Cloning and characterization of a novel sequence-specific single-stranded-DNA-binding protein

Dashzeveg BAYARSAIHAN¹, Ricardo J. SOTO and Lewis N. LUKENS²

Department of Molecular Biology and Biochemistry, Wesleyan University, Middletown, CT 06459, U.S.A.

The promoter region of the chicken $\alpha 2(I)$ collagen gene contains a pyrimidine-rich element that is well conserved in different mammalian species. This sequence can also form an unusual DNA structure as shown by its sensitivity to S1 nuclease *in vitro* and it lies in a region that is DNase I-hypersensitive only when this promoter is active. We have recently reported that fibroblast nuclear proteins, including chicken Y-box-binding protein 1, bind to this single-stranded pyrimidine-rich sequence. Here we report the isolation, from a chick embryo fibroblast cDNA expression library, of a partial cDNA clone encoding a previously unknown protein, designated SSDP (sequence-specific single-stranded DNA-binding protein), that binds this single-stranded sequence. This clone contains 1199 bp of chicken sequence and has a single

long open reading frame that encodes 284 amino acid residues. The affinity-purified recombinant protein encoded by this cDNA binds sequence-specifically to the single-stranded pyrimidine sequence. This cDNA sequence lacks significant similarity to any known gene in the data banks, but it is highly conserved in expressed sequence tags derived from both mouse and human. The corresponding amino acid sequence is remarkably conserved, having 97 % identity with mouse and human expressed sequences. The corresponding mRNA is approx. 1800 nt in length and is expressed in both fibroblasts and chondrocytes. The high affinity of this protein for this conserved pyrimidine-rich region suggests that it might be involved in the transcriptional regulation of the $\alpha 2(I)$ collagen gene.

INTRODUCTION

The expression of the collagen genes in higher vertebrates is precisely regulated during embryogenesis, and a variety of *cis*- and *trans*-acting elements have been shown to be involved in this regulation (reviewed in [1,2]). In the chicken $\alpha 2(I)$ collagen fibroblast promoter, the DNA segment from –712 to –85 bp upstream of the transcription start site has been shown to be important for promoter activity, and DNA-binding assays suggested that the region from –283 to –85 bp was responsible for the major complex formed between this DNA sequence and proteins present in chick embryo fibroblast nuclear extracts [3].

As part of an effort to identify the DNA sequences in the $\alpha 2(I)$ collagen promoter that are responsible for the induction of transcription from the fibroblast transcription start site that occurs when chondrocytes are treated with 5-bromo-2'-deoxyuridine, we have been investigating the possible regulatory role of a polypyrimidine–polypurine sequence located around –200 bp. This sequence is conserved between higher vertebrates and has been proposed to be important in the transcriptional regulation of the $\alpha 2(I)$ collagen gene [4]. Further evidence that this region is likely to have a regulatory role comes from the finding that it contains a DNase I-hypersensitive site associated with the transcriptional activity of the fibroblast-specific promoter [5–7]. This hypersensitive site is absent from the chromatin of chondrocytes that do not initiate transcription at the fibroblast start site, but it reappears when bromodeoxyuridine-treated chondrocytes shift to the use of the fibroblast-specific start site [7].

In vitro, in negatively supercoiled plasmids, this polypyrimidine–polypurine sequence forms a secondary structure different from canonical B-DNA, as shown by its sensitivity to S1 nuclease [8,9]. These S1 nuclease-sensitive sites have been found in the genomes of various organisms and they frequently, as in this case, contain homopyrimidine–homopurine tracts with mirror repeat symmetry (see the sequence of CT oligonucleotide in the Materials and methods section). Such sequences are capable of forming H-DNA structures composed of intramolecular triple-helical and single-stranded regions [10–12]. In eukaryotes these homopyrimidine–homopurine tracts of at least a few dozen base pairs constitute up to 1 % of the entire genome [13]. The pyrimidine–purine tracts often occur in the 5'-flanking regions of eukaryotic genes, and a number of these tracts have been shown to adopt a supercoil-dependent H-DNA structure *in vitro* (reviewed in [14]).

We have recently reported that the polypyrimidine sequence derived from the region of the chicken $\alpha 2(I)$ promoter from –205 to –187 inclusive is bound sequence-specifically by factors present in nuclear extracts from 10-day-old chick embryo fibroblasts [15]. Unexpectedly, the single-stranded sequence is bound with much higher affinity than the double-stranded sequence, and the single-stranded polypurine sequence is not bound at all. To identify the nuclear proteins responsible for this binding, we used an expression-cloning strategy to isolate the cDNA species encoding these proteins from a cDNA library derived from 10-day-old chick embryo fibroblasts. We chose the two largest clones for further study. One of these was found [16] to code for the chicken Y-box-binding protein 1 (YB-1), a member of the

Abbreviations used: YB-1, Y-box binding protein 1; SSDP, sequence-specific single-stranded DNA-binding protein.

¹ Present address: Department of Biology, Kline Biology Tower, Yale University, New Haven, CT 06520, U.S.A.

² To whom correspondence should be addressed.

The nucleotide sequence data reported will appear in DDBJ, EMBL and GenBank Nucleotide Sequence Databases under the accession number U68380.

cold-shock domain family of proteins [17]. The other cDNA clone is the subject of this paper. This cDNA encodes a previously unknown protein that binds with high affinity to the single-stranded polypyrimidine sequence. We propose to call this protein SSDP, for sequence-specific single-stranded DNA-binding protein. This protein seems to be highly conserved in mouse and human.

MATERIALS AND METHODS

Oligonucleotides

Sequences of oligonucleotides used in the present work were as follows, with mutated nucleotides underlined: CT, TCTTCCT-CCCTTCCTCCCTCCCTC; GA, GGGCGAGGGAGGGAG-GAAGGGAGGG; ssM1, TCTTCCGTGGCTTCCTCCCTC-CCTC; ssM2, TCTTCCCTCCCTTCCTCGGTCACTC; ssM3, TCTTCCGTGGCTTCCTCGGTCACTC; ssM4, TCTTCCCT-CCCTAGGTCCCTCCCTC; ssSp1, TTCAGACGTTCCCG-CCCCTC; EA1, GATCCACATCATACCTAGCAG.

The single-stranded CT oligonucleotide was radiolabelled with T4 polynucleotide kinase (New England Biolabs, Beverly, MA, U.S.A.) and [γ - 32 P]ATP (Amersham, Arlington Heights, IL, U.S.A.). Labelled nucleic acids were purified with a Microcon-3 spin column (Amicon, Beverly, MA, U.S.A.).

cDNA library screening

A 10-day-old chicken embryo cDNA library in λ gt11 (Clontech, Palo Alto, CA, U.S.A.) was screened by using the standard procedure [18], as described previously [16]. DNA inserts from positive clones were directly amplified from phage lysates by PCR with λ gt11 forward and reverse primers and then subcloned into pGEM-T vector (Promega, Madison, WI, U.S.A.). Plasmid DNA was purified with the Qiagen (Chatsworth, CA, U.S.A.) Midi Kit and the insert was partly sequenced with Sequenase 2.0 (United States Biochemical Corporation, Cleveland, OH, U.S.A.). The entire cDNA sequence on both strands was then determined by the automated sequencing service from Sequetech Corporation (Mountain View, CA, U.S.A.). Sequence analysis was performed with the BLAST program.

Production of β -galactosidase and His-tagged fusion proteins

β -Galactosidase fusion proteins were prepared from the recombinant λ gt11 phage designated CT-3, as described previously [16]. For production of the CT-3 encoded protein as a His-tagged and S-peptide-tagged fusion protein in a bacterial system, the entire *Eco*RI cDNA fragment from the CT-3 recombinant pGEM-T plasmid was cloned into the *Eco*RI site of the pET-30c vector (Novagen, Madison, WI, U.S.A.). The subsequent steps, including purification of the fusion protein on a nickel column and Western blot detection by using the S-peptide tag and the Novagen S-tag CL-HRP Western Blot kit, were performed as described in Novagen protocols.

Northern blots and reverse transcriptase-mediated PCR

Total RNA from fibroblasts of 10-day-old chick embryos and from sternal chondrocytes from 14-day-old embryos was isolated by using the Promega kit, and mRNA was purified with an oligo(dT) column in accordance with the manufacturer's protocol (Qiagen). For Northern blots, 10 μ g samples of total RNA were separated on 1.0% (w/v) agarose/2.2 M formaldehyde gels by standard procedures [19], and blotting and hybridization was performed with a DIG/Genius[®] 4 RNA Labeling Kit and

a DIG/Genius[®] 7 Luminescent Detection Kit (Boehringer Mannheim, Indianapolis, IN, U.S.A.). The template used for synthesis of the digoxigenin-labelled RNA was the entire CT-3 sequence present in the pGEM-T vector.

For reverse transcriptase-mediated PCR the first-strand cDNA synthesis was performed with 0.20 μ g of poly(A)⁺ RNA with random hexamer primers, with a First Strand cDNA Synthesis kit (Novagen). For amplification, the following primers were used: F1, ATCCCACGCGACAGCAAG corresponding to the CT-3 sequence from 253 to 270; R0, GGAGTGCCCCGGGG-ATTGCTAATGCC corresponding to the complementary strand from 778 to 753. The PCR conditions were as follows: 94 °C denaturing for 1 min, 60 °C annealing for 1 min, 72 °C extension for 1 min; 25 cycles. The amplified product was electrophoresed on a 1.0% (w/v) agarose gel.

Electrophoretic mobility shift assay

Protein-DNA complexes were identified by running samples on non-denaturing 5% (w/v) PAGE (29:1, acrylamide to bisacrylamide) [15]. The labelled oligonucleotide probe (20000 c.p.m./ng) was incubated with protein samples (1–2 μ g of β -galactosidase fusion protein or 0.01–0.1 μ g of affinity-purified protein) for 15 min at 22 °C in a reaction buffer containing 20 mM Tris/HCl, pH 7.9, 2 mM MgCl₂, 50 mM KCl, 1 mM dithiothreitol, 8% (v/v) glycerol and 1 μ g of poly(dI-dC). In competition assays the binding reaction was performed in the presence of an excess amount, as indicated, of unlabelled competitor oligonucleotide for an additional 10 min before the addition of the radiolabelled probe. Gels were dried and analysed by autoradiography.

RESULTS

Isolation of the cDNA clone encoding chicken SSDP

To identify the nuclear factors that bind the single-stranded polypyrimidine (CT) sequence present in the chicken α 2(I) collagen promoter, we screened a λ gt11 cDNA expression library derived from 10-day-old chick embryo fibroblasts with the labelled single-stranded CT oligonucleotide. After three rounds of screening, several positive plaques were obtained. One of them, designated CT-3, was selected for further study because the β -galactosidase fusion protein that it produced had the ability to bind the labelled CT probe sequence-specifically. This binding was abolished in the presence of excess unlabelled CT oligonucleotide but was not decreased by an excess of unlabelled mutant M3-oligonucleotide (results not shown), which was previously shown not to bind the factors present in fibroblast nuclear extracts [15].

Amino acid sequence and structure of SSDP

The insert in clone CT-3 was amplified by PCR with λ gt11 forward and reverse primers; it was then subcloned into pGEM-T to yield the plasmid pGEM-CT3. The 1.2 kb cDNA insert in this vector was completely sequenced from both ends by primer walking (Figure 1). This clone contains a single long open reading frame coding for 284 amino acid residues, starting after the 5' flanking *Eco*RI site (GAATTC) with the GAC codon for aspartate. The first ATG codon in this reading frame at bp 108–110 (Figure 1) almost certainly codes for an internal methionine residue, because translation seems to begin in a more 5' region of the SSDP mRNA that is missing from the CT-3 clone. This mRNA contains approx. 1800 nt (see below); approx. 600 bp are therefore missing from the CT-3 clone, which contains 1199 bp after subtraction of its flanking *Eco*RI sites. The missing

1 D T C E H S S E A K A F H D Y S A A
 19 GAATTCGGGACATTCGGAGCATTCGAGCGAAGCCTTCCATGACATAGCGCAGC
 62 A A P S P V L G N I P P N D G M P G G P
 123 TC6CGCCCGAGTCCCGTGTGGAAACATTC6CGGACGACGGGATGCCCCGGGGCCCC
 19 I P P G F F Q P F M [S] P R Y A G G P R P P
 123 ATCCCGCAGGTTTCTTCAGCCTTTTCATGTACCCGCGCTACGACAGGAGTCCCGGCCCC
 60 I R M G N Q P P G A V P G T Q P L L P N
 184 CCATCCGGATGGGAAACCCAGCCTCCGGGGGCGCTCCCTGGTACACAGCCATTCGTGCCCCAA
 80 S M D P T R Q Q G H P N M G G P M Q R M
 245 TTCGATGGATCCACGCGACAGCAAGGCGACCCGAAACATGGCGGACCCATGCAAGGATG
 100 N P P R G M G P M G P G P Q N Y G S G M R
 306 AATCCTCCGAGGATGGGCGCCCATGGGTCCCGGTCCGAGCAATATGGCAGCGGCATGA
 121 P P P N S L G P G M P G I N M G P G A G
 367 GACCTCCGCCCCAATCTTAGGCCCTGGCATGCCCGGAATTAACATGGGGCGGGAGCTGG
 141 R P W P N P S S A N S I P Y S S S P G
 428 CAGACCATGGCCCAACCCAGCAGTGTCACTCAATCCCATCTCTTCTTCATCGCCTGGT
 161 T Y V G P P G G G G G P P G T P I M P <S> P A
 489 ACATATGTGGTCTCCCGCGCGGGCGGCGCTCCGGGACACCCATGATGCCCGACCTCG
 182 D S T(N) S S D N I Y T M I N P V P P A G
 550 CAGATTCAACAAATCAAGTACACATCTACACAATGATTAATCCAGTACCGCCCGCAGG
 202 S R S N F P M G P G S D G P M G G M G G
 611 CAGTCGATGCAATTC6CGATGGGCGCGGCTCCGACGGCCCATGGGTGGCATGGTGGC
 222 M E P H H M (N) G S L G S G D I D G L P K N
 672 ATGGAGCCCATCATGATGATGATCTAGGTCAGGACATAGATGACATTCACAAAAA
 243 S P N (N) I S G I S N P P G <[T]> P R D D G E
 733 ATTCTCCAAACACATAAGTGGCATTAGCAATCCCGGGGCACTCAAGAGACGCGGGA
 263 L G G N F L G Q N D N Y S P S M T M
 794 GcTGGGAGGCAACATTCTccATTCcTTccAAAAATGacAATATTcCCCCAGcATGacGATG
 284 S V STOP
 855 AGTGTGTGATTTCCTCCCTCCCTCTCTCCAGGATCAAGAGAGTCAGCTGCCGTTCCGGA
 916 GGATCTCAAAAGAAATTATGCAAGAAAGAGCTAGGTGTATGGGCGAGGAGACGCTGGCGGG
 977 CCAGCACCOCAGGTTTATGTTTCATGCAATAAAAGGCTGAACCTTTTATTCATAAAACAT
 1038 GAAGGACAAAACCTAAATATTTCAAGACATGATGACAAGATCCCTTCTTGTGCGAAGGG
 1099 TTTATATATGTACATGACACTTCTTTTGGAAACAAATGGAGGCTCTAGCACCCAACTCC
 1160 GATCTCTTTGCTTTGTCTTTTAAATAAGACAGAAACAGAAACGGAATTC

Figure 1 Nucleotide sequence and deduced amino acid sequence of cDNA clone CT-3

The cDNA has a stop codon (underlined) and a 3' untranslated region with a consensus polyadenylation signal (underlined). The first six and last six bases in the above sequence come from the added *EcoRI* sites, not from the chicken sequence, which encodes 284 residues. The first ATG codon (bold and underlined) is believed to represent an internal methionine codon. Amino acid residues shown in curly brackets are potential N-glycosylation sites; those in square brackets are potential protein kinase C phosphorylation sites, and those in angle brackets are potential casein kinase II phosphorylation sites.

sequence is likely to be at the 5' end because the 3' end has both a translation stop codon and a polyadenylation signal. In addition, the open reading frame of SSDP is extremely conserved in human and mouse (see below) and it extends considerably 5' to the CT-3 sequence in these species. The 284 residues encoded by CT-3 would represent an acidic polypeptide region with a calculated PI of 5.7. This sequence contains no commonly occurring structural motifs but has two consensus sites for phosphorylation by casein kinase II and two potential sites for protein kinase C. Several potential glycosylation sites are also present.

Computer analysis revealed that the CT-3 sequence does not share significant similarity to any known gene in a data bank. Its sequence is, however, highly conserved in many shorter, but overlapping, expressed sequence tags from human and mouse that together span the entire open reading frame and almost all of the 3' untranslated sequence. When overlapping expressed sequence tags are aligned to give a continuous sequence (Figure 2), the polypeptide encoded by these expressed sequence tags is found to be remarkably conserved with the chicken sequence encoded by CT-3. The chicken amino acid sequence is 97% identical with both the human and mouse sequences. The mouse and human sequences are 98% identical with each other.

Affinity-purified recombinant SSDP binds specifically to the single-stranded CT sequence

To produce SSDP in *Escherichia coli*, we used the T7 RNA polymerase *in vivo* expression system developed by Novagen.

chicken	DTCEHSSEAKAFHDYSAASAPVLGNIPNDGMPGGP IPPGFFQPFMSPRYAGGPRPPI	60
mouse	E-----1-----P-----	
human	E-----AS-----	
chicken	RMGNQPPGAVPTQPLLPNSMDPTQQGHMNGMPPRMPPRGMPGPPQNYGSGMR	120
mouse	-----G-----S-----I-----	
human	-----G-----S-----1-----	
Chicken	PPFNSLGPMPGINMGPAGRPWPNPSSANSIPYSSSPGTIVGPPGGGPPGTPIMPSP	180
mouse	-----N-----	
human	-----A-----N-----	
Chicken	ADSTNSDNIYTMINPVPFAGSRNFPMPGSDGPMGGMGMEPHHMGSLGSGDIDGLP	240
mouse	-----G-----3-----	
human	-----G-----	
Chicken	KNSPNNISGISNPPGTFRDDGLGGLNLSHFQNDNYSPTMTSV Stop (TGA)	284
mouse	-----4-----5	
human	-----2-----3	

Figure 2 Comparison of the amino acid sequence of chicken SSDP and homologues in mouse and human

The sequence begins with the GAC codon for aspartate in Figure 1 and contains 284 residues. For most of the chicken sequence, several mouse and human expressed sequence tags (ESTs) matched the mouse and human sequences shown, but the specific EST sequences chosen are identified by the numbers under the last residue encoded by each segment. For mouse: segment 1, three mouse ESTs encode residue 1, and the subsequent sequence is from AA 415680, with guanines inserted after the base pairs that align with CT-3 guanine 117 and guanine 132 (Figure 1) because the other three mouse ESTs in this region have these guanines; 2, the first residue, Q, comes from three ESTs, and the remaining sequence from AA462935; 3, from AA267561; 4, from AA058029; 5, the TGA stop codon is from three ESTs. For human: segment 1, the first three residues are encoded by 3 ESTs, and the subsequent sequence by AA306675; 2, from AA311609, except that five unidentified base pairs (N) were replaced by the base pairs present in other equally similar ESTs; 3, from N79881.

Figure 3 His-tagged SSDP fusion protein of the expected size is produced from an isopropyl β -D-thiogalactoside-inducible expression vector (pET-30c) in *E. coli*

(A) Total proteins corresponding to 40 μ l of bacterial culture were resolved by SDS/PAGE (10% gel). Whole cell extracts were prepared by removing aliquots from the culture at different times after the addition of isopropyl thiogalactoside to induce expression. The gel was stained with Coomassie Blue. (B) The gel was loaded with a 1:100 dilution of the same samples and the fusion proteins were detected on a Western blot by binding of the S-peptide tag to labelled S-protein. Lane M, molecular mass protein markers (Bio-Rad) (sizes are indicated at the left in kDa); lane 1, non-induced transformant culture; lanes 2, 3 and 4 are 20, 40 and 60 min isopropyl thiogalactoside-induced transformant cultures respectively.

The entire sequence of CT-3 was cloned into the *EcoRI* site downstream of the histidine tag in the pET-30c expression vector. Bacterial clones that carried the CT-3 sequence (pET-CT3) were selected and induced with isopropyl β -D-thiogalactoside. A new protein band of approx. 36 kDa appeared on PAGE gels of these bacterial extracts (Figure 3A), and the same protein was revealed by a Western blot with the S-Tag protein from

Figure 4 Affinity purification on a Ni^{2+} column of the His-tagged SSDP fusion protein expressed in *E. coli*

Protein samples (10 μl) were resolved by SDS/PAGE (10% gel) and stained with Coomassie Blue. Lane 1, wash-through of column with binding buffer; lane 2, wash-through with wash buffer; lane 3, eluate of the bound fusion protein with eluate buffer; lane M, molecular mass markers (sizes are indicated at the left in kDa).

Novagen to detect the S-peptide tag present in the His-tagged fusion protein (Figure 3B).

To test whether the purified SSDP specifically binds the single-stranded polypyrimidine sequence of the $\alpha 2(\text{I})$ collagen gene, the His-tagged fusion protein was affinity-purified and shown by PAGE to consist of a single protein of the expected size (Figure 4). Electrophoretic mobility shift assays with the affinity-purified SSDP and end-labelled CT oligonucleotide gave a single retarded

Figure 5 Purified SSDP fusion protein expressed in *E. coli* binds specifically to the single-stranded pyrimidine-rich CT sequence, as shown by electrophoretic mobility shift assay

Affinity-purified SSDP protein was incubated with a ^{32}P -labelled CT oligonucleotide in the presence of a 20-fold molar excess of the indicated single-stranded unlabelled competitor DNA. The arrow indicates the specific DNA-protein complex. Lane 1, free DNA; lane 2, CT-3 fusion protein without competitor; lanes 3–10, as in lane 2 but with the addition of unlabelled oligonucleotide noted above the lane as competitor.

Figure 6 Expression of SSDP mRNA in chick embryo cells

Upper panel: total RNA (10 μg) from chick embryo fibroblasts (CEF) and sternal chondrocytes was subjected to Northern blot analysis by probing with a digoxigenin-labelled cRNA probe derived from the pGEM-CT3 plasmid. Lower panel: staining of the gel with ethidium bromide showed that approximately equal amounts of 27 S and 18 S ribosomal RNA were present in both lanes.

band representing the CT oligonucleotide-SSDP complex (Figure 5). Competition with the single-stranded unlabelled CT oligonucleotide completely chased away the complex, whereas the complementary single-stranded GA-oligonucleotide did not compete for binding. The unrelated single-stranded oligonucleotides ssSp1 and EA1 also competed poorly for binding. A mutated version of the CT sequence with an altered 3 bp spacer sequence between the two mirror repeats (M4) competed efficiently. However, three base changes in both mirror repeats drastically decreased the binding efficiency (M3). M1 and M2, with respectively the mutated 5' and 3' mirror repeat, had intermediate binding efficiency, with M2 binding with higher affinity than M1 (Figure 5).

Size and expression of SSDP mRNA

Northern blots of total RNA revealed approximately equal expression of the mRNA encoded by the CT-3 clone in 10-day-old chick embryo fibroblasts and 14-day-old chick embryo chondrocytes (Figure 6). A single band of approx. 1800 nt was seen on blots of both fibroblast and chondrocyte RNA. Similar results were obtained with the poly(A)⁺ RNA fraction (results not shown). As noted above, the CT-3 clone has 1199 bp of cDNA, compared with the 1800 bp predicted from the size of the mRNA, so that about 600 bp of the complete cDNA sequence are missing from CT-3. The expression of the SSDP mRNA in fibroblasts and chondrocytes was confirmed by reverse tran-

scriptase-mediated PCR, which gave a single band of the expected size of 526 bp after reverse transcription of RNA from fibroblasts or chondrocytes and amplification with the primers F1 and R0 (results not shown).

DISCUSSION

This report describes the cloning and partial characterization of a previously unknown single-stranded DNA-binding protein that is present in fibroblast nuclei and that binds a polypyrimidine sequence present in the promoter of the chicken $\alpha 2(I)$ collagen gene. The sequence of this polypyrimidine–polypurine tract indicates that it should be capable of forming a triple-helical H-DNA conformation [14], and it is clearly capable of forming a non-B-DNA structure, because it becomes SI-sensitive *in vitro* in negatively supercoiled plasmids [8,9]. This sequence is located within a DNase I-hypersensitive site that is present only in the chromatin of cells that contain the active form of this promoter [5–7]. This sequence is also conserved between vertebrate species [4] and between the $\alpha 2(I)$ and the $\alpha 1(I)$ collagen promoters [20].

To examine the possible regulatory role of this sequence, a first step was to ask whether it was bound by factors present in nuclear extracts from 10-day-old chick embryo fibroblasts, a cell type that rapidly synthesizes type I collagen. Electrophoretic mobility shift assays showed, surprisingly, that the single-stranded polypyrimidine sequence was bound with much higher affinity than the double-stranded sequence or the single-stranded polypurine sequence [15]. To identify the nuclear factors that bind the single-stranded polypyrimidine sequence, this sequence was used to screen a cDNA expression library derived from these fibroblasts [16]. This screening identified the YB-1 protein as previously reported [16], and also the novel SSDP protein described here. Because these two proteins are encoded by mRNA species present in these fibroblasts, and because they bind the polypyrimidine single-stranded sequence with a sequence specificity similar to that of crude nuclear extracts, they are strong candidates for the factors that bind this sequence *in vivo*.

The nucleotide sequence of the CT-3 cDNA clone (Figure 1) reveals that it encodes a previously undescribed protein. This sequence contains an open reading-frame coding for 284 amino acid residues, although, as discussed above, it is likely that the full-length SSDP protein contains more N-terminal amino acids. The 284-residue open reading frame in the CT-3 cDNA reveals, by computer analysis, no common secondary structure motifs. The proline and glycine contents are very high, representing respectively 20.2% and 17.1% of the total residues. The three next most abundant amino acids are serine, asparagine and methionine, representing respectively 10.8%, 8.0% and 7.3% of the total residues. The calculated pI of this 284-residue sequence is 5.7. This protein seems to be very conserved in higher vertebrates because the chicken sequence is 97% identical with the amino acid sequences from human and mouse (Figure 2), as deduced from expressed sequence tags. It is clear that this RNA encodes a protein rather than having some other function, because a strikingly small fraction of the base changes that occur between the cDNA sequences of chicken, mouse and human result in a change in the amino acid that is encoded. The exceptional degree of conservation of the SSDP protein sequence present in clone CT-3 is close to that found in chicken YB-1, which also binds single-stranded polypyrimidine sequences [17]. The mouse and human SSDP sequences are 98% identical (Figure 2), compared with 99% for YB-1. As discussed for YB-1 [17], such exceptional conservation suggests that SSDP might engage in numerous and specific protein–nucleic acid and protein–protein interactions.

The affinity-purified, His-tagged recombinant SSDP binds sequence-specifically to the single-stranded CT sequence derived from the $\alpha 2(I)$ collagen promoter (Figure 5). The recombinant SSDP shows the same relative binding affinities for different polypyrimidine-rich sequences, including mutated versions of the wild-type CT sequence as fibroblast nuclear extracts [15]. It is possible that, like YB-1 [17] and related proteins [21], SSDP will be found to have a 'generic' DNA sequence specificity, recognizing a variety of sequences that can form a particular secondary structure, such as H-DNA.

The presence of SSDP mRNA in chondrocytes, in which this $\alpha 2(I)$ promoter is inactive [3], as well as in fibroblasts (Figure 6) could indicate that SSDP is not involved in the activation of this promoter in fibroblasts. Chondrocytes, however, lack the DNase I-hypersensitive site present in fibroblast chromatin in the DNA region containing the polypyrimidine sequence. It is therefore possible that the chromatin structure in chondrocytes would not allow access of SSDP to this sequence.

Earlier studies suggested that this polypyrimidine–polypurine region in the chick $\alpha 2(I)$ collagen promoter was likely to have a functional role because it was conserved between species [4] and was located within a DNA region that became hypersensitive to DNase in chromatin only when this promoter was in a transcriptionally active form [5–7]. More recently, evidence has been obtained from transient transfections and transcription *in vitro* that the region in the mouse $\alpha 2(I)$ collagen promoter (–180 to –136) that contains the conserved polypyrimidine–polypurine sequence acts as an activator of transcription [22]. Moreover a number of proteins seem to bind to this region [22], including SP1, SP2, c-Krox [23] and a newly discovered protein, BFCOL1 [24]. The DNA-binding sites of proteins such as YB-1 and SSDP, which bind single-stranded DNA, have frequently been found to overlap the binding sites for a number of double-stranded DNA-binding transcription factors [17,25,26]. In these cases it has been suggested that the binding of the protein that recognizes single-stranded DNA influences transcription either by preventing access of the transcription factors that bind double-stranded DNA or by stabilizing a DNA secondary structure that prevents their binding. We are currently evaluating whether the interaction of SSDP with this region of the collagen $\alpha 2(I)$ promoter has a role in regulating the activity of this promoter.

This work was supported by grant no. 93-37206-9495 from the United States Department of Agriculture.

REFERENCES

- de Crombrughe, B., Karsenty, G., Maity, S., Vuorio, T., Rossi, P., Ruteshouser, E. C., McKinney, S. H. and Lozano, G. (1990) *Ann. N.Y. Acad. Sci.* **580**, 88–96
- Vuorio, E. and de Crombrughe, B. (1990) *Annu. Rev. Biochem.* **59**, 837–872
- Wang, S. and Lukens, L. N. (1994) *Matrix Biol.* **14**, 313–321
- Schmidt, A., Yamada, Y. and de Crombrughe, B. (1984) *J. Biol. Chem.* **259**, 7411–7415
- Merlino, G. T., McKeon, C., de Crombrughe, B. and Pastan, I. (1983) *J. Biol. Chem.* **258**, 10041–10048
- McKeon, C., Pastan, I. and de Crombrughe, B. (1984) *Nucleic Acids Res.* **12**, 3491–3502
- Beck, K. M., Seekamp, A. H., Askew, G. R., Mei, Z., Farrell, C. M., Wang, S. and Lukens, L. N. (1991) *Nucleic Acids Res.* **19**, 4975–4982
- Finer, M. H., Fodor, E. J. B., Boedtker, H. and Doty, P. (1984) *Proc. Natl. Acad. Sci. U.S.A.* **81**, 1659–1663
- McKeon, C., Schmidt, A. and de Crombrughe, B. (1984) *J. Biol. Chem.* **259**, 6636–6640
- Mirkin, S. M., Lyamichev, V. I., Drushlyak, K. N., Dobrynin, V. N., Filippov, S. A. and Frank-Kamenetskii, M. D. (1987) *Nature (London)* **330**, 495–497
- Behe, M. J. (1995) *Nucleic Acids Res.* **23**, 689–695
- Schroth, G. P. and Shing Ho, P. (1995) *Nucleic Acids Res.* **23**, 1977–1983
- Manor, H., Rao, S. and Martin, R. G. (1988) *J. Mol. Evol.* **27**, 96–101

-
- 14 Mirkin, S. M. and Frank-Kamenetskii, M. D. (1994) *Annu. Rev. Biophys. Biomol. Struct.* **23**, 541–576
 - 15 Bayarsaihan, D. and Lukens, L. N. (1996) *Biochem. J.* **314**, 293–296
 - 16 Bayarsaihan, D., Enkhmandakh, B. and Lukens, L. N. (1996) *Biochem. J.* **319**, 203–207
 - 17 Grant, C. E. and Deeley, R. G. (1993) *Mol. Cell. Biol.* **13**, 4186–4196
 - 18 Singh, H., Clerc, R. G. and LeBowitz, J. H. (1989) *BioTechniques* **7**, 252–261
 - 19 Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*, 2nd edn., pp. 7.43–7.45, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York
 - 20 Karsenty, G. and de Crombrughe, B. (1991) *Biochem. Biophys. Res. Commun.* **177**, 538–544
 - 21 Kolluri, R., Torrey, T. A. and Kinniburg, A. J. (1992) *Nucleic Acids Res.* **20**, 111–116
 - 22 Hasegawa, T., Zhou, X., Garrett, L. A., Ruteshouser, E. C., Maity, S. N. and de Crombrughe, B. (1996) *Nucleic Acids Res.* **24**, 3253–3260
 - 23 Galéra, P., Musso, M., Ducy, P. and Karsenty, G. (1994) *Proc. Natl. Acad. Sci. U.S.A.* **91**, 9372–9376
 - 24 Hasegawa, T., Takeuchi, A., Miyaishi, O., Isobe, K.-i. and de Crombrughe, B. (1997) *J. Biol. Chem.* **272**, 4915–4923
 - 25 Sapru, M. K., Gao, J. R., Walke, W., Burmesiter, M. and Goldman, D. (1996) *J. Biol. Chem.* **271**, 7203–7211
 - 26 Coles, L. S., Diamond, P., Occhiodoro, F., Vadas, M. A. and Shannon, M. F. (1996) *Nucleic Acids Res.* **24**, 2311–2317
-

Received 28 October 1997/8 December 1997; accepted 6 January 1998