ORIGINAL PAPER

Keri A. Multerer · L. Courtney Smith

Two cDNAs from the purple sea urchin, *Strongylocentrotus purpuratus*, encoding mosaic proteins with domains found in factor H, factor I, and complement components C6 and C7

Received: 10 December 2003 / Revised: 26 February 2004 / Published online: 16 April 2004 © Springer-Verlag 2004

Abstract The vertebrate complement system is composed of about 30 serum and cell surface proteins that make up three activation pathways, a lytic pathway, and a set of proteins that regulate complement. Regulatory proteins are required for host protection against autologous complement attack and to control the amplification feedback loop of the alternative pathway. Purple sea urchin, Strongylocentrotus purpuratus, homologues of complement C3 (SpC3) and factor B (SpBf) have been identified, suggesting the presence of an alternative complement pathway. This implies that echinoderms require a complement regulatory system for the same reasons that it is required in higher vertebrates. Two cDNAs, Sp5 and Sp5013, have been characterized from coelomocytes and the deduced structures of the encoded mosaic proteins, SpCRL (S. purpuratus complement related protein, long form) and SpCRS (short form), have domains that are also found in regulatory proteins such as factor H and factor I and the terminal pathway components C6 and C7. These domains include multiple short consensus repeats, a fucolectin domain, Ser/Thr/Pro-rich regions, a Cys-rich region, and a factor I-membrane attack complex domain. The genes are constitutively expressed in all tissues of the sea urchin and are not induced in response to immune challenge. Multiple bands of varying intensity on both genome blots and RNA

K. A. Multerer · L. C. Smith ()∞) Graduate Program in Genetics, The Institute of Biomedical Sciences, George Washington University, Washington, DC 20052, USA e-mail: csmith@gwu.edu Tel.: +1-202-9949211 Fax: +1-202-994-6100

L. C. Smith
Department of Biological Sciences,
George Washington University,
340 Lisner Hall, 2023 G Street NW, Washington, DC 20052, USA

Present address: K. A. Multerer, Human Biology Division, Fred Hutchinson Cancer Center, Seattle, WA 98109, USA blots suggest that *Sp5* and *Sp5013* are members of a small gene family and that they might undergo alternative splicing. Based on the domains present in SpCRL and SpCRS, they might be either examples of complement regulatory proteins or members of the terminal pathway of complement.

Keywords Evolution · Innate · Echinoderm · Complement

Introduction

All multicellular organisms have some form of innate immunity that functions in the identification of, and protection against, invading pathogens and parasites. The complement system, which is composed of about 35 serum and cell surface proteins (Volanakis 1998), is an important component of the more ancient innate immune system. It has been identified throughout the lineage of deuterostome animals (Smith et al. 1999; Nonaka 2001; Azumi et al. 2003) and is being discovered in the rest of the animal kingdom as well, such as in a gorgonian (accession no. AAN86548) and a squid (M. McFall-Ngai, personal communication). There are three activation pathways-classical, alternative, and lectin-that are activated by different types of molecules and lead to the activation of the terminal or lytic pathway. The activation pathways culminate in the formation of two distinct C3convertase enzymes: the alternative pathway convertase (C3bBb) and the classical pathway convertase (C4bC2b) (reviewed by Nonaka et al. 1998; Xu et al. 2001). The formation of C3-convertases leads to the formation of C5-convertases (Pangburn and Rawal 2002) that activate the lytic pathway. C3-convertases function to cleave and activate additional C3, creating a positive feedback loop within the alternative pathway that accelerates the activation of the entire complement system (Liszewski et al. 1996), resulting in quick and efficient opsonization and lysis of foreign cells (Lambris 1988; Becherer et al. 1989).

Because complement proteins C3 and C4 can form covalent thioester bonds with amines and hydroxyls on any molecule, they have the ability to bind any surface, including self, which would lead to inappropriate cell lysis and inflammatory reactions. Covalent bond formation between C3 or C4 and non-self is an important mechanism for identifying pathogens. Therefore, a complement regulatory system is necessary to both protect self-cells against autologous complement attack and, consequently, to direct the attack towards foreign pathogens. Tight regulation of active C3-convertases is also essential to prevent unnecessary depletion of complement proteins from uncontrolled activation, as well as to inhibit complement activities after a pathogen has been cleared (Liszewski et al. 1996). One important mechanism of regulation is based on dissociating the C3-convertase complexes and by degrading C3. Some of the proteins involved are membrane cofactor protein (MCP), complement receptors 1 and 2 (CR1 and CR2), decay accelerating factor (DAF), C4 binding protein (C4BP), factor H, and factor I (Medof et al. 1987; Krych et al. 1991; Liszewski et al. 1996; Arlaud et al. 1998; Kirkitadze and Barlow 2001; Barilla-LaBarca et al. 2002).

Sea urchin immune response

The sea urchin defends itself against pathogens with its innate immune system. Coelomocytes are the immune effector cells (Boolotian and Geise 1958; Johnson 1969; Gross et al. 2000) that carry out many protective functions including encapsulation, phagocytosis, chemotaxis, and expression of putative cytotoxic agents in response to invasion of foreign cells (Smith et al. 1992; Smith and Davidson 1994). There are four morphologically distinct classes of coelomocytes (Johnson 1969; Edds 1993) of which the phagocytes appear to be the major immune effector cells. Phagocytes express Sp064, which encodes SpC3, a homologue of complement component C3 (Al-Sharif et al. 1998; Gross et al. 2000), and Sp152, which encodes SpBf, a homologue of complement component factor B (Bf) (Smith et al. 1998; unpublished data). True thioester activity, similar to that characterized for C3 in higher vertebrates, has been demonstrated chemically for SpC3 through autolytic fragmentation and thioester inactivation with methylamine (Smith 2002). In biological assays, SpC3 functions as an opsonin and augments phagocytosis of yeast by coelomocytes (Smith 2001; Clow et al. 2004). The complement system in the sea urchin has been proposed to function like an alternative pathway with the formation of a C3-convertase complex from SpC3 and SpBf, resulting in a feedback loop to augment the rate of opsonization (Smith et al. 1999; Smith 2001). This suggests that the thioester binding activity and the convertase functions would require a mechanism for protection against autologous attack by controlling or limiting the convertase activity that would avoid depletion of the components. Initial evidence for a complement regulatory system in the sea urchin was obtained from the sequence of SpC3, which has two conserved cleavage sites for factor I-like activity that are located in conserved positions (Al-Sharif et al. 1998). This predicts that additional complement components function in the sea urchin.

In the present study, we show the analysis of two cDNAs, Sp5 (accession no. AY494840) and Sp5013 (accession no. AY494841), which encode proteins SpCRL (S. purpuratus complement related protein, long form) and SpCRS (short form), respectively. Deduced amino acid sequences from both proteins show multiple short consensus repeats (SCRs) (18 in SpCRL and four in SpCRS), two Ser/Thr/Pro (S/T/P)-rich regions, and a factor I-membrane attack complex (FIMAC) domain. In addition, SpCRL has a fucolectin domain and a Cys-rich region. Both genes are expressed in coelomocytes, gut, gonad, pharynx, esophagus, and axial organ, and the level of message accumulation in coelomocytes before and after immune challenge suggests that expression is constitutive. Northern blots for both messages show multiple bands of varying intensity, suggesting the presence of a small family of genes with similar sequences in addition to possible alternative splicing. Genome blots for both genes also show multiple bands of varying intensity consistent with gene structure of multiple exons and perhaps with cross-hybridization with other members of the gene family. Structural and phylogenetic analyses of the deduced amino acid sequences for these two proteins indicate that they share domains with a number of complement proteins from higher vertebrates including factor H, factor I, C6 and C7. These results imply functions within the complement system, possibly as complement regulatory proteins and/or within a primitive terminal pathway in the sea urchin.

Materials and methods

Animals

Purple sea urchins, *Strongylocentrotus purpuratus*, were obtained and housed as previously described (Gross et al. 2000; Shah et al. 2003).

RNA isolation

Total RNA from gut, gonad, esophagus, pharynx, axial organ, and coelomocytes was isolated using the RNAeasy Midi kit (Qiagen, Valencia, Calif.). Animals were sacrificed by removing Aristotle's lantern (mouth parts), and the coelomic fluid was poured from the body cavity through sterile cheese cloth and mixed in a 1:1 ratio with ice-cold, Ca²⁺- and Mg²⁺-free sea water containing 70 mM EDTA and 50 mM imidazole (pH 7.4) (CMFSW-EI) according to Gross et al. (1999). Coelomocytes were pelleted by centrifugation at 6,500 g for 5 min at 4°C. To harvest the internal organs, the test was cut open and tissues were removed from the coelomic cavity and placed in RNAlater (Ambion, Austin, Tex.). Pelleted coelomocytes and approximately 0.24 g of each solid tissue were lysed in guanidine isothiocyanate lysis buffer (Qiagen). The lysate was bound to a silica gel-based column membrane, digested with 27 U DNaseI (Qiagen), washed, eluted in RNase-free water, and quantified with a DU 640 spectrophotometer (Beckman Instruments). RNAsin (Promega, Madison, Wis.) was added to each sample $(0.8 \text{ U}/\mu\text{l})$ prior to storage at -70°C .

Reverse transcriptase polymerase chain reaction

Reverse transcriptase (RT) reactions were performed with 1–3 μ g total RNA and 5 μ M random hexamer primer with Superscript II reverse transcriptase (Invitrogen, Carlsbad, Calif.) according to the manufacturer's instructions. To identify samples with contaminating genomic DNA, approximately 150 ng of each RNA sample was used directly as a template for PCR. All primer sets employed in RT-PCR reactions (see below) were used to identify amplification from genomic DNA. Samples showing bands larger than expected for cDNA amplification were assumed to contain genomic DNA and were discarded.

The cDNA generated from the RT reactions (1 μ l) was mixed with 0.5 U Taq DNA polymerase (Invitrogen), 3 mm of each deoxynucleotide, 1 μ M each primer, 1× company-supplied buffer (Invitrogen), and 1.5 mM MgCl₂ in a volume of 20 μ l. Primers included Sp5: 5for: 5' CCC TGG ACA GTA TGT GTT GCA TGG TAG, 5rev: 5' TAT CCC TGG TTG CAT CCT ATG AGC ACA); Sp5013: 5013for: 5' TCG ATG GGT GTT CCG AGT GGG TCT, 5013rev: 5' TCT ACA TCT AGC AAC TAG CAG GGT GCC; *Sp056*: 056for: GCA CAG CCA GCA ACC AGC ACT ACA AT, 056rev: ACG CCG ATG GGT TCT ACA GTG AAG GT; and SpL8: L8for: CAG CGT AAG GGA GCG GGA AGC GTC TT, L8rev: GTT TGC CGC AGA AGA TGA ACT GTC CCG TGT A. Reactions were heated to 95°C for 5 min; followed by 25 cycles of 94°C for 30 s, 52°C for 30 s, and 72°C for 10 min; followed by 72°C for 2 min and 4°C hold. The amplified fragments were electrophoresed on a 0.8% agarose/0.4% NuSieve gel (BioWhittaker Molecular Application, Rockland, Md.) containing 0.5 µg/ml ethidium bromide in TAE buffer (40 mм Tris base, 20 mм glacial acetic acid, 1 mM EDTA, pH 8.3). Gels were imaged with a DC120 digital camera and 1D digital software (Eastman Kodak, Rochester, N.Y.) followed by image optimization in Photoshop (Adobe Systems, Seattle, Wash.).

Cloning PCR fragments

Fragments amplified by PCR were cloned into the pCRII-TOPO vector according to the manufacturer's instructions (Invitrogen) and transformed into TOP10 bacteria (Invitrogen).

Arrayed cDNA libraries

An arrayed cDNA library was constructed in the pBK-CMV vector (Stratagene, La Jolla, Calif.) using activated coelomocytes from five sea urchins 24 h post-injection of 1 ml of heat-killed bacteria (for details of bacterial isolation and library construction, see Pancer et al. 1999; Cameron et al. 2000; Rast et al. 2000). A similar arrayed cDNA library was constructed using non-induced coelomocytes in the pSPORT vector (Life Technologies, Rockville, Md.) using coelomocytes from non-activated sea urchins (for details, see Smith et al. 1996; Al-Sharif et al. 1998; Cameron et al. 2000). Each library was arrayed into 240 plates of 384 wells each for a total of 92,160 clones per library (Al-Sharif et al. 1998; Cameron et al. 2000; Rast et al. 2000). The insert from every clone was amplified by PCR and spotted in duplicate onto five 22×22-cm Hybond-N⁺ (Amersham) filters for each library, which were used for screening.

cDNA library screens

Filters were prehybridized with 10 ml of hybridization solution [0.1% BSA (w/v), 1 mM ethylenediaminetetraacetic acid (EDTA), 0.25 M phosphate buffer (pH 7.4), 7% sodium dodecylsulfate (SDS) (w/v), and 50% formamide (v/v)] at 42°C for 2 h by rotation in a hybridization oven (Robbins Scientific, Sunnyvale, Calif.). The

hybridization solution was replaced, a riboprobe (see below) was added to the filters and then rotated overnight at 42°C. Filters were washed twice in 4× SSC (20× SSC is 0.3 M sodium citrate, 3 M NaCl, pH 7.0) with 1% SDS, twice in 2× SSC with 1% SDS, and twice in 1× SSC with 1% SDS at 65°C for 30 min. Each wet filter was sealed into a plastic bag and exposed to X-OMAT AR film (Eastman Kodak) without an intensifying screen. After exposure, filters were stripped by washing once in 0.4 M NaOH at 45°C for 3 min, twice in stripping buffer [0.1× SSC, 0.1% SDS, 0.2 M Tris (pH 7.5)] at 65°C for 30 min, and twice in stripping, filters were dried and stored at -20°C.

Riboprobe synthesis

Clones that served as templates for riboprobe synthesis (see Fig. 1) were linearized with either *Bam*H1 at the 5' end or *Xba*I at the 3' end (Promega), and 170 ng to 515 ng was labeled using 50 μ Ci to 70 μ Ci ³²P rUTP (ICN, Irvine, Calif.); 0.5 mM each of rATP, rCTP, and rGTP; 15 U of T7 RNA polymerase (Promega); 1x company-supplied transcription buffer (Promega) in a total volume of 20 μ l. Samples were incubated at 37° for 1 h, followed by the addition of 46 μ g yeast tRNA, 24 U of RNAsin, and digested with 1 U RQ1 DNAse (Promega) at 37°C for 15 min. Unincorporated nucleotides were removed by passing the probe through a G-50 fine Sephadex column (Amersham Pharmacia Biotech AB, Piscataway, N.J.) spun at 2,000 g for 1 min. Incorporation of ³²P was analyzed with an LS6500 liquid scintillation counter (Beckman Instruments).

Clone blots

Individual colonies were dispersed in sterile water (50 μ l) and 5 μ l was used as the template in PCR reactions using a 9600 thermal cycler (Perkin Elmer, Wellesley, Mass). Reactions included 0.5 U *Taq* DNA polymerase (Invitrogen), 0.25 mM each deoxynucleotide, 1 μ M T3 and T7 primers (Qiagen), 1× company-supplied buffer (Invitrogen), and 1.5 mM MgCl₂ in a volume of 20 μ l. Reactions were heated to 95°C for 5 min, followed by 25 cycles of 94°C for 30 s, 52°C for 30 s, and 72°C for 10 min, with a final step of 72°C for 10 min. Amplified products were electrophoresed on an agarose gel (as above), and images were captured using UV illumination with a digital camera (Eastman Kodak). Gels were double blotted by capillary action onto two Genescreen Plus membranes (NEN Life Science Products, Boston, Mass.) by standard procedure (Sambrook et al. 1989).

Non-radioactive DNA probes

Templates for Sp5 and Sp5013 were generated using clones that were amplified by PCR using T3 and T7 primers for clones matching to Sp5013 and T7 and Sp6 primers for clones matching to Sp5. Amplified products were purified using the Geneclean Turbo kit for PCR (Bio101, Carlsbad, Calif.) according to the manufacturer's instructions, and DNA fragments (100 ng) were labeled using alkaline phosphatase non-radioactive GeneImages Alkphos Direct kit (Amersham) according to the manufacturer's instructions. The blots were hybridized with the probe and washed according to the manufacturer's instructions.

Clone maps

Analysis by PCR

Plasmid insert sizes were analyzed from single colonies by PCR as described above, using a combination of a gene-specific primer with either T3 or T7 primers. A touch-down thermal cycling program was used because of the significant difference in annealing



Fig. 1a, b Maps and sequencing passes for Sp5013 and Sp5. **a** Sp5013. An overlapping series of 12 clones was analyzed by touchdown PCR reactions (see Materials and methods) using the 5013 for primer and the T7 primer in addition to restriction digests. Three of the six clones that were used to generate the sequence for Sp5013are shown. The sequencing passes are indicated as *arrows*. The number associated with each arrow corresponds with a primer listed in Tables 1 and 2. Clone 199N18 served as the template to generate the riboprobe for Northern-blot analysis (Fig. 9). **b** Sp5. An overlapping series of 39 Sp5 clones was analyzed using 5for (5' CCC

temperatures between the two primers (gene specific = $62-64^{\circ}$ C; T3 and T7 = 55° C). The PCR program (described above) with a total of 30 cycles was changed so that the initial annealing temperature of 62° C was decreased by 1°C each cycle for eight cycles to 55° C, where it was maintained for 22 cycles, followed by 72° C for 10 min and 4°C hold. Amplified products were separated on an agarose gel and imaged as above.

Restriction-enzyme digests

Digests were done with *Eco*R1, *Xho*I, and *Bam*HI, and fragments were separated on agarose gels.

Northern blots

Poly(A)⁺ RNA was isolated from 100 μ g total RNA from coelomocytes using the Oligotex mRNA mini kit (Qiagen) according to the manufacturer's instructions. Poly(A)⁺ RNA was electrophoresed through a 1% agarose gel containing 2.2 M formaldehyde in 1× MOPS buffer [10× MOPS is 20 mM 3-(*N*-morpholino) propanesulfonic acid, 5 mM NaOAc, 1 mM EDTA, pH 7] and blotted by capillary action onto Genescreen Plus (NEN Life Science Products) with 10× SSC. Filters were hybridized and washed as described above for library screens, and exposed to X-OMAT AR film (Eastman Kodak). Transcript sizes were estimated from RNA standards (Ambion) as well as coelomocyte rRNA. Filters were stripped using the protocol described above for library filters. TGG ACA GTA TGT GTT GCA TGG TAG) with T7 and reverse RceR1 (5' AAC ACA GCT TGG TTG CAC TCC TGT CC) with T3 by touch-down PCR reactions (see Materials and methods). Eleven of 17 clones that were used to generate 5.65 kb of sequence are shown. The length and overlap of the sequencing passes are indicated as *arrows* and the number associated with each arrow corresponds with primers listed in Tables 1 and 2. The area of the figure shown as a *tiled pattern* denotes the sequence provided by Z. Pancer. Clone 231K2 served as the template to generate the riboprobe for Northern-blot analysis (Fig. 9)

Sequencing

Plasmids were isolated using the Wizard Plus SV Miniprep kit (Promega) according to the manufacturer's instructions, quantified with the DU 640 spectrophotometer (Beckman Instruments), and sequenced with BigDye terminator cycle sequencing ready reaction kits (Applied Biosystems, Foster City, Calif.) using either T3, T7, or internal primers. Cycle sequencing reactions were done in a 9600 thermal cycler (Perkin Elmer) with 25 cycles of 96°C for 30 s, 50°C for 30 s, and 60°C for 4 min. Unincorporated nucleotides were removed by passing the samples through a G-50 fine Sephadex (Pharmacia) spin column and dried in a centrivap concentrator (Labconco, Kansas City, Mo.). Dried samples were either dissolved in 15 μ l template suppression buffer and loaded onto an ABI prism 310 Capillary Sequencer or 3 μ l sequencing loading buffer [5:1 ratio of deionized formamide to 25 mM EDTA (pH 8.0), plus bromophenol blue] and loaded onto a 377 Automated Sequencer (Applied Biosystems).

Results and discussion

Sp5 and Sp5013 clones

Two clones, *Sp5* and *Sp5013*, were identified during a library screen using a probe designed to find sequences specific for scavenger receptors with Cys-rich (SRCR) domains (Pancer et al. 1999; Pancer 2000). However, the initial analysis of these two clones only revealed regions that encoded SCRs, and therefore they were kindly provided to us by Z. Pancer. Further analysis was pursued

Table 1 Sequencing primers for *Sp5013* listed in Fig. 1a. All annealing temperatures were between and 55.3°C and 57.2°C

Primer number in Fig. 1a	Primer	Sequence
3 2 4 5 6 7	199N18T3F 199N18T3F3 199N18T3F4 199N18T7R2 199N18T7R3 199N18T7R T3	5' TTGCTAGATGTAGAGACC 5' GGGAGGATACTAGATGG 5' GCAATGGGTCGTCATCA 5' TACACAATTTTGCCCACC 5' TCAGGAATTAAAGCCTCC 5' TTCCTTCAATCCCCACC 5' ATTAACCCTCACTAAAGGGA
8	T7	5' TAATACGACTCACTATAGGG

because many complement proteins have SCRs, and the sequences obtained did not match to *Sp152*, which encodes SpBf with five SCRs (Smith et al. 1998).

Messages and deduced proteins

Two arrayed cDNA libraries were screened using riboprobes made from clones generated by RT-PCR with primers specific for Sp5 and Sp5013. Sixty Sp5 clones and 30 Sp5013 clones were characterized by mixed-primer PCR analysis and restriction digests to identify overlapping regions and to optimize sequencing strategy (Fig. 1). Both standard and internal primers were used to sequence 17 Sp5 clones and 6 Sp5013 clones (Tables 1, 2). Some of the sequence near the 3' end of Sp5 was provided by Z. Pancer. The sequenced region of the Sp5 transcript was 5.96 kb and was composed of 5'UTR [36 nucleotides (nt)], 3'UTR (432 nt), and 5,472 nt of open reading frame (ORF) (Fig. 2). The sequenced region of the Sp5013 transcript was significantly shorter, 2.751 kb, and composed of 3'UTR (42 nt), and 5'UTR (924 nt), with an ORF of 1,785 nt (Fig. 3). Although the 5'UTRs for both of these cDNAs were quite short, the entire 5'UTR for Sp5013 may have been obtained because six clones had identical sequences at the 5' end. However, it is possible that significant secondary structure in this region of the message may have blocked complete reverse transcriptase activity.

The deduced amino acid sequence encoded by *Sp5* and called "SpCRL" consisted of 1,829 amino acids (Fig. 2).

The deduced amino acid sequence encoded by Sp5013 and called "SpCRS" consisted of 595 amino acids (Fig. 3). The absence of a Kozak sequence surrounding the start codon for Sp5 and the presence of a very short 5'UTR made a reliable identification of the correct start codon difficult. However, the choice of the probable start codon was based on (1) the presence of one in-frame stop codon 21 nt upstream in the 5'UTR in addition to a second stop that was out-of-frame and (2) an ORF following the ATG that began with a leader region (underlined, Fig. 2) as predicted by the pSignal program (Nielson et al. 1997). The cleavage site for the removal of the leader was predicted to occur after Ser27, which was directly followed by Cys30 of the first SCR. Sp5013 had an imperfect Kozak sequence [GCC (A/G)CC ATG G] (Kozak 1987) (Fig. 3, underlined), two stop codons in the 5'UTR, and a hydrophobic leader putatively cleaved after Ser27, which was followed by Cys30 of the first SCR (Fig. 3). Neither sequence revealed the presence of a transmembrane region (TMPredict database, Hofmann and Stoffel 1993). Both messages had numerous stop codons in all reading frames in the 3' UTR and both appeared to have polyadenylated stretches at the 3' end of the sequence (however, see below). Sp5 did not have a conserved polyadenylation signal sequence, but four were identified in the 3'UTR of Sp5013. Two of these were located just 5' of the $poly(A)^+$ tail (Fig. 3, underlined). One AU-rich element (ATTTA, Asson-Batres et al. 1994) was identified in the 3'UTR of Sp5013 (Fig. 3, underlined).

SpCRS and SpCRL domain structure

Analysis of the deduced amino acid sequences of SpCRL and SpCRS indicated that both were mosaic proteins with a variety of domains, many of which were typically found in complement regulatory proteins and some complement components. Overall, both proteins contained SCRs (18 in SpCRL and four in SpCRS), a FIMAC domain, two S/T/ P-rich regions, and numerous N-linked and O-linked glycosylation sites (Figs. 2, 3, 4). In addition, SpCRL contained a fucolectin domain and a Cys-rich region. The S/ T/P-rich regions were present in similar locations of both SpCRL and SpCRS (Fig. 4). The composition of Ser and Thr plus Pro in the N-terminal region was 17% in SpCRL

s 1 e	Primer number in Fig. 1b	Primer	Sequence
2°C	2	215A3w42C5R-2	5' CCATAGTCGTCGCC
	3	42C5RT3	5' ACGGCTATCCACATTCC
	4	42C5T3F	5' AGGAGAGTGGCGATAC
	5	33C8RT3	5' ACAGTCATAATGTACTCTGG
	6	211H15RT3	5' GATTGTGTGCACGTAGG
	7	229H19T3F	5' CACCCGACGCCAAAT
	8	210I5R1	5' CCTGCTGTGTGGCCAA
	9	36C1RT3	5' GGAGTGAGCAACGAGAA
	10	CR.R1	5' TCTACGTGCAGTTTGCTGAG
	1	T3	5' ATTAACCCTCACTAAAGGGA
	11	T7	5' TAATACGACTCACTATAGGG

Table 2 Sequencing primers for *Sp5* listed in Fig. 1b. All annealing temperatures were between and 55.3°C and 57.2°C

Sp5 and SpCRL

GTTTGGTCGAGAGTC TAGCTAG TATATCGA <u>CAAGCAATG</u> GAGGGAAAGAGTTATCATTATTTAGTCCTGTTCACAGGACTGTTTCTTTGCCTGTTACACAGTGGCGGG	108
MEGKSIHILVLFTGLFLCLLHSGG leader	∠4
GCATCATCGAGCTGCGATAACATCAATCCGATTGATTTTGCGGGAAATCACGTATGATGAACCGGGAGGGTTAGCATCCCACCCTGACGGGACGGTTGCGAATATTGGG	216
ASSCDNINPIDFAEITYDEPEGLASHPDGTVANIG	60
TGTGGTGTCTTTTATCAGATCGTGCCCACGGATTTTGAGAGAAACAACATGCACCAATGGATCGTGGACAGAACCGTTGCCCCGATGCAGACATATTCCGAGATACTGT	324
C G V F Y Q I V P T D F E R T T C T N G S W T E P L P R C R H I P R Y C CHO(0)^ CHO(N) SCR 1	96
CAACGTCATGTAGCCCTGGTCGAGAGGCTACGGACGGGATGTCGAGGCCCCATGCAGGCAG	432
0 R H V A L V E S Y G R D C 0 T P C R 0 R P N G G C P P N K R C H C D G	132
Cysteine rich region	
${\tt GTTTGTGGATGGAGTTGTATTTCAATTTACGAGGACAATTTCTGCCCAGAAGTGGCTCACCAAGCAGGTCTTTTGGTGACCTACGACACCCCTGAACGGCGATTCAATTTCTGCTGACGACGACGACGACGACGACGACGACGACGACGACGACG$	340
$\frac{V C G W S C I S I Y E D N F C P E V A H Q A G L L V T Y D T P E R F N}{SCR 2}$	168
GGCTACGCCAATTTCTCTTGTGGTGAGGGTTACATTTATCATTCTGGGAGCAGCAGGCTACGGTGTATGAGCAACAGACAG	648
G Y A N F S C G E G Y I Y H S G S S R L R C M S N R Q W G G G N D F V C ^CHO(N) SCR 2 ◀━━━	204
ATCCCAAACGTCGTCTGCAGAGATCCACCCAGCGCAGAACACGCCTCGCTGCAGCCAAGCGGGAAGCAATATTTCCTACCCGGTGACACCCGCTCATATACATGCAAT	756
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	240
${\tt CTCGGCTACATAATCCGTGGAAGTCGGGACATTACTTGTGGCGACGACGACTATGGGTGGTCGGAACCAGACTTCACCTGCAGGCCTCGCCCCTGCACCTGGTAGACCAGACTTCACCTGCAGGCCTCGCCCCTGCACCTGCACCTGGTAGACCAGACTTCACCTGCAGGCCTCGCCCCTGCACCTGCCACCTGCACCCTGCACCTGCACCTGCACCTGCACCTGCACCTGCACCTGCACCTGCACC$	864
L G Y I I R G S R D I T C G D D Y G W S E P D F T C R P R P C T Y P G R	276
SCR 3 \leftarrow SCR 4	
ATATCTAACGCTGATCTCACGACTCAACTATTCCAATTTAAAGAACGGGCAATATACGTTTGTAGGGAGGG	972
I S N A D L T T Q <i>L F</i> Q F K E R A I Y V C R E G Y E N P P Y T L P Y R T	312
TGTCAAGGGAATGGCCAATGGACTCAAATCTTACCAATTTGTGACGCTATTCGTGTCCAGCAATCCTTGACATTACTAATGGGATGGCGATGGCGGTGGCGATGGT	1080
C Q A N G Q W T Q I L P I C E P I Q C P A I L D I T N G N V D S R G N D	348
	1 1 1 0 0
	1 204 1 204
FDSQIIFICNDGIRDDGIARRVCQGDRIWSGQBAVC	304
	. 1296
TEILICEDPGVPVNGYMENEKOVYHIDDVVYHCNRC	420
SCR 6	
${\tt AAGACAATAGATGGAAGTATACTCAACTCTTGCACAGAGTCAGGAGAGTGGCGATACCCGGTGCCCGTATGTGGTGGACCTTGCATTGTTCCTCCTTATCCAAGAGATGGAAGTAGACTGTGGTGGACCTTGCATTGTTCCTCCTTATCCAAGAGATGGCGATACCCGGTGCCCGTATGTGGTGGACCTTGCATTGTTCCTCCTCTTATCCAAGAGATGGCGATACCCGGTGCCCGTATGTGGTGGACCTTGCATTGTTCCTCCTCCTTATCCAAGAGATGGCGATACCCGGTGCCGTATGTGGTGGACCTTGCATTGTTCCTCCTCCTTATCCAAGAGATGGCGATACCCGGTGCCGTATGTGGTGGACCTTGCATTGTTCCTCCTCCTTATCCAAGAGATGGCGATACCCGGTGCCGTATGTGGTGGACCTTGCATTGTTCCTCCTCCTTATCCAAGAGATGGCGATGGCGATACCCGGTGCCGTATGTGGTGGACCTTGCATTGTTCCTCCTCCTTATCCAAGAGATGGCGATACCCGGTGCCGTATGTGGTGGACCTTGCATTGTTCCTCCTCTTATCCAAGAGATGGCGATACCCGGTGCCGTATGTGGTGGACCTTGCATTGTTCCTCCTCTTATCCAAGAGATGTGGCGATACCCGGTGCCGTATGTGGTGGACCTTGCATTGTTCCTCCTCTTATCCAAGAGATGTGGTGGCGATACCCGGTGCCGTGGCGATGCGATGTGGTGGACCTTGCATTGTTGTCCTCCTTATCCAAGAGATGTGGTGGCGATGCCGTGCCGTGGCGATGTGGGGACGTGGCGATGCGGTGGCGATGTGGGGGGGG$	1404
K T I D G S I L N S C T E S G E W R Y P V P V C G G P C I V P P Y P R D	456
SCR 6 SCR 7	
GGTTGGTGGCAAAATGGTAATGAATATCCACCAGAAACAAGTGTACCCCCACAATACCAGACTGCAGTTGACATGTAGAAGCTGGCGATTCAATAAGAGACGAAGCAG	1512
G W W Q N G N E Y P P E T S V P H N T R L Q L T C R S W R F N K R R S S ^CHO(O)	492
GTGAAATGTAATGGAGGGGGGGGGCCGATGGTGGTGATGGCGTAGACGTCGCCTTTGTCGAGGAACTCCATGCGGGGGGGG	. 1620
V K C N D G V W S D S D D V H R L C R G T P C G V R W S I D V A L N I T	528
	1 1 7 0 0
	; 1728 564
	504
TCAAGATGCGAGGAAGGAAGATATAACAATAATATCCCCAGATGCGAGCTAGTCCCATGCTCAGCGCCGGAAGATGTATCCCATGGGAGATTAACTTACACAAACCCT	1836
S R C E Q G R Y N N N I P R C E L V P C S A P E D V S H G R L T Y T N P	600
SCR 8 SCR 9	
${\tt GATGGAGTTCCTCATGAAAAACCCACTTCATGGAGACACCCGGCTCCTTCAATGCGGTTTTGGTTACAGATCAAGAAGCTTCAATAGCAGTCGATGTATAATGGAGTGAGTGAGTGAGTGAGTGAGTGAGTGAGTGAGTGAGTGAGTGAGTGAGTGAGTGAGTGAGTGAGTGAGGAG$	1944
DGVPHENPLHGDTRLLQCGFGYRSRSFNSSRCDNGV ^CHO(N)	636
${\tt TGGGTCGAAGGCAGTCATGATATCCGCTGTTATCCAAAGCCGTGCGACCCTATCCCAGACACAATGGGTCATGTTAACTACCACCAGAACAGCTAAAGCAAATGGGAAATGGGAAAGCAATGGGAAATGGGAAAGCAAATGGGAAAGCAAATGGGAAAGCAAATGGGAAAGCAAATGGGAAAGCAAATGGGAAAGCAAATGGGAAAGCAAATGGGAACAGCAAATGGGAACAGCAAATGGGAACAGCAAATGGGAACAGCAAATGGGAAAGCAAATGGGAAAGCAAATGGGAACAGCAAATGGGAACAGCAAATGGGAACAGCAAATGGGAACAGCAAATGGGAACAGCAAATGGGAAAGCAAATGGGAAAGCAAATGGGAAAGCAAATGGGAAAGCAAATGGGAAAGCAAATGGGAAAGCAAATGGGAACAGCAAATGGGAAAGCAAATGGGAAAGCAAATGGAAAAGCAAATGGAAAAGCAAATGGAAAAGCAAATGGAAAATGGAAAATGGAAAATGGAAAATGGAAAATGGAAAATGGAAAATGGAAAATGGAAAATGGAAAATGAAATGAAATGGAAAATGAAATGGAAAATGAAATGGAAATGAAATGAAATGAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAATGAAATGAAAATGAAAATGAAAATGAAATGAAATGAAATGAAATGAAATGAAATGAAATGAAATGAAATGAAATGAAATGAAATGAAATGAAATGAAATGAAAATGAAAATGAAATGAAATGAAATGAAATGAAAATGAAATGAAATGAAATGAAAATGAAAATGAAATGAAATGAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAAATGAAATAAT$	2052
W V E G S H D I R C Y P K P C D P I P D T M G H V N Y T R T A K A N G K	672
SCR 9 \leftarrow SCR 10 $^{CHO}(N)$	
${\tt TACATCCATGGAACAGAGGTAACTGTCAACTGTAATTCTGGATACCTTCCTGCATATGGTAATGGAACTGCAGTCTGCAATGCGTCACAGTGGTTGACAGTTATTCCTTCC$	2160
Y I H G T E V T V N C N S G Y L P A Y G N G T A V C N A S Q W L T V I P	708
^ CHO (N) ^ CHO (N) SCR 10	1
ACGTGCACCACTCACGTAACATAATCCTAGACTCACATACCACACTCTCTCACAACACCACCACCTAAGAGAAATCCAAGCTTTGATAGGCAATGCATGGTTGGCCAGGGAT	2268
$\begin{array}{c} T C T Q S R N T T L D S H T N V S Q I P P K R N Q A L I G N A W L A R D \\ \hline \\ $	/44
	· 2276
G N T D S A P O Y C S R T K V T N P W W K A V M K D T Y N F T D V K T	780
$\frac{1}{(CHO(0)^{\circ} fucolectin domain} \qquad (CHO(N))$,00
${\tt TACAACCGTCTTGACAGGGAAGAACGCAGATATGATCTGGAGGGTGCAGAAATAAGAGTGGGATTGAATGACAATTACACCACAAACCTGTTGTGCGGAGAACCAGTGACAATTACACACAC$	2484
<u>Y N R L D R E E R R Y D L E G A E I R V G L N D N Y T T N L L C G E P V</u>	816
fucolectin domain ^CHO(N)	
ACCCGACGCCAAATCGAGAGTAGTACCAGGGCAAACCATGGGATCCCCATAAGATGTGTGGGAGAAGGTGTCACCGGTATCAGGGGTAATGTCATCAGTGTGCACATT	2592
<u>T R R Q I E S S T R A N H G I P I R C V G E G V T G I R G N V I S V H I</u>	852
rucolectin domain	

Fig. 2 cDNA sequence of Sp5 and the deduced protein sequence of SpCRL (*Strongylocentrotus purpuratus* complement related protein, long form). Overlapping sequencing passes were used to generate the cDNA sequence and the deduced amino acid sequence was translated using DNASIS (Hitachi, Cascade, Colo.). The signal sequence was predicted using the Expasy SignalP V1.1 program. The start codon and stop codons, including those in the 5'UTR and

3'UTR, are indicated in *boldface*. The Cys in the Cys-rich region and the Ser, Thr, Pro in the Ser/Thr/Pro (S/T/P)-rich regions are also indicated in *boldface*. All domains are labeled and/or numbered, and the limits of each SCR is indicated with an *arrow*. The N- and O-linked glycosylation sites are noted with *CHO(N)* and *CHO(O)*, respectively

CCAACAA	ATAP	CTA	ACAA	GAAG	CGA	GAAC	TCA	AGT	CTGT	GCG	BAAG	TTGi	AGG	TCTA	TCAP	ACAA	GGA	CTT	TCTA	TAT	CAGI	GAAC	GGT	TCC	GAG	AAT	GAA	CAC	GGGI	'GGGA	2700
<u>P T</u>	Ι	1 T	<u> </u>	K	R	<u>E</u>	L	S	L	С	Ε	V I	E	<u>V</u> Y	Q	Q	G	L	S	IS	s v	/ N	G	S	Ε	Ν	G	T	R V	G G	888
mamama		tuc		tin		nain	1700		DOON.	003		2000	TTTC	0010	7070	1070	amm	CITE A I	mmaa	770	3003			N)	2000	D (1)	CH		() 1999 1997		2000
	V	P F	S A	DADE E	MACC	A	D	V	S	GCA S	T T	AIG. NV	V	ΑΤ	ACAG	DAJE O	V	LIA L	TIGG. Ti	AAG(E (GIGA G F	IGGAC	V	RGA.	T	S S	C C	AIG H	AR	H H	2808 924
<u> </u>	→		SCR 1	11			2	•	0	-	-		•		×	×	•			_ `	0 1		•	10	-	0	C				201
GTCCTC	AGAG	GCT	CGGA	CACT	AAC	CATA	TTG	JAT	CTTA	CCI	GTC	TCG	GTA	ATTC	ATCO	CTGG	GAC	CAA	GATA	AAC	CCGI	GTGI	GAG	CCA	GAA	ACG	GCA	CTG	TTGA	TCGC	2916
V L	R	G S	5 D	Т	Ν	Η	I	D	L	Т	С	L (G	N S	S	W	D	Q	D	K I	ΡV	C C	Е	Ρ	Е	Т	С	Т	VΙ	R	960
														^CHC	(N)		SCI	R 1	1	←			1			_ L			≻	SCR	12
CTATTTA	ATG	GCG	TTTT:	ICCA	AAT(JGTA	'AAA'	ATC:	FTAT.	ACA	ACC	ACG	GTG	AAAA	CATC	CACT	TTC	ACT'	TGTA	ATC	CTGG	CTAJ	GAG		GAA	CAT	ATA	GGT	ACTO	GTGT	3024
LF	IN	G	÷ F.	Р	IN	G	ĸ	Т	Ь	Y	IN	н	Ċ	E N	⊥ С₩О /	(NT)	F.	.T.	C	N	ΡG	τt	E	ĸ	E	н	D	R	X N	I C	996
GATAGG	AGAZ	TTG	TAGTO	CCA	CGCZ	AGTC	CAP	AGG	ГGTA	TTC	TAAG	CAT	стт	GCAA	TGCT		GAC	CTT	CCAG	ATC	атат	GGTZ	ACT	GCA	CAA	TCA	TAAC	CTG	ACTT	TCCT	3132
DR	R	I Z	A V	P	R	S	P	R	сI	I	Q.	A S	s	C N	A	P	D	L	P	DI	ΗМ	1 V	T	A	Q	S	Q	P	DF	' P	1032
			SCR	12		←							L		\rightarrow		SCI	R 1	3		c	но (с) ^								
CATGGT	ACGI	TGCI	CGAI	FGTC	TCA	IGTG	JAAG	BAC	GGGT	TTG	BAGC	TTA	GTA	CGAA	.CCAF	AGAA	CAG	CTA	AGAT	GCTI	ATAG	SAGGI	GGC	TGG	AAC	ACN	CCT	TAA	.CGGC	AACA	3240
H G	Т	ΓI	D	V	S	С	Е	D	G	F	Ε	L S	S	T N	Q	Ε	Q	L	R	C	YR	e G	G	W	Ν	Т	Ρ	L	ΤA	T	1068
mamaaaa	77.07	amar			2020	maar		17.00		amo	1001	m 7 <i>a</i>	a 1 a		~~~~			a	aaam	a 1 a								man	0000	плаа	2240
C LO	AGF	GIGI GIGI	ICATC	alGG W	ACA.	a l	.GIG	BAG(JUDB	D	DCCA	TAG	CAC 7	CATI	GAAF	ATAC V	V	CAG	GGGT	CAC	AAGG	ATAI V V	.GIG W	CAT U	GGT	ACC2		TCA V	GGTA D V	TAGC	3348
	Q	5	/ 11	VY	Т	5	<u> </u>	Б	-G ->	Г	SCR	14	n.	г <u>г</u>	I. I.	т	T	Q	G	U (2 6	т т	v	11	G	T	ĸ	v	K 1	0	1104
TGCCGT	JAAG	GCTO	GGA	ATC	GAA	GGGA	4TCG	GCAG	GAGA	GAG	AGT	GCA.	ГСА	ATAG	ACAA	ATGG.	ACA	GGT'	TCTA	CGC	CAGC	ATGO	CAGA	GTA	GCA	GCC	CTC	CTC	GATO	CATG	3456
C R	Е	G V	ΙE	I	Е	G	I	А	Е	R	Е	C I	Ι	N R	Q	W	Т	G	S	ΤI	P A	С	R	V	А	А	Ρ	Ρ	R C	M	1140
																S	CR :	14		←			1								•
CTCCCA	JATA	GAT	CAG	CGGA	CAT(CGTT	'ACT	CA?	ATAA	GAG -	GTT	TTA	CAG	GGAA	TAT	CTTT	GAT(GAG	GGTT	TGC	CCAI	TGGI	'GAG	CGT	GTA	rcc <i>i</i>	ATTT	CTG	TTAG	TTGT	3564
LP	D	RS	S S	G	Н	R	Y	S	I	R	G	F' '.	Г	G N	I	F,	D	Е	G	LI	ΡI	. G	E	R	V	S	I	S	VS	С	1176
AATCAAC	1 5 10.07	מידאי	1200	rcad	ccar	TCAC	TCC	יאמי	ACAC	דממ	CTT	таса	nan	cccc	ACTZ	TCC	TCA	2770	acma	TAC	аалт	יאיימי		ACC	ATC	2777	ACCT	TCT	CTAC	CAAC	3672
N O	G	Y 1	' A	0	P	S	V	0	T T	E	C	L	ron E	R G	V	W	S	V	A	V I	P T	C C	Ιv	R	M	E	R	T. I	CI	' K	1212
2	•			~		-		~			-					s	CR :	15		<u> </u>			1.					- L			
CCTGGA	FACF	TTTC	CTCA	CGTT	GTC	CAGT	ATG	GTGI	AACG	GAC	TAG	AGA	CGA	ACAG	ACTI	'GAT	TCT	TAC	CCTC	ATG	ATGA	GCTI	CCI	'GAA	GGG	ACG	TTC	TAG	TGTC	TCGT	3780
ΡG	Y	ΙS	3 H	V	V	Q	Y	V	Ν	G	L	E :	Г	N R	L	D	S	Y	Ρ	ΗI	DE	L	Ρ	Е	G	Т	F	L	V S	R	1248
SCR 1	L6																						_								
TGCTCA	TCC	CTGC	JACA	JTAT	GTG.	TTGC	'ATG	GTI	AGTG		ATA	GAA	CCT	GTTC	TGAG	BAGC.	AGC'.	rgg.	ACAG	GAG.	TGCA	ACCA	AGC	TGT	GTT(GAA(CAG	ACA		GATC	3888
C D	ш	P	, Q	T	v	Ш	п	G	э.	A	^ Сн	Γ. Γ. (N	1	C B	E	5	G	vv		CR.	16.	į P	5	C	ľ	Б	А	D	_ F		1204
AGCTTT	CAAG	ACA	ACCAC	acco	CTT	GACG	TCA	AGGI	AGTG.	ATG	GAA	CAA	, TCG	TCAT	TCAT	CCA	CGC	AGT	cggc	TCC	TCAT	TCTI	TGT	'CAC	GTG	CCT	CCT	ATA	GCGI	'AGCA	3996
S F	Q	Dľ	1 Q	P	L	D	V	R	S	D	G	T :	I	VI	Н	P	R	s	R	LI	L I	. L	С	Н	V	P	s	Y	S V	7 A	1320
															S/1	C/P	ric	h r	egio	n											
AGATTT	FAA	CAG	AAA	rgga	.CCG(GATG	CTG	HTT:	FATT	GGG	GCC	TAT	CAA	.CAAT	GGTI	ATG.	AGT	CTA.	AACC	CAC	CACA	TACO	TCA	CAA	TCT	GGG	CATT	TCA	CTTG	CCGT	4104
<u>R_F</u>	.E	SI	<u> N</u>	G	P		A	_V	Y	W	<u>G</u>	L	S	T №	V	M	S	_L	N	P 1	P H		S	Q	S O	G	H	F	T (<u>R</u>	1356
TOTATO	2000	ימידיר	ייייריייי	r T T T		TOTO	1 TOT	יארי	атас	ъст	יידירי	CTC		ጥልጥጥ	5/1 דירויידי		CGTO	n ro CCT	agio acca	CAC			יםמיד		U) TTC	~~ ~ ~ ~		מיתמי	ACTA	CTAC	1212
S N	D	R	5 L	S	H	S	V	Y	V V	0	F	A 1	E	I F	ΪC	D	R	P	T	T I	P I	N	G	A	F	0	D	H	DY	Y Y	1392
																	→		SCR	17	^	CHO	(0)								
GGTTGGA	AAA	ATG	GACA	\TAT	TAC	ATGG	GCA	AA	JTGA	TCA	CAT	TTG	CTT	GTAA	CGAI	'GGC'	TAC	ATT	CTTG	ATG	GGGA	AAGO	AGA	ATT	ACC'	TGT(FTGC	TTG	GGAA	GTGG	4320
G W	K	N (3 Q	Y	Y	М	G	К	V	Ι	Т	FΪ	A	C N	D	G	Y	Ι	L	D	G E	R	R	Ι	Т	С	V	L	G K	W	1428
	1000					1000	17.01		2011		ama			ma (13)	maaa	1001	1 0 0	a				11 001							aman	aama	4400
TCACATO	D	N I	CAAGA	4TGC		AGGC	ACA.	IGA(JCAA 7	CA1 T	GTG.	AAGA T 1	aac F	TACA	TCCC	DCCA.	ACCO	CAT(GGAA	CCAA T 1	AGA'I K T	AGGA	1666 C	iaac N	AGA. D	ATAC	GTG	ACT	CTG1	GCTC 7 T.	4428
SCR 1	,		- K	C]×	IX.	11	CHO	ົດໂດ	÷	C			S	CR 1	18	т	11	G	т 1	IC I	. G	G	IN	IC.	Т	G	D	5 1	Ц	1404
ATAGGA	rgca	ACCA	AGGGZ	\TAT	CAG.	TTAC	CAAG	GAG	GAGC	CTT	TCT	TGG	ATT	GTCA	GGAA	AGT	GGA	AAC'	TGGT	CCC	ATCC	TCT	ACCI	GCT	TGC	ATTO	SAAA	TCA	TTGA	GCCG	4536
I G	С	Nζ) G	Y	Q	L	Q	G	Е	Ρ	F	LI	D	СÇ	Е	S	G	Ν	W	S I	H F	ь Г	P	А	C	I	Е	Ι	I E	P	1500
																		^C	HO (N)	SCF	18	•								
GAGAGG	CCCI	GTT	ATTCO	GTC	AGT.	IGCG	;GGG	GTG:	FGGC	AGA	AAT	GCGi	AGA	CCGA	.CTCF	ATCT	GGG	GTT	GGTG	TCT	GTAG	GTGI	TATC	AGT	CCC	AAT:	CCT	GCC	CTGI	AACG	4644
ER	Р	C	2 S	<u>V</u>	<u></u>	<u>C</u>	. <u>G</u>	V	_W	Q	<u>K</u>	<u>C</u>	Ľ	TMAC	S don	S	<u>G</u>	V		V	<u>C R</u>	2C.	<u>+</u>	<u>S</u>	<u>P</u>	- <u>N</u>	<u>S</u>	<u>C</u>	P/	<u>T</u>	1536
AATCAC		12001	יידירי	raaa		anca		וממי				۵CTT	r T	GCCC			CCTO	CTC	CONT	GTA	TTCA	CAR	יאמי	aac	GTA	220		CTT		CACT	4752
N E	T	E V	7 C	G	T	D	G	R	N	Y	Т	N I	F	C R	L	K	A	L	A	C :	I C) N	T	G	V	E	V	A	S F	T	1572
^CHO (1	1)		~~~~~	~~~~~		~~~~~			~~~~~	^CH	IO (N)	~~~~~	~~~~~	FIN	IAC ·	dom	ain	~~~~~	~~~~~			~~~~~	~~~~~~	~~~~~		~~~~~	~~~~~		~~~~~	
TGGTTC	rgca	TTA	ATGGI	ſGTG	CCT	GAGC	GTG	GCG.	FTGC	TAA	ACAC	CAA	CGC	CATT	CATI	[GAA	ATC	GAA	GAGC	CAA	GCCA	GTT	ATCC	TCC	TTT	GAG	AGCA	CTT	ATGA	GGAA	4860
<u>WF</u>	. <u>C</u>	1I	IG.	V	<u>P</u>	_ <u>E</u>	. <u>R</u>	<u>A</u>	_L	L	.T	P	ŗ	<u>P</u> F	I.	<u>E</u>	<u>I</u>	E	E	P	SC)L	<u>S</u>	<u>S</u>	F	E	. <u>S</u>	T	<u>Y E</u>	<u>E</u>	1608
							-			a	CHO	(0)	^ ~ ~ ~		FIMA	AC d	oma	in		a						- am					
TCACCTO	D	CTA:	. ССС. Г. – П.	L'GAA E	CCAP	ACTC	TCC	D	JCAA	CTG T	JAAG T	GAC	CTG D	TACI V T	AGAG	CCA	GACI	ACA	D	CATO		ATCI	. I'A'I V	TCA	TTC:	rGro	LATC	DCCG	ATGA	ATGC	4968
SE	-E	^CH() (O)	G	Ē		10 (C	- <u>-</u>	- 13	⁺ CH	10 (O))	F	-×∺ FI	MAC	dom	ain	~^C	HO (O	2		0)			£			<u>F</u>	ĽĘ	hunde	T044
CCTAGT	AGCI	CAG	GGCI	AGAA	TTG	AGCC	LAGC	GCO	GCAA	GAG	CAA	TCG:	ГCА	TTGC	TTAC	GCC	GAC	AAT'	TGGG	ATG(CAGA	ACCTI	GGC	ATC	TTA	GGGI	ATTC	AAA	TCTA	TGAA	5076
P. S	s	S A	A A	Е	L	s	Q	G	А	R	А	I	V	IA	Y	А	D	Ν	W	D 2	A E	L	G	I	L	G	I	Q	I Y	E	1680
СНО (О)	~	^CHO	0)									_ /	S	/т/р	ric	ch r	egi	on													
GTGATG	CGG	GAG	GCTO	GCCA	TCG.	IGGA	'CAG	SAA(GCT	CTA	ACC	AAG	ACG	TTCT	AATI	GAT	GTG	GGT	GCTG	CCT	CAGA		GGA	TGT	CTT	IGC	CAT	CCT	TTCA	AGCA	5184
<u>v_M</u>	. <u>A</u>	G (±	P	ន	W	T	E	G	5	N	<u>v</u> I	U	V I /m/5	I		V	<u> </u>	<u>A</u>	A .	S E	R	G	<u>C</u>	Ц	C	Р	S	н (<u>A</u>	1716
ACAGACO	ירייר	TCAT	ירארי	100	ירידי	AGTC	מממי	AC	TTG A	GGC	יאממ	ATCO	а ар	A A G G		CTC	CTA(CAC	ACCG	ACA	TTGT	יארידר	מרחי	ጥልጥ	TCT	40'D/	מידבן	ТАТ		AGAA	5292
T D	P	V	I I	R	F	S	Q	N	L	R	Q	D (з. "А	K G	Q	L	L	H	T	D :	I V	7 Ц	P	Y	S	T	D	I	Y C) E	1752
													S	/T/P	ric	ch r	egi	on													

Fig. 2 (continued)

ATA	AGA	CTG	CTA	TTA	GAA	GAG	CCT	GTT	CGT	GTA	CAG	ACC	CAA	GCA	CCC	CAG	GCA	ССТ	CAC	CCA	CAA	GAT	GAA	rcg.	AGC	GAC	'AGC'	TTT	AGC	AGC	CTT	AAC.	AGC	CGT	AAT	5400
Ι	R	L	L	L	Ε	Ε	Ρ	V	R	V	Q	т	Q	А	Ρ	Q	Α	Ρ	Η	P	Q	D	Е	S	s	D	s	F	s	s	L	Ν	S	R	Ν	1787
										CH	0(0) ^				S/	T/P	ri	ch	reg	ion	CH	0 (0)) ^	^C	HO (0)									
GGT	CCA	GAG	IGGA	CAA	'TAA	TCA	TGG	TTT	CAA	ICT	GAT'	TCA	GAC.	ATC.	ATT.	ACT.	ACC.	ACG	GCC	CGC	ACC	AAC	ACA	GAG	ATC	GTT	'GAT'	TCG	CCG	CGT	TCA	TCA	GAA	rcc:	ГСТ	5508
G	Р	Е	G	Q	Ν	S	W	F	Q	S	D	S	D	I	I	т	т	т	Α	R	т	Ν	т	Е	I	V	D	S	Р	R	S	S	Е	S	S	1824
															CH	0(0) ^	^C	HO (0)	S/S	T/P	ri	ch :	reg	ion			CH	0) 01)^	CH	0(0))^		
TTT	GAC	ACA	TAA	TCT	TGT	CTA	A GA	.GTG	GAT	ΓGT.	ACA	GGT	GGC'	TCA	TAC.	AGA	CTT	TTT	TTT	GTT	GTT	ATT'	TTA	CAA	TGC	CAT	CAA	ATT	GCT	TTC	TCT	TCT	TCC	ACT	ГСТ	5616
F	D	т	STO	Ρ																																1829
TTT	TGT	CTT	CCT	CTT	CTT	CTT	CCT	TCC	CTT	CTT	CAC	CTT'	TCT	CCT.	ACT'	TAT	CAT	TAT	TAT	ATT.	ATT	TTC	TCT	ACT	CTT	TAT	TT T .	AAT	TTT	TCT	TTC	CTT	CTC	AAA	ГТС	5724
ACT	CTC	CTT	GTA	CGT	CTC'	TAG	AAT	CTC	ACA	TTA	TAC'	TAC	TCC.	TTA	CTC	GCC'	TCC	TAT	'AAT	ACC	TTT	TAA	ATTO	CAG	ACC	TAC	TAT.	ATT	GTG	TTA	ATC	AGT	ACG	GGA	GCC	5832
ТАТ	TCA	CAT	ACA	AGC	ТАТ	GCT	ттт	TAT	ידידאי	GGC	TCC	CCA	CAT	ттт		TCG	TTT	тат	CTA	CAT	АТТ	TCA	AAA	TTT	TGT	ACT	TAT	TTT	TCG	TAC	АТТ	АТА	CCTC	GAT'	TTG	5940

ΑΑΑΤCTT**TAA**TTAAAAAAAAAAAAAAAA 5965 3'

Fig. 2 (continued)

and 27% in SpCRS, while the same composition in the C-terminal region was 24% and 38%, respectively. In SpCRS the first S/T/P-region was located between SCRs 2 and 3, the FIMAC domain was located C-terminal of the SCR domains, and the second S/T/P-rich region was located C-terminal to the FIMAC domain. The 18 SCRs in SpCRL were spread throughout the protein and the Cysrich region, which had eight Cys in a span of 57 amino acids, was located between SCRs 1 and 2 (Fig. 4). A fucolectin domain in SpCRL was positioned between SCRs 10 and 11, with the first S/T/P-rich region between SCRs 16 and 17. The FIMAC domain in SpCRL was located between SCR 18 and the second S/T/P-rich region.

Glycosylation sites in each protein were identified by searching for the conserved amino acid sequences NXS or NXT, and two prediction programs for O-linked glycosylation sites were used to identify O-linked sites (Hansen et al. 1997, 1998). Both proteins had many conserved sites for both O-linked and N-linked oligosaccharides (Figs. 2, 3). There were 22 O-linked glycosylation sites in SpCRL and 17 in SpCRS, while SpCRL had 17 N-linked glycosylation sites and SpCRS had seven (Fig. 4). In both proteins, many (40% for SpCRL and 64% for SpCRS) of the conserved glycosylation sites were located in the Cterminal S/T/P-rich region (Fig. 4). The rest of the sites in both proteins were located throughout the sequence. In general, both proteins showed similar distributions of the conserved glycosylation sites, and both may be highly glycosylated, particularly at the C-terminal end.

Phylogenetic analysis of domains

An approach for predicting functions of domains and proteins as a whole when the only available data are sequences is to align the amino acids with other sequences from known proteins and to use phylogenetic analysis to identify similarities which can be used to infer function. For the SCR, FIMAC, and fucolectin domains, there were enough sequences available from other proteins that have been characterized previously, some with known function, to make this approach feasible. On the other hand, this approach was not feasible for the S/T/P-rich regions and the Cys-rich region. It is noteworthy that an S/T/Prich region is present at the C-terminus of MCP, which is involved in protecting self-cells from autologous complement attack; however, the functions mediated by the S/ T/P region of MCP are not clear (Liszewski and Atkinson 1992).

Short consensus repeats

SCRs are found in many proteins that function in the complement system; however, many are also present in immune-related proteins that are not complement components, and others that are present in proteins that are not involved in immune responses. The conserved amino acids within SCRs include four Cys, three Gly, two Pro, two Tyr or Phe, and one Trp (Chou and Heinrickson 1997). An alignment of the SCRs from both SpCRL and SpCRS was done using ClustalX (Thompson et al. 1997) and demonstrated that they were typical SCRs, with four Cys and at least 9 of the 12 conserved amino acids present in each domain (Fig. 5). When the alignment was used to generate an unrooted neighbor-joining tree (Saitou and Nei 1987), results showed that the four C-terminal SCRs in SpCRL (SCRs 15–18) were most similar to the four SCRs in SpCRS, and that they were positioned in the same relative order within the two proteins. (This similarity in SCR sequence and domain organization is indicated in Fig. 4 by fill patterns.) This region, which included the four C-terminal SCRs from SpCRL, an Nterminal S/T/P-region, and the FIMAC domain but excluded the C terminal S/T/P-region (which was quite different in sequence and length between the two deduced proteins), was 37% identical and 56% similar between the two proteins (Fig. 4). For the remaining SCRs in SpCRL, four (SCRs 1, 8, 9, and 13) clustered together, with SCR 8 being most similar to SCR 13, and SCR 1 being most similar to SCR 9. Two additional pairs of similar SCRs were identified in SpCRL; SCR 4 was most similar to SCR 14, and SCR 7 was most similar to SCR 12 (denoted with similar patterns in Fig. 4). Although the relative orders of the SCRs at the C-terminus of SpCRL were the same as those in SpCRS, the two sets of SCRs within SpCRL that show sequence similarities were not positioned in the same order (1, 4, 7, 8 vs 9, 14, 12, 13; see Fig. 4).

The phylogenetic analysis was expanded to identify sequence similarities between the SCRs from SpCRS

Sp5013 and SpCRS

	110111		AAC	IGA	AGC	AG	I'CAA	AC	CAG	TTG	GCC.	ACG.	ATG	GAA	GGA.	AAG	CGA	ГТТG	AAG	ACA	TTGI	CCG	AAGi	ATTO	GATC	TAT:	A.L.L.	CAA	CTA	TTC	CIC	AC'I''.	ГСА	108
													М	E	G	Κ	R	F	E J	D	ΙV	7 R	R	L	I	Y	I	Q	L	F	L	т	S	22
																							Le	eade	er									
GTTTG	CGTI	GGC	TCA	rcgc	CAT	GCF	ATGO	СТ	AAG.	ACC	GAC	CGT	CTG	GGG	CGC	TAC	ACG	GTCA	TAG	GCG	GTGG	CGG'	TAA	CCTO	CTGG	TCG	ATG	GGT	GTT	CCG.	AGT	GGG.	ГСT	216
V C	V	G	S	S	Р	С	М	Ρ	K	т	D	R	L	G	R	Y	т	V	I	G (G	G G	Ν	L	W	S	М	G	V	Ρ	s	G	S	58
								~		SC	R 1																							
AGGAT	TCAA	AGTC	TCA	3TAA	GAT	GCC	JATO	G	GGT	GCT.	ATC	GCA	CAA	CAA	TCT	GTG	ACG	гсас	AGT	GCT	CTAA	TGG	AAC	ATGO	JATC	CCI	CCC	TGG	CCT	CGA	TGT	GAGA	AAA	324
RI	0	v	S	V	R	С	D	A	G	A	I	A	0	0	S	v	т	S	0 0	C	S N	I G	Т	W	Ι	Р	Р	W	P	R	cl	Е	K	94
	~	•	~	•		-	-				_		~	~	~	•	- ⁷ Cl	но C	ñ	-	~ ,	СНО	(N)		_	- 9	CR	1 4	Ē		Ŭ.	-		-
ССТТС	TTCZ	GTA	cca		מדמ		CAD	ING	GTT	CAC	CAC	ጥጥጥ	ата	שמע	сст	GAT	aca		TGC		המת	TAG	ידע. מידענ	ግጥልባ		 דיים בי	TCC		• ጥጥ አ	CCT	GAT	aaa	ACC	432
	C C	V	D D	g .	N	K K	G	F	W	н	0	- F	V	N	с С	D	20001 2	т Т	M	D	т п	, G	v	v	TT T	л 10 г Т	g c	T	т.	D	П	C C	T	130
	0	č	-	0.	d G G D D	<u>`</u> >	u	-	v	11	~	1	×.	Τ.	0	D	11	-		-			-	-	ĊĦ	ທ່ີດ	۰×	~~	н <u>о</u> (ດ້	D	0	1	100
CACCT			202	DOT N			2000		770	ሞለሞ		mma	<u>م</u> ب ب		CAR	<u> </u>	7070	707 7	CTTT	-	7 7 C C	17.00	N (7 N				, , , , , ,	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~		() ()	maa	0770	<u>م</u> م	E40
CAGCI	AGII	.GCI	AGA.	C	GAG	D		756	AAG V	IAI.	AGA D	TIG.			GAI		AGA(D	.CII(a .	AAGO		ACA	JDJE M		.9999. D	GAA	GAA		GAA		CAA		10
Qμ	V	А	ĸ	C .	ĸ	D	Р	G	ĸ	ĭ	ĸ	Ц	Ŧ	G	D	Q	R	ĸ	1 (Q	J G	Q	W	1	G	<u>ٿ</u>	E	. Р	E	C	Q	н	100
annas			ana				~~ ~ ~					naa	aam	a m		aa.	- am			a. a	maan	aam	~~~	nam	2011	CR	4	•	(amm	~~~~	<i>с</i> 1 0
GITTCA	AGCA	ACAA	GTC:	1I.Y.I.	TCI	ACC			AAC.	AA'I'	'AAA'	TGG	CGT	GAT	GTA	GCA	AGTZ	AATG	GGA	CAG	TCGI	GGT	JTA.	rGre	-CAA	CG.I	GGA	AA.I.	CTC	AGG	CGT	CTTC	JAT.	648
νQ	A	Q	v	Г	F.	Y	Q	ĸ	N	N	ĸ	W	R	D	V	А	S	N A mur	G	Τ.	V V	v v	Y	V	Q	R	G	N	Ц	R	R	Ц	D	202
																		СН	O (N)														
GTAGT	CTGC	TAA	GAC	ATAA	GTG	GAI	rggo	CTT	AGA	TAT	ACT	CAT	CTG.	ACA	ACT	CCT	CTA	CTGA	ATG	ACG	GAGA	ACTT	ATG	GCA1	rtgg	CAC	AGG	CAT	AAC	TAT	CAT	GGA	AAA	756
V V	C	N	D	I.	S	G	_W	ь.	R	_Y	T	<u>H</u>	L	_T	<u>T</u>	P	_L		N]	D	<u>G</u> [)L	W	<u>H</u>	W_	H	R	<u>H</u>	_N_	_Y	H	G	_K	238
																	s/:	Г/Р	ric	h r	egic	n												
CGGTT	GTAI	CCT	ATC	FCTA	CTG	AT	[TC]	CT	GGC.	ATG	TAT.	ACA	TGC	AAT	GGG	TCG	TCA	ГСАТ	CCC	ACA	GCGI	'ACA'	FGT (GCT	JTTT	AAA	GCT	GCT	GAA	TGC	GAC	GTC	CCT	864
R L	Y	P	I	S	T	D	F	s	G	Μ	_Y	T_	C	_N	G	S	S	S	S 1	H I	<u>s</u> v	7 Н	V	L	F	K	А	А	Е	С	D	V	Ρ	274
							s/1	Г/Р	ri	ch	reg	ion		^C	HO (0)	^CI	HO (N	ſ)													→so	CR 3	
GCGGC	TCCA	AGAG	CAC	GGGA	GGA	TAC	CTAG	JAT	GGT	TAC	AGA	CAG	CAG	GCT	CGA	GAT	ATC	TTTC	CAT	CTG	GAAC	TTC	GGT.	TTC	ATTC	GAA	TGT	GAA	GAT	GAT	TAC	CGC	TTG	972
A A	P	Е	Н	G	R	I	L	D	G	Y	R	Q	Q	A	R	D	I	F	P :	s (GΊ	: s	V	S	F	Е	С	Ε	D	D	Y	R	L	310
GTGGG	ATCI	AGG	AGA	ATCA	CAT	GTC	JAAF	٩AG	GGT	CAA	TGG	TCG	GAT	GAT	CCT	CCC	GTA	TGTA	CAG	TGA	GCAG	GTG	CCC	GGA	GCTA	ACC	ATA	CCA	GAC	CAT	GCA	ATC	ATG	1080
VG	S	R	R	I	т	С	Е	К	G	0	W	S	D	D	Ρ	Ρ	v	сΙ	Т	v	S F	lC	Р	Е	L	т	I	Ρ	D	Н	А	I	М	346
										~		SC	R 3						^CHO	0(0)			~		sc	'R 4							
CACCC	COTTO	יידי אידיוי		r a m ca	007	000	77.07	mm	amm	a a *		a. a				•						-		_								mma	D & F	1100
GAGGG	CGTC	JAI	JUD.	LAIG	GGA	וססנ	JAUP	7 T T	CII	CGA	GAT!	GAG	AGG	ATC	TTC	TAC	TGTZ	AAAG	AGG	GTT.	ATGI	TCT	ACGZ	AGGT	rgat	ACC	'AAC.	AAC	CTG	CGA	TGC	TTGO	JAG	TT 8 8
E G	V	Y Y	G	Y Y	GGA G	R	D	T	L L	R	GAT D	GAG. E	AGG. R	ATC T	TTC' F	TAC Y	TGTZ C	AAAG K	AGG E (GTT. G	ATGI Y V	TCT2	ACG2 R	AGGI G	rgat D	'ACC T	'AAC. N	AAC N	CTG L	CGA' R	TGC C	TTGC	JAG E	382
E G	V	Y	G	Y	GGA	R	D	I	L	R	D	GAG. E	R R	ATC I	TTC' F	TAC Y	TGTZ C	AAAG K	AGG(E (GTT. G	ATGI Y V	TCT2 7 L	ACG2 R	AGG: G	rgat D	'ACC T	'AAC. N	AAC N	CTG L	CGA' R	TGC C	L	E	382
E G	V	Y	G	Y	GGA G	R R	D	I	L	R R	GAT D GCT	GAG E GTT	AGG. R GTT	ATC I CAT	TTC F	TAC Y GTG	TGTZ C TCA	AAAG K	AGG E (GTT. G	ATGI Y V GCGI	TCT L	ACG2 R	AGGI G TGAZ	rgai D	T T	AAC. N	AAC N	CTG L TGT	CGA R GAG	TGC C GTG		E	1188 382
E G	GGTC	Y TGG	G	Y GATC	GGA G CCT	R TAC	D	I ACC	TGT	R R GAA	GAT D .GCT	GAG E GTT	AGG. R GTT	ATC I CAT	TTC' F CAA	TAC Y GTG	TGTZ C TCA	AAAG K CCGT	AGG E (GTG	GTT. G CGA	ATGI Y V GCGI	TCT L GCG	ACGI R CTG	AGGI G IGAZ	IGAI D ATCA	T T TGG	'AAC. N CAG	AAC N CGA	CTG L TGT	CGA' R GAG	TGC C GTG	ACG(EGGA	1188 382 1296
E G AGTGG S G	GGTC V	Y TGG W	G	Y GATC	GGA G CCT P	R TAC L	D CCCF P	I I ACC T	L TGT C	R GAA E	GAT D .GCT A	GAG E GTT V	R R GTT V	ATC I CAT H	TTC' F CAA Q	TAC Y GTG V	TGTZ C TCA S	AAAG K CCGI P	AGG E (GTG C	GTT. G CGA	ATGT Y V GCGT <u>S V</u>	TCT L GCG <u>R</u>	ACG2 R CTG1	AGG: G IGAZ	IGAT D ATCA	T T TGG	AAC. N CAG	AAC N CGA R	CTG L TGT 	CGA R GAG E	TGC C GTG	ACG	E GGA	1188 382 1296 418
E G AGTGG S G	GGTC V	Y TGG W	G TCG S SCI	Y GATC D R 4	GGA CCT P			I I ACC T		GAA E	GAT D .GCT A	GAG E GTT V	R R GTT V	ATC I CAT H	TTC F CAA Q	TAC Y GTG V	TGTZ C TCAO S CI	AAAG K CCGI P HO(C	AGG E GTG C D	GTT. G CGA	ATGI Y V GCGI <u>S V</u>	TCT L GCG R R	ACG2 R CTG7	AGG G IGAZ FIN	IGAT D ATCA S IAC	TGG M don	AAC. N CAG	AAC N CGA R	CTG L TGT 	CGA R GAG E CH	TGC C GTG V O (O	ACGO	E GGA	1188 382 1296 418
E G AGTGG S G CCAAA	GGTC V CACI	Y STGG W CACA	G TCG S SCI	Y GATC D R 4 IGTG	GGA CCT P GCT	R TAC L	D CCCF P	I I ACC T AAT		R GAA E ACA	GAT D GCT A .TCA	GAG E GTT V TGC	AGG. R GTT V CTG.	ATC I CAT H ACG	TTC' F CAA Q TCA	TAC Y GTG V GGG	TGTZ C TCA(S ^CI CCGZ	AAAG K CCGI P HO(C ACTG	AGG E GTG C C) TCT U	GTT. G CGA A GCG	ATGI Y V GCGI <u>S V</u> GAAC	TCT L CGCG L CAAA	ACGI R CTG CCGGI	AGG G IGAZ <u>E</u> FIN AGAC	IGAT D ATCA <u>S</u> IAC CTAT	T T TGG M dom	CAG	AAC N CGA R AGT	CTG L TGT C GAA	CGA R GAG E CH TGC	TGC C GTG V O (O CAT	ACGO	E E E G G AGG	1188 382 1296 418 1404
E G AGTGG S G CCAAA P	GGTC V CACT	Y FTGG W TACA	G TCG S SCI GTC	Y GATC D R 4 IGTG	GGA CCT P GCT G	R TAC L GTZ			CII I TGT C C C C C C C C	R GAA E ACA T	GAT D GCT A TCA	GAG E GTT V TGC	AGG. R GTT V CTG.	ATC I CAT H ACG	TTC' F CAA Q TCA	TAC Y GTG V GGG	TGTZ C TCA(S CCGZ	AAAG K CCGI P HO (C ACTG T	AGG E C C T C V V	GTT. G CGA A GCG C	ATGI Y V GCGI <u>S V</u> GAAC <u>G</u> I	TCT CGCG CRCG CAAA	ACGI R CTG C CGGI G	AGG G IGAZ FIN AGAC	IGAT D ATCA S MAC CTAT Y	TGG TGG M dom TAC	AAC N CAG Q Nain AGC S	AAC N CGA R AGT S	CTG L TGT GAA 	CGA R GAG E CH TGC C	TGC C GTG O (O CAT 	L ACG()^ CTCZ	E EGA G AGG R	1188 382 1296 418 1404 454
E G AGTGG S G CCAAA <u>P N</u>	GGTC V CACI CACI	Y TGG W TACA T (N)	GGC G TCG S SC GTC V	Y BATC D R 4 IGTG	GGA CCT P GCT G	R TAC L GTZ CHC	D CCCF P ATCF I D (N)		TGT C C C C C C C C C	R GAA E ACA	GAT D GCT A TCA S C	GAG E GTT V TGC	AGG. R GTT V CTG. CTG.	ATC I CAT H ACG T	TTC' F CAA' Q TCA' <u>S</u>	TAC Y GTG V GGG GGG HO (TGTZ C TCA(S ^CI CCGZ P O)	AAAG K CCGT P HO (C ACTG	AGG E C C T C C T C C C C C C C C C C C C C	GTT. G CGA A GCG C	ATGI Y V GCGI S V GAAC G I doma	TCT L CGCG M CAAA L I N	ACGI R CTG C CGGI G	AGGI G IGAZ FIN AGAC D	rgat D ATCA S IAC CTAT	TGG TGG M dom TAC	AAC. N CAG Q Nain AGC. S	AAC N CGA R AGT S	CTG L TGT GAA E	CGA R GAG E CH TGC C	TGC C GTG O (O CAT H	L ACGO T)^ CTCZ	E EGA _G AGG _R	1188 382 1296 418 1404 454
E G AGTGG S G CCAAA P	GGTC V CACT CACT CHO(TGCC	Y TGG W TACA T (N)	GGC G S GTC GTC V TTA	GATC D R 4 IGTG C CATG	GGA CCT P GCT G ACC	R TAC L GTZ CHC	D CCCF P ATCF I CAGO	ACC T AAT	CTGT C C CCCC	R GAA E ACA T ATC	GAT D GCT A TCA S C	GAG E GTT V TGC C HO (AGG. R GTT V CTG. <u>L</u> O) AAC	ATC I CAT H ACG T GAC	TTC F CAA Q TCA TCA <u>S</u> C TAT	TAC Y GTG V GGG HO (GGA	TGTZ C TCA(S ^CI CCGZ P O) TGCZ	AAAG K CCGT P HO (C ACTG T ATTA	AGG(E C C C C C C C C C C C C C C C C C C	GTT. G CGA A GCG C GGG	ATGI Y V GCGI <u>S V</u> GAAC <u>G I</u> dom a CTCC	TCT2 CGCG(R CAAA(CAAA) CAAA(CAAA)	ACGI R CTG C CGGI CGGI IGA(AGGI G IGAZ FIN AGAC D GGAC	rgat D ATCA MAC CTAT Y CCAG	TGG TGG M dom TAC	AAC. N CAG Q AGC. S ACG.	AAC N CGA R AGT S ATA	CTG L TGT GAA E CAA	CGA R GAG E CH TGC C CTT.	TGC C GTG O (O CAT ACT	L ACG(E EGA _G AGG _R GAT	1188 382 1296 418 1404 454 1512
E G AGTGG S G CCAAA <u>P_N</u> GCTTA A_Y	GGTC V CACI T CHO (TGCC	Y TGG W TACA T (N) CTGC	GGC G TCG(S GTC GTC TTA(Y BATC D R 4 IGTG C C TATG	GGA CCT P GCT G ACC	R TAC L CGTZ CHC L	D CCCZ P ATCZ I D (N) CAGC	I ACC T AT	CTGT CCCC CCCC 	R GAA E ACA T ATC	GAT D GCT A TCA S C GCA	GAG E GTT V TGC C HO (GAG	AGG. R GTT V CTG. L O) AACO	ATC I CAT H ACG T GAC	TTC F CAA Q TCA S C TAT	TAC Y GTG V GGG HO (GGA	TGTZ C TCA(S CCGZ CCGZ O) TGCZ	AAAG K CCGI P HO (C ACTG T ATTA	AGG(E (C) (C)	GTT. G CGA A GCG C C GGG GGG	ATGI Y V GCGI S <u>V</u> GAAC G <u>I</u> doma CTCC A <u>I</u>	TCT2 TCC3 TCC3 TCC3 TCC3 TCC3 TCC3 TCC3	ACGI R CTG C CGGI C G I G A C G E	AGGI G IGAZ FIN AGAC D GGAC D	rgat D ATCA S MAC CTAT Y CCAG Q	TACC T ATGG dom TAC Y GCCI	CAAC. N CAG Nain AGC. S YACG.	AAC N CGA R AGT S ATA I	CTG L TGT GAA E CAA Q	CGA R GAG E TGC C C C TT C C TT C	TGC C GTG • V • O (O CAT • H • ACT • T	ACGO T)^ CTCZ GCTCZ	E EGA G AGG R GAT D	1188 382 1296 418 1404 454 1512 490
E G AGTGG S G CCAAA P A GCTTA	GGTC V CACT CACT CHO (TGCC	TACA TACA T (N) TACA	GGC G S GTC GTC V TTA	GATC D R 4 FGTG C C ATG	GCT P GCT G ACC D	R TAC L CGTZ CHC TCC L	D CCCF P ATCF I CAGC		CTGT C C CCCC GAT GAT	GAA E ACA T ATC I C d	GAT D GCT A TCA S C GCA GCA	GAG E GTT V TGC HO(GAG E in	AGG. R GTT V CTG. L O) AAC	ATC I CAT H ACG T GAC GAC	TTC F CAA Q TCA S C TAT	TAC Y GTG V GGG HO (GGA	TGTZ C TCA(S CCGZ P O) TGCZ	AAAG K CCGT P HO(C ACTG T ATTA	AGG(E (C C C TCT(V FIM ATG(N N	GTT. G CGA A GCG C C C GGG GGG G G G	ATGT Y V GCGT S V GAAC G I doma CTCC A F	TCT2 CGCG(CAAA(CAAA(CAAA) CTGA CTGA CTGA CTGA	ACGI R CTG C CGGI C G G I GAO E	AGGI G IGAZ FIN AGAC D GGAC D	rgat D ATCA S MAC CTAT Y CCAG Q	TACC T ATGO dom TAC TAC Y CCT	CAAC. N CAG AGC. S VACG. - T C	AAC N CGA AGT S ATA I HO(CTG L TGT GAA E CAA Q O)	CGA R GAG E TGC CH CTT. L	TGC C GTG O(O CAT H ACT T	ACGO T)^ CTCZ GCTO A	E GGA G AGG R GAT D	1296 418 1404 454 1512 490
E G AGTGG S G CCAAA P GCTTA A GAGAG	GGTC V CACI CACI CHO (TGCC TCCC	TACA TACA (N) TGC	GGC G S GTC GTC TTA GCC	GATC D R 4 FGTG CATG H	GCT P GCT G ACC D CAT	R TAC L GTZ CHC TCC L CAT	D CCCF P ATCF I D (N) CAGC	I I ACCC T AAT N CTA CTA F FCG	CTGT C C C C C C C C C C C C C C C C C C	GAA E ACA T ATC I GAC	GAT D GCT A TCA S C GCA GCA	GAG E GTT V TGC GAG GAG In ACT	AGG. R GTT V CTG. <u>L</u> AAC <u>N</u> CTC.	ATC I CAT H ACG T GAC GAC	TTC' F Q TCA' TCA' <u>S</u> TAT' Y GGT'	TAC Y GTG GGG HO (GGA GCT	TGTZ C TCA(S CCGZ CCGZ P O) TGCZ CAGZ	AAAG K CCGI P HO(C ACTG T ATTA L AGCG	AGG E C C TCT V FIM ATG N N TTC	GTT. G CGA A GCG C AC GGG GGG CTA	ATGI Y V GCGI S V GAAC G I doma CTCC A F	TCTA CGCG(ZAAA) CAAA) CAAA) CTGA CTGA CTGA CTGA	ACG2 R CTG7 CGG2 G IGA0 E ICC2	AGGI G IGAZ FIN AGAC D GGAC	rgai D ATCA MAC CTAI Y CCAG Q CGAA	TACC T ATGO dom TAC Y CCT CCT P	CAGC Q Q Q AGC S C ACG C ACG	AAC N CGA AGT S ATA I CCG	CTG L TGT GAA E CAA Q CCG	CGA' R GAG CH TGC C C CTT. L ACA	TGC C GTG 0 (0 CAT H ACT T GAG	L ACGO T OCTCZ GCTCZ A TCAC	E EGA G AGG R EAT D ECA	1188 382 1296 418 1404 454 1512 490 1620
E G AGTGG S G CCAAA P GCTTA AY GAGAG ES	GGTG V CACI T CHO C TGCC TCCG	Y TACA W TACA (N) CTGC C C C C C C C C C C C C C C C C C	GGC G S GTC GTC TTA GCC A	GATC D R 4 FGTG C CATG H FATT	GCT P GCT G G ACC D CAT S	R TTAC L CGTA C CHC TTCC L CAT S	D CCCF P ATCF I CAGC O N) CAGC	L T L T T T T T T T T T T T C G	CCCC CCCC GAT GAT TTT F	GAA E ACA T ATC I GAC	GAT D GCT A TCA C GCA GCA GCA GCA	GAG E GTT V TGC C HO(GAG E In ACT T	AGG. R GTT V CTG. D AAC N CTC.	ATC I CAT H ACG T GAC GAC D ATT I	TTC F Q TCA TCA TCA C TAT GGT GGT	TAC Y GTG GGG HO (GGA GGA GCT A	TGTZ C TCA(S CCGZ CCGZ O) TGCZ CAGZ	AAAG K CCGT P HO(C ACTG T ATTA I AGCG S	AGG E C C TCT V FIM ATG N TTC V	GTT. G CGA A GCG C GGG GGG G GGG CTA P	ATGI Y V GCGI S V GAAC G I doma CTCC A F CTGA	TCTA CGCG(Z_R CAAA(Z_N CTGA CTGA CTGA CTGA C D	ACG2 R CTG7 CGG2 GG2 IGA0 E ICC2 P	AGGI G IGAZ FIN AGAC D GGAC D ACAC	rgat D ATCA MAC CTAT CTAT CCAG Q CGAA		AAC. N CAG AGC. AGC. YACG. T C. GAC	AAC N CGA AGT AGT ATA I CCG P	CTG L TGT GAA CAA CAA O) CCG	CGA' R GAG TGC CH TGC C CTT. L ACA T	TGC C GTG O(O CAT H ACT T GAG	L ACGO T CTCZ GCTO A TCAO S	E EGA G AGG R EAT D GCA	1188 382 1296 418 1404 454 1512 490 1620 526
E G AGTGG S G CCAAA P_N GCTTA A_Y GAGAG E S	GGTG V CACT T CHO TGCC A TCCG	TACA T (N) CTGC C GTA	GGC S SCI GTC TTAC TTAC GCC A	GATC D R 4 FGTG C CATG H FATT	GCT P GCT G ACC D CAT S	TAC L CTAC CTCC CTCC L S CAT	D CCCF P ATCF I CAGC Q CCCI S	L T L T T AAT T T T T C T C T C G S	GAT GAT C CCCC CCC CCC CCC C CCCC C CCCC C C C	R GAA E ACA T ATC I GAC GAC D	GAT D GCT A TCA C GCA GCA GCA GCA C GCA C GCA C GCA C T/P	E GTT V TGC C HO(GAG E in ACT T ri	AGG. R GTT V CTG. D AAC N CTC. L CTC.	ATC I CAT H ACG GAC D ATT I reg	TTC F Q TCA TCA TCA C TAT GGT GGT	TAC Y GTG V GGG GGA GGA GCT A	TGTZ C TCAG S CCGZ P O) TGCZ C CAGZ	AAAG K CCGI P HO(C ACTG T ATTA I AGCG S 	AGG E (C C TCT V FIM ATG N V V V	GTT. G GCGA GCG G GGG G G G G G G G G G G G	ATGI Y V GCGI S V GAAC G I doma CTCC A F CTCA T E	TCTI GCG(CAAA) CAAA) CTGA CTGA CTGA CTGA CTGA CTGA CTGA CTGA	ACGJ R CTG C CGGJ C G G IGA E IGA E ICC 2 P	AGG G IGAZ FIN AGAC D GGAC D ACAC	rgat D ATCA MAC CTAT CTAT CCAG Q CGAA E	T T T T T T T C C C T T C C C T C C T C C T C C T C T C C C T	AAC. N CAG AGC. S CACG. S C ACG. C ACG. C ACG. C ACG.	AAC N CGA R AGT S ATA I HO (CCG P	CTG L TGT C GAA E CAA Q O) CCG P	CGA R GAG CH TGC CTT. L ACA T C	TGC C GTG V O(O CAT H ACT T GAG GAG HO(L ACG()^ CTCZ GCTC A TCA(S O)	E E GGA G AGG R GAT D GCA A C A C C A C C C C C C C C C C C C C	1188 382 1296 418 1404 454 1512 490 1620 526
E G AGTGG S G CCAAA PN GCTTA AY GAGAG E S GCCGA	GGTC V CACT TGCC TGCC TCCC P TATC	Y FTGG W TACA T (N) CTGC CTGC CGTA	GGC S SCI GTC TTAC TTAC GCC A	GATC D R 4 IGTG CATG H IATT Y CCTG	GCA G GCT G G ACC D CAT S CTG	R TAC L GTA C C C C C C C C C C C C C C C C C C C	D CCCCF P ATCF I D (N) CAGC Q FCC7 S FAT7	L T T T T T T T T T T C G T C G T A C T A C T A C C C T A C C T T T T T	TGT C C C C C C C C C C C C C C C C C C	R GAA E ACA T ATC C d GAC D S/	GAT D GCT A TCA C GCA GCA GCA GCA GCA GCA GCA C T/P	GAG E GTT V TGC C HO(GAG E In ACT T TCC	AGG. R GTT V CTG. L O) AACO N CTC. L CTC. CTG	ATC I CAT H ACG GAC GAC D ATT reg GTC	TTC F CAA Q TCA S C TAT GGT G G I On ATG	TAC Y GTG V GGG GGA GGA GCT A GAT	TGTZ C TCA(S CCGZ P CCGZ P O) TGCZ CAGZ CAGZ	AAAG K CCGI P HO(C ACTG T ATTA I AGCG S CCAG	AGG(E (C) TCT(V) FIM ATG(N) TTC(V) CTTC(V)	GTT. G GCGA GCG G GGG G G G G G CTA P - CTA ATG.	ATGI Y V GCGI S V GAAC G I doma CTCC A F CTCA T F ^ CHC	TCTI GCG(CAAA) CAAA) CTGA CTGA CTGA CTGA CTGA CTGA CTGA CTGA	ACG2 R CTG' C CGG2 CGG2 IGA0 E ICC2 P CGA2	AGGT G IGAA FIN AGAC D GGAC D ACAC H AGAC	IGAT D ATCA S MAC CTAT Y CCAG Q CGAA E CACT	TACC T ATGG M dom TACC TACC Y ACCT P ACCT P ACCA A CAAA	AAC. N CAG Q AGC. S ACG. C ACG. C C C C C C C C C C C C C C C C C C	AAC N CGA AGT S ATA I HO (CCCG P TCA	CTG L TGT C GAA E CAA Q O) CCG P CCA	CGA R GAG CH TGC CTT. L CTT. L CTT. CTT. CTT.	TGC C GTG V O(O CAT H ACT T GAG GAG E HO(C CCT	ITGO L ACGO T O CTCZ GCTO A TCAO S O) CCCZ	E GGA G AGG R GAT D GCA ACA	1188 382 1296 418 1404 454 1512 490 1620 526 1728
E G AGTGG S G CCAAA PN GCTTA AY GAGAG E S GCCGA A D	GGTC V CACT TGCC TGCC TGCC TCCC P TATC I	Y FTGG W CACA T (N) CTGC CGTA CGAT D	GGCC GTC SCI GTC TTA GCC A TCT S	GATC D R 4 IGTG CATG H IATT Y CCTG P	GCT G GCT G ACC D CAT S CTG A	R TAC CHC CHC CHC CHC CHC CHC CHC CHC CHC C	D CCCCF P ATCF I D CAGC Q CCCT S CCCT S CCCT Y	I ACCC T N T T T T T T T T T T T T T T T T	TGT CCCC. P GAT. D IMA TTT F TCT S	R GAA E ACA T ATC GAC GAC D S/ TTC F	GATI D GCTI A TCA S C GCA GAG GAG E T/P ATTI I	GAG E GTT V TGC C HO(GAG E In ACT T TCC S	AGG. R GTT V CTG. L O) AACO N CTC. L CTG L	ATC I CAT H ACG GAC D GAC I I I GTC V	TTC' F CAA Q TCA S C TAT GGT G G G G I On ATG M	TAC Y GTG GGG G HO (GGA GCT A GAT D	TGT2 C TCA(S ^CI CCG2 P O) TGC2 C C C C C C C C C C C C C C C C C C	AAAG K CCGI P HO(C ACTG T ACTG T ACTG S ACTG S CCAG	AGG(E (GTG(C)) (TCT(V (FIM) ATG(N (V) (CTTC(V) (CTTC(A)	GTT. G GCGA GCG GGG G GGG G G CTA P P I ATG. D	ATGI Y V GCGI S V GAAC G I doma CTCC A CTGA T E ^ CTGA T E ^ CTGA T E N AAAA E N	TCTI GCG(CAAA(CAAA) CTGA CTGA CTGA CTGA CTGA CTGA CTGA CTGA	ACG2 R CTG? CGG2 G IGA0 E ICC2 P CGA2 E	AGG G FID FID AGAC D GGAC D ACAC H H AGAC D	IGAI D ATCA MAC CTAI Y CCAG Q CGAA E CGAA E CACI T	TACC T ATGG M don TACC TACC Y SCCT P AGCA A A CAAA K	AAC. N CAG CAG CAG CAG CAG CAG CAG CAG CAG CAG	AAC N CGA AGT S ATA I CCG P TCA S	CTG L TGT C GAA E CAA Q O) CCG P CCA P	CGA R GAG CH TGC CTT. L ACA T C CT P	TGC C GTG V O(O CAT H ACT T GAG GAG CCT P	ACGO T CTCZ GCTCZ GCTCZ A TCAC S O) CCCZ P	E GGA G AGG R GAT D GCA A ACA T	1188 382 1296 418 1404 454 1512 490 1620 526 1728 562
E G AGTGG S G CCAAA P GCTTA AY GAGAG E S GCCGA A D	GGTG V CACI TGCO TGCO TGCO P TATCO I	Y FTGG W TACA T (N) TTGC CTGC CGTA CGAT D	GGCC SCI GTC ⁻ V TTA(L GCC ⁻ A TCT(S	PATC D R 4 IGTG CATG H IATT Y CCTG P	GCT G GCT G G ACC D CAT S CTG A	R TTAC L GTZ C CHC CTCC L CCA1 S CA1 S AC1 D	D CCCCF P ATCF I D CAGC O CAGC C CCCT S CCCT S CCCT Y	I ACC T AAT N T T S TAC Y S	TGT CCCC. P GAT. D IMA TTT F TCT S	GAA E ACA T ATC GAC GAC D TTC F P r	GATI D GCTI A TCA S C GCA Oma GGAG. E T/P ATT I I tch	GAG E GTT V TGC C HO(GAG E In ACT T TCC S re	AGG. R GTTY V CTG. L CTC. CTC. CTC. L CTG. L GIO	ATC I CAT H GACG D GAC L J STC V N	TTC F CAA Q TCA S C TAT G G T TAT G G I O N	TAC Y GTG GGG G HO (GGA G GCT A GAT D	TGT2 C TCA(S ^CI CCG2 P CCG2 C O) TGC2 C C C C C C C C C C C C C C C C C C	AAAAG K CCGT P HO (C ACTG T ATTA I AATTA S CCAG P	AGG(E (C) C) TCT(V (FIM) ATG(V) CTTC(V) CTTC(A)	GTT. G GCGA GCG GGG G GGG CTA CTA CTA D	ATGT Y V GCGT S V GAAC G I doma CTCC A A A A A A A A A A A A A A A A A	TTCTI TCTI TCCCGCG TCCCCCC TCCCCCCCCCCCCCCCCCCC	ACGJ R CTG CGGJ CGGJ G IGAG E ICCJ E CGAJ E	AGG G IGAA FIN AGAC D GGAC D C GGAC H H AGAC D D	IGAI D ATCA MAC CTAI Y CCAG Q CGAA E CGAA E CACI T	TACC T ATGG W TACC TAC Y CCT P AGCA A A GCA A K TAAA K	ZAAC. N GCAG Q Main ZAGC. ZACG. C GACC. D GACC. N N O)	AAC N CGA AGT S ATA I CCG CCG P TCA	CTG L TGT C GAA E CAA Q O) CCG P CCA P HO (CGA R GAG CH TGC C CTT. L ACA T C CTT P O)	TGC C GTG V O(O CAT H ACT T GAG GAG GAG CCT P CH	TTG(L ACG(T)^ CTC2 L GCT(A TCA(S O) CCC2/ P O(O)	E GGA G AGG R GAT D GCA A ACA T	1188 382 1296 418 1404 454 1512 490 1620 526 1728 562
E G AGTGG S G CCAAA PN GCTTA AY GAGAG ES GCCGA A D ACCCG	V GGTC V CACT TCCC P TTCCC I TTCCC I	Y TGG W TACA T (N) TTGC CTGC CGTA CGAT D	GGCC SCI GTC ^C V TTA(L GCC ^C A TCT(S	Y GATC D R 4 IGTG C CATG H IATT Y CCTG P GCTA	GCA G CCT G G ACC D CAT S CTG A GAA	R TTAC L GTA C C C C C C C C C C C C C C C C C C C	D CCCCF P ATCF I D (N) CAGC Q CCCT S S FATT Y	I ACC T AAT CTA F CG S TAC Y S	TGT C C C C C C C C C C C C C C C C C C	GAA E ACA T ATC GAC GAC D TTC F F P r ACA	GAT D GCT A TCA S C GCA OMA GAG. C T/P ATT I I CA	GAG E GTT V TGC C HO(GAG E In TCC S CAG GAG	AGG. R GTT: V CTG. L CTG. L CTG. L GTG. L STD.	ATC I CAT H ACG GACG D GTC V T C GTC V N ACC	TTC' F CAA' Q TCA' S ^C: TAT' GGT' G G Ion ATG M	TAC Y GTG C GGG G G G G G G C T T T T T T	TGT2 C TCA(S ^CI CCG2 P CCG2 C P TGC2 C C C C C C C C C C C C C C C C C C	AAAG K CCGT P HO(C ACTG 	AGG(E (E (C (C V V V ATG(CTTC(CTTC(CTTC(CTTC(GTT. G GCGA GCGG G GGG G G CTA P P ATG. D	ATGT Y V GCGT S V GAAC G 1 doma CTCC A A A A A A A A A A A A A A A A A	TTCTA CGCGG CAAAAA CAAAAAA CAAAAA CAAAAA CAAAAAA CAAAAAA CAAAAAA CAAAAAA CAAAAAA CAAAAAAA CAAAAAAA CAAAAAAAA	ACG2 R CTG? CGG2 G G IGA0 E ICC2 E CGA2 E	AGG G G FIN FIN GGA0 D GGA0 D ACA0 D ACA0 D	IGAI D ATCA S MAC CTAI Y CCAG Q CCAAG Q CCAAG CCAAG	TACC T ATGG M M TAC TAC TAC P AGCA A CAAA K CAAA K CHO (AAC. N CAG AGC. S CACG. C C GACC. D GAC N N O)	AAC N CGA AGT S AGT I CCG P TCA S ^C	CTG L TGT GAA E CAA Q O) CCG P CCA P HO (CCT	CGA R GAG CH TGC C C TT L C C C C C C T P O)	TGC C GTG V O(O CAT H ACT T GAG GAG E HO(CCT P CH	TTG(L ACGG T)^ CTCZ L GCTC A TCAG S O) CCCC2 P O(O) GCTC	E GGA G AGG R GAT D GCA A ACA T)^ TT	1188 382 1296 418 1404 454 1512 490 1620 526 1728 562 1836
E G AGTGG S G CCAAA P_N GCTTA A_Y GAGAG E S GCCGA A D ACCCGG T R	GGTC V GGTC V CACT T CHO (TGCC P TATCCC I TATCC I TATCC I	Y GTGG W CACA T (N) CTGC CGTA CGAT D CCAA	GGC G S SCI GTC V TTA GCC A S CCA P	FATC Y GATC C C CATG C CATG H FATT Y S CCTG P GCTA	GCAT P GCT G ACC D CAT S CAT S CAT S A CAT S A CAT S A CAT S A CAT S A CAT S CAT S CAT S CAT S CAT S C C C C C C C C C C C C C C C C C C	R TTAC L GTF C CHC TTCC L CA1 S CA1 S CA1 S CA1 S CA1 S	D CCCCF P ATCF I D (N) CAGC Q CCCT S CCCT S FATT Y	L T L T T T T T T T T T T T T T T T T T	TGT C C C C C C C C C C C C C C C C C C	R GAA E ACA T C d GAC J GAC D S/ TTC F P r AACA T	GAT D GCT A TCA S C GCA GCA GAG GAG GAG T/P ATT I I I C T/P	GAG E GTT V TGC C HO(GAG E T TCC S re GAG E	AGG. R GTT: V CTG. L CTG. L CTG. L GIO: TTA. L	ATC I CAT H ACG GACG D GACG V Feg GTC V ACC T	TTC F CAA Q TCA S C TAT TAT G G I O A TG M C C T P	TAC Y GTG V GGG G G HO (GGA G G G TTT F	TGT/ C TCA(S ^CI CCG/ D TGC/ CAG/ CAG/ CAG/ CAG/ CAG/ CAG/ CAG/ C	AAAG K CCGT P HO(C ACTG T ATTA AGCG S CCAG P ICAA	AGGG E (C C C C C C C C C C C C C C C C C C C	GTT. G CGA GCG C GGG G G CTA P P ATG D	ATGT Y V GCGT S V GAAC G I doma CTCC A A A A F CTCGA A A A A A A A A A A A I I I	TTCTA TCTAAA CAAAA CAAAA CAAAA CAAAA CTGA CTGA CTGA CTGA CTGA CTGA CTGA CTGA CTGA CTGA CAAAA CTGAC CAAAA CTGAC CAAAA CTGAC CAAAA CTGAC CAAAA CTGAC CAAAA CTGAC CAAAA CTGAC CAAAA CTGAC CAAAA CTGAC CAAAA CTGAC CAAAA CTGAC CAAAA CTGAC CAAAA CTGAC CAAAA CTGAC CAAAA CTGAC CAAAA CTGAC CTGAC CAAAA CTGAC CTGAC CTGAC CTGAC CAAAA CTGAC CTGA	ACG2 R CTG ² CGG2 CGG2 E IGA0 E CGA2 E T	AGG G IGAA FIN FIN GGAA D GGAA D ACAA D ACAA D D ACAA D L	IGAI D ATCA S MAC CTAI Q CCAG Q CCGAA E CCGAA CCGAA CCGAA CCGAA CCGAA CCGAA	TACC T UTGG UM CTACC TACC Y CCCT P CCCT P CCCT A CCCT CCCT CCCT CCC	AAC. N CAG Q Nain CAGC. S CACG. D CACG. D N O O TTA	AAC N CGA AGT S AGT I NO(CCGG P TCA S ^C ATT	CTG L TGT GAA E CAA Q O) CCG P CCA P HO(CCT P	CGA' R GAG CH TGC CTT. L CTT. L ACA ACA C CT P O) GAA	TGC C GTG V O(O CAT H ACT T GAG GAG CCT P CH TAA	L ACGG T O CTCZ A CTCZ A TCAG S O O CCCCZ P O O O CCCZ P O O O CCCZ	E GGA GGA AGG AGG BAT D GCA ACA T ACA T)^	1188 382 1296 418 1404 454 1512 490 1620 526 1728 562 1836 595
E G AGTGG S G CCAAA P	V GGTC V CACT T CHO (TGCC P TATCC I TATCC I TATCC	Y GTGG W CACA T (N) CTGC C GGTA D CCAA	GGC G S SCI GTC V TTA GCC A S CCA CCA CCA	Y GATC D R 4 IGTG C CATG H IATT Y S CCTG P GCTA A	GCAT G GCT G ACC D CAT S CTG A GAA R	R TTAC L GTF C CHC TTCC L TTCC L S CAC D L ATC N	D CCCCF P NATCF I D CAGC Q CCCT S CCCT S S CCCT S S CCCT Y Y	I ACC T AAT N CTA F CG S ICG S ICG S IAC Y S AAT	L TGT C C C C C C C C C C C C C C C C C C	R GAA E ACA T C d GAC D S/ TTC F P r ACA T	GAT D GCT A TCA S C GCA GCA GAG GAG GAG T/P ATT I I C ATT I I C C T/P	GAG E GTT V TGC GAG E TCC S re GAG GAG	AGG. R GTT V CTG. L O) AACC CTG. CTC. L CTG. CTG. L CTG. L CTG. L CTG. L	ATC I CAT H ACG GAC D GAC I I Feg GTC V N ACC	TTC F CAAA Q TCA S C TAT G G I O A TG M C C T T T T T T C S C T A T C S C S C S C S C S C S C S C S C S C	TAC Y GTG GGG G HO (GGA GCT A GAT D TTT	TGT/ C TCA(S ^CI CCG/ D TGC/ CAG/ C GCT(A AACC	AAAG K CCGT P HO(C ACTG T ATTA AGCG S CCAG P ICCAA S KO(N		GTT. G CGA A GCG GGG G G CTA P P A TG D TTT F	ATGT Y V GCGT S <u>V</u> GAAC G <u>T</u> doma CTG2 CTG2 CTG2 T <u>E</u> CTG2 CTG2 CTG2 CTG2 CTG2 CTG2 CTG2 CTG2	TCT. TCT. TCT. TCT. TCT. TCAAAA TN TCAAAA TTGA TCAAA TCAAAA TCAAAA TCAAAA TCAAAA TCAAAA TCAAAAAAAAAA	ACG2 R CTG ² CGG2 G IGA0 E ICC2 P CGA2 E AAAC2 T	G G FIN AGAG D GGGAG D ACAG D ACAG D AAGAG D	IGAT D ATCA S CTAT CTAT CCAG Q CCAA C CCAA C CCAA C CCAA C CCAA C CCAA C CCAA C CCAA C CCAA C CCAA C CCAA C CCAA C CCAA C CCAA C CCAA C C CCAA C	TACC T UTGG UM CTACC Y CCCT P CCCT P CCCT A CCCT CCCT CCCT CCC	AAC. N CAG Q Aain CAGC. S CACG. C C C C C C C C C C C C C C C C C C	AAC N CGA AGT AGT ATA I CCG P TCA S ^C ATT I	CTG L TGT C GAA C CAA Q O CCG P CCG P HO (CCT P	CGA' R GAG CH TGC CTT. L CTT. L ACA ACA CT P O) GAA	TGC C GTG O(O CAT H ACT <u>F</u> GAG E E HO(CCT P CH TAA ST	TTG(L ACG(T T CTC2 A GCTC2 A TCA(S O) CCC2 P O (O) GCTC OP	E GGA GGA AGG AGG BAT D GAT ACA T ACA T)^	1188 382 1296 418 1404 454 1512 490 1620 526 1728 562 1836 595
E G AGTGG S G CCAAA P GCTTA AY GAGAG ES GCCGA A D ACCCG TR CHO	GGTC V GGTC V CACT TGCC TGCC TGCC TGCC TACT TACT TACT	Y GTGG W CACA T (N) CTGC C GGTA D CCAA C CCAA	GGC G G TCG(S SC) GTC ^T V TTA(L GCC ^T A TCTC S CCA(P O (O)	Y GATC J A A CATG CATG CATG L L CATG P GCTA A J A A J A J A J A J A J A J A J A A J A J A A A A A A A A A A A A A	GCT G G G G G A CCT G CAT S CTG A GAA R CTT	R TAC L CTCC CTCC CTCC L CTCC S CAJ S ACJ D ATC N	D CCCCF P ATCF I CCCT S CCCT Y CCCT Y	I I I I I I I I I I I I I I	TGTT C C C C C C C C C C C C C C C C C C	R GAA E ACA T ACC I GAC C d GAC C C d GAC F F F T T C A A T	GATI D GCT A TCA C GCA C GCA C GCA C GAG G A T T P C T T P T T T T T T T T T T T T T	GAG E GTT V TGC C HO(GAG E TCC S C GAG GAG GAG GAG	AGG. R GTT V CTG. C	ATC I CAT H ACG GAC D GAC V reg GTC V n ACC V n ACC T	TTC F CAA Q TCA C C TAT G G G G G G G G G G G C C T T T T T T	TAC Y GTG GGG GGG GGA GCT A GGT D TTT F On	TGT2 C TCA(S CCG2 CCG2 CCG2 CCG2 CCG2 CCG2 CCG2 CC	AAAG K CCGT P HO (C ACTG T ATTA AGCG S CCAG P ICAA S CCAG CCAG CCAG CCAG CCAG CCAG CCAG	AGGG E (C C TTCTC V (V TTCC V V C TTCC C V C TTCC C C T C T	GTT. G CGA A GCG G GGG G CTA P P A TG D TTT F F	ATGT Y V GCGT S V GAAC G 1 GAAC CTGZ A CTGZ T E CTGZ T E CTGZ A AAAZ E N TAGZ	TCT. TCT. TCT. TCT. TCT. TCAAAA TN TCAAAA TTGA TCAAA TCAAAA TCAAAA TCAAAA TCAAAA TCAAAA TCAAAAAA TCAAAAAA TCAAAAAA TCAAAAAA TCAAAAAA TCAAAAAA TCAAAAAA TCAAAAAA TCAAAAAAA TCAAAAAAAAA TCAAAAAAAAA TCAAAAAAAAAA	ACGA R CTG CGGA CGGA E CGGA E CGGA E CGGA C CGA C CGA C CGA C CGA C CGA C CGA C CGA C CGA C CGA C C C C	G G FIN AGAG D GGGAG D ACAG D ACAG D ACAG C D	IGAI D ATCA S CTAI CTAI CCAG Q CCGAA CCGAA CCGAA CGGAA CGGAA CCGAA	TACC T ATGG M dom TACC TACC Y AGCCT P AGCA A CAAA K CHO (GGCT A	AAC. N CAG Q Aain AGC. S C C C C C C C C C C C C C C C C C C	AAC N CGA AGT AGT ATA I CCG CCG P TCA S ^C ATT	CTG L TGT C GAA C CAA Q O CCG P CCG P HO (CCT P	CGA R GAG CH TGC CTT. L CTT. L CTT. C CTT. P O) GAA E	TGC C GTG O(O CAT HO CAT ACT E HO CAT CA ST	TTG(L ACG(T TCA(A TCA(S O) CCCCA P O (O) GCT(OP	E E E E GGA AGG \mathbb{R} E C AGG \mathbb{R} E C AGG \mathbb{R} E C AGG \mathbb{R} C C C C C C C C C C C C C	1188 382 1296 418 1404 454 1512 490 1620 526 1728 562 1836 595
E G AGTGG S G CCAAA PN GCTTA AY GAGAG E S GCCGA A D ACCCG GATCT	GGTC GGTC V CACT TCHO (TGCC P TACT I TACT I TACT I CHO (O) C	TACAA T CACAA T CTGC CTGC CGAT CCAAA CCAAA	GGC G S SCI GTC' V TTA(L GCC' S CCA(P O (O CAC)	ATTC	GCT P GCT G ACC D CAT S CTG A GAA R CTT	R TAC L CTCC CTCC CTCC CTCC CTCC CTCC CTC	D CCCF P TCF TCF CCCF CCCFF	ACC T AAT CTA F CCG S C S C	TGT C C C C C C C C C C C C C C C C C C	R GAA E ACA T C d GACC D S/ T T C A C A C A T T T T T T T T T T T T	GAT D GCT A TCA C GCA GCA GAG GAG C GAG S T T T T T T T T T T T T T	GAG E GTT V TGC C HO(GAG E T TCC GAG GAG GAG GAG T TTCC	AGG. R GTTY V CTG. L CTG. CTC. L CTG. L CTG. L CTG. L CTG. L CTG.	ATC I CAT H ACG GAC D GAC D ATT I Feg GTC V N ACC T T T T T	TTC F CAA Q TCA C C TAT G G G G G G G G G G G C C T T T T T T	TAC Y GGG G GGG G GGA GGA T TTT F O O A A TTT	TGT2 C TCA(S CCG2 P O) TGC2 CAG2 Q CAG2 Q CAG2 CAG2 CAG2 CAG2 CAG2	AAAG K CCGT P HO(C ACTG J ATTA AGCG S CCAG P ICAA S HO(N S CAA	AGG(E (GTG(C)) TTT(V 1 FIM2 ATG(N (V 1) STTC(V 1 STTC(A 1 CTT(A 1) AAA((CTA)	GTT. G G G G G G G G G G G G G G G G G G	ATGT Y V GCGT SV GAAC G I doma CTCC AE CTGZ TE CTGZ TE NGZ	TCT CGCGG CGCGG CGCG CGCG CGCG CGCG CGCG CGCG CGCG CGCG CGCG CGCG CGCG CGCG CGCG CGCG CGCC CGCG	ACGA R CTG CCGGA CCGGA CCGGA E CCGAA E CCGAA E AAACZ	AGG G FIRAGA D G G G G G G G G G G G G G G G G G	IGAI D ATCA S AAC CTAI Y CCAC Q CCAC Q CCAA CCAC CCAC CCAC CCA	TACC T TGG M dom TACC TACC TACC TACC CACC P CACC TACC CACC TACC CACC TACC T	AAC. N CAG Q Aain AGC. S C C C C C C C C C C C C C C C C C C	AACC N CGA AGT S AGT I S ATA CCCG CCG CCG CCG TCA S C TTA	CTG L TGT GAA E CAA Q O) CCG P HO (CCT P GTT	CGA R GAG CH TGC CTT. L CTT. L CTT. C C T P O) GAA E ACT	TGC C GTG O(O CAT H ACT T GAG CCT P CH TAA ST CAA	ACGC T T CTCZ CTCZ CTCZ CTCZ CTCZ CCCZ P CCCCZ P O (O) GCTC OP GAGT	E E E E E AGG AGG AGG AGG AGG	1188 382 1296 418 1404 454 1512 490 1620 526 1728 562 1836 595
E G AGTGG S G CCAAA P N GCTTA A S GCCGA A D ACCCG T R CHO GATCT ATATT	GGTC V CACT TCCHO TCCC P TATCC I TATCC I TATCC I GAAC GTCC		GGC G SCI GTC GTC TTA GCC A TTA GCC S CCA S CCA CCA CCA CCA CCA CCA CCA C	Y SATC D A GGTA C CATG H IATT Y GCTA A A AATA TTG C C C C C C C C C C C C C	GGA G CCT P G CCT G CCT G CCT G CCT G CCT G CCT CCT	R TAC L GTA C CHC C TCC L C CHC C TCC L S C C C C C C C C C C C C C C C C C	D D CCCF P D TCF D TCF D D D D D D D D D D D D D D D D D D D	ACC T AT T AT T T AT T T C G S T T C G A T T C G A T T	L TGT C C GAT. D GAT. TTT F TCT S /T/ GAA E HO(I TAT. AAC	R GAA E ACA T ATC GAC GAC D S/ TTC F F F ACA T N N AAT	GATI D GCTI A TCA C: GCA GCA Oma GAG. E T/P A TT/P I A CA T TATT TCT.	GAG E GTT V TGC GAG GAG GAG GAG E C C F C C C C C C C C C C C C C C C C	AGG. R GTT' V CTG. L O) AACC CTC. L CTC. L CTG GIO TTA L L TTA TTC AATT' CACC	ATC I CAT H ACG GAC D GAC D ATT I Feg GTC V N ACC T T T GTC T	TTC' F Q TCA' <u>S</u> C' TAT' <u>S</u> GGT' <u>G</u> ion ATG M CCTT P egi TTT.	TAC Y GGG GGG GGG GGA GGA TTT F GGG GGA TTT F GGG	TGT2 C TCA(S ^CC P CCC P P TGC2 C P C C C C C C C C C C C C C C C C C	AAAG K CCGT P HO(C ACTG T ATTA AGCC S CCAG P ICAA S HO(N S GCAG	AGG0 E (GTG0 C) TTCT(V TTCT(V TTCT(V TTCT(V TTCT(A TCTC(V AAG0 CAA	GTT. G G G G G G G G G G G G G G G G G G	ATGT Y V GCGT S V GAAC G I doma CTCC A E CTCA A E CTCA CTCA CTCA CTCA CTCA CTCA CTCA CTC	TCT CGCGG CAAAA CAAAA CAAAA CAAAA CAAAA CAAAA CAAAA CAAAA CAAAA CAAAA CAAAA CAAAA CAAAA CAAAAAA CAAAAA CAAAAAA CAAAAAA CAAAAAA CAAAAAA CAAAAAA CAAAAAA CAAAAAA CAAAAAA CAAAAAA CAAAAAAA CAAAAAA CAAAAAA CAAAAAA CAAAAAA CAAAAAA CAAAAAAA CAAAAAA CAAAAAAA CAAAAAAA CAAAAAAAA	ACG2 R CTG ⁷ CCGG2 G G IGA0 E E CGA2 E AAC2 T AATTC	AGG G FIRAGA D GGA D GGA CA D GGA ACA D D ACA D D ACA CA D C ACA CA CA CA CA CA CA CA CA CA CA CA	IGAT D ATCA S CTAT CCAG Q CGAA CCAG CGAA CCACT T CGGAG CGGAG CATAA	TACC T ATGG M dom TACC TACC TACC CAAAA K CAAAA K CHO (GCCI A CHO (GCCI A CHO (GCCI A CHO (CHO (CH	AAC. N CAGG Q D AAGC. S ACCG. D C C C C D N O O V TTA S C C C C C C C C C C C C C C C C C C	AACC N CGA AGT S AGT ATA I CCCG CCCG CCCG CCCG CCCG ATT TCA TCA	CTG L TGT GAA E CAA Q O) CCG P HO (CCT P GTT TTC	CGA' R GAG CH TGC CTT. L CTT. L CCT CCT P O GAA E ACT C	TGC C GTG O(O CAT H ACT T GAG CCT P CH TAA ST CAA	TTG(L L ACGC T CTCZ L CTCZ A CCTCZ A CCCZ P O (O) CCCZ P O (O) GCTC OP GAGT	E GGA G AGG R SAT D GCA ACA T)^ GTT TCA AGG	1188 382 1296 418 1404 454 1512 490 1620 526 1728 562 1836 595 1944 2052
E G AGTGG S G CCAAA P	GGTC V GGTC V CHO(TGCC A TCCC P TATC I TATC I TATC I GAAC GTCC GAAC	TACA W TACA T (N) TTGC CGAT CCAA CCAA CCAA CCAA CCAA CCAA	GGC G S SCI GTC V TTA GCC A TCT G CCA C P O (O CAC TTA C CA C	GCTA GCTA GCTG CATG. H IATT Y GCTG AATA TTTG. FAGTA	GGA G CCT P G CCT G A CTT G A CTT G TA	R TAC C GTA C C C C C C C C C C C C C C C C C C C	D CCCF P NTCF I CCCT S S CCTF CCTF CCTF CCTF	ACC T AAT N CTA F CG S CTA F CG S CA S AAT S GG G	TGT C C GAT TTT TTT TTT S /T/ GAA E 	R GAA E ACA T ATC GAC GAC D S/ TTC F F F ACA T N N AAT TAT	GATI D GCTI A TCA S C GCA Oma GAG. E GCA Oma GAG. I T/P A TT T T T T T T T T T T T T T T T T T	GAG E GTT V TGC GAG GAG GAG C S C C F C C C C C C C C C C C C C C C	AGG, R GTT V CTG, L O) AAC(CTC, L CTC, L CTG, CTG, L TTA, TTA, TTA, CACC, TTG, CACC, TTG, CACC, CTC, C	ATC I CAT H ACG GAC D GAC T GAC T GTC N A ACC T T GTC T GTC T T GTT T GTT T GTT T	TTC' F CAAA Q TCA S C' TAT GGT' G G TAT GGT' S C' TAT GGT' TAT GGT' TAT GGT' TAT GGT' TAT GGT' TAT GGT' TAT GGT' TAT	TAC Y GGG GGG GGA GGA GGA TTT F ON AAT GGGA	TGT2 C S ^CC P O TGC2 C C C C C C C C C C C C C C C C C C	AAAG K CCGT P HO (C ACTG T ATTA AGCG S CCAG P TCAA S HO (N GCAA IGGG IATT	AGG(E (C) TTCT(V (V) FIM(ATG(N (V) SCTG) A) AAAA(CAAA)	GTT. G G G G G G G G G G G G G G G G G G	ATGT Y V GCGT GAAC GAAC GAAC GAAC GAAC CTGA TCA CTGA TA	TCT. CGCGG CAAAA CAAAA CAAAA CAAAA CAAAA CAAAA CAAAA CAAAA CAAAA CAAAA CAAAA CAAAAA CAAAAA CAAAAA CAAAAAAA CAAAAAAA CAAAAAAA CAAAAAAA CAAAAAAA CAAAAAAA CAAAAAAA CAAAAAAA CAAAAAAA CAAAAAAAA	ACG2 R CTG ⁷ CCGG2 G G IGA0 E IGC02 E CGA2 E AAAC2 AAAC2 AAAC2	AGG G FIN AGAG D GGAG D GGAG D ACAG D ACAG D ACAG C D ACAG C GTAG	IGAI D ATCA S CTAI CTAI CCAG Q CCAG Q CCAG Q CCAG C CCAG Q CCAG C CCAG Q CCAG C CCAG Q CCAG C CCAG Q CCAG C CCAG Q CCAG C CCAG Q CCAG C CCAG Q CCAG C CCAG Q CCAG C CCAG Q CCAG C CCAG Q CCAG C CCAG Q CCAG C CCAG Q CCAG C CCAG Q CCAG C CCAG Q CCAG C CCAG Q CCAG C CCAG Q CCAG C CCAG Q CCAG C CCAG Q CCAG C CCAG C Q CCAG C CCAG C CCAG C Q CCAG C CCAG C CCAG C CCAG C Q CCAG C C CCAG C C CCAG C C CCAG C C CCAG C C CCAG C C C C		AAC. N CAGG CAG C C C C C C C C C C C C C C C	AACC N CGA AGT S ATA ATA CCCG P TCA S C C ATT I TTA GATA	CTG L TGT GAA CAA Q O) CCG P CCA P CCA P CCA P GTT TTC TTC	CGA' R GAG CH TGC CTT. L CTT. L CTT. C CCT P O GAA E ACT CACT	TGC C GTG O(O CAT H ACT T GAG CCT P CH TAA ST CAA ST CAA TGT	ACG(T TCA(S O) CTC2/ A TCA(S O) CCC2/ P O(O) GCT(OP GAG TTA/ AAT	E GGA G AGG C A AGG C A AGG C A AGG C A A A A A A A A A A A A A	1188 382 1296 418 1404 454 1512 490 1620 526 1728 562 1836 595 1944 2052 2160
E G AGTGG S G CCAAA P GCTTA AY GAGAG ES GCCGA A D ACCCG GATCT ATATT AAAAT GATAA	GGTC V GGTC V CACT T CHO (T T CHO (T T C C T T C C C T T C C T T C C T C T C	TAGE W TACA TACA (N) TGC CGTA CGATA CGATA CCAA TTT CCGI CCGI	GGC GTC S SCI GTC TTA GCC A TCT S CCA C P O (O) CA C CA C TTA C CA C	ATT BATC D R 4 IGTG C CATG H IATT Y CCTG P GCTA A ATA TTG IAGT CTCT	GCAT G GCT G ACC D CAT S CTG A GAA R CTT GTA GTA	R TTAC L CGTZ CTCC TTCC L CCAI S CAI S CCAI S CAI S C CCAI S CCAI S CCAI S CCAI S CCAI S CCAI S CCAI S CCAI S CCAI S CCAI S CCAI S CCAI S CCAI S CCAI S CCAI S C CCAI S CCAI S CCAI S C CCAI S C CCAI S C C C C C C C C C C C C C C C C C C	D CCCF P ATCF I CAGC Q CAGC Q CCTF CCTF CCTF CCTF CCTF CCTF CCTF C	ACC T AAT N CTA F CG S CTA F CG S CA S C S C S C S C S C S C S C S C S	TGT C C C C C C C C C C C C C C C C C C	R GAA E ACA T ACC GAC GAC F F F F T ACA T AACA T AACA T AACA T AACA T AACA	GAT D GCT A TCA C GCA GCA GCA GCA GCA C GCA T/P I I C A TT T T T T T T T T T T T T T T T T	GAG E GTT V TGC C HO(GAG J I ACT TCC S E GAG GAG GAG E TTC TTC ATTA AACA	AGG. R GTTT V CTG. L CTG. CTG. CTG. L CTG. CTG. L TTA. CTG.	ATC I CAT H ACG GAC D GAC D ATT I GAC T T GTC T T GTT T GAA	TTC' F CAAA Q TCA TCA TTCA GGT' ion ATG M CCT' F egi TTT. TCA ATG GTC'	TAC Y GGG G GGG G GGA GGA T T T T GGA GAT D T T T T T T T T GGG A CA	TGT2 C TCA(S CCG2 CCG2 C O) TGC2 CAG2 Q CAG2 AAC' CAG2 CAG2 CAG2 CAG2 CAG2 CAG2 CAG2 CAG2	AAAG K CCGT P HO (C ACTG T ATTA AGCG S CCAG P ICAA S CCAG C CAG C CAG C CAG C CAG C CAG C CAG C CAG C CAG C CAG C CAG C C CAG C C C C	AGG(E (C) () () () () () () () () () () () () ()	GTT. G G G G G G G G G G G G G C TA C TA C T	ATGT Y V GCGT S V GAAC G I doma CTCC G I doma CTCC A AAAAA T E N TAGA L I AGGC TGCC TCCC AAAT	TCT: TCT:	ACGA R CTG ⁷ CGGA CGGA G IGA E IGA E CGAA E CGAA CA A CGAA CA T CGAA CA CGAA CA CGAA CA CGAA CA CGAA CA CCGAA CA CGGAA CA CGGAA C CGGAA C CGGAA C C C C	AGG G G FIN AGAC D GGAC D ACAC D ACAC D ACAC C D GGAC C D GGAC C C C GGAC C C C C C C C C C C C C	IGAI D ATCA AAC CTAI CTAI CCAG Q CGAA E CCAC T T CCAG Q CGAA E CACT T CACT T CACT CACT CACT CACT C	ACC T ATGG M M TAC TAC CCT P CCT P CCT P CCT CCT A CCT A CTAC CTAC	AAC. N CAG Q AGC. S CACG. S CACG. D CACG. D CACG. D C C C C C C C C C C C C C C C C C C	AACC N CGA AGT S ATA ATA CCCG CCG CCG CCG CCG CCG CCG CCG CCG TCA TCA	CTG L TGT GAA CAA Q O) CCG P CCA P CCA P CCA P GTT TTC TTC	CGA' R GAGG CH TGC CTT. CTT. CTT. CTT. CTT. P O) GAA' E ACT CAC' AAT.	TGC C GTG O(O CAT H ACT T GAG E CAA ST CAA ST CAA TGT AGT	ACGC T CTC2 CTC2 CTC2 CTC2 CCCC A TCAC S CCCC P O CCCC P O CCCC P O CCCC TTAI AATT TTGC	E GGA G AGG C GAT D GGCA A C GGCA T C GGCA T C GGCA T C GGCA T C GGCA C C C C C C C C C C C C C	11183 382 1296 418 1404 454 1512 490 1620 526 1728 562 1836 595 1944 2052 2160 2268
E G AGTGG S G CCAAA P GCTTA A GAGAG ES GCCGA A D ACCCG TR CHO GATCT ATATT AAAAT GATAA GGAGT	GGTC V CACT T CHO (T TCCC T TATC I TATC I TATC I GAAC GAAT CCAAT	Y TGG W CACA CTGC CGAT CCAA CCCAA CCAAA CCAAA CCAAA CCAAA CCAAA CCAAA CCAAA CCAAA CCAAA CCAAA CCAAA CCAAA CCAAA CCAAA CCAAA CCAAA CCAAA CCAAA CCAAAA CCAAAA CCAAAA CCAAAA CCAAAAA CCAAAAA CCAAAAAA CCAAAAAAAA	GGC GTCG S SCI GTC TTTA GCC A TCT S CCCA C CAC TTAA CAT AAT AAG	ATC Y BATC D R 4 FGTG C C C C C C C C C C C C C	GGAA G GCT G ACC D CAT S CTG A GAA R CTT ATT GTA GTT ATT	R TAC L CGTZ CHC TCC TCC TCC TCC TCC TCC TCC TCC TCC	D CCCCF P ATCF L D CAGC O CCTS S GTCF V CCTS CCTS CCTS CCTS CCTS CCTS CCTS CC	I I I I I I I I I I I I I I I I I I I	TGT C C C C C C C C C C C C C C C C C C	R GAA E ACA T ATC C d GAC J C d GAC F F F F R A A T T T T T T T T T T T A A T T A A T A A T A A T T A T	GATI D GCTI A TCA S C GCA GCA GCA GAG GAG GAG C A TTP I I I C A TTTT T TTT T C T A TG C A C C A	GAG E GTT V TGC C HO(C GAG GAG GAG GAG GAG GAG TTCC S re GAG GAG TTCC S TTCC S TTCC C ATT TTCC TTCC C GAG TTCC C TTCC C C HO(C) C C C HO(C) C C C C C C C C C C C C C C C C C C	AGG. R GTTT V CTG. L CTG. CTG. CTG. L TTG. GAC TGA	ATC I CAT H ACGG GAC D GAC V n ACC V n ACC V T TGT TGT TGT TGT ATT GAA TCA	TTC F CAA Q TCA TAC TAT GGT TAT GGT M CCT Egi TTT. TCA ATG. GTC AAC	TAC Y GGG GGG GGG GGA GGA TTTT F ON AAT GGG ACA TGT	TGT/ C TCA(S CCGJ D TGC/ CAGJ CAGG CAGG CAGG CAGG CAGG CAGG CAGG	AAAG K CCGT P HO (C ACTG T ATTA AGCG P CCAG P ICAA S S CCAG P ICAA S IGGG IATT CGCT GACT	AGG E () GGTG () () GTCT() V () GCTT() T () GCTGJ () () GCTT() T ()	GTT. GCGA GCGG GCGG GGGG G CTA P P CTA F F GAG GCG	ATGT Y V GCGT S V GAAC G 1 doma CTCC G 1 doma CTCC A AAAA E N TAGA E N TAGA E N TAGA CHC AAAAA E N TAGA CHC CHC AAAAA E N TAGA CHC CHC CHC CHC CHC CHC CHC CHC CHC CH	TCT. TCT. TCT. TCT. TCAAAA TN TCAAAA TTGA TCAAA TCAAAA TCAAAA TCAAAA TAAAAAAAAAA	ACGA R CTG' CGGA CGGA CGGA IGA E IGA E CGAA CGAA TTC AATTC AATTC AGAG GGG	AGG G IGAA FIN FIN AGA0 D GGA0 D ACA0 D ACA0 D ACA0 D ACA0 D GGA ACA0 C ACA0 C ACA0 C ACA0 C ACA0 C ACA0 C ACA0 C ACA0 C ACA0 C ACA0 C D C C C C C C C C C C C C C C C C C	IGAI D ATCA ATCA IS. CTAI CTAI CTAI CTAI CTAI CTAI CTAI CTAI		AAC. N CAG Q Cacina CAGC. S CACG. S CACG. D CACG. D CACT. N O) TTA. C GGT TTCC. CGTA. CCTT. ACA	AAC N CGA AGT AGT ATA I CCG CCG CCG CCG S CCG S CCG TCA S CTTA GATT TTA GATT TGA	CTG L TGT GAA CAA CCAA CCAA CCAA CCAA P CCCA P CCCA P CCCA P GTT TTC AAT TTT GGA	CGA' R GAG CH TGC CTT. CTT. CTT. CTT. CCT P O) GAA E ACT CAC AACT CAC AACT	TGC C GTG QTG CAT H ACT T GAG CAT T CAT ST CAA TGT CCA TGT AGG	ACG(ACG(T CTC2/ CTC2/ CCTC/ A CCTC/ A CCTC/ A CCTC/ A CCCC/ P O (O) CCCC/ P O (O) CCCC/ P O (O) CCCC/ CCC/ CCC/ A CCCC/ CCC/	E GGA G AGG R GAT D GCA ACA T CA GTT ICA AAGG ICA	11183 382 1296 418 1404 454 1512 490 1620 526 1728 562 1836 595 1944 2052 22160 2268 2376
E G AGTGG S G CCAAA P N GCTTA A S GCCGA A D ACCCG GATCT AAAAT GATAA GGAGT AGGAGT	GGTC V CACT T TCCO TGCC TGCC TATCC I TATCC I TATCC GACT CAAT TCCA	Y TACA W TACA W TACA W TTGC GGTA CGAT CGAA TCAA GGTA TCAA CGAA TTTT CGAA TCAA	GGC GTC S SCI GTC TTTA(L GCC A TCT S CCCA CAC TAA CAA CAA CAA CAA CAA CAA CA	ATC P AATC D R 4 IGTG CATG CATG IATT Y SCCTG P AATA AATA AATA TTTG ITTG ITTG ITCT ICCC JACG	GGA G GCT G G ACC D CAT S CTG A CTT ATT GTA GTT ATT TTA	TAC TAC CTAC CTAC CTCCC CTCC CTCC CTCC CTCC CTCC CTCC CTCC CTCC CTCC C	D D CCCF P ATCF I D CAGC O CCT S CCTF S CCTF S CCTF CCTF CCTF CCTF	I I I I I I I I I I I I I I I I I I I	TGT C C C C C C C C C C C C C C C C C C	R GAA E ACA T ATC GAC J GAC T T T C d GAC F F F R ACA T T T T T T T T T T T T T T T T T T	GAT D GCT A TCA S C GCA C GCA GGAG E T/P ATT I ich ACA T T T T T T T T T T T T T T T T T T	GAG E GTT V TGC GAG GAG E TCC S GAG GAG E TTCC S C GAG TTCC S TTCC C ATT AAAA	AGG. R GTT V CTG. C	ATC I CAT H ACG GAC D ATT GAC I I Teg GTC V N ACC I I TGT ATT GAA TCA	TTC' F CAAA Q TCA CT TAT' GGT' G GGT' G TTT TCA ATG GTC' AACC	TAC Y GGG G G GGG G GGT A GGT A GGT TTT T TTT TT	TGT/ C TCA(S CCG2 P O) TGC2 CAG2 Q CAG2 A A CCAG2 GGCT CAG2 GGCT A A CCGA	AAAG K CCGT P HO(CCG T ATTA AGCG S CCAG P TCAA S CCAG P TCAA S CCAG C CCAG P TCAA S CCAG C TCAA S CCAG TCAA S CCAG TCAA S CCAG T CCAG CCAG	AGG E (GGG) GGG GGGG GGG GGGG GGG GGG GGG GGGG GGG GGGG GGG GGG GGG GGG GGG GGG GGGG GGG GGG GGG GGG GGGG GGGG GGGG GGGGG GGGGG GGGGG GGGGGG GGGGG GGGGG <td>GTT. G G G G G G G G G G G G G G G G G G</td> <td>ATGT Y V GCGT S V GAAC G I doma CTCC A A CTCA CTCA CTCA CTCA CTCA CTCA</td> <td>TCT. TCT. TCT. TCT. TCCCC. TCCCC. TCCCC. TCCCC. TCCCC. TCCC. TCCCC. TCCCCC. TCCCC. TCCCC. TCCCC. TCCCC. T</td> <td>ACG2 R CTG? CGG2 CGG2 CGG2 E IGA3 E CGA2 E CGA2 E CGA2 E AAC22 E AAC22 E AAC22 E CGA2 CGA2 E CGA2 E CGA2 E CGA2 E CGG2 E C CGG2 E CGG2 E C CGG2 E C CGG2 E C CGG2 E C CGG2 E C C CGG2 E C C CGG2 E C C C C C C C C C C C C C C C C C C</td> <td>AGG G IGAA E FIL AGAC D AGACAC D ACAC D ACAC D ACAC D ACAC C C C C</td> <td>IGAI D ATCA S MAC TTAI CTAI CCAG Q CGAA CCAG CGAA T CCACI T CCACI CCACI CCACI CCACI CCACI CCACI CCACI CCACI</td> <td></td> <td>AAC. N CAG Q Cain AGC. S C C C C C C C C C C C C C C C C C C</td> <td>AAC N CGA AGT AAGT AATA I CCCG CCCG CCCG CCCG S C ATT I CCCG S C C ATT I CCCG S C C ATT AATT</td> <td>CTG L TGT GAA CAA O CCG P CCG P CCG P CCG P CCG CCT TTC CCT TTT TTC TTT TTT GGA TTT</td> <td>CGA' R GAG CH TGC CTT. L CTT. L CTT. C C C C C C C C C C C C C C C C C</td> <td>TGC C GTG C CAT H ACT T GAG E CAT T CCT P CH TAA ST CAA TGT TCC AGG ATT</td> <td>TTGG L ACGG T TCZ CTCZ CCCZ P O(O) GCTC OP GAG' TTAZ AAT' TTG' GGG' AGGG</td> <td>E GGA G AGG AGG AGG AGG AGG AGG</td> <td>11183 382 1296 418 1404 454 1512 490 1620 526 1728 562 1836 595 1944 2052 2160 2268 2376 2484</td>	GTT. G G G G G G G G G G G G G G G G G G	ATGT Y V GCGT S V GAAC G I doma CTCC A A CTCA CTCA CTCA CTCA CTCA CTCA	TCT. TCT. TCT. TCT. TCCCC. TCCCC. TCCCC. TCCCC. TCCCC. TCCC. TCCCC. TCCCCC. TCCCC. TCCCC. TCCCC. TCCCC. T	ACG2 R CTG? CGG2 CGG2 CGG2 E IGA3 E CGA2 E CGA2 E CGA2 E AAC22 E AAC22 E AAC22 E CGA2 CGA2 E CGA2 E CGA2 E CGA2 E CGG2 E C CGG2 E CGG2 E C CGG2 E C CGG2 E C CGG2 E C CGG2 E C C CGG2 E C C CGG2 E C C C C C C C C C C C C C C C C C C	AGG G IGAA E FIL AGAC D AGACAC D ACAC D ACAC D ACAC D ACAC C C C C	IGAI D ATCA S MAC TTAI CTAI CCAG Q CGAA CCAG CGAA T CCACI T CCACI CCACI CCACI CCACI CCACI CCACI CCACI CCACI		AAC. N CAG Q Cain AGC. S C C C C C C C C C C C C C C C C C C	AAC N CGA AGT AAGT AATA I CCCG CCCG CCCG CCCG S C ATT I CCCG S C C ATT I CCCG S C C ATT AATT	CTG L TGT GAA CAA O CCG P CCG P CCG P CCG P CCG CCT TTC CCT TTT TTC TTT TTT GGA TTT	CGA' R GAG CH TGC CTT. L CTT. L CTT. C C C C C C C C C C C C C C C C C	TGC C GTG C CAT H ACT T GAG E CAT T CCT P CH TAA ST CAA TGT TCC AGG ATT	TTGG L ACGG T TCZ CTCZ CCCZ P O(O) GCTC OP GAG' TTAZ AAT' TTG' GGG' AGGG	E GGA G AGG AGG AGG AGG AGG AGG	11183 382 1296 418 1404 454 1512 490 1620 526 1728 562 1836 595 1944 2052 2160 2268 2376 2484
E G AGTGG S G CCAAA P	GGTC V GGTC V CACT TCCO P TACC I I TACC I I I I I I I I I I I I I I I I I I	Y TACA W TACA W TACA W TGC GTA CGAA CGAA TTT CGAA TTT CGAA CGAA CGAA CGAA CGAA	GGC GTC S SCI GTC TTA GCC A TCT S CCA C CA C CA C CA C CA C CA	GCTA GCTA GCTA CATG CATG CATG CATG F AATA CCTG P AATA CTTG CTCT FAGTA CTTCT CTCTT CTCTCT CCCC GACG GATT	GGA G GCT G GCT G A CAT S CTG A GAA GTT ATT GTA GTT ATT TTA GAA	TACC TACC	D CCC2 P L L D CCC3 S S CCC3 Y S TCC4 Y S TCC4 Y S TCC4 Y S TCC4 Y S TCC4 Y S TCC4 S CCC4 Y S CCC4 P L D CCC2 P L D CCC2 P L D CCC2 P L D CCC4 P L D CCC4 P L D CCC4 P L D CCC4 P L D CCC4 P L D CCC4 P L D CCC4 P L D CCC4 P L D CCC4 P L D CCC4 P L D CCC4 P L D CCC4 C N D CCC4 C N D CCC4 C N C CCC4 C N C CCC4 C N C CCC4 C C C C	LAT LAT LAT F C C C C C C C C C C C C C	TGT C C C C C C C C C C C C C C C C C C	R GAA E ACA T ATC GAC D TTC F F P r ACA TAA ACG ACGA CAT	GAT D GCT A TCA C GCA GCA GCA GCA GCA GCA GCA TTP TTT TCT TCT ATG CCAC CCAC	GAG E GTT V TGC GAG GAG E TCC S GAG GAG C TTC TTC C ATT AAAA ATC GAAG	AGG. R GTT V CTG: CTG: CTG: CTC: CTG:	ATC I CAT H ACG GAC D GAC T I GAC T C ATT GAA TCA TCA TCA	TTC F CAAA Q TCAA S C TTAT G T G T T C T T T T T T T T	TAC Y GGG G G GGA GGA T TTT F ON AAT GGG ACA TGT TTC	TGT/ C TCAC S ^CI CCCG/ D TGC/ CAG/ CCAG/ GGCT CAG/ GGCT CAG/ GGCT CAG/ GGCT CCAG/ CCAG/ GTGT	AAAG K CCGT P HO(CCG T ATTA AGCG S CCAG P ICCAA S CCAG IGGG IATTI GCGA IGGG IATTI GCGATI GGA ITGAA	AGG(E () GTG(C () GTCT(V () TTCT(V () TTCT(N () TTCT(N () TTCC(N () GCTT(T () AAAA(AAAA(() GCTT(T) TAAAA(() GCTT(T) TTCA(() TTCC(GAAG(()	GTT. GCGA AC GCG GC GC GC GC GC GC GC GC GC GC AC GC GC AC GC GC AC GC GC GC GC GC GC GC GC GC GC GC GC GC	ATGT Y V GCGT S V GAAC G I GAAC CTCC A I CTGZ CTCZ AAAZ F N TAGZ CTCC AAAZ L I AGGG TCCZ AAAT TCCT	TCT: TCT:	ACGI R CTG ² CGGI CGGI E IGAG E ICCI E CGAI E CGAI E AACI CGAI A ACAG	AGG G IGAA FIN AGAC D GGAC D ACAC D ACAC D ACAC A CACAC CCCC CCCC	IGAI D ATCA S GAC CTAI CCAG Q CGAA CCAG CGAA CCAC CCAC CACT ATGG ATAA AGCI ACTI CCAC	ACC T ATGG M dom TAC TAC CCI P AGCA A CCI CCI CAC TAC CCI CAC TAC CCI CAC TAC CCI CCI CCI CCI CCI CCI CCI CCI CCI C	AAC. N CAG Q CAG C C C C C C C C C C C C C C C C C C	AAC N CGA AGT AAGT ATA I CCCG CCCG CCCG CCCG S C C ATT I GAT TTA AGT AGT ACA	CTG L TGT GAA CAA O O CCG P CCA P CCA P CCA P CCA CCA CCA CCA CCA	CGA' R GAG CH TGC CTT. L CTT. L CTT. P O) GAA' E ACT CAC' AACT CAC' CAC' CAC' CAC' CAC'	TGC C GTG CAT H ACT T GAG <u>E</u> HO (CCT P CHA ST CAA TGT TCC AGG AGT	ACG(ACG(T CTC2) CTC2 CTC2 CCC2 CCC2 P CCCC2 P CCCC2 P CCCC2 P CCCC2 CCCCC2 CCCCCC2 CCCCC2 CCCCC2 CCCCCC2 CCCC2 CCCC2 CCCC2 CCCC2 CCCC2 CCCCC2 CCCCCC	E GGA <u>G</u> AGG <u>A</u> GAGG <u>A</u> <u>A</u> <u>A</u> <u>A</u> <u>A</u> <u>A</u> <u>A</u> <u>A</u>	11183 382 1296 418 1404 454 1512 490 1620 526 1728 562 1836 595 1944 2052 22160 2268 2376 2484 2592
E G AGTGG S G CCAAA P		Y GTGG W TACA T CTGC CTGC CTGC CGAT D CGAT D CGAT CGAT CGAT CGAT CGAT CGAT CGAT CGAT CGAT CGAT	GGC GTC SC GTC TTA GCC CA S CCA CA S CCA CA CA S CCA CA CA S CCA CA CA S	A TO GATC D CATG CATG CATG H IATT Y CCTG P CCTG P CCTG CCG CC	GGA G GCT G GCT G A CAT S CTG A GAA GAA GTT ATT GTA GTT ATT A GAA CGA	R TTAC GTA C GTA C TTCC TTCC TTCC TTCC TT	D CCCF P ATCF J MATCF J MATCF CAG CAG CAG CAG CAG CAG CAG CAG CAG CAG	LCC T LAT LAT F CCG S CAT S CAT S C C S C C S C C C C C C C C C C C C	TGT C C C C C C C C C C C C C C C C C C	R GAA E ACA T ATC C d GAC D TTC F F F F ACA T AACA TAA AACG TAA TAA GGAA CAT CCA	GAT D GCT A TCA S C GCA GCA GAG GAG GAG S T T T T T T T T T T T T T T T T T T	GAG E GTT V TGC GAG GAG F TTC S C GAG GAG GAG TTC C S TTC C S C GAG GAA AAC GAA GCG	AGG. R GTTT: V CTG. L CTG. CTC. L CTG. CTG. GAC. TTG. GACC CCCC ATG TTA.	ATC I CAT H ACG GAC D T GAC V ATT I J GTC V ACC T T GTC ATT GAA TCA TCA TCA	TTC F CAAA Q TCAA S C TTTAT GGT G G G C C TTT CAA S C C TTTT TCA M GGT G G C C TTTT TCA M C C C C C C C C C C C C C	TAC Y GGG GGG GGA GGAT D TTT GGG GAT T TTT TCC TCA	TGT/ C TCA(S CCG2/ P O) TGC2/ CAG2/ CCAG2/ GGTC GGGC GGGC GGGC GGGC	AAAG K CCGI P HO (C ACTG T ATTA AGCG S CCAG P ICAA S CCAG P ICAA S CCAG C P ICAA S CCAG C ICAA C CCAG C P ICAA S C CCAG I C C C C C C C C C C C C C C C C C C	AGGG E (GTGG C C TTCT(V C FIM ATG(ATG(ATTCC V C C C C C C C C C C C C C C C C C	GTT. GGA GCGA GCGA GGG GGG GGG GGG GGG GGG G	ATGT Y V GCGT S V GAAC G 1 doma CTCC A CTCC A CTCC A CTCC A AAA7 E N TAGA E N TAGA E N TAGA CTCC AAA7 AGGC TCCC AAA7	TCT. TCT. TCCC. TCC. TCCC. TCC. TCC. TCC. TCC. TCC. TCC. TCC. TCCC	ACG2 R CTG ² CGG2 CGG3 IGA(E ICC2 E CGG4 CGA4 CGA4 CGA4 CGA4 CGA4 CGG ² CGA4 CGA4 CGA4 CGA4 CGA4 CGA4 CGA4 CGA4	AGG G FIN AGAG D GGAG D GGAG D ACCAG D ACCAG GTA CCAG CCCCG	IGAI D ATCA S AAC CTAI CCAG Q CCAA CCAA CCAA CCAA CCAA CCAA CC		AAC. N CAG Q aain AGC. S AGC. S AGC. C C C C C C C C C C C C C C C C C C	AAC N CGA AGT S ATA HO (CCGG CCG CCG CCG CCG S ^C ATT S ATT TGA ATT AGT AGT AGT AGT AGA AAT	CTG L TGT GAA C CAA Q O CCG P CCCA P CCCA P GTT TTTC GGAT TTT TTTC	CGA' R GAGG CH TGC CTT. CTT. CTT. CTT. CTT. CO O) GAA' E ACT. CACT. TGG. CTT. GTT.	TGC C GTG V O(O CAT H T GAG E HO(C C TAA ST CAA ST CAA ST CAA TGT TCC AGG AGT CCT	ACG(T ACG(T CTC2) CTC2) CCC2 A TCA(A CCC2 P O O O CCC2 P O O O CCC2 P O O O CCC2 TTAM AAT' TTG' CGG' GGG' GGG' AGG(ATG(TTG')	E GGA AGG AGG AGG AGG AGG AGG AG	11183 382 1296 418 1404 454 1512 490 1620 526 1728 562 1836 595 1944 2052 2160 2268 2376 2484 22592 2700
E G AGTGG S G CCAAA P GCTTA AY GAGAG ES GCCGA A D ACCCG GACCG GACCG GACCG GAGGT AGAGT AGAGT		Y GTGG W TACA TACA CTGC CTGC CTGC CTGC CCAA	GGCC ^G SCC SCC TTA(L GCC ^C S CCA(S CCA(S CCA(CAC TAA' CAT' AAA' GGGG GGGC GAC(TTG'	A TOTAL	GGA G CCT G G CAT S CTG A GAA GAA GTT GTA GTA GTA GTA GTA GTA	TAC TAC TAC TAC TAC TAC TAC TAC	D CCCZF P L L CCCZF CCCZF S CCCZF S CCCZF S CCCZF CCCCZF CCCZF CCCZF CCCZF CCCZF CCCZF CCCZF CCCZF CCCZF CCCZF CCCZF CCC	I I I I I I I I I I I I I I	TGT C C C C C C C C C C C C C C C C C C	R GAA E ACA T ATCC C d GAC C C d GAC T T C C d S/ T T T T T T T T T A A T T A A T A A T A A C A C	GAT D GCT A TCA C GCA GCA GCA GCA T/P T T T T T T T T T T T T T T T T T T	GAG E GTT V TGC GAG GAG E TTC S C GAG GAG C GAG C GAG C GAG C GAG C GAG C GAG C GAG C GAG C GAG C GAG C GAG C GAG C C C C	AGG. R GTTT V CTG. L CTG.	ATC I CAT H ACG GAC D ATT I GAC V I I Teg GTC V ATT GAA TTG GAA TTG GAA TTG TCA TTG	TTC' F CAA Q TCA <u>C</u> C TAT <u>C</u> GGT <u>G</u> C TTT GGT C C TTT TCA ATG GTC C C TTT AA C C TTT TCA AAC	TAC Y GGG GGG GGG GGA GGA T GGA GGA T GGA T T T T	TGT/ C TCA(S CCG/ CCG/ TGC/ CAG/ CCGA GCT(CAG/ GGG/ AAT AAG(CGA GGG/ GCA/ CCTC	AAAG K CCGT P HO(C ACTG T ATTA AGCG P CCAG P CCAG P CCAG S CCAG P CCAG C CCAG C CCAG C CCAG C C CCAG C C CCAG C C CCAG C C C C	AGGG E (GTGG C) TTCTO FIMI ATGG CTTTCO CTTCO CTTCO GCT GCT GCT GCT GCT TTCO GCT TTCO GCT TTCO GCT TTCO GCT TTCO GCT GCT GCT GCT GCT CO GCT TTCO GCT TTCO GCT TTCO GCT TTCO GCT TTCO GCT TTCO GCT TTCO GCT TTCO GCT TTCO GCT TTCO GCT TTCO GCT TCO GCT TTCO TTCO	GTT. GGA GGG GGG GGG GGG GGG GGG CTA F F GA GGG GGG GGG GGG GGG GGG GGG GGG	ATGT Y V GCGT S V GAAC G I GOMA CTCC A A CTGZ CTCC A AAAZ E N TAGA E N TAGA E N TAGA C TCCZ AAAT TCCZ AAAT	TCT: TCT: TCT: TCT: TCAAA TAAAA TCAAA TCAAA TCAAA TCAAA TCAAA TCAAA TCAAA TCAAA TCAAA TCAAA TCAAA TCAAA TCAAAA TCAAAA TCAAAAA TCAAAAAA	ACGI R CTG ² CGGJ G IGAQ E IGAQ E CGAJ E CGAJ E CGAJ CGAA C A AACJ CGAJ CGAA CACAC AAAJ	AGG G FIN AGAC D GGAC D ACAC D ACAC D ACAC D ACAC C C GCGZ ACAZ C C C C C C C C C C C C C C C C C	IGAI D ATCA AC CTAI CCAG Q CCAG Q CCAC Q CCAC Q CCAC CCAC Q CCAC Q CCAC CCAC Q CCAC CCAC CCAC Q CCAC CCAC Q CCAC CCAC CCAC CCAC Q CCAC CCAC CCAC Q CCAC CCAC Q CCAC CCAC Q CCAC CCAC Q CCAC CCAC CCAC CCAC Q CCAC CCAC CCAC CCAC CCAC Q CCAC		AAC. N CAG CAG CAG ACG. AAC. CAC. CAC. CAC. CA	AAC N CGA AGT S ATA HO (CCGG CCG CCG CCG CCG CCG ATT S ATT S ATT TTA AATT AATT AATT AAGT AATT AAGT	CTG L TGT GAA CAA Q CCA Q CCA P CCA P CCA P CCCA P CCCA P TTT TTC GTTT TTTG GTTT TTTC AA	CGA' R GAG CH TGC CTT. L CTT. ACA CT C CT C CT C C C C C C C C C C C	TGC C GTG V O(O CAT H T GAG E E HO(C C C T TAA ST CAA ST CAA ST CAA ST CAA ST CAA ST CCAA CAA CAA ST CCAA CCAT C C C C C C C C C C C C C C	L ACG(T T CTC2 CTC2 C CC2 A CCC2 P CCC2 P CCC2 P CCC2 P CCC2 P CCC2 CCC2 P CCC2 CCC2 P CCC2 CCC2 P CCC2 CCCCC2 CCCC	E GGA AGG AGG AGG AGG AGG AGG AG	11183 382 1296 418 1404 454 1512 490 1620 526 1728 562 1836 595 1944 2052 2160 2268 2376 22484 2592 2700 2795

Fig. 3 cDNA sequence of *Sp5013* and the deduced protein sequence of SpCRS (*S. purpuratus* complement related protein, short form). The nucleotide and amino acid sequences were generated as described for *Sp5* (see legend for Fig. 2). The leader, domains, and conserved sites are labeled as in Fig. 2. An almost-perfect Kozak

and SpCRL with SCRs from both complement and non-complement proteins. GenBank was queried using BLASTX with the entire SpCRL and SpCRS protein sequences to identify proteins that contained SCRs that were similar to those in SpCRL and/or SpCRS. This search identified SCRs from both complement and noncomplement proteins (see footnote to Table 3) and all sequence (GCC(A/G)CCATGG) surrounds the start codon and is *underlined* (Kozak 1987). Three $poly(A)^+$ signal sequences (AA-TAAA) and one non-consensus-type $poly(A)^+$ signal sequence (ATTAAA) plus one AU-rich element (ATTTA) are present in the 3'UTR and are *underlined*

SCRs (n=549 from 87 proteins) were aligned and used to generate a neighbor-joining tree (not shown). Clustering of SCRs implied sequence similarities, which were used to infer putative functional similarities (Table 3) and have been indicated with symbols in Fig. 4. Many SCRs from proteins not involved in complement or immune function that were included in this analysis did not cluster with the



Fig. 4 Overview of the domain structures of SpCRL (a) and SpCRS (b). The structure for SpCRL and SpCRS are positioned to illustrate the similarities between the C-terminal end of SpCRL and the structure of the entire SpCRL. Domains are indicated as follows: The leader is shown as a jagged line at the N-terminal end of both proteins. Short consensus repeats (SCRs) are numbered and indicated with a black rectangle, and similar fill patterns indicate SCRs with similar sequences. The Cys-rich region in SpCRL is shown with a *dark gray circle*, and the fucolectin domain is indicated with a variegated gray rectangle. The S/T/P domains are shown in both proteins as a *black diamond* and the factor I-membrane attack complex (FIMAC) domain as a variegated gray oval. Inferred functions of some SCRs, based on phylogenetic analysis, are shown above the SCR as a *star* to indicate putative binding to SpC3, a *plus sign* to indicate putative binding to polyanionic structures on cell surfaces such as sialic acid or heparin, and a (Zodiac) Cancer symbol to indicate cell adhesion function. Conserved N-linked glycosylation sites are shown with a black circle/ black line and O-linked sites are shown with a white circle/black line

SCRs from SpCRL and SpCRS (see footnote to Table 3). The most significant result of this analysis suggested that SCRs 5, 6, 7, and 12 from SpCRL have sequence similarities with SCRs in vertebrate Bf or C4BP, which are known to bind C3b or C4b, respectively (Hourcade et al. 1995; Blom et al. 2001), and are also known to have cofactor activity (Gordon et al. 1995; Kuhn et al. 1995) (Table 3). SCRs 1, 8, 9, 13, and 17 from SpCRL clustered with SCRs in factor H that bind sialic acid (Gordon et al. 1995) and/or C3b (Zipfel et al. 1999; Jokiranta et al. 2000). Because the C-terminal SCRs in both SpCRL and SpCRS were similar to each other and were positioned in the same relative order within the proteins (see above), it was not surprising that they clustered with each other and with the same SCRs from other proteins. The similarities identified for the C-terminal SCRs in both sea urchin proteins implied functions such as maintenance of structural conformation (SCR 1 from SpCRS, SCR 15 from SpCRL), cofactor activity and binding functions for sialic acid and C3b (SCR 3 from SpCRS, SCR 17 from SpCRL; Zipfel et al. 1999), and putative cell-adhesion activities (SCR 4 from SpCRS, SCR 18 from SpCRL; Fries et al. 1993). Overall, these results suggested the possibility that some of the SCRs in SpCRL and SpCRS may be involved in functioning within the complement system of the sea urchin and perhaps through interactions with SpC3. Furthermore, because cell surface sialic acid in mammals is important in blocking the alternative and lectin pathways of complement through binding factor H (Jarvis et al. 1987; Jack et al. 2001), SCRs in the sea urchin proteins that show sequence similarities to SCRs with sialic acidbinding activities suggest that SpCRL may also associate with cell surfaces and perhaps function in complement regulation.

FIMAC domains

Two types of proteins are known to contain FIMAC domains: the heavy chain of factor I and complement components C6 and C7 (Minta et al. 1996). The modular structure of factor I heavy chain consists of a FIMAC domain, a SRCR domain, and two LDL-receptor domains, while the light chain is a serine protease domain (reviewed by Arlaud et al. 1998). The FIMAC domain in factor I is also thought to be involved in protein-protein interactions (DiScipio 1992). The domain organization of complement components C6 and C7 have one or two thrombospondin domains, a type A LDL-receptor domain, a perforin-like domain, another thrombospondin domain, and two SCRs followed by two FIMAC domains (DiScipio and Hugli 1989; Haefliger et al. 1989). The SCR domains and the FIMAC domains in C6 are known to bind C5b (DiScipio 1992; DiScipio et al. 1999). Initial analysis of the C-terminal region in both SpCRL and SpCRS identified a single Kazal or FIMAC domain based on BLASTX searches of GenBank. This sequence has been identified previously as sea urchin coelomocyte expressed sequence tags (ESTs) [EST132 (accession no. R61984), EST351 (accession no. R62091), and EST202 (accession no. R62023), Smith et al. 1996]. Kazal domains are part of the follistatin serine protease family and function as serine protease inhibitors (Schlott et al. 2002). FIMAC domains are a subset of that family; however, they function as complement cofactors by binding C3 convertase, which promotes the cleavage of C3b by factor I (Terado et al. 2002), or by binding C5b in the terminal pathway (DiScipio 1992; DiScipio et al. 1999).

To determine whether the domains from SpCRL and SpCRS were similar to FIMAC domains from factor I and C6/C7 or if they were more like Kazal domains in protease inhibitors, they were used in a phylogenetic analysis to determine with which set of domains they clustered. The data set included 79 Kazal and FIMAC domains from 44 proteins as identified by BLASTX searches of Gen-Bank. The sequences were analyzed by parsimony using the ratchet algorithm (Nixon 1999) in PAUP (Sikes and Lewis 2001; Swofford 2002), and a portion of the 50% strict consensus tree is shown in Fig. 6a. All of the FIMAC domains, plus those from SpCRL and SpCRS and two Kazal domains, clustered together in a clade defined by node A. Within that clade, a subclade defined by node B included all factor I FIMAC domains plus the sea urchin domains. The FIMAC domains from complement components C6 and C7, plus the zebrafish FIMAC domains (from a protein with a structure consistent with a C6 homologue, accession no. AAH57429.1), clustered in two clades defined by C (Fig. 6a). This result suggested that the SpCRL and SpCRS domains were more similar to FIMAC domains than Kazal domains, and perhaps more similar to factor I FIMAC domains than to C6/C7 FIMAC Fig. 5 SCR alignment. The SCRs from SpCRS and SpCRL were aligned using ClustalX (Higgins et al. 1996) and edited in Word (Microsoft). The consensus amino acids that have been defined for SCRs in general, and most of which are found in the SCRs shown here (between 33% and 100%), are indicated in *boldface* and on the bottom line. Some linkers between SCRs were truncated for the purposes of improving this alignment and are indicated with an *asterisk*

Sp5013 SCR1	C M	- P KTDRL	GRYTVIG-	- GGGNLWSM	GV	PS	-GSRI-		Q'	VSVR C DI	⊾- G	QAIA-
Sp5013 SCR2	CSV	- P SNKGE	VHQFV	NGDATMPIT	SY Y TTSTLE	5 1	TDC		Q	lvar c ri	0PG-K Y	RLI
Sp5013 SCR3	CDV	- P AAPE	HGRILD	- G	YRQQARI	DIFPS	- G T		S	VSFE C EI)D Y	RLV
Sp5013 SCR4	C	- PELTIE	- AH- IAHD	- G	-V Y GYGRDI	LRD	<i>ERI</i> .	E X -		C KI	GY	VLR
Sp5 SCR1	CDNIN	ATDIAN	<i>TT</i>		YDEPEGI	IA HZAL	T D C	AV-	NIG	C-	- GVFY	T G V I Q
Sp5 SCR2	C	- A- V3 4 -	QA	- G LLVT	Y DTPE	RRF1	3G	- Y AI	NFS	C GI	L-GYIY	HS
Sp5 SCR3	CRD	- PSAEHA	SLQPSGKÇ	2	X ET		GDTRS	- Y -		T C NI	- G Y	IIR
Sp5 SCR4	CTY	- P GRISN	QTTIDA	5PE	<i>Q</i> F¥	(E	RA	- Y I		V C RH	2-GY	ENDDA
Sp5 SCR5	C	- PA-IL-	-DIT	NGN-VDSRG	ND F D-SQI-			- Y -		FT C NI) - G Y	RLD
Sp5 SCR6	CED	- P GV - PV		NG	– – A MENEKČ	2VYI	ITDDAA.	Y Y – I	<i>H</i>	C NF	e – G KTI	D
Sp5 SCR7	CIV	- P PY - PR	-DG-WW-Ç	0N G NE	Y- PPETS	SVPH-NTF	RLQ			LT C R-	SW	RFNK-
Sp5 SCR8	C GV	RWSIV	TINJA		Y DPPISI	DEDRFNH	- CHQIS		TRVHY	D C AV	1-GY	RLQ—-
Sp5 SCR9	CSA	- P ED V	SH	- G RLT	$ \mathbf{Y}TNPDG$	Abhenri	G DTRL			LQ C GI	- G Y	RSR—-
Sp5 SCR10	C D	- P I PD	MT	- G HVN	Y TRTA	-KAJ	1 G K	- Y I)	H G TEVTVI	N C NS	6 - G Y	- YAGJ
Sp5 SCR11	CVV	- P RV	ENADVSST		VP	ATQQVLLF	EGEDVR			IS C HI	-RHVL	RGSDT
Sp5 SCR12	C TVD -	<i>RLF</i>		N G G	F PNGK		I.	$L\mathbf{Y}N$	H G ENITF'	Г С МІ	9 - G Y	EKE
Sp5 SCR13	CAN	- P DL - PD	QATV−−MH	<u>2</u> S	QP-	-DF		bj	HGTLLDV	3 C EI) - G F	ELST-
Sp5 SCR14	CEG	- P PIAPI	K		Y YQGS	3Q	-G	- Y V!	H G TRVRY	5 C RI	2 - G W	$EIE^{}$
Sp5 SCR15	CML	- PDR - SS		- G HR—	Y SIR-GF	TGNIFDE	E G LPI-		-GERVSI	SVS - C NQ	2-GY	-—QAT
Sp5 SCR16	CTK	- PGYISH	NVYQVV	- GLETNRLD	S-YPHDE	<i>L</i> PF	E G TFLV			SR C SI	PGQ-Y	VLH—-
Sp5 SCR17	CDR	- P T - PT	M	- GAFQDHD -	Y YGWK	· 1	4G −Q−−	- Y Y	MGKVITF	A C NI) - G Y	ILD
Sp5 SCR18	CEELH	TT P T		H G TKIG			- G	-NR	I G DSVLI	3 C N(2-GY	QLQ—-
consensus	С	Р		G	Y		G	Y	G	С	G	
					F							
Sp5013 SCR1	QSVTS	3Q- C	SNGIWIPE	$W \mathbf{P} \mathbf{R} \mathbf{C} \mathbf{E} K$	P							
Sp5013 SCR2	$\mathbf{G}DQRF$	R-T C	QG G Q W TGE	$E = \mathbf{P} E C Q H$	-Y7LVQAQV	*						
Sp5013 SCR3	G S-RF	RITC	EK G Q W SDE	0P P V C TV	SR							
Sp5013 SCR4	GDTM	NLR C LES	G V W SDE	$T \mathbf{b}L \mathbf{C}EY$	VVHQVSP							
Sp5 SCR1	DFER	Г-Т С	TN G S W TE-	PLPRCRH	IPRYCQRHV] - *						
Sp5 SCR2	G SSR-	-lr c msn	RQ WGGG	SNDFV GID	NVV	-						
Sp5 SCR3	G SRD]	I – T C GDE	Y - G - W SE -	PDFT C RP	RP							
Sp5 SCR4	TTDAL	R-T C -Q-	-QT W Q D IAA	ILPICEP	IQ	-						
Sp5 SCR5	GTARE	R-V C -Q-	GDKT W SG-	QEAV CTE	II							
Sp5 SCR6	GSILM	N-SC-TE	-SGEWRY-	PV P V C GG	<i>b</i>	-						
Sp5 SCR7	RRSS	J-KC	ND G V W SDS	DDVHRLCRG	TP							
Sp5 SCR8	G VTES	S-R C	EQ G RYNN-	NIPRCEL	VP							
Sp5 SCR9	SFNSS	S-R C -	DN GVWVEC	S-HDIR C YP	Kb							
Sp5 SCR10	GNGTI	A-V C	VTJ W QZAN	I P T C TQ	SRNIILDSF	TF						
Sp5 SCR11	NHIDI	L-TCL	GNSS W DQ-	DK P V C EP	ET							
Sp5 SCR12	-HDRY	Y-W C -DR	VAIR.	-PRSPRCIQ	AS							
Sp5 SCR13	NQEQI	L-R C -	YR G G W NT -	-PLTAT CQQ	SVMWTS							
Sp5 SCR14	GIAER	R-E C	INRQ W TG-	STPACRV	<i>R</i> 44AA	-						
Sp5 SCR15	PSVQ	T-E C I	ER G V W SV-	AVPICVR	MERL							
Sp5 SCR16	GSANF	R-T C	SESS W TG-	VQPSCVE	ADTRIS	*						
Sp5 SCR17	GERR	I – T C – – –	VL G K W SH-	PA P R C QR	HRAT							
Sp5 SCR18	$\mathbf{G} E P F I$	L-D C- Q-	ES GNWSH-	PL P ACIE	IIEPERPCY	IS						
aonaonaua	G	C	CW	ъ <i>с</i>								

domains. Bremer support (Bremer 1988) was calculated for several of the internal nodes, and is shown below the branch near the node (for details, see legend to Fig. 6). Bremer support indicates the number of extra steps that is required in a longer tree before a clade is lost from the consensus tree (Kitching et al. 1998). Consequently, support was quite good for nodes A and B, within which SpCRL and SpCRS clustered. Bootstrapping analysis was conducted, but results were uninformative. This was due to the short length of the domain (84 informative positions) and the large number of sequences (79) that were analyzed. Bootstrapping is optimal when the data matrix has at least 1,000 informative characters (Kitching et al. 1998). Bootstrapping results may also have been confounded by the presence of several sequences, such as the FIMAC domain from carp factor I-A, which did not align well (see Fig. 6b). The carp sequence did not cluster repeatedly within a given clade but "jumped" to a variety of locations within clade A in different, but equally parsimonious, trees. An alignment of the sequences that clustered in the clades defined by nodes B and C, plus the carp sequence, is presented in Fig. 6b. For most sequences, including those from SpCRS and SpCRL, the conserved amino acids are present including 10 Cys, which are involved in forming the disulfide bonds within the domain (Terado et al. 2002), in addition to Trp, Gln/ Glu, Arg/Lys, Pro, Val and Ala/Ser. The carp sequence does not align well, particularly in the C-terminal end of the domain.

Although the domains of both SpCRL and SpCRS cluster with FIMAC domains from factor I proteins, the domain structure of the sea urchin proteins, in which the FIMAC domains were preceded by two SCRs (Figs. 2, 3, 4), was similar to the domain structure of complement proteins C6 and C7 (DiScipio 1992). Factor I, on the other hand, has the FIMAC domain located at the N-terminal end of the protein and is not associated with adjacent SCRs (Arlaud et al. 1998; Terado et al. 2002). Together, the combination of sequence similarity to factor I and the structural similarity to C6 and C7 implies that the FIMAC domains in SpCRS and SpCRL may have protein-protein binding functions similar to that known for FIMAC domains in factor I (DiScipio 1992) or for C6 (DiScipio et al. 1999), both of which interact with C3 or C5-members of the thioester-containing protein family. It is interesting that the C6 homologues identified in Amphioxus

	Protein source of SCR	Similar SCR	Protein function, SCR function	Species	Accession nos.
SpCRL	SCRs ^a				
1, 9,	Factor H-related	SCR 6, SCR	Complement regulation, SCR13 binds	Human, pig,	XP_037279.1, CAC81999.1,
8, 13	protein-5, factor H	13	sialic acid	mouse, rat	NP_034018.1, NP_569093.1
3	IL2 receptor	SCR 2	Binds IL-2, SCR function unknown	Mouse, rat	P01590, P26897
5	Factor B	SCR 1, SCR 2	Alternative pathway, SCRs bind C3b	Pig, mouse	Q03710, 67613
6	C4 binding protein	SCR 6, SCR 8	Complement regulation, SCR6 has cofactor activity, SCR8 binds C4b	Rat, human, mouse	Q63514, P04003, P08607
7, 12	Factor B	SCR 1	Alternative pathway, SCR1 binds C3b	Mouse	67613
10	MASP 2, C1r	SCR 1	Lectin and classical complement pathways, SCR function unknown	Human, mouse	NP_006601.2, NP_075632.1
15	Complement	SCR 8, SCR	Complement receptor type 2, binds	Mouse, sheep	XP 129684.1, AAB92375
	receptor type 2	11	C3d, SCRs function in structural conformation?		_ ,
17	Factor H, factor H-like protein 1	SCR 3	Complement regulation, SCR3 binds sialic acid. C3b	Mouse, rat, pig, human	NP_034018.1, NP_569093.1, CAC81999.1, NP_002104.1
18	P selectin	SCR 2, SCR 3	Cell adhesion, SCR function un-	Mouse, rat, human	NP_035476, NP_620234.1, NP_002996.1
SpCRS	, SCRs ^a				
1	Complement	SCR 8, SCR	Receptor for C3d, SCR function	Mouse, sheep	CAB03143.2, XP_129684.1,
	receptor type 2	11	structural conformation?		AAB92375
2	Complement	SCR 6, SCR	Complement regulation, SCR func-	Human	I73012
	receptor type 1	13, SCR 20,	tion structural conformation?		
		SCR 27, SCR			
		34			
3	Factor H, factor	SCR 3	Complement regulation, SCRs bind	Mouse, rat, pig,	NP_150094, NP_444401,
	H-like protein I		stalic acid and C3b	human	NP_034018.1, NP_569093.1,
					CAC81999.1,
					NF_002104.1.1, ND_027270_1
4	D coloctin	SCD 1 SCD 2	Call adhesion SCP function up	Mouse rot	Ar_U3/2/9.1 ND 025476 ND 620224 1
4	r seleculi	SCK 2, SCK 3	known	human	NP_002996.1

 Table 3 Sequence similarities between short consensus repeats (SCRs) from <u>Strongylocentrotus purpuratus complement related protein</u>, long form (SpCRL) and <u>S. purpuratus complement related protein</u>, short form (SpCRS) and SCRs from other proteins

^a SCRs used in this analysis were obtained from the following proteins: factor B (accession nos. NP_571413, AAA31021, P81475, NP_032224.1), mouse complement component C2 (XP_123064.1), C4 binding protein (NP_000706.1, NP_000707.1, P08607NP_036648.1, NP_058691.1, NP_031602.1), factor H (NP_569093.1, NP_034018.1, CAC81999, gil2135094), factor H-related proteins (XP_037279.1, CAA66980.1, NP_002104.1, CAA48639.1), MASP (NP_006601.2, XP_029605.1, XP_148328.1), C1s (13787045), C1r (NP_075632.1), selectins (NP_002996.1, NP_035477.1, NP_000646.1, NP_037246.1), type 1 and 2 complement receptors (21536276, A46458, XP_002008.7, AAB92375, XP_129684.1), chondroitin sulfate proteoglycan (NP_113841.1), versican (NP_04376.2, XP_127448.1), aggrecan (NP_071526.1), neurocan (XP_125051.1), brevican (NP_031555.1, NP_068767.2), MCP (NP_002380.2, NP_062063.1, NP_034908.1), DAF (NP_000565.1, NP_034146.1, NP_031853.1, P49457), haptoglobin (XP_042621.1, NP_059066.1), scavenger receptor cysteine-rich protein (AAB40715.1, T17405), apolipoprotein receptor (AAA30994.1), apolipoprotein H (NP_038503.1), seizure gene 6 (XP_114203.1, XP_126232.1), pregnancy protein A (NP_002572.1, XP_131437.1), polydomain protein (NP_078776.2, NP_073725.1), sushi domain protein (NP_081114, NP_150094.1, NP_032384.1, NP_058607.1), thyroid peroxidase (NP_062226.1, NP_033443.1, NP_000538.2), interleukin receptors (NP_037295.1, NP_032384.1, NP_032393.1), complement component C7 (NP_000578), C3b/C4b receptor (173012), protein X (NP_006298.1), human KIAA1884 (XP_055539.6), Drosophila scavenger receptor (NP_477102, NP_524747), Drosophila hikaru genki (Q09101), β -2 glycoprotein (AB20668), pox virus complement inhibitor (gil10120606), GABA B (XP_165689.1), p100 serine protease (BAA03944.1), factor XIII β (NP_112441.1), zona pellucida 3 receptor (NP_477102, NP_524747), Drosophila hikaru genki (Q09101), β -2 glycoprotein (AB20668), pox virus complement (NP_033607.1), *Caenorhabditis elegans* K07E12 (AAA50715.1), *C. elegans* LDL receptor (CAB03143.2), *C. elegans*

(Suzuki et al. 2002) and *Ciona* (Azumi et al. 2003) have a domain organization that is similar to the N-terminal half of vertebrate C6 and C7, including the thrombospondin domains, a type A LDL-receptor domain, and a perforinlike domain, but lack the SCR and FIMAC domains. Consequently, the C6-like sequences in these two lower chordates share no domains with SpCRL and SpCRS, which also appear to be C6-like. Determination of whether the chordate or echinoderm C6-like proteins might actually function in a terminal pathway will require additional analysis of complement activities from these invertebrates.

Fucolectin domain

A BLASTX search of the region between SCR 10 and SCR 11 in SpCRL revealed a significant sequence similarity with a family of fucolectins that have been characterized from the serum of the European eel, *Anguilla anguilla*, and have been denoted the *A. anguilla* agglutinins (AAA) (Bianchet et al. 2002). Fucolectins are a class of lectins that recognize fucose on the surface of cells and play an important role in the innate immune functions of both vertebrates and invertebrates as patternrecognition receptors (Bianchet et al. 2002). An alignment between the SpCRL fucolectin domain and the seven eel fucolectin proteins revealed a number of conserved amino acids in addition to several positions where the eel and sea urchin sequences differed (Fig. 7). The fucolectin structure, which is diagnostic of the F-lectin family, is a β barrel with jellyroll topology of eight antiparallel β strands oriented in two sheets consisting of five and three strands (Bianchet et al. 2002). The domain is locked into its conformation with two disulfide bonds and two salt bridges. The disulfide bonds were conserved in both the eel fucolectins and the SpCRL fucolectin domain (Fig. 7). The salt bridges were formed between Arg41-Glu140 and Asp64-Arg131 in the eel sequences that interact with a cation-either Na⁺ or Ca²⁺. In SpCRL, neither Arg involved in salt bridges were present in the conserved positions, but were located within two to four positions of the site conserved in the eel sequences. In the AAA fucolectins, there are five loops connecting the β strands of the β barrel that protrude like complementarity-determining regions (CDRs) and surround the fucose-binding site. The CDRs in SpCRL showed conserved amino acids that defined the borders of the regions and the sizes of the regions were similar to that in AAA. The fucose-binding site, composed of the motif $H(X)_{24}RGDCC(G/E)ER$, showed conservation of the significant amino acids in SpCRL except for His52, which was replaced with an Arg, and the double-Cys motif, which was missing in SpCRL. Fucolectins have been shown to have hemagglutination activity mediated by the conserved residues AIDGN located between CDR 1 and CDR 2, a motif that was partially conserved in SpCRL, ARDGN. The detailed structural analysis of fucolectins reported by Bianchet et al. (2002) has enabled a detailed analysis of the fucolectin domain of SpCRL and overall, the alignment suggests that the sea urchin protein has a fucolectin domain, but comparisons revealed a divergent structure from that of the AAA proteins (G. Vasta, personal communication). Variations in amino acids required for salt-bridge formation to stabilize the barrel structure and the missing His52 which is involved in fucose binding in AAA, make it possible that the sea urchin domain may bind a different ligand than that characterized for AAA. The location of the ligand, perhaps on self-cells or on the pathogen, may provide a clue as to how SpCRL might be involved in the complement system of the sea urchin.

Transcript sizes and alternative splicing

Sp5

The mRNA length for both cDNAs were determined by $poly(A)^+$ Northern blots (Fig. 8). Results for *Sp5* (Fig. 8a) showed two pairs of bands (7.4 kb and 7.9 kb, 11.9 kb and 12.7 kb) with the smaller of each pair being more intense. All major bands were larger than the sequenced *Sp5* cDNA (5.965 kb) suggesting that parts of the 5'UTR and 3'UTR were missing. The 5'UTR was very short, and it is feasible that much of this region was missing. However,

the blot and the cDNA sequence both suggested that the short stretch of As at the 3' end was not the actual $poly(A)^+$ tail, indicating that some of the 3'UTR was also missing. This was supported by the absence of a polyadenylation signal sequence located 5' of the $poly(A)^+$ stretch (Fig. 2). Taking this into consideration, the known sequence for *Sp5* may correspond to either the 7.4-kb or 7.9-kb band. Longer exposures of the same blot showed five less prominent bands ranging in size from 1.2 kb to 3.8 kb (Fig. 8b). These bands may not correspond directly to the *Sp5* sequence because they were significantly shorter than the known sequence of the cDNA and did not hybridize well to the probe (see below for a discussion of small gene family and alternative splicing).

Sp5013

The Northern blot for *Sp5013* revealed three major bands of 6.0 kb, 4.6 kb, and 3.8 kb after a short exposure (Fig. 8c), plus six weakly hybridizing bands of 12.7 kb, 11.4 kb, 9.3 kb, 7.4 kb, 2.9 kb, and 2.2 kb that appeared after a longer exposure (Fig. 8d). The known sequence of Sp5013 (2.795 kb) was not exactly the same size as any of the bands on the Northern; however, the cDNA might correspond with either the 2.9-kb or 3.8-kb band if parts of both the 5'UTR and 3'UTR were missing from the cDNA sequence. The 3'UTR was probably complete because the $poly(A)^+$ tail was preceded by two polyadenylation signal sequences located 100 nt to 150 nt upstream. Although we argue above that the 5'UTR was complete based on duplicate cDNAs with the same sequence at the 5' end, the shortness of this region is not typical for sea urchin mRNA.

Similar band sizes and alternative splicing

Size comparisons between the array of bands for Sp5 and Sp5013 in Fig. 8 indicate that a few may be the same. However, bands of the same size never appeared as major bands for both probes. The best example of message sizes present on both blots was the 12.7-kb band, which was one of the major Sp5 doublet bands and was a minor band for Sp5013 (Fig. 8b, d). Another example was the 3.8-kb band, which was a major band on the blot for Sp5013 but was a minor band for Sp5. Finally, the 7.4-kb band, which appeared on both blots, was a major band for Sp5 and a minor band for Sp5013. In addition, there was a set of minor bands of 2.8 kb/2.9 kb that was present on both blots. The probes used on the Northern blots were produced from clones that included the entire Sp5013 sequence, and that spanned the 3' half of the Sp5 sequence (the fucolectin domain to the 3' end, see legend to Fig. 1). These were the regions of the two messages that showed the greatest similarity and may have resulted in some cross-hybridization, even at high stringency. Because the small bands did not correspond to sizes estimated from the known sequences of the cDNAs, this suggests that the







b

SpCRS FIMAC	CASVRCESWQRCEVTGPNT	TVCGCINPTS-C-LTSGPTVC
SpCRL FIMAC	IIEPERPCYSVSCGVWQKCETDSSGV0	GVCRCISPNS-CPVTNETEVC
Human Factor I	SCDKVFCQPWQRCIE	GTCVCKLPYQ-CPKNG-TAVCA
Mouse Factor I	SCNKVFCQPWQRCIE	GTCICKLPYQ-CPRAG-TPVCA
Xenopus Factor I	SCHKVFCAPWQRCVA	GVCRCKLPYQ-CPKNATTEVCI
Shark Factor I	KSCQKVFCQPWEKCIN	GRCECKLPYQ-CPKQV-NEVCS
Fugu Factor I	RRYTRQSCDLVFCPPWERCLD	GQCLCKVPYQ-CPSENVTAVC
Human C6 FIMAC 1	TKLKGHCQLGQKQSGS	SECICMSPEEDCSHHS-EDLC
Human C6 FIMAC 2	SCGYDTCYDWEKCSASTS	SKCVCLLPPQ-CFKGGNQLYC
Human C7 FIMAC 1	TQAVPKCQRWEKLQNS	SRCVCKMPYE-CGPSLDVCA
Human C7 FIMAC 2	PASAEKACGACPLWGKCDAESS	SKCVCREASE-CEEEG-FSIC
Zebrafish FIMAC 1	PDSSCKPGEINDG	TKCVCMTKER-CRGYR-EDLC
Zebrafish FIMAC 2	EPCGSDTCYEWETCSVSH	KTCECKMPRE-CPKDGKKIYCI
Carp Factor I-A	MRAVFYFMCLLFQTALN	- QPKV P DEDFLG-PAQ C I
conserved position	ns CVCWQKC	CCPC VC
	ER	
SpCRS FIMAC	T-NGDYYS-SECHLRAYACLHDLO	QLDIAENDYG C INGAP
SpCRL FIMAC	T-DGRNYT-NFCRLKALACIQNT	GVEVASRTWFCINGVP
Human Factor I	T-NRRSFP-TYCQQKSLECLHP	GT- -
Mouse Factor I	M-NGRSYP-TYCHQKSFECLHPE-	IKFSHNGT- C AA
Xenopus Factor I	D-GKRKLQ-SYCQLKSVECSNPLM	NSKYRFSSEAP-CTE
Shark Factor I	S-RGKKYR-SYCQLKSIECIRGLE	ESFSHFGM- C SMGT
Fugu Factor I	R-DGRNYR-SYCQVMAVSCRTKSI	PKFSHFGQN C AVVRVFVP
Human C6 FIMAC 1	FDTDSND-YFTSPACKFLAEKCLNNQ	QLHFLHIGS- C QD
Human C6 FIMAC 2	K-MGSSTSEKTLNICEVGTIRCANRK	MEILHPGK- C LA
Human C7 FIMAC 1	Q-DERSKRILPLTVCKMHVLHCQGRN-	YTLTGRDS- C TL
Human C7 FIMAC 2	EVNGKEQTMSECEAGALRCRGQS-	ISVTSIRP- C AA
Zebrafish FIMAC 1	YDAGKET-AIMMSLCAFHADRCHGDRI	LYFMNNGP- C KSD
Zebrafish FIMAC 2	KIVRTQTTRSMN-LCFMAAMKCSSIE	FELQHEGP- C AGS
Carp Factor I-A	DQKYTR-LSCSKVF	
conserved position	IS CAC	С
-	ŝ	

Fig. 6a, b Phylogenetic analysis and alignment of the FIMAC domains from SpCRL and SpCRS. a An alignment using 79 FI-MAC and Kazal domains from 44 proteins was done using ClustalX (Thompson et al. 1997), edited in BioEdit sequence alignment editor (Hall 1999), formatted in Winclada (obtained from http://www.cladistics.com), and analyzed in PAUP (Sikes and Lewis 2001; Swofford 2002) to generate a phylogenetic tree. Multiple

FIMAC and Kazal domains from the same protein are *numbered* and treated as independent sequences. Accession numbers are indicated in *brackets*. The most parsimonious tree was calculated using the parsimony ratchet algorithm, PAUPRat, version 1 (Nixon 1999), with 1,000 iterations, each with 20 random additions of taxa and a limit of 20 trees saved and swapped for each random addition. A portion of the 50% strict consensus tree is shown that was



Fig. 7 Fucolectin alignment. The fucolectin domain from SpCRL was aligned with the seven fucolectins from the European eel, *Anguilla anguilla* [denoted *A. anguilla* agglutinins (AAA)], using ClustalX (Higgins et al. 1996) and edited in Word (Microsoft). The conserved amino acids are shown in *boldface* within the alignment. Amino acids directly involved in binding functions are shown in *boldface* in the *top line* in addition to the fucose-binding site and the hemagglutination site. The complementarity-determining regions (CDRs) are labeled and identified with *brackets*. The two Arg located within the fucose-binding site and the His located towards the N-terminus (in *boldface, top line*) are the triad of amino acids

that are involved in the polar interactions with fucose. The first residue of this triad in SpCRL is Arg, which does not match the conserved His in the eel fucolectins. The Cys are identified with numbers on the top line and disulfide bonds are formed between Cys1–4 and Cys2–3 in AAA sequences (Bianchette et al. 2002). Four amino acids involved in stabilizing the domain through salt bridges in the AAA sequences are indicated with *arrows*. Salt bridges are formed between Arg41 and Glu149 and between Asp64 and Arg131 (numbering is based on Bianchette et al. 2002). Amino acids involved in cation binding are shown in *italics* on the *top line*

generated from 919 equally parsimonious trees. Bremer support (Bremer 1988) was calculated for several internal nodes using PAUP and TreeRot, version 2 (Sorneson 1999), and is indicated near the relevant nodes below the branches. Kazal and FIMAC domains (54) from 18 other proteins are not shown on the tree. They include bovine acrosin inhibitor [P01000], C. elegans protease inhibitor [CAB01753.1], chick agrin [P31696], chick follistatin [Q90844], chick ovoinhibitor [P10184], chick ovomucoid [P01005], crayfish proteinase inhibitor [CAA56043.1], Dipetalogaster thrombin inhibitor [CAA10384.1], dog double-headed protease inhibitor [P01002], eel pancreatic proteinase inhibitor [P11706], Eurasian badger double-headed protease inhibitor [P16226], herring sperm activating protein [BAA14008.1], human acrosin trypsin inhibitor [P20155], human agrin [AAC39776.1], human serine protease inhibitor [CAB40839.1], human prostacyclin stimulating factor [AAB32370.1], human testican 3 [NP-058646], human transmembrane protein [AAA64622.1], Japanese quail QR1 protein [P23499], lion double-headed protease inhibitor [P08481], leech protease inhibitor [AAK58688.1], mouse testican [CAA63448.1], mouse serine protease inhibitor Kazal type 4 [NP035593.1], pig sperm-associated acrosin inhibitor [P00999], rat pancreatic trypsin inhibitor [P09655_IPK1], sea anemone elastase inhibitor [P16895], turkey ovomucoid [P01004]. b An alignment of FIMAC domains from SpCRL and SpCRS with FIMAC domains that clustered in a. The alignment was done with ClustalX (Thompson et al. 1997) and edited in Word (Microsoft). Consensus amino acids are shown in *boldface* and on the *bottom line*

probes may have cross-hybridized to messages other than *Sp5013* and *Sp5*, implying the presence of a small gene family with similar sequences.

Multiple bands on the Northern blots may also be interpreted as evidence of alternative splicing. The pattern of size differences among the bands for Sp5013 was present in multiples of 0.7-0.8 kb (Fig. 8e), which corresponds with regions of the message that would encode pairs of SCRs including the linker between them. For example, if the length of the message encoding SCR 1 and SCR 2 plus the linker were doubled, it would be 810 nt. Similarly, if the sequence encoding SCR 3 and SCR 4 plus the linker were doubled, it would be 738 nt (Fig. 8e). These sizes correspond with regular size differences noted between the bands for Sp5013 (Fig. 8c, e) and could be interpreted as the result of alternative splicing of multiples of SCR exons. If this set of messages of varying lengths were transcribed, it would result in a set of proteins with many more SCRs and perhaps fewer SCRs than were identified from the deduced sequence of SpCRS. In support of this interpretation, several clones for Sp5013 were identified with identical sequences at the 5' and 3' 104



Fig. 8a-d Expression of Sp5 and Sp5013 in coelomocytes. Poly(A)⁺ RNA was isolated from coelomocytes from an animal within 1 day of shipment from California and therefore assumed to be activated. Between 1 μ g and 1.8 μ g mRNA was loaded per lane on a Northern blot and analyzed with riboprobes made from Sp5 and Sp5013. Clones used as riboprobe templates for this purpose included the entire sequence for Sp5013 and approximately the 3' half of Sp5 (see legend to Fig. 1). The blot probed for Sp5 was exposed for 30 min (a) and 5 h (b). The blot probed for Sp5013 was also exposed for 30 min (c) and 5 hr (d). Expression of SpL8, the sea urchin homologue of the human ribosomal gene L8, was used as a control, and a single band of 1.6 kb is shown. Sizes of the bands are indicated between the two sets of blots. e Differences in observed sizes between bands for Sp5013 (d) and the approximate number of deleted SCRs which might have resulted in the observed differences are indicated to the right (see text)

ends to the sequence provided here, but with proteincoding regions of variable sizes (data not shown). The deletion of SCRs by alternative splicing has recently been demonstrated for another sea urchin cDNA, *Sp152*, which encodes the homologue of complement factor B, SpBf (Terwilliger et al. 2000; unpublished data). Splice variants of *Sp152* were identified in which the first and/or the fourth SCR were/was deleted from the total of five.

Genome analysis

Genome blots using sperm DNA from two sea urchins that was digested with *Hin*dIII, *Eco*RI, and *Pst*I were analyzed with probes for each gene. Results showed multiple bands of variable intensity in both blots that would be consistent with gene structure of multiple exons plus cross-hybridization among members of a small gene family (data not shown).

Expression patterns of *Sp5* and *Sp5013* in sea urchin tissues

Expression patterns of *Sp5* and *Sp5013* were characterized by RT-PCR using several sets of gene-specific primers and total RNA from several sea urchin tissues (Fig. 9). The tissues examined included coelomocytes, gut, gonad, pharynx, esophagus, and axial organ and were



Fig. 9 Expression of Sp5 and Sp5013 in sea urchin tissues. Total RNA was isolated from coelomocytes, gut, gonad, pharynx, esophagus, and axial organ and used in reverse transcriptase (RT)-PCR. The RNA was analyzed for expression of Sp5 and Sp5013, and Sp056 (an LPS-inducible C-type lectin, accession no. AY336600, unpublished data). SpL8, a constitutively expressed homologue of the human ribosomal gene L8, was used as the control. a A sea urchin was injected with LPS for 2 days and coelomocytes and other tissues were collected for RNA isolation 24 h after the last injection. **b** The sea urchin used for analysis was assumed to be upregulated as a result of shipping stress. c Tissues were collected from an immunoquiescent sea urchin which had been kept in a closed system aquarium for 2 years prior to being used in this experiment (see Gross et al. 1999). d Coelomocytes were taken from an immunoquiescent sea urchin 15 min prior to a single LPS injection and 24 h post-injection. C Coelomocytes, G gut, OT ovary or testis, E esophagus, P pharynx, AO axial organ. Control PCR reactions employed cloned templates with the appropriate primers (positive) or reactions in which the cloned template was omitted (negative)

isolated from sea urchins that were either activated by injection of LPS (Fig. 9a), were in an immune activated state (Fig. 9b), or were immunoquiescent (Fig. 9c). Results showed that all tissues expressed both Sp5 and Sp5013, and that expression was not altered in coelomocytes from immunoquiescent animals after injection with LPS (Fig. 9d). In comparison, expression of Sp056 (accession no. AY336600), which encodes a C-type lectin that is reliably inducible with LPS (unpublished data), was restricted to LPS-activated coelomocytes. A sea urchin homologue of the human ribosomal gene L8, SpL8 (EST219, accession no. R62029; Smith et al. 1996) was used as the control to ensure that approximately equal amounts of cDNA template were used in the PCR reactions. Overall, the RT-PCR analysis indicated that Sp5 and Sp5013 were ubiquitously expressed and were not induced by immune challenge.

Summary

Analysis of the two mosaic proteins, SpCRL and SpCRS, indicated that they share domains with factor H, factor I, C6 and C7. Similarities among SCRs from a variety of complement proteins (Table 3) indicated that many of the SCRs in both SpCRL and SpCRS may function in interactions with other proteins such as SpC3, or with cell surfaces perhaps through binding carbohydrates (Fig. 4). The FIMAC domains in SpCRL and SpCRS show sequence similarities with FIMAC domains from complement proteins rather than with Kazal domains from protease inhibitors (Fig. 6). Functions of SpCRL might involve interactions with SpC3 as a regulatory protein to modulate complement activation, either in fluid phase or in association with self-cell surfaces, either through some of the SCRs and/or through the fucolectin domain. This is the first evidence of proteins in the sea urchin that might function in a putative complement regulatory system and/ or perhaps in a putative terminal pathway. Knowledge of whether an expanded complement system exists in the sea urchin, as has been identified from the Ciona genome (Azumi et al. 2003), will also most likely come from the analysis of the sea urchin genome when it is completed. An understanding of all the components and pathways of the sea urchin complement system will provide a better understanding of the evolution of this system that has culminated in the complex complement system that is essential to immune functions in the higher vertebrates today.

Acknowledgements The authors would like to thank Dr. Sham V. Nair, David P. Terwilliger, and Dr. Zeev Pancer for their assistance with this research. We are greatly indebted to Fernando Alvarez-Padilla for his assistance in the phylogenetic analysis. We appreciated the helpful suggestions from the anonymous reviewers. This research was supported by funding from the National Science Foundation (MCB-0077970).

References

- Al-Sharif WZ, Sunyer O, Lambris JD, Smith LC (1998) Sea urchin coelomocytes specifically express a homologue of the complement component C3. J Immunol 160:2983–2997
- Arlaud GJ, Volanakis JE, Thielens NM, Narayana SV, Rossi V, Xu Y (1998) The atypical serine proteases of the complement system. Adv Immunol 69:249–307
- Asson-Batres MA, Spurgeon LS, Diaz J, DeLougherty TG, Bagby GC (1994) Evolutionary conservation of the AU-rich 3' untranslated region of messenger RNA. Proc Natl Acad Sci USA 91:1318–1322
- Azumi K, De Santis R, De Tomaso A, Rigoutsos I, Yoshizaki F, Pinto MR, Marino R, Shida K, Ikeda M, Ikeda M, Arai M, Inoue Y, Shimizu T, Satoh N, Rokhsar DS, Du Pasquier L, Kasahara M, Satake M, Nonaka M (2003) Genomic analysis of immunity in a Urochordate and the emergence of the vertebrate immune system: "waiting for Godot." Immunogenetics 55:570– 581
- Barilla-LaBarca ML, Liszewski MK, Lambris JD, Hourcade D, Atkinson JP (2002) Role of membrane cofactor protein (CD46) in regulation of C4b and C3b deposited on cells. J Immunol 168:6298–6304

- Becherer JD, Alsenz C, Servis B, Myones B, Lambris JD (1989) Cell surface proteins reacting with activated complement components. Complement Inflamm 6:142–165
- Bianchet MA, Odom EW, Vasta GR, Amzel LM (2002) A novel fucose recognition fold involved in innate immunity. Nat Struct Biol 9:628–634
- Blom AM, Kask L, Dahlback B (2001) Structural requirements for the complement regulatory activities of C4BP. J Biol Chem 276:27136–27144
- Boolootian RA, Giese AC (1958) Coelomic corpuscles of echinoderms. Biol Bull 15:53–63
- Bremer K (1988) The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. Evolution 42:795–803
- Cameron RA, Mahairas G, Rast JP, Martinez P, Biondi TR, Swartzell S, Wallace JC, Poustka AJ, Livingston BT, Wray GA, Ettensohn CA, Lehrach H, Britten RJ, Davidson EH, Hood L (2000) A sea urchin genome project: sequence scan, virtual map, and additional resources. Proc Natl Acad Sci USA 97:9514–9518
- Chou K-C, Heinrikson R (1997) Prediction of the tertiary structure of the complement control protein module. J Protein Chem 19:765–773
- Clow LA, Raftos DA, Gross PA, Smith LC (2004) The sea urchin complement homologue, SpC3, functions as an opsonin. J Exp Biol (in press)
- DiScipio RG (1992) Formation and structure of the C5b-7 complex of the lytic pathway of complement. J Biol Chem 267:78087– 78094
- DiScipio RG, Hugli TE (1989) The molecular architecture of human complement component C6. J Biol Chem 267:16197– 16206
- DiScipio RG, Linton SM, Rushmere NK (1999) Function of the factor I molecules (FIMs) of human complement component C6. J Biol Chem 274:31811–31818
- Edds KT (1993) Cell biology of echinoid coelomocytes. I. Diversity and characterization of cell types. J Invert Pathol 61:173– 178
- Fries JW, Williams AJ, Atkins RC, Newman W, Lipscomb MF, Collins T (1993) Expression of VCAM-1 and E-selectin in an in vivo model of endothelial activation. Am J Pathol 143:725–737
- Gordon DL, Kaufman RM, Blackmore TK, Kwong J, Lublin DM (1995) Identification of complement regulatory domains in factor H. J Immunol 155:348–356
- Gross PS, Al-Sharif WZ, Clow LA, Smith LC (1999) Echinoderm immunity and the evolution of the complement system. Dev Comp Immun 23:429–442
- Gross PS, Clow LA, Smith LC (2000) SpC3, the complement homologue from the purple sea urchin, *Strongylocentrotus purpuratus*, is expressed in two subpopulations of the phagocytic coelomocytes. Immunogenetics 51:1034–1044
- Haefliger JA, Tschopp J, Vial N, Jenne DE (1989) Complete primary structure and functional characterization of the sixth component of the human complement system. Identification of the C5b-binding domain in complement C6. J Biol Chem 264:18041–18051
- Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41:95–98
- Hansen JE, Lund O, Tolstrup N, Gooley AA, Williams KL, Brunak S (1997) O-glycbase, ver 2.0—a revised database of O-glycosylated proteins. Nucleic Acids Res 25:278–282
- Hansen JE, Lund O, Tolstrup N, Gooley AA, Williams KL, Brunak S (1998) NetOglyc: prediction of mucin-type O-glycosylation sites based on sequence context and surface accessibility. Glycoconj J 15:115–130
- Higgins DG, Thompson JD, Gibson TJ (1996) Using Clustal for multiple sequence alignments. Methods Enzymol 266:383–402
- Hofmann K, Stoffel W (1993) TMbase—a database of membrane spanning proteins segments. Biol Chem 374:166
- Hourcade DE, Wagner LM, Oglesby TJ (1995) Analysis of the short consensus repeats of human complement factor B by sitedirected mutagenesis. J Biol Chem 70:19716–19722

- Jack PL, Klein NJ, Turner MW (2001) Mannose-binding lectin: targeting the microbial world for complement attack and opsonophagocytosis. Immunol Rev 180:86–99
- Jarvis GA, Vedros NA (1987) Sialic acid of group B Neisseria meningitidis regulates alternative complement pathway activation. Infect Immun 55:174–180
- Johnson PT (1969) The coelomic elements of sea urchins (*Strongylocentrotus*). I. The normal coelomocytes: their morphology and dynamics in hanging drops. J Invertebr Pathol 13:25–41
- Jokiranta TS, Hellwage J, Koistinen V, Zipfel PF, Meri S (2000) Each of the three binding sites on complement factor H interacts with a distinct site on C3b. J Biol Chem 275:27657–27662
- Kirkitadze MD, Barlow PN (2001) Structure and flexibility of the multiple domain proteins that regulate complement activation. Immunol Rev 180:146–161
- Kitching IJ, Forey PL, Humphries CJ, Williams DM (1998) Cladistics: the theory and practice of parsimony analysis, 2nd edn. Oxford University Press, New York, N.Y.
- Kozak, M (1987) An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res 15:8125–8148
- Krych M, Hourcade D, Atkinson JP (1991) Regulator of complement activation interact with their ligands via SCRs. Proc Natl Acad Sci USA 88:4353–4357
- Kuhn S, Zipfel PF (1995) Mapping of the complement regulatory domains in the human factor H-like protein 1 and in factor H. J Immunol 155:5663–5670
- Lambris JD (1988) The multifunctional role of C3, the third component of complement. Immunol Today 9:387–393
- Liszewski MK, Atkinson JP (1992) Membrane cofactor protein. Curr Top Microbiol Immunol 178:45–60
- Liszewski MK, Farries TC, Lublin DM, Rooney IA, Atkinson JP (1996) Control of the complement system. Adv Immunol 61:201–283
- Medof EM, Lublin DM, Holers M, Ayers DJ, Getty RR, Leykam JF, Atkinson JP, Tykocinski ML (1987) Cloning and characterization of cDNAs encoding the complete sequence of decayaccelerating factor of human complement. Proc Natl Acad Sci USA 84:2007–2011
- Minta JO, Wong MJ, Kozak CA, Kunnath-Muglia LM, Goldberger G (1996) cDNA cloning, sequencing and chromosomal assignment of the gene for mouse complement factor I (C3b/C4b inactivator): identification of a species specific divergent segment in factor I. Mol Immunol 33:101–112
- Nakao M, Hisamatsu S, Nakahara M, Kato Y, Smith SL, Yano T (2003) Molecular cloning of the complement regulatory factor I isotypes from the common carp (*Cyprinus carpio*). Immunogenetics 54:801–806
- Nielsen H, Engelbrecht J, Brunak S, Von Heijne G (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 1:101–106
- Nixon KC (1999) The parsimony ratchet: a new method for rapid parsimony analysis. Cladistics 15:407–414
- Nonaka M (2001) Evolution of the complement system. Curr Opin Immunol 13:69–73
- Nonaka M, Kuroda N, Naruse K, Shima A (1998) Molecular genetics of the complement C3 convertase in lower vertebrates. Immunol Rev 166:59–65
- Pancer Z (2000) Dynamic expression of multiple scavenger receptor cysteine-rich genes in coelomocytes of the purple sea urchin. Proc Natl Acad Sci USA 97:13156–13161
- Pancer Z, Rast JP, Davidson EH (1999) Origins of immunity: transcription factors and homologues of effector genes of the vertebrate immune system expressed in sea urchin coelomocytes. Immunogenetics 49:773–786
- Pangburn MK, Rawal N (2002) Structure and function of complement C5 convertase enzymes. Biochem Soc Trans 30:1006– 1010

- Rast JP, Amore G, Calestani C, Livi CB, Ransick A, Davidson EH (2000) Recovery of developmentally defined gene sets from high-density cDNA macroarrays. Dev Biol 228:270–286
- Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425
- Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning. A laboratory manual, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
- Schlott B, Wohnert J, Icke C, Hartmann M, Ramachandran R, Guhrs KH, Glusa E, Flemming J, Gorlach M, Grosse F, Ohlenschlager O (2002) Interaction of Kazal-type inhibitor domains with serine proteinases: biochemical and structural studies. J Mol Biol 318:533–546
- Shah M, Brown KM, Smith LC (2003) The gene encoding the sea urchin complement proteins, SpC3, is expressed in embryos and can be upregulated by bacteria. Dev Comp Immunol 27:529–538
- Sikes DS, Lewis PO (2001) Beta software, ver 1. PAUPRat: PAUP* implementation of the parsimony ratchet. Distributed by the authors. Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Conn.
- Smith LC (2001) The complement system in sea urchins. Adv Exp Med Biol 484:363–372
- Smith LC (2002) Thioester function is conserved in SpC3, the sea urchin homologue of the complement component C3. Dev Comp Immunol 26:603–614
- Smith LC, Davidson EH (1994) The echinoderm immune system. Characters shared with vertebrate immune systems and characters arising later in deuterostome phylogeny. Ann NY Acad Sci 712:213–226
- Smith LC, Britten RJ, Davidson EH (1992) SpCoel1: a sea urchin profilin gene expressed specifically in coelomocytes in response to injury. Mol Biol Cell 3:403–414
- Smith LC, Chang L, Britten RJ, Davidson EH (1996) Sea urchin genes expressed in activated coelomocytes are identified by expressed sequence tags. Complement homologues and other putative immune response genes suggest immune system homology within the deuterostomes. J Immunol 156:593–602
- Smith LC, Shih C-S, Dachenhausen SG (1998) Coelomocytes express SpBf, a homologue of factor B, the second component in the sea urchin complement system. J Immunol 161:6784–6793
- Smith LC, Azumi K, Nonaka M (1999) Complement systems in invertebrates. The ancient alternative and lectin pathways. Immunopharmacology 42:107–120
- Sorenson MD (1999) TreeRot, ver 2. Boston University, Boston, Mass.
- Suzuki MM, Sato M, Nonaka M (2002) C6-like and C3-like molecules from the cephalochrodate, *Amphioxus*, suggest a cytolytic complement system in invertebrates. J Mol Evol 54:671–679
- Swofford DL (2002) PAUP*. Phylogenetic analysis using parsimony (*and other methods), ver 4. Sinauer, Sunderland, Mass.
- Terado Ť, Nonaka MI, Nonaka M, Kimura H (2002) Conservation of the modular structure of complement factor I through vertebrate evolution. Dev Comp Immunol 26:403–413
- Terwilliger DP, Clow LA, Gross PS, Smith LC (2000) Expression of SpBf, the homologue of complement factor B, and evidence for alternative splicing in the purple sea urchin, *Strongylocentrotus purpuratus*. Dev Comp Immunol 24:S28
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 24:4876–4882
- Volanakis JE (1998) Overview of the complement system. Immunol Med 20:9–32
- Xu Y, Narayana SV, Volanakis JE (2001) Structural biology of the alternative pathway convertase. Immunol Rev 180:123–135
- Zipfel PF, Jokiranta TS, Hellwage J, Koistinen V, Meri S (1999) The factor H protein family. Immunopharmacology 42:53–60