Nucleic Acids Research, Vol. 19, No. 19 5237–5245

Cloning and sequencing of PYBP, a pyrimidine-rich specific single strand DNA-binding protein

Franck Brunel, Pedro M.Alzari[1], Pascual Ferrara[2] and Mario M.Zakin*
Laboratoire d'Expression des Gènes Eucaryotes and [1]Unité d'Immunologie Structurale, Institut Pasteur, 28 rue du Dr Roux, 75724 Paris Cedex 15 and [2]Sanofi Elf-Biorecherches, Labège-Innopole, Voie no.1, BP 137, 31328 Labège Cedex, France

Received July 18, 1991; Revised and Accepted September 17, 1991 EMBL accession nos X60789 and X60790

ABSTRACT

In the human transferrin gene promoter, PRI and DRI are positive *cis*-acting elements interacting respectively with two families of proteins, Tf-LF1 and Tf-LF2. In this paper, we report the purification from rat liver nuclei, of one of these factors, PYBP, as well as the cloning and the sequencing of its cDNA. PYBP is a DNA-binding protein, purified as a 58 kDa doublet which binds only to single strand pyrimidine-rich DNA present for example in PRI and DRI. The protein binds also to a similar polypyrimidine tract present in one of the two strands of a DNA regulatory element of the rat tyrosine aminotransferase gene enhancer. PYBP gene is transcribed ubiquitously as a roughly 2,8 kb RNA which is likely to be subject to an alternative splicing. PYBP is highly homologous to a mouse nuclear protein, as well as to PTB, its human version, which interacts specifically with the pyrimidine tracts of introns. Primary structure information and predicted secondary structure elements of the protein indicate that PYBP contains four sequence repeats. Each of these repeats appears to exhibit the typical RNA recognition motif found in several proteins interacting with RNA or single strand DNA. Finally several hypotheses concerning the biological function of PYBP are presented.

INTRODUCTION

DNA- and RNA-protein interactions are some of the molecular mechanisms used by the cell to achieve the physiological process necessary for its survival. Indeed, these interactions are essential for DNA replication and transcription, and for mRNA splicing. Recent observations allowed the identification of DNA and RNA sequences necessary for normal and abnormal development of cellular functions. These sequences are generally targets for proteins which interacts specifically with them; the purification of these proteins and the cloning of their cDNAs are the subject of intensive activity (reviewed in 1 and 2 and references therein).

In our laboratory, we are studying the regulation of the human transferrin gene (hTf[1]) expression. We have previously determined that several positive and negative *cis*-acting DNA elements are involved in the regulation of transferrin gene expression (3, 4, 5). Among these elements, PRI and DRI lie respectively in the proximal promoter and in a distal regulatory region. PRI and DRI contain in their core the same decanucleotide ($_{5'}$TCTTTGACCT$_{3'}$). Despite this, two families of transcription factors interact with the PRI and DRI elements (6). In this paper, we present the purification of one of these proteins, which is refered hereafter as PYBP (for polypyrimidine binding protein). The cloning and the sequencing of its cDNA is also described, and electrophoresis mobility shift assays (EMSA) allowed us to find surprisingly that PYBP binds only to the strand of the PRI and DRI elements which contains stretches of pyrimidine (C or T nucleotides). This protein does not bind to the complementary purine-rich DNA strand, or to the double strand DNA. In this paper, several hypotheses are analysed in order to understand the biological function of PYBP.

MATERIALS AND METHODS

DNA probes

Synthetic oligonucleotides were purified on a 20% polyacrylamide-7M urea gel (7). Their concentration were determined by UV absorbance. Complementary oligonucleotides were annealed and 5'-end labelled with T4 polynucleotide kinase and ATP [γP^{32}]. In the experiments involving double strand DNA shown in Figure 8B, one strand was 5'-end labelled and annealed with a two-fold molar excess of the complementary strand; asymetrical labelled double strand was then purified from residual labelled single strand on acrylamide gel. EMSA was performed according to Garner and Revzin (8) as modified by Ochoa *et al* (6).

The oligonucleotides used in this study are listed in Table 1.

Purification of PYBP

Liver nuclear proteins were prepared from 80 rats according to the method described by Gorski *et al* (9), modified by Brunel *et al* (3). Proteins (1.4 g) were loaded on to a DEAE-Sephacel column (volume: 30 ml; diameter: 1.5 cm; flow rate: 20 ml/h)

* To whom correspondence should be addressed

in buffer Z (Buffer Z is: 20 mM Hepes-K^+ pH 7.5; 0.2 mM EDTA; 12.5 mM $MgCl_2$; 0.5 mM DTT; 0.5 mM PMSF; 10% glycerol; 0.1% Nonidet P-40) containing 300 mM KCl. PYBP was collected in the flow-through, diluted to 200 mM KCl with buffer Z without KCl and loaded on a heparin-Sepharose column (volume: 100 ml; diameter: 2 cm; flow-rate: 20 ml/h; 5 ml fractions were collected). The column was washed with buffer Z containing 200 mM KCl (five column volumes, flow-rate: 20 ml/h, 5 ml fractions were collected) and eluted with a linear KCl gradient from 200 mM KCl to 700 mM KCl (five column volumes, flow-rate 40 ml/h, 5 ml fractions were collected). A 2 M KCl step was then applied (five column volumes, flow-rate: 40 ml/h). Fractions containing PYBP were pooled, dialyzed against buffer Y (Buffer Y is: 20 mM Tris-HCl pH 7.5; 0.2 mM EDTA; 12.5 mM $MgCl_2$; 0.5 mM DTT; 0.5 mM PMSF; 10% glycerol; 0.1% Nonidet P-40) containing 50 mM NaCl and loaded on a S-Sepharose Fast Flow (Pharmacia) column (volume: 2 ml; diameter: 0.5 cm, flow rate: 15 ml/h; 1 ml fractions were collected). The column was washed with five column volumes of buffer Y containing 50 mM NaCl. Proteins were then eluted with a 50 mM NaCl to 250 mM NaCl linear gradient (seven column volumes). Fractions containing PYBP were pooled, dialyzed against buffer Y containing 50mM NaCl and loaded on a Q-Sepharose Fast-Flow column (volume 1 ml; diameter 0.5 cm; flow-rate 15 ml/h; 4 ml fractions were collected). Column was washed with buffer Y containing 50mM NaCl and eluted stepwise with the same buffer containing increasing concentrations of NaCl. PYBP was collected in the flow-through of this column.

50 μg of protein from the Q-Sepharose fractions were TCA-precipitated and fractionated on a 12.5% SDS discontinuous-polyacrylamide gel (10) along with pre-stained molecular weight markers (Biorad). After suitable separation, proteins were visualized by ice-cold KCl shadowing (11) and eluted from the gel slices in buffer X (30 mM Tris pH 7.4; 50 mM NaCl; 12,5 mM $MgCl_2$; 0.1% Nonidet P-40; 100 μg/ml BSA; 8 M urea) at 37°C during 18 h. Renaturation was carried out by exhaustive dialysis against X buffer without urea.

Micro-sequencing of PYBP

9 ml prepurified Q-Sepharose fractions (4 ml of fraction 12, 4 ml of fraction 13 and 1 ml of fraction 14) were dialyzed against 1 l of 20 mM Tris-HCl pH 7.4, 50 mM NaCl, for 24 h. at 4°C. Then the sample was divided in three fractions and each was loaded onto a reverse phase HPLC C4 column (0.2×10 cm, Brownlee) and eluted with a linear gradient from 10% to 80% acetonitrile, 0.1% trifluoroacetic acid over 15 min., at a flow rate of 0.3 ml/min. Elution was monitored at 216 nm and eluting peaks were collected manually. The fractions containing the major peak from each run, eluting at 15 min., were pooled and concentrated under vacuum on a Speedvac centrifuge down to 30 μl; 3 μl were analyzed on a 12.5% SDS-PAGE using a Phast-Gel System (Pharmacia) and the rest ($\approx$ 75 μg) was adjusted at pH 8.8 with 15 μl of 0.2 M ammonium carbonate and digested with 6μg of trypsin (TPCK-treated trypsin, Cooper Biochemical) for 24 h. at room temperature. At the end of the incubation, the resulting peptides were separated on a reverse phase C18 column (0.2×25 cm, Chromegabond) with a 120 min 1 to 60% linear gradient of acetonitrile, 0.1% trifluoroacetic acid, at a flow rate of 0.23 ml/min. Elution was monitored at 216 nm and eluting peptides were collected manually. Automated Edman degradation of selected peptides was performed on a Model 470A gas-phase sequencer with a on-line 120A PTH analyser (Applied Biosystems).

Production of an unambiguous probe using polymerase chain reaction

Two degenerated primers coding for the N- and C-terminal moieties of the T54 peptide were synthesized (all the degeneracy of the genetic code was taken into account). Their sequences were: ON1: $_{5'}$ATGGCNYTNATZCARATGGG$_{3'}$; ON2: $_{5'}$YTTRTGRTGRTTYTCNCC$_{3'}$ (the abbreviations for the codon degeneracy are: $Y=\frac{C}{T}$; $R=\frac{A}{G}$; Z= A or C or T; N=A, C, T or G). 35 pmoles of ON1 and ON2 oligonucleotide primers were mixed with 10^7 λgt10 recombinant bacteriophages in a total volume of 100 μl of a buffer containing 200 μM each dNTP, 2 mM $MgCl_2$, 0.1% gelatin, 50 mM KCl and 20 mM Tris-HCl pH 7.4. Two rounds of successive cycles were performed: 5 soft cycles (30'' at 95°C, 1' at 37°C, 1' at 72°C) followed by 35 more stringent cycles (30'' at 95°C, 1' at 42°C, 1' at 72°C). The expected amplified 90 bp DNA fragment was phosphorylated, purified on a low melting point 2% agarose gel, flush-ended with the Klenow fragment and subcloned in the SmaI site of the pBluescript phagemid (Stratragene). Nucleotide sequence was performed on both strands of single strand template by the chain termination inhibition method (12) with the modified T7 DNA polymerase (Sequenase from UBS).

Cloning and sequencing of PYBP

A cDNA library was prepared from size selected mRNAs of Fao hepatoma cells treated with cycloheximide (a generous gift of S. Cereghini and M. Blumenfeld). This cDNA library was screened with the random-primed labelled T54 cDNA fragment (specific activity 10^9 cpm/mg) at high stringency under standard procedures (7). Hybridization was performed in 50% formamide, 6× SSPE (1× SSPE: 180 mM NaCl, 10 mM NaH_2PO_4 pH 7; 1mM EDTA), 0.1% SDS, 1× Denhardt's solution (0,02% BSA; 0,02% polyvinyl pyrrolidone; 0,02% Ficoll), 100 μg/ml salmon sperm DNA ; the specific activity of the probe was 10^6 cpm/ml. Washes were as follows: 2×10' in 2× SSC (1× SSC: 150 mM NaCl; 15 mM Na Citrate pH 7) at room temperature, 1h in 1× SSC at 65°C and 1h in 0.2 SSC at 65°C. Inserts of positive clones were PCR-amplified with primers flanking the cloning site of λgt10 and subcloned in the pBluescript phagemid. Nucleotide sequence were performed on both strands of single strand template of nested deletions (cyclone kit, IBI) by the chain termination inhibition method (Sanger *et al*, 1978) using the modified T7 DNA polymerase (Sequenase, USB). Ambiguous sequences were resolved with the dITP labelling mix (USB). A cloning artefact leading to a frame-shift in the λ22 cDNA 5' end was resolved by PCR amplification of the 5' end of mRNA using primers derived from λ22 cDNA sequence, subcloning and sequencing of the amplified product (data kindly provided by P. Jansen-Dürr).

Northern blot analysis

A membrane containing transferred polyA^+ RNAs was kindly provided by S. Cereghini (13) Hybridization was performed with a 900 bp random-primed labelled EcoRI/XbaI fragment (nucleotide 1029 to nucleotide 1930) in 50% formamide, 5× SSPE, 5× Denhardt's solution for 24 h at 42°C. Membrane was washed at 60°C in 0.1×SSC, 0.25%SDS.

Antibody preparations

A peptide extending from H_{511} to R_{522} amino acid residues of the deduced protein sequence of the λ22 cDNA clone was chemically synthesized and linked to the keyhole limpet hemocyanin (KLH) using the *m*-maleimidobenzoyl-N-hydroxysuccinimide ester method (14). 200 μg of peptide-KLH conjugates were injected into rabbits (15).

Computer analysis of nucleotide and aminoacid sequences

Searches in the EMBL nucleotide database were performed via electronic mail with the FASTA program (16). Secondary structure predictions (17), dot search plots (18) and pairwise alignments (19) were performed with computer software distributed by the Protein Identification Resource (PIR) of the National Research Foundation, Washington (20). Similarity scores for binary comparisons were calculated with the Mutation Data Matrix (21). All pairwise comparisons were assessed for significance by comparing the similarity score of the real comparison with the mean value determined for pairs of randomized sequences of the same length and composition. If the real alignment score is more than 3 standard deviations (σ) above the mean, the odds of obtaining it by chance are less than one in a thousand, assuming a normal distribution of randomized scores.

RESULTS

Purification of a rat liver protein interacting with the DRI element of the human transferrin gene promoter

DNA-binding activity was monitored through the purification with EMSA using as a probe a double strand oligonucleotide which bears the DRI sequence (Table 1). In the presence of crude rat nuclear liver proteins, such an assay allows the detection of several DNA-protein complexes, and the proteins involved in these complexes were referred to under the generic name of Tf-LF2 (3). We fractionated these liver nuclei proteins by heparin-Sepharose chromatography. The proteins were loaded at 200 mM KCl and the resin washed with the same salt concentration. Proteins were then released by a linear 200 mM KCl to 700 mM KCl salt gradient. Figure 1 shows the DNA-binding activity analysis with the DRI oligonucleotide of all the relevant chromatographic fractions . In the wash fractions one could detect a broad activity peak which reaches a maximum at fraction 87. The material eluted in these fractions leads to a complex with the DRI oligonucleotide which has the same electrophoretic mobility as one of the complexes detected with the crude nuclear extract (compare lane RNLE and lane F87 in Figure 1). The linear salt gradient releases at least two group of proteins (fractions 192 to 227 and fractions 232 to 262). Cross-competition experiments (data not shown) allowed us to infer that the protein eluted at 500 mM (fractions 232 to 262) is LF-A1. This transcription factor interacts with several regulatory regions of hepatic genes (22). Our finding was supported by the report of Rangan and Das (23)and confirm our previous papers that a family of liver transcripiton factors shares DNA-binding specificities and can interact with various *cis*-acting DNA elements (6)

In light of these results, it was not possible to keep the generic name of Tf-LF2 to designate proteins that interact with the DRI oligonucleotide but present different chromatographic properties. We then named PYBP (Pyrimidine Binding Protein; see below) the protein eluted in the wash of the heparin-Sepharose. We purified PYBP to near homogeneity through two subsequent ion-exchange chromatographies (see Materials and Methods). A total of 200 μg of PYBP was purified from 1.4 g of crude rat liver nuclei proteins. Figure 2A shows a silver-stained SDS-PAGE of the proteins of the purified fraction from the Q-Sepharose column.

TABLE 1: Nucleotidic sequence of synthetic oligonucleeotides used in this study

OLIGONUCLEOTIDE	SEQUENCE
DRI sense	5'TGCTGAGTCTG**TCTTTGACCT**TGAGCCCAGCT3'
DRI antisense	5'AGCAAGCTGGGCTCAAGGTCAAAGACAGACTC3'
PRI sense	5'ACAAACACGGGAGGTCAAAGATTGCGCCCAG3'
PRI antisens	5'CTGGGCGCAA**TCTTTGACCT**CCCGTGTTTGT3'
TAT sense	5'GATCTCCTGCTGC**TCTTTGATCT**GTATACCTG3'
TAT antisense	5'GATCCAGGTATACAGATCAAAGAGCAGCAGGA3'
DRI sense mut1	5'TGCTGAGTCTGGAGGAGACCTTGAGCCCAGCT3'
DRI sense mut2	5'TGCTGAGTATGTAAGGGACAGTGAGACGAGCT3'

The nucleotide sequence of the synthetic oligonucleotides used is shown. The $_{5'}$TCTTTGA$\frac{C}{T}$CT$_{3'}$ core common to the DRI, PRI and TAT *cis*-elements is shown in bold cases. Mutated sequences are underlined. The polarity of the strands of the oligonucleotides is indicated with the following conventions: the antisense strand sequence is colinear to the transcribed strand soquence, so the sense strand displays a sequence colinear to the genetic message sequence.

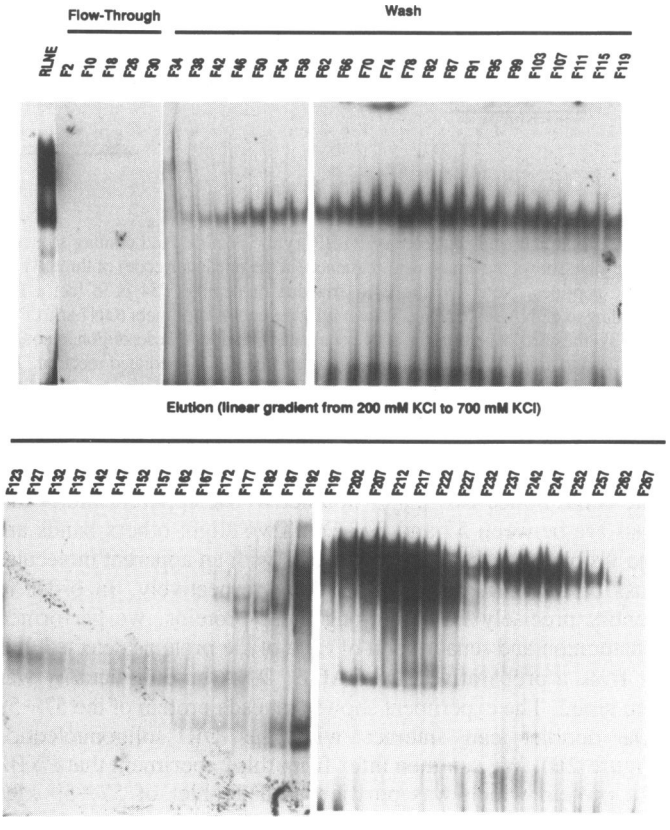

Figure 1. Heparin-Sepharose chromatography of rat liver nuclear proteins. The figure shows the analysis of the heparin-Sepharose chromatography with a double strand oligonucleotide bearing the Tf DRI element. The top of each lane indicates the fraction number. The salt concentration is indicated above the fraction numbers. The lane labelled RNLE shows the complexes detected with crude rat liver nuclei extracts.

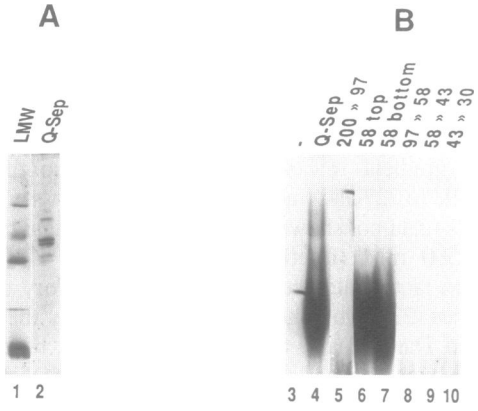

Figure 2. Identification of the active PYBP polypeptides species. A: silver stained SDS-PAGE of purified PYBP. Lane 1: molecular weight markers (97 kDa, 67 kDa, 43 kDa, 30 kDa, 20 kDa and 14 kDa). Lane 2: 10 μl of purified PYBP (Q-Sepharose fraction). B: analysis of DNA-binding activity of the polypeptides eluted from slices of a preparative SDS-PAGE gel. Lane 3: no protein. Lane 4 purified PYBP (Q-Sepharose fraction). Lane 5: proteins eluted from the 200 kDa to 97 kDa gel slice. Lane 6: protein eluted from the upper band of the 58 kDa doublet. Lane 7: protein eluted the lower band of the 58 kDa doublet. Lane 8: proteins eluted from the 97 kDa to 58 kDa gel slice. Lane 9: protein eluted from the 58 kDa to 43 kDa gel slice. Lane 10: protein eluted from the 43 kDa to 30 kDa gel slice.

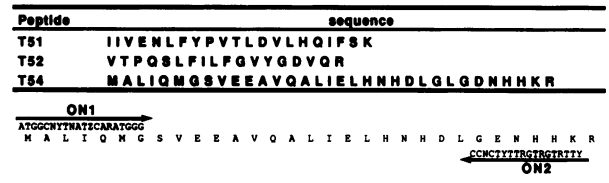

Figure 3. Amino acid sequence of PYBP tryptic peptides and cloning strategy. A: the table shows the amino acid sequence (in the one letter code) of three PYBP tryptic peptides. Note that the Lys_{30} residue of peptide T54 is in fact a Leu according to cDNA sequencing. B: two fully degenerated primers (ON1 and ON2) were synthesized based on the T54 sequenced peptide. These oligonucleotides were used to amplify the intervening sequences in a PCR-mediated reaction. The amplified DNA was sequenced and used as a long non-degenerated probe (90 bp) to screen a cDNA library.

One could notice two major doublet whose apparent molecular mass are between 57 and 59 kDa. Two slight others bands are also visible corresponding to proteins with an apparent molecular mass of about 40 kDa and 85 kDa respectively. In order to identify precisely PYBP among these proteins, we performed denaturation and renaturation of each of the proteins detected thus far from a preparative SDS-PAGE. DNA-binding activity was then tested. The experiment shows that each protein of the 57–59 kDa doublet can interact with the DRI oligonucleotide (Figure 2B). We can then infer from this experiment that PYBP is a protein which was purified as a doublet of 57–59 kDa apparent molecular mass.

Tryptic peptides of PYBP were prepared, purified and some of them sequenced (see Materials and Methods). Figure 3A shows the primary structure of three peptides which yielded unambiguous sequences. The sequence of the T54 peptide allowed us to design a non degenerated probe to screen a cDNA library. Indeed, T54 is a 30-amino acid residues peptide bearing at its

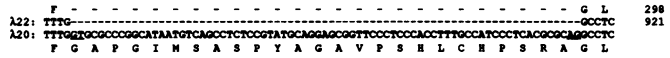

Figure 4: Comparison of λ22 and λ20 cDNA.clones. The nucleotide sequence of λ22 cDNA, as well as the predicted amino acid sequence is aligned with λ20 cDNA sequence in the position where there is a variation between the two cDNAs. Outside this zone, the nucleotide sequence of λ22 and λ20 cDNAs clones are identical. Note that the 75 nucleotide insertion does not shift the reading frame of λ20 cDNA. The putative donor/acceptor splice sites are underlined.

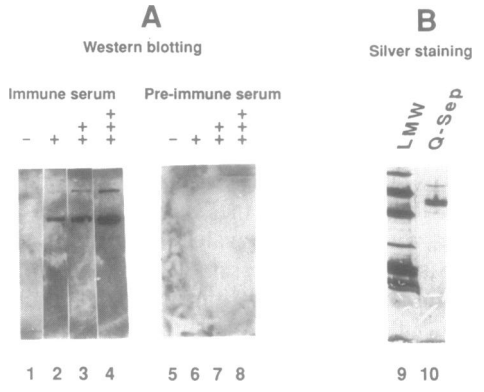

Figure 5. An antiserum against a synthetic peptide whose sequence was deduced from PYBP cDNA reacts with the doublet of the purified PYBP. A: Western blot analysis of purified PYBP using peptide antiserum. Lanes 1 and 5: no proteins. Lanes 2 and 6: 1 μl of purified PYBP. Lanes 3 and 7: 5μl of purified PYBP. Lanes 4 and 8: 10 μl of purified PYBP. The blot shown on the left was incubated with the peptide antiserum described in Materials and Methods. The blot shown on the right was incubated with a pre-immune serum from the same rabbit. B: Silver-stained SDS-PAGE of the purified PYBP. Lane 9: molecular weight markers (97 kDa, 67 kDa, 43 kDa, 30 kDa, 20 kDa and 14 kDa). Lane 10: 10 μl of purified PYBP.

C- and N-terminal moieties residues encoded by the less degenerated codons of the genetic code (Figure 3A). A 90 bp cDNA fragment was amplified using two degenerated primers (Figure 3B). The nucleotide sequence of this 90 bp fragment confirmed that it encoded the T54 peptide.

Cloning and sequencing of a cDNA coding for PYBP

An hepatoma size-selected cDNA library was screened with the previously described probe. This cDNA library was obtained by treatement of Fao cells with cycloheximide in order to increase the stability of potentially unstable mRNAs (24). 50 cDNA clones were isolated from 500,000 recombinant λgt10 bacteriophages using standard methods (7). Among these positive phages, two clones, λ22 and λ20, with the longest inserts were sequenced. The λ22 clone insert is 2,697 nucleotides long and ends with a polyA tail (data not shown, the nucleotide sequence is available in EMBL database under the accession no. X60789). We notice that the first AUG codon lies at position 27 and its surrounding sequences match the consensus sequence for translation initiation (25). Moreover, this codon is not preceded by an in frame translation stop codon. Assuming this AUG codon as the translation initiation codon, the λ22 open reading frame (ORF) encodes a 530 amino acid residues protein whose calculated molecular weight is 56,872. It is notable that this molecular weight agrees well with the apparent molecular mass of PYBP determined by denaturation-renaturation experiments (Figure 2). The T51, T52 and T54 PYBP tryptic peptides are also encoded

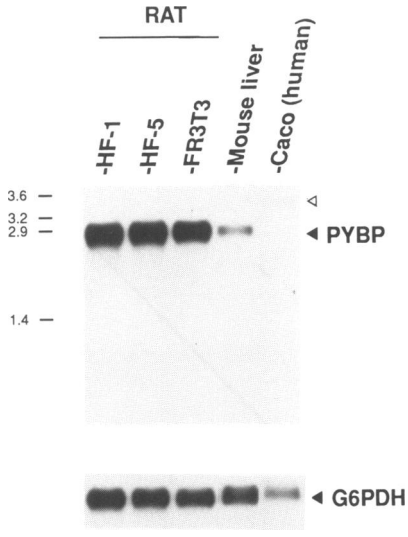

Figure 6. Northern blot analysis. High stringency hybridization of 5 μg of poly(A)$^+$ RNA isolated from rat and human cell lines, and from a mouse tissue were performed with a DNA probe coding for the C-terminal moiety of PYBP. Numbers at the left of this autoradiogram refer to the electrophoretic mobility of RNA molecular weight markers. The size of the detected PYBP RNA is about 2.8 kb in rat and mouse cells (indicated with a pointing arrow), and 3.4 kb in human cells (indicated with an empty arrow). The blot was also hybridized with a glyceraldehyde-6-phosphate-dehydrogenase probe (G6PDH) and its mRNA is indicated with a pointing arrow. HF-1 and HF-5 are rat hepatoma cells respectively differentiated and dedifferentiated, FR3T3 is a rat fibroblast cell line and CaCo-2 is a human cell line derived from a colon carcinoma.

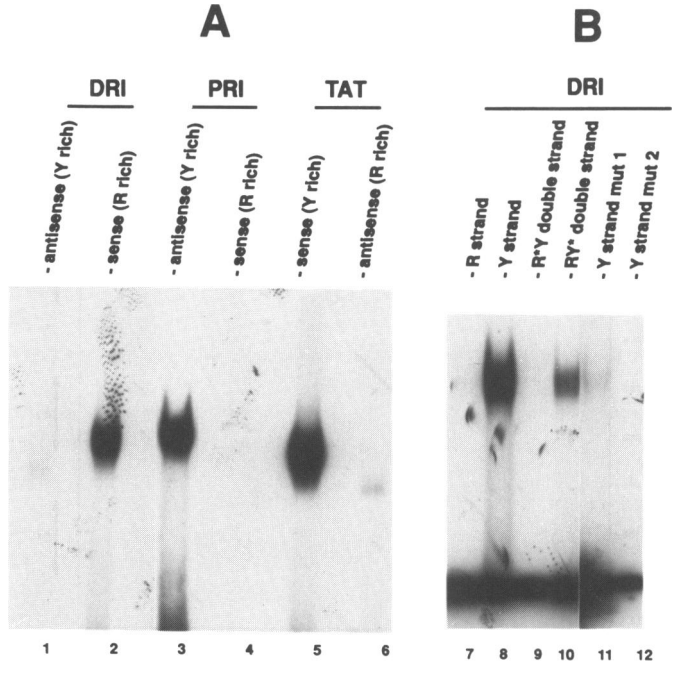

Figure 7. Analysis of PYBP DNA-binding activity. A: PYBP binds only to the pyrimidine-rich strand of the PRI, DRI and TAT oligonucleotides. The oligonucleotides used in the EMSA are indicated above each lanes. Y and R mean respectively pyrimidine and purine. B: PYBP is able to disrupt the double strand. EMSA were performed with the 5'-end labelled oligonucleotides indicated above each lane. When a double strand oligonucleotide was used, the star (*) indicates which strand was 5'-end labelled. Refer to panel A for the Y and R abbreviations. The sequence of the two DRI pyrimidine strand mutants is shown in Table 1.

by this ORF. The other clone, λ20, is 2,297 bp long and lacks the first 500 nucleotides of the λ22 clone (data not shown, the nucleotide sequence is available in EMBL database under the accession no. X60790). However, the nucleotide sequence of the λ20 insert is identical to the corresponding sequence of the former clone except for an insertion of 75 nucleotides between positions 916 and 917 which does not disrupt the reading frame (Figure 4). Thus, despite the fact that λ20 clone lacks the 5' end of the ORF, it probably encodes a protein identical to the one encoded by the λ22 clone with an insertion of a 2,432 Da peptide sequence. This molecular weight agrees well with the fact that PYBP was identified after denaturation-renaturation from SDS-PAGE slices as a doublet of 57−59 kDa. Furthermore, the insertion is flanked by donor/acceptor splice sites (Figure 4) (26). It is thus likely that these two cDNA clones derive from an alternatively spliced mRNA.

Antibodies directed against a synthetic peptide whose sequence was deduced from the λ22 clone recognize purified PYBP 58 kDa protein

An epitope prediction algorithm (27) was used to select a peptide from the λ22 deduced amino acid sequence in order to obtain antibodies. A 13-amino acid peptide was synthesized and linked to the keyhole limpet hemocyanin (see Materials and Methods). The conjugated peptide was injected into rabbits, leading to the production of antibodies directed against the synthetic peptide. As shown in Figure 5, the antiserum is able to react against each protein of the PYBP doublet in a Western blot assay. As a control, a pre-immune serum from the same rabbit failed to interact with PYBP (Figure 5). These results indicate that antibodies elicited towards a region of the λ22 cDNA deduced amino-acid sequence, recognize a similar sequence in purified PYBP. This strongly underlines the correspondence between purified PYBP and the protein encoded by our cDNA clones.

PYBP mRNA is ubiquitously expressed

To determine steady-state levels of PYBP mRNA, a PYBP cDNA fragment (described in Material and Methods) was used as a probe in Northern blot analysis. The probe hybridizes to a mRNA of approximately 2,800 nucleotides, that is detected in all cell lines and tissues analyzed (Figure 7). The size of this detected mRNA is consistent with the length of the 2,697 bp λ22 cDNA clone. This results indicate that PYBP mRNA is ubiquitously transcribed. Moreover, it is possible to detect it in mouse tissues and human cell lines, suggesting a structural homology between human, mouse and rat PYBP mRNAs since the blot was hybridized and washed under the most stringent conditions. One could also notice that the size of the human mRNA detected with the PYBP probe is higher than the mouse or rat mRNAs.

Rat PYBP is highly homologous to a mouse protein and to the human PTB protein

A computer search in the EMBL database with the Pearson and Lipman algorithm (16) shows a striking homology between PYBP and a recently cloned murine plasmocytoma protein (A. Bothwell, personal communication), hereafter referred to under its database accession number X52101 (28). When the analysis of PYBP and X52101 cDNAs is confined to their ORF, the nucleotide sequences are 95% identical. As recently indicated elsewhere (29), X52101 is the murine version of the human PTB factor

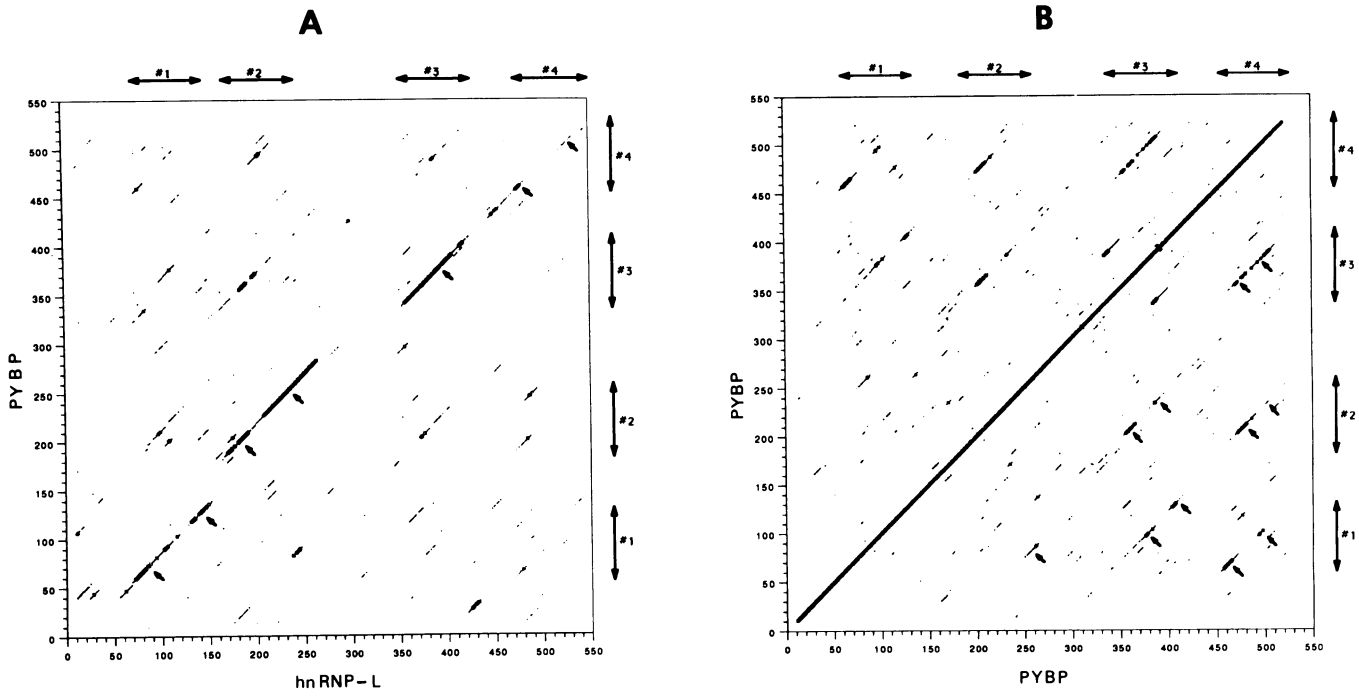

Figure 8. Dot matrix analysis of the aminoacid sequences of PYBP and hnRNP-L. Scores (S) were calculated with the Mutation Data Matrix (21) using a search window length of 25 residues. The peak values are expressed as the number of standard deviations (σ) above the matrix mean and plotted for the central window position. The path corresponding to the alignment shown in Figure 9 is delimited by small arrows. A: homology between PYBP and hnRNP-L. Blackened bars indicate score value higher than 4.1 σ; thin lines refer to $3.1\sigma < S < 4.1\ \sigma$. B: internal repeats in PYBP. For this comparison, blackened bars refer to score values higher than 3.5 σ, thin lines to $2.7\sigma < S < 3.5\ \sigma$.

(30). The PTB protein sequence was deduced from cDNAs clones obtained after PTB purification from HeLa cells and sequencing of some tryptic peptides (30). Comparison between PTB and PYBP nucleotide sequences shows that the cDNAs are 88% identical in their ORF (data not shown). The identity score at the protein level jumps to 97% identity, because of mutations of the wobble bases. The structural conservation between PTB, X52101 and PYBP is worth noticing and underlines a selective pressure to maintain a particular structure of the proteins.

It was previously reported that p62/PTB interacts with the polypyrimidine tracts of mammalian introns (31). The structural homology between PYBP and PTB led us to test if PYBP is also able to interact with single strand nucleic acid sequences.

PYBP binds to single strand pyrimidine-rich DNA

In order to test if PYBP would be able to interact with single strand sequences, we performed EMSA with the labelled sense strand and antisense strand of the oligonucleotides bearing the DRI element sequence (Table 1). As shown in Figure 7A, only the sense strand which contains pyrimidine tracts and the $_{5'}$TCTTTGACCT$_{3'}$ core, is recognized by purified PYBP (compare lanes 1 and 2). In the transferrin gene promoter, the proximal region PRI essential for the hepatic transcription of the gene (4, 32) contains pyrimidine tracts encompassing the same decanucleotide core but on the antisense strand (3, 6). Purified PYBP interacts only with this pyrimidine-rich strand of the PRI element (compare lanes 3 and 4 in Figure 7A). In the enhancer of the rat tyrosine aminotransferase (TAT) gene, the B domain contains a regulatory DNA element (33) which contains a sequence ($_{5'}$CTCTTTGATCT$_{3'}$) homologous to the transferrin PRI and DRI regions. Figure 7A shows again that purified PYBP interacts only with the pyrimidine-rich strand of the TAT B domain (compare lanes 5 and 6 in the Figure 7A). We thus conclude that PYBP is a DNA-binding-protein that binds specifically to the pyrimidine-rich strand of at least three regulatory *cis*-elements of hepatic genes. This was confirmed by experiments using DRI oligonucleotides whose pyrimidine tracts has been changed by purine sequences (Table 1). Indeed, PYBP interacts weakly with DRI mut 1, where only one pyrimidine tract has been mutated, and does not interact with the DRI mut2 oligonucleotide, where all pyrimidine tracts have been removed (compare lane 8 with lanes 11 and 12 in Figure 7B). The question arises then, how PYBP binding activity could be monitored through purification utilizing binding to a double strand oligonucleotide. To answer this question, we performed gel mobility shift assays and Figure 7B presents strong evidence indicating that PYBP may disrupt the DRI double strand in order to interact only with the pyrimidine-rich sequence. When the DRI sense strand is labelled, annealed with a two-fold molar excess of the unlabelled complementary strand, and the double strand subsequently purified (see Materials and Methods), a DNA-PYBP complex is detected by EMSA (lane 10 in Figure 7B). On the contrary, when the purified DRI double strand oligonucleotide is only labelled on the antisense strand, no DNA-protein complex is detected (lane 9 in Figure 7B). These observations suggest that, PYBP is able to disrupt the two strands of the DNA double helix, despite the energy constraints implicated in such an interaction.

PYBP contains four internal repeats whose sequence is compatible with the U1-A RRM folding

Database sequence comparisons also detected homologies of PYBP with the human heterogenous ribonucleoprotein L

(hnRNP-L) (34) , and to a lesser extent with the *Drosophila* neuronal protein *Elav*, required for normal visual system development (35). Pairwise comparisons between PYBP and hnRNP L revealed that the homology is largely confined to four distinct regions (Figure 8A). The alignment score of the entire sequences based on the homology matrix results is highly significant, about 20 standard deviations (σ) above the mean value obtained from random shuffling of the sequences (see Materials and Methods). In contrast, the significance of the overall homology between PYBP and *Elav* sequences is less clear, with an alignment score lower than 3 σ.

The existence of discrete homology regions between PYBP and hnRNP L moved us to search for internal repeats in the PYBP protein. Indeed, internal repeats which span approximately the same sequence fragments similar to hnRNP-L are detected, and allow us to propose the existence of four putative domains in PYBP (Figure 8B). The optimal alignment of these domains, as well as those of hnRNP-L, PTB and X52101 (respectively the human and mouse versions of PYBP, see above) are displayed in Figure 9. The internal repeats of PYBP are significantly homologous to each other, except for two pairwise comparisons (regions #1-#3 and #2-#4) which scored only 2.5 σ above the mean ; all other pairwise alignments scored higher than 4 standard deviations. The closest similarity between PYBP and hnRNP-L sequences is observed for the two central repeats, with respectively 32 and 25 identical residues for the alignment shown in Figure 9. The C-terminal repeats (#4) are the less similar to each other (16 identical residues, mostly in the N-terminal half of the sequence). In general terms, PYBP repeats seem to be more similar to the equivalent hnRNP-L domains than to each other, suggesting that the genetic duplication events leading to these repeats occurred earlier in evolution than the differenciation between PYBP and hnRNP-L.

A large family of nucleic acid-binding proteins that shares an RNA recognition motif (RRM) has been identified (29). Most members, but not all, of these proteins bind RNA. Some of them are also involved in DNA-binding (29). RRM consists of a conserved primary structure of about 80 residues containing two highly conserved short peptides (RNP-1 and RNP-2). Aromatic and basic residues present in these two peptides are thought to be involved in the interaction with RNA (36, 37). Recently, the three dimensional (3D) structure of an RNA binding domain from the U1-A small nuclear ribonucleoprotein (snRNP) has been determined by NMR spectroscopy (38) and X-ray diffraction (37). Availability of the tertiary structure of an RRM domain allows the discussion of the structural and functional implications of the conserved features in the amino acid sequences of this family of nucleic acid-binding proteins. The four PYBP internal repeats appear to exhibit the typical RRM found in several single strand nucleic acid-binding proteins (Figure 9). A limited degree of homology ($\approx$20%) is observed between PYBP repeats and the other RRMs described in the literature. However, secondary structure prediction (data not shown) suggests that all the sequences shown in Figure 9 share the ($\beta\alpha\beta-\beta\alpha\beta$) pattern observed in the 3D structure of the U1-A subunit. Moreover, hydrophobic residues buried in the tertiary structure, that are important for protein folding are mostly conserved (Figure 9). The first three PYBP RRMs have been recently described as such for both PTB (the human version of PYBP) and hnRNP-L (29, 30). The fourth putative RRM of PYBP (residues 451 to 527) is clearly homologous to the three others (Figure 8B and Figure 9), in particular to the second one (25% of identical

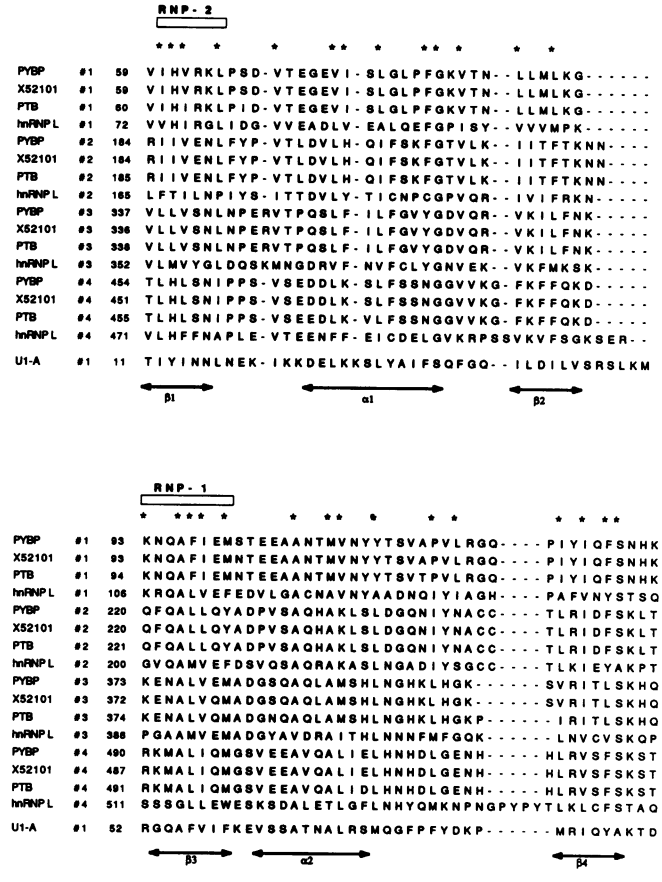

Figure 9. Alignment of the internal repeats of PYBP, PTB, X52101 and hnRNP-L with the first RNA-binding domain of the ribonucleotide U1-A (29). Multiple alignments are based on the dot search plots; minor modifications were required to generate the final optimal alignment. Asterisks indicate either buried residues in the U1-A structure or conserved positions among RRM-containing proteins. The limits of the secondary structure for the U1-A N-terminal domain are indicated with arrows, and the consensus RNP-1 and RNP-2 with bars.

residues). The identification of the fourth internal repeat in hnRNP-L as a RRM domain is more difficult, as only the N-terminal half of this sequence displays the expected pattern of conserved residues (Figure 9) The last 40 residues of this domain are not significantly homologous to other RRM-containing proteins.

It is interesting to note that the joining region between PYBP repeats #2 and #3 can accomodate a large insertion like the 25 residues deduced from the λ20 clone (Figure 4). The amino acid composition of this joining region in both PYPB and hnRNP-L proteins suggest a flexible (random coil) conformation in the respective tertiary structure. Finally, we can notice the poorly conserved RNP-1 and RNP-2 motifs in PYBP protein; this was already observed with the PTB and hnRNP-L proteins (29).

DISCUSSION

Experimental data now support the notion that different DNA-binding proteins can interact with *cis*-acting regulatory DNA elements which share sequences homologies (39, 40, 13). DRI is a positive *cis*-regulatory element that lies in the distal region of the human transferrin gene promoter (3, 4). We have

previously reported that different DNA-binding proteins interact with this element (6). Our previous proposal is now supported in this paper by the fractionation of these proteins by heparin-Sepharose chromatography (Figure 1). We further purified to near homogeneity (Figure 2A) one of these proteins which is referred as PYBP (PYrimidine Binding Protein). By denaturation and renaturation of proteins eluted from SDS-PAGE slices, we show that PYBP is a doublet of two polypeptides of about 57 and 59 kDa (Figure 2B). Cloning and sequencing of PYBP cDNAs allowed us to find that these two polypeptides likely arise from a differential spliced mRNA (Figure 4) that is ubiquitously expressed (Figure 6). Antibodies directed against a synthetic peptide whose sequence was deduced from the PYBP cDNAs clones recognize the purified protein (Figure 5). This result, together with the calculated molecular weights of the proteins encoded by the λ22 and λ20 cDNAs highlights the relationships between the proteins encoded by the cDNA clones and the purified proteins. Surprisingly, purified PYBP interacts specifically with the polypyrimidine-rich single strand DNA sequences, present in several *cis*-elements located in the regulatory regions of some hepatic genes (Figure 7).

Sequence comparisons show that the deduced PYBP peptide sequence is highly homologous to a recently cloned mouse plasmocytoma protein (28) and to PTB, its human version (30) (98% and 97% identity respectively). PTB, previously referred to as p62, binds specifically to pyrimidine tracts of introns (31). Furthermore, PYBP presents an homology (Figure 8A) with the hnRNP-L protein (34) and pairwise comparison shows that the homologies are confined to four distinct regions (Figure 8A). Internal sequence comparisons indicate that PYBP is organized in four repeats, each matching with the regions previously identified in the PYBP/hnRNP-L comparisons (Figure 8B). This finding strongly suggests that the actual PYBP gene is the result of several genetic duplication events. Optimal alignment of the sequences (Figure 9) further indicates that each PYBP repeat is more homologous to its hnRNP-L counterpart than to each other, suggesting that the divergent evolution between PYBP and hnRNP-L occurred after the genetic duplications. The four internal repeats of PYBP appear to exhibit the typical RNA recognition motif (RRM) identified in a wide variety of RNA or single strand DNA-binding proteins (29). Indeed, primary structure information and secondary structure predictions of the PYBP repeats match well with the three-dimensional model of the RRM deduced from cristallographic and spectroscopic data (37, 38).

We would like also to point out the capacity of the joining region which links the second and third RRM repeats to accomodate large insertions, as the 25 amino acid residues deduced form the l20 cDNA sequence (Figure 4). This observation is consistent with our four sequence domains model and suggests that the joining region might serve as a spacer between two separate structural domains. Indeed, it was recently shown (A. Bothwell, personal communication) that the C-terminal half of the murine X52101 protein possesses by itself the ability to bind RNA.

Early criteria for inclusion of a protein in the RRM family was heavily dependent on the high conservation of the two consensus sequences, RNP-1 and RNP-2. Following the determination of an RRM three-dimensional structure, different criteria based on tertiary rather than primary structure can now be considered. In this paper, we propose that the internal repeats of PYBP, its human PTB and mouse X52101 analogs, and hnRNP-L proteins should be considered as members of the RRM family (29). These are probably the first examples of putative RRM- containing proteins which exhibits poorly conserved RNP-1 and RNP-2 elements.

Until now, it has not been possible to assign an unambiguous function for PYBP, as well as for its human or murine version. PTB has been proposed to be a critical component of the splicing reaction (31). However, preliminary *in vitro* splicing experiments performed with purified PYBP, the rat PTB version, have thus far failed to demonstrate a function of PYBP in pre-splicing complex formation (A. Krämer, personal communication). In our laboratory, preliminary *in vitro* transcription reactions were performed (D. Mendelzon, unpublished results). In these assays, the −620/−9 transferrin promoter was linked to a G-free cassette (41). Using a limited amount of crude liver nuclear proteins, transcription of the cassette is strongly enhanced when the purified PYBP protein is added to the system. A similar result was obtained when using as a template the adenovirus 2 major late promoter upstream of a G-free cassette. This kind of assay is now being refined using liver proteins immunodepleted with polyclonal antibodies raised against cloned PYBP. In any case, the biological role of PYBP remains to be defined; indeed, this protein is abundant, highly conserved during evolution and its mRNA subjected to differential splicing. This protein binds to single strand DNA sequences rich in pyrimidine (this paper) and in a similar manner to mRNA introns as it was reported for the human version of PYBP (30). More recently a protein refered as ssARS-T has been purified (42). ssARS-T binds to the T-rich strand of the consensus core sequence of the yeast autonomous replicating sequences (ARS). Its properties point out to an essential function in the replication of the ARS. The unsual binding properties of PYBP (Figure 7) allow us to conceive that a localized unwinding of specific nucleic acid sequences may be concomitant with the binding of the protein whose role in transcription, replication and/or splicing remains to be investigated in detail.

ACKNOWLEDGEMENTS

We are very grateful to P. Jansen-Dürr, G. Shütz, A. Gil, A. Krämer and A. Bothwell for generously providing unpublished results. We thank G.N. Cohen, S. Cereghini, T. Chouard and O. Barzu for helpful discussions. We also thank J.C. Guillemot for his contribution in the protein sequencing experiments, N. Vita for the synthesis of peptides and I. Petropoulos for her generous help in nucleotide sequencing. The expert technical assistance of M.C. Py is acknowledged. We thank E. Croullebois for typing the manuscript. This work was supported by the Centre National de la Recherche Scientifique (U.R.A. 1129).

REFERENCES

1. Johnson, P.F. and MacKnight, S.L. (1989) *Ann. Rev. Biochem*, **58**, 799−839
2. Mitchell, P.J. and Tjian, R.T. (1989) *Science*, 245, 371−378.
3. Brunel, F., Ochoa, A., Schaeffer, E., Boissier, F., Guillou, Y., Cereghini, S., Cohen, G.N. and Zakin, M. M. (1988) *J. Biol. Chem.*, 263, 10180−10185.
4. Schaeffer, E., Boissier, F., PY, M.C. and Zakin, M.M. (1989) *J. Biol. Chem.*, **264**, 7153−7160.
5. Boissier, F., Augé-Gouillou, C., Schaeffer, E. and Zakin, M.M. (1991) *J. Biol. Chem.*, **266**, 98−22−9828
6. . Ochoa, A., Brunel, F., Mendelzon, D. and Zakin, M.M. (1989) *Nuc. Acids Res.*, **17**, 119−133.

7. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning: a laboratory manual. Cold Spring Harbor University Press, Cold Spring Harbor.
8. Garner, M.M. and Revzin, A. (1981) *Nucleic Acids Res.*, 9, 3047–3060.
9. Gorski, K., Carneiro, M. and Schibler, U. (1986) *Cell*, **47**, 767–776.
10. Laemmli, U. (1970) *Nature*, **227**, 680–685.
11. Hagard, D.A. and Burgess, R.B. (1980).*Anal. Biochem.* **109**, 76–86.
12. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA*, 74, 5463–5467.
13. Rey-Campos, J., Chouard, T., Yaniv, M. and Cereghini, S. (1991) *EMBO J.*, **10**, 1445–1457.
14. Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A. and Struhl, K. (1989) *Current protocol in molecular biology*. Greene Publishing Associates and Wiley Interscience, New York.
15. Zakin, M.M., Garel, J.R., Dautry-Varsat, A., Cohen, G.N. and Boulot, G. (1978) *Biochemistry*, **17**, 4318–4323.
16. Pearson, W.R and Lipman, D.J. (1988) *Proc. Natl. Aca Sci. USA*, **85**, 2444–2448.
17. Chou, P.Y. and Fasman, G.D. (1978) *Ann. Rev. Biochem.*, **47**, 254–276.
18. Staden, R. (1982) *Nucleic Acids Res.*, **10**, 2951–2961.
19. Needleman, S.B. and Wunsch, C.D. (1970) *J. Mol. Biol.*, **48**, 443–453.
20. The Protein Identification Resource (PIR), National Biochemical Research Foundation, Georgetown University Medical center. Washington D.C..
21. Dayhoff, M.O., Barker, W.C. and Hunt, L.T. (1983) *Meth. Enzymol*, 91, 524–545.
22. Hardon, E.M., Frain, M., Paonessa, M., and Cortese, R. (1988) *EMBO J.*, **7**, 1711–1719.
23. Rangan, V.S. and Das, G.C. (1990) *J. Biol. Chem.*, 265, 8874–8879.
24. Deschatrette, J. and Weiss, M. (1974) *Biochimie* , **56**, 1603–1611.
25. Kozak, M. (1987) *Nucleic. Acids Res.*, **15**, 8125–8133.
26. Breathnach, R. and Chambon, P. (1981) *Ann. Rev. Biochem.*, **50**, 349–383.
27. Dootlittle, R.F., (1982) *J. Mol. Biol.*, **157**, 105–132.
28. Bothwell, A.L.M., Ballard, D.W. and Philbrick, W.M. (Unpublished). X52101.
29. Kenan, D.J., Query, C.C. and Keene, J.D. (1991) *Trends Biochem. Sci.*, **16**, 214–220.
30. Gil, A., Sharp, P.A., Jamison, S.F. and Garcia-Blanco, M.A. (1991) *Genes Dev* , **5**, 1224–1236.
31. Garcia-Blanco, M.A., Jamison, S. and Sharp, P.A. (1989) *Genes Dev.*, **3**, 1874–1886.
32. Mendelzon, D., Boissier, F. and Zakin, M.M. (1990) *Nucleic Acids Res.*, **18**, 5717–5721.
33. Boshart, M., Weih, F., Schmidt, A., Fournier, R.E.K. and Schütz, G.(1990). *Cell*, **61**, 905–916.
34. Piñol-Roma, S., Swanson, M.S., Gal, J.G. and Dreyfuss, G. (1989) *J. Cell Biol.* , **109**, 2575–2587.
35. Robinow, S., Campos, A.R., Yao, K.M. and White, K. (1988). *Science*, **242**, 1570–1572.
36. Merril, B.M., Stone, K.L., Cobianchi, F., Wilson S.H. and Williams, K.R. (1988). *J. Biol. Chem.*, **263**, 3307–3313.
37. Nagai, K., Oubridge, C., Jessen, T.H., Li, J. and Evans, P.R. (1990). *Nature*, **348**, 515–520.
38. Hoffman, D.W., Query, C.C. , Golden, B.R., White, S.W. and Keene, J.D. (1990) *Proc. Natl. Acad. Sci. USA*, **88**, 2495–2499.
39. Hai, T., Liu, F., Coukos, W.J. and Green, M. (1989) *Genes Dev.*, **3**, 2083–2090.
40. Schaffner, W. (1989). *Trends Genet.*, **5**, 37–39.
41. Sawadogo, M.S. and Roeder, R.G. (1985). *Proc. Natl. Acad. Sci. USA*, **82**, 4394–4398.
42. Schmidt, A.M.A., Herterich, S.U. and Krauss, G. (1991) *EMBO J.*, **10**, 981–985.