Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 7876–7879, September 1992
Medical Science

Structure of the CD59-encoding gene: Further evidence of a relationship to murine lymphocyte antigen Ly-6 protein

(human late complement inhibitor gene)

JOHN G. PETRANKA, DONALD E. FLEENOR, KATHRYN SYKES*, RUSSEL E. KAUFMAN, AND WENDELL F. ROSSE†

Division of Hematology/Oncology, Department of Medicine, Duke University Medical Center, Durham, NC 27710

Communicated by W. K. Joklik, May 22, 1992

ABSTRACT **The gene for CD59 [membrane inhibitor of reactive lysis (MIRL), protectin], a phosphatidylinositol-linked surface glycoprotein that regulates the formation of the polymeric C9 complex of complement and that is deficient on the abnormal hematopoietic cells of patients with paroxysmal nocturnal hemoglobinuria, consists of four exons spanning 20 kilobases. The untranslated first exon is preceded by a G+C-rich promoter region that lacks a consensus TATA or CAAT motif. The second exon encodes the hydrophobic leader sequence of the protein, and the third exon encodes the amino-terminal portion of the mature protein. The fourth exon encodes the remainder of the mature protein, including the hydrophobic sequence necessary for glycosyl–phosphatidylinositol anchor attachment. The structure of the CD59 gene is very similar to that encoding Ly-6, a murine glycoprotein with which CD59 has some structural similarity. The striking similarity in gene structure is further evidence that the two proteins belong to a superfamily of proteins that may also include the urokinase plasminogen-activator receptor and a squid glycoprotein of unknown function.**

CD59 is an 18-kDa membrane protein (1–4) that inhibits the lysis of blood cells by activated complement (5–7) through an inhibition of the formation of the polymeric C9 complex (8–10). CD59 antigen is affixed to the membrane by a glycosyl–phosphatidylinositol anchor (11) and, like all other proteins affixed in this way, is deficient on the blood cells of patients with paroxysmal nocturnal hemoglobinuria (12). The primary structure of the protein has been deduced from the cDNA sequence (13–17). The structure is characterized by a hydrophobic signal sequence of 25 amino acid residues, a cysteine-rich sequence with at least two potential N-glycosylation sites, and a hydrophobic carboxyl terminus characteristic of proteins that posttranslationally attach the glycosyl-phosphatidylinositol anchor.

The amino acid sequence of CD59 has 24–30% similarity to a family of murine proteins called Ly-6 (13, 15–17). In particular, the amino- and carboxyl-terminal sequences of the mature proteins show an even greater regional similarity. In addition, all 10 cysteines of the mature proteins are placed in comparable positions; these cysteines appear to form intramolecular disulfide bonds (18), the disruption of which inactivates CD59 (19).

The function of Ly-6 proteins is not known (20). Reaction with monoclonal antibodies to epitopes on the proteins results in activation of the cell characterized by Ca^{2+} influx and depolarization (21). Similar responses have been found when monoclonal antibodies are bound to CD59 (4, 22).

We report the cloning and characterization of the CD59 gene.‡ The gene consists of four exons spanning ≈20 kilobases (kb). The promoter region is G+C rich and does not contain consensus CAAT or TATA motifs. Northern (RNA) and cDNA analyses indicate that the gene encodes four mature transcripts that arise from alternative use of different polyadenylylation signals. Three polyadenylylation sites had previously been identified; we have identified a fourth site corresponding to a longer species of mRNA. The CD59 gene organization closely parallels that of the murine *Ly-6* genes (23), thereby providing further evidence that CD59 and Ly-6 are closely related.

METHODS

Isolation and Characterization of cDNA and Genomic Clones. A HeLa cDNA library constructed in Λ ZAPII (Stratagene) was probed with a 1.0-kb CD59 cDNA probe generated by gene-specific PCR amplification of HeLa cDNA. Positively hybridizing clones were subcloned by *in vivo* excision into the pBluescript SK^- phagemid for restriction analysis and sequencing. Dideoxynucleotide sequencing was done by using the Sequenase system (United States Biochemical) with universal, reverse, and/or specific oligonucleotide primers.

A human placental DNA genomic library prepared in Lambda FixII (Stratagene) was screened with the 1.0-kb CD59 cDNA probe. Positively hybridizing, overlapping genomic clones were isolated, subcloned into plasmid vectors, and characterized by restriction analysis, Southern blotting, and dideoxynucleotide chain-termination sequencing. Analysis of the DNA sequence was facilitated by the Genetics Computer Group programs.

Primer-extension analysis. Primer-extension analysis was done as described by Townes *et al.* (23) with 5 μg of poly(A)$^+$ mRNA isolated from HeLa cells.

RESULTS AND DISCUSSION

Extension of cDNA Sequence. Four species of CD59 mRNA have been identified with mobilities corresponding to 600 base pairs (bp), 1.2 kb, 1.9 kb, and 2.2 kb (13, 15); the last of these is variably present. Consensus polyadenylylation signals have been located for the first three forms, and cDNA clones containing a terminal poly(A) sequence have been isolated for each (13, 14). To identify cDNA clones that correspond to the largest mRNA, a HeLa cDNA library was probed with a 1.0-kb CD59 cDNA clone generated by reverse transcription–PCR. Twelve positive clones were identified, of which four hybridized with an oligonucleotide correspond-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

*Present address: Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, TX 75235.
†To whom reprint requests should be addressed at: Box 3934, Duke University Medical Center, Durham, NC 27710.
‡The sequence reported in this paper has been deposited in the GenBank data base (accession no. M82840).

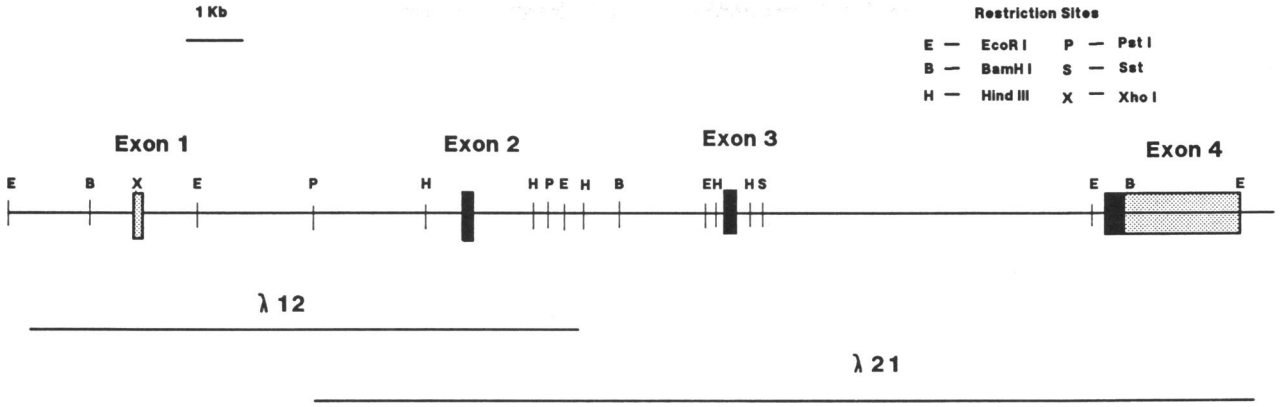

FIG. 1. Structure of CD59-encoding gene. Exons are indicated by boxes; translated portions are solid, and untranslated portions are shaded. The restriction sites used in assessment of the structure are indicated.

ing to the 3′ end of the cDNA sequence described by Davies *et al.* (13). Two of the clones share identical 3′ termini and extend the previously reported sequence of Sawada *et al.* (14) for an additional 177 bp. The clones contain a potential polyadenylylation signal (AAATAGAA) just upstream, although poly(A) sequences were not found on either clone. These clones likely represent the fourth and largest species of CD59 mRNA.

Determination of CD59 Gene Structure. The gene for CD59 was found contained within two overlapping clones from a human placental genomic library that span a 20-kb region of genomic DNA. Southern analysis with oligonucleotides derived from the known cDNA sequence showed that the gene is partitioned into four exons separated by three lengthy introns (Fig. 1); the exons are flanked by appropriate consensus acceptor or donor splice sites (Fig. 2). The first exon is 45 bp in length, is entirely untranslated, and is separated from exon 2 by an intron of 5.4 kb. Exon 2 is 85 bp in length and contains a short (18 bp) untranslated region followed by the sequence coding for most of the hydrophobic-signal

```
ggatcccctggtcttgctcttctgtccctgcctctggtggctgctggggaacagtagctaccagctaagttgatagagacactcgaagaggttcattcaaaacactggtccccgagcagc -549
ccagtaagcttttaaacgcataacattttgtttgggtctcaaacaagtcacaaggcacggatacggatgcggctgtggctgatggctatttggatgctgcccaagcctcgcagtaggaag -429
cagcttcagactgcagcagggccgcatagcccgatgccaggccctggggcacccagggcaccagcaagacgcacagaaacgttttgaaaaacattttaatatacttaaaaaaaaagtaaa -309
agggaaaatcagaagtctttggaagtttcctactgtttttatgtcccatagcaaatccgaggaggacgccccaaactttcagttccctggggttttgaaggtgctcattgggtcctggcc -189
                                                    AP-1
acccggccttctcagaacctgggccaggaggctgagctccgcgcgggggtggagggagaggaggaggttcctgccgaggtgcggctgcgcggtgcgggagccgggagcctgggagggcaa  -69
                Sp1         Sp1                 *       *       *              *  ****
gggcatcctgagggggcggggccgggggcggagccttgcgggctggagcgaaagaatgcgggggctgagCGCAGAAGCGGCTCGAGGCTGGAAGAGGATCTTGGGCGCCGCCAGgtaagaa   52
ggcccgaaggcctgctggggtttgggtgagccgagccgaggtggcggcggcagtctgggcccgggcggagggtctgtgggggcatccccgtcctcgctttccactcgacccctgcc...   169
                                                                                                          *   *
.......aaatgttcctctcagaagccttcttggccttccagcccttggtttttgagacaaccagcagtcatttgttcgttcctgacattccttcctgtccctccttccagGTTCTGTG  282
          M  G  I  Q  G  G  S  V  L  F  G  L  L  L  V  L  A  V  F  C  H  S                                          22
GACAATCACAATGGGAATCCAAGGAGGGTCTGTCCTGTTCGGGCTGCTGCTCGTCCTGGCTGTCTTCTGCCATTCAGgtgagtgctccagtctcaggacatggaatctaggctgccttgg  402
ccatgaactcccttcttaagcctcagtttcagccccagctgctcctccaggcctggatttggttctttt.....aagcttcctaattgcacagttagatggatattcaaatgtgg  512
atatttcctttgtgcttgcatcagggtccaaaaatgttgctggaactgtagtttgaatttggaaagatataactgctatacatttgtgcgggagtggaatgataccacaagttgctgact  632
                                                                                  G  H  S  L  Q  C  Y  N  C  P  N  P   34
ttgggcccatattaatggagatggtgggctgccaggggacaagtcagtgctgctttaagagatcctgactttcttcctgattctagGTCATAGCCTGCAGTGCTACAACTGTCCTAACCC  752
  T  A  D  C  K  T  A  V  N  C  S  S  D  F  D  A  C  L  I  T  K  A                                              56
AACTGCTGACTGCAAAACAGCCGTCAATTGTTCATCTGATTTTGATGCGTGTCTCATTACCAAAGCTGgtaagagcctcccctgtctgtctcctaaatgtaatggggtaataagtgcctg  872
ggaaaaaaattgtgccactgtaatcctcattaggcttgcatcaagtaatagctaccatttattagcgtcaggtgtgtattgggctaagcacttcttcatacattttaaatttaatgtttt  992
gca.....tgttcctcctccgtcccacccccataactatactggctctgatgagaccttggttttctgtaaaagctctatttagaggtgtatcattatttacttaattgttctcc 1102
tttacaacccacctgggatgagcatcttgcctagaagtctctacttgcacaggatacatacgaaatagattgaggattcaaagcagatacagaactcttcccacttactttcttaccctg 1222
             G  L  Q  V  Y  N  K  C  W  K  F  E  H  C  N  F  N  D  V  T  T  R  L  R  E  N  E  L  T  Y  Y  C  C  K  K    91
tgtgtctccccacagGGTTACAAGTGTATAACAAGTGTTGGAAGTTTGAGCATTGCAATTTCAACGACGTCACAACCCGCTTGAGGGAAAATGAGCTAACGTACTACTGCTGCAAGAAGG 1342
D  L  C  N  F  N  E  Q  L  E  N  G  G  T  S  L  S  E  K  T  V  L  L  L  V  T  P  F  L  A  A  A  W  S  L  H  P  *         129
ACCTGTGTAACTTTAACGAACAGCTTGAAAATGGTGGGACATCCTTATCAGAGAAAACAGTTCTTCTGCTGGTGACTCCATTTCTGGCAGCAGCCTGGAGCCTTCATCCCTAAGTCAACA 1462
                                                                                                                  #
CCAGGAGAGCTTCTCCCAAACTCCCCGTTCCTGCGTAGTCCGCTTTCTCTTGCTGCCACATTCTAAAGGCTTGATATTTTCCAAATGGATCCTGTTGGGAAAGAATAAAATTAGCTTGAG 1582
#  #
CAACCTGGCTAAGATAGAGGGGCTCTGGGAGACTTTGAAGACCAGTCCTGTTTGCAGGGAAGCCCCACTTGAAGGAAGAAGTCTAAGAGTGAAGTAGGTGTGACTTGAACTAGATTGCAT 1702
GCTTCCTCCTTTGCTCTTGGGAAGACCAGCTTTGCAGTGACAGCTTGAGTGGGTTCTCTGCAGCCCTCAGATTATTTTTCCTCTGGCTCCTTGGATGTAGTCAGTTAGCATCATTAGTAC 1822
ATCTTTGGAGGGTGGGGCAGGAGTATATGAGCATCCTCTCTCACATGGAACGCTTTCATAAACTTCAGGGATCCCGTGTTGCCATGGAGGCATGCCAAATGTTCCATATGTGGGTGTCAG 1942
TCAGGGACAACAAGATCCTTAATGCAGAGCTAGAGGACTTCTGGCAGGGAAGTGGGGAAGTGTTCCAGATAGCAGGGCATGAAAACTTAGAGAGGTACAAGTGGCTGAAAATCGAGTTTT 2062
                                                                                                                  #
TCCTCTGTCTTTAAATTTTATATGGGCTTTGTTATCTTCCACTGGAAAAGTGTAATAGCATACATCAATGGTGTGTTAAAGCTATTCCTTGCCTTTTTTTTTATTGGAATGGTAGGATAT 2182
CTTGGCTTTGCCACACACAGTTACAGAGTGAACACTCTACTACATGTGACTGGCAGTATTAAGTGTGCTTATTTTAAATGTTACTGGTAGAAAGGCAGTTCAGGTATGTGTGTATATAGT 2302
ATGAATGCAGTGGGGACACCCTTTGTGGTTACAGTTTGAGACTTCCAAAGGTCATCCTTAATAACAACAGATCTGCAGGGGTATGTTTTACCATCTGCATCCAGCCTCCTGCTAACTCCT 2422
AGCTGACTCAGCATAGATTGTATAAAATACCTTTGTAACGGCTCTTAGCACACTCACAGATGTTTGAGGCTTTCAGAAGCTCTTCTAAAAAATGATACACACCTTTCACAAGGGCAAACT 2582
TTTTCCTTTTCCCTGTGTATTCTAGTGAATGAATCTCAAGATTCAGTAGACCTAATGACATTTGTATTTTATGATCTTGGCTGTATTTAATGGCATAGGCTGACTTTTGCAGATGGAGGA 2662
   #   #   #         #                  #
ATTTCTTGATTAATGTTGAAAAAAAACCCTTGATTATACTCTGTTGGACAAACCGAGTGCAATGAATGATGCTTTTCTGAAAATGAAATATAACAAGTGGGTGAATGTGGTTATGGCCGA 2782
                                                                                                                  #
AAAGGATATGCAGTATGCTTAATGGTAGCAACTGAAAGAAGACATCCTGAGCAGTGCCAGCTTTCTTCTGTTGATGCCGTTCCCTGAACATAGGAAAATAGAAACTTGCTTATCAAAACT 2902
TAGCATTACCTTGGTGCTCTGTGTTCTCTGTTAGCTCAGTGTCTTTCCTTACATCAATAGGTTTTTTTTTTTTTTTTGGCCTGAGGAAGTACTGACCATGCCCACAGCCACCGGCTGAG 3022
CAAAGAAGCTCATTTCATGTGAGTTCTAAGGAATGAGAAACAATTTTGATGAATTTAAGCAGAAAATGAATTTCTGGGAACTTTTTTGGGGGCGGGGGGGTGGGGAATTC          3132
```

FIG. 2. Nucleotide sequence of CD59-encoding gene. Exons are denoted by uppercase letters; translated sequence is denoted by boldface type. Consensus donor and acceptor sites within introns are underlined. The 5′ end of known cDNAs are indicated by stars (boldface when more than one is found at the same site; the 3′ ends of known cDNAs are indicated by the symbol # (boldface when more than one is found at the same site). The 3′-terminal sequence derived from genomic clones is indicated by italics. Potential polyadenylylation signals are indicated by double underlining. The translated protein sequence is indicated by single-letter amino acid designations.

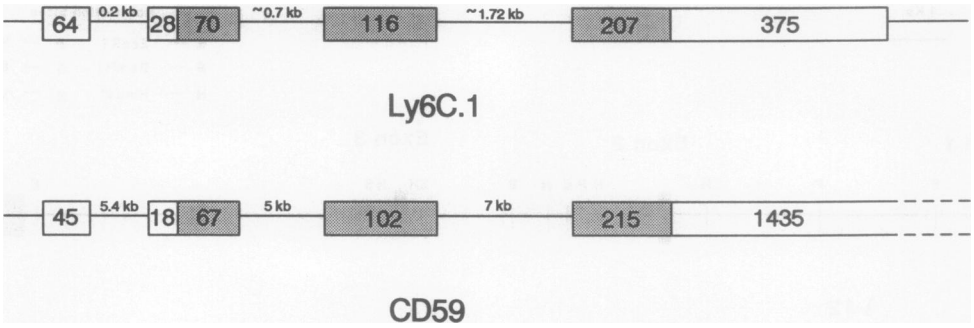

FIG. 3. Similarities in the structure of the gene for CD59 and that for Ly6C.1 (23). The number of nucleotide base pairs in each exon is given; translated exons and portions of exons are shaded.

sequence of the protein. Exons 2 and 3 are separated by an intron of ≈5 kb. Exon 3 is 102 bp in length and codes for a small portion of the leader sequence and the amino-terminal 31 amino acids of the mature protein. Exon 3 is separated from exon 4 by an intron of 7 kb. Exon 4 encodes the remainder of the translated sequence as well as, at least, 1438 bp of untranslated sequence.

The presumed promoter region 5′ to exon 1 is G+C rich (78% of the first 100 bases) and does not contain TATA or CAAT boxes but does contain duplicate transcription factor Sp1 sites 55 and 43 bases upstream from the beginning of the first exon. The 5′ end of exon 1 was determined by primer extension to correspond to that of the cDNA described by Davies *et al.* (13). One primer-extension study suggested that mRNA transcription could alternatively be initiated at the beginning of exon 2, a conclusion supported by the fact that several cDNA clones previously described extend only to the beginning of exon 2; however, no such initiation site was found in other studies.

Two differences were found between the genomic sequence and the sequence of previously published cDNA. In the first exon, the published cDNA contains a *Bam*HI restriction site at bp 27–33; this site is not present in the genomic sequence, as GGATCC is replaced by GGATCT. *Bam*HI did not cleave the clone from genomic DNA at this site; this change apparently represents polymorphism at this position. Further, at base 2106 of the cDNA described by Davies *et al.* (13), a duplicated TCCAGAT sequence is indicated. This sequence is not duplicated in the genomic sequence in our studies [or in the cDNA sequence reported by Sawada *et al.* (14)].

The amino acid sequence of CD59 has some similarity to that of the Ly-6 series of murine proteins; ≈25% of the residues are identical. This similarity is especially strong in the leader sequence (exon 2 of CD59), the amino terminus of the mature protein (exon 3), and the hydrophobic region near the site of glycosyl–phosphatidylinositol-anchor addition (exon 4). Further, both proteins contain 10 cysteines that are found in comparable positions in the protein (13–17, 22).

The organization of the genes of the murine Ly-6 proteins is strikingly similar to that of CD59 antigen (22). The genes are each divided into four exons that are remarkably similar in size to their counterparts, the first exon is untranslated, and the second exon encodes for the respective leader sequences (Fig. 3). When the proteins are aligned according to the positions of the cysteine residues, the exon splice sites occur at nearly the same positions. Further, similarity (52%) not previously noted occurs in the nucleotide sequence in the first, untranslated exon. Finally, in both genes, mRNA transcription may alternatively commence at the beginning of exon 2 (29).

On the other hand, there are considerable differences in the promoter region of the two genes. The promoter region of CD59 is G+C rich, whereas that of Ly-6 is not. The *Ly-6* promoter contains consensus sequences related to inducibility by interferon, whereas that of CD59 does not; this difference might account for the fact that CD59 is apparently not induced by interferon (15).

From these data, CD59 clearly belongs to the same family of proteins that includes the murine Ly-6 proteins. Other members of the family have been identified—the human urokinase plasminogen-activator receptor (which has three repeats of the homologous, cysteine-rich structure) (24) and a squid glycoprotein of unknown function (25). All members of the family are linked to the membrane by the glycosyl–phosphatidylinositol anchor (24, 26). Where the function of these proteins is known, it involves protein–protein interactions. Thus, the cysteine-rich structure of this superfamily may provide the basis for repeating units that mediate protein binding, analogous to the short consensus repeats of the complement-binding proteins (27) and the "kringles" of the coagulation-system proteins (28).

1. Stefanova, I., Hilgert, I., Kristofova, H., Brown, R., Low, M. G. & Horejsi, V. (1989) *Mol. Immunol.* **26**, 153–161.
2. Hadam, M. R. (1989) in *Leucocyte Typing IV: Handbook of Experimental Immunology*, ed. Knall, W. (Oxford Univ. Press, Oxford), pp. 720–722.
3. Groux, H., Huet, S., Aubrit, F., Tran, H. C., Boumsell, L. & Bernard, A. (1989) *J. Immunol.* **142**, 3013–3020.
4. Okada, N., Harada, R., Fujita, T. & Akada, H. (1989) *J. Immunol.* **143**, 2262–2266.
5. Sugita, Y., Nakano, Y. & Tomita, M. (1988) *J. Biochem.* **104**, 633–637.
6. Holguin, M. H., Frederick, L. R., Bernshaw, N. J., Wilcox, L. A. & Parker, C. J. (1989) *J. Clin. Invest.* **84**, 7–17.
7. Okada, N., Harada, R., Fujita, T. & Okada, H. (1989) *Int. Immunol.* **1**, 205–208.
8. Rollins, S. A. & Sims, P. J. (1990) *J. Immunol.* **144**, 3478–3483.
9. Meri, S., Morgan, B. P., Davies, A., Daniels, R. H., Olavesen, M. G., Waldmann, H. & Lachmann, P. J. (1990) *Immunology* **71**, 1–9.
10. Whitlow, M. B., Iida, K., Stefanova, I., Bernard, A. & Nussenzweig, V. (1990) *Cell. Immunol.* **126**, 176–184.
11. Holguin, M. H., Wilcox, L. A., Bernshaw, N. J., Rosse, W. F. & Parker, C. J. (1990) *Blood* **75**, 284–289.
12. Rosse, W. F. (1990) *Blood* **75**, 1595–1601.
13. Davies, A., Simmons, D. L., Hale, G., Harrison, R. A., Tighe, H., Lachmann, P. J. & Waldmann, H. (1989) *J. Exp. Med.* **170**, 637–654.
14. Sawada, R., Ohasi, K., Anaguchi, H., Okazaki, H., Hattori, M., Minato, N. & Naturo, M. (1990) *DNA Cell Biol.* **9**, 213–220.
15. Philbrick, W. M., Palfree, R. G. E., Maher, S. E., Bridgett, M. M., Serlin, S. & Bothwell, A. L. M. (1990) *Eur. J. Immunol.* **20**, 87–92.
16. Sugita, Y., Tobe, T., Oda, E., Tomita, M., Yasukawa, K., Jamaji, N., Takemoto, T., Furuichi, K., Takayama, M. & Yano, S. (1989) *J. Biochem.* **106**, 555–557.
17. Okada, H., Nagami, Y., Takahashi, K., Okada, N., Hideshima, T., Takizawa, H., Kondo, J. (1989) *Biochem. Biophys. Res. Commun.* **162**, 1552–1559.
18. Tomita, M., Tobe, T., Choi-Miura, N., Nakano, Y., Kusano, M. & Oda, E. (1991) *Complement Inflammation* **8**, 233 (abstr.).

19. Ezzell, J. L., Wilcox, L. A., Bernshaw, N. J. & Parker, C. J. (1991) *Blood* **77**, 2764–2773.
20. Shevach, E. M. & Korty, P. E. (1989) *Immunol. Today* **10**, 195–200.
21. Malek, T. R., Ortega, G., Chan, C., Kroszek, R. A. & Shevach, E. M. (1986) *J. Exp. Med.* **164**, 709–722.
22. Korty, P. E., Brando, C. & Shevach, E. M. (1991) *J. Immunol.* **146**, 4092–4098.
23. Townes, T. M., Lingrel, J. B., Chen, H. Y., Brinster, R. I. & Palmiter, R. D. (1985) *EMBO J.* **4**, 1715–1723.
24. Roldan, A. L., Cubellis, M. V., Masucci, M. T., Behrendt, N., Lund, L. R., Danø, K., Appella, E. & Blasi, F. (1990) *EMBO J.* **9**, 467–474.
25. Williams, A. F., Tse, A. G.-D. & Gagnon, J. (1988) *Immunogenetics* **27**, 265–272.
26. Ploug, M., Rønne, E., Behrendt, N., Jensen, A. L., Blasi, F. & Danø, K. (1991) *J. Biol. Chem.* **266**, 1926–1933.
27. Campbell, R. D., Law, S. K. A., Reid, K. B. M. & Sim, R. B. (1988) *Annu. Rev. Immunol.* **6**, 161–195.
28. Park, C. H. & Tulinsky, A. (1986) *Biochemistry* **25**, 3977–3982.
29. Bothwell, A., Pace, P. E. & LeClair, K. P. (1988) *J. Immunol.* **140**, 2815–2820.