Volume 11 Number 9 1983 Nucleic Acids Research

Gene structure of human apolipoprotein Al

Carol C.Shoulders, Alberto R.Kornblihtt, B.Sean Munro and Francisco E.Baralle

Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, UK

Received 9 March 1983; Accepted 6 April 1983

ABSTRACT

Apolipoprotein Al is the major polypeptide of the human plasma high density lipoprotein (HDL). The structure and function of the apo Al gene are of interest because of the inverse correlation shown between HDL levels and coronary heart disease. We have determined the nucleotide sequence of the apo Al gene previously isolated. Its coding sequence is interrupted by two introns of 185 and 588 base pairs. As there may be one or more introns in the 5' non coding region, the 5' end of the gene could not be precisely located. The human apo Al has an unusual propeptide segment very similar to its rat counterpart. The data reported here provide an essential basis for future studies of structural and functional alleles of the apo Al gene.

INTRODUCTION

The lipoprotein transport system is central to the mechanism by which genes, diet and hormones interact to regulate the cholesterol and triglycerides plasma levels and their tissue distribution. There are four main classes of plasma lipoproteins: chylomicrons, very low density, low density and high density lipoproteins (CM, VLDL, LDL and HDL respectively). For transport in plasma, triglycerides and cholesteryl esters are packaged into these lipoprotein particles in which they form a hydrophobic core surrounded by a surface monolayer of polar phospholipids. The surface coat also contains unesterified cholesterol in relatively small amounts together with proteins called apoproteins. At least nine apoproteins have been identified: apo AI, AII, AIV, B, CI, CII, CIII, D and E. They are not only responsible for solubilising the lipids, but also some of them can interact with many of the enzymes involved in lipoprotein metabolism and with the specific cell receptors (for review, see refs. 1 and 2).

The apoproteins of HDL may be of particular importance as the plasma levels of HDL appear to be inversely correlated to the incidence of arterial disease and thus in many cases to coronary heart disease (3-9). We have recently isolated cDNA and genomic clones of human apolipoprotein Al (10).

© IRL Press Limited, Oxford, England.

Apo A1 is the major apoprotein of human HDL. It serves as activator for the plasma enzyme LCAT (lecithin cholesterol acyl transferase), which catalyses the conversion of cholesterol to cholesteryl esters (11-13). Apo A1 promotes cholesterol efflux from tissue culture cells (14). Apo A1 may also play a role in the HDL function as an acceptor for the surface components of VLDL and CM during lipolysis in peripheral tissues. Defects in this acceptor function of HDL could impair the catabolism of triglyceride rich lipoproteins increasing the triglyceride plasma levels. Hypertriglyceridaemia occurs in more than 10% of the UK population (2). Only in rare cases has a mutant protein been associated with it and although there is a familial tendency, the disease is apparently not transmitted in a simple mendelian basis (15).

We have recently described a DNA polymorphism adjacent to the apo A1 gene present in 5% of a random sample of the population. When subjects with type IV and V hypertriglyceridaemia were studied, the DNA polymorphism frequency was raised to 44% (16). Whilst it is not clear what relationship, if any, there is between the DNA polymorphism and the apo A1 gene, it is however suggestive that a well defined group of patients has a high frequency of an uncommon genetic marker. Another type of apoprotein A1 gene alteration (as defined by genomic blots) has been reported recently in a rare inherited condition characterised by deficient levels of apoproteins A1 and CIII (22).

We wish to report here the nucleotide sequence of the human apolipoprotein A1 gene. This is a first step towards a rational interpretation of the role of apo A1 gene function in hyperlipidaemias and a necessary groundwork for future studies on mutant apo A1 genes.

MATERIALS AND METHODS

(a) Construction of the recombinant clones used: pBA1, M13A1 and λA1 are cDNA and genomic clones containing apo A1 sequences that were previously described (10). pPA1 1.8 and pPA1 2.0 (see Fig. 1) were obtained by cloning PvuII digested λA1 DNA by blunt end ligation into the PvuII site of a new vector (pAT153/PvuII/8, a pAT153 derivative constructed by Huddleston, Gould and Brownlee, unpublished). Similarly, pRH A1 5.7 was obtained by cloning EcoR1/HindIII digested λA1 DNA into EcoR1/HindIII digested pAT153/PvuII/8. pP A1 2.0 and pRH A1 5.7 were isolated by screening the resultant recombinants with the 48 bp long insert of M13 A1 prepared as previously described (10). The probe used to screen for pPA1 1.8 was the end labelled

Figure 1. Restriction enzyme map of the human apolipoprotein A1 gene.

A. This shows the pBR322 cDNA and pAT153/PvuII/8 genomic subclones used. (1) is the cDNA clone pBA1 (1O), (2), (3) and (4) are respectively the genomic subclones pRHA1 5.7, pPA1 1.8 and pPA1 2.O. The dotted underlining shows the portion of each clone eluted from agarose gels for eventual DNA sequence analysis (see Materials and Methods).

B. This restriction enzyme map represents the 12Kb fragment from λA1. The central SstI indicates two restriction sites 35 nucleotides apart.

C. Shows an expanded map around the apo A1 gene. Only the restriction sites used for sequencing are shown.

D. The extent and direction of each sequence reading are indicated by arrows. Whilst the complete sequence of the cDNA clone pB A1 was also determined, only the strategy used in determining the genomic sequence is shown. No differences in cDNA clone sequence and its genomic counterpart were found.

E. The solid area of the map denotes exon sequences. The clear areas represent 5' and 3' flanking regions and introns. The clear box preceding the AUG initiator codon may contain one or more exons encoding part of or the whole 5' non coding region of the mRNA (see Results and Discussion).

HindIII/PvuII fragment from pRHA1 5.7 (see Fig. 1).

(b) Sequencing methods: The fragments used were obtained either by end labelling whole plasmid digests and fractionating the fragments in polyacrylamide gels or by first purifying selected large molecular weight fragments by agarose gel fractionation and electroelution (17). The sequence was deter-

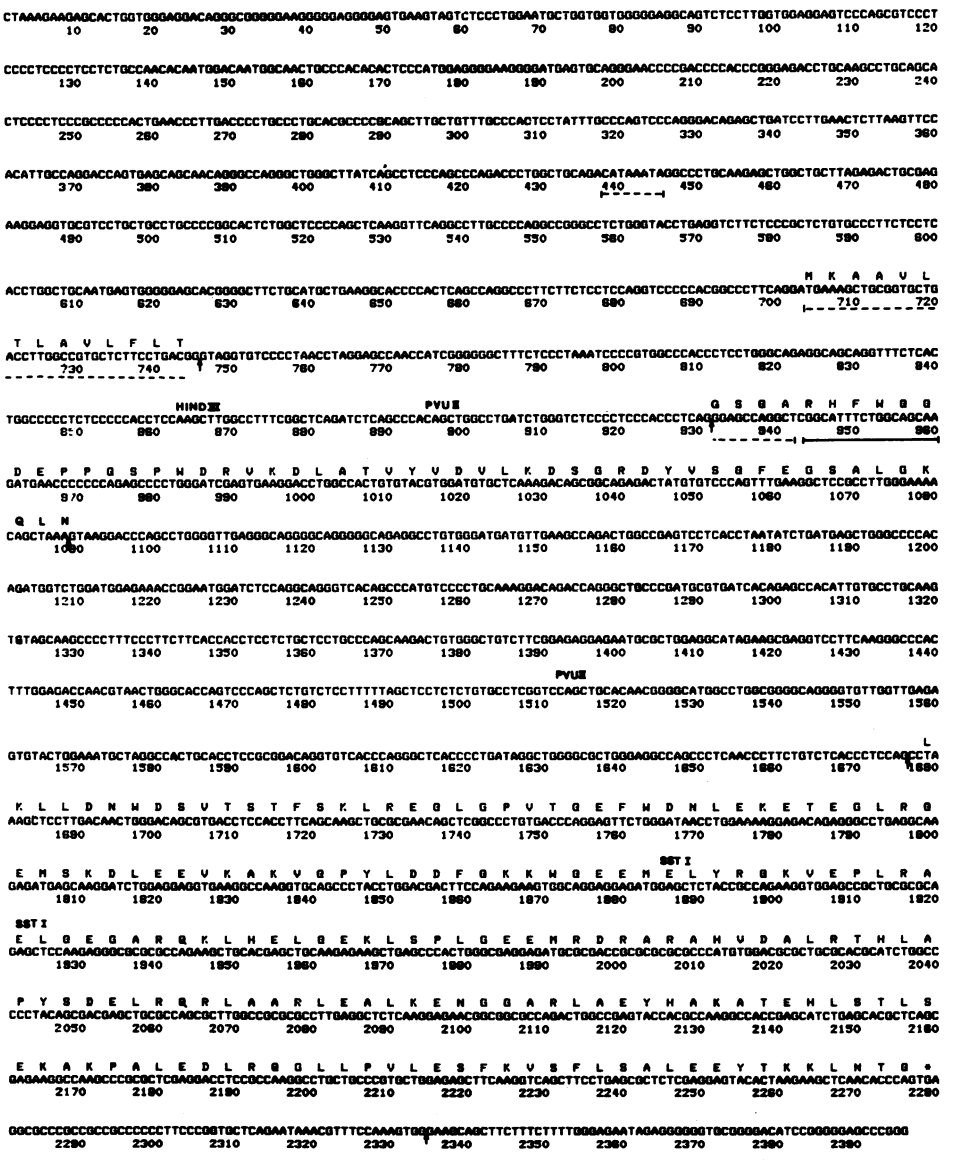

Figure 2. The nucleotide sequence of the apolipoprotein A1 gene. The dotted lines at position 438-445 indicate the only TATA box-like sequence (32) found in the 5' flanking region of the structural gene. The smaller dotted line spanning positions 703-745 and 932-942 indicates the prepeptide signal sequence. The solid line between positions 943 and 960 indicates the propeptide sequence. The mature protein coding region starts at position 961. The arrows after positions 745 and 1088 denote the exon-intron splice junction, the arrows before positions 932 and 1677 denote the intron-exon

splice junction. The arrow at position 2334 indicates the polyA addition site as determined from the cDNA sequence. It should be noted that the glycine residue (G) at position 932 and the asparagine residue (N) at position 1087 are coded by the triplets GGG and AAC generated by the splicing event. Some of the more relevant restriction enzyme sites are indicated. The sequences between positions 450-475, 1505-1515 and 1565-1575 were read in one strand only, the bases between 764 and 774 were difficult to read because of artefactual compressions.

mined by the chemical degradation procedure (18) following the strategy shown in Figure 1. In some cases, the sequencing gels were 80cm long, rather than the standard 40cm ones, to increase the number of nucleotides read from a given fragment.

(c) RNA blotting: Transfer of RNA from electrophoretic gels containing formaldehyde (19) to Gene Screen membranes and subsequent hybridisation were under conditions specified by the manufacturer (New England Nuclear, Boston, USA). The mRNA of apo A1 showed as a single species about 1000 nucleotides long.

RESULTS AND DISCUSSION

The complete nucleotide sequence of the human apolipoprotein A1 gene is shown in Figure 2. The genomic sequence was determined in the λA1 subclones pPA1 1.8, pPA1 2.0 and pRHA1 5.7. The sequence between positions 1681 and 2334 was also determined in the cDNA clone pBA1 (10) and it was in complete agreement with its genomic counterpart. However, there are 3 discrepancies at positions 1984, 2124 and 2282 with the partial cDNA sequence previously published (20). The difference at position 1984 (G for C) results in an amino acid change (see below). The sequence differences were detected in both the cDNA and the genomic clones.

We shall now analyse the features of the different sections of the gene.

The 5' flanking region

As none of the existing cDNA clones extends further than amino acid 44, there is no detailed information about the 5' end of the mRNA and hence it is difficult to locate the capping site in the gene sequence presented in Figure 2. However, our previous oligonucleotide primer extension experiments gave some indications. When the oligonucleotide designed to be complementary to positions 1870-1884 was used to prime reverse transcriptase synthesis of cDNA using liver RNA as template, a series of products was obtained (10). Of the 6 bands analysed, only one 445 $\pm$ 25 nucleotides long proved to be apo A1 cDNA. If we assume that this band was the full length

(a)	664	GGCCCTTCTTCTCCTCCAGG ↓	683
(b)	683	GTCCCCCACGGCCCTTCAGG	702
(c)	913	TCTCCCCTCCCACCCTCAGG	932
(d)	1658	CTTCTGTCTCACCCTCCAGC	1677

Figure 3. The potential splice junction (a,b) in the immediate 5' flanking region of the apo Al gene and the two known to be used during the processing of the mRNA (c,d). The arrow indicates the splicing point. The sequences are all in agreement with the accepted intron-exon junction consensus sequence: $\left(\begin{matrix}C\\T\end{matrix}\right)_{15}N\left(\begin{matrix}C\\T\end{matrix}\right)AG$ (21).

cDNA, we can deduce that the 5' non coding region of apo Al mRNA should be a maximum of 70 nucleotides long, as there are 408 bases from the initiator ATG to the 5' end of the primer. This is also consistent with the mRNA size (about 1000 nucleotides) determined by RNA blotting (see Materials and Methods). However, no TATA box (32) or similar structure is found in the expected -30 position from the putative capping sites (see Fig. 2 positions 550-650).

The only TATA box-like structure is found at position 438-445 that is 264 nucleotides from the initiator ATG. To establish if any sequence in this area forms part of the apo Al liver mRNA, total liver RNA blottings were carried out. The probes used were the following Sau961 fragments (see Fig. 2): A, from position 447-553; B, 555-692; C, 693-804; and D, a PvuII/Kpn fragment extending about 1500 bp 5' of the Kpn site at position 562. A definitive hybridisation to apo Al mRNA was observed only with fragment C. This seems to imply that there is no contiguous sequence of significant length (i.e. over 30 nucleotides) from this 5' flanking region that are present in the messenger. However, a split 5' non coding region could still be encoded in this area. This is further supported by the presence of two stretches of sequences strikingly similar to intron-exon junctions (21) 1 and 20 nucleotides 5' of the initiator ATG. Figure 3 shows these potential splice junctions and the two internal ones found in the gene.

Furthermore, studies on the avian apoprotein II gene (34,35) have shown that a very short exon at the 5' end of the gene is separated by a 1.5Kb intron from the next exon which contains the ATG initiation codon. It is possible that human apo Al has a similar structure. There also exists the possibility of tissue specific 5' non coding regions arising from alternative splicing. This organisation has been demonstrated for mouse liver and saliv-

```
Human  M-K-A-A-V-L-T-L-A-V-L-F-L-T-G-S-Q-A|R-H-F-W-Q-Q
                 * *   * *         *     | * *   *
Rat    M-K-A-A-V-L-A-V-A-L-V-F-L-T-G-C-Q-A|X-E-F-X-Q-Q
```

Figure 4. Comparison of the pre and propeptides of human and rat apolipoprotein A1. The human sequence was predicted from the nucleotide sequence of the gene (Fig. 2), while the rat sequence was previously determined (25) by peptide sequence of the in vitro translation product. The vertical line indicates the boundary between pre and propeptide, X indicates an unknown residue, * denotes an amino acid difference, the unknown residues are counted as non homologous.

ary gland α-amylase mRNAs (23,24). We are currently cloning the cDNA of the 5' ends of liver and intestine apo A1 mRNA. Their eventual sequencing and matching with the genomic sequence will elucidate the organisation of the 5' end of the gene.

The signal sequence

The apoprotein A1 signal sequence from rat was recently determined by sequencing the in vitro translation product of rat intestinal mRNA (25). The first 18 amino acids of this sequence can be operationally defined as a prepeptide and is followed by a 6 amino acid long propeptide (25). By comparing this sequence with the one predicted by the DNA sequence (Fig. 2), we were able to locate the start of the human prepro apo A1. Figure 4 compares the rat and human prepro sequence. All of the amino acid differences between the signal sequences could be explained by a point mutation of the first base in each codon. They have the standard length (26) and hydrophobicity. Both of them end in an amino acid with a small side chain (alanine). The extent of amino acid homology is very unusual between signal peptides even for those present in proteins synthesised in the same tissue or species (26). In fact, the apoprotein signal peptide homology is shared not only by human and rat apo A1 but also by the rat intestinal apo AIV (27).

However, very little homology exists between these mammalian apoprotein signal peptides and the only other characterised preapoprotein namely chicken VLDL apoprotein II (34,35).

The DNA coding for the signal sequence bears an intron of 185 base pairs. The precise positions of the splicing sites were assigned by locating the characteristic G-T and A-G consensus sequences at the 5' and 3' ends of the exon-intron junctions (21) and by optimizing the resultant apo A1 amino acid sequence homology to the rat apo A1 amino acid sequence. Alternative exon-

intron junctions might exist at positions 750 or 752 instead of 746 (Fig. 2). The use of these splicing sites seems unlikely and, particularly the 752 junction would result in 2 extra amino acids, arginine and tryptophan, being present in the otherwise unchanged signal sequence.

Exon 1 contains all of the central hydrophobic core of the signal sequence and thus supports Gilbert's hypothesis that exons encode functional domains (33).

The Propeptide

The predicted sequence of amino acids 19-24 inclusive code for a propeptide 6 amino acids long. Figure 4 shows this sequence and the rat intestinal propeptide (25). No accurate comparison is possible because of the two uncertainties in the rat sequence. However, they both bear a pair of glutamine residues at their carboxy terminal end. This is unusual; these are the first propeptides to be reported that do not end in a pair of basic amino acids. Gordon et al (25) suggest that this must mean that the post translational processing of propeptide A1 must proceed along an as yet undefined pathway. They proposed that this propeptide might prevent interaction of pro apo A1 with a hydrophobic environment. Its removal could then act as a signal which triggers nascent HDL particle formation.

The other apolipoproteins whose primary translation product amino terminal sequence has been determined are rat apo AIV (27) and avian apo VLDL II (36). No propeptide can be discerned in either of these proteins.

The Mature Protein coding region

Two amino acid sequences are published for apo A1 (28,29). There are considerable differences between them. The amino acid sequence predicted from the apo A1 gene sequence reported here agrees closely with that published by Brewer et al (29). We found only 3 differences with their sequence. They are: amino acid 34, 146 and 147 of the mature protein. In all of them the DNA sequence predicts glutamic acid residues rather than glutamine residues. The amino acid at position 146 is the one where our DNA sequence disagrees with the cDNA sequence previously reported (20). Further detailed studies are needed to clarify if some of these differences are genuine polymorphisms.

The DNA coding for the mature protein bears an intron of 588 base pairs (see Fig. 2). The precise positions of the splicing sites were assigned as described above for the first intron.

The codon usage for amino acids is shown in Table 1.

Table 1. Codon usage in apo A1

Phe	UUU	1	Ser	UCU	0	Tyr	UAU	1	Cys	UGU	0
	UUC	7		UCC	3		UAC	6		UGC	0
Leu	UUA	0		UCA	0	Term	UAA	0	Term	UGA	1
	UUG	3		UCG	0	Term	UAG	0	Trp	UGG	5
Leu	CUU	2	Pro	CCU	1	His	CAU	4	Arg	CGU	0
	CUC	12		CCC	7		CAC	2		CGC	12
	CUA	2		CCA	1	Gln	CAA	5		CGA	1
	CUG	22		CCG	1		CAG	14		CGG	1
Ile	AUU	0	Thr	ACU	2	Asn	AAU	0	Ser	AGU	0
	AUC	0		ACC	6		AAC	5		AGC	13
	AUA	0		ACA	1	Lys	AAA	3	Arg	AGA	2
Met	AUG	4		ACG	3		AAG	19		AGG	1
Val	GUU	0	Ala	GCU	4	Asp	GAU	5	Gly	GGU	0
	GUC	1		GCC	12		GAC	11		GGC	9
	GUA	0		GCA	1	Glu	GAA	4		GGA	1
	GUG	14		GCG	6		GAG	26		GGG	1

From this data, it is clear that usage is not random. For example, the glutamic acid is coded 4 times by GAA and 26 times by GAG. Other preferred codons are UUC (Phe), CUG (Leu), GUG (Val), CCC (Pro), GCC (Ala), UAC (Tyr), CAG (Gln), AAG (Lys), GAC (Asp), CGC (Arg) and GGC (Gly).

The number of times that A, C, G or T are used in the third position of the codon are 22, 106, 120 and 20 respectively. Of the cases where it is possible that any of the 4 nucleotide bases could be used, A is used 9 times, C 75, G 52 and T 9. Indeed, the unusual C and G distribution of the coding region is highlighted between positions 1993 and 2078 (Fig. 2) where there are 18 CG doublets.

The 3' flanking region

The human apo A1 mRNA has a rather short 3' non coding region (54 nucleotides between the termination codon and the polyA addition site, see Fig. 2). The polyA addition site is 14 nucleotides 3' of the AATAAA sequence (30). About 2.8 kilobases beyond the polyadenylation site maps the SstI(SacI) polymorphism that correlates with hypertriglyceridaemia (16). A more detailed study of the 5' and 3' flanking regions may be of interest if the structural apo A1 gene linked to the polymorphic site, currently being isolated, is proved to be identical to the normal gene.

The data presented in this paper provide an essential basis for future studies of structural and functional variants of the apo A1 gene which will hopefully improve our understanding of the lipoprotein transport system and its role in the pathogenesis of atherosclerosis.

ACKNOWLEDGEMENTS

We are grateful to Mr C.R. Sharpe for the RNA blotting experiments and Mrs C.E. Phillips for parts of the early sequencing data. This work was supported by the Medical Research Council and the British Heart Foundation. ARK is a recipient of a research fellowship from the Consejo Nacional de Investigaciones Cientificas y Tecnicas of Argentina.

REFERENCES

1. Brown, M.S., Kovanen, P.T. and Goldstein, J.L. Science 212, 628-635 (1981)
2. Galton, D.J., Stocks, J., Dodson, P.M., Clin.Biochem.Rev. 3, 377-405 (1982)
3. Gotto, A.M., Miller, N.E. and Oliver, M.F., eds. High Density Lipoproteins and Atherosclerosis, Third Argenteuil Symposium, Elsevier, New York (1978)
4. Gordon, T., Castelli, W.P., Hjortland, M.C., Kannel, W.B. and Dawber, T.R. Am.J.Med. 62, 707 (1977)
5. Rossner, S., Kjellin, K.G., Mettinger, K.L., Siden, A. and Soderström, C.E., Lancet (1), 577 (1978)
6. Miller, G.J. and Miller, N.E. Lancet (1), 16 (1975)
7. Willett, W. et al, N.Engl.J.Med. 303, 1150-1161 (1980)
8. Stein, Y., Glangeaud, M.C., Fainaru, M. and Stein, O. Biochim.Biophys. Acta 380, 106 (1975)
9. Miller, N.E., Weinstein, D.B., Carew, T.E., Koschinsky, T. and Steinberg, D., J.Clin.Invest. 60, 78 (1977)
10. Shoulders, C.C. and Baralle, F.E. Nucl.Acids Res. 10, 4873-4882 (1982)
11. Fielding, C.J., Shore, V.G. and Fielding, P.D., Biochem.Biophys.Res. Commun. 46, 1943-1949 (1972)
12. Soutar, A., Garner, C., Baker, H.N., Sparrow, J.T., Jackson, R.L., Gotto, A.M. and Smith, L.C., Biochemistry 14, 3057-3064 (1975)
13. Yokoyama, S., Fukushima, D., Kupferberg, J.P., Kezdy, F.J. and Kaiser, E.T., J.Biol.Chem. 255, 7333-7339 (1980)
14. Jackson, R.L., Gotto, A.M., Stein, O. and Stein, Y., J.Biol.Chem. 250, 7204-7209 (1975)
15. The Metabolic Basis of Inherited Disease, eds. J.B. Stanbury, J.B. Wyngaarden and D.S. Frederickson, McGraw-Hill, New York (1978)
16. Rees, A., Shoulders, C.C., Stocks, J., Galton, D.J. and Baralle, F.E., Lancet (1), 444-446 (1983)
17. Girwitz, S.C., Bacchetti, S., Rainbow, A.J. and Graham, F.L., Anal. Biochem. 106, 492-496 (1980)
18. Maxam, A.M. and Gilbert, W., Proc.Natl.Acad.Sci.USA 74, 560-564 (1977)
19. Lehrach, H., Diamond, D., Wezney, J.M. and Boldtber, H., Biochemistry 16, 4743-4751 (1977)
20. Breslow, J.L., Ross, D., McPherson, J., Williams, H., Kurnit, D., Nussbaum, A.L., Karathanasis, S.K. and Zannis, V.I., Proc.Natl.Acad.Sci. USA 79, 6861-6865 (1982)
21. Mount, S.M., Nucl.Acids Res., 10, 459-472 (1982)
22. Karathanasis, S.K., Norum, R.A., Zannis, V.I. and Breslow, J.L., Nature 301, 718-720 (1983)
23. Hagenbüchle, O., Tosi, M., Schibler, U., Bovey, R., Wellauer, P.K. and Young, R.A., Nature 289, 643-646 (1981)
24. Young, R.A., Hagenbüchle, O. and Schibler, U., Cell 23, 451-458 (1981)
25. Gordon, J.I., Smith, D.P., Andy, R., Alpers, D.H., Schonfeld, G. and Strauss, A.W., J.Biol.Chem. 257, 971-978 (1982)

26. Davis, B.D. and Tai, P.C., Nature 283, 433-438 (1980)
27. Gordon, J.I., Smith, D.P., Alpers, D.H. and Strauss, A.W., J.Biol.Chem. 257, 8418-8423 (1982)
28. Baker, H.N., Gotto, A.M. and Jackson, R.L., J.Biol.Chem. 250, 2725-2738 (1975)
29. Brewer, H.B., Fairwell, T., Larne, A., Ronan, R., Hauser, A. and Bronzert, T.J., Biochem.Biophys.Res.Commun. 80, 623-630 (1978)
30. Proudfoot, N.J. and Brownlee, G.G., Nature 263, 211-214 (1976)
31. Dayhoff, M.O., Atlas of Protein Sequence and Structure, Nat.Biomed.Res. Foundn., Washington (1976)
32. Goldberg, M., Ph.D. Thesis, Stanford University, Stanford, California (1979)
33. Gilbert, W., Nature 271, 501 (1978)
34. Frits, C.P.W.M., van het Schip, A.D., Arnberg, A.C., Wieringa, B., Geert AB and Gruber, M., J.Biol.Chem. 256, 9668-9671 (1981)
35. Wiskocil, R., Goldman, P. and Deeley, R.G., J.Biol.Chem. 256, 9662-9667 (1981)
36. Chan, L., Bradley, W.A. and Means, R.A., J.Biol.Chem. 255, 10,060-10,063 (1980)