569

Structure of the human sialophorin (CD43) gene

Identification of features atypical of genes encoding integral membrane proteins

C. Simon SHELLEY,*†‖ Eileen REMOLD-O'DONNELL,*†§ Fred S. ROSEN*†‡
and Alexander S. WHITEHEAD†‡¶
*Center for Blood Research, †Division of Immunology, Children's Hospital, Departments of ‡Pediatrics and
§Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, U.S.A.

A human sialophorin (CD43) specific genomic clone was isolated, and a 6.5 kb fragment containing the 4.6 kb
sialophorin gene was sequenced. The promoter region contains no TATA or CAAT boxes, but is highly enriched in G
and C nucleotides and contains short repeat sequences similar to those found in the promoters of 'housekeeping' genes.
S1-nuclease protection and primer-extension experiments established that the sialophorin gene has two major
transcription initiation sites. There is a single intron of 378 bp that interrupts the sequence specifying the mRNA 5'
untranslated region. The gene is therefore unusual in that the discrete extracellular, transmembrane and intracellular
regions of the protein, including repeat sequences in the extracellular region, are not encoded by separate exons.
Utilization of alternative polyadenylation signals was previously shown to generate two sialophorin mRNAs of 1.9 and
4.3 kb, which differ in the length of their 3' untranslated regions. Sequence analysis of the gene establishes that a single
polyadenylation signal 2301 bp downstream of the first major transcription initiation site and five overlapping
polyadenylation signals beginning a further 2290 bp downstream define the 3' termini of the 1.9 and 4.3 kb mRNA species
respectively. The gene contains potential Z-DNA structures, *Alu* sequences, and elements that may be involved in
regulating mRNA stability.

INTRODUCTION

Sialophorin, which has also been called gpL115 [1,2], large
sialoglycoprotein [3], leukosialin [4,5], CD43 [6] in human and
W3/13 antigen in rat [7], is a heavily glycosylated surface
molecule on T-lymphocytes and thymocytes, some B-
lymphocytes, monocytes, neutrophils and platelets [8,9].
Sialophorin is phenotypically defective on lymphocytes in the
inherited immunodeficiency Wiskott–Aldrich syndrome [1,2]. It
has been postulated that the maintenance of lymphocytes in
the circulation [9], as well as their surface morphology [10],
is dependent on the presence of intact negatively charged
sialophorin molecules. Sialophorin functions *in vitro* as the cell-
surface component of a T-cell activation pathway [3,8,11], which
is independent of the T-cell receptor–CD3 complex [12]. Anti-
bodies to sialophorin also activate monocytes [13] and increase
the activity of NK cells [14].

Sialophorin is a transmembrane protein; the overall com-
position of the molecule purified from lymphoid cells is approx.
60% carbohydrate/40% protein [7,15]. The extracellular region
of lymphoid sialophorin contains 80–90 mucin-type carbohydrate
units, which were characterized as primarily sialylated Gal-
GalNAc units [2,7,16]. Studies with phorbol 12-myristate 13-
acetate indicate that the intracellular region of sialophorin is
phosphorylated by protein kinase C [17]. Cell-specific and stage-
specific glycosylation patterns give rise to sialophorin species
with different apparent molecular masses; co-electrophoresis [4]
and peptide-mapping [18] studies indicate that the different
species share a common polypeptide core.

Human [19,20] and rat [5] sialophorin cDNA clones have been
isolated and sequenced. The human cDNA sequences predict a
primary translation product of 400 amino acids comprising an

apparent 19-residue leader peptide and a mature polypeptide of
381 amino acids (M_r 38444), which includes a single 23-residue
transmembrane region [19,20]. The 235-residue *N*-terminal
extracellular sequence contains five tandemly arranged 18-amino-
acid repeats proximal to the membrane, and is extremely rich
in serine, threonine and proline residues, consistent with the
attachment of 80–90 *O*-linked carbohydrate units. The *C*-
terminal intracellular region comprises 123 residues and contains
several potential phosphorylation sites. Pallant *et al.* [19] have
estimated the sialophorin mRNAs to be 2.3 and 8.0 kb.

Sialophorin is encoded by two mRNAs which are both present
in all lymphoid cell lines analysed to date [19,20]. Our previous
studies indicate that the sialophorin mRNA species are 1.9 and
4.3 kb; the size difference arises primarily from utilization of
different polyadenylation signal sequences [20].

Southern-blot analysis has established that sialophorin is
encoded by a single gene on chromosome 16 [20]; hybridization
in situ has further localized the gene to band p11.2 [19]. There-
fore, although sialophorin is phenotypically defective on T-
lymphocytes in the Wiskott–Aldrich syndrome, its gene cannot
be the causative locus, since the disease is X-chromosome-linked.

Here we report the isolation and nucleotide sequence of the
human sialophorin gene and its immediate flanking regions.
Several structural and potential control elements that are not
typical of genes encoding integral membrane proteins have been
identified.

MATERIALS AND METHODS

Isolation and analysis of the sialophorin gene

An EMBL-3 human genomic library constructed from a partial
endonuclease-*Sau*3AI digest of human lymphoblast DNA (gift

from Dr. S. Orkin, Children's Hospital, Boston, MA, U.S.A.) was screened in duplicate with a human sialophorin cDNA probe (CEM-E/N; [20]) as described by Benton & Davis [21]. One positive clone λS20, was identified among 10^6 plaques, isolated and grown in bulk for DNA preparation [22].

Restriction-endonuclease mapping

DNA prepared from human peripheral-blood lymphocytes [23] and cloned λS20 DNA were digested with various restriction endonucleases (New England Biolabs, Beverly, MA, U.S.A.). DNA fragments were subjected to agarose-gel electrophoresis and transferred to nitrocellulose (Schleicher and Schuell, Keene, NH, U.S.A.) by the method of Southern [24]. Prehybridization, hybridization of the cDNA fragment CEM-E/N [20] and washing of filters were as described by Maniatis et al. [22].

Subcloning of phage genomic DNA

A 6.5 kb HindIII fragment of λS20 containing the sialophorin gene was isolated by gel electrophoresis followed by electroelution using a Model-UEA electroeluter (International Biotechnologies, New Haven, CT, U.S.A.). This fragment was ligated with the plasmid vector pAA-PZ618 (Gold Biotechnology, St. Louis, MO, U.S.A.) that had been digested with HindIII and treated with calf intestinal phosphatase (Boehringer Mannheim Biochemicals, Indianapolis, IN, U.S.A.). The ligated products were used to transform Escherichia coli strain DH5α (Bethesda Research Laboratories, Gaithersburg, MD, U.S.A.), and bacterial colonies containing recombinant plasmids were identified by restriction-endonuclease mapping of isolated plasmid DNA. One recombinant plasmid, pSH6.5, was prepared in bulk [22]. Subfragments of the pSH6.5 insert, specifically PvuII fragments of 2.1 and 2.0 kb, PvuII–BamHI fragments of 1.1 and 0.8 kb and PvuII–BstXI fragments of 1.0 and 0.9 kb, were cloned into pAA-PZ618 to generate the plasmids pSP2.1, pSP2.0, pSPB1.1, pSPB0.8, pSPBX1.0 and pSPBX0.9 respectively.

DNA sequencing

Double-stranded DNA was sequenced by the dideoxynucleotide-chain-termination/extension method [25] using Sequenase (United States Biochemical, Cleveland, OH, U.S.A.). The extreme 5' and 3' ends of the inserts of pSH6.5, pSP2.1, pSP2.0, pSPB1.1, pSPB0.8, pSPBX1.0 and pSPBX0.9 were sequenced using the reverse and universal M13 sequencing primers (New England Biolabs) respectively. A series of specific oligonucleotide primers based on the sialophorin cDNA sequence [20], and the sialophorin gene sequence as it became known, were synthesized on an Applied Biosystems 380B DNA synthesizer and used to sequence the remaining regions of the insert of pSH6.5. The entire nucleotide sequence of the insert of pSH6.5 was determined on both DNA strands.

S1-nuclease protection experiments

The transcription-initiation sites of the sialophorin gene were identified by S1-nuclease protection using a 164 bp PvuII–MspI fragment spanning the first exon and 5' flanking sequence. The fragment was isolated by polyacrylamide-gel electrophoresis and passively eluted. The 5' end of the non-coding strand was radiolabelled using polynucleotide kinase (New England Biolabs) and [γ-^{32}P]ATP (Amersham Corp., Arlington Heights, IL, U.S.A.). The 3' end of the sialophorin gene was analysed for introns using a 2.4 kb Taq1 fragment. This fragment was purified by agarose-gel electrophoresis, followed by electroelution, and the 3' end of the non-coding strand was radiolabelled using the large (Klenow) fragment of DNA polymerase I (New England Biolabs) and [α-^{32}P]dCTP (Amersham Corp.). The S1-nuclease protection experiments were performed as described by Berk &

Sharp [26]. Briefly, total RNA was isolated [27] from the human lymphoblastoid cell line CEM. A 20 μg sample of polyadenylated [poly(A)$^+$] RNA was purified by oligo(dT) chromatography [28] and ethanol-precipitated in the presence of 1000 c.p.m. of radiolabelled DNA probe. The resulting RNA/DNA pellet was dried, dissolved in 30 μl of hybridization buffer [80% (v/v) formamide/40 mM-Pipes buffer (pH 6.4)/400 mM-NaCl/1 mM-EDTA], denatured at 85 °C for 10 min and annealed at 45 °C for 24 h. After annealing, 300 μl of ice-cold S1-nuclease buffer [50 mM-sodium acetate buffer (pH 4.6)/250 mM-NaCl/4.5 mM-ZnSO$_4$/herring sperm DNA (20 μg/ml)], containing 50 units of S1-nuclease (Bethesda Research Laboratories), were added. The reaction mixtures were incubated for 1 h at 30 °C before the addition of 100 μl of S1-nuclease stop solution [15 mM- EDTA/ yeast tRNA (0.3 mg/ml)] and ethanol precipitation. After a 70% (v/v) ethanol rinse, the DNA/RNA product was resuspended in 10 μl of 'formamide dye' (95% formamide/20 mM-EDTA/0.05% Xylene Cyanol FF/0.05% Bromophenol Blue), heated at 85 °C for 10 min, and electrophoresed through a denaturing polyacrylamide gel.

Primer-extension experiments

The sites of transcription initiation were determined by primer extension using an 18-nucleotide primer (5'-CGGCTCGGC-TGATAAGGA-3') complementary to the sequence at the extreme 3' end of exon 1. The 5' nucleotide of this primer corresponded to the labelled 5' nucleotide of the non-coding strand of the PvuII–MspI fragment used to determine the transcription initiation sites by S1-nuclease protection. The primer was 5'-end-labelled by using polynucleotide kinase and [γ-^{32}P]ATP, and 1000 c.p.m. of radiolabelled primer was ethanol-precipitated in the presence of 20 μg of CEM-cell poly(A)$^+$ RNA. The primer/RNA mixture was dissolved in 25 μl of 10 mM-Pipes buffer, (pH 6.4)/400 mM-NaCl, denatured at 85 °C for 10 min, slowly cooled to 42 °C and hybridized at that temperature overnight. After hybridization, 60 μl of ice-cold reverse-transcriptase buffer [50 mM-Tris/HCl buffer (pH 8.2)/10 mM-dithiothreitol/6 mM-MgCl$_2$/actinomycin D (25 μg/ml)/0.5 mM-dATP + dCTP + dGTP + dTTP], and 5 units of reverse transcriptase (Bethesda Research Laboratories) was added on ice and incubated at 42 °C for 1 h. RNAase (20 μg in 2 μl) was added for 10 min at 42 °C before addition of 133 μl of water. The product of primer extension was extracted once with 200 μl of phenol/chloroform (1:1, v/v) and three times with diethyl ether before ethanol precipitation. The product was dried, resuspended in 10 μl of 'formamide dye', heated at 85 °C for 10 min, and co-electrophoresed through a denaturing 8%-polyacrylamide gel with genomic DNA sequencing reactions generated by using the same oligonucleotide as was used to generate the primed product.

Computer analysis

Sequence analyses were carried out using the Bionet Resource (funded by National Institutes of Health Grant P41RR01685) accessed via PCGENE software (Intelligenetics, Mountain View, CA, U.S.A.). Sequence comparisons were by the method of Pearson & Lipman [29].

RESULTS AND DISCUSSION

Isolation of the human sialophorin (CD43) gene

A human genomic library was screened with an EcoRI–NcoI fragment containing part of the coding region of the human sialophorin cDNA clone CEM1.7 [20]. Approx. 10^6 recombinants were screened, and one, namely λS20, hybridized strongly to the sialophorin cDNA probe. Southern-blot analysis of λS20
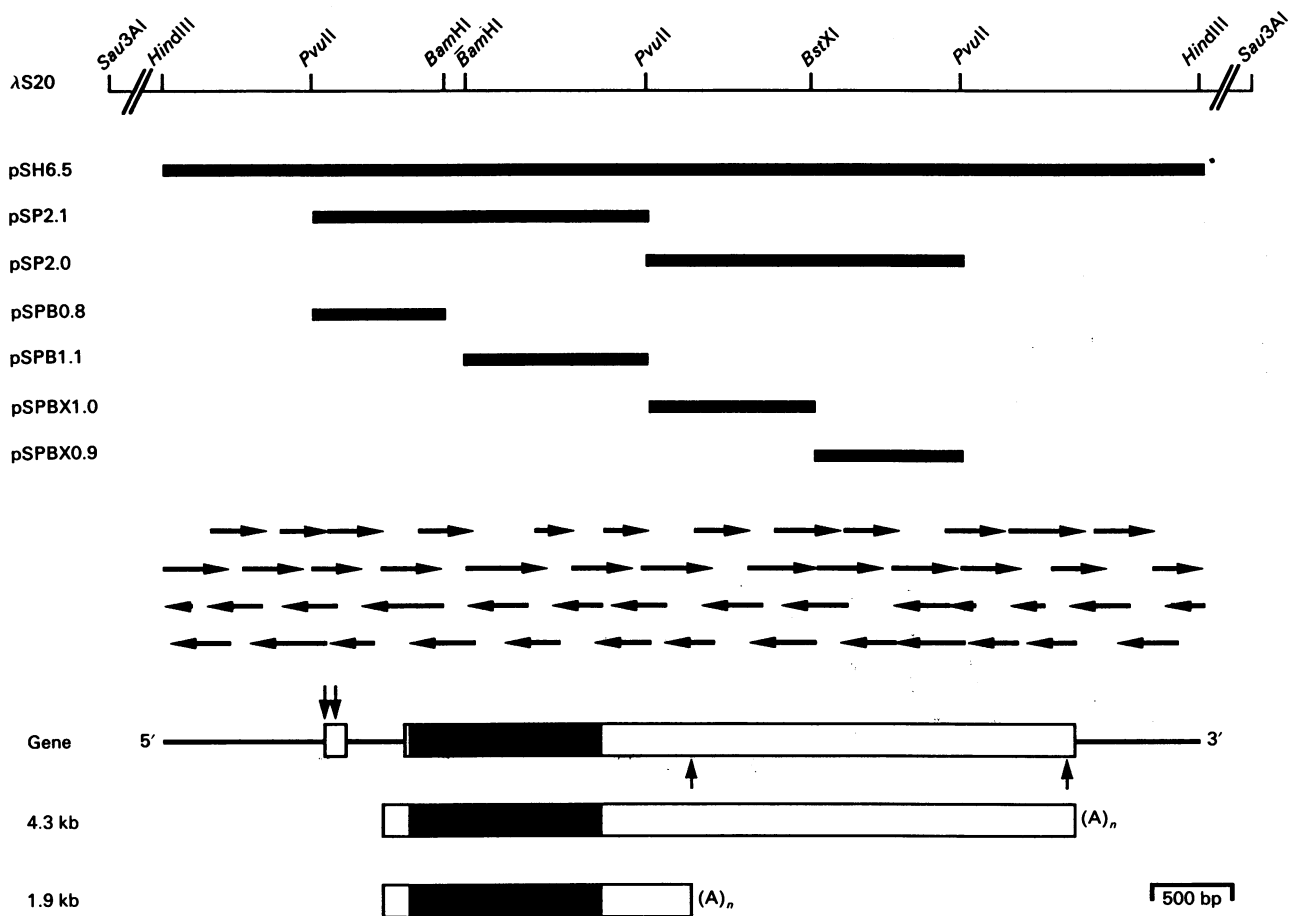
Fig. 1. Partial restriction map, sequencing strategy and organization of the human sialophorin gene

The partial restriction map of the insert of the EMBL-3 phage clone λS20 shows only those sites used in subcloning fragments of the sialophorin gene. The subcloned fragments are indicated below the restriction map. The strategy used in sequencing is depicted by arrows denoting the length and direction of individual sequencing reactions. The organization of the sialophorin gene is represented together with its relationship to the mature sialophorin transcripts. ■ and ☐ represent respectively the coding and non-coding regions of the exons or the transcripts they specify. The intron and flanking sequences are shown by thick black horizontal lines. Above exon 1 at the 5′ end of the gene, the arrows mark the major sites of transcription initiation. Below exon 2 the two arrows mark the polyadenylation signal sequences. The 4.3 kb and 1.9 kb sialophorin mRNA species generated by the differential use of the polyadenylation signal sequences are aligned below the gene. The 5′ end of each mRNA species is drawn arbitrarily using the position of the 5′ major transcription initiation site; the true relationship of the mRNA species with particular transcription initiation sites is not known.

digested with a number of restriction endonucleases generated sialophorin-specific hybridization patterns comparable with the simple patterns observed in similar analysis of human chromosomal DNA (results not shown), indicating that λS20 contains the single authentic human sialophorin gene.

A 6.5 kb HindIII fragment of λS20 containing the entire sialophorin gene was subcloned into pAA-PZ618 for detailed restriction mapping and sequence determination. The strategy used to sequence the 6.5 kb HindIII fragment is illustrated in Fig. 1; the sequence is shown in Fig. 2. As described below, the gene is approx. 4.6 kb and comprises two exons and one intron.

Transcription initiation sites and first exon

Primer extension and S1-nuclease protection experiments were performed to identify the sites at which sialophorin-gene transcription is initiated. Both analyses yielded the same two major products (Fig. 3b, lanes 3 and 4). These findings establish that transcription is initiated principally at two sites (indicated by arrows in Figs. 1 and 3), the thymidine at nucleotide position 1 and the adenosine at nucleotide position 56, which give rise (respectively) to 125 and 70 nucleotides of 5′ untranslated region

specified by exon 1. Upstream of each major product are several minor products, possibly representing minor transcription initiation sites. Whether the two major transcription initiation sites are used differentially to define the specific 5′ termini of the 1.9 and 4.3 kb transcripts is unknown.

Intron

The sialophorin gene contains a single intron of 378 bp that interrupts the sequence specifying the mRNA 5′ untranslated region. Sequences at its 5′ and 3′ ends are in general agreement with the consensus sequences derived from equivalent positions in other eukaryotic genes [30].

Second exon

The second exon is approx. 4.1 kb long and specifies 34 nucleotides of the 5′ untranslated region and the entire coding and 3′ untranslated regions. A single AATAAA polyadenylation signal sequence beginning at nucleotide 2302 (1799 bp from the 5′ end of the second exon) is used to specify the 3′ terminus of the 1.9 kb sialophorin mRNA [20], whereas five overlapping polyadenylation signal sequences between nucleotides 4592 and

The following position markers appear along the sequence: -832, -673, -514, -355, -196, -37, 123, 282, 441, 600, 759, 918, 1077, 1236, 1395, 1554, 1713, 1872, 2031, 2190, 2349, 2508, 2667

Predicted amino acid translation (single-letter code):

M A T L L L L G V L L V V S P D A L G S T
T A V Q T P T S G E P L V S T S E P L S S K M Y T T S I T S D P K A D S T G D Q T S A L P P S T S I N E G
S P L W T S I G A S T G S P L P E P T T Y Q E V S I K M S S V P Q E T P H A T S H P A V P I T A N S L G S
H T V G G T I T T N S P E T S S R T S G A P V T T A A S S L E T S R G T S G P P L T M A T V S L E T S K
G T S G P P P V T M A T D S L E T S T G T T G P P V T M T T G S L E P S S G A S G P Q V S S V K L S T M M S
P T T S T N A S T V P F R N P D E N S R G M L P V A V L V A L L A V I V L V A L L L W R R R Q K R R T G
A L V L S R G G K R N G V V D A W A G P A Q V P E E G A V T V V G G S G G D K G S G F P D G E G S R R
P T L T T F F G R R K S R Q G S L A M E E L K S G S G P S L K G E E E P L V A S E D G A V D A P A P D E P
E G G D G A A A P

Positions (right margin, top to bottom): 2826, 2985, 3144, 3303, 3462, 3621, 3780, 3939, 4098, 4257, 4416, 4575, 4734, 4893, 5052, 5211, 5370, 5513

Fig. 2. Nucleotide sequence of the human sialophorin gene

The exons are underlined. The derived sialophorin amino acid sequence is depicted in single-letter code above the coding portion of the gene; the putative leader peptide and transmembrane region are in italics. Arrows mark the major transcription initiation sites, the first of which is designated nucleotide +1. An inverted repeat between the transcription initiation sites is marked with converging arrows. The potential AP-2 binding site is indicated by dots above the nucleotides. Polyadenylation signal sequences are double underlined, and the polyadenylation site used by the 1.9 kb mRNA is indicated (*). Sequences specifying the potential regulatory elements AUUU(A), UAUU, and UAUUUAU in the 3' untranslated region are overlined. The *Alu* repeat sequences are bracketed; the repeat elements $(CT)_{11}(CA)_{13}(CG)_{12}$ between nucleotides 3312 and 3383, and $(TGAA)$, between nucleotides 5231 and 5258 are in italics. The sequences of sialophorin cDNA clones CEM0.8 and CEM1.7 [20] are identical with the corresponding region of the gene sequence presented here. Relative to the gene sequence, the cDNA clones HPB1.9 and HPB2.5 [20] contain one neutral C-to-T substitution in the coding region (nucleotide 1416). In addition, the 3' non-coding region of HPB2.5 contains four A-to-G substitutions (nucleotides 1921, 1994, 1998 and 2003), one T deletion (nucleotide 1877) and one T insertion (nucleotide 1894).

Fig. 3. Identification of the transcription initiation sites of the human sialophorin gene by S1-nuclease mapping and primer-extension analysis

(a) The 5' region of the gene is represented at the top; the open box indicates exon 1, specifying mRNA 5' non-coding sequence and the line represents 5' flanking sequence. Arrows mark the major transcription initiation sites identified by both primer-extension and S1-nuclease protection analysis. These sites are located 70 and 125 nucleotides upstream of the GT dinucleotide which defines the 5' end of the intron. The 164 bp PvuII–MspI genomic DNA fragment employed as a probe in the S1-nuclease protection analysis is represented below as are the major fragments protected from S1-nuclease digestion. The oligonucleotide primer used in the primer-extension analysis represents the 18 nucleotides of the anti-sense strand at the 3' end of the probe used in the S1-nuclease protection assay. It is aligned with the major products of extension with reverse transcriptase. This primer was also used to generate reference dideoxy sequencing reactions of the sialophorin gene. (b) The products of S1-nuclease protection analysis (lanes 1 and 3) and primer-extension analysis (lanes 2 and 4) using 10 μg of yeast tRNA (lanes 1 and 2) and 10 μg of human CEM poly(A)$^+$ RNA (lanes 3 and 4). Lanes marked T, C, A and G represent the reference dideoxy sequencing reactions. This panel is aligned to allow direct comparison of the genomic sequence with the transcription start sites depicted in lanes 3 and 4 of (b). Arrows on the sequence depicting the sense strand derived from the sequencing reactions mark the major transcription initiation sites located 70 and 125 nucleotides upstream of the 5' end of the intron.

4616 (i.e. 2285 and 2309 bp further downstream) are used to specify the 3' terminus of the 4.3 kb mRNA.

The genomic sequence contains only one intron (described above) in the region corresponding to the cDNAs previously sequenced by us ([20]; C. S. Shelley, E. Remold-O'Donnell, F. S. Rosen & A. S. Whitehead, unpublished work). Since these cDNA clones lack the 1.3 kb at the extreme 3' end of the 4.3 kb mRNA, the 3' end of the gene was examined for intron content by S1-nuclease protection assays. The 2.4 kb genomic DNA fragment used as a probe overlapped the 79 nucleotides at the 3' end of HPB2.5, the longest partial cDNA clone derived from the 4.3 kb mRNA (Fig. 4). A 1.5 kb protected product was generated, establishing that there are no introns in the portion of the 3' untranslated region that is unique to the 4.3 kb mRNA. Therefore a combination of S1-nuclease mapping and genomic DNA sequencing, in addition to Northern-blot analysis and cDNA sequencing [20], independently predict sialophorin mRNAs of 1.9 and 4.3 kb.
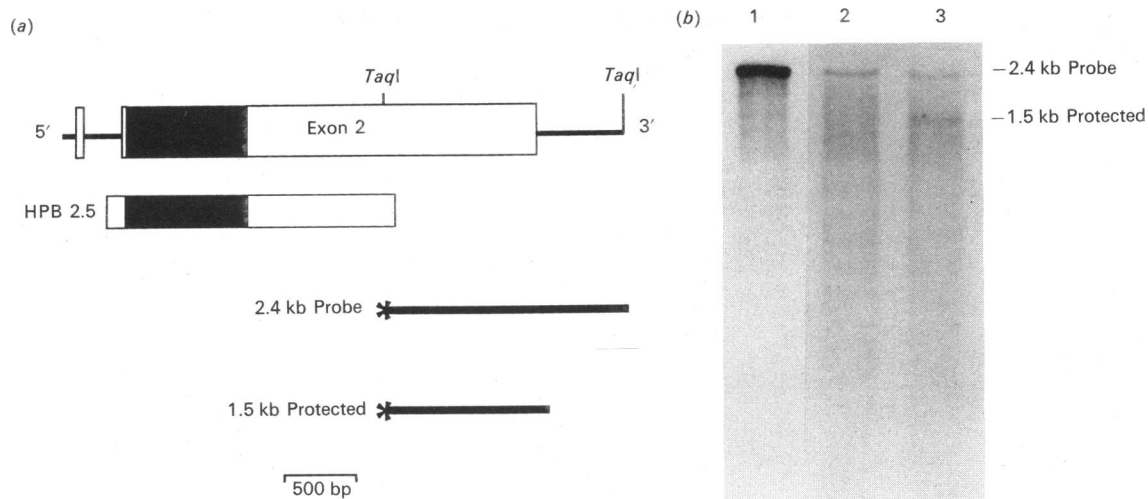
3' Non-coding region

The 3' untranslated region common to the two sialophorin mRNAs contains an inverted Alu repeat sequence between nucleotides 1883 and 2164 (Fig. 2). The 3' untranslated region unique to the 4.3 kb mRNA contains the 3' half of an Alu repeat between nucleotides 3045 and 3180 and a complete Alu sequence between nucleotides 4310 and 4587, which ends four nucleotides before the clustered polyadenylation signals (Fig. 2). The 3' untranslated region of the 4.3 kb mRNA also contains a repeat

element, $(CT)_{11}$ $(CA)_{13}$ $(CG)_{12}$, between nucleotides 3312 and 3383 (Fig. 2), the last 50 nucleotides of which have the potential to form Z-DNA.

The presence of AUUU(A)[31], UAUU[32] and UUAUUUAU [33] sequences in 3' untranslated regions may mediate mRNA instability. The 1.9 kb sialophorin mRNA has only two AUUU sequences and one UAUU sequence in its 3' untranslated region, whereas the 4.3 kb mRNA has eight additional AUUU sequences and three additional UAUU sequences. The only AUUUA motif, contained within the sequence UAUUUAU, is positioned at the start of the 3' untranslated region unique to the 4.3 kb mRNA. Therefore differential polyadenylation coupled with different stabilities of the 1.9 and 4.3 kb sialophorin mRNAs may provide a means of modulating sialophorin expression.

5' Flanking region

Upstream from the 3' major transcription initiation site at nucleotide 56, 1045 nucleotides have been sequenced. Between nucleotide −841 and −560 there is an Alu repeat. In the presumed promoter region there is a potential AP-2 binding site [34] between nucleotides −98 and −91, but there is no TATA box and only a degenerate CAAT box with the sequence CCACT. This latter motif begins 34 bp upstream from the 3' major transcription initiation site rather than the usual CAAT box location 70–80 bp upstream from a transcription initiation site [35]. However, the 100 nucleotides upstream of the 3' major transcription initiation site have a high content (71 %) of G and C nucleotides and contain a number of short G- and C-rich

Fig. 4. S1-nuclease analysis of the 3′ untranslated region of the 4.3 kb sialophorin mRNA

(a) The sialophorin gene is represented at the top; ☐ indicates regions of the gene specifying non-coding regions of the mRNA and ▰ the region specifying the coding region. Lines represent 5′ and 3′ flanking sequences and the single intron. The sialophorin cDNA clone HPB2·5 is aligned below the gene; ☐ and ▰ represent non-coding and coding regions respectively. The 2.4 kb genomic DNA TaqI fragment used as a probe in the S1-nuclease protection analysis is represented as is the extent to which it is protected from S1-nuclease digestion. (b) lane 1, the undigested TaqI fragment used as probe; lane 2, product of protection of the TaqI fragment from S1-nuclease digestion using 10 μg of yeast tRNA; lane 3, product of protection of the TaqI fragment from S1-nuclease digestion using 20 μg of total human CEM RNA. The 2.4 kb undigested probe and the 1.5 kb fragment protected from S1-nuclease digestion are marked.

repeats. One such repeat, GGGCCC, which is present in forward and reverse orientations between the major transcription initiation sites, has dyad symmetry and represents six out of eight bases that could form the stem of a hairpin/loop structure (nucleotides 16–47; Fig. 2). The region between nucleotides −31 and +14 is dominated by four repeat motifs (GCAGG, GGGTGGG, TGGAG, and GGGG) that comprise 96% of the sequence. This region on the antisense strands also contains five copies of the sequence CACCC, which constitutes an important element of the promoters of adult β-globin genes [36]. The absence of TATA- and CAAT-box motifs and the presence of GC-rich repeats is typical of the promoter regions of 'housekeeping' genes [37]. However, whereas the 5′ flanking sequences of 'housekeeping' genes usually contain large numbers of CG dinucleotides [38], no such region is present in the proximal 1045 nucleotides of the 5′ flank of the sialophorin gene. The structure of the sialophorin gene promoter therefore lacks the features typical of the promoters of other well-characterized categories of genes.

3′ Flanking region

Downstream of the five overlapping AATAAA sequences defining the 3′ end of the 4.3 kb mRNA, 897 nucleotides have been sequenced. Between nucleotides 4695 and 4851 there is an extensive region of pyrimidine/purine (principally TG) repeats capable, in theory, of forming Z-DNA. Alternating TG sequences may act as signals for gene conversion and may constitute 'hot spots' of recombination and/or gene rearrangement [39, 40]. Such elements may also weakly inhibit or enhance transcription [41,42]. Alternatively, as UG repeats, this region may be involved in processing of precursor RNA at the 3′ end [43]. There is an additional repeat, (TGAA)₇, present between nucleotides 5231 and 5258, that is surrounded by two inverted partial *Alu* sequences (Fig. 2).

Organization of the gene

The unusually large number of *Alu* sequences associated with the sialophorin gene suggests that retroposition may have played

a role in the evolution of its structure: replacement of the bulk of an ancestral gene with a partially processed sialophorin transcript may account for the relative lack of introns. The lack of introns in the coding region, although unusual, is not unprecedented for integral membrane proteins. Other examples include the hamster β-adrenergic receptor [44] and human thrombomodulin [45]. In particular, the organization of the sialophorin gene is similar to that of platelet glycoprotein Ibα [46]. Each of these genes contains a single intron interrupting the mRNA 5′ untranslated region and a long second exon specifying the remaining 5′ untranslated sequence, the extracellular, transmembrane and intracellular portions of the protein and the mRNA 3′ untranslated region. Each gene contains multiple *Alu* sequences and promoter regions with potential AP-2 binding sites, but lacking TATA and CAAT elements. Although their primary amino acid sequences are unrelated, sialophorin and glycoprotein Ibα are major sialic-acid-carrying O-glycosylated proteins on the surface of lymphocytes and platelets respectively.

REFERENCES

1. Parkman, R., Remold-O'Donnell, E., Kenney, D. M., Perrine, S. & Rosen, F. S. (1981) Lancet ii, 1387–1389
2. Remold-O'Donnell, E., Kenney, D. M., Parkman, R., Cairns, L., Savage, B. & Rosen, F. S. (1984) J. Exp. Med. 159, 1705–1723
3. Axelsson, B., Youseffi-Etemad, R., Hammarström, S. & Perlmann, P. (1988) J. Immunol. 141, 2912–2917
4. Carlsson, S. R. & Fukuda, M. (1986) J. Biol Chem. 261, 12779–12786
5. Killeen, N., Barclay, A. N., Willis, A. C. & Williams, A. F. (1987) EMBO J. 6, 4029–4034
6. Borche, L., Lozano, F., Viella, R. & Vives, J. (1987) Eur. J. Immunol. 17, 1523–1526

7. Brown, W. R. A., Barclay, A. N., Sunderland, A. & Williams, A. F. (1981) Nature (London) **289**, 456–460

8. Mentzer, S. J., Remold-O'Donnell, E., Crimmins, M. A. V., Bierer, B. E., Rosen, F. S. & Burakoff, S. J. (1987) J. Exp. Med. **165**, 1383–1392

9. Remold-O'Donnell, E., Zimmerman, C., Kenney, D. & Rosen, F. S. (1987) Blood **70**, 104–109

10. Kenney, D. M., Cairns, L., Remold-O'Donnell, E., Peterson, J., Rosen, F. S. & Parkman, R. (1986) Blood **68**, 1329–1332

11. Bierer, B. E., Peterson, A., Park, J., Remold-O'Donnell, E., Rosen, F. S., Seed, B. & Burakoff, S. J. (1988) Immunol. Allergy Clin. N. Am. **8**, 51–67

12. Silverman, L. B., Wong, R. C. K., Remold-O'Donnell, E., Vercelli, D., Sancho, J., Terhorst, C., Rosen, F. S., Geha, R. S. & Chatila, T. A. (1989) J. Immunol. **142**, 4194–4200

13. Nong, Y-H., Remold-O'Donnell, E., LeBien, T. W. & Remold, H. G. (1989) J. Exp. Med. **170**, 259–267

14. Vargas-Cortes, M., Axelsson, B., Larsson, A., Berzins, T. & Perlmann, P. (1988) Scand. J. Immunol. **27**, 661–671

15. Remold-O'Donnell, E., Davis, A. E., III, Bhaskar, K. R. & Rosen, F. S. (1986) J. Biol. Chem. **261**, 7526–7530

16. Piller, F., Piller, V., Fox, R. I. & Fukuda, M. (1988) J. Biol. Chem. **263**, 15146–15150

17. Chatila, T. A. & Geha, R. S. (1988) J. Immunol. **140**, 4308–4314

18. Remold-O'Donnell, E., Kenney, D. M. & Rosen, F. S. (1987) Biochemistry **26**, 3908–3913

19. Pallant, A., Eskenazi, A., Mattei, M.-G., Fournier, R. E. K., Carlsson, S. R., Fukuda, M. & Frelinger, J. G. (1989) Proc. Natl. Acad. Sci. U.S.A. **86**, 1328–1332

20. Shelley, C. S., Remold-O'Donnell, E., Davis, A. E., III, Bruns, G. A. P., Rosen, F. S., Carroll, M. C. & Whitehead, A. S. (1989) Proc. Natl. Acad. Sci. U.S.A. **86**, 2819–2823

21. Benton, W. D. & Davis, R. W. (1977) Science **196**, 180–182

22. Maniatis, T., Fritsch, Z. F. & Sambrook, J. (1982) in Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

23. Bell, G. I., Karem, J. H. & Rutter, W. J. (1981) Proc. Natl. Acad. Sci. U.S.A. **78**, 5759–5763

24. Southern, E. M. (1975) J. Mol. Biol. **98**, 503–517

25. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) Proc. Natl. Acad. Sci. U.S.A. **74**, 5463–5467

26. Berk, A. J. & Sharp, P. A. (1977) Cell (Cambridge, Mass.) **12**, 721–732

27. Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. & Rutter, W. J. (1979) Biochemistry **18**, 5294–5299

28. Aviv, H. & Leder, P. (1972) Proc. Natl. Acad. Sci. U.S.A. **69**, 1408–1412

29. Pearson, W. R. & Lipman, D. J. (1988) Proc. Natl. Acad. Sci. U.S.A. **85**, 2444–2448

30. Breathnach, R. & Chambon, P. (1981) Annu. Rev. Biochem. **50**, 349–383

31. Shaw, G. & Kamen, R. (1986) Cell (Cambridge, Mass.) **46**, 659–667

32. Reeves, R., Elton, T. S., Nissen, M. S., Lehn, D. & Johnson, K. R. (1987) Proc. Natl. Acad. Sci. U.S.A. **84**, 6531–6535

33. Caput, D., Beutler, B., Hartog, K., Thayer, R., Brown-Shimer, S. & Cerami, A. (1986) Proc. Natl. Acad. Sci. U.S.A. **83**, 1670–1674

34. Mitchell, P. J., Wang, C. & Tjian, R. (1987) Cell (Cambridge, Mass.) **50**, 847–861

35. Benoist, C., O'Hare, K., Breathnach, R. & Chambon, P. (1980) Nucleic Acids Res. **8**, 127–142

36. Myers, R. M., Tilly, K. & Maniatis, T. (1986) Science **232**, 613–618

37. Martini, G., Toniolo, D., Vulliamy, T., Luzzatto, L., Dono, R., Viglietto, G., Panonessa, G., D'Urso, M. & Persico, M. G. (1986) EMBO J. **5**, 1849–1855

38. Bird, A. P. (1986) Nature (London) **321**, 209–213

39. Miesfeld, R., Krystal, M. & Arnheim, N. (1981) Nucleic Acids Res. **9**, 5931–5947

40. Kilpatrick, M. W., Klysik, J., Singleton, C. K., Zarling, D. A., Jovin, T. M., Hanau, L. H., Erlanger, B. F. & Wells, R. D. (1984) J. Biol Chem. **259**, 7268–7274

41. Nordheim, A., Tesser, P., Azorin, F., Kwon, Y. H., Moller, A. & Rich, A. (1982) Proc. Natl. Acad. Sci. U.S.A. **79**, 7729–7733

42. Hipskind, R. A. & Clarkson, S. G. (1983) Cell (Cambridge, Mass.) **34**, 881–890

43. Birnstiel, M. L., Buslinger, M. & Strub, K. (1985) Cell (Cambridge, Mass.) **41**, 349–359

44. Dixon, R. A. F., Kobilka, B. K., Strader, D. J., Benovi, J. L., Dohlman, H. G., Frielle, T., Bolanowski, M. A., Bennett, C. D., Rands, E., Diehl, R. E., Mumford, R. A., Slater, E. E., Sigal, I. S., Caron, M. G., Lefkowitz, R. J. & Strader, C. D. (1986) Nature (London) **321**, 75–79

45. Jackman, R. W., Beeler, D. L., Fritze, L., Soff, G. & Rosenberg, R. D. (1987) Proc. Natl. Acad. Sci. U.S.A. **84**, 6425–6429

46. Wenger, R. H., Kieffer, N., Wicki, A. N. & Clemetson, K. J. (1988) Biochem. Biophys. Res. Commun. **156**, 389–395