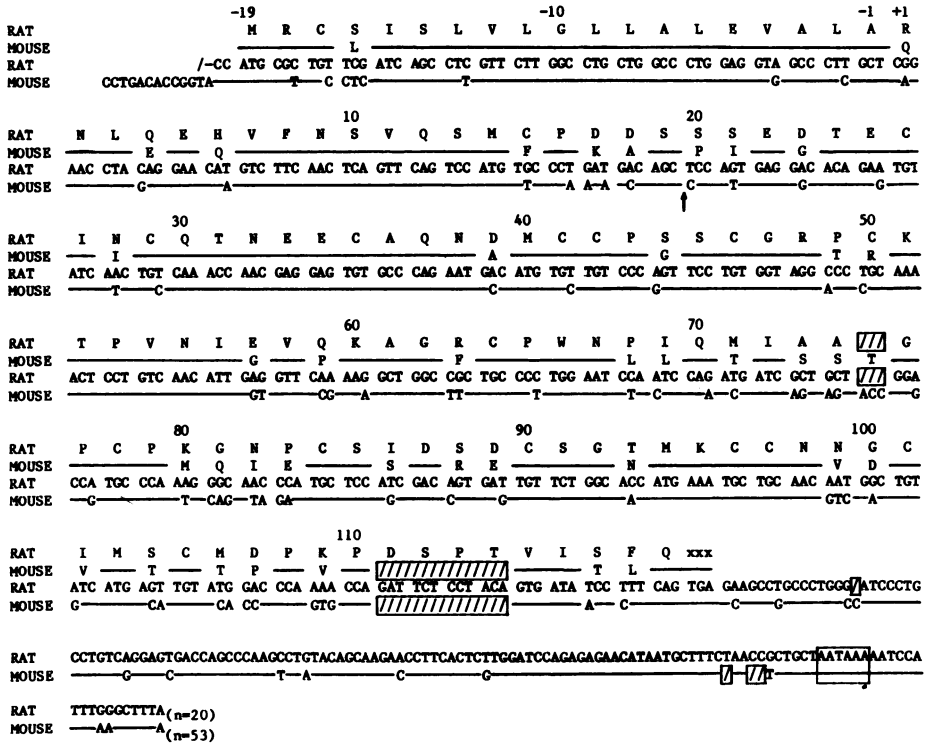

Figure 2

Comparison of the Nucleotide Sequences of the cDNAs and the Deduced Amino Acid Sequences for Rat and Mouse Whey Acidic Proteins. Only differences between the two sequences are shown for clarity. The hatched areas indicate introduced gaps, while the arrow after amino acid 19 indicates the end of the cDNA clone pX32.

the mouse and rat, respectively, which is consistent with the sizes of the homologous in vitro translation products (4). The amino terminal amino acid residues are mainly hydrophobic as it is typical of the signal peptides of other secretory proteins (16-18). By comparison with the signal peptides of other milk proteins (19,2o) we assume that the signal peptide of the whey acidic proteins consists of residues -19 to -1. This conclusion has recently been confirmed by direct sequence analysis of the amino terminal portion of the mature rat WAP (K.E. Ebner, personal communication). The amino acid compositions for the deduced mature protein sequences also agree closely with the amino acid compositions of the respective mouse and rat WAPs determined previously (1,2,6,7).

In order to maximize the homology between the two nucleotide sequences, it was necessary to introduce several gaps. Two of these, NTs 295-297 in the rat, and NTs 4o3-414 in the mouse, occurred within the coding region, but did not change the reading frame. The introduction of three one- and two-nucleotide gaps within the 3' non-coding region was necessary to maximize the homology within this region.

Two domain structure

The most outstanding feature of the amino acid sequences is the conservation of the positions of the relatively large number of cysteine residues within the mature protein (shown diagrammatically in Fig. 3). Many of the cysteine residues appear to be arranged into two domains (aa 29-5o and 84-1o5) with an identical cysteine pattern. This pattern resembles that of several other small cysteine rich proteins, the neurophysins, snake venom neurotoxins and wheat germ agglutinin (21,22). The relationship between these proteins

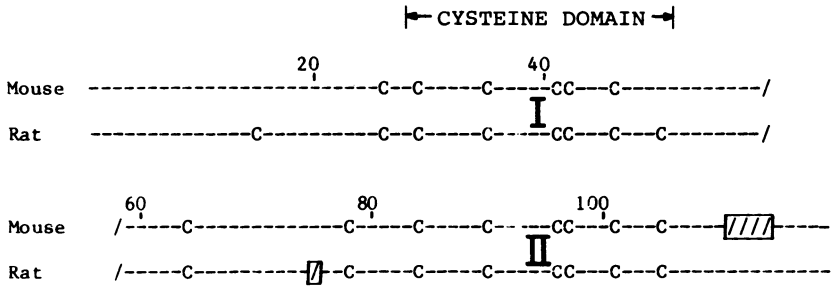

Figure 3

Alignment of the Cysteine Domains of Rat and Mouse Whey Acidic Protein. Only the cysteine residues are shown. Each dash represents one amino acid.

and mouse WAP has been discussed recently (7). The identical distributions of cysteine residues within the two cysteine domains suggest that the WAP gene in mouse and rat could have evolved by an intragenic duplication. However, a comparison of the two cysteine domains at both nucleotide and amino acid levels revealed no apparent homology apart from the cysteine residues. In the mouse, one of the cysteine residues of the first domain has been replaced by an arginine. The existance of allelic mouse WAPs with an additional cysteine and one less arginine has been reported (6).

It is interesting to note that while the mouse whey acidic protein does not contain phosphate (6) the rat WAP is a phosphoprotein (2, K.E. Ebner, personal communication). The phosphorylation may now be explained by the sequence analysis of the two proteins. In the bovine caseins it has been shown that, in general, the phosphorylated serine residues are followed two residues downstream by either a glutamic acid or a phosphoserine (23). In the rat whey acidic protein, the sequence Asp-Ser-Ser-Ser-Glu occurs (18-22) in which the central serine residue could be phosphorylated. In the mouse protein at the same position, the sequence is Ala-Ser-Pro-Ile-Gly-, in which the serine residue cannot be phosphorylated.

Evolutionary Comparisons

In Fig. 2, the preproteins have been aligned to achieve maximum homology. From the 133 overlapping amino acid residues, 7o% are identical between the mouse and rat preprotein. However, the sequence homology is not equally distributed. The signal peptides show 95% conservation of the amino acid sequence while cysteine domain I (Fig. 3) shows 81% conservation. All other regions including cysteine domain II have retained only 6o-65% homology. To compare this level of conversation with that of other proteins, the absolute rate of divergence between the mature rat and mouse WAP was estimated using the method of Dayhoff (24). The difference of 39 amino acids out of a total of 114 overlapping residues (34%) corresponds to an evolutionary distance of 45 PAMs (accepted point mutations per 1oo residues). Assuming that rat and mouse diverged 3o million years ago (24), the murine WAPs are diverging at the rate of 64 PAMs per 1oo million years. This value can be compared to o.1-o.9 for histones, and 27 and 37 for the rapidly diverging proteins α-lactalbumin and κ-casein respectively (24). Thus, the WAPs are among the more rapidly diverging proteins known. The amino acid homology is reflected by an extensive homology of the coding region (82%) at the nucleic acid level. The extent and distribution of this homology can be seen readily in Fig. 4 where deviations above the baseline represent

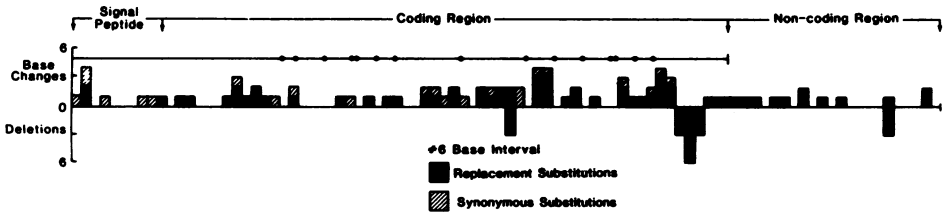

Figure 4

Comparison of the Rat and Mouse Whey Acidic Protein mRNA. Each sequence was compared over 6 nucleotide intervals. The number of differences per interval are shown above the baseline. Solid squares represent replacement substitutions while hatched squares represent silent substitutions. The half solid, half hatched squares represent the indeterminate substitutions. Deletions are indicated below the line. The positions of the cysteine residues are indicated by the dots above the baseline.

nucleotide changes while those below the line represent deletions. The substitutions are not distributed evenly over the coding region, but are clustered toward the 3'end of the coding region.

By examination of each substitution it was possible to deduce that 18 were definitely silent while 5o caused an amino acid replacement (Table I). In five cases this designation could not be made since it depends upon the order in which the multiple substitutions ocurred. When these results were analyzed it was found that 75% (6/8) of the substitutions in the signal

Table I

REGION	TOTAL NUCLEOTIDE CHANGES		SYNONYMOUS SUBSTITUTIONS				REPLACEMENT SUBSTITUTIONS			
	NO.	%	NO.	%	FACTOR	CORRECTED %	NO.	%	FACTOR	CORRECTED %
ENTIRE CODING REGION	73/399	18.3	18-23/132	14-17	1.59	22-28	50-55/333	15-17	1.19	18-20
SIGNAL PEPTIDE	8/57	14.0	6/21	28	1.26	36	2/40	5	1.09	5
MATURE PROTEIN	65/342	19.0	12-17/111	11-15	1.68	18-26	48-53/293	16-18	1.20	20-22
CYSTEINE DOMAIN I	6/66	9.1	2/22	9	1.94	17	4/57	7	1.04	7
CYSTEINE DOMAIN II	11/66	16.7	1-2/20	5-10	1.87	9-19	9-10/59	15-17	1.14	17-19

Nucleotide Substitution Rate for the Coding Region of the Whey Acidic Protein mRNAs. The differences between the sequences were classified as synonymous, replacement or indeterminate. The values for the replacement and synonymous substitutions were calculated as in Methods, both including and excluding the number of indeterminate changes. Thus the two values indicated should represent the minimum and maximum for each value. No correction was made for the possibility of two mutations at the same position.

peptide were silent while only 18-26% of nucleotide differences in the region coding for the mature protein were silent.

In order to compare the percentages or relative rates of silent and replacement substitutions, a variation of the procedure of Kafatos et al. (15) was used. In their comparison of β-globin sequences, only substitutions in potentially silent sites were considered, and not whether the substitutions were actually silent. Also no correction was performed for those sites which are both potential replacement sites as well as potential silent sites when calculating the percent replacement substitutions. To eliminate these difficulties, a modified method was used (see Methods). No attempt was made to correct for the probability of two mutations having occurred at the same site. In the present analysis most of the results showed less than about a 25% substitution rate. Thus it was decided that this comparison would probably not be affected by corrections for double mutations. However, comparison of sequences showing much greater or variable extents of divergence would require a further correction to take this factor into account. These calculations are similar to those used more recently for the comparison of the preproinsulin and globin sequences (25).

After calculation of the true rates of synonymous and replacement substitutions, the relative rates of nucleotide substitution for different regions of these mRNAs were compared (Table I). Not surprisingly, in the signal peptide region the relative rate of silent substitutions was almost seven-fold higher than that of the replacement substitutions. However, the coding region for the mature protein shows an unusual distribution of changes, with the percent of replacement substitutions being approximately the same as that of silent substitutions. This pattern is observed in the region coding for cysteine domain II, but not in the region coding for cysteine domain I, which is not only conserved more highly at the amino acid level, but shows a replacement substitution rate only slightly higher than for the signal peptide. Thus, only this domain appears to be under evolutionary pressure to retain its amino acid sequence. The silent substitution rate for the region coding for the mature protein is only half that of the rate for the signal peptide, but this is not a significant difference using a χ^2 significance test.

All eukaryotic mRNAs sequenced so far contain a 3' non-coding region. Only one feature common to all of these sequences has been observed, which is the hexanucleotide AAUAAA (or variations) (26,27). This sequence, which is involved in polyadenylation (28), is also found in the whey acidic protein

mRNAs, 17 nucleotides 5' from the poly(A) tail. However, comparison of the WAP mRNAs also shows a considerable conservation of the entire 3' non-coding region. In contrast, comparative studies of mRNAs from the globin gene family revealed a rapid evolutionary change in the 3' untranslated region (29,3o). This observation suggested that most of the 3' non-coding region was non functional other than to accept the poly(A), and was consistent with in vitro translation experiments in which blocking or eliminating the 3' non-coding region had no apparent effect on the in vitro synthesis of β-globin (31). A more extensive examination of several mammalian mRNAs indicated that some 3' non-coding regions were conserved to a significant extent (29). With respect to the extent of sequence divergence, the 3' non-coding regions could be divided into two distinct blocks: the 5' portions which evolved at a rate equivalent to the synonymous codon positions, and the 3' portions which diverged at a slower rate, equivalent to that of the 5' non-coding region and the replacement substitution sites. This conclusion is also applicable to some extent to WAP mRNAs. The block of 57 overlapping nucleotides at the very 3' end of the WAP mRNAs shows only a 5% divergence (3/57) which is equal to the frequency of replacement substitutions in the signal peptide, the most conserved portion of the coding region. The 5' block of 71 overlapping nucleotides shows a 12.5% divergence which is greater than for the 3' portion, but is still less than for the extent of synonymous substitutions. Overall, this region shows a 91% conservation which is significantly greater than for the entire coding region (82%) ($\chi^2 = 4.76$, $p < o,o5$). This result supports the previous conclusions that the 3' non-coding regions of some mRNAs must be under significant selection pressure. The apparently unusually low rate of divergence of this region in WAP mRNAs may indicate a yet unknown functional significance either on the level of RNA or gene DNA.

As mentioned previously, the structure of two identical cysteine domains of WAP suggests that the gene evolved by intragenic duplication of a region coding for a single domain. Support for this hypothesis could come from the analysis of the exon structure of the WAP gene. If such a duplication has occured, then the lack of any homology at either the amino acid or nucleotide levels between the two domains would suggest that such an event occured well before divergence of rat and mouse. An examination of other mammalian species for the presence of homologous proteins and genes and their subsequent characterization are necessary prerequisites for determining the evolutionary history of the WAP gene. The hybridization of the mouse WAP cDNA to a specific hamster DNA fragment observed in previous chromosomal

localization studies (12) suggests that the WAP gene may be sufficiently conserved for identification of homologous mRNAs and genes in other mammalian species.

ACKNOWLEDGEMENTS

Part of this work was supported by a grant of the Deutsche Forschungsgemeinschaft (SFB 74/E 5) to A.E.S.

REFERENCES

1. Mc Kenzie, R.M. and Larson, B.L. (1978) J. Dairy Sci. 61, 714-722.
2. Mc Kenzie, R.M. and Larson, B.L. (1978) J. Dairy Sci. 61, 723-728.
3. Green, M.R. and Pastewka, J.V. (1976) J. Dairy Sci. 59, 207-215.
4. Hennighausen, L.G. and Sippel, A.E. (1982) Eur. J. Biochem. 125, 131-141.
5. Zamierowski, M.M. and Ebner, K.E. (1980) J. Immun. Methods 36, 211-220.
6. Piletz, J.E., Heinlen, M. and Ganshow, R.E. (1981) J. Biol. Chem. 256, 11509-11516.
7. Hennighausen, L.G. and Sippel, A.E. (1982) Nucleic Acids Res. 10, 2677-2684.
8. Richard, D.A., Rodgers, J.R., Suppowitt, S.C. and Rosen, J.M. (1981) J. Biol. Chem. 256, 526-532.
9. Qasba, P.K. Nakhasi, H.L. (1978) Proc. Natl. Acad. Sci. USA 75, 4739-4743.
10. Chakrabartty, P.K. and Qasba, P.K. (1977) Nuclei Acids. Res. 4, 2065-2074.
11. Hobbs, A.A., Richards, D.A., Kessler, D.T. and Rosen, J.M. (1982) J.Biol. Chem. 257, 3598-3605.
12. Gupta, P., Rosen, J.M., D'Eustachio, P. and Ruddle, F.H. (1982) J. Cell Biol. 93, 199-204.
13. Hill, D.F. and Peterson, G.B. (1980) J. Virology 34, 40-50.
14. Maxam, A.M. and Gilbert, W. (1980) Methods Enzymol. 65, 499-560.
15. Kafatos, F.C., Efstratiadis, A., Forget, B.G. and Weisman, S.M. (1977) Proc. Natl. Acad. Sci. USA 74, 5618-5622.
16. Blobel, G. and Dobberstein, B. (1975) J. Cell Biol. 67, 835-851.
17. Blobel, G. and Dobberstein, B. (1975) J. Cell Biol. 67, 852-862.
18. Davis, B.D. and Tai, P.C. (1980) Nature 283, 433-439.
19. Gaye, P., Gautron, J-P., Mercier, J-C. and Haze, G. (1977) Biochem. Biophys. Res. Comm. 79, 903-911.
20. Hennighausen, L.G., Steudle, A. and Sippel, A.E., submitted.
21. Drenth, J., Low, B.W., Richardson, J.S. and Wright, S.C. (1980) J. Biol. Chem. 255, 2652-2655.
22. Drenth, J. (1981) J. Biol. Chem. 256, 2601-2602.
23. Brignon, G., Ribadeau-Dumas, B., Mercier, J-C. and Pellisier, J-P. (1977) FEBS lett. 76, 274-279.
24. Dayhoff, M.O. (1978) in Atlas of Protein Sequence and Structure, Vol. 5 Suppl. 3 (M.O. Dayhoff, ed.) National Biomedical Research Foundation Md.
25. Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. and Dodgson, J. (1980) Cell 20, 555-566.
26. Proudfoot, N.J. and Brownlee, G.G. (1974) Nature 252, 359-362.
27. Proudfoot, N.J. and Brownlee, G.G. (1976) Nature 263, 211-214.
28. Fitzgerald, M. and Shenk, T. (1981) Cell 24, 251-260.
29. Miyata, T., Yasunaga, T. and Nishida, T. (1980) Proc. Natl. Acad. Sci.

USA 77, 7328-7332.
30. Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connel, C., Spritz, R.A., DeRiel, J.K., Forget, B.W., Baralle, F.E., Shoulders, C.C. and Proudfoot, N.J. (1980) Cell 21, 653-668.
31. Kronenberg, H.M., Roberts, B.E. and Efstratiadis, A. (1979) Nucleic Acids Res. 6, 153-166.