The EMBO Journal vol.3 no.2 pp.429-437, 1984

Isolation and characterization of the human prolactin gene

Anh T. Truong, Colette Duez, Alexandra Belayew, André Renard, Raymond Pictet[1], Graeme I. Bell[2] and Joseph A. Martial*

Laboratoire central de génie génétique, Institut de Chimie, B6, Université de Liège, 4000 Sart-Tilman, Belgium, [1]Institut de recherche en biologie moléculaire, CNRS, Université Paris VII, Tour 43, 2, Place Jussieu, 75221 Paris cedex 05, France, and [2]Chiron Corporation, 4560 Horton Street, Emeryville, CA 94608, USA

*To whom reprint requests should be sent
Communicated by W. Fiers

Prolactin (PRL) and growth hormone (GH) genes derive from a common ancestor and still share some sequence homologies. Their expression in the pituitary gland is regulated in opposite directions by most of the many hormones acting on them. This provides an interesting system to study sequences involved in gene expression. Using a human PRL cDNA clone as a probe, we screened a human genomic DNA library in λ phage and isolated a single recombinant comprising the whole hPRL gene. It was characterized by restriction endonuclease mapping and cDNA hybridization, by DNA heteroduplex analysis and by nucleotide sequencing. The hPRL gene is present as a single copy per haploid genome, is ~10 kb long and contains four introns, three of which interrupt the coding sequence at the same locations as in the known GH and PRL genes. The origin of transcription was determined by S1 mapping on prolactinoma mRNAs. The search for direct and inverted repeats, as well as dyad symmetries was carried out in the 900-bp sequenced in the 5'-flanking region. Sequence homologies between hPRL, hGH and rPRL were derived from computer drawn matrices for these upstream regions.

Key words: DNA sequence homologies/exon-intron junctions/ dyad symmetries/gene family/methylation

Introduction

Prolactin (PRL), growth hormone (GH) and chorionic somatomammotropin (CS, also known as placental lactogen) constitute a family of polypeptide hormones related by structural and biological properties (Niall *et al.*, 1971). PRL and GH are synthesized in the anterior pituitary by the mammotrophs and the somatotrophs, respectively, while CS is produced by the placenta syncytio-trophoblast.

PRL and GH gene expression is under complex hormonal control. Studies carried out on primary pituitary cultures as well as on established rat pituitary cell lines have shown that thyroid hormones, glucocorticoids, estrogens, thyrotropin-releasing hormone, epidermal growth factor, dopamine etc., directly regulate the levels of PRL and/or GH mRNAs (reviewed by Tashjian, 1979; Gourdji *et al.*, 1982; Dobner *et al.*, 1981; Evans *et al.*, 1982; Wegnez *et al.*, 1982; Spindler *et al.*, 1982; Maurer, 1982a, 1982b; Murdoch *et al.*, 1982; Laverrière *et al.*, 1983). PRL and GH transcription is generally regulated in opposite directions by most of the above cited hormones, providing a very interesting system to study the mechanisms of hormonal control of gene expression.

PRL and GH have 39% sequence homology at the mRNA level in the rat, 42% in humans (Cooke *et al.*, 1980, 1981). The PRL - GH - CS family most probably arose by consecutive duplications of a common ancestral gene (Niall *et al.*, 1971), the first of which took place ~400 million years ago (Cooke *et al.*, 1980).

The cloning and analysis of the rat GH and PRL genes have been reported: they both seem to be present only once per haploid genome in that species (Barta *et al.*, 1981; Chien and Thompson, 1980; Gubbins *et al.*, 1980; Page *et al.*, 1981). In humans, as many as seven genes (including pseudogenes) have been identified for GH and CS (Fiddes *et al.*, 1978; Lawn *et al.*, 1978; Moore *et al.*, 1982; Seeburg, 1982) all of which are located on chromosome 17 (Owerbach *et al.*, 1980). The evolutionary origin of this large set is unclear: either CS could have arisen by duplication of a different related gene in several species (convergent evolution) or an ancestral CS gene could have been modified after rats and humans diverged, by recombination or gene conversion (Moore *et al.*, 1982). Human PRL sequences have been located on chromosome 6 (Owerbach *et al.*, 1981).

We report here the cloning and partial characterization of the human PRL gene. It is present as a single copy per haploid genome, is ~10 kb long and contains four introns, three of which interrupt the coding sequence at the same locations as in the known GH, CS and PRL genes. We also determined, by S1 mapping, the origin of transcription of the hPRL mRNA produced *in vivo*.

Results

Isolation of the cloned hPRL gene

Some of us have previously described the cloning of a human PRL cDNA in pBR322 (phPRL cDNA) that comprises the entire amino acid coding sequence and most of the 3'-untranslated region (Cooke *et al.*, 1981). This cloned hPRL cDNA was first used as a probe to analyse the distribution of PRL sequences in the human genome. Total human DNA was digested with the restriction endonuclease *Eco*RI; the resulting fragments were separated by electrophoresis on agarose gel and transferred to nitrocellulose filters where they were hybridized to radiolabelled hPRL cDNA. The autoradiography (Figure 1) displays four bands corresponding to DNA fragments of 0.7, 3.4, 3.9 and 4.1 kb. A shorter exposure of this blot or a longer agarose gel electrophoresis clearly shows the 4.1- and 3.9-kb *Eco*RI fragments (data not shown). As the hPRL cDNA sequence is cut only once by *Eco*RI (Cooke *et al.*, 1981) the PRL gene contains at least two intervening sequences. The pattern generated in Figure 1 is identical for genomic DNA extracted from HeLa cells (lane 1) and placenta (lane 2). However, white blood cell DNA (lane 3) does not present either the 3.4- or the 0.7-kb fragments and this has been observed in 20 different individuals (data not

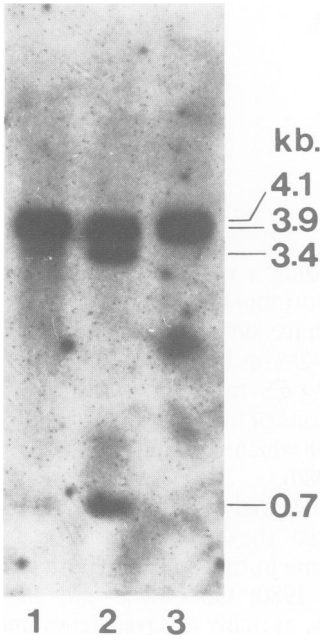

Fig. 1. Hybridization of radiolabelled cloned hPRL cDNA, to human genomic DNA. 10 μg of DNA isolated from cultured HeLa cells (**lane 1**), placenta (**lane 2**) and white blood cells (**lane 3**) was digested with *Eco*RI and the fragments separated by electrophoresis in a 1% agarose gel. The sizes of the *Eco*RI fragments hybridizing in each of these DNA preparations are indicated.

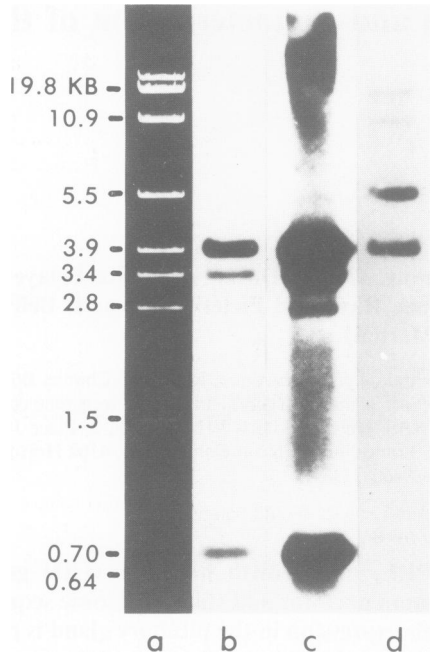

Fig. 2. Analysis of the human prolactin gene-containing recombinant phage, λ Ch4A hPRL. An *Eco*RI digest was separated by electrophoresis on a 1% agarose gel. **Lane a**: ethidium bromide staining. **Lanes b** and **c**: hybridization of Southern blots of **lane a** to hPRL cDNA clone; autoradiography was overexposed in **lane c. Lane d**: hybridization of Southern blot to a human repetitive sequence probe derived from the human insulin gene 3′-flanking sequence (see text).

shown, see Discussion).

The hPRL cDNA probe was then used to screen by *in situ* plaque hybridization (Benton and Davis, 1977) the human genomic library in bacteriophage λ provided by T. Maniatis (Lawn *et al.*, 1978). From the 1.25 x 10^6 plaques screened, only one hybridized to the ^{32}P-labelled phPRL cDNA. This recombinant phage λ Ch4A hPRL was purified and its DNA characterized. Digestion by *Eco*RI generates seven fragments besides the vector arms (Figure 2, lane a). The DNA was then transferred onto nitrocellulose and hybridized to radiolabelled phPRL cDNA: a short (lane b) and a longer (lane c) exposure of the autoradiography reveal that in the λ recombinant the coding sequences of the hPRL gene span four fragments of, 3.9, 3.4, 2.8 and 0.7 kb, respectively. This DNA blot was also probed with a DNA fragment extending from 6 to 8 kb downstream from the human insulin gene sequence where an Alu repeat and other short repeated elements have been identified (Bell *et al.*, 1980). The autoradiography (Figure 2, lane d) clearly shows that the 5.5-and 3.9-kb fragments of the hPRL recombinant phage contain repeated sequences. These same fragments, and no others, also hybridize with ^{32}P-labelled total human genomic DNA under conditions (Bell *et al.*, 1980) where only repeated sequences hybridize (data not shown).

Analysis of heteroduplexes by electron microscopy

To identify the PRL coding regions in λ Ch4A hPRL, heteroduplexes were formed between the recombinant phage DNA, the phPRL cDNA plasmid and native λ DNA. The plasmid had been previously linearized by digestion with *Bam*HI which does not cleave within the inserted cDNA and leaves pBR322 tails of ~1.1 and 3.2 kb on the 5′ and 3′ ends of the insert, respectively. Since the cDNA orientation within the phPRL plasmid had been determined (Cooke *et al.*, 1981), we were able to assign the 5′ and 3′ sides on the heteroduplex

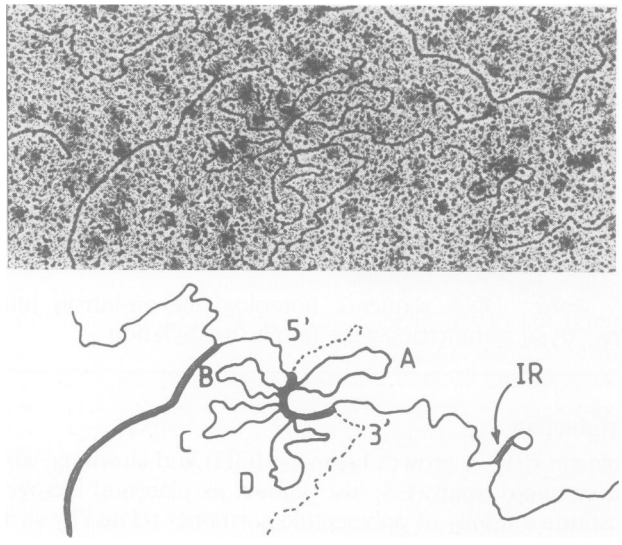

Fig. 3. **(A)** Electron microscopy picture of hPRL cDNA-genomic DNA heteroduplex. The recombinant phage λ Ch4A hPRL was mixed with non-recombinant λ Ch4A and a 10-fold molar excess of phPRL cDNA plasmid that was asymmetrically linearized by *Bam*HI. They were denatured and allowed to hybridize as described in Materials and methods. **(B)** Interpretative drawing. Single- and double-stranded DNA are represented by thin and thick solid lines, respectively, except for single-stranded pBR322 sequences which are shown by dashed lines. The orientation 5′ to 3′ is that of the cDNA clone (see text), introns are lettered A to D and an inverted repeat is denoted IR.

structures visualized on the electron micrographs. One of these is presented in Figure 3, showing four intervening sequence loops (A – D) and five homologous coding regions.

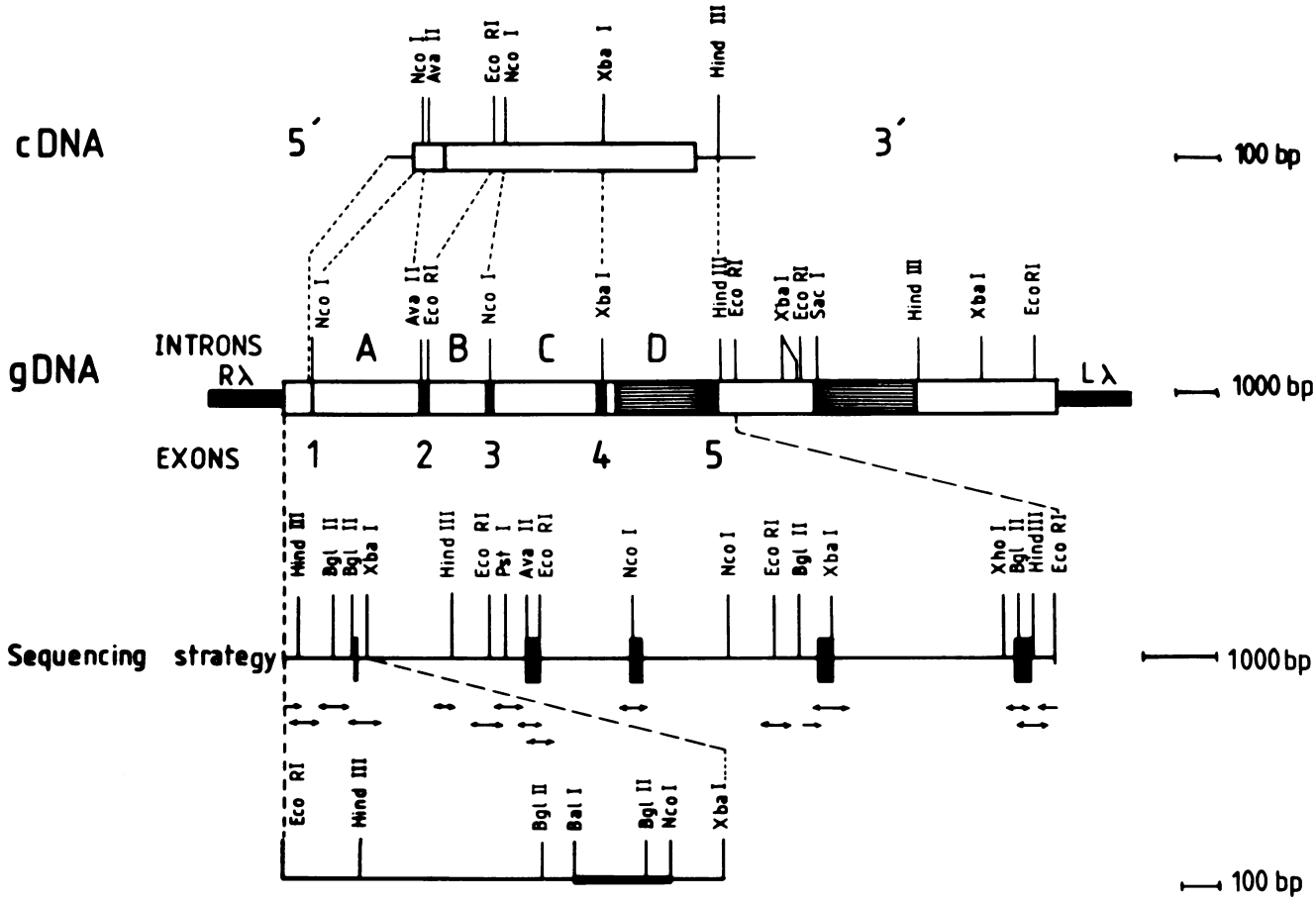

Fig. 4. Restriction map of the hPRL gene. **First line**: alignment of the cloned hPRL cDNA restriction sites over the gene ones. **Second line**: genomic DNA clone λ Ch4A hPRL. The exons are the black boxes numbered 1–5 and the introns are lettered A–D. The two grey boxes are regions containing Alu family dispersed middle repetitive sequences. λ arms are drawn as a thick horizontal line. The orientation of the cloned gene is 5′ to 3′ from the right arm (R) to the left (L). **Third line**: enlargement of the hPRL gene itself with the restriction sites used for sequencing. Horizontal arrows show the direction and extent of sequence readings. **Fourth line**: further enlargement of the hPRL gene 5′ region showing the probe used for S1 mapping (thick line).

The following sizes have been derived from the electron microscopic analysis. The exons range in length from 70 to 370 bp while the four introns have sizes of 2.4, 1.8, 2.4 and 2.6 kb from A to D, respectively. The intervening sequences expand the hPRL gene to ~10 kb for an mRNA 10 times shorter (Cooke *et al.*, 1981).

A stem and loop structure is formed by the hPRL gene recombinant phage itself (IR in Figure 3) ~3 kb downstream from the 3′ end of its homology to the cDNA insert of the phPRL plasmid. This structure characterizes two inverted homologous segments of ~500 bp each, located ~600 bp from one another. These inverted repeats are found in the 5.5-kb *Eco*RI fragment that hybridizes to a middle repetitive sequence probe, as shown in Figure 2, lane d.

Restriction map and sequence analysis of the hPRL gene

A detailed restriction map of the cloned hPRL gene is shown in Figure 4. The positions of the coding regions were deduced from Southern blots of λ Ch4A hPRL DNA digests hybridized with different probes derived from the hPRL cDNA recombinant plasmid (Figure 2, and data not shown). From the known nucleotide sequence of the hPRL cDNA clone (Cooke *et al.*, 1981) several restriction sites found in the genomic DNA could be attributed to exons, taking into account the sizes of the intervening sequences deduced from the heteroduplex analysis (Figure 3).

The four *Eco*RI fragments spanning the hPRL gene were subcloned into pBR325 and used for detailed restriction mapping and sequencing. The strategy used to locate precisely the four introns was the following: sequences were determined (Maxam and Gilbert, 1980) in both directions from each of five unique restriction sites attributed to exons 2 to 5 by comparison with the cDNA sequence. A stretch of ~350 bp downstream from the last exon as well as the 1200 bp located at the 5′ end of the cloned genomic DNA were also sequenced. The latter region includes exon 1 and 900 bp of 5′-flanking DNA. The sequencing strategy is indicated on the third line of Figure 4 and the sequence is displayed in Figure 5. We found no difference between the five exons and the previously published hPRL cDNA sequence (Cooke *et al.*, 1981), implying that we had cloned a functional gene. As one single PRL recombinant phage was picked out of 1.25×10^6, this PRL gene is most likely the only one present in the human haploid genome.

Exon 1 comprised 56 bp corresponding to the mRNA 5′-untranslated sequence, nine codons of the signal peptide and the first nucleotide of the tenth codon. Exon 2 begins with the last two nucleotides of the tenth codon followed by the 18 last triplets coding for the signal peptide and the first 40 codons of the mature protein. Exons 3 and 4 code for amino acids 41–76 and 77–136, respectively. Exon 5 includes the last 63 codons (amino acids 137–199), an 'ochre' termination triplet (TAA) and 118 bp before the AATAAA polyadenylation signal located 20 nucleotides 5′ to the

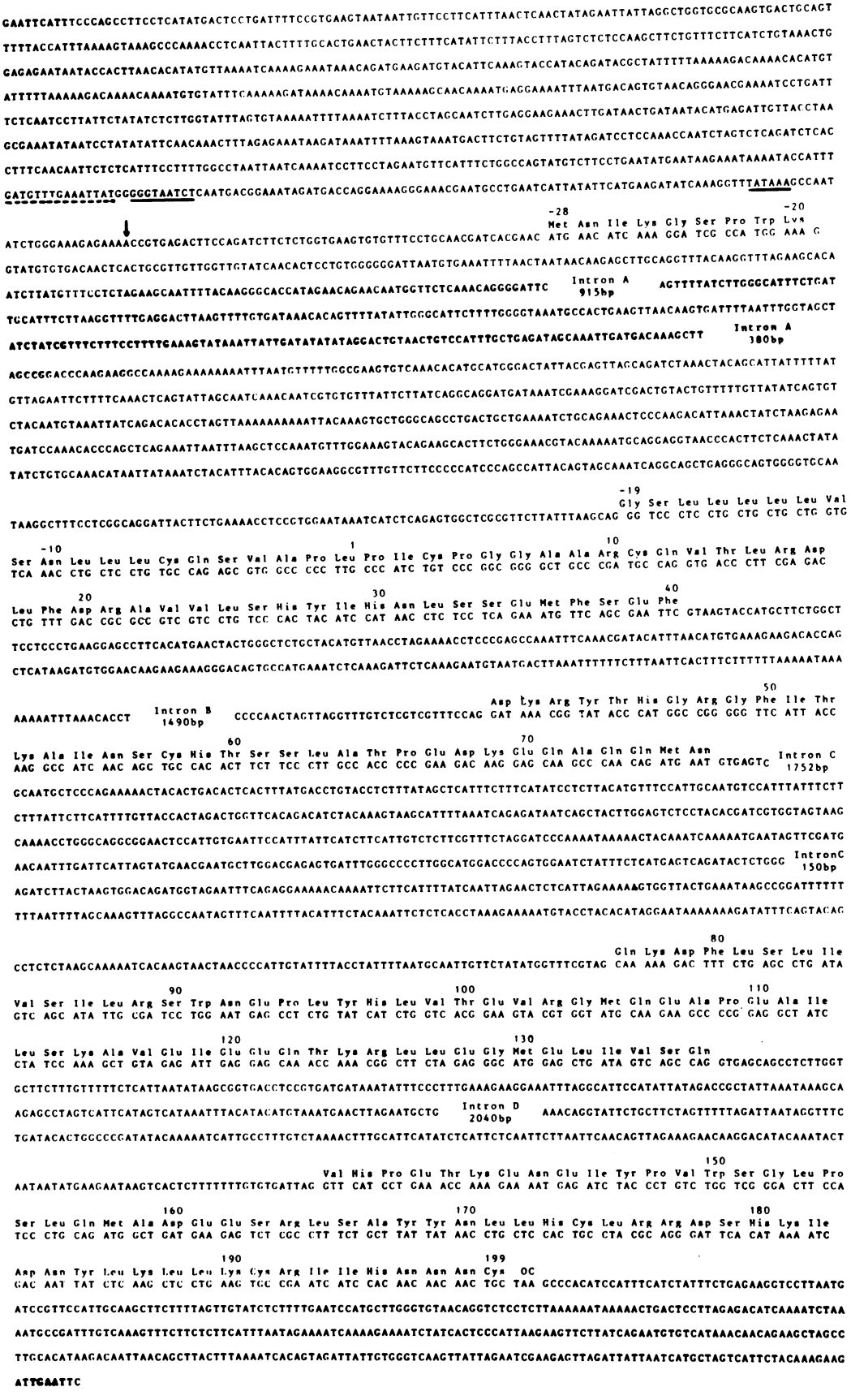

Fig. 5. Sequence of the hPRL gene. The experimentally determined origin of transcription is shown by a vertical arrow. The TATA and putative CAAT boxes are underlined. Sequences conserved in the PRL - GH - CS gene family are underlined with a dashed line. Positive numbers refer to mature prolactin amino acids, negative numbers to the signal peptide.

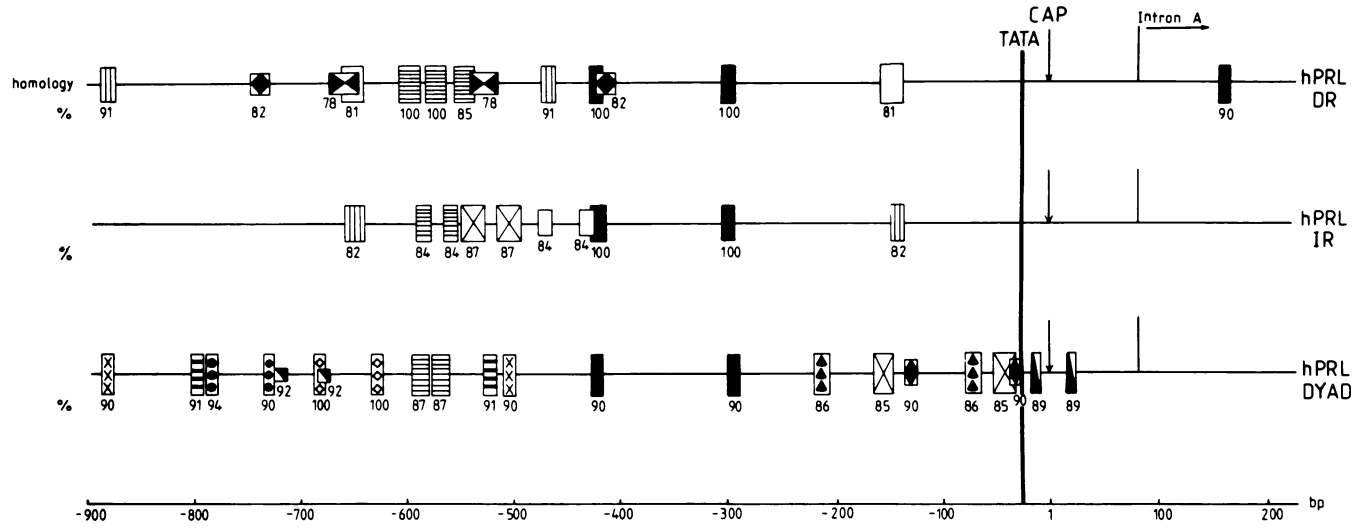

Fig. 6. Analysis of the 5′-flanking sequence of hPRL gene. Homology matrices were computer-drawn according to Pustell and Kafatos (1982). Direct repeats **(top line)**, inverted repeats **(middle line)** outlined here are >77% homologous. Dyad symmetries **(bottom line)** are >84% homologous. Boxes with different patterns correspond to different sets of repeats. The coordinates refer to distances in base pairs from the origin of transcription (CAP).

poly(A) addition site found by Cooke *et al.* (1981) on the phPRL cDNA.

The intron splice sites were unambiguously located following the GT-AG rule (Breathnach *et al.*, 1978). Intron A is the only one that interrupts an amino acid coding triplet (class I splice site; Sharp, 1980); introns B, C and D are located between triplets (class 0 splice sites; Sharp, 1980). Several potential splice sites can be found in the hPRL intron A. The most 5′ of these (90 bp downstream from exon 1) would allow an open reading frame for 32 amino acids including the nine codons of exon 1.

The 5′-flanking region of the hPRL gene comprises two conserved sequences involved in the promotion of eucaryotic gene transcription by RNA polymerase II (reviewed by Breathnach and Chambon, 1981). The so-called TATA box (underlined in Figure 5) is located 29 bp upstream of the origin of transcription. As far as the CAAT box is concerned, there is some controversy in the literature about its presence in the rPRL gene. Maurer *et al.* (1981) as well as Cooke and Baxter (1982) consider that rPRL lacks this sequence. On the other hand, Miller and Eberhardt (1983) describe a CAAT box 65 bp 5′ to the TATA box in the hGH, hCS, rGH and rPRL genes. The sequence AATAGATGA that they find in the rPRL gene is in the same location in hPRL presenting 33% homology to the CAAT consensus, GG$^{C}_{T}$CAATCT (Benoist *et al.*, 1980). However, there is a sequence (GGGTAATCT underlined in Figure 5) ~20 bp upstream from this putative CAAT box in the hPRL and rPRL genes, which is 78% homologous to the consensus.

Searches for direct and inverted repeats as well as dyad symmetries (including palindromes) in the 5′-flanking sequence of the hPRL gene were carried out according to Pustell and Kafatos (1982). The homology matrix plots gave intricate overlapping patterns for this very AT-rich region. The sequences with >77% homology found with these computer programs have been drawn on Figure 6. Direct and inverted repeats are especially frequent between 400 and 600 bp upstream of the origin of transcription. Three repeated regions are remarkable for their length and high degree of homology. One of them is 21 bp long and is repeated twice between coordinates −535 and −607; the second (10 bp) appears 300 and 420 bp upstream of the origin of transcription and the third repeat, which has only 80% homology but is 21 bp long, is located at −139 bp and at −666 bp. Dyad symmetries cluster mainly between coordinates −400 and −600 and around the TATA box. Among these, a 20-bp sequence located 28−8 bp upstream of the TATA box is 85% homologous to a sequence at −169 to 149 bp. The upstream part of the dyad is next to the 15-bp sequence conserved in all the genes of the GH - CS - PRL family (see below).

Determination of the origin of transcription of the hPRL gene

Total RNA extracted from a human pituitary tumor (prolactinoma) was used as a source of PRL mRNA. We used as a probe a 253-bp *Ball-Nco*I DNA fragment (Figure 4, bottom line) uniquely labelled at the 5′ end of the *Nco*I site located in exon 1. This piece of DNA, extending upstream of the TATA box, was hybridized to the prolactinoma total RNA. After digestion with S1 nuclease, the DNA fragment protected in the hybrid formed with the PRL mRNA was analysed by electrophoresis on sequencing gel (Figure 7). Its exact size was found to be 80 bp by comparison with the sequencing ladders of another known piece of DNA. This distance of 80 nucleotides 5′ to the *Nco*I site assigns the origin of transcription to an A residue located 57 bp upstream of the initial ATG of hPRL and 29 bp downstream of the TATA box.

Comparison of the hGH, hPRL and rPRL genes

We have compared the hPRL gene with other members of the GH - CS - PRL gene family. Within the same species, i.e., human, the study has been made only on the GH gene characterized by Seeburg (1982), because it has a very high nucleotide sequence homology to the other GH and CS genes. For an interspecies comparison, the only PRL gene known to date is that of the rat whose analysis has been published by Gubbins *et al.* (1980), Maurer *et al.* (1981) and Cooke and Baxter (1981).

To compare intron locations, the cDNA sequences for hGH, hPRL and rPRL were aligned and gaps were introduced to maximize homology, according to Cooke *et al.* (1980, 1981). Intron A separates the first and second nucleotides of a codon (class I splice site; Sharp, 1980) in each case. In rPRL

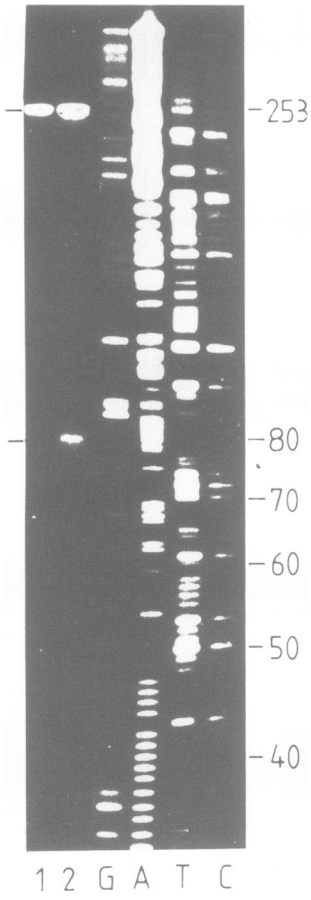

Fig. 7. S1 mapping of hPRL mRNA start site. **Lane 2**: S1-resistant fragment obtained from hybrid between total human prolactinoma RNAs and the 253-bp *BalI-NcoI* DNA fragment extending upstream from the 5′-labelled *NcoI* restriction site within exon I (Figure 4, bottom line). **Lane 1**: same as **lane 2** but no RNA added. **Lanes G, A, T, C**: sequencing ladders of a known fragment run on the same gel to provide size markers. Numbers to the right refer to fragment sizes in bp.

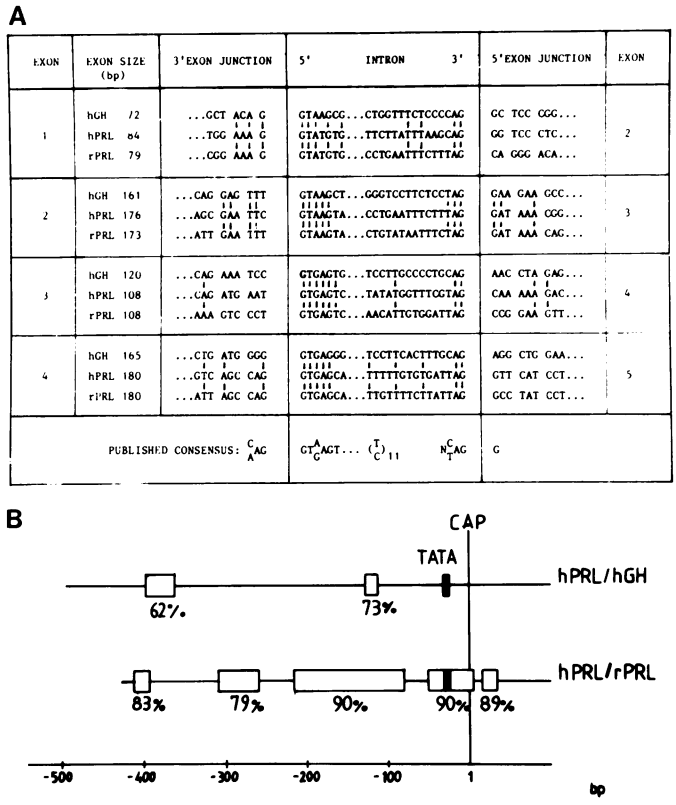

Fig. 8. Comparison of the hPRL, hGH and rPRL genes. **(A)** Sequence homologies of the exon-intron junctions. The positions of the exon-intron junctions are indicated by the long vertical lines, and each intron is flanked by its adjacent exon. The sequences of the last two or three complete in-frame codons at the 3′ exon junction are followed in each instance by the first seven nucleotides of the adjacent intron. The last 15 nucleotides of that intron and the first two or three complete in-frame codons of the next exon follow the dotted line which represents the intron internal sequences. Identical nucleotides in an identical position in the three genes are linked by short vertical lines. The consensus sequence of Mount (1982) for the splice junction is shown at the bottom. **(B)** Sequence homologies in the 5′-flanking sequences. Homology matrices were computer-drawn according to Pustell and Kafatos (1982). The sequences homologous to hPRL are drawn on the known hGH (top line, sequence from Seeburg, 1982) and rPRL (bottom line, sequence from Cooke and Baxter, 1982, origin of transcription from Maurer *et al.*, 1981) genes that have been aligned relative to their TATA boxes. The coordinates refer to distances in bp from the origin of transcription (CAP).

and hGH it is three and nine nucleotides, respectively, upstream of its hPRL location. Introns B and C interrupt the three sequences at the same place, between codons. Although intron D also creates a class 0 splice site (Sharp, 1980), it cuts the hGH coding sequence one triplet of nucleotides upstream of its location in the two PRL DNAs. Exons 2 to 5 have similar sizes in the three genes; exon 1 in human and rat PRL is 15 and 12 bp larger, respectively, than in the GH gene family.

A comparison of the exon-intron junctions of the hPRL with the rPRL and hGH genes is displayed in Figure 8A, together with the consensus sequences published by Mount (1982). A very good conservation of the 5′ donor splice site is found among the three genes, with a high (80%) homology to the consensus. Within the 3′ acceptor splice site, the sequence N$^{C}_{T}$AG/ of the consensus is present in all junctions. The remaining sequences are less conserved having a 70% average homology to the consensus.

As could be expected, there is very little sequence conservation between the hGH and hPRL introns, the latter being five times larger than the former as was already described in the rat genes. The known intron sequences of the two PRL genes also differ except for the following regions: an identical 9-bp stretch of DNA starts at 87, 71 and 81 bp from the 3′ end of exon 1 in intron A of hPRL, rPRL and hGH, respectively. This sequence, TGCAG/GTTT, is a potential alternative splice site; if it were used, exon 1 would be extended by an open reading frame of 23, 15 and only 4 amino acids for hPRL, rPRL and hGH, respectively. There is no evidence that any of these extra splice sites is used. Another conserved sequence is found on the 3′ side of intron A in the two PRL genes. The 13 base pairs present 80% homology and start at 84 and 57 bp upstream of the 5′ end of exon 2 in hPRL and rPRL, respectively.

The 5′-flanking sequences of hPRL, rPRL and hGH were compared, by computer drawing the homology matrices according to Pustell and Kafatos (1982). Very little conservation was found for the hPRL and hGH genes. Figure 8B (top line) shows that over an average coding sequence homology of 42%, two regions of higher conservation can be found. The 15-bp sequence GANNTTNNAAATTAT (underlined with a dashed line in Figure 5) centered at −122 bp is present in the same position relative to the origin of transcription in the

hGH, hCS, rGH, bovine GH (Woychick *et al.*, 1982), hPRL and rPRL genes. It is located just upstream of our putative CAAT box in the hPRL and rPRL genes. Another lower but significant homology (62%) is found for a 37-bp sequence starting at coordinate −400 in hPRL and −388 in the hGH gene.

The overall homology for coding sequences between hPRL and rPRL is 73% (Cooke *et al.*, 1981). In the 5′-flanking region we could identify five regions of 79–90% homology between the rat and human PRL (Figure 8B, bottom line). The largest one spans 136 bp, it has 90% homology and includes the 15-bp sequence conserved in all the genes of the family (see above). The two 20-bp sequences (90% homology) that form a dyad symmetry at −57 and −169 bp in the hPRL gene have two corresponding 12-bp sequences (83% homology) in the rPRL at the same relative positions although both dyad symmetries are not homologous to each other. In the two genes, those sequences surround the 15-bp region conserved in the GH - CS - PRL gene family.

Discussion

We have screened a human genomic library in phage λ with a cloned hPRL cDNA and isolated a recombinant phage containing the whole prolactin gene. It spans ~10 kb and is interrupted by four introns. We established a restriction map of the hPRL gene and sequenced ~5000 bp including 900 bp of 5′-flanking region, 350 bp of 3′-flanking region and all the exon-intron junctions. Comparison with the hGH gene confirms the structural and evolutionary relationship between GH and PRL. Besides the TATA and CAAT boxes, potential regulatory sequences have been identified by homology study with the rPRL and hGH 5′-flanking regions.

When we looked at the PRL gene organization (Figure 1) by hPRL cDNA hybridization to *Eco*RI-digested human genomic DNA, we found fragments of different sizes depending on the source of DNA. We do not favor an explanation involving point mutations at an *Eco*RI restriction site because the pattern was identical for white blood cells DNA of 20 different individuals. A more likely interpretation is a modification of the C residues around the *Eco*RI site. Methylation of Cs in the CGAATTCG sequence would not be surprising since most Cs in CG dinucleotides are methylated (Ehrlich and Wang, 1981). If this is the case, it is different from the usual *Eco*RI site methylation, CGA5meATTCG, which prevents digestion (Greene *et al.*, 1975). If this C modification inhibits hydrolysis at the *Eco*RI site, it would explain the partial digestion pattern observed in white blood cells DNA. DNA from placenta, HeLa cells and human hepatoma cell line can be cleaved at this site (Figure 1 and Edman *et al.*, 1980). Owerbach *et al.* (1981) present a similar pattern with DNA from somatic cell hybrids of human and mouse cells although they overestimate the sizes of the two larger fragments, 4.8 and 4.0 kb instead of 4.1 and 3.4 kb, the shorter 0.7-kb fragment having most probably run out of their gel.

The frequency at which we found human prolactin recombinant phages in this library is quite low ($1/1.25 \times 10^6$) compared with what some of us (Bell *et al.*, 1980) obtained with another unique gene (e.g., insulin, $6/1.25 \times 10^6$). Taken together with the analysis of human DNA digest using the hPRL cDNA probe, this finding seems to indicate that in humans the PRL gene is present as a single copy per haploid genome, as it is in rat. Another argument is the fact that the sequence we determined for the five exons is identical to the previously published hPRL cDNA sequence (Cooke *et al.*, 1981).

Control of gene expression can be achieved through the interaction of a regulatory protein with a specific DNA sequence. The experimental evidence for such a mechanism has to show the *in vitro* binding of the regulatory protein to a DNA sequence which is necessary *in vivo* for the gene control to take place. Such evidence has accumulated over the past few years about the induction of MMTV DNA transcription by glucocorticoids through binding of the hormone-receptor complex to two DNA sequences located 150 and 100 bp upstream of the mRNA start site (e.g., Scheideret *et al.*, 1983; Hynes *et al.*, 1983). If such mechanisms are involved in the complex hormonal control of PRL gene expression, it should be possible to define important DNA sequences by comparing the 5′-flanking region of the same gene in different species where a similar regulation is observed. By computer plotting the homology matrix between rat and human PRL we could identify five highly conserved DNA sequences (79–90% homology *versus* 73% for the coding regions). We also compared hPRL with a gene regulated in the opposite way, hGH, with which it has 42% homology for the coding regions. These PRL highly conserved sequences (except for 15 bp, see below) are not found in hGH nor rGH, suggesting that they might be involved in a PRL specific gene regulation.

All the genes of the PRL - GH - CS family comprise a high homology 15-bp segment centered at about −122 bp. This region could be necessary for transcription to take place and may not be involved in regulation. Breathnach and Chambon (1981) mention such sequences located >100 bp upstream from the mRNA start site, that do not seem to be conserved in different genes but cannot be deleted without strongly inhibiting transcription. However, if a regulatory protein does bind to this 15-bp segment in both GH and PRL genes, the specific activation or inhibition of transcription observed for each of these genes might proceed through a second protein-DNA interaction at a different location.

A dyad symmetry is found in the same relative position in both hPRL and rPRL genes (coordinates −57 and −169). Although the human and rat dyads are not homologous to each other, these sequences are arranged in an identical pattern in both genes: the distance between them corresponds to about one turn of DNA around a nucleosome. Several authors have found other sequence structures that might be involved in the rPRL gene regulation (Maurer *et al.*, 1981; Cooke and Baxter, 1982). We believe that only systematic comparison of several genes representative of the PRL - GH - CS family could provide such information. Moreover, caution should be used in interpreting computer-drawn homology matrices, especially when analysing AT-rich sequences such as the 5′-flanking region of the hPRL gene. All the sequence homologies presented in this paper have been selected for their length and high conservation. *In vitro* studies of chromatin proteins binding to hPRL 5′-flanking DNA as well as *in vivo* assay of various deletion mutants are currently under way to find out the biological role, if any, of all these conserved sequences.

Materials and methods

Screening of the human genomic library

The partial *Hae*III-*Alu*I library of human fetal liver DNA in bacteriophage lambda Ch4A constructed by Lawn *et al.* (1978) was kindly provided by Dr.

T. Maniatis. 1.25 x 10^6 phages (24 000/15 cm dish) were screened as described in Benton and Davis (1977) except that 0.8% agarose was used in the top agar solution to prevent adherence of the agar to the nitrocellulose filter. Duplicate fiters were prepared from each plate. The filters were hybridized with nick-translated (Maniatis *et al.*, 1975; 2–4 x 10^8 c.p.m./μg) phPRL cDNA insert (Cooke *et al.*, 1981) according to Wahl *et al.* (1979). The radioactive concentration of the probe was 2.4 x 10^5 c.p.m./ml. Filters were washed for 1 h at 50°C in 0.1 x SSC, 0.1% SDS. Autoradiography for the initial screen was for 3 days with a single intensifying screen and pre-flashed X-ray film. The phages containing prolactin sequences were identified and plaque purified. Purified phages were propagated in *Escherichia coli* DP50SupF liquid culture and DNA prepared as described by Blattner *et al.* (1977) with an infection multiplicity of 0.1.

Recombinant DNA safety producedures

The screening and propagation of recombinant bacteriophage were conducted in accordance with the NIH Guidelines for Recombinant DNA Research. The experiments were performed in a P2 facility using the EK2 host/vector system lambda Charon 4A/*E. coli* DP50SupF.

Heteroduplex analyses

Heteroduplex structures were formed by mixing equimolar amounts of native and recombinant λ Charon 4A with a 10-fold molar excess of phPRL cDNA linearized at its unique *Bam*HI site. The DNA was denatured in 0.1 M NaOH/10 mM EDTA for 10 min at 20°C. The medium was adjusted to pH 8.4 with Tris HCl (0.1 M final concentration) and made 50% in deionized formamide. The mixture was incubated for 6–18 h at 25°C so as to obtain ~50% of double-stranded DNA. Heteroduplexes between λ Charon 4A, λ-hPRL gene and phPRL cDNA as well as between native and recombinant λ, and between λ-hPRL gene and phPRL cDNA were photographed and measurements made on the negatives using a H.P. System.

Knowing the orientation of the cDNA in the plasmid (5′→3′ clockwise), the orientation of the gene in the recombinant phage could be determined by measuring the arms of the unhybridized pBR322 corresponding to the *Pst*I-*Bam*HI short and long fragments. The conversion of length to base pairs was done by using as standard the middle piece and λ arms of the heteroduplex for single- and double-stranded DNA, respectively.

Restriction endonuclease digestion, agarose gel electrophoresis, blotting and hybridization

Restriction endonuclease obtained from New England Biolabs (Beverly, MA), Bethesda Research Labs (Gaithersburg, MD) and Boehringer (Mannheim, FRG) were used according to the manufacturers' specifications. Agarose gel electrophoresis was as described (Southern, 1979). DNA was transferred to nitrocellulose filters by the procedure of Southern (1975). Hybridization was in 10% dextran sulfate (Wahl *et al.*, 1979). Following hybridization, filters were washed in 0.1 x SSC, 0.1% SDS at 50°C for 1 h and then autoradiographed.

Determination and analysis of DNA sequence

DNA was sequenced using 5′[^{32}P]kinase-treated fragments according to the method of Maxam and Gilbert (1980).

Sequence homology matrices were computer-drawn with the program established by Pustell and Kafatos (1982). In the search for direct and inverted repeats, as well as dyad symmetries, the range was raised from 3 to 6 with a minimum value plotted of 70–90%. In the comparison of 5′-flanking regions of hPRL with hGH and rPRL, the range was 3 and the minimum value plotted was 75–80%.

RNA extraction and nuclease S1 mapping

A human prolactinoma obtained from surgery was immediately frozen in liquid nitrogen. Total RNA was extracted by the guanidinium thiocyanate procedure described by Chirgwin *et al.* (1979). The RNA pellet obtained after centrifugation through a CsCl solution was further extracted with phenol before final ethanol precipitation. Presence of PRL-specific mRNA was confirmed by hybridization to phPRL cDNA (data not shown).

S1 nuclease mapping of the mRNA 5′ end was done according to Berk and Sharp (1977) with the following modifications. The most 5′ *Eco*RI genomic subclone was cut in exon 1 with *Nco*I, treated with bacterial alkaline phosphatase and 5′ end-labelled with [γ-^{32}P]ATP using T4 polynucleotide kinase. After recutting 5′ with *Bal*I, the 253-bp fragment extending from coordinates −173 to +80 bp was uniquely labelled on its 3′ side. About 75 000 Cerenkov c.p.m. of this *Bal*I-*Nco*I* probe were mixed with 10 ng of total prolactinoma RNA, dried under vacuum, solubilized in 10 μl 80% twice-recrystallised formamide, 40 mM PIPES pH 6.4, 400 mM NaCl, 1 mM EDTA, and transferred to a capillary tube. After denaturation for 15 min at 80°C, they were allowed to hybridize for 5 h at 57°C. After addition to 9 volumes of S1 buffer (30 mM NaAc pH 4.75, 300 mM NaCl, 3 mM $ZnCl_2$, 20 μg/ml denatured salmon sperm DNA), the hybrids were digested with S1 nuclease (PL Biochemicals) at a final concentration of 100 U/ml for 30 min at 30°C. S1-resistant material was directly ethanol precipitated with *E. coli* tRNA as a carrier, and analysed by electrophoresis on an 8% sequencing gel together with a 334-bp known fragment submitted to the Maxam and Gilbert chemical reaction. The gel was exposed to a Kodak RP royal XOmat film with an intensifying screen for 1 month.

Acknowledgements

We are grateful to T. Maniatis for providing the human library and to F. Kafatos for the homology search computer programs. This work was supported by a grant 'Actions concertées génie génétique 81/86' from the 'Services de Programmation de la Politique Scientifique' of the Belgian Government. G.I.B. was partially supported by a Research and Development Award from the American Diabetes Association.

References

Barta,A., Richards,R., Baxter,J.D. and Shine,J. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 4867-4871.
Bell,G.I., Pictet,R. and Rutter,W.J. (1980) *Nucleic Acids Res.*, **8**, 4091-4109.
Benoist,C., O'Hare,K., Breathnach,R. and Chambon,P. (1980) *Nucleic Acids Res.*, **8**, 127-142.
Benton,W.D. and Davis,R.W. (1977) *Science (Wash.)*, **196**, 180-182.
Berk,A.J. and Sharp,P.A. (1977) *Cell*, **12**, 721-732.
Blattner,F.R., Williams,B.G., Blechl,A.E., Deniston-Thompson,K., Faber, H.E., Furlong,L.A., Grunwald,D.J., Kiefer,D.O., Moore,D.D., Schumm, J.W., Sheldon,E.L. and Smithies,O. (1977) *Science (Wash.)*, **196**, 161-169.
Breathnach,R., Benoist,C., O'Hare,K., Gannon,F. and Chambon,P. (1978) *Proc. Natl. Acad. Sci. USA*, **75**, 4853-4857.
Breathnach,R. and Chambon,P. (1981) *Annu. Rev. Biochem.*, **50**, 349-383.
Chien,Y.H. and Thompson,E.B. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 4583-4587.
Chirgwin,J.M., Przybyla,A.E., MacDonald,R.J. and Rutter,W.J. (1979) *Biochemistry (Wash.)*, **18**, 5294-5299.
Cooke,N.E. and Baxter,J.D. (1982) *Nature*, **297**, 603-606.
Cooke,N.E., Coit,D., Weiner,R.I., Baxter,J.D. and Martial,J.A. (1980) *J. Biol. Chem.*, **255**, 6502-6510.
Cooke,N.E., Coit,D., Shine,J., Baxter,J.D. and Martial,J.A. (1981) *J. Biol. Chem.*, **256**, 4007-4016.
Dobner,P.R., Kawasaki,E.S., Yu,L.Y. and Bancroft,F.C. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 2230-2234.
Edman,J.C., Gray,P., Valenzuela,P., Rall,L.B. and Rutter,W.J. (1980) *Nature*, **286**, 535-538.
Ehrlich,M. and Wang,R.Y.H. (1981) *Science (Wash.)*, **212**, 1350-1357.
Evans,R.M., Birnberg,N.C. and Rosenfeld,M.G. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 7659-7663.
Fiddes,J.C., Seeburg,P.H., De Noto,F.M., Hallewell,R.A., Baxter,J.D. and Goodman,H.M. (1978) *Proc. Natl. Acad. Sci. USA*, **76**, 4294-4298.
Greene,P.J., Poonian,M.S., Nussbaum,A.L., Tobias,L., Garfin,D.E., Boyer,H.W. and Goodman,H.M. (1975) *J. Mol. Biol.*, **99**, 237-261.
Gourdji,D., Tougard,C. and Tixier-Vidal,A. (1982) in Ganong,W.F. and Martini,L. (eds.), *Frontiers in Neuroendocrinology*, Vol. **7**, Raven Press, NY, pp. 317-357.
Gubbins,E.J., Maurer,R.A., Lagrimi,M., Erwin,C.R. and Donelson,J.E. (1980) *J. Biol. Chem.*, **255**, 8655-8662.
Hynes,N.E., Van Doyen,A.J.J., Kennedy,N., Herrlich,P., Ponta,M. and Groner,B. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 3637-3641.
Laverriere,J.N., Morin,A., Tixier-Vidal,A., Truong,A.T., Gourdji,D. and Martial,J.A. (1983) *EMBO J.*, **2**, 1493-1499.
Lawn,R.M., Fritsch,E.F., Parker,R.C., Blake,G. and Maniatis,T. (1978) *Cell*, **15**, 1157-1174.
Maniatis,T., Jeffrey,A. and Kleid,D.G. (1975) *Proc. Natl. Acad. Sci. USA*, **72**, 1184-1188.
Maurer,R.A. (1982a) *Endocrinology*, **110**, 1507-1524.
Maurer,R.A. (1982b) *J. Biol. Chem.*, **257**, 2133-2136.
Maurer,R.A., Erwin,C.R. and Donelson,J.E. (1981) *J. Biol. Chem.*, **256**, 10524-10528.
Maxam,A.M. and Gilbert,W. (1980) *Methods Enzymol.*, **65**, 39-43.
Miller,W.L. and Eberhardt,N.L. (1983) *Endocrine Rev.*, **4**, 97-130.
Moore,D.D., Walker,M.D., Diamond,D.J., Conkling,M.A. and Goodman, H.M. (1982) *Recent Prog. Horm. Res.*, **38**, 197-225.
Mount,S.M. (1982) *Nucleic Acids Res.*, **10**, 459-472.
Murdoch,G.H., Potter,E., Nicolaisen,A.K., Evans,R.M. and Rosenfeld, M.G. (1982) *Nature*, **300**, 192-194.
Niall,H.D., Hogan,M.L., Sauer,R., Rosenblum,I.Y. and Greenwood,F.C. (1971) *Proc. Natl. Acad. Sci. USA*, **68**, 866-869.

Owerbach,D., Rutter,W.J., Martial,J.A., Baxter,J.D. and Shows,T.B. (1980) *Science (Wash.),* **209**, 289-291.
Owerbach,D., Rutter,W.J., Cooke,N.E., Martial,J.A. and Shows,T.B. (1981) *Science (Wash.),* **212**, 815-816.
Page,G.S., Smith,S. and Goodman,H.M. (1981) *Nucleic Acids Res.,* **9**, 2087-2104.
Pustell,J. and Kafatos,F.C. (1982) *Nucleic Acids Res.,* **10**, 4765-4782.
Scheideret,C., Geisse,S., Westphal,H.M. and Beato,M. (1983) *Nature,* **304**, 749-752.
Seeburg,P.H. (1982) *DNA,* **1**, 239-249.
Sharp,P.A. (1980) *Cell,* **23**, 643-646.
Southern,E.M. (1975) *J. Mol. Biol.,* **98**, 503-517.
Southern,E. (1979) *Methods Enzymol.,* **68**, 152-176.
Spindler,S.R., Mellon,S.H. and Baxter,J.D. (1982) *J. Biol. Chem.,* **257**, 11627-11638.
Tashjian,A.H,.Jr. (1979) *Methods Enzymol.,* **58**, 527-534.
Wahl,G.M., Stern,M. and Stark,G.R. (1979) *Proc. Natl. Acad. Sci. USA,* **76**, 3683-3687.
Wegnez,M., Shachter,B.C., Baxter,J.D. and Martial,J.A. (1982) *DNA,* **1**, 145-153.
Woychick,R.P., Camper,S.A., Lyons,R.H., Horowitz,S., Goodwin,E.C. and Rottman,F.M. (1982) *Nucleic Acids Res.,* **10**, 7197-7210.

Received on 14 November 1983