The EMBO Journal vol.6 no.13 pp.4029–4034, 1987

# The sequence of rat leukosialin (W3/13 antigen) reveals a molecule with O-linked glycosylation of one third of its extracellular amino acids

**Nigel Killeen, A.Neil Barclay, Antony C.Willis[1] and Alan F.Williams**

MRC Cellular Immunology Unit, Sir William Dunn School of Pathology, University of Oxford, Oxford, OX1 3RE and [1]MRC Immunochemistry Unit, Department of Biochemistry, University of Oxford, Oxford, OX1 3QU, UK

Communicated by A.F.Williams

**Leukosialin is one of the major glycoproteins of thymocytes and T lymphocytes and is notable for a very high content of O-linked carbohydrate structures. The full protein sequence for rat leukosialin as translated from cDNA clones is now reported. The molecule contains 371 amino acids with 224 residues outside the cell, one transmembrane sequence and 124 cytoplasmic residues. Data from the peptide sequence and carbohydrate composition suggest that one in three of the extracellular amino acids may be O-glycosylated with no N-linked glycosylation sites. The cDNA sequence contained a CpG rich region in the 3′ coding sequence and a large 3′ non-coding region which included tandem repeats of the sequence GGAT.**

*Key words:* leukosialin/mucin-like glycoprotein/O-linked carbohydrate/surface antigen/T lymphocyte

## Introduction

Much of the carbohydrate of T lymphoid cells is displayed by two sets of molecules, namely the leucocyte-common antigen (T200) and the leucocyte sialoglycoprotein (henceforth referred to as leukosialin. The names leucocyte sialoglycoprotein (Brown *et al.*, 1981), leukosialin (Carlsson and Fukuda, 1986) and most recently sialophorin (Mentzer *et al.*, 1987) have been used for this molecule. The term leukosialin seems best as it is brief and indicates both the predominant tissue origin and the molecular type. The leukosialins of various species and cell types have apparent $M_r$s in the range 100 000–150 000 and are notable for a very high content of sialic acid that is mainly found on O-linked carbohydrate structures containing the core disaccharide Galβ1–3GalNAc. The carbohydrate structures account for ~60% by weight of leukosialin (Gahmberg *et al.*, 1976; Standring *et al.*, 1978; Conzelmann *et al.*, 1980; Saito and Osawa, 1980; Brown *et al.*, 1981; Axelsson *et al.*, 1985; Remold-O'Donnell *et al.*, 1986; Carlsson and Fukuda, 1986; Carlsson *et al.*, 1986).

Heterogeneity of leukosialins is seen in size and antigenicity. For example the L10 monoclonal antibody (MAb) detects a human lymphocyte leukosialin of 115 000 apparent $M_r$ compared with 135 000 $M_r$ for the molecule from neutrophils (Remold-O'Donnell *et al.*, 1987). In the rat, the W3/13 MAb binds only half of the thymocyte leukosialin band (Brown *et al.*, 1981) while other MAbs, including MRC OX-56 which has been used for purification in this study, can bind some of the leukosialin that is not W3/13$^+$ (unpublished results). Thus it has been unclear whether leukosialin is encoded by one gene or whether a set of related genes may exist.

Studies with the W3/13 MAb showed that rat leukosialin with an $M_r$ of ~100 000 is found on thymocytes, T lymphocytes, neutrophils, plasma cells and myelomas but not on B lymphocytes or most other tissues (Williams *et al.*, 1977; Brown *et al.*, 1981). Antigenic activity was also found in brain extracts but preliminary studies on affinity-purified brain antigen identified a heavily glycosylated band of >200 000 apparent $M_r$. In addition a number of new MAbs against rat leukosialin did not identify antigenicity in the brain (W.R.A.Brown, unpublished). Thus it may be that the activity identified in brain with W3/13 MAb is not encoded by the same gene as the antigen from leucocytes. Studies on human haemopoietic cells with the L10 MAb showed a similar distribution to that seen in the rat (Mentzer *et al.*, 1987).

Leukosialin is of interest with regard to aspects of cell surface behaviour. These include: (i) Cell surface charge; the molecule must play a key role in determining the physicochemical properties of the T cell surface and its expression on T but not B cells may account for the surface charge difference between these cells (reviewed in Brown *et al.*, 1981). (ii) Lectin binding; the specificity of peanut lectin for thymocyte subsets is determined by the relative lack of terminal sialic acid on cortical thymocytes in contrast to medullary cells or T lymphocytes. Leukosialin appears to account for much of the peanut lectin binding. (Brown and Williams, 1982; De Maio *et al.*, 1986). (iii) The Wiskott-Aldrich immunodeficiency syndrome; expression of leukosialin on leucocytes is reduced in this syndrome and some of the molecules that are detected have abnormal apparent $M_r$s. It has been suggested that defects in leukosialin lead to a shortened half-life for Wiskott-Aldrich T lymphocytes (Remold-O'Donnell *et al.*, 1984; Kenney *et al.*, 1986). (iv) Triggering of T lymphocyte division; the L10 MAb is mitogenic for human T lymphocytes in the presence of accessory cells (Mentzer *et al.*, 1987). This potential for cell triggering may indicate that leukosialin has functions other than that of contributing to the physicochemical properties of the cell surface.

Genetic probes for leukosialin are essential for future work on this molecule and we now describe the cDNA cloning and sequence of the rat molecule.

## Results

### Protein purification

Rat leukosialin was purified from detergent extracts of thymocytes using an affinity column containing both W3/13 and OX-56 MAbs. After two cycles of affinity chromatography and a gel filtration step, analysis by SDS–PAGE revealed a final product consisting mainly of material with an apparent $M_r$ of ~100 000 (data not shown). Minor bands were also seen at 82, 70, 56 and 45 kd and all these bands reacted strongly with the periodic acid Schiff carbohydrate stain and also with W3/13 and OX-56 MAbs by immunoblotting. Rat leukosialin is susceptible to proteolysis (Brown *et al.*, 1981) and the lower $M_r$ forms are assumed to be proteolytic products from the 100 kd form. Amino acid analysis of the purified material gave the expected high values

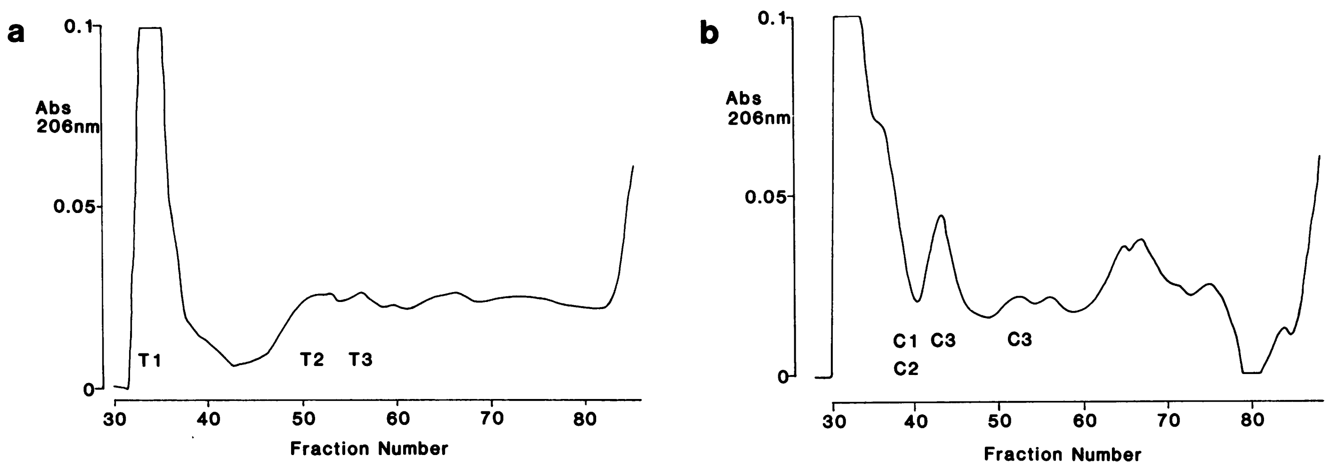


**Fig. 1.** Gel filtration of peptides on Biogel P-30. The traces show absorbance at 206 nm from peptides run on a Biogel P-30 column (150 × 1 cm) in 0.1 M $NH_4HCO_3$. The peptides resulted from tryptic (**a**) and chymotryptic (**b**) digests and the derivation of those that were sequenced (Figure 2) is shown on the traces.

| $NH_2$-Terminal sequence | | | Position in Predicted Sequence | Sequences Used for Oligonucleotides |
|---|---|---|---|---|
| | ***PN-M-MLPF-PN-E-P---EAL--V--IA- | | 1-33 | |
| Peptide sequences | | | | |
| TCN3 | LPF-PN-E | | 10-17 | |
| TCN1 | A-G-LGP-KET-GL-A-IA | Double sequence resolved using cDNA sequence. | 166-184 | |
| TCN2 | -SGVA-DPPVTI-NPA--- | | 101-119 | |
| C1 | *GSVALEELK-GTG-NLKG | | 321-340 | EELKP (14/48)<br>KGEEEP (17/64) |
| C2 | EELKPGTGPNLKGEEEPLVGSEDEAVET | | 327-354 | PNLKG (14/96) |
| C3 | AGPARVPDEEATTASGSGGN | | 274-293 | PDEEA (14/32) |
| T2 | *SGAPETDGSGQ-PTL | | 295-310 | PETDG (14/64) |
| T3 | DGAAPQSL | | 364-371 | |

Peptide derivation:T,tryptic;C,chymotryptic;TCN,tryptic followed by cyanogen bromide cleavage.
* Position with multiple PTH-residues precluding an unambiguous assignment.
- Position where no PTH-residue was detected.

**Fig. 2.** $NH_2$-terminal and peptide sequences. Protein and peptides were sequenced on an Applied Biosystems gas phase sequencer. In the $NH_2$-terminal run ~2 nmol of protein were used and the yield of Pro at residue 4 was 0.45 nmol. The peptide sequences were obtained at levels of 0.1–3.3 nmol. The numbers in brackets after the peptide sequences used for oligonucleotide construction indicate the length and redundancy of the mixed oligonucleotide.

for serine, threonine, proline and galactosamine. Thus the properties of material purified in this study were fully in accord with the previous purification data of Brown *et al.* (1981).

### *Peptide purification and sequence*

Material from the affinity column was digested with trypsin or chymotrypsin and the digest was fractionated by gel filtration on a Biogel P-30 column (Figure 1) followed by reverse phase HPLC (not shown). In each case a large amount of material ran at the front of the Biogel P-30 column and this contained a fragment of ~45 000 $M_r$ that was resistant to further digestion with a number of proteases. This fragment stained intensely for carbohydrate and gave a high value for galactosamine in amino acid analysis. To obtain peptides from this glycosylated region, the large tryptic fragment T1 (Figure 1(a)) was cleaved further by cyanogen bromide treatment and peptides were isolated by HPLC. In most cases HPLC peaks were subjected to amino acid analysis and these data were used to select peptides for sequencing.

Figure 2 shows all of the protein sequence data. In an $NH_2$-terminal sequence run clear data were obtained after the first three sequencing cycles for which unambiguous assignments could not be made. In the $NH_2$-terminal sequence and in some other peptide sequences no phenyl-thio-hydantoin (PTH)-derivatives were seen at a number of positions. On the basis of the predicted protein sequence it seems likely that these were glycosylated (see below). A number of peptide sequences were obtained that were suitable for synthesis of mixed oligonucleotides for use in screening cDNA libraries and these are given at the right in Figure 2.

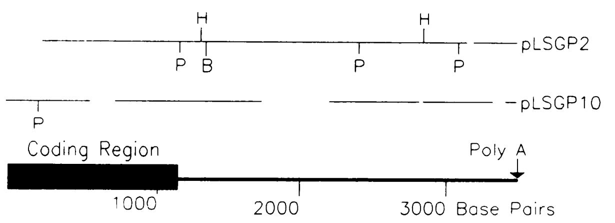
### *Isolation and sequence of cDNA clones and Northern blot analysis*

The oligonucleotides were used to screen a rat thymocyte cDNA library prepared by the S1 nuclease loopback procedure and the clone pLSGP2 was chosen for sequencing because it reacted with all five oligonucleotide probes. This clone was sequenced to identify a coding region accounting for the protein sequence in Figure 2 with the exception of the $NH_2$-terminal sequence and TCN3. Thus further clones were isolated from a thymocyte cDNA library prepared by the RNase H method and pLSGP10 accounted for the sequence in pLSGP2 plus the missing protein sequence. pLSGP10 lacks a full leader sequence but provides enough information to deduce the full sequence of the mature glycoprotein.

Figure 3 shows the composite DNA sequence from pLSGP2 and 10. Sequence in both clones was determined at 71% of the positions and all of the determined nucleotides were identical. The sequence in the coding region was verified by accounting for all of the peptide data.

Unusual features of the nucleotide sequence were a high con-

A.Restriction Maps and Regions Sequenced of pLSGP10 and pLSGP2.



B.Nucleotide Sequence of Rat Leukosialin cDNA.

```
ctgggcccaggtggtgagccaagaaaatctgccgaatacgatgacgatgttgccatttac   60
tccgaatagcgagtccccaagtacctctgaagccttgagtacctactcatcaattgctac  120
agtgccagtgacagaggaccctaaggagagtatcagcccctgggggcagaccactgcccc  180
agcttcctcaatccccttgggaactccagaattgtcttctttttttttacatcagctgg  240
tgccagcgggaacaccccagtacctgagcttacaacctctcaggaagtttccaccgaggc  300
atcattagtactgtttccaaaatcatcaggtgtagccagtgaccctcctgtcactataac  360
taatcctgcgacaagcagtgctgtggcatctacctctttagaaactttaaagggaccag  420
tgcacctcctgtcactgtgacaagctcgaccatgaccagtggaccctttgtggctacaac  480
tgtaagctccgagaccagtggtcctccagtcactatggctactgggtctctggggccctc  540
caaagagacacatggactctctgccaccatagcaaccagttctggtgagagttctagtgt  600
ggccggtggcaccccagtcttcagtacaaaaatatccacaacgtctaccccaaatcccat  660
aaccaccgtgccaccacgcccaggatcgagtggcatgttgctggtttccatgctcattgc  720
cttgacggtggttttggtccttgtggcgctactgctgctgtggcgccagaggcagaagcg  780
gaggactggggccctgaccctgagcagaggtggaaaacgaaatggtacggtggatgcctg  840
ggctgggccagctcgggttcctgacgaagaggccaccaccgcatcagggtcagggggcaa  900
caagagctccggagcgccggagacggatggctccgggcagcggcccacgctcaccacttt  960
cttcagcagacggaagtctcgccagggctcggtagcactagaagagctgaagcctgggac 1020
gggtcccaacctgaagggggaggaagagcctctggtaggcagtgaggatgaagctgtgga 1080
aaccccaacttctgacgggccacaagccaaagatggggctgcacctcaatccctatgagt 1140
aacggtattcatctagtctgctccttgaccccagatcctgctgtcagctgcctctcactc 1200
tgcatttgtgaatatttgtgggccagatccattcgttcattcattattcagccattcatt 1260
cctctccaacccccgctgttttccacctccccttcctctgtgctccctttgaatctccag 1320
aagataagcttagttgacccacattttccaagcatccccccttggaatgctgggatccgac 1380
aagaagaatgaaaatgataagccagaaaactccggaggaggggccccctgaggactgcct 1440
ctcaactaccatttgctttggggacagcttagagctggatgtagccattcagtttgctct 1500
tgggtgaactgtgacctgtgacttttcccagagcactccgatttgggctggccggtagc 1560
tgcttggggtagcctgccagagatcagtcacttctcctggaggtggggatggggagactt 1620
cctggctccatgtggtccctgcacctgttttgtccactcggaagtatggtttatgtgcaa 1680
gataggacaatttaggcgatggctgtgaacagtactggtgagtatagaccatggagagag 1740
caggggagatgcaaaaggagggcacgggagctgccgagccctcagatcccttcattgtg 1800
ctcagacaggatatggctgaaaacctaagtttgtctgtagttccccaaaagaatgcaaac 1860
atgatggggactaaatgtctccatacttgtgccatcaagaatgagttcttcagcttagct 1920
gccacagcctttcctcaaaggctccctcagaccagcacacaccaagcacccatgcaagaa 1980
ggggagaagaatcagcccaatgcagggactgggtggtgcaaaaccactgcttttgatgat 2040
gctttccctacaggctcaatttgtccgtttcctctgaagtccttcccatctctggaataa 2100
gaaagagttaaatgaaagacaggctcaggggggtggggcatggatgcccaaagggtctgag 2160
gtagagaggggacagaccacaattgtgctgttttctgactctgataccagcccggctcgt 2220
ggtccaaagctggttgctggaagagctgtgcaaggggaagtcgttgagtcagaggacact 2280
gtccttgagaacagcctgacctacaaaaagatgtgtgtgtgcggctgtgtggcctgggtg 2340
gaaaaggggaaggtgggtaggagggaggggacgaggagtggaaagggacaaggatcagct 2400
gccaccacatcatcatctcctccccaagtctgtgggaagaggaagggccaagatgggcac 2460
acctgagccataccgaacccatttcttataggcgcagtgtggagcagctctaacagatgg 2520
ctagtgagttgcaagtgtcccttaaggaacataacaactgttgtggctgatttttgccaa 2580
accgaatgtccctatttcaggcttagaggagaatgttaccaagctgagaggtcaatggct 2640
ctacccatcatgttccactgacactcctgaggtattacaggtttccaggtacagggagga 2700
agacaaggaaatggggtagggatggtggatggcacgcaataggacctcagccctgacaga 2760
caaaatgtcacaagaggctaggacatgattggtattccacctccgaaggaggacaaggaa 2820
gcttaatgtttgtcccagggtcttggccatatctctctttctagtggctccagtacctgg 2880
caagcagcaaggactgagtgcatgctccgatggatggatggatggatggatggatgttgg 2940
atggatgaatggatggatgaatggacagatggatagatgaggtattggcagctagcaaga 3000
gagaaggcatttaacacaaagattagtaatgcaaaaatatataagtgagtggaacagaga 3060
tcatggtggatacacacagctgtctttcagctcaggagagcgagcacccagtaagcaata 3120
gaaaagcaaaaatagaagaagaaaaagaagaaaaaagcaggcaaacagaaacagggaaat 3180
gtaaaacacagagaccgtttgaactccgtcagttatgaaaccatagtaagtggcttgtgt 3240
ttgtgacatcctggccttgtcttcagaaatgtgtaatgacttcagaaatgcctgcggctg 3300
acacggggtagggaagtaaatgatctgtaaatacatgtgtgttattgcatttgtgatagc 3360
agctacatcctgaatgtttagtacggtcctcctaaaggggaatttaagtgtgattttttt 3420
cacctaacttgtctttgttttacacaataaatgtctttacttttacagttaaaaaaaaa  3479
```

C.The CpG Rich Region.

```
                                  aaacgaaatggtacggtggatgcctg  840
ggctgggccagctcgggttcctgacgaagaggccaccaccgcatcagggtcagggggcaa  900
caagagctccggagcgccggagacggatggctccgggcagcggcccacgctcaccacttt  960
cttcagcagacggaagtctcgccagggctcggtagcacta                    1000
```

D.The 3' Repeating Sequence.

```
                                  ctctctttctagtggctccagtacctgg 2880
caagcagcaaggactgagtgcatgctccgatggatggatggatggatggatggatgttgg 2940
atggatgaatggatggatgaatggacagatggatagatgaggtattggcagctagcaaga 3000
```

**Fig. 3.** Sequence of cDNA from pLSGP2 and pLSGP10. **A**. shows the regions in each clone that were sequenced plus the coding and non-coding regions. B, H, and P indicate *Bam*HI, *Hin*dIII and *Pvu*II restriction enzyme cleavage sites. The cDNA inserts are in pAT153/*Pvu*II/8 and thus are flanked at the 5′ end by *Eco*RI and *Hin*dIII sites and at the 3′ end by a *Bam*HI site. **B**. shows the nucleotide sequence derived from the two clones. Underlined are the stop codon starting at nucleotide 1136 and the poly(A) recognition site starting at 3446. In the CpG rich region (**C**) and the 3′ repeating sequence (**D**), the relevant features of the sequence are underlined.

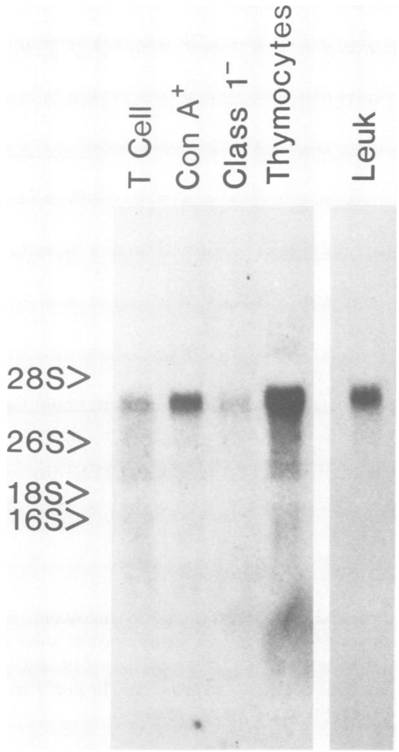


**Fig. 4.** Northern blot analysis. RNA was analysed from peripheral T lymphocytes (T cell); thymocyte blasts prepared by stimulation with Con A *in vitro* (Con A$^+$); cortical-type thymocytes (Class I$^-$); unfractionated thymocytes (Thymocytes) and a rat CD4$^+$ leukaemic cell line (Leuk). In all cases one distinct mRNA band is seen at ~4.5 kb. In tracks for RNA from T cells and cortical type-thymocytes there is a smear to smaller size material and this is believed to be due to partial degradation of the RNA. The data are not suitable for quantitative comparison.

tent of CpG dinucleotides in the 3′ part of the coding region and the presence of a repeating motif in the 3′ non-coding region. Between nucleotides 815 and 1000 the content of C plus G is 65% (Figure 3C) and the level of the dinucleotide CpG is 80% of that expected on a random basis. Thus this region does not show marked suppression of the CpG dinucleotides as is commonly seen in eukaryotic DNA (Gardiner-Garden and Frommer, 1987). The repeating motif in the 3′ non-coding region is GGAT which is found 11 times between nucleotides 2912 and 2974 (Figure 5D). This tetranucleotide is identical to repeats found in the 5′ region of the human myoglobin gene (Weller *et al.*, 1984) and in the gene cluster for rat transfer RNA genes (Shibuya *et al.*, 1985). Gene segments containing repeats of this type are believed to arise through unequal crossing-over in genetic recombination (Smith, 1976).

Northern blot analysis was consistent with the cDNA data in showing an mRNA of ~4.5 kb and this was the same size in thymocytes, T lymphocytes, concanavalin A-activated thymocytes and a lymphoma cell line (Figure 4).

### Protein sequence

The translated protein sequence is shown in Figure 5A and includes 371 residues that are predicted to constitute the fully processed sequence. A full leader sequence was not identified since full length clones at the 5′ side were not obtained. The identification of the $NH_2$-terminal sequence as shown in Figure 2 provides the evidence that a full-length processed sequence has been determined and the only doubt about this is the possibility that the protein sequence may have come from a fragment of the molecule with the real $NH_2$-terminus being blocked. However, this seems unlikely since no known protease would readily cleave

A.Predicted Protein Sequence of Rat Leukosialin.

```
-7           +1                                    20                                      40
W A Q V V S Q E N L P N T M T M L P F T P N S E S P S T S E A L S T Y S S I A T V P V T E D P K E S I S P W G Q T T A P

           60                                      80                                     100
A S S I P L G T P E L S S F F F T S A G A S G N T P V P E L T T S Q E V S T E A S L V L F P K S S G V A S D P P V T I T

          120                                     140                                     160
N P A T S S A V A S T S L E T F K G T S A P P V T V T S S T M T S G P F V A T T V S S E T S G P P V T M A T G S L G P S

          180                                     200                                     220
K E T H G L S A T I A T S S G E S S S V A G G T P V F S T K I S T T S T P N P I T T V P P R P G S S G M L L V S M L I A

          240                                     260                                     280
L T V V L V L V A L L L L W R Q R Q K R R T G A L T L S R G G K R N G T V D A W A G P A R V P D E E A T T A S G S G G N

          300                                     320                                     340
K S S G A P E T D G S G Q R P T L T T F F S R R K S R Q G S V A L E E L K P G T G P N L K G E E E P L V G S E D E A V E

          360
T P T S D G P Q A K D G A A P Q S L *
```

B.Probable Glycosylation Sites Deduced From Peptide and cDNA Sequences.

```
E N L P N Ⓣ M Ⓣ M L P F Ⓣ P N Ⓢ E Ⓢ P Ⓢ Ⓣ Ⓢ E A L Ⓢ Ⓣ Y Ⓢ Ⓢ I A Ⓣ (Residues 1-33)
  ▲                 △
Ⓢ S G V A Ⓢ D P P V T I Ⓣ N P A Ⓣ Ⓢ Ⓢ (Residues 101-119)
                    △
A Ⓣ G Ⓢ L G P Ⓢ K E T H G L Ⓢ A Ⓣ I A (Residues 166-184)

○ Glycosylated S or T
▲ Non-glycosylated S or T
△ Glycosylation state unknown
```

**Fig. 5. A.** The predicted protein sequence for leukosialin. Similarities within the sequence are underlined and a putative transmembrane sequence is boxed. **B**. The probable glycosylation sites are determined as sites where Ser or Thr is predicted from the cDNA sequence but where no residue was obtained in the protein seauence. The sequences at residues 101–119 and 166–184 were from a double sequence in which coincidence of residues occurs at positions 11 and 17. At 11 PTH-Thr was detected and thus the Thr could be unglycosylated in one or both of the sequences whilst at 17 no PTH-Thr was obtained and thus these residues are thought to be glycosylated.

between Gln and Glu residues as would be required to produce such a fragment. Also, the partial sequence for the putative leader does not conflict with the signal sequence consensus of von Heijne (1986). The peptide sequences in Figure 2 are all present in the sequence from the cDNA and confirm the predicted sequence throughout its full length. The amino acid composition from the sequence in Figure 5 mostly accords with that previously determined by Brown *et al.* (1981) as shown by a comparison of experimental values for percent amino acids with those from the sequence: Cys (0,0), Asx (5.3,4.8), Thr (13.7,14.3), Ser (13.8, 15.9), Glx (9.4,9.2), Pro (9.4,8.6), Gly (9.5,9.2), Ala (8.7, 8.6), Val (7.5,7.5), Met (1.4,1.6), Ile (2.5,2.2), Leu (8.1,7.8), Tyr (0.5,0.3), Phe (2.7,2.7), His (0.6,0.3), Lys (3.6,3.2), Arg (3.5,3.2), Trp (0,1.1). The only discrepancy is for Trp which was not detected in the analysis but is present at four residues/molecule in the sequence in Figure 5. Trp is an unstable amino acid and presumably was destroyed during hydrolysis even though this was done in paratoluene sulphonic acid which should have allowed recovery of Trp residues (Brown *et al.*, 1981).

The 371 residue sequence in Figure 5 is bisected by a hydrophobic sequence of 23 amino acids at residues 225–247 and this is assumed to cross the lipid bilayer. This divides the protein into extracellular and intracellular domains of 224 and 124 residues respectively.

The extracellular domain contains no potential N-linked glycosylation sites but there are very large proportions of the residues that are commonly found in sequences that are heavily O-glycosylated namely Ser/Thr, 38%; Pro, 12%; Gly/Ala, 14.7%. In the peptide sequences in Figure 2 Ser and Thr residues were commonly not detected in extracellular peptides yet in no case was a Ser or Thr assignment missed for intracellular peptides. It thus seems likely that all unassigned Ser or Thr residues are glycosylated and on the basis of this assumption the probable glycosylation for the peptides from the external sequence can be suggested as shown in Figure 5B. Amongst the relevant sequences only two or three out of 27 Ser or Thr residues were not glycosylated. If this applies throughout the whole extracellular part, then the molecule would contain ~75 O-linked carbohydrate structures and 1/3 of extracellular residues would be glycosylated. This is in accord with the estimate based on composition that one in five amino acids is O-glycosylated (Brown *et al.*, 1981) given that only 2/3 of the sequence is extracellular. Within the extracellular sequence a repeat of PPVT is seen in three places and this is flanked by other similarities as shown in Figure 5. In a search of the NBRF database, version 12, the PPVT tetrapeptide was seen only five times and it thus seems likely that the repeat of this in the leukosialin sequence marks three internal homology units each of ~26 residues overall.

The putative cytoplasmic domain of leukosialin contains 124 residues and is typical in beginning with a cluster of basic amino acids immediately after the transmembrane sequence. Overall, the cytoplasmic domain contains 17% basic residues and this is notably different from the extracellular part in which His, Lys and Arg account for only 3% of all residues.

## Discussion

The peptide sequences clearly identified the cDNA clones as coding for leukosialin and the predicted sequence was in accord with biochemical properties previously determined (Brown *et al.*, 1981). This included sequences typical of a molecule heavily glycosylated with O-linked oligosaccharides and the absence of Cys residues and any N-linked glycosylation sites. The only discrepancy was the presence of four Trp residues (1.1% of amino acids) in the sequence that was not detected by amino acid analysis but this was presumably due to susceptibility of Trp residues to destruction during hydrolysis. The sequence length of 371 amino acids which predicts a peptide backbone of 37 627 $M_r$ is in good accord with a protein of apparent $M_r$ of 100 000 in analysis by SDS–PAGE that contains 60% by weight of carbohydrate. However, this concordance is probably fortuitous since sizes determined by SDS–PAGE can be anomalous for heavily glycosylated molecules (Segrest and Jackson, 1972). Human leukosialin is predicted to contain 500–520 amino acids based on the size of an unglycosylated precursor detected in biosynthesis studies (Carlsson and Fukuda, 1986; Remold-O'Donnell *et al.*, 1987) and this is considerably larger than the rat molecule. The mature human molecule from T lymphocytes is also estimated by SDS-PAGE to be larger than the rat molecule with an $M_r$ of 115 000 compared with ~100 000 but some of this difference may be accounted for by the presence of 1-2 N-linked carbohydrate structures in the human molecule.

The question of whether there is one or more genes for leukosialins is not resolved by this study but only one size of mRNA was seen in a variety of different T lymphocyte types. Biosynthetic studies on human leukosialins have shown that precursor forms for molecules from different cell types are the same size even though the processed forms can differ markedly in apparent $M_r$ (Carlsson and Fukuda, 1986; Remold-O'Donnell *et al.*, 1987). It may thus be that the heterogeneity in size and antigenicity of leukosialin is due to differences in glycosylation and Carlsson *et al.* (1986) have characterized some of the different carbohydrate structures that can be found on this molecule.

In database searches we found no clear match between the leukosialin sequence and other protein sequences. However, within the rat leukosialin sequence there is evidence for three internal repeats and this suggests that the protein may have evolved by duplication of short segments. It is of note that in other large structures with a high content of O-linked carbohydrate exact repeats of short stretches of amino acids can occur. This is the case for polysialoglycoproteins of Rainbow trout eggs which have a 13 residue repeat (Kitajima *et al.*, 1986) and for the Ca-1 cell surface glycoprotein that is selectively expressed on cancers of epithelial cells in which the repeat size is 20 amino acids (Swallow *et al.*, 1987). In the case of leukosialin a sequence of identical repeats may have diverged to the present form where the repeats are only indistinctly seen.

The cytoplasmic domain of leukosialin contains 124 amino acids and this is much larger than is needed to anchor the molecule in the membrane. For example, cell surface IgM has only three cytoplasmic amino acids (Kehry *et al.*, 1980). There is the possibility that this region interacts with intracellular molecules and this could be important in the triggering of mitogenesis by the cross-linking of leukosialin with the L10 MAb (Mentzer *et al.*, 1987). It is also of interest that leukosialin can spontaneously cap to one pole of the cell on thymocytes (De Petris, 1984) and an interaction linked to the cytoskeleton may play a role in this. If the cytoplasmic domain does interact with other sequences then conservation of sequence in this region may be expected between species. For example, the leucocyte-common antigen sequence is notable in that its large intracellular domain is 85% identical between rat and human while the extracellular sequence is conserved only to the level of 45% (Thomas *et al.*, 1985; Ralph *et al.*, 1987).

## Materials and methods

### *Purification of leukosialin*

Two MAbs W3/13 and MRC OX-56 were used to purify leukosialin by antibody affinity chromatography. The MRC OX-56 MAb was obtained from a fusion with spleen cells from mice immunized with activated T cells (Jefferies *et al.*, 1985) and binds a larger fraction of the leukosialin than does W3/13 (unpublished data). Detergent extracts of rat thymocytes were prepared either by solubilization in sodium deoxycholate from crude membrane (Sunderland *et al.*, 1979) or by lysis of cells in Brij 96 followed by addition of sodium deoxycholate (Brown *et al.*, 1981). In most cases extracts that had been used in the purification of other antigens and stored at -40°C were used. Proteolytic inhibitors were added at various stages throughout the purification (Brown *et al.*, 1981). Two cycles of affinity chromatography on a 35 ml column containing equal volumes of Sepharose 4B-CL coupled with W3/13 IgG (10 mg/ml beads) or MRC OX-56 IgG (10 mg/ml) and one gel filtration step as in Brown *et al.* (1981) gave pure but partially degraded antigen (see Results). Sodium deoxycholate was removed by dialysis against 0.1 M $NH_4HCO_3$.

### *Preparation and sequencing of leukosialin peptides*

Protein succinylation and peptide isolation by gel filtration and reverse phase HPLC were as in Johnson *et al.* (1985). In tryptic digests ~1 mg (10 nmol) of succinylated glycoprotein was used in 0.5 ml 0.1 M $NH_4HCO_3$ plus 2% trypsin added in two aliquots for 28 h at 37°C. In a chymotryptic digest 15 nmol of glycoprotein were digested as above but with $\alpha$-chymotrypsin. A CNBr digest was performed on 15 nmol of pooled tryptic fragment T1 [Figure 1(a)] as in Thomas *et al.*, 1985. The resultant peptides were fractionated by HPLC. Peptides were sequenced on an Applied Biosystems gas phase sequencer using the 01tfav programme.

### *Isolation of leukosialin cDNA clones*

cDNA clones were isolated by screening rat thymocyte cDNA libraries (Thomas *et al.*, 1985; Barclay *et al.*, 1987) plated on nitrocellulose using five degenerate oligonucleotide probes essentially as described in Johnson *et al.* (1985), with other standard procedures as in Maniatis *et al.* (1982). Hybridization was at 4°C below the Th temperature as calculated for the oligonucleotide in each mixture with lowest G+C content [Th = 4(G+C) + 2 × (A+T)°C]. Filters were washed at the Th temperature for 1 h. cDNA clones were fragmented by sonication or restriction enzyme digestion and cloned into M13 for sequencing by the dideoxy method (Messing, 1983; Biggin *et al.*, 1983). Nucleic acid sequences were analysed with the programs of Staden (1986).

### *Northern blot analysis*

Total RNA from thymocytes, cortical thymocytes, activated T cells and a $CD4^+$ leukaemia cell line (Clark *et al.*, 1987) was prepared by the guanidine isothiocyanate method (Chirgwin *et al.*, 1979) and 10 $\mu$g samples were electrophoresed on a 1.2% agarose gel in formaldehyde. RNA was transferred to a Genescreen membrane (New England Nuclear, USA) and hybridized with probes labelled with [$^{32}$P]dATP by random hexanucleotide priming as described (M.Morris *et al.*, in preparation). Similar results were obtained with a coding region probe (*Bam*HI-*Eco*RI 1.1 kb fragment from pLSGP2) or the complete insert (*Bam*HI-*Hin*dIII 3.5 kb from pLSGP10).

## Acknowledgements

We thank Michael Morris for Northern blots, Tony Gascoyne for amino acid analyses, Keith Gould for help with the synthesis of oligonucleotides, William Brown, David Jackson and Pauline Johnson for helpful discussions, Stan Buckingham and Catherine Lee for photography and Denise Roby for manuscript preparation. Nigel Killeen was supported by the Leukaemia Research Fund.

## References

Axelsson,B., Hammerstrom,S., Finne,J. and Perlmann,P. (1985) *Eur. J. Immunol.*, **15**, 427-433.

Barclay,A.N., Jackson,D.I., Willis,A.C. and Williams,A.F. (1987) *EMBO J.*, **6**, 1259-1264.

Biggin,M.D., Gibson,T.J. and Hong,G.F. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 3963-3965.

Brown,W.R.A., Barclay,A.N., Sunderland,C.A. and Williams,A.F. (1981) *Nature*, **289**, 456-460.

Brown,W.R.A. and Williams,A.F. (1982) *Immunology*, **46**, 713-726.

Carlsson,S.R. and Fukuda,M. (1986) *J. Biol. Chem.*, **261**, 12779-12786.

Carlsson,S.R., Sasaki,H. and Fukuda,M. (1986) *J. Biol. Chem.*, **261**, 12787-12795.

Clark,S.J., Jefferies,W.A., Barclay,A.N., Gagnon,J. and Williams,A.F. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 1649-1653.

Chirgwin,J.M., Przybla,A.E., MacDonald,R.J. and Rutter,W.J. (1979) *Biochemistry*, **18**, 5294-5299.

Conzelman,A., Pink,R., Acuto,D., Mach,J.-P., Dolivo,S. and Nabholz,M. (1980) *Eur. J. Immunol.*, **10**, 860-868.

De Petris,S. (1984) *Exp. Cell Res.*, **152**, 510-519.

De Maio,A., Lis,H., Gershoni,J.M. and Sharon,N. (1986) *FEBS Lett.*, **194**, 28-32.

Gahmberg,C.G., Häyry,P. and Andersson,L.C. (1976) *J. Cell Biol.*, **68**, 642-653.

Gardiner-Garden,M. and Frommer,M. (1987) *J. Mol. Biol.*, **196**, 261-282.

Jefferies,W.A., Green,J.R. and Williams,A.F. (1985) *J. Exp. Med.*, **162**, 117-127.

Johnson,P., Gagnon,J., Barclay,A.N. and Williams,A.F. (1985) *EMBO J.*, **4**, 2539-2545.

Kehry,M., Ewald,S., Douglas,R., Sibley,C., Raschke,W., Fambrough,D. and Hood,L. (1980) *Cell*, **21**, 393-406.

Kenney,D., Cairns,L., Remold-O'Donnell,E., Peterson,J., Rosen,F.S. and

Parkman,R. (1986) *Blood*, **68**, 1329–1332.

Kitajima,K., Inoue,Y. and Inoue,S. (1986) *J. Biol. Chem.*, **261**, 5262–5269.

Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) *Molecular Cloning, A Laboratory Manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

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

Messing,J. (1983) *Methods Enzymol.*, **101**, 20–78.

Ralph,S.J., Thomas,M.L., Morton,C.C. and Trowbridge,I.S. (1987) *EMBO J.*, **6**, 1251–1257.

Remold-O'Donnell,E., Kenney,D.M., Parkman,R., Cairns,L., Savage,B. and Rosen,F.S. (1984) *J. Exp. Med.*, **159**, 1705–1723.

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

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

Segrest, J.P. and Jackson R.L. (1972) *Methods Enzymol.*, **28**, 54–63.

Saito,M. and Osawa,T. (1980) *Carbohydrate Res.*, **78**, 341–348.

Shibuya,K., Noguchi,S., Yamaki,M., Nishimura,S. and Sekiya,T. (1985) *J. Biochem.*, **97**, 1719–1725.

Smith,G. (1976) *Science*, **191**, 528–535.

Staden,R. (1986) *Nucleic Acids Res.*, **14**, 217–231.

Standring,R., McMaster,W.R., Sunderland,C.A. and Williams,A.F. (1978) *Eur. J. Immunol.*, **8**, 832–839.

Sunderland,C.A., McMaster,W.R. and Williams,A.F. (1979) *Eur. J. Immunol.*, **9**, 155–159.

Swallow,D.M., Gendler,S., Griffiths,B., Corney,G., Taylor-Papadimitriou,J. and Bramwell,M. (1987) *Nature*, **328**, 82–84.

Thomas,M.L., Barclay,A.N., Gagnon,J. and Williams,A.F. (1985) *Cell*, **41**, 83–93.

von Heijne,G. (1986) *Nucleic Acids Res.*, **14**, 4683–4690.

Weller,P., Jeffreys,A.J., Wilson,V. and Blanchetot,A. (1984) *EMBO J.*, **3**, 439–446.

Williams,A.F., Galfre,G. and Milstein,C. (1977) *Cell*, **12**, 663–673.

*Received on August 17, 1987; revised on September 25, 1987*

## Note added in proof

These sequence data have been submitted to the EMBL/GenBank Data libraries under the accession number Y00090