Ribonucleoprotein SS-B/La belongs to a protein family with consensus sequences for RNA-binding

Edward K.L.Chan, Kevin F.Sullivan and Eng M.Tan

W.M.Keck Autoimmune Disease Center, Department of Molecular and Experimental Medicine, Scripps Clinic and Research Foundation, 10666 North Torrey Pines Road, La Jolla, CA 92037, USA

Received December 2, 1988; Revised and Accepted February 15, 1989

ABSTRACT

Autoantibodies from systemic rheumatic disorders have become useful reagents in molecular biology. SS-B/La, a major target of autoantibodies in lupus and Sjogren's syndrome, has been identified as a 46 kDa protein component of a ribonucleoprotein (RNP) particle implicated in the maturation of RNA polymerase III transcripts. This report describes the complete sequences of human and bovine SS-B/La and the identification of RNA-binding protein consensus sequences RNP1 and RNP2 in the N-terminal region previously shown to be complexed with RNA in UV-crosslinking experiments. Segments of about 95 residues from the RNA-binding domain of SS-B/La and from 29 RNA-binding domains of several other proteins are analysed with respect to the frequency of amino acids and their hydrophobicity at each position. The data suggest that SS-B/La belongs to a large family of RNA-binding proteins which includes heterogeneous nuclear RNPs, nucleolin, mRNA polyadenylate binding protein, and small nuclear RNPs.

INTRODUCTION

SS-B/La is a ubiquitous protein transiently associated with RNA polymerase III transcripts (1-8). SS-B/La also binds to U1 RNA (9), an RNA polymerase II transcript, and to the vesicular stomatitis virus leader RNA (10). The 3'-oligouridylylate residues of these RNAs are involved in the interaction with the SS-B/La protein (4,6,7). Since SS-B/La complexes only with the precursors of 5S RNA and tRNA and not with their corresponding mature species, there may be a role for SS-B/La in the maturation of these small RNAs (2). SS-B/La was originally identified as a target antigen of autoantibodies found in the sera of patients with systemic lupus erythematosus and Sjogren's syndrome (11,12). It is anticipated that further elucidation of the structure and function of SS-B/La might add to the understanding of these diseases. Since it has been shown that the antigenic determinants recognized by human autoantibodies are conserved among several mammalian species (13), efforts in the identification of these determinants are being directed to the cloning of the SS-B/La protein from different species. In this report, the cDNA clones encoding SS-B/La from two species, human and bovine, are described. Comparison of the full-length sequences reveals that SS-B/La is highly conserved between the two species. An analysis of RNA-binding regions of SS-B/La and 13 other RNA-binding proteins demonstrates, in addition to the presence of consensus sequences, further interesting similarities among these RNA-binding proteins.

MATERIALS AND METHODS***SS-B/La cDNA Clones***

A bovine pituitary λ gt11 cDNA library II was obtained from Professor Richard A.Maurer,

University of Iowa. Human T cell lymphoblastoma (MOLT-4) λ gt11 cDNA library was constructed by Drs. K.Ogata and D.J.Noonan, Scripps Clinic and Research Foundation. The bovine cDNA library was screened with human serum containing anti-SS-B/La antibodies by the method of Young and Davis (14). The MOLT-4 library was screened for human SS-B/La cDNA clones by DNA hybridization with a 5'-cDNA probe La6 (400-bp) kindly provided by Dr. J.D.Keene, Duke University. DNA probes were labeled by the method of Feinberg and Vogelstein (15). All screenings were carried out with duplicate filters and positive phages were plaque-purified. The resulting cDNAs were subcloned into Bluescript vectors (Stratagene, La Jolla, CA) for further analysis. Plasmids were purified by the Chen and Seeburg method (16) and DNA sequencing was performed by the dideoxy chain termination method of Sanger et. al. (17) using commercial SK, KS, T3 and T7 primers (Stratagene).

RNA Preparation

Human cells HeLa S3 (cervical epithelioid carcinoma), KB (oral epidermoid carcinoma), and A253 (submaxillary gland epidermoid carcinoma), and bovine kidney cells MDBK were obtained from American Type Culture Collection (Rockville, MD) and cultured as described (13). Sal-1, a human salivary gland-derived cell line (18), was kindly provided by Dr. R. I. Fox, Scripps Clinic and Research Foundation. Total RNA was isolated using the guanidinium thiocyanate solubilization method (19) and used directly in Northern blot analysis.

Antibodies

High titer human autoimmune SS-B/La serum Ca (20) was adsorbed with wild type λ gt11/RY1090 bacterial lysate prior to screening the bovine cDNA library. Five monoclonal antibodies to SS-B/La (A1 – A5) were produced from BALB/c mice hyperimmunized with SS-B/La and their characteristics were described in detail elsewhere (13).

SS-B/La Fragments and Protein Chemistry Methods

A crude preparation containing SS-B/La was obtained by ammonium sulfate fractionation

Figure 1: Sequencing strategy of both the bovine and human SS-B/La cDNAs. Restriction sites are labeled as follows: Bg, *Bgl*II; Bm, *Bsp*MI; Bs, *Bst*EII; K, *Kpn*I; S, *S*tyI; X, *Xba*I; Xh, *Xho*I; Xo, *Xho*II; Xm, *Xmn*I. The shaded boxes indicate the open reading frames. Arrows indicate the direction and the length of the region sequenced.

Human (Hu) -72 Bovine (Bo) -18 AACTGAAAGACGCCAAA
CGAGTCCTTCTGTTGCTCTTTGTGACCCCTGTCGGGGGCTTCTGTGCGCCGA C T T GC

Bo 1 M A E N G D N E K M A A L E A K I C H Q
Bo 1 ATGCCTGAAAAATCGTATAATGAAAAATGGCTGCTCGGAGGCCAAAATCTGTGATCAA
Hu 1 G C
Hu 1

21 I E Y Y F G D F N L P R D K F L K E Q I
61 ATTGAGTATTATTTGGAGACTTCAATTTGCCACGGGACAAAATTTTAAAGGAACAGATC
61 C G C A
21

41 K L D E G W V P L E I M I K F N R L N R
121 AAATCGATGAAGCTGGCTACCTTTGGAGATAATGATAAAGTTTAAAGTTAAACCGT
121 [(A C C G
41

61 L T T D F N V I V E A L S K S E A E L M
181 TTAACGACAGACTTTAATGTAATAGTAGAGGCCCTGAGCAAAATCAGAGGCAGAACTCATG
181 C A T G A AT CA
61 K

81 E I S E D K T K I R R S P S K P L P E V
241 GAAATAAGTGAAGATAAACTAAAATAGAAATCTCCAAGCAAACTCTCCCTGAAGTG
241 C C G C A
81

101 T D E Y K N D V K N R S V Y I K G F P T
301 ACTGATGAGTATAAAAAATGATGTA AAAACAGATCTGTTTATATTAAAGGCTTCCGCGAC
301 A T
101

RNP2

121 D A A L D D I K E W L E D K G Q V L N I
361 GATGCAGCTCTTGATGACATAAAAGAATGGTTAGAAGATAAAGGTCAACTACTAAATATT
361 A
121 T

141 Q M R R T L H K A F K G S I F A V F D S
421 CAGATGAGAAGAACCTTGACACAAAGCATTTAAGGGTTCAATATTGCTGTATTGATAGT
421 A T A T T G C
141 V

RNP1

161 I E S A K K F V E T P G Q K Y K D T D L
481 ATTGAATCAGCTAAAAAGTTTGTGAGACCCCTGGCCAGAACTACAAGGACACAGACCTG
481 T G A A A
161 E

181 L I L F K E D Y F T K K N E E R K Q N K
541 TTAATACCTTTCAAGGAAGATTATTTACCAAAAAAATGAAGAAAGAAAGCAAAATAAA
541 C C C TG A
181 D A

201 M E A K L R A K Q E Q E E K Q K L A E N
601 ATGGAAGCTAAATTACGAGCTAAAACAGAGCAAGAGAAAAACAAAGCTAGCAGAAAAAT
601 G A G C T A G
201 V A A E D

Bo peptide: M K S L E E K I G C L L K F S G D L
721 A E M K S L E E K I G C L L K F S G D L
661 CCTGAAATGAAATCTCTAGAAGAAAAGATTGGCTGCTGCTGAAATTTTCAGGAGACTTA
661 A G T T
221

241 D D Q T C R E D L H T L F S N H G E I K
721 GATGATCAGACCTGTAGAGAAGATTTCACACCCCTTTTCTCAAAATCATGCTGAAATAAAA
721 C A TA
241 I

261 W I H F V R G A K E G I I L F K E K A K
781 TCGATACACTTTGTGAGAGCCGCCAAAGAGGCAATAATTCTATTTAAGAAAAAGCTAAG
781 G C A A G C
261 D

281 E A L D K A K E A N N G N L Q L R N K E
841 GAAGCACTGGATAAAGCAAAAGAGGCAATAATCGTAACCTACAACCTAAGCAACAAAGAA
841 T G C A T T
281 G D

301 V T W E V L E G D V E K E A L K K I I E
901 GTCACCTGGGAAGTACTAGAAGGAGATGCTGAAAAAGCACTGAAAAAATAATAGAA
901 G T G E
301

321 D Q Q E S L N K W K S K G R R F K G K G
961 GATGAACAAGAATCTCTAAACAAATGGAAGTCAAAAGTCAAGATTAAAGAAAAAGGA
961 C C T
321

341 K G N K A A Q A G S A K G K V Q F Q G K
1021 AAGGGAATAAAGCTGCCAGGCTGGGTCTGCTAAAGAAAAAGTACATTTTCAGGCCAAG
1021 T C G
341 P G

361 K T K F D S D - - - D E R D E N G A S R
1081 AAAACGAAATTTGATAGTGAT-----GATGAACGTGATGAAAAATGGTGCATCTCGA
1081 C GATGAACAT A A G
361 A D E H H T G

378 A V K R A R E E T D K - E P P S K Q Q K
1132 GCAGTAAAAAGAGCAAGAGAGAAACAGACAAA---GAACCTGCATCAAAAACACAGAAA
1141 C T G GAA G C
381 P E A

397 T E N G A G D Q * 404
1189 ACAGAAAAATGCTGCCGAGACCACTGCTTTAATAACAAATTTTTTATTCATTAAATTA
1201 T G C T A
401 * 408

Bo 1249 GATTTTAAGCTGCTTTTGTCTTTGGAGGCTCTTAAAAAGAAACCGAATTAGATCCACTT
Hu 1261 G A GA - GC G T G

1309 CGATGCTACCTGTAAGAAAAAGAGATTTTTTGTGTTGTTGTAACCTGTCTTTTGT
1320 A C G G A A - - - - - T TG

1369 TATCGAA---ATACGTTTTTTAATGTGCAGTCTGCTTTGTGTTCTTCGGATTATTCAA
1375 TAT C AATG G TT C G ATT A A G

1425 CTATCAAAAGAAACATTCTCCGATTAAGTTGGCTTTGTAATATGAGAATCTATTAGTACA
1435 A G G A

1485 ---AAGTAATAAATGTATACGTACATAAAATAAACTAAGTTTATTTTTTTTAAATTA
1495 AACT C A AT G G GC--poly(A) 1530)

Bo 1541 CCAGCATATTGGCTCTTGTGATTATAGGCCCATCTGTTTACGTAAATATTTATCTTTT
1601 ATGCTGGCTTTCTGATTACTCAAGCAACTTGACAACATTAGAAATAATTTATCTGAAG
1661 TTTTTCAGTACTGAATGTGGATAAACAGCAATCAAGAGACTTACTGATTTTTTTTATGA
1721 AACTTGTATGATGCTGGCAGGTCGTATCGCAGCTCTATAAAAAACAAAGCCACTAAGGAT
1781 GAAAAATTGACTTCCACACATTTACGACCTTGACTTAAATATGCTAGTACCTTCCTATT
1841 TGTAATTTCTAATAATCTATAGCTACCAATTTGTAATAATCATGATTATCTTTTCAGG
1901 TTTGATTTTAGTAACGGTACTAGTAGCTAATTATCTTCTACTAAGCACTCTCTAAAT
1961 CACAGCTCTCAAAATTTAATGTGCAAAACAAATCACCTTCAAACTCTAATCTGTTCAAATG
2021 CAGATTGTAATTCAGTTTCTCTGGCTGCGAGCTGAATCCTGCAAGCTGCAACAGCTCCCA
2081 AGGGATGGTCAGGCCGATGCTCCATACTTGGCAGCACACTTGGAGTAGCAAGGCCCTGGC
2141 TTGGTACTACTATTGGAAGCTCTTCAAGGCTCCATGTTGCAATATTATATCTTAAACT
2201 AATAAGGAATGGCACTACTGAATAGGAATACAGAAGGTGACAGTTCTAGTTATGTTGTA
2261 AATCCAATCCAAAAAATATCCCCCTTATTGTATATCTTATTTCAGACTTTT
2321 CAGAATTAATTGGCTTAGAGTCAGCTTCTAGTTATATGCTTTTCCCATTTTTCATGCTCTG
2381 GAGTATTATTATGCAATTTACTAATAAAATGTATCGAACATG--poly(A) 2423

Figure 2: Complete nucleotide sequence of the bovine and human SS-B/La cDNA, 5' and 3' flanking regions, and the corresponding amino acid sequences. Upper two lines show deduced amino acid and nucleotide sequences of the bovine SS-B/La cDNA (EMBL accession number X13698). Nucleotide residues in cDNA are numbered in the 5' to 3' direction, beginning with the first residue of the ATG methionine initiator codon. Nucleotide and amino acid differences found in human SS-B/La (EMBL accession number X13697) are displayed beneath the bovine sequence. Hyphens indicate no comparable nucleotide or amino acid residues. The asterisk indicates a stop codon. All methionine residues, putative polyadenylation signals, and RNA-binding protein consensus sequences are underlined. Residues 223–240 of the bovine sequence are identical to a 18-mer sequence derived from amino

of calf thymus and rabbit thymus extracts as previously described (20). The SS-B/La protein was purified via an anti-SS-B/La antibody-affinity column (13), preparative SDS-PAGE, and electroelution (21). The bovine extract was also partially digested with S.a.V8 protease (1 μ g enzyme per 10 mg extract) at 37°C for 30 min. At the end of incubation, protease activity was quenched by the addition of PMSF to 2 mM final concentration and chilled on ice. Protease-generated SS-B/La fragments were also purified via antibody affinity column at 8°C, preparative SDS-PAGE, and electroelution. N-terminal amino acid sequence was determined using an Applied Bio Systems Model 470A protein sequenator (22). The phenylthiohydantoin amino acid residues were fractionated using an Altex Ultrasphere C18 reverse-phase column.

Sequence Analysis

DNA and protein sequences were analysed by the Genetics Computer Group Sequence Analysis Software Package for VAX computers unless otherwise stated. Alignment of protein sequences were initially achieved with the Gap program which uses the algorithm of Needleman and Wunsch (23). Further refinements in sequence alignments were performed manually. Predictions of secondary structures from protein sequences were obtained in the PeptideStructure program which employs both the Chou–Fasman (24) and the Garnier–Osguthorpe–Robson methods (25).

In considering sequence similarities within RNA-binding regions, the Wu and Kabat variability parameter (26) was adapted and an Invariance Index was defined mathematically as the reciprocal of the Wu and Kabat variability for each position within the RNA-binding domain. The Invariance Index was, therefore, calculated as the frequency of the most common amino acid at a given position divided by the number of different amino acids at that position. The frequency of the most common amino acid was defined as the number of times it occurred divided by the total number of sequences examined (26). The consensus hydrophobicity scale of Sweet and Eisenberg (27) was adapted for the analysis of variations in hydrophobicity among RNA-binding domains. The variance of hydrophobicity was calculated as the standard deviation of hydrophobicity at each residue position in the sequence alignment. Helical wheels were used to represent and identify segments with helical potential based on the clustering of hydrophobic and hydrophilic residues on opposite sides of the helical wheel.

RESULTS AND DISCUSSION

SS-B/La cDNA Clones

Bovine cDNA clones were obtained by screening 700,000 recombinants from the bovine pituitary λ gt11 expression library with a human SS-B/La serum. Human cDNA clones were selected from about 800,000 plaques of the MOLT-4 cDNA library by hybridization with a radiolabeled cDNA probe La6 (28). Nine bovine (B1–9) and 14 human (H1–14) clones were initially identified and were subsequently plaque-purified. All 9 bovine clones and 10 of the 14 human clones were further confirmed to be SS-B/La by their immunoreactivity with three different monoclonal anti-SS-B/La antibodies A1, A2, and A5, that were produced via immunization with bovine SS-B/La and were shown to be

acid sequencing of a 14 kDa bovine SS-B/La peptide (shown above deduced sequence). [] and {} bracket human partial SS-B/La sequences reported by Chambers and Keene (28) and Sturgess *et al.* (30), respectively. There are 3 single-nucleotide substitutions and 1 addition in our nucleotide sequence when compared to the corresponding partial sequence in ref 30. *Arrow Heads* point to the locations with nucleotide differences: at position 160 (G to A), 863 (T to A), 1350 (add T), and 1530 (A to C).

Figure 3: Northern blot analysis of mRNA. A: 10 μ g RNA from human HeLa (lane 1), KB (lane 2), A293 (lane 3), and Sal-1 (lane 4) cells were hybridized to the complete human cDNA probe H9. B: 10 μ g RNA from bovine MDBK cells was hybridized to radiolabeled B5 cDNA. A single mRNA of 1700-bp was detected in all human cell lines tested and an additional mRNA of about 2600-bp was also detected in bovine cells. Locations of 18S and 28S RNA are indicated.

specific for both human and bovine SS-B/La (13). Results of restriction enzyme mapping (Figure 1) showed that the 9 bovine clones represented 3 unique cDNAs with inserts of 1300-bp (B6), 1400-bp (B1, B3, B7), and 1800-bp (B2, B4, B5, B8, B9). Clones B6 and B5 together spanned the complete protein coding sequence. cDNA inserts from the human clones were about 1700-bp. Deletions within cDNA inserts in Bluescript vector were made using various restriction enzymes and the complete overlapping sequences were determined from both strands using standard primers (Figure 1). Complete cDNA and deduced amino acid sequences are shown in Figure 2. There is only one large uninterrupted open reading frame in both the human and bovine cDNA sequences. The designated methionine initiation codons are the first ATG codons found in the complete cDNA sequences and their flanking sequences diverge only in 1 or 2 bases from the consensus eukaryotic translation initiator sequence (29). The stop codons at nucleotide 1213 and 1225 of bovine and human SS-B/La cDNA respectively, are followed by about 300-bp of similar 3'-untranslated sequences with an extra 900-bp untranslated sequence in the bovine cDNA. This is consistent with the results of mRNA blotting (Figure 3) which showed one major mRNA form of 1700 bases among several human cell lines tested and an additional mRNA of 2600 bases in bovine cell probably as a result of alternative polyadenylation. A polyadenylation signal AATAAA is found within 30 bases of the 3'-ends of the human sequence and two AATAAAs are present in the corresponding region of the bovine sequence (Figure 2).

Two partial human SS-B/La cDNA sequences have been reported (bracketed in Figure 2 for comparison). The first published sequence by Chambers and Keene (28) was 165 nucleotides long and corresponds to nucleotides 134–291. Sturgess *et al.* (30) reported a second cDNA sequence of 1400-bp which corresponds to nucleotides 160–1530 of our human cDNA sequence. There are 3 single-nucleotide differences and 1 single-nucleotide deletion compared to our sequence as shown in Figure 2 (30). Two of these differences,

Protein	Domain	1	50	95	Ref.					
snRNP 70kDa		DPNAQGDAPK TLFVARVNYD TTESKLRR E FEVYGPQIKRI	HMVYSKRSGCK PRGYAFIEYE HERDMHSAYK	HADGK KIDG RRLVLVDVERG RTVKG	37					
snRNP A	I	NAVPTETPRNH TIYVNNLNK IKKDELKSL YALFSQFQI	LDLILVSRSLK MRQQAIVFK EVSSATNALR	SMQFPF FYD KPMRIQYAKT DSDII	38					
	II	QPLSENPPNH ILFLTNLPEE TNEHMLSM L FNQFPFCFKEV	RLVFCGR HDIAFVEFD	NEQQAAGARD ALQCFKITOM	NAHKISFAKK					
snRNP B*	I	...NDIRPNH TIYVNNLNK IKKDELKSL YALFSQFQI	VDIVALKTHK MRQQAIVFK ELGSSNTALR	QLQCFP FYG KPMRIQYAKT DSDII	39					
	II	PQVPOYPYNY ILFLTNLPEE TNEHMLSM L FNQFPFCFKEV	RLVFCGR HDIAFVEFE	NDQQAAGARD ALQCFKITPS	HANKITYYAKK					
snRNP E		KVQKVKMQPI NLIPRYLQNR SQIQ VMLYBQV	NMRTG CGIFGFNEYT	NLIVIRDAEII HSKTSKRKL	GRIMLKGDNI TLQGS	40				
hnRNP C1		NKTDPRSHMS RVFVGNLNTL VVKKSDVEAI	FSKYGKIVGCC SV	HKGFPAPQVY NERNARAIVA	GEMGRM IAG QVLDINLAAE	PKVNR	41			
hnRNP UP-2		KAMKTEKPVK KIFVGGLSFD TPEEKIRE Y	FGGPGEVESI	ELPMDNKTNK RRGFCFITFN	QEEP		49			
Human	I	ESPKPEQLR KLFIGGLSFE TTDESLSL H	FEQAGTLTDC	VVMRDPTNKR SRGPGFVTYA	TVEEDVAAMN ARPHK . VDG	RVEVPEKAVS	REDSQ	36		
hnRNP A1	II	QKPGAHITVK KIFVGGICKD TEEHILRD Y	FEYQCKLEVI	EINTDRGSKK KRGFAPVTVD	DHDSVDKIVI QKYHT . VNG	HNCVEKRLS	KQEMA			
Drosophila	I	DSITEPEHNR KLFIGGLDVR TTENLKA H	FEKGNIVDV	VVMKDPRTKR SRGPGFITTS	HSKSHDEAQK SBPHK . IDG	RVEVPEKAVP	RQDID	47		
hnRNP A1	II	DSFNAGATVK KLFIGALKDD HDESSIRD Y	FOHEGNIVDI	NIVIDKETCK KRGFAPVEFD	DYDPVKVVL QKQKQ . LMG	KGVVCKALP	KQNDQ			
Rat hnRNP A1	I	ESPKPEQLR KLFIGGLSFE TTDESLSL H	FEQAGTLTDC	VVMRDPTNKR SRGPGFVTYA	TVEEDVAAMN ARPHK . VDG	RVEVPEKAVS	REDSQ	48		
	II	QSPGAHLTVK KIFVGGICKD TEEHILRD Y	FEYQCKLEVI	EINTDRGSKK KRGFAPVTVD	DHDSVDKIVI QKYHT . VNG	HNCVEKRLS	KQEMA			
Human PABP	I	NPSAPSPYNA SLYVGDHLDP VTAAHLYEK	FSPAGPILSI	RVCGRDMITR SLGYAYVNFQ	QPADAERAILD	THNFV IKG	KPVRIHNSQR	DPSLR	44	
	II	DPSLRKSGVG NIFIKNLKDS IDMAKALYNT	FSPAGNILSC	KVVCDEEN G SKGYGVNHF	TQGAERAILD	IKNGML LND	RKVFVGRFES	RKERE		
	III	ELGARAKEFT NVYIKNGED HDEERLKD L	FGPALSVK .	...VMTDE SKC LKGGFVFSFE	RHEDAKAVD	ENQCK ELNG	KELTVYGRAK	KVERE		
	IV	QDRITRYQCV NLYVGNLDDG VDEERLKE	FSPFGTITSA	KVMHEGR . K SKGPGFVCS	SPEEATKAVT	ENQGR I VAT	KELYVALAQK	KEERQ		
Yeast PABP	I	ESQSVENHSA SLYVGDHLDP VSEAHLYD I	FSPIGSVSSI	RVCGRDAITKT SLGYAYVNFN	DHEDAKAIE	QLNYTP IKG	RLCRIMHSQR	DPSLR	45,46	
	II	DPSLRKSGSG NIFIKNLHDP IDMAKALYDT	FSPVGTILSS	KIATDEN CK SKGPGFVHFE	EBGAKAIEID	ALNGML LMG	QELTVYAPHS	RKERE		
	III	QLEETAKHYT NLYVGNHNS TTEDDFE L	PAKPGPIVSA	SLEKDA DCK LKGGFVFSFE	RHEDAKAIE	ALNDS ELNG	KELTVYGRAK	KVERE		
	IV	LEMAKARQCV NLYVGNLDDG VDEERLKE	FAPGTLTISA	KVMHEGR CK SKGPGFVCS	TPEEATKAIT	ENQQL VAG	KELYVALAQK	KVERE		
Mouse	I	...SEPTTF KLFIGGLNFM KSVLEKFA	ISELPAKNDL	AVVDV RTGT NRKPGYVDFE	SAEDLEKALE	LT GL KVPK	NEIKLEKPKG	RDSIK	42	
Nucleolin	I	RDSKVKRAAR TLLAKNLSM ITTEDELKE V	FEDEMEIRLV	SQ DCK SKGLAYIEFK	SEDAEAKNLE	EKQKA EIDG	RSEVLYTYGK	KQBRQ		
	III	KTSVSGESK TLLVSNLSYS ATKETLKE V	FEKATFIK V	PQ ...NPHCK PKGYAFIEFA	SFEDAKAL	NSCRHKEIEG	RTILRELQGS	NSRSQ		
Hamster	I	VEGSESTTF KLFIGGLNFM KSVLEKFA	A ISEFFAKNDL	AVVDV RTGT NRKPGYVDFE	SAEDLEKALE	LT GL KVPK	NEIKLEKPKG	RDSIK	43	
Nucleolin	II	RDSKVKRAAR TLLAKNLSM ITTEDELKE V	FEDEMEIRLV	SQ DCK SKGLAYIEFK	SEDAEAKNLE	EKQKA EIDG	R			
	III	KTSVSGESK TLLVSNLSYS ATKETLKE V	FEKATFIK V	PQ ...NQCK SKGYAFIEFA	SFEDAKAIE	NSCRHKEIEG	RTILRELQGP	RGSP		
	IV	SPHARSQSK TLPVKGSLSD TTETLTKS	FEQSVARIV	T ...DRETS SKGPGFVDN	SEDAEAKAIE	AKHEDGE IDG	NKVTLL DMA	KPEGE		
Human SS-B/La		DEYKNDVKNR SVYIKGPFPT ATLDDIKE W	LEDKGQVLI	QMRRTLHKA	FKGSIPIVVD	SIESAKKFEV	TPQKYKYDT	LILILPKDYTF	AKKHE	
Bovine SS-B/La		DEYKNDVKNR SVYIKGPFPT AALDDIKE W	LEDKGQVLI	QMRRTLHKA	FKGSIPIVVD	SIESAKKFEV	TPQKYKYDT	LILILPKDYTF	TKKHE	
RNP		LFVKKL		KGPGFVXF					33	
CONSENSUS		TYL I		R YAY Y						
SEQUENCE		FVI I								
(Dreyfuss et al.)		VTH								
RNP		LFINKL		KGPAFVXF						
CONSENSUS		IYVGI		R YGYI Y						
SEQUENCE		VVL I		H						
(modified)		L F								
		I								
RNP2				RNP1						

Figure 4: RNA-binding protein domains and consensus sequences. Sequences similar to the RNA-binding protein consensus sequences described by Dreyfuss *et al.* (33) were determined by first scanning sequences visually for Phe and Tyr residues. An X in the consensus sequence designates any amino acid residue. *PABP* is the poly(A)-binding protein. *Underlined* sequences represent sequences that may conform to an amphipathic helix as predicted by helical wheel modeling. RNP1 and RNP2 consensus sequences and the proposed modifications are aligned directly below the composite tabulation of different proteins.

at nucleotides 160 and 863, result in two amino acid differences, Lys to Glu and Asp to Val at positions 54 and 288 respectively, as compared with the previously reported sequence data (30). We have analyzed 7 independent human cDNA clones. No nucleotide changes were observed among these several 5'-end sequences except for a single base substitution in H9 at nucleotide position -62 (C to T) which we suspect may be the result of a reverse transcriptase error. These data suggest that there is only one active SS-B/La gene in MOLT-4 cells.

cDNA Identity Confirmed by Independent Amino Acid Sequence

N-terminal amino acid sequencing of the bovine 46 kDa SS-B/La protein could not be successfully performed suggesting a blocked N-terminal residue. A major autologous degradation product of bovine SS-B/La was a 43 kDa polypeptide (20) and this was also found to contain an N-terminal blocked residue. Fragments were therefore generated from bovine SS-B/La by *S. aureus* V8 protease, purified via anti-SS-B/La affinity column and a 14 kDa immunoreactive peptide was found suitable for amino acid sequence analysis. The 18 amino acid residues obtained from N-terminal analysis of the 14 kDa peptide matched the deduced protein sequence derived from cDNA cloning at residues 223-240 (Figure 2). This provided further evidence for the identity of these clones as SS-B/La cDNAs in addition to their reactivity with monoclonal antibodies A1, A2, and A5.

SS-B/La Domain Structure and Sequence Analysis

We have previously identified 2 protease-resistant domains of SS-B/La from HeLa cells designated X and Y domains corresponding to 28 kDa and 23 kDa respectively (20). Recently, this finding has been confirmed by Sturgess *et.al.* (30) and Bach & Palau (31). Domain X was rich in methionine but lacked phosphorylated amino acid residues whereas domain Y contained all of the detectable phosphorylated amino acids (20). In the sequences reported here, there are a total of 6 and 7 methionine residues in human and bovine SS-B/La protein respectively and they are located entirely in the N-terminal half of the protein. It is now possible to assign domains X and Y to the N- and C-terminal portions of the protein respectively.

The deduced bovine and human SS-B/La proteins contain 404 and 408 amino acids and have calculated molecular masses of 46,482 Da and 46,784 Da respectively. Comparison of the two complete protein sequences reveals a high degree of conservation of amino acid sequence showing only 23 substitutions. Sixteen amino acid substitutions (70%) and all deletions reside in the C-terminal half of the bovine SS-B/La when compared to human while there is only 1 substitution (4%) within the N-terminal 25% of the molecule. The difference in the number of amino acids is not surprising since we have shown that human HeLa SS-B/La migrated slower than bovine SS-B/La in SDS-PAGE gels and apparent molecular masses of 48 kDa and 46 kDa, respectively were assigned (20). Three of the four extra amino acids in the human sequence are contained in the tripeptide Asp-Glu-His which is located within a repeating sequence of Asp-Glu-His-Asp-Glu-His-Asp-Glu at residues 368–375. The other extra amino acid, Glu, is also found in a direct repeat of Glu-Glu at residues 392–393. These duplicated sequences are present in the C-terminal 10% of the protein and are within the so-called 'PEST' regions which have been suggested to be common to rapidly degraded proteins (32). PEST regions have been defined as sequences rich in Pro, Glu (or E), Ser, Thr, and to a lesser extent Asp residues, and are found to begin and end with positively charged residues (32). PEST regions in SS-B/La may explain the previous finding that SS-B/La is fairly sensitive to endogenous proteases in analytical purification from tissue and cell extracts (20). For example, two PEST regions close to the C-terminus (residues 383–388 and 388–393 in the bovine sequence and residues 386–391 and 391–397 in the human sequence) may account for the frequently observed major degradation fragment of about 43 kDa (20) and PEST regions at residues 216–229 may explain the protease sensitivity previously observed in the generation of domains X and Y (20). The mapping of the residue at the N-terminus of the bovine 14 kDa peptide to residue 223 agrees with the hypothesis that the PEST region at residue 216–229 of SS-B/La is indeed sensitive to protease and if the 14 kDa peptide is a N-terminal fragment of domain Y, the boundary between domains X and Y will be around residues 216–229.

RNA-Binding Protein Consensus Sequences

A recent report from Dreyfuss *et.al.* (33) described 2 RNA-binding protein consensus sequences called RNP1 and RNP2 which were stretches of 8 and 6 amino acids respectively with RNP1 located about 30 amino acids downstream of RNP2. These consensus elements contained one or more residues of Phe or Tyr. Recently Merrill *et.al.* (34) showed that in experiments where RNA and the A1 hnRNP protein were crosslinked by *in vitro* UV-irradiation, four Phe residues were identified as the sites of covalent adduct formation. Two RNP1 and 2 RNP2 sequences were identified in the A1 hnRNP protein and the 4 Phe residues in question were located within the RNP consensus sequences (1 Phe per consensus). Taken together these findings suggested that RNP1 and RNP2 consensus

Figure 5: Invariance at different amino acid positions within the RNA-binding domain. The Invariance index is defined as the mathematical reciprocal of the variability parameter defined by Wu and Kabat (26).

sequences found in RNA-binding proteins might participate directly in the functional interactions with RNA.

A single copy of each RNP1 and RNP2 consensus-like sequence is found within the N-terminal half of SS-B/La. This supports the previous finding that UV-irradiation crosslinked RNA to domain X alone (35). The RNP1 and RNP2-like sequences of SS-B/La do not fit perfectly with the consensus proposed (Figure 4) and within the RNP1-like sequence of SS-B/La there is an amino acid substitution between the human and bovine protein (Figure 2). To analyze these findings further, we have examined published sequences in several RNA-binding proteins and snRNP proteins (36–49). With the current understanding of RNA-protein interaction as discussed above, we identified RNP1 and RNP2-like sequences in putative RNA-binding proteins by searching for regions rich in Phe and Tyr. In both the A and B'' snRNP proteins, we could locate two sets of RNP1 and RNP2-like sequences instead of just one identified previously (38) and thus there may be two RNA-binding domains within these two proteins. The two sets of RNP1 and RNP2-like sequences of these two proteins are displayed in Figure 4.

Based on some of the previous sequence alignments from reference 33 and additional sequences identified in other RNA-binding proteins, 95-residue segments were aligned as described in *Materials and Methods* and listed in Figure 4. The general sequence alignment in Figure 4 correlated well with the previous alignment (33) with the noted exception in aligning the human poly(A) binding protein sequence at the region corresponding to the RNP2 sequence. The previous alignment resulted in the inclusion of Thr and Asn in the RNP2 consensus sequence. Based on the alignment as shown in Figure 4, we propose two modified RNP1 and RNP2 consensus sequences. The modified consensus sequences are defined by including only residues that occur 4 or more times in a particular position and their closely related conservative substitutions.

Invariance Index and Hydrophobicity

To evaluate the constant and variable residues within these RNA-binding sequences, the method of Wu and Kabat (26) originally described for studies of immunoglobulin antigen-binding sequences was employed. A modification was made by defining an Invariance Index as the reciprocal of the Wu and Kabat variability parameter for each amino acid position in the sequence. Plotting Invariance Index against position for the 95 residues

Figure 6: *A. Average hydrophobicity and variance in hydrophobicity at different amino acid positions within the RNA binding domain. Average hydrophobicity is defined as the average of hydrophobicity values assigned to amino acids by Sweet and Eisenberg (27). Variance is calculated as the standard deviation of hydrophobicity values at each position. Arrows illustrated positions with high average hydrophobicity and low variance and together may predict a segment of amphipathic helix. B. Helical wheel representation of residues 124 to 139 of the SS-B/La sequence. This diagram illustrates the amphipathic distribution of amino acid residues within the putative helical structure. The display sequence is located in between the two RNP consensus sequences in the RNA binding domain of SS-B/La.*

of the RNA-binding domain (Figure 5) highlights positions that are relatively invariant. An absolutely invariant residue would have an index of 1 by definition. The most conserved residues are still the two aromatic residues located at 56th and 59th residues corresponding to the fifth and eighth positions of RNP1 consensus respectively. An additional relatively invariant region coincided with the RNP2 consensus sequence and also at positions 26 and 31. This analysis confirms the existence of the RNP1 and RNP2 consensus elements and it was observed that many of the less variant positions had hydrophobic residues.

Based on the consensus hydrophobicity scale of Sweet and Eisenberg (27), the average hydrophobicity and variance at different positions of the RNA-binding domain is shown in Figure 6A. This analysis essentially rounds off the actual differences among hydrophobic residues in conserved positions and provides an overall view of the hydrophobic structure of the RNA-binding domain. High average hydrophobicity value at a given position generally corresponds to relative low variance. The point is clearly shown in the RNP1 and RNP2 sequences where the most hydrophobic residues are highly conserved. The overall

conserved structure of these RNA-binding protein domains may depend on the hydrophobic interactions among different elements within the domain. The segment including positions 31, 34, 37, and 40 (marked by arrows, Figure 6A) shows a conserved pattern of hydrophobic residues separated by hydrophilic residues and appears to be a characteristic for a conserved amphipathic helical structure. Using the helical wheel model, the SS-B/La sequence from residues 124 to 139 (equivalent to positions 23 to 39 of the RNA-binding segment) is depicted as an amphipathic α -helix in Figure 6B. Individual sequences spanning positions 31 to 40 are fitted into amphipathic helices with some helices being longer than others; these putative amphipathic helical sequences are underlined in Figure 4. Since they are quite variable in sequence (Figure 5) and are strategically located between the 2 RNP consensus elements, these putative amphipathic helices might structurally delineate the specificities of the RNA interaction. The authenticity of the helical structures clearly requires experimental verification, nevertheless the similarities as analysed further confirms these proteins as members of a distinct family of RNA-binding proteins.

While this manuscript was being reviewed for publication, Chambers *et al.* (50) reported the genomic structure and the complete cDNA sequence of human SS-B/La. No difference was found between this coding sequence of human SS-B/La and the one reported here. A comparison of the two sequences flanking the initiator methionine showed a single substitution of C to T at the -2 position. A putative G/C-rich promotor was described upstream of the SS-B/La mRNA start site (50). A total of 11 exons were found in the human SS-B/La gene with the RNP1 consensus sequence split into two exons by the fifth intron (50). We would like to comment that this feature is also present in mouse nucleolin, a nucleolar protein consisting of 4 similar RNA-binding domains, in which each of the 4 RNP1 consensus sequences was also split by an intron (42). In the human SS-B/La gene, the putative amphipathic helix discussed above was found within a single exon (exon 5) consisting of 36 amino acid residues (50).

ACKNOWLEDGEMENTS

We are grateful to Kevin Ferreri and Dr. T.E.Hugli for the determinations of amino acid sequences. We thank Drs. D.A.Carson and T.Nobori for access to the bovine pituitary cDNA library and Drs. J.D.Keene and J.C.Chambers for providing a useful DNA probe and discussion of their work prior to publication. We acknowledge Dr. John P.Huff for generous help during the initial cDNA library screening and Dr. K.Mike Pollard for informing us of the PEST Hypothesis. This is publication 5396BCR from the Research Institute of Scripps Clinic. This work is supported by Grants AM32063 and AI10386 from the National Institutes of Health. E.K.L.C. is a recipient of an Arthritis Foundation Postdoctoral Fellowship. K.F.S. is a recipient of an Arthritis Foundation Investigator Award.

REFERENCES

1. Lerner, M.R., and Steitz, J.A. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 5495–5499.
2. Rinke, J., and Steitz, J.A. (1982) *Cell* **29**, 149–159.
3. Francoeur, A.M. and Mathews, M.B. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 6772–6776.
4. Reddy, R., Henning, D., Tan, E.M., and Busch, H. (1983) *J. Biol. Chem.* **258**, 8352–8356.
5. Chambers, J.C., Kurilla, M.G., and Keene, J.D. (1983) *J. Biol. Chem.* **258**, 11438–11441.
6. Mathews, M.B. and Francoeur, A.M. (1984) *Mol. Cell. Biol.* **4**, 1134–1140.
7. Stefano, J.E. (1984) *Cell* **36**, 145–154.
8. Rinke, J. and Steitz, J.A. (1985) *Nucleic Acids Res.* **13**, 2617–2629.
9. Madore, S.J., Wieben, E.D., and Pederson, T. (1984) *J. Biol. Chem.* **259**, 1929–1933.

10. Wilusz, J., Kurilla, M.G., and Keene, J.D. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 5827–5831.
11. Alspaugh, M.A., Talal, N., and Tan, E.M. (1976) *Arthritis Rheum.* **19**, 216–222.
12. Mattioli, M., and Reichlin, M. (1974) *Arthritis Rheum.* **17**, 421–429.
13. Chan, E.K.L., and Tan, E.M. (1987) *J. Exp. Med.* **166**, 1627–1640.
14. Young, R.A., and Davis, R.W. (1983) *Science* **222**, 778–782.
15. Feinberg, A.P., and Vogelstein, B. (1983) *Anal. Biochem.* **132**, 6–13.
16. Chen, E.Y., and Seeburg, P.H. (1985) *DNA* **4**, 165–170.
17. Sanger, F., Coulson, A., Barrell, B., Smith, J., and Roe, B. (1980) *J. Mol. Biol.* **143**, 161–169.
18. Fox, R.I., Bumol, T., Fantozzi, R., Bone, R., and Schreiber, R. (1986) *Arthritis Rheum.* **29**, 1105–1111.
19. Chirgwin, J., Przybyla, A., MacDonald, R., and Rutter, W. (1979) *Biochemistry* **18**, 5294–5299.
20. Chan, E.K.L., Francoeur, A.-M., and Tan, E.M. (1986) *J. Immunol.* **136**, 3744–3749.
21. Hunkapillar, M.W., Lujan, E., Ostrander, F., and Hood, L.E. (1983) *Methods Enzymol.* **91**, 227–236.
22. Hewick, R.M., Hunkapillar, M.W., Hood, L.E., and Dreyer, W.J. (1981) *J. Biol. Chem.* **256**, 7990–7997.
23. Needleman, S.B., and Wunsch, C.D. (1970) *J. Mol. Biol.* **48**, 443–453.
24. Chou, P.Y., and Fasman, G.D. (1978) *Adv. Enzymol.* **47**, 45–147.
25. Garnier, J., Osguthorpe, D.J., and Robson, B. (1978) *J. Mol. Biol.* **120**, 97–120.
26. Wu, T.T. and Kabat, E.A. (1970) *J. Exp. Med.* **132**, 211–250.
27. Sweet, R.M. and Eisenberg, D. (1983) *J. Mol. Biol.* **171**, 479–488.
28. Chambers, J.C., and Keene, J.D. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 2115–2119.
29. Kozak, M. (1987) *Nucleic Acid Res.* **15**, 8125–8148.
30. Sturgess, A.D., Peterson, M.G., McNeilage, L.J., Whittingham, S., and Coppel, R.L. (1988) *J. Immunol.* **140**, 3212–3218.
31. Bach, M., and Palau, J. (1987) *Eur. J. Biochem.* **165**, 117–123.
32. Rogers, S., Wells, R., and Rechsteiner, M. (1986) *Science* **234**, 364–368.
33. Dreyfuss, G., Swanson, M.S., and Pinol-Roma, S. (1988) *TIBS* **13**, 86–91.
34. Merrill, B.M., Stone, K.L., Cobianchi, F., Wilson, S.H., and Williams, K.R. (1988) *J. Biol. Chem.* **263**, 3307–3313.
35. Chan, E.K.L., and Tan, E.M. (1987) *Mol. Cell. Biol.* **7**, 2588–2591.
36. Buvoli, M., Biamonti, G., Tsoulfas, P., Bassi, M.T., Ghetti, A., Riva, S., and Morandi, C. (1988) *Nucleic Acids Res.* **16**, 3751–3770.
37. Theissen, H., Etzerodt, M., Reuter, R., Schneider, C., Lottspeich, F., Argos, P., Luhrmann, R., and Philipson, L. (1986) *EMBO J.* **5**, 3209–3217.
38. Sillekens, P.T.G., Habets, W.J., Beijer, R.P., and van Venrooij, W.J. (1987) *EMBO J.* **6**, 3841–3848.
39. Habets, W.J., Sillekens, P.T.G., Hoet, M.H., Schalken, J.A., Roebroek, A.J. M., Leunissen, J.A.M., van den Ven, W.J.M., and van Venrooij, W.J. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 2421–2425.
40. Wieben, E.D., Rohleder, A.M., Nenninger, J.M., and Pederson, T. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 7914–7918.
41. Swanson, M.S., Nakagawa, T., LeVan, K., and Dreyfuss, G. (1987) *Mol. Cell. Biol.* **7**, 1731–1739.
42. Bourbon, H.-M., Lapeyre, B., and Amalric, F. (1988) *J. Mol. Biol.* **200**, 627–638.
43. Bugler, B., Bourbon, H., Lapeyre, B., Wallace, M.O., Chang, J.-H., Amalric, F., and Olson, M.O. (1987) *J. Biol. Chem.* **262**, 10922–10925.
44. Grange, T., Martins de Sa, C., Oddos, J., and Pictet, R. (1987) *Nucleic Acids Res.* **15**, 4771–4787.
45. Adam, S.A., Nakagawa, T., Swanson, M.S., Woodruff, T.K., and Dreyfuss, G. (1986) *Mol. Cell. Biol.* **6**, 2932–2943.
46. Sachs, A.B., Bond, M.W., and Kornberg, R.D. (1986) *Cell* **45**, 827–835.
47. Haynes, S.R., Rebbert, M.L., Mozer, B.A., Forquignon, F., and Dawid, I.B. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 1819–1823.
48. Cobianchi, F., SenGupta, D.N., Zmudzka, B.Z., and Wilson, S.H. (1986) *J. Biol. Chem.* **261**, 3536–3543.
49. Lahiri, D.K., and Thomas, J.O. (1986) *Nucleic Acids Res.* **14**, 4077–4094.
50. Chambers, J.C., Kenan, D., Martin, B.J., and Keene, J.D. (1988) *J. Biol. Chem.* **263**, 18043–18051.

This article, submitted on disc, has been automatically converted into this typeset format by the publisher.