Volume 12 Number 6 1984 Nucleic Acids Research

Sequence of the pS2 mRNA induced by estrogen in the human breast cancer cell line MCF-7

Sonia B.Jakowlew, Richard Breathnach, Jean-Marc Jeltsch, Piotr Masiakowski and Pierre Chambon*

Laboratoire de Génétique Moléculaire des Eucaryotes du CNRS, Unité 184 de Biologie Moléculaire et de Génie Génétique de l'INSERM, Institut de Chimie Biologique, Faculté de Médecine, Strasbourg, France

Received 30 January 1984; Accepted 17 February 1984

ABSTRACT

We present the complete sequence of an mRNA which is induced by estrogen in the human breast cancer cell line MCF-7 [pS2 mRNA, Masiakowski et al., Nucleic Acids Res. 10, 7895-7903 (1982)]. Primer extension and cloning of double-stranded cDNA (ds-cDNA) into a vector designed to make full-length cDNA were used to determine the sequence of the fifteen 5'-terminal nucleotides which were not present in the original pS2 ds-cDNA clone. The mRNA sequence has a major open reading frame encoding 84 amino-acids, flanked by a 40 nucleotide 5'-untranslated region and a 198 nucleotide 3'-untranslated region preceding the polyA tail. The 3'-untranslated region contains a polyadenylation signal, AUUAAA, 14 nucleotides upstream from the polyA tail. The derived protein sequence contains a putative signal peptide region suggesting that the protein may be secreted. The nucleotide and derived amino-acid sequences were compared to previously determined sequences, particularly to those of hormone-regulated proteins and growth factors, and no obvious similarities were observed.

INTRODUCTION

Considerable work has been done in recent years to determine how steroid hormones modulate gene expression. The mouse mammary tumor virus (MMTV) provides an excellent model system to study the molecular mechanisms underlying induction of transcription of a specific gene by glucocorticoids (for refs see 1-4). The human pS2 gene whose double-stranded cDNA (ds-cDNA) has been cloned recently by our group (5) may constitute a counterpart of the MMTV system for studying the transcriptional regulation of gene expression in mammals by estrogen. Indeed, we have shown that the induction of expression of the pS2 gene by estrogen in the human breast cancer cell line MCF-7 is a primary transcriptional event (6). In addition, pS2 RNA has been found only in MCF-7 cells and in biopsies of some breast cancers (7,8). It cannot be detected in normal breast tissue. Thus the pS2 gene may be useful for clinical studies as well as basic studies aimed at elucidating how estrogens work at the molecular level. The MCF-7 cell line and pS2 gene provide a more convenient system for *in vitro* genetics studies on estrogen-

induced genes than the popular chick oviduct, which has been extensively used to study the molecular mode of action of steroid hormones (for discussion of this point, see ref. 7).

We report here the complete nucleotide sequence of the pS2 mRNA and show that it contains an open reading frame coding for a small protein which might be secreted and presents no obvious similarities to already known proteins.

MATERIALS AND METHODS

Cell culture

MCF7 cells were a gift from the Michigan Cancer Foundation and were maintained in Dulbecco's modified Eagle's medium supplemented with 10% fetal calf serum (Gibco) and 0.6 μg/ml insulin (Sigma).

Preparation of RNA

Total RNA was prepared by homogenization of cells in urea-SDS and precipitation with LiCl according to the procedure of Auffray and Rougeon (9) except that following overnight precipitation with LiCl, the RNA was extracted twice with phenol-chloroform-iso-amyl alcohol (50/50/1).

Cloning of full-length pS2 double-stranded cDNA.

The construction of recombinant plasmids for cloning full-length ds-cDNA (10) was as described by Breathnach and Harris (11). As outlined in Fig. 4, a T-tailed primer fragment is used to prime cDNA synthesis from polyadenylated RNA. The cDNA-mRNA hybrid is then tailed with dC residues and is annealed to a linker fragment derived from the plasmid pSVE (11) carrying a tail of dG residues. Following ligation, the RNA is then replaced by DNA using RNase H and DNA polymerase I. The advantage of this cloning technique is that S1 nuclease is not used, leading to a greater retention of sequences representing the 5'-ends of mRNAs.

This technique was used to construct a ds-cDNA library from polyadenylated RNA of estrogen-treated MCF-7 cells. The ds-cDNA library was screened by hybridization with the nick-translated PstI fragment A (Fig. 1A) of the pS2 ds-cDNA previously cloned by Masiakowski et al. (5). Two clones were detected. Their cDNA inserts were approximately the same length as that of the original pS2 ds-cDNA. These recombinant plasmids, designated pSVES1 and pSVES2 (Fig. 4), were sequenced from the unique BalI site of the ds-cDNA. In practice, the BalI sites of pSVES1 and pSVES2 were converted into SalI sites by cleavage with BalI and SalI, followed by repair with DNA polymerase I and ligation (BalI cleaved DNA very inefficiently in our

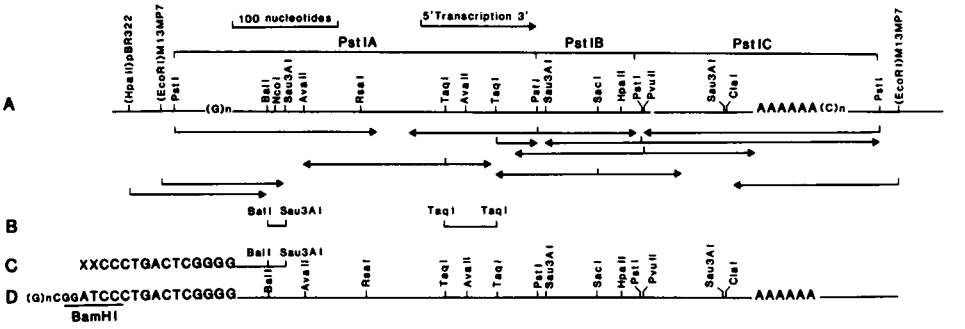

Figure 1 : (A) Restriction enzyme map of the "original" pS2 ds-cDNA insert (5) and diagram of the restriction fragments used for sequencing. The horizontal arrows indicate the orientation and length of the sequences obtained. The vertical bars on the arrows indicate the point of labeling. The C's and the G's correspond to the tails used for inserting the ds-cDNA into the pBR322 vector and the A'S correspond to the mRNA polyA tail. (HpaII) pBR322 indicates the nearest HpaII site in pBR322 upstream from the unique PstI site. (EcoRI) M13MP7 indicates the 2 EcoRI sites in M13MP7 that are on either side of the PstI site. PstIA, B and C represents the 3 pS2 ds-cDNA fragments obtained upon digestion with PstI. The map is drawn to scale up to the homopolymer tails. (B) ds-cDNA fragments prepared from the original pS2 ds-cDNA and used as primers for extension with reverse transcriptase after hybridization to MCF-7 RNA (Materials and Methods). (C) The extension product obtained using the BalI-Sau3A fragment (coding strand) as primer. Only the newly synthesized nucleotides past the 5'-end of the originally cloned pS2 ds-cDNA are shown (non coding-strand). The nucleotides represented by X could not be determined with certainty. (D) Restriction enzyme cleavage map of the full-length ds-cDNA insertion present in pSVES1 (Fig. 4 and text) and nucleotide sequence (non-coding strand) of the 5'-terminus.

hands). The resulting plasmids were designated pSVESS1 and pSVESS2. After digestion with SalI, the newly generated SalI extremities were end-labeled with T4 polynucleotide kinase and [^{32}P]-γ-ATP, the DNAs cleaved with KpnI (see Fig. 4) and the small SalI-KpnI fragments of pSVESS1 and pSVESS2 sequenced.

Preparation of DNA fragments.

Restriction enzyme fragments of plasmids were fractionated by polyacrylamide gel electrophoresis and eluted by diffusion (12). DNA fragments were also isolated by electrophoresis on low melting agarose gels (Sea-Plaque) and DNA recovered by heating and repeated phenol-chloroform extraction (13). Restriction enzymes were used as recommended by the commercial suppliers.

DNA Sequencing

Labeling of the 5'-ends of polyacrylamide gel purified DNA restric-

tion fragments was performed using T4 polynucleotide kinase and [γ-^{32}P]-ATP (Amersham, 3000 Ci/mmole) as described by Maxam and Gilbert (12). The sequences of the 5' end-labeled DNA restriction fragments were determined following a secondary cleavage with another restriction enzyme or strand separation. The 3'-ends of DNA fragments were labeled by incubating 1-5 μg of restricted DNA in 50 mM Tris-HCl, pH 8.2, 10 mM $MgCl_2$, 5 mM dithiothreitol, 50 mM KCl, 3 μM [α-^{32}P]-dATP (New England Nuclear, 400 Ci/mmole) and all other unlabeled dNTPs at 4 μM, with 20 units of AMV reverse transcriptase (Life Sciences, Inc., St Petersburg, FL) at 37°C for 40 min (14). Sequencing of the end-labeled DNA fragments was as described by Maxam and Gilbert (12) using 8% and 15% polyacrylamide gels (0.3 mm thick and 40 cm long) containing 8.3 M urea.

mRNA-templated primer extension.

The 23 nucleotide BalI-Sau3AI and the 48 nucleotide TaqI-TaqI fragments (Figs. 1B and 5) were labeled with [γ^{32}P]-ATP and T4 polynucleotide kinase on both 5'-ends. The DNA strands were separated as described (12), and the strand complementary to mRNA was hybridized to total RNA from estrogen-treated MCF-7 cells using 1 pmole of primer DNA fragment per 100 μg of total MCF-7 RNA in 166 mM Tris-HCl, pH 8.2, 166 mM KCl, and 33 mM $MgCl_2$ (in a volume of 15 μl) at 42°C for 60 min. Following hybridization, 0.5 mM each of dATP, dCTP, dGTP and TTP were added along with 5 mM dithiothreitol and 48 units of AMV reverse transcriptase in a volume of 50 μl and incubation was continued for 2.5 hrs. After addition of NaOH to 0.3 M and RNA digestion overnight at 37°C, the reaction was neutralized with acetic acid and precipitated with ethanol. The DNA products were electrophoresed on a 8% sequencing gel.The strong stop band visualized by autoradiography was excised from the gel and sequenced by the Maxam and Gilbert (12) technique.

RESULTS

Sequencing of the original pS2 double-stranded cDNA.

The previously cloned pS2 ds-cDNA, corresponding to a polyadenylated RNA whose level in MCF-7 cells is rapidly increased by addition of estradiol to the culture medium (5), was sequenced directly using the procedure of Maxam and Gilbert (12). The restriction map shown in Fig. 1A indicates the presence of restriction endonuclease sites as determined by a combination of direct multiple restriction endonuclease analysis and sequence analysis. This pS2 ds-cDNA contains two internal PstI sites yielding three fragments, A, B and C, of approximately 320, 100 and 150 nucleotides in length (dC-

tailed ds-cDNA was annealed to pBR322 dG-tailed at its PstI site; so the ds-cDNA is flanked by PstI sites). These PstI sites, together with a PvuII and a SacI site, were the most useful sites for 5' end-labeling the fragments and subsequent sequence determination. Some TaqI sites were also used as labeling sites to confirm sequence determinations and for overlapping the sequences of adjacent fragments. The sequence of pS2 ds-cDNA was determined for the most part on both DNA strands to ensure accuracy as outlined in Fig. 1A. The extreme 5' part of the sequence was determined, after subcloning PstI fragment A into the PstI site of M13mp7 (15), by sequencing from the EcoRI and BamHI sites of the replicative form of the M13 ds-cDNA recombinant. The 5'-end of the pS2 ds-cDNA was also sequenced from an upstream HpaII site of pBR322 (Fig. 1A) using a HpaII-RsaI fragment. The 3'-end region of the original pS2 ds-cDNA was also confirmed by sequencing from the 5' end-labeled EcoRI site of a M13mp7-pS2 ds-cDNA recombinant containing the PstI fragment C of the original pS2 ds-cDNA.

All of the restriction sites determined from mapping data were accounted for by sequence analysis. Some of the restriction sites predicted by the sequence were not cut by the appropriate enzyme. These include the AvaII site at position 76 and the ClaI site at position 482 (see Fig. 5). This is due to a C and an A methylation within the AvaII and ClaI sites, respectively (16). The ClaI site was cleaved (A. Brown, personal communication) when the pS2 ds-cDNA clone was transfected into a dam$^-$ strain of E.coli where adenine methylation does not occur.

The sequence data (Fig. 5) indicates that the original pS2 ds-cDNA clone contains a 559 nucleotide insert including 20 and 14 dCMP residue tails at the 5' and 3'-ends, respectively, and a 3'-end 50 nucleotide-long polydA tail. Two approaches were used to determine the 5'-terminal sequence of pS2 mRNA.

Determination of pS2 mRNA 5'- end sequence by primer extension

Previous estimates of the size of the mRNA corresponding to pS2 ds-cDNA by electrophoresis on denaturing methylmercury hydroxide-agarose or formaldehyde-agarose gels were approximately 600 nucleotides, using an AluI digest of pBR322 as size marker (5). Since this length was not very different from that of the ds-cDNA insert present in the original pS2 ds-cDNA clone, it appeared feasible to determine the number of any additional nucleotides at the 5'-end of the pS2 mRNA using primer extension. Two primers, the 23 nucleotide BalI-Sau3AI fragment and the 48 nucleotide TaqI-TaqI fragment were prepared as described in Materials and Methods (see Figs. 1B and

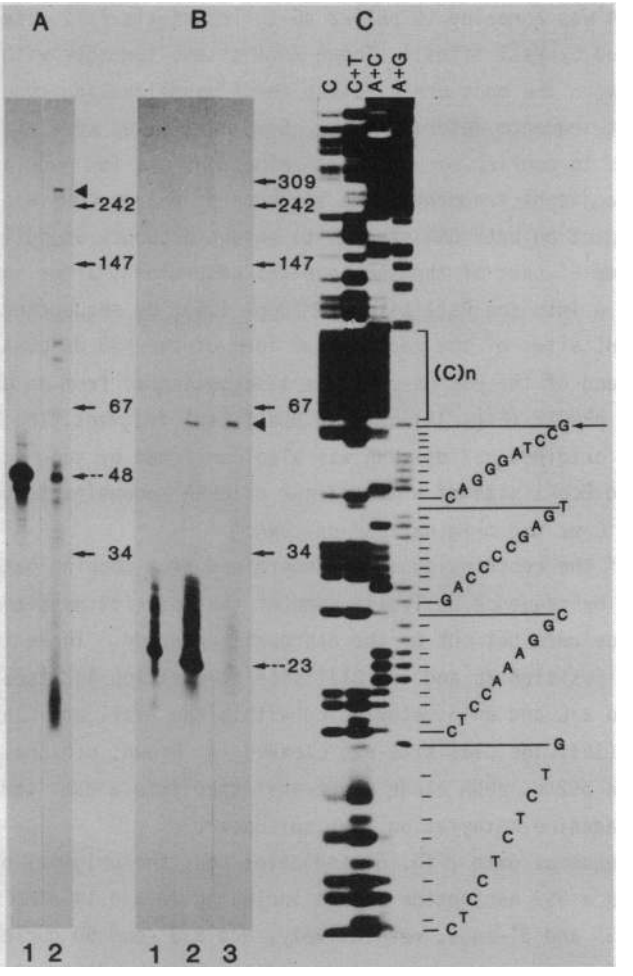

Figure 2 : Synthesis by primer extension of cDNA complementary to the 5'-end of the pS2 mRNA. (A) The cDNA was synthesized in the presence of total RNA from MCF-7 cells grown in the presence of estradiol using the coding strand of the TaqI-TaqI fragment (48 nucleotides in length, Fig. 1B) of the original pS2 ds-cDNA as primer, and was electrophoresed on a 8% polyacrylamide/8.3 M urea gel. The sizes of marker fragments from a MspI digest of pBR322 are given. Lane 1 : TaqI-TaqI fragment before primer extension; Lane 2 : TaqI-TaqI fragment after extension with reverse transcriptase followed by S1 nuclease digestion (Materials and Methods). (B) As in (A), but the BalI-Sau3AI fragment (23 nucleotides in length, Fig. 1B and 5) was used as primer. Lane 1 : BalI-Sau3AI fragment before extension; Lane 2 : BalI-Sau3AI fragment after extension followed by NaOH digestion (Materials and Methods); Lane 3 : BalI-Sau3AI fragment after extension followed by S1 nuclease digestion. (C) Maxam and Gilbert sequence analysis of the 5'-end of the ds-cDNA present in pSVES1 (see Fig. 4, Materials and Methods and text). The sequence shown on the right of the ladder corresponds to the coding strand.

5). The 5'-terminally labeled coding strand of each primer was hybridized with the mRNA and elongated by incubation with reverse transcriptase. Following digestion with either NaOH or S1 nuclease, the elongated DNA was analyzed on a denaturing urea/polyacrylamide gel. The autoradiographs are shown in Fig. 2, A and B.

When the 48 nucleotide long TaqI-TaqI fragment (dashed arrows, Fig. 2A) was used as a primer for extension with reverse transcriptase, a band of approximately 260 nucleotides was observed upon gel electrophoresis after digestion of the elongated product with S1 nuclease (lane 2, arrowhead) together with some prematurely terminated cDNA. Taking into consideration that this transcript included the 48 nucleotide primer and 197 nucleotides previously sequenced at the 5'-end of the original pS2 ds-cDNA (see Fig. 5), this result indicated that the original clone was missing approximately 15 bp of the 5'-end of the mRNA. We then used as primer the 23 nucleotide BalI-Sau3AI fragment (dashed arrow, Fig. 2B) that is closer to the 5'-end to determine the sequence of the extended transcript (as seen in Fig. 2B lane 1, the primer fragment preparation contained some additional non-identified labeled DNA fragments). A single band was observed after elongation with reverse transcriptase, followed by NaOH or S1 nuclease digestion (Fig. 2B, lanes 2 and 3, arrowhead). This band was estimated to be approximately 65 nucleotides long on the basis of relative mobility using a MspI digest of pBR322 as size marker. Therefore, the primer was elongated by about 42 nucleotides. Taking into consideration that these 42 nucleotides include 29 nucleotides previously sequenced at the 5'-end of the original pS2 ds-cDNA (see Fig. 5), this result indicated that approximately 13 nucleotides were missing at the 5'-end of the original ps2 ds-cDNA. The "65 nucleotide" band was eluted from a preparative gel and sequenced (Figs. 1C, 3A and results not shown). The fidelity of the primer extension technique was demonstrated by the identity of the sequence obtained using this method with that determined in the original pS2 ds-cDNA for the 29 nucleotide segment located upstream from the BalI site. However, using this technique, it was not possible to determine conclusively the nature of the last two nucleotides at the 5'-end of the mRNA.

Determination of the pS2 RNA 5'-end sequence by cloning full-length double-stranded cDNA.

Eukaryotic expression vectors have been constructed that are compatible with the Okayama and Berg (10) cloning technique for producing full-length ds-cDNA in high yield in the correct orientation relative to eukaryo-

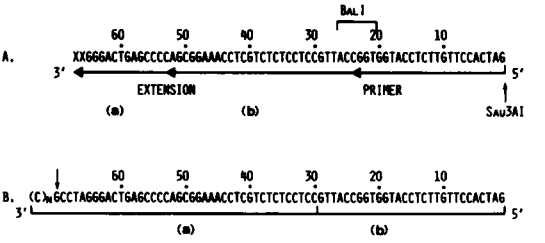

Figure 3 : A) Nucleotide sequence (coding strand) of DNA fragment obtained by reverse transcriptase extension of the BalI-Sau3AI primer (see Fig. 2B and text). The arrows indicate the extension direction of the primer. Extension (a) (see Fig. 2B and text) indicates the nucleotide sequence which could be determined solely from primer extension and was not present in the original pS2 ds-cDNA. Extension (b) indicates the nucleotide sequence which was also determined by direct sequencing of the original pS2 ds-cDNA. X represents those nucleotides that could not be determined with certainty. (B) Nucleotide sequence (coding strand) of the 5'-end region of the "full-length" ds-cDNA insert present in pSVES1. (a) is the sequence shown in Fig. 2C (the G which precedes the homopolymer dC tail is indicated by an arrow, see text); (b) is taken from (A) (original pS2 ds-cDNA sequence).

tic promoters and processing signals (11). One of these expression vectors was used to construct a ds-cDNA library from polyadenylated RNA of MCF-7 cells grown in the presence of estradiol (Fig. 4 and Materials and Methods). Two clones (pSVES1 and pSVES2) were detected by hybridization with the nick-translated original pS2 ds-cDNA. Their ds-cDNA insertions were approximately the same length as the original pS2 ds-cDNA and contained the same major restriction enzyme sites. The 5'-end region of the pS2 ds-cDNA inserts present in pSVES1 and pSVES2 were sequenced from the BalI site as described in Materials and Methods.

The result of the sequence determination for pSVES1 is shown in Figs. 2C, 1D and 3B. Five nucleotides preceeding the dC homopolymer tail which could not be determined from the primer extension sequencing gel, were characterized. However, although both pSVES1 and pSVE2 clones gave the same 5'terminal sequence, the G nucleotide shown by the arrow in Figs. 2C and 3B was present only in pSVES1. That this extra G corresponds in fact to a contamination of the dCTP used for dC tailing during the construction of the ds-cDNA library was confirmed by sequencing the 5'-end region of a cloned pS2 "genomic" gene, which indicated the absence of a BamHI site (Fig. 1D) at the 5'-end of this gene (17). From all of these results, it is clear that the original pS2 ds-cDNA was missing 15 base pairs at its 5'-end and that the first base of pS2 RNA is an A.

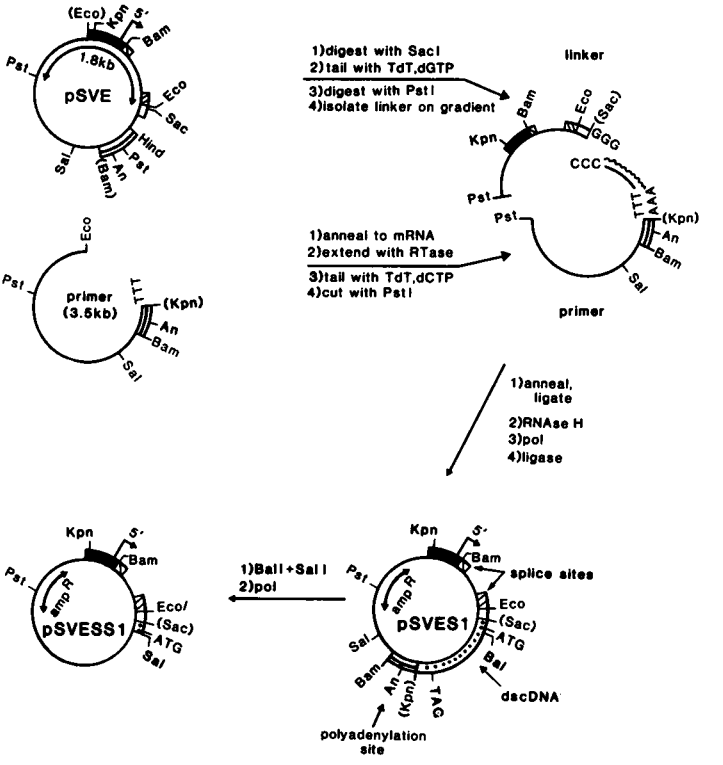

Figure 4 : Construction of plasmid pSVES1 containing a full-length pS2 double-stranded cDNA inserted into the eukaryotic expression vector pSVE (Materials and Methods and text). Abbreviations are as follows : Eco = EcoRI, Bam = BamHI, Pst = PstI, Sal = SalI, Sac = SacI, Kpn = KpnI, Hind = HindIII, Bal = BalI, pol = DNA polymerase I. TdT = terminal deoxynucleotidyl transferase, RTase = reverse transcriptase. The filled-in box is the SV40 early gene promoter region, the boxes with diagonal lines are β-globin exons, the box with a horizontal line is an SV40 fragment carrying the early messenger polyadenylation site (An) and the open box a small part of a chicken ovalbumin gene exon (see ref. 11). The mRNA is shown as a wavy line, and the ds-cDNA as a box with a dotted line. The bla gene of plasmid pBR327 is shown as amp^R. The arrow 5' on pSVE, pSVES1 and pSVESS1 indicates the approximate location of the SV40 early startsites. Restriction enzyme sites which are bracketed have been destroyed during the construction of the recombinants. Drawings are not to scale.

Nucleotide sequence of the pS2 RNA.

The complete nucleotide sequence of the pS2 RNA is shown in Fig. 5. There is a 5'-untranslated region consisting of 40 nucleotides, a 252 nucleotide-long coding region extending from nucleotide 41 to 292 and a 3'-untranslated region of 198 nucleotides preceding the polyA tail. Thus the

length of the complete unpolyadenylated pS2 RNA is 490 nucleotides, in good agreement with the estimated length of the bulk of the polyadenylated pS2 RNA as approximately 600 nucleotides (5). The 5'-terminal base in the mRNA sequence is adenine. This is consistent with the capping sites of other eukaryotic mRNAs where A often follows the 7-methyl G cap (18).

Two regions of complementarity with the 3'-end sequence of 18S rRNA of higher eukaryotes occur in the 5'-untranslated region of pS2 RNA. The first sequence, 5'-CCUU-3', which occurs 19 nucleotides downstream from the RNA 5'-terminus, is complementary to nucleotides 3'-GGAA-5' of the conserved sequence 3'-AUUACUAGGAAGGCGUCC-5', present at the 3'-end region of the 18S rRNAs of higher eukaryotes (19). In addition, the sequence, 5'-GCAG-3', which is located five nucleotides downstream from the CCUU sequence, is complementary to nucleotides 3'-CGUC-5' at the 3'-end of the 18S rRNA. This is consistent with the observation that there is some base sequence complementarity between the 5'-untranslated region of a large number of eukaryotic mRNAs and the 3'-end region of 18S rRNAs, which might play a role in initiation of translation (for refs., see 19-21).

The 3'-untranslated region consists of 195 nucleotides between the translation stop signal and before the polyA tail. It has been shown that the hexanucleotide A-A-U-A-A-A found 15-25 nucleotides upstream from the polyadenylation site in most eukaryotic mRNAs serves as part of the signal for polyadenylation (22-24). In pS2 mRNA, the sequence AUUAAA is present 14 nucleotides upstream from the polyA tract. This somewhat atypical sequence has also been reported for other eukaryotic genes (for refs., see 25, 26).

Sequence of the putative pS2 protein.

An amino acid sequence derived from the sequence of the pS2 mRNA is shown in Fig. 5. It corresponds to a possible 84 amino acids long polypeptide with a molecular weight of 9140 daltons. The amino acid sequence shown here represents the longest open reading frame. The translational initiation site was tentatively assigned to the methionine codon at position 41 of the mRNA sequence, rather than to the two AUG located further downstream (positions 50 and 89) for the following reasons. According to the translational "scanning model" (27), eukaryotic ribosomes bind to the 5'-terminus of mRNA and migrate along the mRNA sequence until they encounter the first AUG triplet, which, by virtue of its position, is the initiation codon. An extended version of the scanning model (28, 29) was proposed in which the nucleotides flanking the AUG initiation codon play an important role in the recognition by eukaryotic ribosomes, the most favorable sequence for favorable sequence

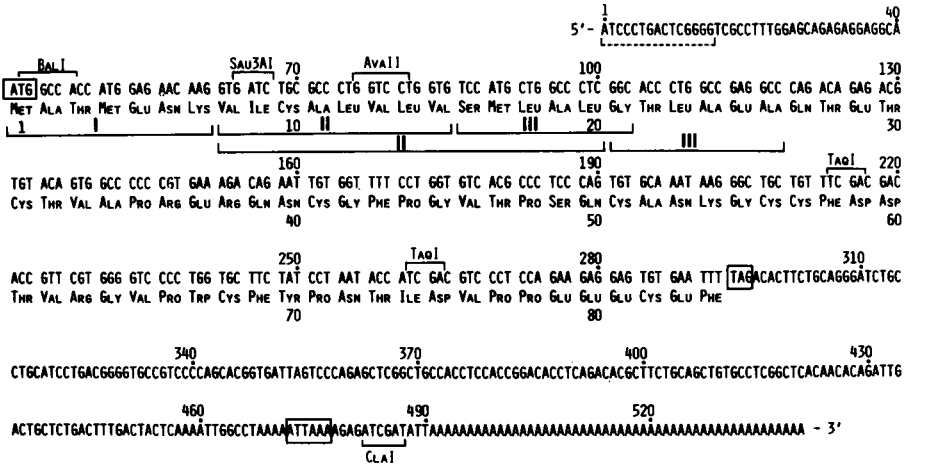

Figure 5 : The nucleotide sequence of pS2 full-length cDNA. The putative start and stop codons and the polyadenylation signal are boxed. The portions of the sequence labeled I, II and III represent sections of the putative signal peptide (see text). Some key restriction enzyme sites are indicated (see text). The fifteen 5'-end terminal nucleotides which are missing in the original pS2 ds-cDNA are underlined with a dashed line.

for initiation being $\begin{matrix}A\\G\end{matrix}$-N-N-A-U-G-G. The AUG at positions 41 and 50 have such a flanking sequence, while the third AUG codon further downstream in the sequence does not.

An examination of the N-terminal sequence of the presumptive pS2 protein suggests that it contains a signal peptide typical of proteins which are secreted (30). A number of signal peptides have been analyzed in prokaryotes and eukaryotes in terms of both amino acid sequences and functions (31-35). These peptides consist of 15-30 amino acid residues with a highly hydrophobic central core. According to the loop model of secreted proteins suggested by Inouye and Halegoua (31), and modified recently by Von Heijne (34) and Perlman and Halvorson (35), the signal peptide is composed of three sections. Section I is the N-terminal section having one or more basic amino acid residues. Section II, the central portion, the core, is composed on the average of 12 hydrophobic amino acid residues. The end of the core section is defined by a charged residue, a sequence of amino acid residues which may introduce a β-turn in a polypeptide or an interruption in potential α-helix or β-extended strand structure. In section III, following the hydrophobic core and preceding the cleavage site, there is a highly non-random and

localized distribution of alanine within the initial 8 positions. The most frequent sequence preceding the signal peptidase cleavage site has been observed to be Ala-X-Ala and the cleavage site is preferentially located after the sixth amino-acid following the core sequence. As indicated in Fig. 5, the sequence starting with the first methionine exhibits features that are characteristic of a signal peptide. There is a basic lysine residue 6 amino acids away from this methionine (Section I). The sequence then contains 12 non-polar hydrophobic amino-acids interrupted by only 2 polar amino-acid residues (cysteine and serine; Section II). According to the model, the signal peptidase cleavage site is likely to occur following alanine at position 26 (Section III). This would allow a signal peptide sequence consisting of 26 amino-acid residues and a secreted protein of 58 residues. Alternatively, the signal peptidase cleavage site could be located after the glycine at position 21. Such a location, which would also be in agreement with the model, would reduce the size of the core section (II) to 8 residues, instead of 13, and increase the length of the secreted protein to 63 residues (see Fig. 5). Thus the molecular weight of the secreted protein would be either 6450 daltons or 6970 daltons depending on where the signal peptidase cleavage occurs.

Besides the putative hydrophobic signal peptide sequence present a the amino terminus of pS2 protein, there is another stretch of 8 very hydro phobic residues near the carboxyl terminus (positions 64-71) and a moderately hydrophobic stretch of 9 residues in the central portion of the coding sequence (positions 41-49). The overall percentage of hydrophobic amino-acids in the sequence is 30%. The carboxyl terminus is very hydrophilic, consisting of a stretch of 10 hydrophilic residues, 4 of which are glutamic acid and 1 aspartic acid. These amino-acid arrangements are reminiscent of transmembrane proteins, such as is observed for the heavy chains of the major histocompatibility complex antigens (36, 37) glycophorin (38) and membrane IgM (39). An examination of the charged residues in the putative unprocessed protein indicates that there are 11 acidic residues (8 glutamic acid and 3 aspartic acid residues), while there are 5 basic residues (2 lysine and 3 arginine residues). The putative 58 amino-acid processed polypeptide would contain 9 acidic and 4 basic residues.

Codon usage.

Table I shows the codon usage for pS2 mRNA The codon usage is non-random and very similar to that found in other eukaryotic mRNAs (see for instance : 40-44). It reflects a high preference for G or C in the third

TABLE I : CODON USAGE FOR pS2 mRNA

	T	C	A	G
T	Phe TTT 2 Phe TTC 2 Leu TTA Leu TTG	Ser TCT Ser TCC 2 Ser TCA Ser TCG	Tyr TAT 1 Tyr TAC TAA TAG 1	Cys TGT 5 Cys TGC 3 TGA Trp TGG 1
C	Leu CTT Leu CTC 1 Leu CTA Leu CTG 4	Pro CTT 3 Pro CCC 3 Pro CCA 1 Pro CCG	His CAT His CAC Gln CAA Gln CAG 3	Arg CGT 2 Arg CGC Arg CGA Arg CGG
A	Ile ATT Ile ATC 2 Ile ATA Met ATG 3	Thr ACT Thr ACC 4 Thr ACA 2 Thr ACG 2	Asn AAT 3 Asn AAC 1 Lys AAA Lys AAG 2	Ser AGT Ser AGC Arg AGA 1 Arg AGG
G	Val GTT 1 Val GTC 4 Val GTA Val GTG 3	Ala GCT Ala GCC 6 Ala GCA 1 Ala GCG	Asp GAT Asp GAC 3 Glu GAA 3 Glu GAG 5	Gly GGT 2 Gly GGC 2 Gly GGA Gly GGG 1

position of the codons : 33 codons (39%) terminate in C, 24 (28%) in G, 19 (23%) in T and 8 (10%) in A. There are two ACG triplets at positions 30 and 47 coding for threonine. Codons ending in CG are rare in eukaryotes (44).

DISCUSSION

We report here the complete sequence of pS2 mRNA, a mRNA whose level is rapidly increased by the addition of estradiol to the culture medium of the MCF-7 human breast cancer cell line (see Introduction). The original pS2 ds-cDNA clone (5) was obtained by classical cloning techniques involving an S1 nuclease digestion step. In the present study, pS2 ds-cDNA clones were also constructed by direct cloning into a eukaryotic expression vector (11), a procedure in which S1 nuclease is not used and where the likelihood of obtaining a full-length ds-cDNA product is greater. Both cloning techniques yielded the ds-cDNA with complete coding regions. The size of the specific cDNA transcripts formed by primer extension with reverse transcriptase indicated that not more than 15 nucleotides were missing from the originally cloned pS2 ds-cDNA sequence. These 15 nucleotides were present in the ds-cDNA obtained by direct cloning in the eukaryotic expression vector.

The pS2 mRNA sequence is characterized by a high overall G + C content (56% compared to the average 40% G + C content of the human genome). The G + C content decreases drastically as one enters the last 60 nucleo-

tides of the 3' untranslated region : 62.5% G + C in the 5'-untranslated region, 58% in the coding region, 62% and 31% in the first 37 nucleotides and in the last 60 nucleotides of the 3'-untranslated region, respectively. In the coding region, this bias is due to an assymetry in the base choices for the third codon position : 68% for G + C and 32% for A + T. A similar bias has been observed for other mammalian mRNAs (for refs., see 45). In addition, there is no bias against the CpG dinucleotide. The ds-cDNA sequence has a high frequency of restriction sites for the "CpG" restriction enzymes such as HhaI, HpaII, FnuDII and TaqI. The CpG/GpC ratio is 0.77 in the protein coding region. This ratio is similar to that observed for rat skeletal α-actin (0.78), human α-globin (0.80) and human proopiomelanocortin (0.79), and very different from that observed for human cardiac α-actin (0.51), human growth hormone (0.45) and human β-globin (0.14) (for refs., see 45). The high CpG content of the sequence of pS2 mRNA is also reflected in the presence of two rare ACG threonine codons (44) at positions 30 and 47. It has been observed by Hanauer et al. (45) that in human mRNA sequences, a large excess of G + C at the third codon position is correlated with an absence of bias against CpG dinucleotides. These authors have suggested that the particular codon usage found in some mRNAs reflects more the presence of a gene in a specific genomic region which differ in its base composition over long distances, than some selective pressure related to protein synthesis mechanims.

The initiation codon was tentatively assigned to the methionine triplet at positions 41-43, which is the first AUG codon within the deduced mRNA sequence. The sequence of the 21-26 amino acid residues starting with this initial methionine exhibits features that are characteristic of the signal peptides present at the amino terminus of secretory proteins (see Results). Possible sites for cleavage of the signal peptide are located after the glycine residue at position 21 or after alanine residue at position 26. Thus after removal of the signal peptide, the putative 9120 dalton preprotein might be processed to a polypeptide of 58 or 63 amino-acid residues, and the molecular weight reduced to 6450 daltons or 6970 daltons. After removal of the signal peptide, the remaining protein would contain 7 cysteine residues which are distributed throughout the sequence. These residues may form disulphide bonds, as is the case for peptide hormones such as insulin and growth hormone (46). There is also a hydrophobic stretch of 8 amino-acid residues near the carboxyl terminus of the sequence extending from the glycine residue at position 64 to the proline residue at position

71. The presence of this second hydrophobic region suggests that in addition to being secreted, the protein could conceivably be a transmembrane protein. Experiments are currently in progress to investigate in vivo and in vitro the protein product corresponding to pS2 mRNA.

The molecular weight of the putative pS2 protein does not correspond to the size of any of the various polypeptides which, up to now, have been shown to be induced by estrogen in MCF-7 cells (see 47 and 48). The pS2 protein does not correspond either to the product of the transforming gene which has been characterized in the genome of MCF-7 cells (49), since this gene is not cut by BamHI and EcoRI restriction enzymes, whereas the pS2 gene is cut by these two enzymes within its first intron (17). The sequence of the pS2 ds-cDNA and the derived polypeptide were compared with the nucleic acid and protein sequence databases at the National Biomedical Research Foundation, Georgetown University Medical Center, Washington, D.C. by Dr. M.O. Dayhoff. No closely related DNA or amino-acid sequence homologies were found. However, some characteristics of the amino-acid sequence of pS2 protein (low molecular weight, high cysteine content, putative signal peptide sequence) prompted us to compare its structure to that of known growth factors. Some overall structural similarities were observed with the amino-acid sequences of the growth hormone-dependent insulin-like polypeptides termed somatomedins, and particularly with human insulin-like growth factor I and II (IGF I and II). The length of the amino-acid sequence of pS2 protein is similar to that of the human insulin-like growth factors; 58 or 63 amino-acids for the pS2 protein after cleavage of the putative signal peptide sequence compared to 67 and 70 amino-acids, respectively for human IGF I and II (50, 51). The putative pS2 protein and IGFs are also similar in their richness in cysteine residues. The IGFs contain 6 half-cystine residues which can form 3 disulphide bonds. The amino-acid sequence of pS2 protein contains 8 half-cystine residues, 7 of which are located after the putative signal peptide sequence, and thus also has the capacity to form 3 disulphide bonds. In addition, the arrangement of some cystine residues appears to be similar in the sequence of the pS2 proteins and IGFs. In both cases the sequence Cys-Cys-Phe (positions 56-58 in pS2 protein) is present and followed by two additional cysteine residues located further downstream. Cysteine residue clusters have also been found in other growth factors, such as nerve growth factor (52, 53), epidermal and tumor growth factor (see 54) and plalelet-derived growth factor (55), but there is no obvious amino-acid sequence homologies between the putative pS2 protein and these growth

factors. However, this absence of homology does not rule out that the pS2 protein might be an estrogen-inducible growth factor, since there is no sequence similarity between some of these growth factors either. Whether or not the pS2 protein is a member of the estromedin family (56), a group of estrogen-inducible growth factors which are still poorly characterized, is an interesting question with implications for both the mode of action of estrogens and the etiology of human breast cancer.

ACKNOWLEDGEMENTS

We thank Dr. R. Heilig for his kind help with DNA sequencing and gratefully acknowledge the excellent technical assistance of M.C. Gesnel. and the expert help of B. Boulay and C. Werlé with the elaboration of the figures. This work was supported by grants from the CNRS (ATP 3582), the INSERM (PRC 118012), the Fondation pour la Recherche Médicale, the Association pour le Développement de la Recherche sur le Cancer and the Fondation Simone et Cino del Duca. S.B. Jakowlew was a post-doctoral fellow of the American Cancer Society (grant number PF-2156).

*To whom correspondence should be sent

REFERENCES

1. Buetti, E., and Diggelmann, H. (1983) EMBO J. 2, 1423-1429.
2. Groner, B., Hynes, N.E., Rahmsdorf, U. and Ponta H. (1983) Nucleic Acids Res. 11, 4713-4725.
3. Ostrowski, M.C., Richard-Foy, H., Wolford, R.G., Berard, D.S. and Hager, G.L. (1983) Mol. and Cell. Biology 11, 2045-2057.
4. Payvar, F. DeFranco, D., Firestone, G.L., Edgar, B., Wrange, O., Okret, S., Gustafsson, J.A. and Yamamoto, K.R. (1983) Cell 35, 381-392.
5. Masiakowski, P., Breathnach, R., Bloch, J., Gannon, F., Krust, A. and Chambon, P. (1982) Nucleic Acids Res. 10, 7895-7903.
6. Brown, A.M.C., Krust, A., Jeltsch, J.M., Roberts, M. and Chambon, P. (1984), in preparation.
7. Chambon, P., Dierich, A., Gaub, M.P., Jakowlew, S.B., Jongstra, J., Krust, A., LePennec, J.P., Oudet, P. and Reudelhuber, T. (1984) in "Recent Progress in Hormone Research", Vol. 40, in press.
8. Krust, A. and Chambon, P. (1984) in preparation.
9. Auffray, C. and Rougeon, F. (1980) Eur. J. Biochem. 107, 303-314.
10. Okayama, H. and Berg, P. (1982) Mol. Cell. Biol. 2, 161-170.
11. Breathnach, R. and Harris, B.A. (1983) Nucleic Acids Res. 11, 7119-7136.
12. Maxam, A.M. and Gilbert, W. (1980) in Methods in Enzymolgy, Colowick, S.P. and Kaplan, N.O., Eds., Vol. 65, pp 499-560.
13. Weislander, L. (1979) Anal. Biochem. 98, 305-309.
14. Smith, D.R. and Calvo, J.M. (1980) Nucleic Acids Res. 17, 2255-2273.
15. Messing, J., Crea, R. and Seeburg, P.H. (1981) Nucleic Acids Res. 10, 309-321.
16. McClelland, M. (1983) Nucleic Acids Res. 11, r169-r173.

17. Jeltsch, J.M., Brown, A.M.C., Garnier, J.M., Schatz, C., Roberts, M. and Chambon, P. (1984) in preparation.
18. Breathnach, R. and Chambon, P. (1981) Ann. Rev. Biochem. 50, 349-383.
19. Hagenbuchle, O., Santer, M., Steitz, J. and Mans, R.J. (1978) Cell 13, 551-563.
20. Kozak, M (1978) Cell 15, 1109-1123.
21. Cochet, M., Gannon, F., Hen, R., Maroteaux, L., Perrin, F. and Chambon, P. (1979) Nature 282, 567-574.
22. Proudfoot, N.J. and Brownlee, G.G. (1976) Nature 263, 211-214.
23. Fitzgerald, M. and Shenk, T. (1981) Cell 24, 251-260.
24. Montell, C., Fisher, E.F., Caruthers, M.H. and Berk, A.J. (1982) Nature 305, 600-605.
25. Gerlinger, P., Krust, A., LeMeur, M., Perrin, F., Cochet, M., Gannon, F., Dupret, D. and Chambon, P. (1982) J. Mol. Biol. 162, 345-364.
26. Scott, J., Selby, M., Urdea, M., Quiroga, M., Bill, G.I. and Rutter, W.J. (1983) Nature 302, 538-540.
27. Kozak, M. (1981) in Current Topics in Microbiology and Immunolgy, Shatkin, A.J., Ed. (Springer Verlag - Berlin), Vol. 293, pp 81-123.
28. Kozak, M. (1981) Nucleic Acids Res. 9, 5233-5252.
29. Kozak, M. (1982) J. Mol. Biol. 156, 807-820.
30. Blobel, G. and Dobberstein, B. (1975) J. Cell Biol. 67, 835-851.
31. Inouye, M. and Halegoua, S. (1979) Crit. Rev. Biochem. 7, 339-371.
32. Davis, B.D. and Tai, P.C. (1980) Nature 283, 433-438.
33. Muller, F. and Clarkson, S.G. (1980) Cell 19, 345-353.
34. Von Heijne, G. (1983) Eur. J. Biochem. 133, 17-21.
35. Perlman, D. and Halvorson, H.O. (1983) J. Mol. Biol. 167, 391-409.
36. Coligan, J.E., Kindt, T.J., Vehara, H., Martinko, J. and Nathenson, S.G. (1981) Nature 297, 35-39.
37. Korman, A.J., Knudsen, P.J., Kaufman, J.F. and Strominger, J.L. (1982) Proc. Natl. Acad. Sci. USA 79, 1844-1848.
38. Tomita, M., Furthmayr, H. and Marchesi, V.T. (1978) Biochem. 17, 4756-4770.
39. Rogers, J., Early, P., Carter, C., Calames, K., Bond, M., Hood, L. and Wall, R. (1980) Cell 20, 303-312.
40. Ulrich, A., Shine, J., Chirgwin, J., Pictet, R., Tischer, E., Rutter, W. and Goodman, H. (1977) Science 196, 1313-1318
41. Efstratiadis, A., Kafatos, F.C. and Maniatis, T. (1977) Cell 10, 571-585.
42. Seeburg, P.H., Shine, J., Martial, J.A., Baxter, J.D. and Goodman, H.M. (1977) Nature 270, 486-494.
43. Nakanishi, S., Inoue, A., Kita, T., Nakamura, M., Chang, A.C.Y., Cohen, S.N. and Numa, S. (1979) Nature 278, 423-427.
44. Grantham, R., Gautier, C., Gomy, M., Jacobzone, M. and Mercier, R. (1981) Nucleic Acids Res. 9, r43-r74.
45. Hanauner, A., Levin, M., Heilig, R., Daegelen, D., Kahn, A. and Mandel, J.L. (1983) Nucleic Acids Res. 11, 3503-3516.
46. Dayhoff, M.O. (1972) in Atlas of Protein Sequence and Structure, Vol. 5, pp. 173-227 (Natl. Biomedical Research Foundation, Washington D.C.).
47. Adams, D.J., Edwards, D.P. and McGuire, W.L. (1983) in Regulation of Gene Expression by Hormones, McKerns, K.W. Ed., (Plenum Press, N.Y.), pp. 1-25.
48. Rochefort, H. (1983) in Regulation of Gene Expression by Hormones, McKerns, K.W., Ed. (Plenum Press, N.Y.), pp. 27-37.
49. Lane, M.A., Sainten, A. and Cooper, G.M. (1981) Proc. Natl. Acad. Sci. USA, 78, 5185-5189.
50. Rinderknecht, E. and Humbel, R.E. (1978) J. Biol. Chem. 253, 2769-2776.

51. Rinderknecht, E. and Humbel, R.E. (1978) FEBS Lett. 89, 283-286.
52. Angeletti, R.H., Hermodsen, M.A. and Bradshaw, R.A. (1973) Biochem. 12, 100-115.
53. Hogue-Angeletti, R.H., Frazier, W.A., Jacobs, J.W., Niall, H.D. and Bradshaw, R.A. (1976) Biochem. 15, 26-34.
54. Marquardt, H., Hunkapiller, M.W., Hood, L.E., Twardzik, D.R., De Larco, J.E., Stephenson, J.R. and Todaro, G.J. (1983) Proc. Natl. Acad. Sci. USA 8, 4684-4688.
55. Waterfield, M.D., Scrace, G.T., Whittle, N., Strooban, P., Johnson, B., Heldin, C-H., Huang, J.S. and Deuel, T.F. (1983) Nature 304, 35-39.
56. Ikeda, T., Liu, Q.F., Danielpour, D., Officer, J.B., Masayoshi, I., Leland, F.E. and Sirbasku, D.A. (1982) In Vitro 18, 961-979.