

University of St Andrews



Full metadata for this thesis is available in
St Andrews Research Repository
at:

<http://research-repository.st-andrews.ac.uk/>

This thesis is protected by original copyright

**Ancestral Genes of the Complement System in the Ascidian,
*Ciona intestinalis***

John Anthony Hammond

University of St Andrews

**A thesis submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy**

September 2002



Th E 326

Declarations

I, John A Hammond, hereby certify that this thesis, which is approximately 40 000 words in length, has been written by me, and that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

date 20/9/02 signature of candidate

I was admitted as a research student in October 1998 and as a candidate for the degree of Doctor of Philosophy in October 1998; the higher study for which this is a record was carried out in the University of St Andrews between 1998 and 2002.

date 20/9/02 signature of candidate

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of Doctor of Philosophy in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

date 20/9/02 signature of supervisor

In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and abstract will be published, and a copy of the work may be made and supplied to any bona fide library or research worker.

date 20/9/02 signature of candidate

Acknowledgements

I would like to thank my supervisors, Dr Valerie Smith and Dr Graham Kemp, for giving me the opportunity to carry out this research and their supervision and support over its duration.

I would like to acknowledge Dr Miki Nakao for his friendship and tuition in the early stages of this work, Dr Masaro Nonaka and Dr Sylvia Smith for their advice concerning the complexities of complement, John Bishop for his gift of *C. intestinalis* larvae and the staff at Croabh Haven for allowing me to collect animals whenever it suited me.

Special thanks are due to members of the Comparative Immunology Group past and present, especially to Dr June Chisholm and Dr Alison Walton who started me off in research; Ralph, for collecting specimens with me all year round and Jorge for coffee breaks. Thanks also to the others in the Gatty who have helped in various ways.

Thanks also to my family for their support and optimism and finally to Jan who has made the last three years so special.

This work was supported by a BBSRC special studentship (Ref. 98/B1/S/04524).

Abstract

Recent identification of complement component homologues in deuterostome invertebrate species has shown that complement has its origins in innate immunity. This study used the ascidian, *Ciona intestinalis*, to show that the complement system is likely to have arisen from a species very close to *C. intestinalis*. Several novel genes were isolated from the same gene families as complement components, some of which are likely to be involved in pathways thought to be ancestral to the complement system. One of these, a serine protease, shares domains with serine proteases of the complement system but has a unique structure. The structure of this protein points to a role in the innate immune system. A thiolester protein gene was isolated that appears to represent a stage in the evolution of the central complement component C3 and its ancestor alpha2-macroglobulin. The domain organisation and motifs within these domains provides evidence that this protein is an important stage in the evolution of the complement system.

Contents

Declarations	i
Abstract	ii
Acknowledgements	iii
Contents	iv
List of figures	ix
List of Tables	xi
Abbreviations	xii
Chapter 1: Introduction	1
1.1 Introduction	2
1.2 Activation of the Mammalian Complement System	3
1.2.1 Activation of the Classical Pathway of Mammalian Complement	4
1.2.2 Activation of the Alternative Pathway of Mammalian Complement	5
1.2.3 Activation of the Lectin Pathway of Mammalian Complement	7
1.2.4 The Lytic Pathway or Membrane Attack Complex	8
1.3 Complement System Activation of the Lower Vertebrates	11
1.3.1 Bony Fish (Osteichthyes)	11
1.3.2 Cartilaginous Fish (Chondrichthyes)	12
1.3.3 Jawless Fish (Cyclostoma)	12
1.4 Invertebrate Complement	15
1.5 Evolution and Structure	19
1.5.1 Gene Duplication	19
1.5.2 The Central C3 Molecule	20
1.5.3 The Serine Protease factor B of the Alternative Activation Pathway	21
1.5.4 Mannan Associated Serine Proteases (MASPs) of the Lectin Activation Pathway	22
1.5.5 Origins of the Complement Molecules	23
1.6 Minimum requirements for a primitive Complement System	27
1.7 Experimental Animal: the Ascidian, <i>Ciona intestinalis</i>	29
1.7.1 The Phylogeny of the Ascidians	29
1.7.2 Phylogeny of <i>Ciona intestinalis</i>	30
1.7.3 Biology of the Ascidians	32
1.7.4 Evidence for Complement in <i>Ciona intestinalis</i>	35

1.8	Specific Aims	37
 Chapter 2: Total RNA isolation		38
2.1	Introduction	39
2.2	Materials and Methods	41
2.2.1	Specimens	41
2.2.2	LPS Treatment of Adult Animals	41
2.2.3	Tissue Preparation and Homogenisation	42
2.2.3.1	Whole Animal	42
2.2.3.2	Hepatopancreas	42
2.2.3.3	Mixed Blood Cells	43
2.2.3.4	Separated Blood Cells	44
2.2.3.5	Larvae	44
2.2.4	Total RNA Extraction from Homogenised samples	45
2.2.5	Total RNA Isolation using Glassmax RNA Micro-Isolation System	46
2.2.6	Quantification of Total RNA	48
2.2.7	Total RNA Quality	48
2.3	Results	50
2.3.1	Larvae	50
2.3.2	Total Cell Population	52
2.3.3	Separated Cell Pools	52
2.3.4	Hepatopancreas	53
2.3.5	Whole Animal	53
2.4	Discussion	57
 Chapter 3: Reverse Transcription-Polymerase Chain Reaction to Amplify Complement Genes		60
3.1	Introduction	61
3.2	Materials and Methods	63
3.2.1	RNA Samples for Reverse Transcription	63
3.2.2	Reverse Transcription	64

3.2.3	Degenerate Primer Design	66
3.2.4	PCR	69
3.2.5	Re-amplification or Nested PCR	70
3.2.6	Analysis of Results	71
3.2.7	Cloning of PCR Amplified Products	72
3.2.8	DNA sequencing	75
3.3	Results	76
3.3.1	Serine Protease RT-PCR for Bf and MASP	76
3.3.2	Thiolester RT-PCR for C3	83
3.4	Discussion	90

Chapter 4: Rapid Amplification of cDNA ends from

	Candidate Complement Gene Fragments	94
4.1	Introduction	95
4.2	Materials and Methods	97
4.2.1	Total RNA Samples	97
4.2.2	5' RACE Template	99
4.2.3	3' RACE Template	99
4.2.4	Gene Specific Primer Design	100
4.2.5	Primary RACE PCR	105
4.2.6	Nested RACE PCR	108
4.2.7	Analysis of Results	108
4.2.8	Cloning and Plasmid Extraction	109
4.2.9	Plasmid Quantification and Quality Control	110
4.2.10	Sequencing	111
4.3	Results	112
4.3.1	SP1 RACE	112
4.3.2	SP2 RACE	113
4.3.3	SP3 RACE	114
4.3.4	SP4 RACE	115
4.3.5	SP5 RACE	117
4.3.6	SP6 RACE	118
4.3.7	SP7 RACE	119
4.3.8	Thiolester 1 RACE	120

4.3.9	Thiolester 2 RACE	122
4.3.10	Summary	122
4.4	Discussion	145
Chapter 5: Bioinformatic Analysis		151
5.1	Introduction	152
5.2	Materials and Methods	154
5.2.1	Similarity Searching	154
5.2.2	Protein Domains and Motifs	155
5.2.3	Further Analysis	158
5.2.4	Multiple Alignment and Phylogenetic Trees	159
5.3	Results	162
5.3.1	Serine protease 1 (SP1)	162
5.3.2	Serine Protease 2 (SP2)	164
5.3.3	Serine Protease 3 (SP3)	166
5.3.4	Serine Protease 4 (SP4)	168
5.3.5	Serine Protease5 (SP5)	170
5.3.6	Serine Protease 6 (SP6)	172
5.3.7	Serine Protease 7 (SP7)	173
5.3.8	Thiolester 1 (Thiol1)	176
5.4	Discussion	186
5.4.1	Serine Proteases	186
5.4.1.1	SP1	187
5.4.1.2	SP2	187
5.4.1.3	SP3	188
5.4.1.4	SP4	189
5.4.1.5	SP5	190
5.4.1.6	SP6	191
5.4.1.7	SP7	191
5.4.2	Thiolester 1	194
Chapter 6: General Discussion		199
6.1	General Discussion	200
6.1.2	Thiolester-Containing Genes	204

6.1.3	Serine Proteases	208
6.2	Future Work	211
6.3	Conclusion	214
Appendices		215
Appendix 1	Degeneracy Code	216
Appendix 2	Oligonucleotide Melting Point Formula	216
Appendix 3	Multiple Alignment of Serine Protease Sequences used to Design Degenerate Primers	217
Appendix 4	Multiple Alignment Thiolester Sequences used to Deign Degenerate Primers	219
Appendix 5	Topo Vector Map	220
Appendix 6	SP1 Multiple Alignment	221
Appendix 7	SP2 Multiple Alignment	224
Appendix 8	SP3 Multiple Alignment	230
Appendix 9	SP5 Multiple Alignment	233
Appendix 10	SP6 Multiple Alignment	235
Appendix 11	SP7 Multiple Alignment	238
Appendix 12	Thiol1 Multiple Alignment	251
References		263

List of Figures

Chapter 1: Introduction

1.1	The Three Activation Pathways of Complement	10
1.2	Phylogeny of Complement Proteins from Different Animal Groups	18
1.3	The Domain Structure of Complement Components	26
1.4	<i>Ciona intestinalis</i>	31
1.5	The Anatomy of <i>Ciona intestinalis</i>	34

Chapter 2: Total RNA Isolation

2.1	Total RNA from Larvae	51
2.2	Total RNA from Hepatopancreas	54
2.3	Total RNA from Whole Animal	55
2.4	Total Hepatopancreas RNA from Stimulated Animals	56

Chapter 3: RT-PCR

3.1	Serine Protease RT-PCR Amplified Products	78
3.2	Serine protease DNA Sequences	80
3.3	Multiple Alignment of the Serine Protease Fragments	82
3.4	Thiolester RT-PCR Amplified Products	85
3.5	Thiolester DNA Sequences	87
3.6	Multiple Alignment of the Thiolester Fragments	88

Chapter 4: RACE

4.1	RACE Primers Based on RT-PCR Fragments	103
4.2	Complete SP1 cDNA Sequence	124
4.3	Complete SP2 cDNA Sequence	126
4.4	Complete SP3 cDNA Sequence	128
4.5	Complete SP4 cDNA Sequence	130
4.6	Complete SP5 cDNA Sequence	133
4.7	Complete SP6 cDNA Sequence	134
4.8	Complete SP7 cDNA Sequence	136
4.9	Complete Thiolester 1 cDNA Sequence	140

Chapter 5: Bioinformatic Analysis		
5.1	Serine Protease 1 (SP1) Domain Schematic	162
5.2	Serine Protease 2 (SP2) Domain Schematic	164
5.3	Serine Protease 3 (SP3) Domain Schematic	166
5.4	Serine Protease 4 (SP4) Domain Schematic	168
5.5	Serine Protease5 (SP5) Domain Schematic	170
5.6	Serine Protease 6 (SP6) Domain Schematic	172
5.7	Serine Protease 7 (SP7) Domain Schematic	173
5.8	Thiolester 1 (Thiol1) Domain Schematic	176
5.9	Multiple Alignment of Thiolester Specificity Defining Residues	181
5.10	Phylogenetic Tree of Thiolester Proteins	183
5.11	Domain Schematic of all RACE Sequences	185
 Chapter 6: Discussion		
6.1	Schematic Domain Diagram of Thiolester Genes	210

List of Tables

Chapter 3: RT-PCR

3.1	Degenerate Primers for the Serine Proteases	68
3.2	Degenerate Primers for the Thiolester Proteins	68
3.3	Serine Protease Primer Success	79
3.4	Thiolester Primer Success	86

Chapter 4: RACE

4.1	RACE Gene Specific Primers	102
-----	----------------------------	-----

Abbreviations

bp	base pairs
Bf	complement factor B
cDNA	complementary DNA
DEPC	diethyl pyrocarbonate
Df	complement factor D
DNA	deoxyribonucleic acid
EDTA	ethylenediamine-tetraacetic acid
EST	expressed sequence tag
HCl	hydrochloric acid
IPTG	isopropyl thiogalactoside
KCl	potassium chloride
LB	Luria Bertani
LPS	lipopolysaccharide
MAC	membrane attack complex
MOPS	3-(N-morpholino)propanesulfonic acid
mRNA	messenger RNA
MS	marine saline
NaCl	sodium chloride
NaOAc	sodium acetate
NP-40	nonident P-40
NUP	nested universal primer
PCR	polymerase chain reaction
RACE	rapid amplification of cDNA ends

RNA	ribonucleic acid
RNase	ribonuclease
rRNA	ribosomal RNA
RT	reverse transcriptase
RT-PCR	reverse transcription-polymerase chain reaction
TEP	thiolester protein
T _m	melting point
UPM	universal primer mix
UTR	untranslated region
UV	ultra-violet
X-Gal	5-bromo-4-chloro-3-indolyl- β -D-galactopyranoside

Chapter 1

General Introduction

1.1 Introduction

The mammalian immune system is very complex containing thousands of different components involved in hundreds of different systems. In the absence of conserved fossil records, investigating the evolution of these components and the systems they work within can only be done using comparative studies. By comparing the defence mechanisms of extant vertebrate and invertebrate species from different phyla, links in the evolution of vertebrate defence systems can be discovered. Identifying and understanding the most simple complement systems allows an insight into the vast range of mechanisms and effects it possesses. This can then provide insights into the extremely complex mammalian system, which is still not completely understood.

Evidence indicates that adaptive immunity occurred during the evolution of the jawed vertebrates as the principal components immunoglobulins (Schluter *et al.*, 1997), the T-cell receptor (Rast and Litman, 1994), major histocompatibility complex classes 1 and 2 (Saltercid and Flajnik, 1995) and recombination-activating genes have been identified only from sharks and higher vertebrates (Kasahara *et al.*, 1997). Complement appears to have more ancient origins as the principal genes, C3 and factor B (Bf), have been identified from the sea urchins (Smith *et al.*, 1998; Gross *et al.*, 1999b) ascidians (Ji *et al.*, 1998; Ji *et al.*, 2000), branchiostomes (Suzuki, 2000; unpublished; Acc No AB050668), and cyclostomes (Nonaka and Takahashi, 1992). Complement has therefore evolved as part of the innate immune system. The genomes of *Caenorhabditis elegans* (nematode) (The *C. elegans* sequencing consortium, 1999) and *Drosophila*

melanogaster (arthropod) (Adams *et al.*, 2000) do not contain any complement genes, so the most primitive complement system appears to function as part of the deuterostomes immune system, post dating the protostomes.

This thesis is concerned with a comparative analysis of the complement system and its evolution from invertebrate origins. The phylogenetically older innate immune system is still of fundamental importance to any animal that has acquired an adaptive immune response (Fearon, 1997a; 1997b; 1999). It is complement that links these systems (Carroll and Prodeus, 1998) and has evolved alongside the most complex immune mechanisms to be of fundamental importance as the major soluble protein effector of both the innate and adaptive immunity (Janeway and Travers, 1996).

Increasing our knowledge of complement activity in the innate immune system is important, as it is this branch of our immune system that we often come to rely upon (Ezekowitz and Hoffmann, 1998; 2001). Any novel pathogen must be dealt with by the innate immune system as no immune memory can exist, which is especially important for juveniles with limited exposure to potential pathogens. Vaccination is fundamental in boosting our own immune response and, although acting on the adaptive immune system, relies on the innate immune system to maintain an effective defence mechanism while the adaptive immune system responds.

1.2 Activation of the Mammalian Complement System

In mammals, the complement system is composed of a highly sophisticated chain of reactions involving a number of components that act to eliminate pathogens by

mechanisms that enable the host to discriminate between non-self and self (Matsushita *et al.*, 1998b). The thirty or more proteins of the complement system are either soluble or membrane bound and interact with each other after activation from various stimuli (Sim and Dodds, 1997). Complement is the major soluble protein effector of both the adaptive and innate immune systems (Janeway and Travers, 1996), although the complement response to challenge by microorganisms can occur before an adaptive immune response has been developed (Sim and Dodds, 1997). Activation of the mammalian complement system initiates many biological processes, including phagocytosis, lysis, inflammation and the regulation of the adaptive immune response (Janeway and Travers, 1996; Sunyer and Lambris, 1998; Lambris *et al.*, 1999). Indeed, a key function of the complement system is bridging the gap between innate and adaptive immunity (Fearon and Locksley, 1996; Carroll, 1998; Carroll and Prodeus, 1998; Fearon, 1999).

1.2.1 Activation of the Classical Pathway of Mammalian Complement

The classical pathway is activated by the binding of IgG, or less frequently IgM, to a pathogen surface (Janeway and Travers, 1996) and is therefore, part of the adaptive immune system. However, many other substances, including some bacteria and viruses, have also been demonstrated to activate this pathway without the need for antibody (Janeway and Travers, 1996).

The classical pathway primarily involves the complement components C1, C4, C2, and C3 (Frank, 1979; Loos, 1982) (Fig 1.1). Three glycoprotein serine protease sub-units form C1; one molecule of C1q binds to the antigen-antibody complex, inducing a

conformational change in C1, activating the C1s and C1r sub-units, of which there are two in every C1 complex (Nicholson-Weller and Klickstein, 1999). Once activated, C1 forms a complex with C4 where C1s cleaves C4 to C4a and C4b (Law and Reid, 1988) (Fig. 1.1). C4b contains an internal thiolester bond that allows covalent attachment to surfaces via their hydroxyl or amino groups (Law and Dodds, 1997). The first amplification stage of complement occurs here as several C4 molecules surround the C1-antibody site as activated C1s cleaves several molecules of C4 (Sunyer and Lambris, 1998). The serine protease C2 then binds to C4b, which surrounds the surface of the antigen-antibody complex, where it is cleaved by C1s to C2a and C2b (Law and Reid, 1988) (Fig. 1.1). C2a contains a catalytic site and so, once bound to C4b, forms the C3 convertase of the classical pathway (C4bC2a) (Janeway and Travers, 1996). C3 is the central molecule of the complement system and also contains a thiolester site. The C3 convertase binds and cleaves C3 into C3a and C3b, where C3b is deposited in large amounts on the pathogen surface by covalent bonding of the thiolester site (Law and Dodds, 1990; 1996; 1997). The microorganism is thus opsonised to facilitate phagocytosis or C5 is cleaved initiating the membrane attack complex (MAC) (Sim and Dodds, 1997).

1.2.2 Activation of the Alternative Pathway of Mammalian Complement

The alternative pathway is rarely activated by antibody (Law and Reid, 1988), being activated in the majority by microorganisms including viruses, bacteria, fungi and protozoans (Law and Reid, 1988) (Sunyer and Lambris, 1998) (Sim and Dodds, 1997) as part of the innate, non-adaptive immune system.

Activation of the alternative pathway relies on the complement components factors D, B, C3 and properdin (Loos, 1982; Janeway and Travers, 1996; Sim and Dodds, 1997; Smith *et al.*, 1999; Song *et al.*, 2000) (Fig. 1.1). The thiolester group in native C3 in serum is susceptible to hydrolysis, producing a low level of active C3 in serum sometimes referred to as C3(H₂O) (Sim and Dodds, 1997). This C3(H₂O) binds to the serine protease, factor B (homologous to C2 in the classical pathway), in the presence of Mg²⁺, becoming proteolytically active (Sim and Dodds, 1997) (Fig. 1.1). The serine protease factor D (Df), cleaves Bf into Ba and Bb, generating the C3 convertase of the alternative pathway C3bBb (Fig. 1.1). This is homologous to the C3 convertase of the classical pathway, C4bC2a (Janeway and Travers, 1996; Sim and Dodds, 1997; Sim and Laich, 2000) (Fig. 1.1). This C3 convertase can activate native C3 into C3a and C3b (Janeway and Travers, 1996). C3b covalently binds to all available surfaces, including those of the host cells as well as activating pathogen surfaces through the thiolester bond, via ester or amide linkages with –OH or –NH₂ groups (Law and Dodds, 1990; 1996; 1997; Sim and Dodds, 1997). Factor I instantly inactivates this C3b created by the alternative pathway if it becomes deposited on host cells (Sim and Dodds, 1997). If it is deposited on an activating surface of a potential pathogen, C3b cleaves factor B and binds Bb (Law and Reid, 1988) (Fig. 1.1). Properdin stabilises bound C3bBb only on activating surfaces by excluding factor I (Sim and Dodds, 1997). Bound and stable C3bBb converts more native C3 in an amplification loop, opsonising the microorganism with a coating of C3b for phagocytosis or initiating the MAC (Sim and Dodds, 1997). This constant activation of C3(H₂O) and its random binding to any available surface has been described as a surveillance mechanism (Sim and Dodds, 1997), in which all materials in contact with blood are tested for their ability to activate complement (Sim and Dodds, 1997). This cascade reaction is further controlled as the C3 convertase is

bound to the pathogen surface containing the reaction locally. Any C3b that is not bound is readily hydrolysed and becomes inactive (iC3 H₂O) (Sim and Dodds, 1997).

1.2.3 Activation of the Lectin Pathway of Mammalian Complement

The lectin pathway relies on mannans or *N*-acetylglucosamine present on the surfaces of many microorganisms (Matsushita, 1996). The complex of mannose-binding lectin (MBL) and mannan associated serine protease (MASP) binds to mannan enabling MASP to directly cleave and activate C2 and C4 (Matsushita and Fujita, 1992) in the same manner as the classical activation pathway (Fig. 1.1). Consequently the C3 convertase of the lectin pathway is formed, causing deposition of C3b on the pathogen surface for phagocytic opsonisation or MAC initiation.

Mannose-binding lectin is a member of the collectin family of lectins, all of which are related to C1q structurally and functionally (Matsushita, 1996). Each of the serum lectins have opsonic activity but only MBL has been shown to activate complement (Kuhlman *et al.*, 1989; Ohta *et al.*, 1990; Matsushita and Fujita, 1992). *In vitro*, MBL can directly activate C1r and C1s from the classical pathway in the same manner as C1q, after activation by mannan-rich surfaces (Lu *et al.*, 1990; Malhotra *et al.*, 1995). Mannose-binding lectin also shares a similar structure to C1q (Lu *et al.*, 1990; Sim and Dodds, 1997), but it is unclear if this interaction of MBL with C1r and C1s occurs *in vivo* as the levels of MBL and MASP are much lower than C1 components. Attempts to isolate MBL in association with C1r and C1s have been unsuccessful (Matsushita, 1996).

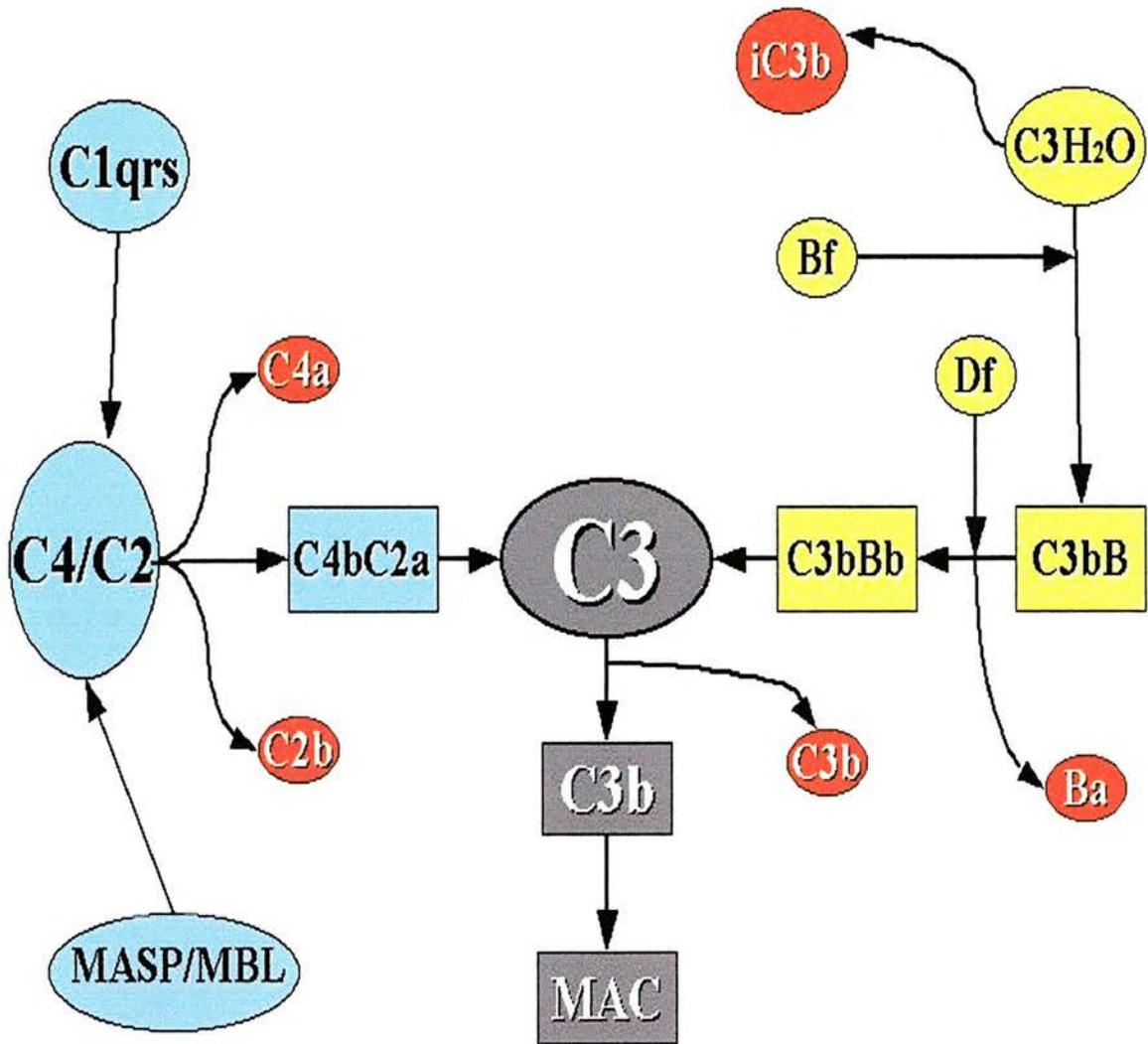
1.2.4 The Lytic Pathway or Membrane Attack Complex (MAC)

All pathways of mammalian complement converge with the production of the C5 convertase (Sunyer and Lambris, 1998) that cleaves C5 to C5a and C5b (Janeway and Travers, 1996). The C3b fragment then initiates the assembly of the MAC. The MAC involves five glycoproteins C5b, C6, C7, C8 and numerous molecules of C9 (Sim and Dodds, 1997). Together these molecules form transmembrane channels in the pathogens, disrupting cellular regulation and causing death (Janeway and Travers, 1996).

C5 is homologous to C3 and C4 but lacks a thiolester site (Law and Dodds, 1990), although its activation occurs in the same way by cleavage to generate C5a and C5b (Janeway and Travers, 1996). C5a is the most potent anaphylatoxin of the complement system (Ember and Hugli, 1997). The cleavage of C5 is the only proteolytic event of the MAC. C5b, loosely bound to C3b, binds to C6 and C7 to form the C5b67 complex with a binding site for membrane surfaces (Janeway and Travers, 1996). This complex disassociates from C3b but, if it does not quickly bind to a membrane surface, its cytolytic activity is lost (Sim and Dodds, 1997). If membrane binding is successful the specific binding site for C8 on the C5b molecule is exposed and binds the three-chain C8 molecule (Sim and Dodds, 1997). This complex is capable of slowly lysing some cells but seems to have a primary function of acting as a receptor for C9 (Janeway and Travers, 1996). The α -chain of C8 binds C9. Once bound, C9 provides high affinity to more molecules of C9 that become inserted in the phospholipid membrane (Janeway and Travers, 1996) forming the pore that leads to cell lysis.

To summarise, each pathway has the ability to interact with another through the serine protease activators and the central C3 molecule as *in vivo* they interact simultaneously (Fig. 1.1). Classical pathway activation occurs as part of the adaptive and innate immune system. The adaptive immune response is activated when the C1 complex binds to antigen-antibody complexes. The C1 molecule and the MBL-MASP complex can bind directly to a microorganism as part of the innate immune response. Alternative pathway amplification can occur *in vivo* via the classical pathway C3 convertase, C4b2a. If C4b2a cleaves native C3 this source of C3b can become bound to factor B, activating the alternative pathway amplification loop. The lectin pathway is only activated by MBL but utilises the same serine protease activators as the classical pathway to generate the C3 convertase. MASP can also cleave native C3 and this active C3b is free to activate the alternative pathway if bound to an activating surface. This ability allows the lectin pathway to influence directly the activation of the alternative pathway.

Figure 1.1 Schematic representation of the three activation pathways of complement and the points at which they interact. The main components of each pathway are labelled. The classical and lectin pathway components are represented by blue, the alternative by yellow. Cleavage products and inactivated products are represented in red. The central C3 molecule which is activated by C3 convertase of each pathway leading to the deposition of C3 on the pathogen surface and initiation of the membrane attack complex is represented in grey.



1.3 Complement System Activation of the Lower Vertebrates

1.3.1 Bony Fish (Osteichthyes)

Complement in teleost or bony fishes has been studied in depth. The classical pathway components C1r (Nakao *et al.*, 2001), C4 (Kuroda *et al.*, 2000) and C3 (Kuroda *et al.*, 2000; Nakao *et al.*, 2000; Zarkadis *et al.*, 2001b) have been isolated. From the innate pathways, the alternative pathway components Bf (Kuroda *et al.*, 1996; Sunyer *et al.*, 1998) and Df (Yano and Nakao, 1994), and the lectin pathway components MASP (Nagai *et al.*, 2000) and MBL (Arason, 1996) (unpublished nucleotide sequence data) have been discovered. Finally, the lytic pathway components C5 (Franchini *et al.*, 2001), C8 (Uemura *et al.*, 1996; Katagiri *et al.*, 1999) and C9 (Tomlinson *et al.*, 1993; Katagiri *et al.*, 1999) have been identified in several different species. For a complete review of the literature describing all the complement components found in the range of teleosts studied see Sunyer and Lambris (1998).

The complement system of bony fish differs from that of mammals in several ways. Teleost fish are ectothermic and rely on their surrounding environmental temperature to maintain their body temperature (Ellis, 1982). Commonly, the water temperature in these environments is much lower than mammalian body temperature, reflected in the optimum and inactivation temperatures of their complement systems. The optimal activation temperature of teleost complement is between 20 and 25 °C in contrast with 37 °C in mammals (Sunyer and Lambris, 1998). Inactivation occurs between 40-45 °C whereas the mammalian alternative pathway is inactivated at 50 °C and the classical pathway is inactivated 56 °C (Sunyer and Lambris, 1998).

The relative importance of the different pathways is reflected in the titre of the molecules of each pathway. Although the classical pathway titres are similar to those of mammals, the alternative pathway components can be up to five times higher (Koppenheffer, 1987). As the antibody response in these organisms is not as advanced as in mammals, the importance of the alternative pathway and antibody independent classical pathway activation may be greater (Nonaka *et al.*, 1981a; Nonaka *et al.*, 1981b).

A unique variance in the complement system of the bony fish is multiple isoforms of C3 and Bf. These are products from different genes that vary in their structure and function (Sunyer and Lambris, 1998). Eight different C3 cDNA clones have been isolated from the carp categorised into five different types (Nakao *et al.*, 2000) and five different forms of C3 have been characterised in the sea bream (*Sparus aurata*) (Sunyer *et al.*, 1997). Several other species are also known to contain multiple forms of C3 and Bf including rainbow trout (Zarkadis *et al.*, 2001b), medaka fish (Kuroda *et al.*, 1996; Kuroda *et al.*, 2000) and the zebra fish (Gongora *et al.*, 1998). Importantly, different isoforms have been found in both diploid and tetraploid fish, so this feature seems to be a phenomenon of all the teleosts and not simply a result of tetraploidy.

The purpose of these different isoforms of C3 and Bf is revealed through the fact they exhibit different binding efficiencies to different activating surfaces (Sunyer *et al.*, 1997). As mammals do not have these different isoforms, it can be postulated their need has arisen because of the less developed adaptive immune system in comparison to mammals. Only one class of antibody is present in fish (IgM), which only appears to be active at temperatures around each species' normal environmental range and not in any

extremes (Ellis, 1982). A range of C3 molecules and a range of their activators could provide a more competent immune reaction able to deal with a wider range of pathogens at lower environmental temperatures.

1.3.2 Cartilaginous Fish (Chondrichthyes)

Early studies on the nurse shark, (*Ginglymostoma cirratum*), (Jensen *et al.*, 1981) revealed a six-molecule functioning complement system. It has been subsequently determined that the shark has all three functioning complement activation pathways after identification of C1q of the classical pathway (Smith, 1998), Bf (Smith and Jensen, 1986) and a putative factor H of the alternative pathway (Smith, 1997) and MASP (gene yet to be cloned) of the lectin pathway (Sunyer and Lambris, 1998). Evidence for a functioning lytic pathway also exists with the isolation of a C8 and C9 (Smith, 1997). Factor B, C3 and C4-like molecules have also been cloned from the Japanese shark, *Triakis* (Takemoto *et al.*, 2000). Adaptive immune molecules appear for the first time in the sharks (section 1.1) (Kasahara *et al.*, 1997) and accordingly sharks appear to contain the earliest example of the classical complement activation pathway.

1.3.3 Jawless Fish (Cyclostoma)

The most primitive vertebrates and the only living class of the Cyclostoma are the agnathans, including the lampreys and the more primitive hagfish. These animals lack the molecular machinery for adaptive molecules (Nonaka *et al.*, 1984) and, consequently, do not possess the classical complement activation pathway (Fig. 1.2). A lamprey C3 homologue has been purified and cloned (Nonaka and Takahashi, 1992) and

has been shown to act as an opsonin (Nonaka *et al.*, 1984). The structure of the lamprey C3 homologue resembles both C3 and C4 indicating a possible early evolutionary origin to the diversification of these thiolester-containing molecules. The lamprey C3 has a three-chain structure similar to C4 (α , β , and γ) but the amino acid sequence shows a greater identity to C3 (Nonaka and Takahashi, 1992). Lamprey MASP shows more identity to the MASP-2 form than MASP-1 (Matsushita *et al.*, 1998b) but no lytic complement components are present (Fujii *et al.*, 1992) (Fig. 1.2).

Hagfish C3 is a two-chain structure (Fujii *et al.*, 1992). Although it has some characteristics of a three-chain molecule, only a two-chain form has been purified. Hagfish C3 functions as an opsonin and is consequently thought to be part of a complement system (Hanley *et al.*, 1992). The lytic pathway also has some representatives in the hagfish. For example, the control protein, CD59, has recently been identified (dos Remedies *et al.*, 1999) in combination with the C5a-like activity observed in the plasma (dos Remedies *et al.*, 1999). Moreover, recent evidence for lymphocyte-like cells (Shintani *et al.*, 2000) indicates the possible presence of a lytic pathway.

The Cyclostoma appear to represent an important intermediate stage in the evolution of the complement system (Fig. 1.2). The lack of antibody and a classical pathway illustrates that the classical pathway was probably the last to develop. The presence of C3, a Bf/C2 and a MASP homologue in this group shows that these pathways have more ancient origins. The Bf/C2 homologue isolated from the lamprey (Nonaka *et al.*, 1994) shows equal identity to both Bf and C2 from higher vertebrates and may represent the predecessor of both these molecules before gene duplication.

1.4 Invertebrate Complement

The first definitive protein homologue isolated from an invertebrate was the C3 homologue SpC3 from a deuterostome, the echinoderm, *Strongylocentrotus purpuratus* (purple sea urchin) (Al-Sharif *et al.*, 1997). This thiolester containing protein has conserved cysteine residues, including those involved in forming the inter-chain disulfide bridge, and two factor I cleavage sites (Al-Sharif *et al.*, 1997). Recently, evidence has been presented that indicates the function of the thiolester site of this protein has been conserved (Smith, 2002). The second component of the sea urchin complement system (SpBf) has also been cloned and is homologous to the C2/Bf family (Smith *et al.*, 1998). A mosaic structure is present like other complement serine proteases but SpBf is unusual as it contains five SCR domain repeats, while all other known Bf/C2 homologues contain three. This may represent the ancestral gene form, as the sea urchins are the most primitive group studied to date. No lectin pathway gene homologues have yet been isolated from echinoderms (Fig. 1.2).

Urochordates occupy a phylogenetic position close to the origin of the vertebrates (Sato and Jeffery, 1995) (Fig 1.2). The complement molecules of three species of ascidian, *Botryllus schlosseri*, *Clavelina picta* and *Halocynthia rorezi* have been closely studied. Strong evidence for complement-regulatory proteins containing SCR domains have been isolated from *B. schlosseri* (Li *et al.*, 2000). A C3 homologue from *C. picta* shows identity to other known C3 sequences (Zarkadis *et al.*, 2001a) and the C3 homologue isolated from *H. rorezi* (AsC3) has all the basic characteristics of C3 with a catalytic

histidine residue and a typical thiolester site (Nonaka and Azumi, 1999; Nonaka *et al.*, 1999).

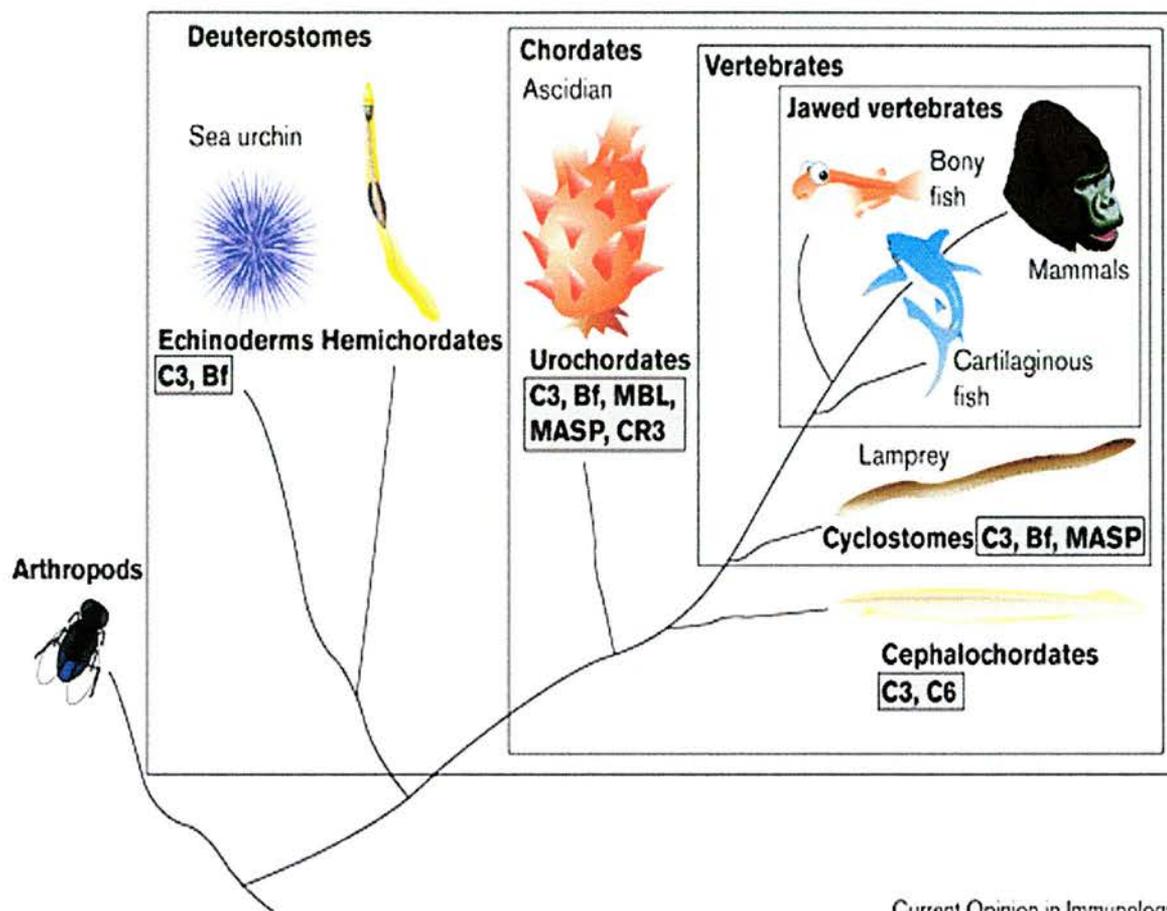
Two MASPs showing similarity to mammalian MASP-1 have also been cloned from *H. rorezi* (Ji *et al.*, 1997), illustrating the likely presence of a lectin pathway in the ascidians. The more recent findings of a MASP-like gene in *C. picta* (Vasta *et al.*, 1999), the cloning of an MBL recognition molecule in *C. picta* (CpMBL) (Vasta *et al.*, 1999), and the reported presence of a C3 receptor in *H. rorezi* (Miyazawa *et al.*, 2001) have reinforced the likelihood of a functioning lectin activation pathway in this group of invertebrates (Fig. 1.2).

Sea urchins occupy a unique phylogenetic position in the deuterostomes between the more ancient protostome arthropods and the deuterostome ascidians (subphylum urochordata). However, a functioning lectin pathway in the echinoderms has as yet to be proven as no MASP or MBL gene homologues have been identified (Fig. 1.2). Both MASP and MBL homologues have been cloned from the urochordates (Quesenberry *et al.*, 1998b; Nonaka and Azumi, 1999) providing strong evidence for a functioning lectin pathway of complement activation in these animals. The lectin and alternative pathways, independent of each other in mammals, may form two linked parts of one activation pathway in the invertebrates (Nonaka, 2001) as several other molecules associated with these pathways emerged in the vertebrate lineage.

From the evidence described above, complement appears to have evolved in the invertebrates as part of the innate opsonic immune system. All of the complement homologues from the invertebrates belong to the same gene super-families as their

vertebrate counterparts. This provides a clue as to the origins of the complement system and the source of its complexity. As the classical pathway relies mainly on activation by antibody (Loos, 1982) and appears to be absent in the cyclostoma, the alternative and lectin pathways appear to have more ancient origins.

Figure 1.2 Representation of a phylogenetic tree showing complement proteins identified in each group. Complement proteins isolated from the jawed vertebrates are not shown. Picture taken from Nonaka (2001), Evolution of the complement system. *Current Opinion in Immunology* 13, 69-73.



Current Opinion in Immunology

1.5 Evolution and Structure

1.5.1 Gene Duplication

The increasing complexity of the complement system from invertebrates to mammals and the homology of the different complement molecules suggest that complement evolved through gene duplication (Nonaka, 2001). This would account for the complex complement cascades found in the higher vertebrates that are composed of molecules belonging to the same gene superfamilies (Bentley, 1988), e.g. the thiolester containing family, the Bf/C2 family and the terminal components family (Gross *et al.*, 1999a). This analysis is based on both sequence and function and the resultant parallels between the activation pathways (Bentley, 1988).

Further evidence can be gained from the presence of ancestral molecules from the same gene families and homologous protein domains in the protostome lineage (Fig. 1.3). Duplications may have been an independent event for each component or may have happened simultaneously during genome duplication (Nonaka, 2001). This tetraploidization is believed to have occurred twice during the early stages of vertebrate evolution (Ohno, 1999). This phenomenon would explain the precise one-to-one relationship of the classical pathway with the alternative and lectin pathways (Nonaka, 2001): C1q with MBL; C1r and C1s with MASP-1 and MASP-2; C2 with Bf and C4 with C3 (Nonaka *et al.*, 1998).

1.5.2 The Central C3 Molecule

The central complement molecule, C3, is homologous to both C4 and C5 of the complement system (Alsenz *et al.*, 1992) and all three appear to have evolved from the serine protease inhibitor α 2-macroglobulin (Sottrup-Jensen *et al.*, 1985). The plasma protein, α 2-macroglobulin, functions as part of the innate immune system in many taxa from humans (Dodds and Law, 1998) to arthropod invertebrates (Asokan *et al.*, 2000) as it binds to proteases of endogenous and exogenous origin (Armstrong and Quigley, 1999). An α 2-macroglobulin homologue has also been reported in the ancient invertebrate phylum Cnidaria (Dishaw *et al.*, 2000).

These four molecules (α 2-macroglobulin, C3, C4 and C5) thus belong to the same α 2-macroglobulin gene family, although it is still unclear as to which complement molecule diverged from α 2-macroglobulin first (Hughes and Yeager, 1997). Mammalian C3 is formed from a single pro-C3 polypeptide that is cleaved at the β - α processing site (Law and Dodds, 1997) (Fig. 1.3). Disulfide bonds join the resulting α and β chains, and the C3a anaphylatoxin peptide is cleaved from the C terminal of the α chain, yielding the final two-chain molecule known as C3b (Hughes, 1994) (Fig. 1.3). Cleavage of the C3a exposes the thiolester site on C3b allowing the covalent attachment of C3b to foreign surfaces (Sottrup-Jensen *et al.*, 1981).

All the C3 homologues share this basic structure with only small variations, all having at least a two-chain structure (Fig. 1.3). Differences occur in the structure of the C3a

region, conserved in the lamprey and all other vertebrates, but not in the invertebrate homologues (Nonaka, 2001). Therefore, only the vertebrate lineage has an anaphylatoxin active C3a fragment.

1.5.3 The Serine Protease Factor B of the Alternative Activation Pathway

The serine proteases Bf and C2 both belong to the same gene family and share a similar structure of (from the N-terminus) three short consensus repeat (SCR) modules, a von Willebrand factor domain and a serine protease domain (Ishikawa *et al.*, 1990) (Fig. 1.3). Both these molecules play crucial roles as the proteolytic subunits of the C3 convertases of the alternative and classical pathways respectively (Nakao *et al.*, 1998). As Bf homologues have been discovered in the agnathans (Nonaka *et al.*, 1994), urochordates (Ji *et al.*, 2000) and echinoderms (Smith *et al.*, 1998), the innate alternative pathway of activation is likely to be present in these animals. Serine protease Bf is thought to have derived from the common ancestor to both Bf and C2 (Nonaka, 2001). As Bf is part of the older alternative pathway, in comparison to the more recent adaptive classical immune pathway, it is recognised as more ancient.

Both the Bf molecules discovered in the sea urchin (SpBf) (Smith *et al.*, 1998) and the ascidian (AsBf) (Ji *et al.*, 2000) differ from the common structure found in the jawed vertebrate homologues: SpBf cloned from *S. papuratus* has two extra SCR domains (Smith *et al.*, 1998) (Fig. 1.3) and AsBf from *H. rozezi* also has two additional SCR domains as well as three extra low-density lipoprotein receptor domains (Ji *et al.*, 2000) (Fig. 1.3). These findings indicate that the ancestral form of Bf was likely to have these additional domains that were lost in the jawed vertebrate lineage. Additionally,

extensive exon shuffling appears to have taken place at an early stage in the evolution of the complement system (Nonaka, 2001). No investigations have been carried out to determine the possible functional significance of these structural differences.

1.5.4 Mannan-Associated Serine Proteases (MASPs) of the Lectin Activation Pathway

The MASPs are proteolytic enzymes that activate C3 through the lectin pathway. Human MASP has two polypeptides; a heavy chain and a light chain joined by disulfide bonds as in C1r and C1s. Human MASP has a similar structure to C1s but contains the unique ability to activate native C3 in serum (Matsushita and Fujita, 1995), which results in the activation of the alternative pathway. In humans there are at least four different MASPs, MASP-1 (Sato *et al.*, 1994), MASP-2 (Thiel *et al.*, 1997), MASP-3 (Dahl *et al.*, 2001) and sMAP/Map19 (Takahashi *et al.*, 1999). Map19 is a form of MASP-2 truncated at its carboxyl terminus and is not a serine protease. Although the difference in function between the different MASPs has not been entirely elucidated, it appears that each of the isoforms has differing ability to cleave C2 and C4 (Sim and Dodds, 1997).

Both MASP-1 and MASP-3 from humans are alternatively spliced from the same gene (Dahl *et al.*, 2001). The non-truncated serine protease MASPs from vertebrates contain six domains: domain one and three are bone morphogenic protein domains (CUB) homologous to each other: Domain two is an epidermal growth factor domain (EGF): Domains four and five are short consensus repeat domains (SCR): Domain six is the serine protease domain (Matsushita *et al.*, 1998b) (Fig. 1.3). The serine protease domains vary in the codon for the active serine. The most common codon is TCN

(where N is A, G, C, or T) (Appendix 1), but the amino acid serine can also be coded for by AGY (where Y is C or T) (Appendix 1). Human MASP-2 and MASP-3 contains the AGY codon rather than the TCN codon for MASP-1.

Endo *et al.* (2000) isolated MASP cDNA from five species of vertebrate including two MASPs from an euteleostomi amphibian, *Xenopus laevis*, and one from the chondrichthian, *Triakia scyllium*. Only one MASP has been isolated from the lamprey (Nonaka and Takahashi, 1992) and this contains the more unusual AGY serine protease codon. The resulting molecular analysis of these and all the known MASP homologues indicates two lineages of the MASP genes; one of which diverged from an ancestral molecule before the vertebrate lineage (Endo *et al.*, 2000).

Both the ascidian MASPs (AsMASPa and AsMASPb) (Fig. 1.3) code for the TCN serine (Ji *et al.*, 1997) and show most identity to human MASP-1 (Matsushita *et al.*, 1998b). They probably arose from a single common ancestor as they are the most ancient molecules of the lectin pathway so far discovered. As yet, no MASP gene has been located in the echinoderms. However, the presence of two isoforms in ascidians shows the lectin pathway also has ancient origins in the invertebrates.

1.5.5 Origins of the Complement Molecules

In humans, there are evolutionary relationships, including activation, between the complement, kinin, coagulation and fibrinolytic systems *in vitro* (Sundsmo and Fair, 1983). The biological significance of their interactions is unclear. Protostomes and deuterostomes contain several defence pathways analogous to complement e.g.

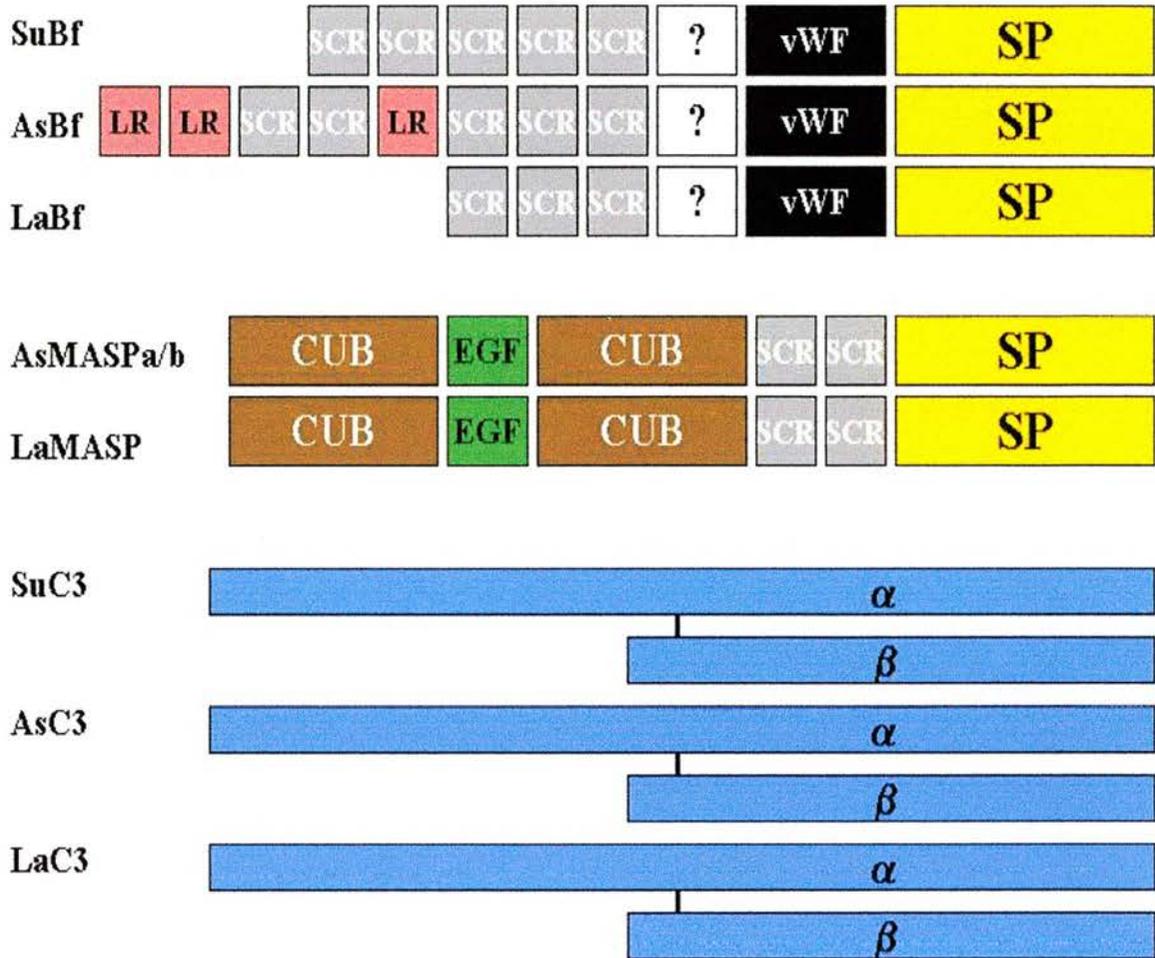
prophenoloxidase pathway (Sundsmo and Fair, 1983), showing that early vertebrates have the cellular complexity and molecular mechanisms to support such systems. These invertebrates have clotting systems (Bohn, 1986) and encapsulation systems (Götz, 1986). Parts of the prophenoloxidase pathway are also present in both protostomes and deuterostomes and is analogous to complement in terms of its activation, cascading amplification and cleavage by serine proteases (Smith, 1996), allowing speculation that mechanisms exist to activate and control a complex immune pathway, such as complement (Cerenius and Söderhäll, 1995).

The plasma protein, $\alpha 2$ -macroglobulin, a conserved arm of the innate immune system and ancestor of C3, is present in many protostome invertebrate groups (Armstrong and Quigley, 1999). The close relationship of $\alpha 2$ -macroglobulin to C3 allows speculation about the role that this molecule played in the evolution of C3 (Sottrup-Jensen *et al.*, 1985). As a serine protease inhibitor, $\alpha 2$ -macroglobulin has the ability to attach to enzymes that cleave it in the same region as C3 (Dodds and Day, 1993). This has led to speculation that apart from an inhibitory function in vertebrates, in invertebrates this molecule can attach to and inhibit novel proteases (Armstrong and Quigley, 1999). A major source of these novel proteases would be from foreign microorganisms linking $\alpha 2$ -macroglobulin to a role in defence (Armstrong and Quigley, 1999). An $\alpha 2$ -macroglobulin receptor has been found and appears unrelated to any of the C3 receptors recognised so far (Walport, 1996). Although it does not appear that this receptor diverged at the same time as the complement components arose from the protease inhibitors, C3 has at least 4 different receptors all with a different function on various cell types (Walport, 1996). However, it is likely that $\alpha 2$ -macroglobulin has other receptors as yet unknown (Dodds and Day, 1993). This combination of molecular and

functional evidence has led to $\alpha 2$ -macroglobulin being considered the evolutionary precursor molecule of C3.

Several other proteins containing domains important in the activation and regulation of complement are also known to be present in several protostome invertebrate species (Iwanaga *et al.*, 1998; Pahler *et al.*, 1998). Lectins, involved in recognition of self and non-self, are present throughout the metazoa (Olafsen, 1986) and can play a key opsonic role in organisms without a true complement system (Hardy *et al.*, 1977). However, complete complement proteins have been isolated only from the deuterostome lineage (Al-Sharif *et al.*, 1997; Ji *et al.*, 1997; Smith *et al.*, 1998; Nonaka and Azumi, 1999; Nonaka *et al.*, 1999; Vasta *et al.*, 1999; Li *et al.*, 2000; Zarkadis *et al.*, 2001a) and the genomes of *Caenorhabditis elegans* (nematode) (The *C. elegans* sequencing consortium, 1999) and *Drosophila melanogaster* (arthropod) (Adams *et al.*, 2000) do not contain any complement genes. Complement postdates the protostomes lineage.

Figure 1.3 Schematic diagram adapted from Nonaka (2001) illustrating the domain structure of the complement components of lower vertebrates and the deuterostomes: Su, sea urchin; As, ascidian; La, lamprey. SP, serine protease domain; vWF, von Willebrand Factor type A domain; SCR, short consensus repeat; LR, low density lipoprotein receptor class A domain; CUB, C1r/C1s/uEGF/bonemorphogenic protein domain; EGF, epidermal growth factor. C3 proteins represented in blue with the alpha (α) and beta (β) chains.



1.6 Minimum Requirements for a Primitive Complement System

From the evidence presented, it can be postulated what the minimum requirements are likely to be for the most primitive complement system and in which taxa these might be found;

- 1, a C3-like protein with a two or three chain structure containing a thiolester site
- 2, a Bf/C2-like protein containing SCRs and a serine protease domain
- 3, a receptor for complement on phagocytic cells
- 4, a regulatory molecule to protect host cells from complement attack.

Such a system is likely to be present early in the deuterostome lineage as no complement components have been found in the protostome line (Zarkadis *et al.*, 2001a). Several ascidians have been shown to possess complement systems, but the most primitive organism studied has been the sea urchin, *S. purpuratus* (Smith *et al.*, 1998; Gross *et al.*, 1999b). The phylogenetic relationship of the echinoderms to the other deuterostome groups (hemichordates, chordates and chaetognaths) based on their pre and post developmental features is unclear, although it is generally accepted that these groups are very closely related because they share many striking features (Barnes *et al.*, 1993). However, there is some doubt as to the position of the chaetognaths in the deuterostome group (Barnes *et al.*, 1993; Wada and Satoh, 1994; Satoh and Jeffery, 1995). Phylogeny based on molecular evolution of 18s RNA (Field *et al.*, 1988) has confirmed the position of these groups, but again, could not resolve the relationship between them.

From the evidence detailed above, the hypothesis for this study is that the more ancient alternative and lectin complement activation pathways exist in the urochordates. The proteins that make up these pathways will be more ancient than the complement homologues so far discovered. As no lectin pathway components have been isolated from the echinoderms (Fig. 1.2), the origin of this pathway is likely to exist in the ascidians. Components of the alternative pathway have been discovered in the echinoderms (Al-Sharif *et al.*, 1997; Smith *et al.*, 1998) (Fig. 1.2) but the phylogeny of this group is unclear. Discovery of complement components from a more ancient species in the urochordates than those studied to date would provide evidence for the evolutionary genetics of the complement system, and the evolutionary positions of the deuterostomes groups.

1.7 Experimental Animal: the Ascidian, *Ciona intestinalis*

1.7.1 Phylogeny of the Ascidians

The aim of this study is to isolate complement molecule homologues in the ascidian, *Ciona intestinalis* (Fig. 1.4). As previously mentioned, ascidians are part of the sub-phylum Urochordata and anatomical, embryological, developmental and molecular phylogenetic investigations place the ascidians close to the origin of the vertebrate line (Jefferies, 1986; Field *et al.*, 1988; Morris, 1993; Satoh and Jeffery, 1995) (Fig. 1.2). It is thought that a form resembling the ascidian tadpole was the foundation of the phylum Chordata (Rogers, 1986; Satoh and Jeffery, 1995). This tadpole-like larva had a notochord, a dorsal nerve cord and pharyngeal gill slits (Jefferies, 1986; Satoh and Jeffery, 1995); all of which are features of the chordates.

A study of the 18s rDNA sequences of urochordates and other protostome and deuterostome species suggests that the ancestors of the vertebrates were from free-living rather than sessile animals (Wada and Satoh, 1994); a theory first proposed in 1971 (Tokioka, 1971). The chordate lineage is believed to have emerged from the larvaceans, followed by sessile ascidians and then salps (Wada and Satoh, 1994). Several other slight differences have been proposed but the ascidians have a key place in them all (Satoh and Jeffery, 1995), establishing them close to the origin of the vertebrate line.

1.7.2 Phylogeny of *Ciona intestinalis*

Several ascidian species have previously been studied in the pursuit of invertebrate complement molecules (Ji *et al.*, 1997; Ji *et al.*, 1998; Nonaka and Azumi, 1999; Azumi *et al.*, 2000; Li *et al.*, 2000; Nair *et al.*, 2000; Miyazawa *et al.*, 2001; Sekine *et al.*, 2001) due to their strategically important phylogenetic position. A principal reason why this study is focused specifically on *C. intestinalis* is because the family Cionidae, to which *C. intestinalis* belongs, is considered to be the most primitive form of surviving ascidians (Berril, 1936; Jefferies, 1986) (Fig. 1.3). Therefore, *C. intestinalis* is probably a more ancient ascidian than the other species from which complement homologues have been identified, e.g. *Halocynthia rorezi* (Nonaka and Azumi, 1999). *C. intestinalis* complement homologues discovered in this study would be the most ancient isolated from the urochordates and represent a very ancient complement gene. The study of an ancient species of ascidian will also allow some comparison with SpC3 and SpBf from the echinoderms (Al-Sharif *et al.*, 1997; Smith *et al.*, 1998) in order to ascertain which contains the most ancient complement gene and thereby resolving some phylogenetic queries about the emergence of the echinoderms and the urochordates.

Figure 1.4 Picture of *Ciona intestinalis* taken by Clare Peddie in 1995 at Croabh Haven marina, Argyll, Scotland.



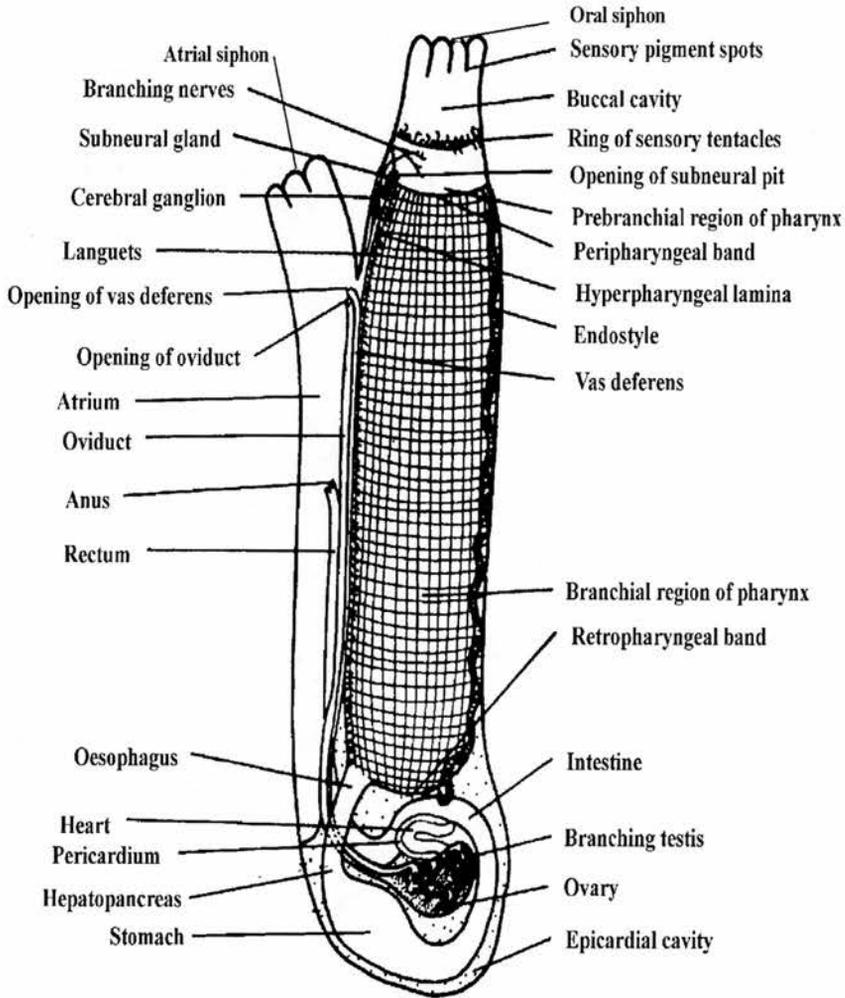
1.7.3 Biology of the Ascidians

Ascidians are in the class Ascidiaceae within the phylum Chordata. *C. intestinalis* is in a member of the family Cionidae, sub-order Aplousobranchia, order Enterogona, class Ascidiaceae, sub-phylum Urochordata, phylum Chordata (Hayward and Ryland, 1995). The biology of *C. intestinalis* has been well characterised (Millar, 1971; Svane and Havenhand, 1993; Hayward and Ryland, 1995) (Fig. 1.5) and a brief summary follows.

Ascidians are all marine with sessile adults forming being either compound or solitary. Most adults are plankton filter feeders and the larvae are non-feeding. The larvae are motile through a tail and can survive from only a few hours to several days in the water column. The adult has no coelomic body cavity, segmentation or bony tissue. Beneath the external surface, a gelatinous tunic containing cellulose, are two distinct body regions. The abdominal area contains a heart, gut and gonads (Fig. 1.5). The anterior pharyngeal region contains the pharyngeal basket into which plankton is passed after feeding (Fig. 1.5). Plankton is filtered from the seawater by gill clefts in the pharyngeal region as water enters the oral siphon, passes over the pharynx and exits via the atrial siphon (Fig. 1.5). A peristaltic heart that can periodically change the direction of the blood flow drives the circulatory system (Fig. 1.5). Blood is driven through two vessels at each end of the heart that branch throughout the body and tunic. These vessels are not closed in the periphery and cells are known to migrate from the blood vessels and move directly through tissues.

C. intestinalis is distributed around the entire British coast, and in many other parts of the world, occupying submerged surfaces between the lower shore and 500 metres (Millar, 1971). Colonies can reach densities of several thousand individuals per m² (Svane and Havenhand, 1993). This species is oviparous, and either spawns freely or through very adhesive mucus strings (Svane and Havenhand, 1993) into the water column. The eggs are fertilised externally and the free-swimming larvae develop and hatch (Hayward and Ryland, 1995).

Figure 1.5 The anatomy of *Ciona intestinalis* (after Bullough 1958)



1.7.4 Evidence for Complement in *Ciona intestinalis*

Although a complement system has not been discovered in this species of ascidian, many investigations have analysed the range of immune reactions possessed by this organism. *C. intestinalis* contains a range of antimicrobial factors (Findlay and Smith, 1995), cytotoxic activity (Peddie and Smith, 1993; 1994a), opsonins (Smith and Peddie, 1992) and at least two components of the phenoloxidase system (Jackson *et al.*, 1993). In combination these produce an efficient humoral defence strategy (Smith and Söderhäll, 1991).

If a functioning complement system were operating in *C. intestinalis*, characteristic immune reactions would be observed. The most important of these is phagocytosis as complement is thought to have evolved primarily as a humoral opsonin to aid pathogen recognition by phagocytes. The vacuolar and granular amoebocytes are phagocytic against bacteria (Smith and Peddie, 1992), and this phagocytic rate increases if the bacteria are incubated with *C. intestinalis* blood cell lysate supernatant (Smith and Peddie, 1992). This phenomenon is abolished if serine protease inhibitors are included (Smith and Peddie, 1992). Protease activity increases in the blood cells after incubation with lipopolysaccharide (Jackson and Smith, 1993) and phagocytic rates increase if the lysate is from cells previously incubated with lipopolysaccharide (Smith and Peddie, 1992). This provides some functional evidence that an opsonic system is present in *C. intestinalis* that relies upon serine proteases for activation, a fundamental process in the activation of complement.

C. intestinalis displays an inflammatory-like reaction in the tunic in response to injury (DeLeo *et al.*, 1996; DeLeo *et al.*, 1997; Di Bella and De Leo, 2000). Such responses are a known function of the anaphylatoxin fragments of several of the complement proteins including C3 and C5 (Morgan and Gasque, 1996; Smith *et al.*, 1997; Franchini *et al.*, 2001). Once the cells have migrated, they form a capsule around the wound and a clotting mechanism is initiated (DeLeo *et al.*, 1997). Clotting pathways are analogous to complement activation mechanisms involving serine proteases, and possibly related to the ancestral molecules of the complement system (Iwanaga, 1989; Opal, 2000).

Investigations into the immune systems of other ascidians have shown that they too rely on opsonic systems for enhanced phagocytic rates (Kelly *et al.*, 1993a; 1993b; Ballarin *et al.*, 1994; Ohtake *et al.*, 1994; Ballarin *et al.*, 1999; Nonaka and Azumi, 1999; Nonaka *et al.*, 1999; Azumi *et al.*, 2000; Ballarin *et al.*, 2000). Several of these ascidian opsonic systems have also been shown to rely, in part, on a complement system (Nonaka and Azumi, 1999; Azumi *et al.*, 2000) or on complement-like activating molecules (Kelly *et al.*, 1993a; Ballarin *et al.*, 1999).

Considering this immune evidence from *C. intestinalis* and other ascidians, a functioning opsonic complement system is likely to be present in *C. intestinalis*. A complement molecule from *C. intestinalis* would represent one of the most ancient complement components providing a powerful tool in analysing the evolution of the complement system.

1.8 Specific Aims

To determine if a complement system is present in *C. intestinalis* this study will investigate the following questions.

1. Is any of the mRNA in *C. intestinalis* transcribed from complement-like genes from the alternative and lectin activation pathways or their ancestors?
2. If there are any complement-like genes, what is the likelihood they are part of a functioning complement system?

Chapter 2

Total RNA isolation

2.1 Introduction

The isolation of stable and competent ribonucleic acid (RNA) from animal tissue is fundamental for the isolation and characterisation of transcribed genes. Gene expression comes from genomic DNA through a messenger RNA (mRNA) to form a functional polypeptide. RNA isolation is the obvious starting point for this study as the main aim is to look for expressed genes coding for complement proteins.

Expression of genes can be tissue specific within the organism or developmental stage. It was first proposed in 1928 (Garstang, 1928) that larvae from urochordates gave rise to the vertebrate line through paedogenesis; a view of the ancestry of the first vertebrates that is now widely accepted. All the invertebrate complement gene homologues isolated so far have been isolated from adult tissue, although this final developmental stage may not have been represented during the evolution of the Chordata. In the present study, RNA was isolated from adult and larval tissue to allow comparison between the adult and an evolutionarily key developmental stage.

Complement gene fragments isolated by reverse transcriptase-polymerase chain reaction (RT-PCR) in other invertebrates have come from a range of adult tissues. For example, C3 and two MASP gene fragments from the Japanese ascidian, *H. rorezi*, were discovered from cDNA synthesised with RNA extracted from the hepatopancreas (Nonaka and Azumi, 1999). In the sea urchin, *S. purpuratus*, C3 is expressed in the blood cells (Clow *et al.*, 2000) while a Bf homologue was isolated in cDNA synthesised from a mix of cellular and adult tissue RNA (Smith *et al.*, 1998). Accordingly, total RNA from different tissues was extracted from various stages in the life cycle of *C.*

intestinalis. This could then be used in RT-PCR and the rapid amplification of cDNA ends (RACE) directly, or mRNA could be purified from the total RNA if necessary.

It has been well recorded that any animal's immune system can be stimulated by lipopolysaccharide (LPS) from microbial origin. The use of LPS as an immune stimulant for invertebrates has been well documented for crustaceans (Söderhäll, 1982; Söderhäll and Hall, 1984; Hammond and Smith, 2002), for ascidians (Smith and Davidson, 1992; Jackson and Smith, 1993) and for sea urchins (Smith *et al.*, 1995). Indeed, in the sea urchin, *S purpuratus*, expression of the complement homologue SpC3 is increased after stimulation with LPS (Clow *et al.*, 2000). Subsets of animals were stimulated using LPS before removal of tissue and extraction of RNA to provide comparison with tissues extracted from non-stimulated animals in the present study.

The aim of this chapter is to compare different methods of total RNA extraction to find the best protocol for each tissue type. Two basic methods of total RNA extraction are commonly used, those that begin with a gentle membrane solubilising buffer (e.g. hypotonic NP-40 based buffers) and those that begin with a chaotropic (biologically destructive) agent that disrupts cell and organelle membranes while simultaneously inactivating RNases (e.g. guanidine isothiocyanate) (Farrell, 1998). In this study, total RNA was extracted by using chaotropic agents. The advantages of these are their speed, immediate RNase inactivation and yield of high quality RNA (Sambrook and Russell, 2001).

2.2 Materials and Methods

2.2.1 Specimens

Adult *C. intestinalis* were collected from Croabh Haven Marina, Argyll, Scotland. Specimens 5 - 10 cm high were taken off the underside of floating pontoons by hand and transported to St Andrews in seawater from the site, maintained at the same temperature. On arrival they were transferred to aquaria with flow through seawater (32 ± 2 ‰; 10 ± 2 °C) pumped directly from St Andrews Bay, St. Andrews, Scotland. All samples were taken from these animals within 72 h of their collection.

2.2.2 LPS Treatment of Adult Animals

LPS (from *E. coli* serotype 0111:B4) (SIGMA, Poole, Dorset, UK) was suspended in marine saline (MS) (940 mOsm kg^{-1}) (12 mM $\text{CaCl}_2 \cdot 6\text{H}_2\text{O}$; 11 mM KCl; 26 mM $\text{MgCl}_2 \cdot 6\text{H}_2\text{O}$; 45 mM Tris; 38 mM HCl; 0.4 mM NaCl; pH 7.4) (Smith and Peddie, 1992) at two stock concentrations of $100 \mu\text{gml}^{-1}$ and $10 \mu\text{gml}^{-1}$. Specimens of *C. intestinalis* were given 25 μl doses of LPS from a 1 ml syringe fitted with a 23 G (0.5 x 16 mm) needle. All injections were made through the outer and inner tests into the epicardial cavity below the stomach and intestine (Fig. 1.5). The stomach is visible through the tests and this small area is free from any organs so no damage was inflicted to the animal besides the puncture wound.

Groups of 5 animals were given doses of LPS at $100 \mu\text{gml}^{-1}$ or $10 \mu\text{gml}^{-1}$, corresponding to 2.5 μg or 0.25 μg LPS per animal respectively. Control animals were given an

equivalent volume injection of sterile MS. Each treatment group was incubated in flow through seawater ($32 \pm 2\%$; 10 ± 2 °C) tanks for 3 h, 12 h or 24 h. Tissues were then removed after the appropriate incubation time using the methodologies described below.

2.2.3 Tissue Preparation and Homogenisation

2.2.3.1 Whole Animal

The outer tests were completely removed to expose the contents of the epicardial cavity. The intestine and stomach were then removed from just below the hepatopancreas to the anus. This was necessary to ensure that there was no contamination of the sample with material filtered by the ascidian from the seawater. The remaining tissues were then frozen in liquid nitrogen before being ground into a fine powder with a sterile pestle and mortar. Due to the large amount of material, it was homogenised in a sterile 50 ml falcon tube containing 7.5 ml TRIzol® or TRIzol® LS by pulse vortexing at 22 °C. Aliquots of this (approx. 0.85 ml) were dispensed into 1.5 ml Eppendorf tubes. This increased the sample processing speed using a faster 1.5 ml Eppendorf centrifuge.

2.2.3.2 Hepatopancreas

Fresh and healthy specimens of *C. intestinalis* were removed from seawater and placed immediately on ice. A longitudinal incision was made along the length of the body to peel back and remove the outer and inner tests. A cut into the epicardial cavity allowed the contents, including the hepatopancreas, to extrude. The hepatopancreas was then removed from above the intestine and snap frozen in liquid nitrogen. Once frozen, the

hepatopancreas was placed in a sterile mortar and ground into a fine powder, adding more liquid nitrogen if necessary. This was then transferred into a sterile 1.5 ml Eppendorf tube containing 0.75 ml of TRIzol®/TRIzol® LS and pulse vortexed until completely homogenised at 22 °C.

2.2.3.3 Mixed Blood Cells

Animals were removed from aquaria and immediately placed in ice. An incision was made through the outer and inner tests to reveal the mantle. The mantle (a clear sack containing the circulatory, reproductive and digestive organs) was gently squeezed through the first incision, and another small incision made through the mantle into the epicardial cavity revealing the heart. Blood was withdrawn directly from the heart by piercing and allowing it to drip into a sterile 1.5 ml Eppendorf containing 0.5 ml of sterile 0.45 µm filtered ice-cold marine anticoagulant (MAC) (985 mOsm kg⁻¹) (0.1 M glucose; 15 mM trisodium citrate; 13 mM citric acid; 10 mM EDTA; 0.45 mM NaCl; pH 7.0) (Smith and Peddie, 1992). Blood was collected from approximately 30 animals (usually 20-30 µl per animal) until a total volume of 1 ml diluted blood was reached (i.e. blood 1:1 MAC). The cells were pelleted by centrifugation at 800 g for 10 minutes at 4 °C. After removal of the supernatant, a small sub-sample of the pellet was taken and re-suspended in MS for determination of cell viability by eosin dye exclusion scrutinised under a Leitz Diaplan microscope. Any samples containing less than 80 % viable cells were discarded. The remaining cells were completely homogenised in 0.75 ml TRIzol® or TRIzol® LS by pulse vortexing for 5 min at 22 °C.

2.2.3.4 Separated Blood Cells

Approximately 2.5 ml of fresh blood/MAC dilution (prepared as above in section 2.2.2.3) were laid onto a continuous gradient of 60 % Percoll (Pharmacia, Upsala, Sweden) in 3.2 % NaCl, preformed at 42000 g for 20 min at 4 °C (Smith and Söderhäll, 1991). These gradients were spun at 1900 g for 10 min at 4 °C and the cells in each of the resulting 6 bands that were removed using sterile Pasteur pipettes. These were classified using the criteria in Smith and Peddie (1992) and Rowley (1981, 1982). Briefly these were; band 1 (top), signet ring cells; band 2, hyaline leucocytes; band 3, phagocytic amoebocytes; band 4, morula cells; band 5, pigment cells; band 6 (bottom), stem cells. Equivalent bands from all the gradients were pooled and washed twice in x 10 volume MS to remove any Percoll and cellular debris (Smith and Peddie, 1992) by centrifugation at 200 g for 5 min at 4 °C. A sub-population of each cell band was removed for cell viability assessment as in section 2.2.3.3. The final cell pellet was transferred to a 1.5 ml Eppendorf tube and homogenised in 0.75 ml TRIzol® LS by pulse vortexing for 5 min at 22 °C. To obtain any cell pellets for all 6 cell bands approximately 1200 animals were bled.

2.2.3.5 Larvae

Five adult *C. intestinalis* specimens producing gametes were collected from Sutton Harbour Marina, Plymouth, UK, in the same manner described in 2.2.1, and left to spawn in a petri dish at 10 °C. Eggs and zygotes were pipetted into 800 ml of 0.45 µm filtered seawater at 16 °C from fertilisation and development to hatching. The proportion of hatched larvae was determined every 12 h under a dissecting microscope by taking a

well mixed sub-sample and counting a minimum of 200 individual embryos per larvae. Once 60 % were hatched, the upper portion of the sample containing the active swimming larvae was removed with a sterile 10 ml pipette (approximately 250 ml) leaving undeveloped embryos in the seawater. The concentration of actively swimming larvae in this 250 ml was $7.76 \text{ larvae ml}^{-1}$ determined by visual discrimination under a dissecting microscope. These larvae were then filtered through plankton gauze ($40 \mu\text{m}$) (Lockertex, Warrington, UK) removing them from suspension before rinsing with seawater to ensure a clean sample. Immediately after rinsing, larvae were removed from the gauze using a sterile scalpel into a tube containing 5 ml TRIzol® LS, and homogenised by pulse vortexing for 5 min at $22 \text{ }^\circ\text{C}$.

2.2.4 Total RNA Extraction from Homogenised Samples

The following protocol has been adapted from the product inserts of TRIzol® and TRIzol® LS. All samples were processed in sterile 1.5 ml Eppendorf tubes with sterile pipette tips autoclaved twice for 15 min at $121 \text{ }^\circ\text{C}$ to prevent RNase contamination.

After complete homogenisation, samples were left at room temperature for a further 5 min to allow complete dissociation of nucleoprotein complexes. Each sample then received 0.2 ml molecular grade chloroform and was shaken vigorously for 15 sec before being left for 15 min at room temperature. Samples were then centrifuged at 12000 g for 15 min at 4°C , separating the mixture into a lower phenol-chloroform phase, an interphase and an upper aqueous phase. Due to the acidic environment, RNA remains exclusively in the upper aqueous phase while the DNA dissolves in the lower phenol-

chloroform phase. The aqueous phase was carefully removed, making sure none of the interphase or lower phase was disturbed, and decanted into a 1.5 ml Eppendorf.

RNA was precipitated using 0.5 ml molecular grade isopropyl alcohol for each sample (i.e. 0.5 ml isopropyl alcohol per 0.75 ml TRIzol® or TRIzol® LS used initially). After 10 min incubation at room temperature, the RNA was pelleted by centrifugation at 12000g for 10 min at 4°C. The supernatant was carefully removed and the RNA pellet washed in 75 % molecular grade ethanol, prepared with DEPC-treated water, by centrifugation at 7500 g for 5 min at 4°C. Any residual ethanol was removed by pipetting and the pellet was left to air dry for approximately 10 min according to the size of the pellet. DEPC-treated water heated to 50°C was used to re-suspend the extracted RNA pellet. The amount of water used varied depending on the size of the pellet but was normally 5-10 µl for the cellular RNA, 30-60 µl for hepatopancreas RNA and over 200 µl for RNA extracted from the whole animal.

After extraction all RNA samples were stored at -70 °C for no longer than one year.

2.2.5 Total RNA Isolation using Glassmax® RNA Micro-Isolation System

Due to the expected small yield of RNA from ascidian cells, another chaotropic method was adopted that exploits the property of nucleic acids to bind to silica in the presence of high salt (Farrell, 1998). This technique enables a small amount RNA to be isolated cleanly. The protocol was taken from the manual supplied with the Glassmax® RNA micro-isolation spin cartridge system (Invitrogen, formerly Life Technologies, Paisley,

UK). All plastic-ware was sterile and RNase-free and all pipette tips were double autoclaved.

Total and separated cell populations were harvested and pelleted as above (2.2.3.3 and 2.2.3.4). Cell pellets in sterile 1.5 ml Eppendorfs were placed on ice immediately before complete homogenisation in 400 μ l ice cold GuSCN/ME buffer (4 M guanidine isothiocyanate; 0.005 % 2-mercaptoethanol) (supplied in the kit) by pulse vortexing. Between pulses the tube was returned to ice to keep the temperature at 4 °C or below. Ice-cold molecular grade absolute ethanol was added (280 μ l) and mixed by inversion. This suspension was then spun at 13000 g for 5 min at room temperature. After carefully removing the supernatant, 450 μ l of binding solution (supplied in the kit) was added, followed by 3 M sodium iodide (NaOAc) (pH 5.5) (supplied in the kit) and vortexed for 1 min. This solution was loaded onto the spin column (supplied in the kit) containing the silica matrix and centrifuged at 13000 g for 20 sec at room temperature. The collection tube was emptied and replaced before the RNA bound to the silica matrix was washed 3 times with 0.5 ml of 4 °C wash buffer (supplied in the kit). The RNA was then washed twice with 0.5 ml 70 % molecular grade ethanol by spinning at 13000 g for 20 sec at room temperature emptying the collection tube each time. Any residual ethanol was removed by a final spin at 13000 g for 1 min at room temperature before the RNA was eluted with 40 μ l of DEPC-treated water pre-heated to 65 °C spun at 13000 g for 20 sec at room temperature.

After extraction all RNA samples were stored at -70 °C for no longer than one year.

2.2.6 Quantification of Total RNA

Extracted total RNA was quantified by spectrophotometry, measuring absorbance at 260 nm. Absorbance was measured using 1 cm path length quartz cuvettes pre-washed in methanol 1:1 HCl, rinsed in DEPC-treated water and dried. Four microlitres of total RNA was pipetted into the cuvette and diluted in 796 μ l DEPC-treated water. The spectrophotometer was blanked using DEPC-treated water.

Nucleic acids absorb UV light maximally at 260 nm allowing the calculation of RNA within an individual sample with the following formula (Farrell, 1998)

$$(\text{RNA}) \mu\text{g/ml}^{-1} = A_{260} \times \text{dilution} \times 40.0$$

where

A_{260} = absorbance (in optical densities) at 260 nm

Dilution = dilution factor (usually 200)

40.0 = extinction coefficient

2.2.7 Total RNA Quality

Total RNA quality was determined by 1% agarose gel analysis in 1 x MOPS buffer (0.2 M MOPS; 0.05 M sodium acetate; 0.01 M EDTA; pH 5.5-7.0) (Sambrook and Russell, 2001). Before loading onto the gel 5 μ l of RNA was added to 5 μ l 2 x denaturing loading buffer (2 x MOPS buffer; 5 % glycerol; 0.25 % bromophenol blue; 50 % formamide; 2 % formaldehyde), adapted from Sambrook and Russell (2001) and Farrell (1998) and heated at 70°C for 5 minutes. RNA in which the 28S and 18S ribosomal RNA (rRNA) bands were defined with minimal smearing above, between and below

was considered of good quality (Farrell, 1998). Transfer RNA (tRNA), 5S tRNA and 5.8S rRNA co-migrate at the leading edge of the gel (Farrell, 1998). Therefore, RNA with no visible banding after agarose electrophoresis was considered degraded and not used.

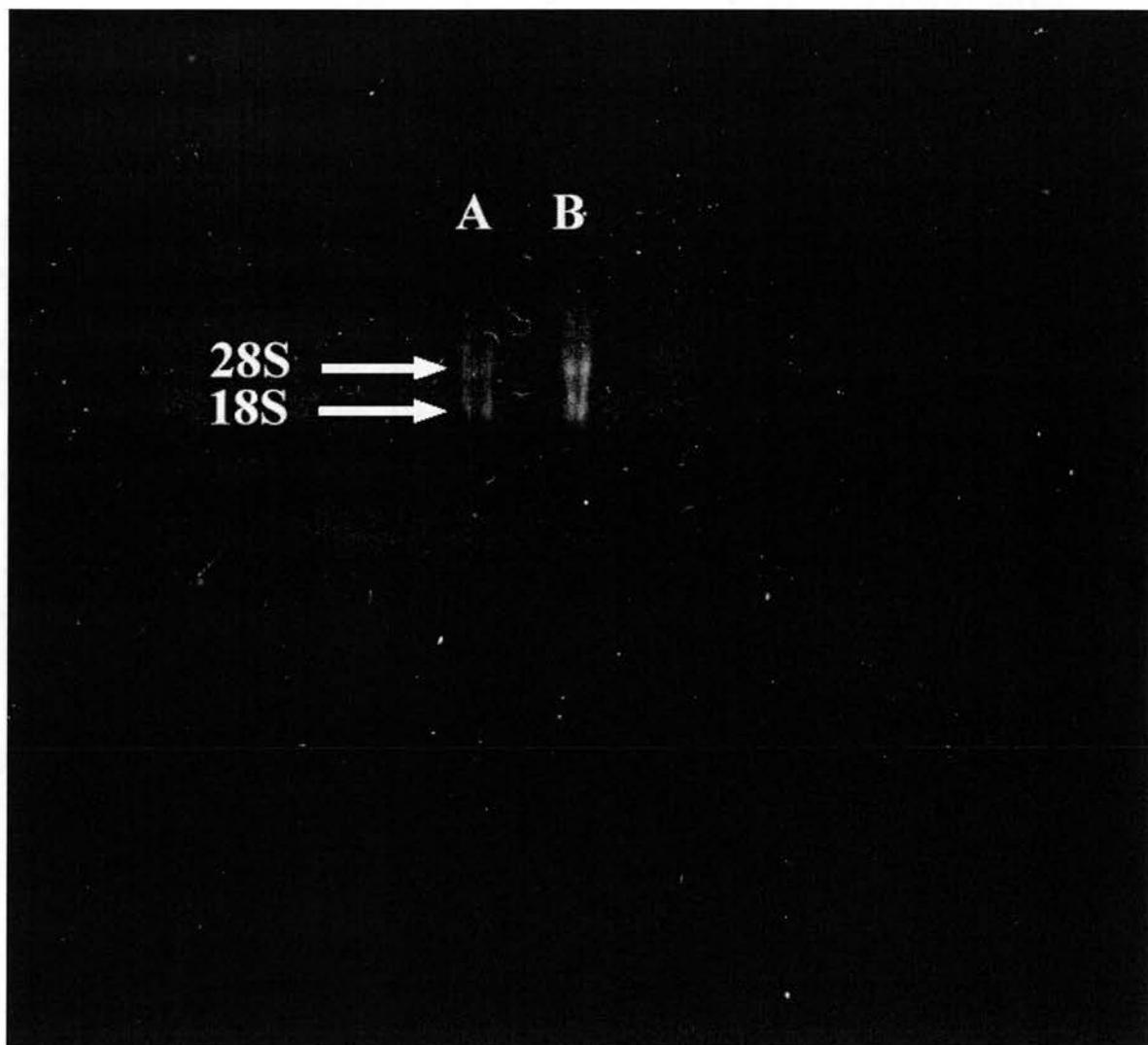
The A_{260}/A_{280} ratio was not calculated as these methods of total RNA extraction leave residual contaminating polysaccharides and proteoglycans (Sambrook and Russell, 2001). These contaminants do not affect any subsequent uses of the RNA but affect the A_{260}/A_{280} ratio so it cannot be relied upon for accurate quality discrimination.

2.3 Results

2.3.1 Larvae

Larvae samples yielded a low concentration of RNA as determined by spectrophotometry at $344 \mu\text{g ml}^{-1}$ after dissolving the final pellet in $10 \mu\text{l}$ DEPC-treated water (i.e. a total RNA yield of $3.44 \mu\text{g}$). Due to the low concentration and volume of this sample, only $1 \mu\text{g}$ was run on an agarose gel to determine quality (Fig. 2.1). Although faint, the 28S and 18S rRNA bands are intact and there is smearing above, between and below these representing the mRNA showing this total RNA was of good quality.

Figure 2.1 Total RNA extracted from *Ciona intestinalis* larvae using TRIzol LS® analysed by agarose gel electrophoresis with 28S and 18S rRNA bands marked with arrows. Sample A is 1 µg RNA extracted from larvae where the 28S and 18S bands are very faint. Sample B is 3 µg of good quality RNA extracted from the hepatopancreas.



2.3.2 Total Cell Population

Despite having clean and viable cell pellets that were immediately homogenised, no intact RNA was recovered. An RNA pellet was obtained after precipitation from TRIzol®, TRIzol® LS and Glassmax® methods but quality assessment consistently revealed degradation. The 28S and 18S ribosomal RNA (rRNA) bands were not visible after agarose gel analysis, although they represent between 80–85 % of cellular RNA (Farrell, 1998), with all the samples appearing as smears along the length of the gel.

2.3.3 Separated Cell Pools

Cell bands 1 and 4 separated on Percoll gradients were often not visible and always composed of fewer cells than the other bands. Cell viability of each band was over 90 % immediately before homogenisation. Only bands 2, 3 and 5 produced an RNA pellet that was very small and dissolved in \leq μ l DEPC-treated water. None of the cell pools, collected from several gradients, produced a quantifiable amount of RNA using TRIzol®, TRIzol® LS or Glassmax® methods. Agarose gel analysis of the entire sample from bands 2 and 5 revealed no visible RNA whilst band 3 revealed a very faint smear. The concentration was too low to determine the quality of this sample.

2.3.4 Hepatopancreas

Total yield varied from 51.36-213.84 μg depending on the size of the hepatopancreas with RNA pellets being dissolved in 30-60 μl DEPC-treated water. Good quality total RNA was consistently recovered from all hepatopancreas tissue used from both TRIzol® and TRIzol® LS with rRNA bands clearly visible (Fig. 2.2).

Total RNA recovered from the hepatopancreas of immune stimulated animals showed no difference in quantity or quality than that of non-stimulated animals (Fig. 2.4).

2.3.5 Whole Animal

Total RNA yield was often in excess of 2000 μg depending on the size of the animal. All RNA was of good quality with rRNA bands clearly visible indicating little or no degradation from TRIzol® LS (Fig. 2.3). RNA extracted using TRIzol® was consistently of a lower quantity for the same amount of starting tissue (Fig. 2.3).

There was no observable difference between stimulated and non-stimulated animals in quantity or quality.

Figure 2.2 Total RNA extracted from 5 g of hepatopancreas from different specimens of *Ciona intestinalis* analysed by agarose gel electrophoresis with 28S and 18S rRNA bands marked with arrows. Samples T were extracted using TRIzol® and samples L were extracted using TRIzol LS®.

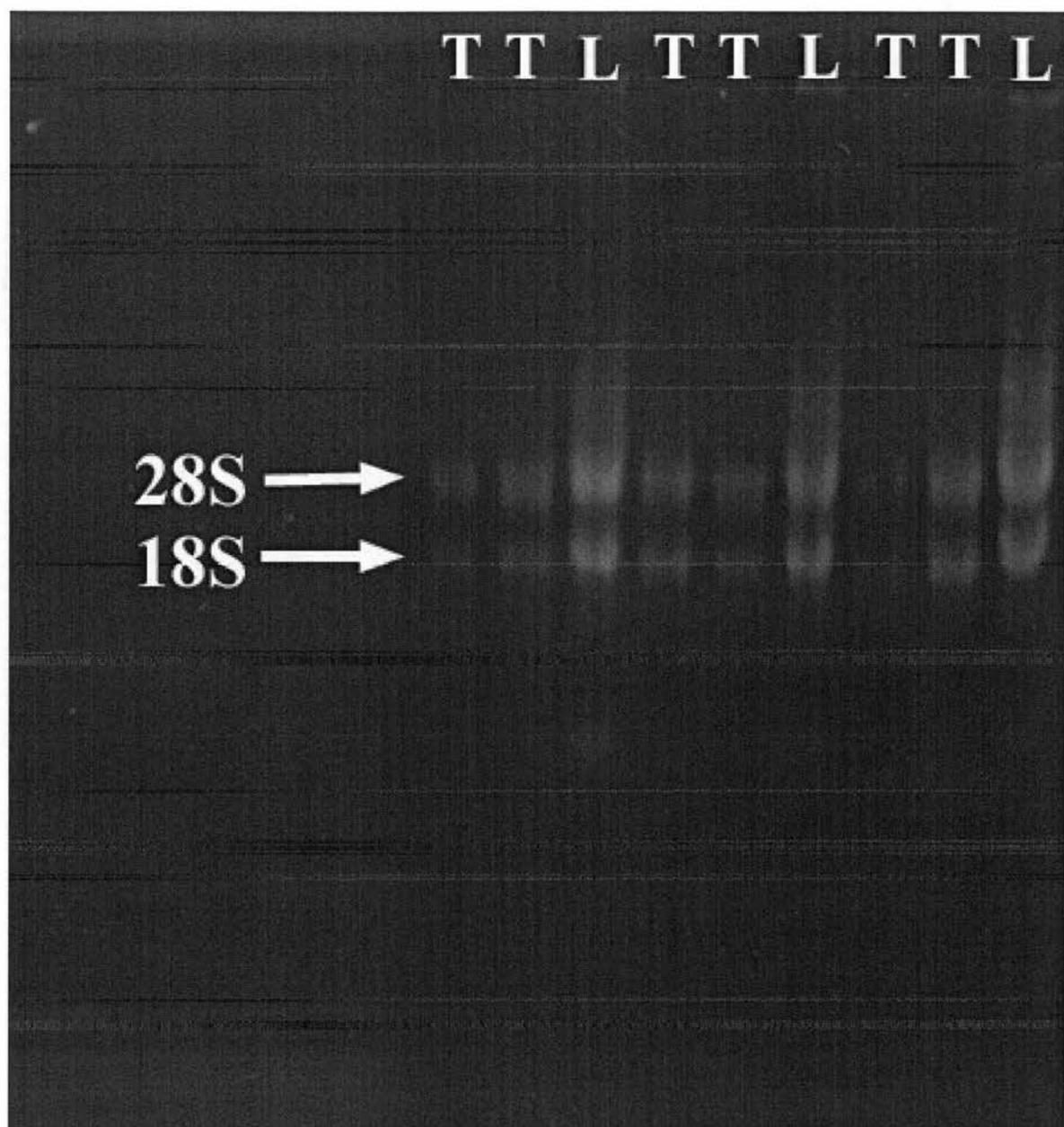


Figure 2.3 Total RNA extracted from whole *Ciona intestinalis* analysed by agarose gel electrophoresis with 28S and 18S rRNA bands marked with arrows. Samples extracted using TRIZOL LS[®] are of a higher concentration good quality. Samples extracted using TRIZOL[®] are less concentrated but the 28S and 18S bands are still visible.

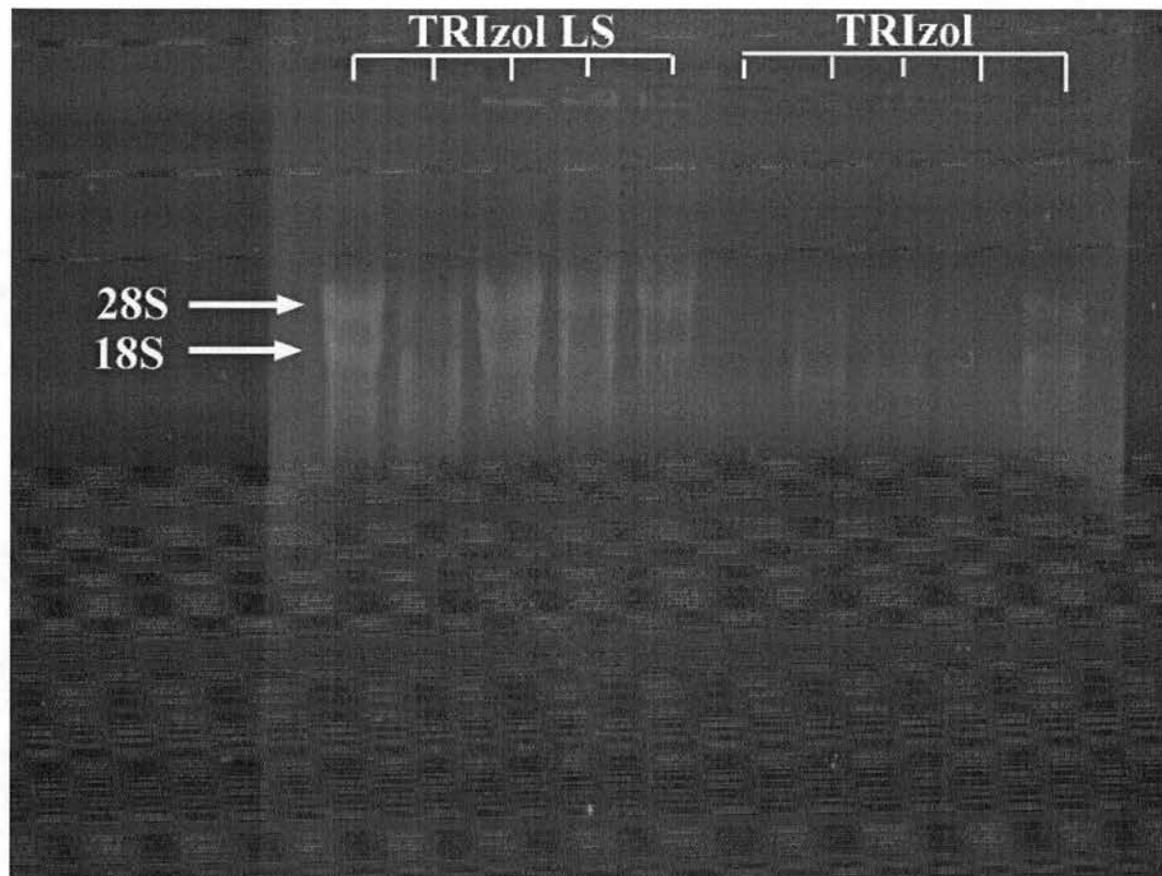
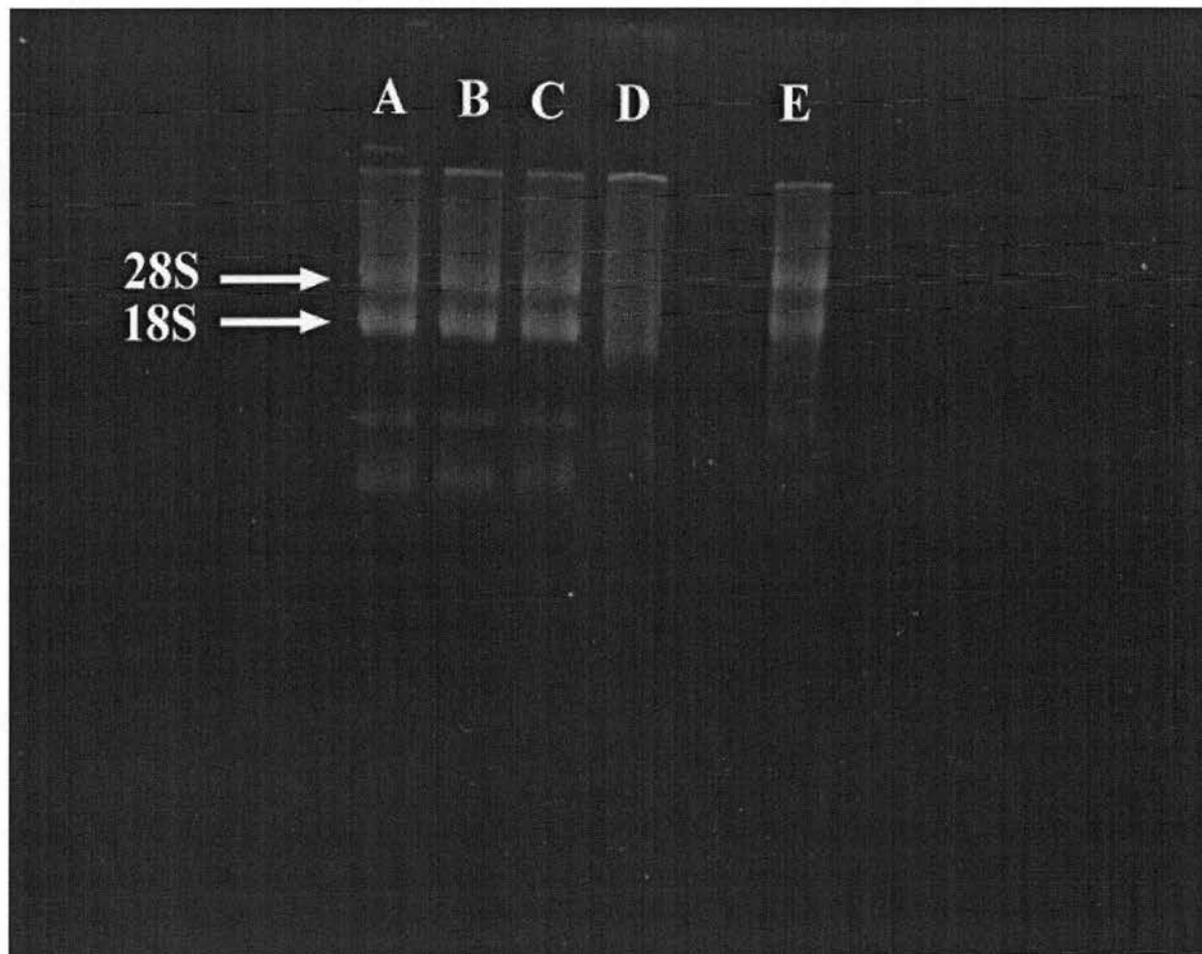


Figure 2.4 Total RNA extracted using TRIzol LS® from the hepatopancreas of *Ciona intestinalis* analysed by agarose gel electrophoresis with 28S and 18S rRNA bands marked with arrows. Sample A is from a non-stimulated animal, sample B is from an animal stimulated with 2.5 µg LPS, sample C is from an animal stimulated with 25 µg LPS, sample D is partially degraded RNA from a non-stimulated animal and sample E is from a saline injected control animal.



2.4 Discussion

As this RNA was to be used in 1st strand synthesis for RT-PCR and RACE, it was of paramount importance it was of high quality. RACE requires cDNA transcribed from full length mRNA if the entire gene sequence is to be amplified. Any mRNA degradation could prevent RACE from working, especially for the 5' end of the gene that is furthest from the cDNA synthesis starting point at the poly A+ tail.

TRIzol® LS, a more concentrated version of TRIzol®, consistently produced the best quality RNA from all the tissue types RNA (Figs 2.2 & 2.3). The higher concentration of guanidine isothiocyanate and phenol in TRIzol® LS, which alters the tertiary folding of RNases and effectively denatures proteins (Farrell, 1998), increased the speed with which samples were homogenised into an RNase free environment, preventing any degradation. As a consequence, all RNA used for subsequent RT-PCR and RACE was extracted using TRIzol® LS

The Glassmax® system used for extraction of cellular RNA produced equivalent results to TRIzol® LS but the RNA was consistently degraded no matter which method was employed.

The successful isolation of RNA from a relatively small number of larvae, adult animal tissue and whole adult ascidian provided a range of RNA samples in which to locate complement genes or their ancestors. Spanning the life cycle of the animal will provide an insight into when these proteins become active and consequently important in the immune system of *C. intestinalis*.

Differences between the RNA from adult animals that had been treated with LPS and control animals were not seen in either quality or quantity using these methods (Fig. 2.4). As other invertebrate species (ascidians, crustaceans or sea urchins) up-regulate the transcription and translation of mRNA during non-self challenge (Jackson and Smith, 1993; Smith *et al.*, 1995; Smith *et al.*, 1996; Hammond and Smith, 2002), the presence of complement molecules may only be detected in the RNA from stimulated animals. Although specimens used in this experiment were constantly under challenge in the marine environment they were collected from, immune stimulation above this background level may be required before increased activity in this arm of the immune system can be detected.

In the present study, good quality RNA was not extracted from the total blood cell extractions using any of the different methods described, including several different attempts using an increasing number of animals. Reasons for this are unclear. Extractions from the mixed blood cell population produced RNA pellets after precipitation, but the RNA was always degraded on analysis. A variety of cells exist in the ascidian circulatory system (Rowley, 1981; 1982; Smith and Peddie, 1992) some are known to have the ability to leave the circulation and migrate through the tissues (Barrington, 1965) indicating they might have multiple functions. There are circulating cells that contain pigments and metabolites, including high levels of vanadium (over 100 mg in one animal), known to be toxic and disruptive to metabolic processes (Kustin *et al.*, 1990). Vanadium contained within cell vacuoles would be liberated during homogenisation along with RNA from other organelles. Vanadium has been reported to damage DNA by nicking (Sreedhara *et al.*, 1996) on account of its reducing and

complexing ability under physiological conditions. It is likely RNA could be damaged in the same way after incubation with vanadium in comparatively high concentrations.

Separated blood cell populations should remove any sub-populations that cause RNA degradation. However, whilst different populations were successfully isolated, the RNA content of even the most numerous sub-population, the phagocytic amoebocytes, (band 3) was too small to be quantified or visualised, even after bleeding from approximately 1200 animals. The only visible RNA pellet was seen in cell band 3 after precipitation following homogenisation in TRIzol® LS. However, the pellet was only just visible without microscopy; a yield too small for use in RT-PCR or RACE. Consequently, isolating RNA from mixed or separated blood cell populations was abandoned.

In summary, after testing different methods of RNA extraction, TRIzol® LS was found to yield the best quality RNA from the whole animal, hepatopancreas and larvae. Good quality RNA was extracted from these tissues using un-treated animals, MS injected control animals and LPS treated animals. RNA extraction from a mixed cell population proved unsuccessful whichever method was adopted as the RNA was always degraded. Isolation from a separated cell population proved unpractical due to the high number of specimens needed to harvest enough of each cell type to extract a working amount of RNA.

Chapter 3

Reverse Transcriptase-Polymerase Chain Reaction to Amplify Complement Genes

3.1 Introduction

Polymerase chain reaction (PCR) enables the isolation and amplification of a specific fragment of DNA within a sample. RT-PCR uses the mRNA extracted from tissues and a reverse transcriptase (RT) enzyme to synthesise complementary DNA (cDNA), which is then used as the template in PCR. The advantage of using RT-PCR is that any DNA fragments amplified must be from transcribed genes i.e. a transcript must have been present in the mRNA. Consequently, synthesising cDNA from a tissue homogenate creates a record of all the genes that were being transcribed at the precise time it was homogenised. This allows direct comparison between samples from different tissues. Also, working with cDNA provides advantages over working directly with RNA. RNA is inherently labile being prone to abundant intrinsic and extrinsic RNases (Farrell, 1998), but the greater stability of cDNA allows the long-term storage of gene transcription information from a specific sample.

RT-PCR was chosen as the preliminary gene location method for complement genes encoding C3, Mannan Associated Serine Protease (MASP) and factor B (Bf) in this study for several reasons. These were that other studies locating complement genes from invertebrates have begun using this technique and have found success in finding several complement gene homologues (Nonaka and Takahashi, 1992; Nonaka *et al.*, 1994; Gross *et al.*, 1999b Nonaka, 1999 #491; Bateman *et al.*, 2002). Invertebrate complement component C3 cDNA has been isolated from the cephalochordate, *Branchiostoma belcheri* (Amphioxus), (Suzuki, 2000; unpublished; Acc No AB050668), C3, Bf and MASP homologues have been found in the Japanese ascidian, *Halocynthia rorezi*, (Nonaka *et al.*, 1999) (Nonaka and Azumi, 1999) (Ji *et al* unpublished; Acc No

AF224491) and homologues to C3 and Bf have been isolated from the sea urchin, *Strongylocentrotus purpuratus*, (Smith *et al.*, 1998; Gross *et al.*, 1999b). The information from these invertebrate studies, and others from vertebrates and lower vertebrates (Nonaka and Takahashi, 1992; Nonaka *et al.*, 1994; Nakao and Yano, 1998; Nonaka *et al.*, 1998) provides information with which to design degenerate primers. Designing good degenerate primers is the key to the success of RT-PCR when the target sequence is not known. Having plenty of information for the design of primers allows an accurate size prediction for the DNA fragment to be amplified. This provides a powerful discriminating tool when analysing results if more than one fragment has been amplified from the template cDNA.

RT-PCR allows relatively rare transcripts in the RNA to be isolated and amplified. Optimally, PCR can amplify one transcript present in a sample of cDNA (McPherson and Møller, 2000). Results from RT-PCR are easily analysed by agarose gel electrophoresis with ethidium bromide to visualise the DNA. The technology of cloning and sequencing DNA from RT-PCR is now well established and reliable (Sambrook and Russell, 2001). The resulting DNA sequence information from RT-PCR can then be used to discover the remaining cDNA sequence from the same pool of RNA or cDNA using either a cDNA library or the rapid amplification of cDNA ends (RACE).

The aim of this chapter was to design degenerate primers using conserved regions in the complement genes Bf, MASP and C3, and use these to find homologous gene fragments in *C. intestinalis* by RT-PCR.

3.2 Materials and Methods

3.2.1 RNA Samples for Reverse Transcription

Only good quality RNA samples as determined in Chapter 2 were used in RT-PCR. These were the stimulated and non-stimulated whole animal and hepatopancreas total RNA samples and the larvae total RNA samples:

(Chapter 2, section 2.2.3.1)

- A:** Whole animal (healthy/untreated)
- B:** Whole animal stimulated $10 \mu\text{g ml}^{-1}$ LPS 3 h incubation
- C:** Whole animal stimulated $100 \mu\text{g ml}^{-1}$ LPS 3 h incubation
- D:** Whole animal stimulated $10 \mu\text{g ml}^{-1}$ LPS 24 h incubation
- E:** Whole animal stimulated $100 \mu\text{g ml}^{-1}$ LPS 24 h incubation
- F:** Whole animal stimulated control (MS)

(Chapter 2, section 2.2.3.2)

- G:** Hepatopancreas (healthy/untreated)
- H:** Hepatopancreas from stimulated animal $10 \mu\text{g ml}^{-1}$ LPS 3 h incubation
- I:** Hepatopancreas from stimulated animal $100 \mu\text{g ml}^{-1}$ LPS 3 h incubation
- J:** Hepatopancreas from stimulated control (MS) 24 h incubation
- K:** Hepatopancreas from stimulated animal $10 \mu\text{g ml}^{-1}$ LPS 24 h incubation
- L:** Hepatopancreas from stimulated animal $100 \mu\text{g ml}^{-1}$ LPS 24 h incubation
- M:** Hepatopancreas from stimulated animal control (MS) 24 h incubation

(Chapter 2, section 2.2.3.5)

N: Approximately 2000 whole larvae

3.2.2 Reverse Transcription

All Eppendorf tubes and pipette tips were sterilised as in Chapter 2 to ensure they were RNase free. Only high quality total RNA determined by agarose gel electrophoresis (section 2.2.7) was used for reverse transcription to ensure that sample degradation could not prevent the discovery of rare transcripts.

The reverse transcriptase enzyme used was Superscript II (Invitrogen, formerly Life Technologies, Paisley, UK) as some of the degenerate priming sites are several thousand bases from the poly (A⁺) end of the mRNA. Superscript II has been engineered to remove the RNase H activity that is present in some other RNA-dependent DNA polymerases (AMV, M-MLV), thus enhancing reverse transcription efficiency (Farrell, 1998). As total RNA was used, first strand synthesis was primed using oligo (dT) rather than random primers. This selected for just mRNA in transcription as it primes from the furthest 3' end of the poly (A⁺) tail present only in the mRNA. Consequently the resulting cDNA is poly (A⁺) cDNA rather than total cDNA.

Sample RNA was removed from storage at -70 °C and kept on ice until used, usually no longer than 15 min. An initial mix of 5 µg total RNA, 1 µl oligo (dT) (500 µg ml⁻¹) and 1 µl dNTP mix (10 mM each) was added to a 0.5 ml Eppendorf and a total volume made up to 13 µl with nuclease-free water. This was heated at 65 °C for 5 min to reduce any secondary structure and then chilled immediately on ice. Contents were collected at the

bottom of the Eppendorf tube by brief centrifugation before adding 4 μ l first strand buffer (250 mM Tris-HCl, 375 mM KCl, 15 mM MgCl, pH 8.3) (supplied with Superscript II) and 2 μ l 200 mM dithiothreitol (DTT). This was incubated at 42 °C for 2 min before the addition of 1 μ l (200 units) of Superscript II reverse transcriptase and further incubation for 1.5 h at 42 °C. After this, a final incubation for 15 min at 70 °C permanently inactivated the enzyme. The resulting cDNA was then diluted with 80 μ l of nuclease-free water (i.e. x 5 dilution) and stored at -20 °C. Controls were made for each template in precisely the same way, except the reverse transcriptase was replaced by nuclease-free water. This control was used as a template in PCR alongside the full templates ensuring amplified DNA came from cDNA rather than any contaminating genomic DNA remaining from the RNA extraction.

A total of fourteen cDNA templates were made corresponding to the fourteen different RNA samples transcribed (section 3.2.1). These were:

cDNA **A**: Whole animal (healthy/untreated)

cDNA **B**: Whole animal stimulated 10 μ g ml LPS 3 h incubation

cDNA **C**: Whole animal stimulated 100 μ g ml LPS 3 h incubation

cDNA **D**: Whole animal stimulated 10 μ g ml LPS 24 h incubation

cDNA **E**: Whole animal stimulated 100 μ g ml LPS 24 h incubation

cDNA **F**: Whole animal stimulated control (MS)

cDNA **G**: Hepatopancreas (healthy/untreated)

cDNA **H**: Hepatopancreas from stimulated animal 10 μ g ml LPS 3 h incubation

cDNA **I**: Hepatopancreas from stimulated animal 100 μ g ml LPS 3 h incubation

cDNA **J**: Hepatopancreas from stimulated control (MS) 24 h incubation

cDNA **K**: Hepatopancreas from stimulated animal 10 µg ml LPS 24 h incubation

cDNA **L**: Hepatopancreas from stimulated animal 100 µg ml LPS 24 h incubation

cDNA **M**: Hepatopancreas from stimulated animal control (MS) 24 h incubation

cDNA **N**: Approximately 2000 whole larvae

Each template was used with the full range of PCR conditions (section 3.2.4) allowing comparison between these tissues and ensuring the amplification of as many transcribed genes of interest as possible, which may be expressed after LPS challenge or at different stages in the life cycle of *C. intestinalis*.

3.2.3 Degenerate Primer Design

All primers were synthesised by MWG Biotech (Ebersberg, Germany)

As C3, MASP and Bf proteins are unknown in *C. intestinalis*, degenerate primers were designed against sequences from lower vertebrates (Nonaka and Takahashi, 1992; Nonaka *et al.*, 1994; Nakao and Yano, 1998; Nonaka *et al.*, 1998) and other invertebrates (Smith *et al.*, 1998; Gross *et al.*, 1999b; Nonaka and Azumi, 1999; Nonaka *et al.*, 1999) (Table 3.1 & 3.2). Degenerate primers are a mix of oligonucleotides containing all the possible sequences that can encode a given amino acid sequence (Sambrook and Russell, 2001). If the amino acid sequence chosen is coded by the sequence of interest in the cDNA, one of the oligonucleotides in the mix will match that DNA perfectly.

Multiple amino acid and DNA sequence alignments were produced, including similar sequences from the same gene families as the proteins of interest (Appendices 3 & 4). This highlighted conserved regions likely to be present in the individual cDNA's of interest, and allowed selective discrimination against regions common to all sequences that would decrease the efficiency of PCR to amplify only the fragments of interest.

Serine protease primers were designed using conserved serine protease domains in most Bf and MASP sequences (Appendix 3). The anti-sense primers were designed against 2 regions, the first specific to Bf and MASP and the second a domain present in many serine proteases including Bf and MASP so nested PCR could be performed (Appendix 3). The expected product from the serine protease primers was approximately 460 base pairs (bp) using anti-sense 3 and 4 primers and approximately 230 bp using 1 and 2 anti-sense primers (Table 3.1). Four forward and four reverse primers were synthesised allowing 16 different combinations (Table 3.1).

C3 primers were designed against the thiolester region, which is highly conserved among all C3, C4 and $\alpha 2m$ proteins (Nonaka and Takahashi, 1992; Nonaka, 1997) (Appendix 4). The expected product size for a DNA fragment amplified by the C3 degenerate primers was approximately 220 bp (Table 3.2). Three forward and five reverse primers were synthesised allowing 15 different combinations (Table 3.2).

Table 3.1 Degenerate primers for the serine proteases Bf, MASP a and MASP b (degeneracy code in Appendix 1). T_m (Appendix 2) represents melting temperature in degrees centigrade of the oligonucleotide primer.

Sense	Target sequence	Degenerate sequence (5'→ 3')	T _m °C
A	VLTAABC	GTNCTNACNGCNGCNCAYTG	61.4
B	VLTAABHV	GTNCTNACNGCNGCNCAYGT	61.4
C	VLTAABL	GTNCTNACNGCNGCNCAYTT	59.4
D	LTAABC	CTNACNGCNGCNCAYTG	59.4
Anti-sense			
1	PICLPCT	GTRCANGGNAGRCADATNGG	59.0
2	PVCLPCT	GTRCANGGNAGRCANACNGG	61.4
3	GDSGGP(TCN)	GGNCCNCCNGARTCNCC	61.2
4	GDSGGP(AGR)	GGNCCNCCRCTRTCNCC	61.2

Table 3.2 Degenerate primers for C3, (degeneracy code in Appendix 1). T_m (Appendix 2) represents melting temperature in degrees centigrade of the oligonucleotide primer.

Sense	Target sequence	Degenerate sequence (5'→ 3')	T _m °C
E	GCGEQNM	GGNTGYGGNGARCARAAYATG	57.3
F	GCAEQNM	GGNTGYGCNGARCARAAYATG	59.8
G	GCGEQNMI	GGNTGYGGNGARCARAAYATGAT	60.6
Anti-sense			
5	TWLTAYV	ACRTANGCNGTNAGCCANGT	58.3
6	WLTAYV	TANGCNGTNAGCCANGT	52.8
7	TWLNQFV	ACRTANGCNGTNAGCCA	52.8
8	GGFISTQ	ACNAANCCRTTNAGCCANGT	56.3
9	WLNQFV	TGNGTNGANATRAANCCNCC	57.3

3.2.4 PCR

PCR was performed using *Taq* DNA polymerase (Amersham Pharmacia Biotech Inc, Piscataway, USA) according to the manufacturer's guidelines using the single stranded cDNA synthesised in section 3.2.2. Briefly, 20 μ l reactions were set up on ice in thin walled 0.2 ml PCR tubes (Axygen, California, USA). To each tube, 0.5 μ l of template was pipetted into the bottom and the appropriate amount of each primer was pipetted onto the inside wall of the tube. Final primer concentrations were initially 2 μ M but were increased to 5 μ M if amplification was consistently unsuccessful. A master mix was then made to include all the remaining reaction components to a final volume sufficient for all the reactions. This prevented any error from the pipetting of several small volumes (e.g. 0.2 μ l) into the same tube. This mix contained 2 μ l of $\times 10$ PCR buffer (500 mM KCl; 15 mM MgCl₂; 100mM Tris HCl; pH 9.0) (supplied in the kit), 10 mM of each nucleotide and 1 unit of *Taq* DNA polymerase, made up to a final volume with nuclease-free water for each tube. When the master mix was added to the PCR tube containing the template and primers, the final volume was 20 μ l.

For every full reaction, comprising of template made with RT with forward and reverse primers, the following control reactions were run:

1. Template made with RT with forward primer only
2. Template made with RT with reverse primer only
3. Template made with RT with no primers
4. Template made without RT with forward and reverse primers
5. No template with forward and reverse primers

Thermal cycling was performed in either a Touchdown or Sprint thermalcycler (Hybaid Ltd, Middlesex, UK). The only variable in the cycling conditions was the primer annealing temperature. Basic cycling conditions were:

Denaturation at 94 °C for 5 min

30 cycles

Denaturation-94 °C for 30 sec

Primer annealing-variable °C for 30 sec

Primer extension-72 °C for 1 min

Elongation-72 °C for 7 min

Storage at -20 °C

Each serine protease primer combination was tested at an annealing temperature of 45, 50 or 55 °C for all the different templates used. For the thiolester primers, annealing temperatures of 40, 42, 45, 50 or 55 °C were used.

3.2.5 Re-amplification or Nested PCR

For faint PCR products or for more specific PCR reactions, re-amplification or nested PCR was performed. Re-amplification uses the initial PCR reaction as a template for a subsequent PCR using the same primers to increase the concentration of the DNA.

Nested PCR uses different 'nested' primers designed against amino acid regions within the primary PCR product. DNA amplified with the serine protease primers 3 and 4 (Table 3.1) was subjected to nested PCR using the same forward primer with reverse primers 1 and 2 (Table 3.1). Primers 1 and 2 were designed against more specific regions in complement proteins and should isolate sequences of interest from the mix of products amplified by primers 3 and 4 (Appendix 3).

DNA was excised from an agarose gel (detailed in 3.2.6) and diluted 20x with nuclease-free water. This reduced the amount of EDTA that can inhibit PCR and reduces the amount of target template to a more efficient level. A 1 μ l volume of this extracted DNA was the template in the secondary PCR using the same mix as the initial PCR (3.2.4) with the appropriate primers. Similar controls were used as section 3.2.4.

3.2.6 Analysis of Results

A sub-sample of 5 μ l was removed from the PCR reactions, mixed with 6x loading buffer (100 mM EDTA; 25 mM Tris-HCl; 25 % glycerol; 0.05 % bromophenol blue; pH 7.0) (Sambrook and Russell, 2001) and run on an agarose gel. For expected products of 200-500 bp a 2 % agarose gel was used containing 1 μ g ml⁻¹ ethidium bromide using 0.5x TBE electrophoresis buffer (45 mM Tris-borate; 1 mM EDTA; pH 8.0) (Sambrook and Russell, 2001) and run at 100 V until the dye front had migrated along half the length of the gel. Ethidium bromide binds to DNA between the stacked base pairs and fluoresces with exposure to ultraviolet (UV) radiation (Sambrook and Russell, 2001) enabling visualisation in the agarose gel. Molecular markers (Bioladder 100; Hybaid

Ltd, Middlesex, UK) were also run on the gels to determine that DNA fragments were of the expected size.

3.2.7 Cloning of PCR Amplified Products

The remaining DNA of interest was run on an agarose gel, as above, and then excised from the gel with a clean scalpel. DNA was recovered from the agarose using a Qiaquick gel extraction kit (Qiagen Ltd, West Sussex, UK) according to the manufacturer's guidelines. DNA was eluted from the spin columns (supplied in the kit) using 30 μ l nuclease-free water. To determine DNA concentration, 1 μ l of eluted DNA was run on an agarose gel as above and compared to the known concentration of the molecular weight markers (Bioladder 100). Recovered DNA was stored at 4 °C until required.

The pCR 2.1 Topo vector (Invitrogen, California, USA) (Appendix 5) was used to clone RT-PCR products. Fragments were ligated into the vector using a slight modification of the manufacturer's guidelines supplied with kit. Recovered DNA (20 ng) was mixed with nuclease-free water to a volume of 4.5 μ l before the addition of salt solution (1.2 M NaCl; 60 mM MgCl₂) (supplied in the kit) and 0.5 μ l of plasmid vector (supplied in the kit) giving a final concentration of 0.83 ng/ μ l. This was incubated at 20 °C for 10 min then placed on ice while a vial of TOP10F' competent cells (50 μ l) (supplied in the kit) was thawed on ice from -80 °C. Once thawed, 2 μ l of ligation mix was added to the cells and mixed gently before incubation on ice for 10 min to allow the vector to attach to the cell surface. The cells were then heat shocked at 42 °C for 30 sec to allow pore formation in the cell membrane so the DNA could penetrate. The cells were again

removed to ice and 250 μl of SOC medium (supplied in the kit) was added before incubation in a shaking water bath at 37 °C for 1 h.

The surfaces of Luria-Bertani (LB) agar (ICN Biomedicals Inc, Ohio, USA) plates containing 50 $\mu\text{g ml}^{-1}$ ampicillin were spread with 40 μl isopropyl thiogalactoside (IPTG) (Promega, Southhampton, UK) and 40 μl X-GAL (5-bromo-4-chloro-3-indolyl- β -D-galactopyranoside) (Promega, Southhampton, UK), using sterile glass spreaders, dried under aseptic conditions and warmed to 37 °C in an air incubator. After incubation, 50 μl and 150 μl of the cells were spread onto two separate plates and incubated at 37 °C for 12 h or overnight.

The plates were then tested for positive recombinant colonies. Blue and white colony screening facilitates the fast visual screening of colonies that are likely to contain the inserted PCR product. The pUC 18 supercoiled plasmid contains the LacZ gene that synthesises β -galactoside, which cleaves X-GAL producing a blue product. In successful transformations DNA is inserted into this gene, disrupting its function so ultimately the blue product is not present and the colony appears white.

Several white colonies were tested for the correct insert by colony PCR. M13 forward and reverse primers (supplied in the kit) corresponding to regions either side of the cloning site were used to amplify a small piece of the vector containing the inserted PCR product. Several control reactions were set up for each white colony to elucidate if the insert was present, and also to ascertain if the PCR amplification had been successful with the forward and reverse degenerate primers, used in RT-PCR, at either end of the sequence:

1. M13 forward and reverse
2. Degenerate forward and reverse
3. Degenerate forward only
4. Degenerate reverse only

General control reactions were also set up using M13 forward and reverse primers with (A) a blue colony that should contain no insert and (B) control DNA (supplied in the kit) as the template that will produce an amplified product if the PCR conditions are correct.

PCR reactions containing the appropriate primers at the right stock concentration were set up in the same way as described in section 3.2.4, except no template was present in the tube. Individual candidate colonies were then touched with a sterile cocktail stick, dipped into the PCR tube and then touched onto a sterile LB/ampicillin plate prepared in the same way as detailed above. A numbered grid was drawn on the reverse side of this plate so these sub-colonies could later be identified once the plate had been incubated for 12 h over night.

PCR was carried out in the same manner as described in section 3.2.3 with modified cycling conditions:

Denaturation at 95 °C for 5 min

30 cycles

Denaturation-95 °C for 30 sec

Primer annealing-55 °C for 30 sec

Primer extension-72 °C for 1 min

Storage at -20 °C

An elongation step was no longer needed as the addition of adenosine (A) overhangs on the 3' end of the PCR product were no longer needed.

Reactions were scrutinised in the same manner as 3.2.4.

3.2.8 DNA Sequencing

The remaining DNA from the M13-primed colony PCR reactions, which identified positive recombinant colonies containing the correct insert, was run on and extracted from an agarose gel and quantified as in section 3.2.4. The majority of the sequencing was then carried out personally on an ABI Prism 377 (Applied Biosystems, Foster City, USA) sequencer. Sequencing was carried out according to the manufacturer's guidelines using Big Dye Terminator Cycle sequence kit (Applied Biosystems). The St Andrews DNA sequencing unit (University of St Andrews, St Andrews, Fife) sequenced the remaining DNA.

3.3 Results

3.3.1 Serine Protease RT-PCR for Bf and MASP

Of the 16 serine protease primer combinations used, 13 amplified DNA bands of appropriate sizes (Table 3.3) using a final primer concentration of 2 μ m. Increasing the primer concentration did not yield any more products or increase the concentration of the DNA band in the agarose gel. Control templates without RT and all forward primer controls were negative. However, both the reverse primers 3 and 4 (Table 3.1) amplified approximately 460 bp DNA fragments when used individually with no forward primer. This meant that DNA fragments amplified with both forward and reverse primers were likely to contain at least a proportion of mis-primed sequences. Using products amplified with a forward primer and reverse primers 3 and 4 as a template, nested RACE using the same forward primer and reverse primers 1 and 2 was successful (Fig. 3.1) (Table 3.3). This provided several pools of DNA from forward primers C and D (Table 3.3) with all nested negative controls containing no amplification of DNA at this 230 bp size.

No DNA was amplified from whole animal template using any of the 16 primer combinations at any concentration or temperature (Table 3.3).

Hepatopancreas template G produced bands from all the 13 successful primer combinations at a final concentration of 2 μ m (Table 3.3). Template from animals treated with LPS produced DNA fragments with fewer primer combinations (Table 3.3).

An increasing dose of LPS and an increasing incubation time decreased the number of successful primer combinations (Table 3.3).

The template synthesised from the larvae yielded products using only forward primer D in combination with the reverse primers 2, 3 and 4 (Table 3.3).

All DNA extracted was successfully cloned. The vast majority of colonies passed all the colony controls (section 3.2.7) but sequencing revealed only PCR products amplified using D primer were fragments from serine protease genes. Several different serine protease gene fragments were isolated using D primer in combination with all the reverse primers. Alignments with the complement sequences used to design the degenerate primers (Fig. 3.3) revealed which were likely fragments of complement gene homologues and those that showed significant identity to other serine proteases.

In total 9 different products were isolated and 7 were selected for further study (Fig. 3.2). SP1 and SP5 amplified using nested PCR D4:D2 from template G, SP2 was amplified using nested PCR with primer combinations D3:D2 from templates G and N, SP3, SP4 SP6 and SP7 were amplified with primers D2 from templates G, H and N (Table 3.3)

Figure 3.1 Serine protease RT-PCR DNA bands amplified using degenerate primer combinations D1 (sample A) and D2 (sample B) using DNA amplified from primer combinations D3 and D4 as a template run on a 2 % agarose gel. 1 is using untreated hepatopancreas RNA as the original template, 2 is using larvae as the original template and 3 is using whole animal as original template. Lane M is Hybaid BioLadder® DNA markers which have bands with a range from top to bottom of 1000 bp to 100 bp at 100 bp intervals. No DNA was amplified with the whole animal RNA template with primer combinations D1 and D2 (lanes A3 and B3).

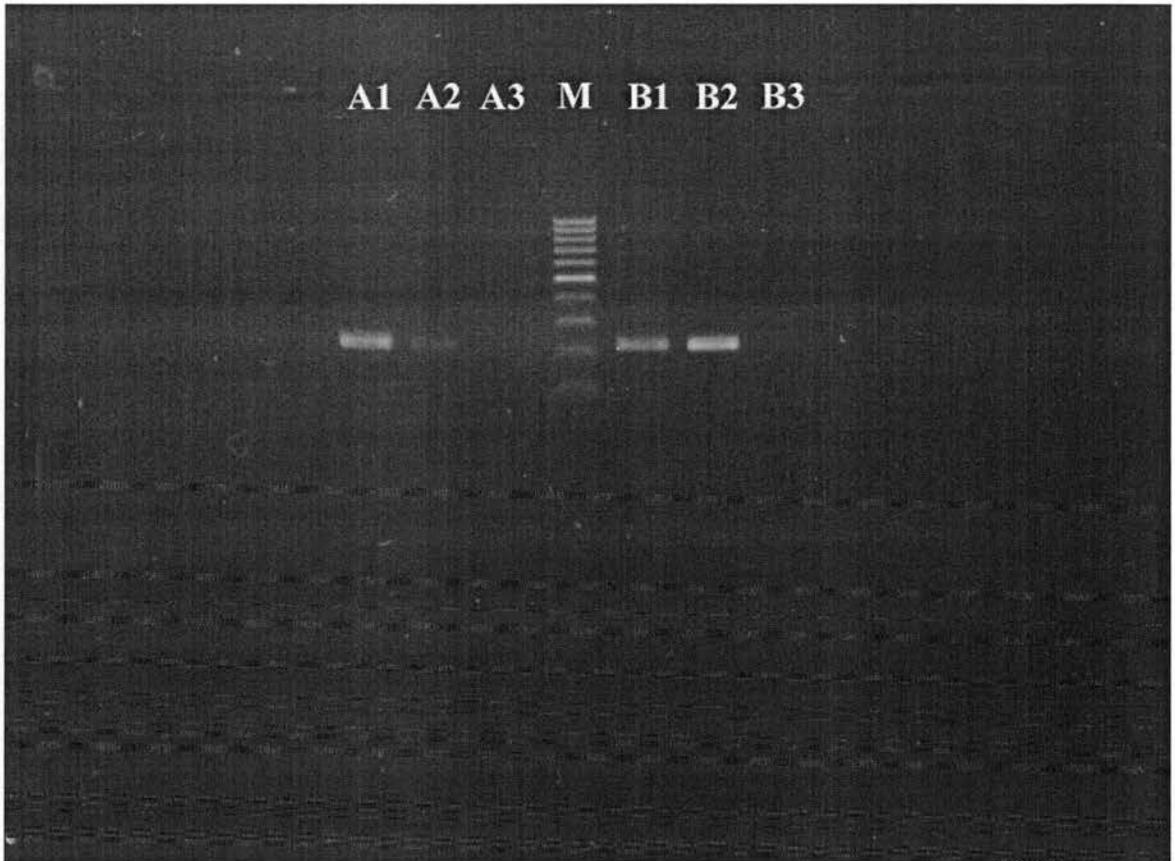


Table 3.3 Serine proteases degenerate primer combinations amplifying DNA fragments of appropriate size.

Primers	Amplification	Successful conditions			
		cDNA template	Nested template	Anneal °C	Primer conc μM
A1	✗				
A2	✓	G & J		50	2
A3	✓	G, H, I & J	✗	50 & 55	2
A4	✓	G, H & J		50 & 55	2
B1	✗				
B2	✓	G & J		50	2
B3	✓	G, H, I & J	✗	50 & 55	2
B4	✗				
C1	✓	✗	C3	50	2
C2	✓	G, H & J	C3	50 & 55	2
C3	✓	G, H, I & J	✗	50 & 55	2
C4	✓	G & I	✗	55	2
D1	✓	G, H & J	D3 & D4	50 & 55	2
D2	✓	G H & N	D3 & D4	50 & 55	2
D3	✓	G, I, J & N	✗	50 & 55	2
D4	✓	G, J & N	✗	55	2

Figure 3.2 Serine protease DNA sequences homologues to complement factors with amino acid translation. Degenerate primer sequences removed. SP1 and SP5 amplified using nested PCR D4:D2 from templates G, SP2 was amplified using nested PCR with primer combinations D3:D2 from templates G, and N, SP3, SP4 and SP6 were amplified with primers D2 from templates G, H and N.

SP1

```

1      TTGCAACAAATAACGGAAAATGAGTACAGCATTACACAAGTTTTTCGGCAGTATTTGGATTG
1      L  Q  Q  I  T  E  N  E  Y  S  I  H  K  F  S  A  V  F  G  L

61     TTTCGATTGAATTTGCAACACAACACACAGAGAATTGGTTTCAAAGAACATTTATTCAT
21     F  R  L  N  L  Q  H  N  T  Q  R  I  G  F  K  R  T  F  I  H

121    TCGGATTTTCAAAGCGCACATTTAACTTTTAGAAAACGATGTCGCATTGATACAATTAGAT
41    S  D  F  Q  S  A  H  L  T  F  R  N  D  V  A  L  I  Q  L  D

181    CGAAAAATACAATGGACAAGCAATATTCGC
61    R  K  I  Q  W  T  S  N  I  R

```

SP2

```

1      ACAAGTAGAGTAAAAAGAGAAAGAAAGAAACACTTTGTTTCGAGTTGGAGATTATTTCAAC
1      T  S  R  V  K  R  E  R  K  K  H  F  V  R  V  G  D  Y  F  N

61     CGAGATAACCTTCCTCATAGTCAGGATTCTATGGTTGAAGAGTCACATGATATAGCAATT
21     R  D  N  L  P  H  S  Q  D  S  M  V  E  E  S  H  D  I  A  I

121    AGCCAAATTTATATTCATGAGGGTTTTACTCAGTACCCTGCAACAAGAAACGATATTGCT
41    S  Q  I  Y  I  H  E  G  F  T  Q  Y  P  A  T  R  N  D  I  A

181    TTAATTAACCTAAGCGAACCGGTGTCGCTAACACGGTTTTGTTCAA
61    L  I  K  L  S  E  P  V  S  L  T  R  F  V  Q

```

SP3

```

1      AGATCCGTATCTTACTCCGGTCTCCTTGTTTACCTCGGAACCACCAGGAGCTCTCATCTT
1      R  S  V  S  Y  S  G  L  L  V  Y  L  G  T  T  R  S  S  H  L

61     ACACATCTTGATACCACTAGGAGGCAACGAAGAGAGGTTGAACAGATCATAGTACACCCC
21     T  H  L  D  T  T  R  R  Q  R  R  E  V  E  Q  I  I  V  H  P

121    GGTTCACCGCTGAGTATTTGAACGACGTTGCATTAATAAAGCTGAGTCGCCCCGTTGTG
41    G  F  T  A  E  Y  L  N  D  V  A  L  I  K  L  S  R  P  V  V

181    TTTAATGACATCATCACC
61    F  N  D  I  I  T

```

SP4

```

1      GCATCTATAACAAACAACAACCCAAGCACCATTAACGTCATATTGGGTGTTGTTGACACA
1      A  S  I  T  N  N  N  P  S  T  I  N  V  I  L  G  V  V  D  T

61     ATTGATTCAGGAAACATACATGAACAATCTTTTTCTGTTACAAGACTTATAATTCATCCA
21     I  D  S  G  N  I  H  E  Q  S  F  S  V  T  R  L  I  I  H  P

121    AACTACAATTTCCCAAACAACGACCTTGCATTGCTACAACCTGGACCATGATGCTCTGATT
41    N  Y  N  F  P  N  N  D  L  A  L  L  Q  L  D  H  D  A  L  I

181    GATGCGGCTTTTGTGAAA
61    D  A  A  F  V  K

```

SP5

1 CTTAAACATGACATCCATCGACATGGTTTGGTCATCGTTACATTAGGAATGCTACGTCAG
1 L K H D I H R H G L V I V T L G M L R Q
61 CATGTCACGTTTGGAGCGTAGCCGGCAGTACAGAATCGACAAAAGGATTGTCATACATCCA
21 H V T F E R S R Q Y R I D K R I V I H P
121 GAATTCGTTTTCCCGCACTATGACGTCGCGTTAATCGAAGTGGATCGCGCTTTTGACGTT
41 E F V F P H Y D V A L I E V D R A F D V
181 ACTGGCGTTTTTGTGTCAGG
61 T G V F V R

SP6

1 GACCATGTGACGTCACAAAAGCAGGTTGATAAAACCATTCTCGGCTTTGGGACTTCCCAG
1 D H V T S Q K Q V D K T I L G F G T S Q
61 CTTTGGCGGGCTGACGCGCCTCTCGCACCTGCTGCGTGATAGTGACGTCACCGTGACGTCA
21 L W R L T R L S H L L R D S D V T V T S
121 ATGGATGACGTCACCTGGTGCGAGCGTGATCCGTTTAAAGTCGCATCTACAGCCACCCTACT
41 M D D V T G A S V I R L S R I Y S H P T
181 TACGGCGAAAACCTGGATAGTGATATCGTATTGATCAAGGTTGCCGAGCCGATCACGTGG
61 Y G E N L D S D I V L I K V A E P I T W
241 TCGTCTCGGGTATTC
81 S S R V F

SP7

1 TTGCAAAATGATGAAATAAATATAACATCAGTCCATGTGTTTGGTGGGAAAGTCTTAACT
1 L Q N D E I N I T S V H V F V G K V L T
61 GATGTTACATTGATTGAACCATAACCAACAACATTCTCTCGTCTCCCATGTTGCATTTTCAT
21 D V T L I E P Y Q Q H S L V S H V A F H
121 GAGAATTACGATCCCATAATTTAAATTCAGATATCGCCATTCTTACGTTATCAACGCAA
41 E N Y D P D N L N S D I A I L T L S T Q
181 ATAGTATTCACCAAAGCAGTGAGCCGCAGCTTAG
61 I V F T K A V S R S L

3.3.2 Thiolester RT-PCR for C3

Of the 15 thiolester primer combinations used, 8 amplified DNA fragments of approximately 220 bp (Fig 3.4) (Table 3.4) at a final primer concentration of 5 μ m. Lower concentrations either produced very faint or no bands.

All controls were negative except for single primer controls using all hepatopancreas cDNA templates for primers E, F, and 6, which amplified a DNA fragment of the same size at 220 bp. To amplify enough DNA for cloning, PCR had to be performed again using the same primer combination with the initial PCR product as a template in all the reactions using a final primer concentration of 5 μ m.

Only the combination of F6 amplified a 220 bp DNA fragment from whole animal cDNA templates A and B. The cDNA templates from the higher level of LPS stimulation (2.5 μ g) and longer incubation time of 24 h produced no results.

Amplification from healthy or untreated animals succeeded for all the successful primer combinations. Templates synthesised from the hepatopancreas LPS treated animals produced bands with fewer primer combinations than hepatopancreas cDNA from non-stimulated and control animals (Table 3.4). The higher the level of LPS stimulation the fewer primer combinations worked regardless of the incubation time (Table 3.4).

Only the primer combination of E6 produced amplified a DNA fragment from the larvae template M (Fig 3.4) (Table 3.4).

All extracted DNA after secondary PCR re-amplification was successfully cloned. Colony controls allowed the discrimination of DNA fragments that had been amplified by the mis-priming of a single primer, indicated by the single primer controls during PCR. After sequencing, only 2 different DNA fragments were successfully isolated, thiolester 1 and 2, both from the primer combination E6 (Fig 3.5). All the other primer combinations had mis-primed and the amplified DNA was either an un-related sequence or nonsense. Thiolester 1 was amplified from all the hepatopancreas templates apart from L and the larvae, but was not amplified from the whole animal. Thiolester 2 was amplified from the hepatopancreas template G (untreated) only. Alignment with other thiolester containing proteins used to design the degenerate primers (Fig. 3.6) (Appendix 4) showed that thiolester 1 and 2 were both homologous and likely fragments of complement gene homologues.

Figure 3.4 Thiolester RT-PCR DNA fragments amplified using degenerate primer combinations E6 (sample A) and E5 (sample B) run on a 2 % agarose gel. 1 is using non-stimulated hepatopancreas cDNA; 2 is using larvae cDNA; 3 is using whole animal RNA and 4 is using hepatopancreas RNA from 2.5 µg LPS injected animal. Lane M is Hybaid BioLadder® DNA markers which have bands with a range from top to bottom of 1000 bp to 100 bp at 100 bp intervals. No DNA of the correct size was amplified with the whole animal RNA template with primer combinations D6 or E5 (lanes A3 and B3).

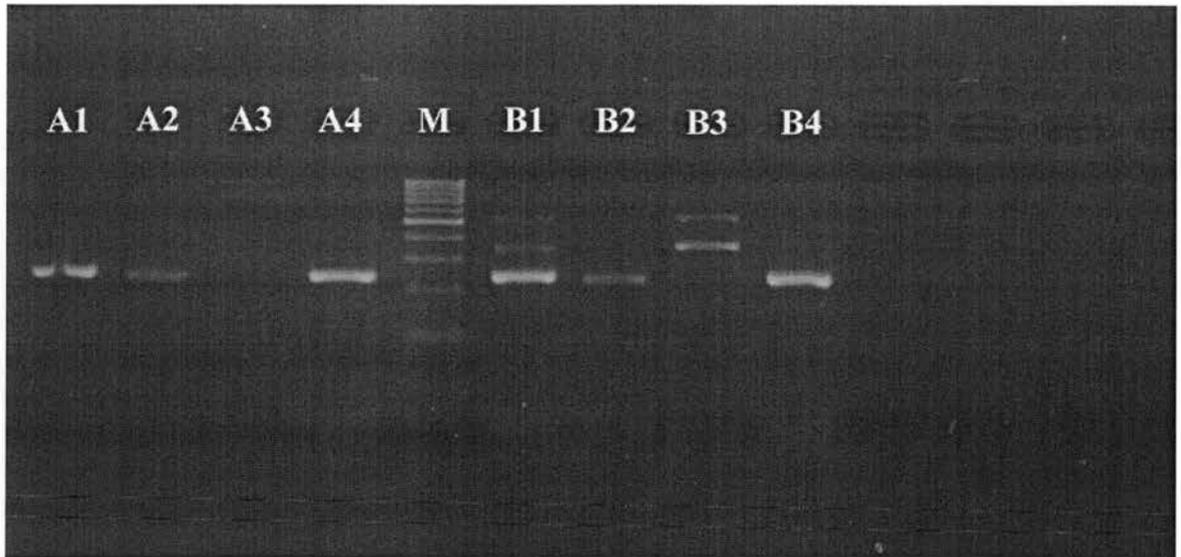


Table 3.4 Thiolester degenerate primer combinations amplifying DNA fragment of appropriate size.

Primers	Amplification	Successful conditions			
		Template	Nested template	Anneal °C	Primer conc µM
E5	✓	G, H, I & J	✗	55	5
E6	✓	G, H, I, J, K & M	E5 & F7	40, 42 & 45	2 & 5
E7	✗				
E8	✓	G	✗	50	5
E9	✓	G	✗	45	5
F5	✓	G, K, L & M	✗	40, 42 & 45	5
F6	✓	A, B, G, K, L & M	✗	40, 42 & 45	5
F7	✓	G	✗	50	5
F8	✗				
F9	✗				
G5	✗				
G6	✗				
G7	✓	G	F7	50	5
G8	✗				
G9	✗				

Figure 3.5 Thiolester DNA sequences with amino acid translation. Primer sequences removed. Thiolester 1 was amplified from all the hepatopancreas templates apart from L and the larvae. Thiolester 2 was amplified from the hepatopancreas template G (non-stimulated) only.

Thiol1

```

1      CTCGGGTTTGCGCCAGATGTGTTTCGTGACTCTCTACCTCCACTCGGC GGGCAAGCTCGAC
1      L G F A P D V F V T L Y L H S A G K L D

61     GCGCAACGAGAGCAAAGCTTTCAAACATTTCCAGACTGGTTACTCTAATGAACTAAAC
21     A A T R A K A F K H F Q T G Y S N E L N

121    TACAAGCACAGAGATGGATCATTTCAGTGCATTTCGGTGAAGGGGACGCCTCAGGCAGCACA
41     Y K H R D G S F S A F G E G D A S G S T

```

Thiol2

```

1      GGATCTGGCATCGGCTTGGTAACCTTGGACTCACTCCTTGGCCCAGTGATGGCTTCTTCA
1      G S G I G L V T L D S L L G P V M A S S

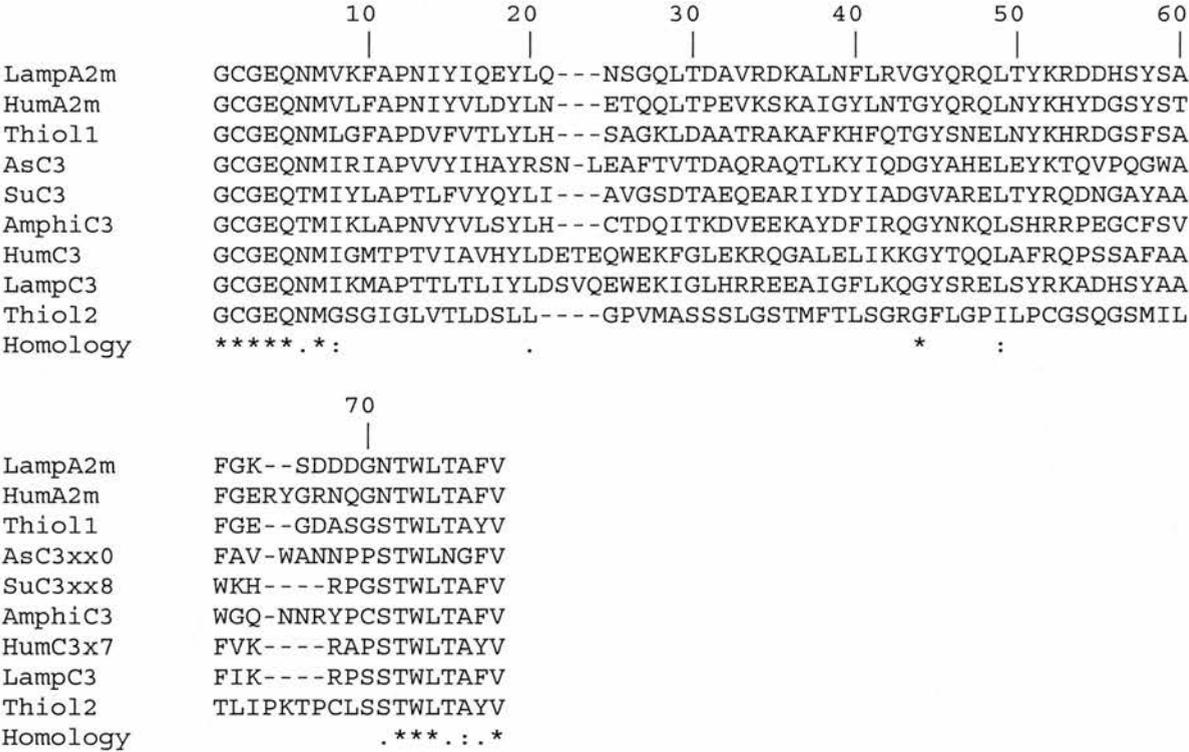
61     TCTTTAGGTTCAACAATGTTTACATTGTCAGGCAGAGGTTTCTTGGGTCCGATTTTACCT
21     S L G S T M F T L S G R G F L G P I L P

121    TGTGGATCCCAGGGAAGCATGATTTTAACTTGGATACCCAAGACACCTTGTCTGAGTAGC
41     C G S Q G S M I L T L I P K T P C L S S

181    ACA
61     T

```

Figure 3.6 Multiple alignment using CLUSTAL W illustrating homology between thiolester 1 and 2 and other corresponding a2m and C3 thiolester sequence fragments. Asterisks below the sequence indicates positions where all the sequences share the same amino acid residue, two dots indicates conserved amino acid substitutions, one dot indicates semi-conserved amino acid substitutions. Degenerate primer sequences remain at the beginning and end of each sequence. Huma2m, human alpha2 macroglobulin; Lampa2m, lamprey alpha2 macroglobulin; thiol1, thiolester 1; thiol2, thiolester 2; AsC3; ascidian C3; AmphiC3, amphioxus C3; SuC3, sea urchin C3, LampC3, lamprey C3; HuC3, human C3.



3.4 Discussion

Templates synthesised from the whole animals proved to be unsuccessful with no sequences isolated from this tissue. The RNA was of good quality but the proportion of specific mRNA within this pool would have been less than the hepatopancreas template. House keeping genes, expressed at a relatively constant level in most cells to maintain basic metabolic activities, are often used as an internal standard to quantify the amount of a specific RNA (Chelly *et al.*, 1990). Consequently, the mix of cells from whole animal tissue in the present study could dilute the amount of specific RNA from those cells transcribing immune proteins, perhaps to a level that RT-PCR could not detect.

Problems occurred with several individual primers amplifying a fragment of DNA for both the serine protease and the thiolester sets of degenerate primers used with *C. intestinalis*. The high level of degeneracy and low annealing temperature, especially for the C3 primer combinations, often allowed mis-priming and non-specific amplification. This accounted for the low level of clones transformed with sequences amplified using the forward and reverse primers. From all the successful primers, clones were detected after controls containing DNA amplified from both primers. In the majority of cases however, the sequence insert was nonsense. This may have been because the precise target sequence is unknown, corresponding to the high level of degeneracy required in the primer design. Also as the annealing temperature had to be decreased often to over 15 °C below the T_m (melting point) of the primers, non-specific annealing became likely. This may have been a consequence of the degenerate sequence not matching completely.

All the serine protease fragments found were isolated from the untreated hepatopancreas template G (section 3.2.2). Although the same fragments were also found in some of the stimulated templates, no novel fragments were isolated from this tissue i.e. hepatopancreas cDNA template G from untreated animals provided a template for all the serine protease gene fragments isolated. Other primer combinations were successful in amplifying good sequences from non-stimulated animal tissue. A comparison of the intensity of the bands obtained with equivalent primers from hepatopancreas tissue using animals treated with different amounts of LPS at different incubation times shows that LPS may have down-regulated the transcription of genes detected by these primers. In *C. intestinalis* LPS challenge is known to activate an immune response by increasing serum protease activity, prophenoloxidase activity and phagocytosis (Smith and Peddie, 1992; Jackson and Smith, 1993). In *S. purpuratus* LPS treatment has been shown to up-regulate genes (Smith *et al.*, 1995). It is also likely such a process involves down-regulation of genes that are not essential, keeping the energetic cost of transcription down. This process may have been observed in the present study as several serine proteases, including those involved in the complement system, contain the conserved regions against which the degenerate primers were designed (Appendices 3 & 4). Serine proteases are not exclusively involved in immunity. They may function, for example, as digestive enzymes (e.g. trypsin) that are likely to present in the hepatopancreas RNA. Such non-immune related serine protease genes might be down-regulated during immune challenge, thus providing less template after cDNA synthesis and consequently, less DNA amplification from fewer of the primer combinations.

An incubation time of 24 h after LPS challenge appears to have reduced transcription of all genes corresponding to the primers at the higher LPS dose of 2.5 µg to a level RT-

PCR could not detect (template L) (Table 3.3 & 3.4). None of the serine protease primer combinations amplified DNA fragments from *C. intestinalis* and the thiolester primers that produced a product had mis-primed. This may indicate that the LPS doses of 0.25 and 2.5 μg were too high, and after 3 and 24 h incubation periods, recovery was insufficient to observe the up-regulation of these genes using RT-PCR. As animals were collected from the wild their previous antigenic experience is unknown, so it is impossible to know if the up-regulation of genes had already occurred in the RNA used for cDNA synthesis. Even at the lower dose of 10 $\mu\text{g ml}^{-1}$ (template K) serine protease amplification was unsuccessful and only thiolester 1 was found.

Template M (section 3.2.2) synthesised from larvae mRNA produced DNA fragments using both serine protease and thiolester primers but none differed from those isolated from the hepatopancreas tissue. Hepatopancreas RNA provided nine serine protease gene fragments, two of which showed significant identity to digestive proteases after alignments, and the remainder were homologous to immune proteins from other animals. This illustrates that the hepatopancreas does transcribe genes involved in the immune response.

All seven serine protease sequences showing homology to immune proteins were amplified using the same forward primer D. Primer D was designed against the Bf sequence from lamprey, the 5' end started with a leucine rather than valine and the 3' end stopped 2 residues after the histidine, a cysteine and a phenylalanine (Table 3.1). The forward primers designed against other invertebrate Bf and MASP sequences failed.

Alignment of all seven serine protease sequences isolated from *C. intestinalis* (Fig. 3.2) shows they all contain a similar serine protease region (D (VI) A L) between the primer sites at each end. However, there is little other conservation. When the conserved primer regions are removed, there are a number of conserved amino acids between invertebrate complement homologues, vertebrate complement homologues and the seven serine proteases isolated from *C. intestinalis*. All the *C. intestinalis* sequences (Fig. 3.2) are homologous to each other and to complement protein homologues (Fig. 3.3), and thus deemed to be from the chymotrypsin protein superfamily that contains the complement serine protease homologues from all species (Nonaka, 1997). This provides sufficient evidence to justify further analysis as these genes are likely candidates for complement protein homologues.

Two thiolester sequences were amplified with the primer combination E6 from the hepatopancreas tissue at a low temperature of 43 °C (Table 3.4). Like the serine protease amplification, E primer was designed against the general thiolester site GCGEQNM (Nonaka and Takahashi, 1992), and not against an invertebrate complement protein homologue thiolester region (Appendix 4; Table 3.2). This sequence indicates that *C. intestinalis* thiolester containing proteins may be different from the Japanese ascidian, *H. rorezi*. Reverse primer 6 was also a more general primer and spanned only 6 amino acids.

Only two individual thiolester containing sequences were isolated from all the cDNA templates from *C. intestinalis*. All genes in the alpha2 macroglobulin (a2m) family contain a thiolester and a2m is likely to be present in *C. intestinalis*. Alignment of thiolester 1 and 2 shows they are homologous to both C3 and a2m (Fig. 3.6), but without

more information from the 3' and 5' end it is not possible to determine which fragment is likely to come from which gene. The 5' end of C3 should have an alpha/beta-chain processing site (Nonaka *et al.*, 1998) and the 3' end should extend further than a2m sequences.

In summary, after alignments and homology searches seven serine protease and two thiolester sequences were isolated as candidates for complement proteins from larvae and hepatopancreas tissue. Although the same sequences were amplified from the hepatopancreas cDNA from LPS treated animals, no novel sequences were found that were not present in untreated animals.

Chapter 4

Rapid Amplification of cDNA ends from Candidate Complement Gene Fragments

4.1 Introduction

Rapid amplification of cDNA ends (RACE) amplifies cDNA regions corresponding to the 3' and 5' end of the mRNA (Frohman *et al.*, 1988). RACE has been successfully used to isolate rare transcripts (McPherson and Møller, 2000) but requires sequence information for a region of a gene to enable the design of gene specific primers corresponding precisely to the gene of interest. Sequence information on nine novel genes from *C. intestinalis* was discovered by RT-PCR in Chapter 3 providing enough information for RACE.

Other studies that initially used RT-PCR to gain sequence knowledge of likely complement gene fragments have used cDNA library construction and screening rather than RACE to isolate the 3' and 5' end of the gene. The only complement homologue full-length cDNA that has been isolated using RACE is in the lamprey (Nonaka and Takahashi, 1992). Complementary DNA (cDNA) libraries were used in preference to RACE as the use of cDNA libraries were pioneered before the advent of RACE as the application has been proven (for a review of the literature see Sambrook and Russell, 2001). Until recently RACE was problematic in that 5' ends were often not full length and the multiple amplified fragments appeared as smears on agarose gels, rather than as distinct bands (McPherson and Møller, 2000; Sambrook and Russell, 2001). These problems can now be overcome (described in section 4.2.5) so the capture of a specific 5' and 3' cDNA fragments and their amplification is now a reliable technique.

RACE can also overcome some of the problems associated with cDNA libraries e.g. where full-length 5' fragments are missing (Sambrook and Russell, 2001). Partial clones

containing incomplete 5' fragments arise when the reverse transcriptase fails to reach the end of the mRNA. This problem increases in likelihood the greater the length of the mRNA template, the more secondary structure the mRNA template contains and the rarer the transcript (Sambrook and Russell, 2001). The only solution is the manufacture and screening of a new library. RACE provides a template for multiple screenings and requires fewer stages than cDNA library construction, thereby increasing the chances of finding rare transcripts and more complete 5' ends of the gene.

The aims of this chapter are to adapt the RACE technique to finding the complete 5' and 3' ends of the seven serine protease and 2 thiolester containing gene fragments isolated from the hepatopancreas cDNA of *C. intestinalis*.

4.2 Materials and Methods

All Eppendorf tubes and pipette tips were sterilised as in the Chapter 2 to ensure they were RNase free. Only high quality total RNA determined by agarose gel electrophoresis (section 2.2.7) was used for reverse transcription (RT) to ensure that sample degradation would not prevent the discovery of rare transcripts.

4.2.1 Total RNA Samples

RACE template was synthesised from mRNA using a reverse transcriptase enzyme, as in section 3.2.1. Only RNA isolated from the hepatopancreas of non-stimulated *C. intestinalis* was used as a template for RACE as this RNA pool contained all the serine protease and thiolester sequences isolated by RT-PCR in Chapter 3 (template G, section 3.2.1). Only high quality samples were selected, as determined in Chapter 2. Total RNA samples rather than purified mRNA samples were used in reverse transcription because of the concentration of mRNA, obtained in preliminary experiments using a total RNA sample of 200 µg, yielded only between 0.5-1 µg mRNA. Also, the probable rarity of the transcripts of interest in the mRNA, deduced from the need to perform secondary PCR re-amplification in RT-PCR (section 3.2.4), meant that RNA quality was paramount. Fewer stages of RNA manipulation conserves RNA integrity prior to reverse transcription in the presence of RNase inhibitor.

4.2.2 5' RACE Template

RACE was performed using the SMART™ RACE cDNA amplification kit (Clontech, Hampshire, UK). Synthesis of the 5' RACE template exploits the terminal transferase activity of reverse transcriptase derivatives to add 3-5 residues (predominantly dC) to the 3' end of the first strand cDNA. The Smart II oligo (supplied in the RACE kit) has a stretch of dG residues that binds to the terminal dC stretch of the cDNA and serves as an extended template for reverse transcription. When the reverse transcriptase switches from the RNA template to the cDNA a complete cDNA copy of the original RNA is synthesised with the Smart II oligo sequence at the end. This sequence corresponds to the forward 5' RACE primer used for 5' RACE PCR cycling along with the reverse gene specific primers.

The 5' RACE protocol used in the present study was adapted from the RACE kit manual supplied by the manufacturer's (Clontech). In a sterile microcentrifuge tube, 1 µg of good quality total RNA (determined from Chapter 2), 1 µl of 5' CDS primer (10 µM) (supplied in the kit) and 1 µl of the Smart II oligo, (supplied in the kit) made up to a volume of 4 µl with nuclease-free water, were heated at 70 °C for 2 min to reduce secondary structure and placed immediately on ice. Once chilled, 2 µl of 5x first strand buffer (250 mM Tris-HCL; 375 mM KCl; pH 8.3), 1 µl DTT (20 mM), 1 µl dNTP mix (10µM each nucleotide), 1 µl RNase inhibitor (RNase OUT Recombinant RNase inhibitor, 40 units/µl) (Invitrogen; formerly Life Technologies, Paisley, UK) and 1 µl reverse transcriptase (200 units) was added to the tube creating a final transcription volume of 10 µl. The reverse transcriptase used was Superscript II (Life Technologies), as used for cDNA synthesis in RT-PCR for the same reasons as previously outlined

above (section 3.2.1). The addition of RNase inhibitor was necessary to guarantee RNA integrity during transcription ensuring full-length cDNAs were synthesised. The contents of the tube were then spun briefly in a microcentrifuge to pool the contents at the tube bottom, before incubation at 42 °C for 1.5 h in an air incubator to reduce the effect of evaporation on the reaction volume. A final incubation at 70 °C for 15 min ensured the reverse transcription enzyme was inactivated. The final reaction (10 µl) was then diluted with 100 µl of tricine-EDTA buffer (pH 8.0) and stored at – 20 °C until use. This 5' template was stored no longer than 6 months.

A control template was also made alongside the 5' template. Exactly the same reaction mix and methods were used with the only difference being the addition of nuclease-free water in the place of the reverse transcriptase. This was used in the RACE PCR cycling with the full template to indicate that DNA fragments amplified are from cDNA and not from any contaminating genomic DNA, which may have remained from the RNA extraction.

4.2.3 3' RACE Template

Synthesis of the 3' RACE template relies on the same principal as reverse transcription of the template in RT-PCR (section 3.2.1). An oligo d(T) primer exploits the poly (A⁺) tail on the 3' end of the mRNA providing the point from which the reverse transcriptase synthesises. The difference between the RACE oligo d(T) primer and the primer used in reverse transcription is the presence of two degenerate oligonucleotides at the 3' end. This ensures that synthesis always starts from the same point, ensuring that the cDNA population of the gene of interest are precisely the same length, producing a discrete

amplified DNA fragment after RACE PCR. The RACE oligo d(T) primer also contains a region on the 5' end, which corresponds to the reverse RACE primer sequence used in the RACE PCR cycling, along with the forward gene specific primers.

The 3' RACE cDNA synthesis protocol was modified from the procedure described in the RACE kit manual. Briefly, 1 µg of good quality total RNA (removed from -70 °C and kept on ice until used) and 1 µl of the 3' cDNA synthesis (CDS) primer (10 µM) (supplied in the kit) were mixed in a 0.5 ml sterile microcentrifuge tube and the volume was made up to 4 µl with nuclease-free water. This mix was heated at 70 °C for 2 min to reduce secondary structure then placed immediately on ice. The transcription procedure for the 5' template was repeated for the 3' template and stored at -20°C for no longer than six months. The template control was as in the 5' template above (section 4.2.2) replacing the reverse transcriptase with nuclease-free water. This ensured that amplified DNA fragments were from the cDNA of transcribed genes and not from contaminating genomic DNA.

4.2.4 Gene Specific Primer Design

Each of the gene specific RACE primers (Table 4.1) was designed based on the gene fragments isolated by RT-PCR (Fig 4.2) except for the thiolester 1 5' primers, where sequence information was obtained from 3' RACE before 5' RACE was performed, maximising the amount information available for optimal primer design and allowing some characterisation of the gene.

Primary and nested primers were designed for both the 5' and the 3' RACE (Table 4.1) (Fig. 4.1) to allow for the further analysis of amplified fragments or the amplification of small amounts of DNA amplified from the primary RACE reaction. If possible, the 5' and 3' primers were designed to amplify fragments with an overlapping region so the 5' and 3' RACE primers could be used as a positive control together, amplifying a fragment of the cDNA of interest and constructing the final cDNA sequence is easier if an overlapping region is present on the 5' and 3' fragments (Fig. 4.1).

Primers GSP1 and NGSP1 (Table 4.1) correspond to the gene specific primer and nested gene specific primer respectively for the 5' RACE amplification. Primers GSP2 and NGSP2 (Table 4.1) correspond to the gene specific primer and nested gene specific primer respectively for the 3' RACE amplification.

Table 4.1 RACE primers (5' to 3') and their melting points (T_m) used for the 5' RACE and 3' RACE of the gene fragments SP1-6 and thiolester 1 & 2 isolated by RT-PCR in Chapter 3 (section 3.3).

Fragment	Primer	Sequence	T _m
SP1	GSP1	GGGCGAATATTGCTTGTCCA	57.3 °C
	NGSP1	CGATCTAATTGTATCAATGCGACA	57.6 °C
	GSP2	CGGAAAATGAGTACAGCATTCAC	58.9 °C
	NGSP2	GCATTCACAAGTTTTTCGGCA	55.3 °C
SP2	GSP1	TGAACAAACCGTGTAGCGAC	57.9 °C
	NGSP1	CGTTTCTTGTTGCAGGGTACTGAG	62.7 °C
	GSP2	CGAGTTGGAGATTATTTCAACCG	58.9 °C
	NGSP2	GGATTCTATGGTTGAAGAGTCACA	59.3 °C
SP3	GSP1	AAACACAACGGGGCGACTCAGCTT	64.4 °C
	NGSP1	CAGCGGTGAAACCGGGGTGTACTATGA	68.0 °C
	GSP2	TCGGGACCACCAGGAGCTCTCATCTTA	68.0 °C
	NGSP2	CCACTAGGAGGCAACGAAGAGAGGTTGA	68.0 °C
SP4	GSP1	CAAAAGCCGCATCAATCAGAGCATC	64.8 °C
	NGSP1	AGCAATGCAAGGTCGTTGTTGGGA	63.0 °C
	GSP2	AACAAACAACAACCCAAGCACCA	58.9 °C
	NGSP2	CGTCATATTGGGTGTTGTTGACACA	61.3 °C
SP5	GSP1	AGACGGGGAATACCCGAGACGA	64.0 °C
	NGSP1	ATCGGCTCGGCAACCTTGAT	59.4 °C
	GSP2	ACATTGCTTCGACCATGTGACG	60.3 °C
	NGSP2	GCAGGTTGATAAAACCATTCTCGG	61.0 °C
SP6	GSP1	CACCATGAGTGAGGTTGCCTGCACAT	66.4 °C
	NGSP1	GGAGATCAACTTGTTGTAGAGACGGCG	66.5 °C
	GSP2	CATCCATCGACATGGTTTGG	57.3 °C
	NGSP2	GAGCGTAGCCGGCAGTACAGAATCGAC	69.5 °C
SP7	GSP1	CACTGCTTTGGTGAATACTATTTGCGTTGAT	64.2 °C
	NGSP1	CAATGTAACATCAGTTAAGACTTTCCCAACAA	63.1 °C
	GSP2	CATCAGTCCATGTGTTTGTGCCAAAGTC	65.3 °C
	NGSP2	CCATACCAACAACATTCTCTCGTCTCCCA	66.7 °C
Thiolester 1	GSP1	CGAGCGAACATAAAGCACTTAGCAGCG	66.5 °C
	NGSP1	TTTGAAAGCTTTTGCTCTCGTTGCGGC	65.0 °C
	GSP2	TTTGCGCCAGATGTGTTTCGTGACTCTC	66.5 °C
	NGSP2	CCGCAACGAGAGCAAAAGCTTTCAAAC	65.0 °C
Thiolester 2	GSP1	-	
	NGSP1	-	
	GSP2	GGATCTGGCATCGGCTTGGTAAACCTTG	68 °C
	NGSP2	TCACTCCTTGCCAGTGATGGCTTC	68 °C

Figure 4.1 Gene fragments isolated by RT-PCR (section 3.3) and the RACE primers designed from them. **GSP1**; **NGSP1**; **GSP2**; **NGSP2**. Outline text represents an area where the primers overlap.

Serine protease 1

1 TTGCAACAAATAACGGAAAATGAGTACAGCATTTCACAAGTTTTTCGGCAGTATTTGGATTG
 61 TTTTCGATTGAATTTGCAACACAACACACAGAGAATTGGTTTCAAAGAACATTTATTCAT
 121 TCGGATTTTCAAAGCGCACATTTAACTTTTAGAAACGATGTCGCATTGATACAATTAGAT
 181 CGAAAAATACAA TGGACAAGCAATATTCGCC

Serine protease 2

1 ACAAGTAGAGTAAAAAGAGAAAGAAAGAAACACTTTGTTTCGAGTTGGAGATTATTTCAAC
 61 CGAGATAACCTTCCATAGTCAGGATTCATGTTGAAGAGTCACATGATATAGCAATT
 121 AGCCAAATTTATATTCATGAGGGTTTTACTCAGTACCCTGCAACAAGAAACGATATTGCT
 181 TTAATTAAC TAAGCGAACCGGTGTCGCTAACACGGTTTTGTTCAA

Serine protease 3

1 AGATCCGTATCTTACTCCGGTCTCCTTGTTTACC TCGGGACCACCAGGAGCTCTCATCTT
 61 ACACATCTTGATACCACTAGGAGGCAACGAAGAGAGGTTGAACAGATCATAGTACACCCC
 121 GGTTCACCGCTGAGTATTTGAACGACGTTGCATTAATAAAGCTGAGTCGCCCCGTTGTG
 181 TTTAATGACATCATCACC

Serine protease 4

1 GCATCTAT AACAAACAACAACCCAAGCACCATTAACGTCATATTGGGTGTTGTTGACACA
 61 ATTGATTCAGGAAACATACATGAACAATCTTTTTCTGTTACAAGACTTATAATTCATCCA
 121 AACTACAATT TCCCAACAACGACCTTGCAATTGCTACAAC TGGACCATGATGCTCTGATT
 181 GATGCGGCTTTTGTGAAA

Serine protease 5

1 ACATTGCTTCGACCATGTGACGTCACAAAAGCAGGTTGATAAAACCATTCTCGGCTTTGG
 61 GACTTCCCAGCTTTGGCGCTGACGCGCTCTCGCACCTGCTGCGTGATAGTGACGTCAC
 121 CGTGACGTCAATGGATGACGTCACCTGGTGCGAGCGTGATCCGTTTAAAGTCGCATCTACAG
 181 CCACCTACTTACGGCGAAAACCTGGATAGTGATATCGTATTGATCAAGGTTGCCGAGCC
 241 GATCACGTGGTCGTC TCGGGTATTCGCCGTCT

Serine protease 6

1 CTTAAACATGACATCCATCGACATGGTTTTGGTCATCGTTACATTAGGAATGCTACGTCAG
 61 CATGTCAGTTT GAGCGTAGCCGGCAGTACAGAATCGACAAAAGGATTGTCATACATCCA
 121 GAATTCGTTTTCCCGCATTATGACGTCGCGTTAATCGAAGTGATCGCGCTTTTGACGTT
 181 ACTGGCGTTTTTGTGAGG

Serine protease 7

1 CTGACTGCCGCGCACTGTTTGCAAATGATGAAATAAATATAACATCAGTCCATGTGT^{TTT}
61 GTTGGGAAAAGTCTTAACTGATGTTACATTGATTGAACCATACCAACAACATTCTCTCGTC
121 TCCCATGTTGCATTTTCATGAGAATTACGATCCCGATAATTTAAATTCAGATATCGCCATT
181 CTTACGTTATCAACGCAAATAGTATTCACCAAAGCAGTGAGCCGCAGCTTAG

Thiolester 1

1 CTCGGGTTTGCGCCAGATGTGTTTCGTGACTCTCTACCTCCACTCGGCGGGCAAGCTCGAC
61 GCCGCAACGAGAGCAAAGCTTTCAAACATTTCCAGACTGGTTACTCTAATGAACTAAAC
121 TACAAGCACAGAGATGGATCATTTCAGTGCATTCGGTGAAGGGGACGCCTCAGGCAGCACA
181 TTGCTCACTGCGTTCGCTGCTAAGTGCTTTATGTTTCGCTCG

Thiolester 1

1 CTCGGGTTTGCGCCAGATGTGTTTCGTGACTCTCTACCTCCACTCGGCGGGCAAGCTCGAC
61 GCCGCAACGAGAGCAAAGCTTTCAAACATTTCCAGACTGGTTACTCTAATGAACTAAAC
121 TACAAGCACAGAGATGGATCATTTCAGTGCATTCGGTGAAGGGGACGCCTCAGGCAGCACA
181 TTGCTCACTGCGTTCGCTGCTAAGTGCTTTATGTTTCGCTCG

Thiolester 2

1 GGATCTGGCATCGGCTTGTAACCTTGACTCACTCCTTGGCCAGTGATGGCTTCTTCA
61 TCTTTAGGTTCAACAATGTTTACATTGTCAGGCAGAGGTTTCTTGGGTCCGATTTTACCT
121 TGTGGATCCCAGGGAAGCATGATTTTAACCTTGATACCAAGACACCTTGTCTGAGTAGC
181 ACA

4.2.5 Primary RACE PCR

RACE PCR was performed using the advantage™ cDNA polymerase mix (Clontech). This mix contains a primary DNA polymerase (KlenTaq-1 DNA polymerase) for amplification, a minor amount of secondary ‘proof reading’ *Taq* DNA polymerase and a *Taq* antibody to provide automatic ‘hot start’ PCR (Kellogg *et al.*, 1994). Inclusion of a proof reading enzyme serves two functions: First, *Taq* polymerase alone may mismatch at a rate of approximately one base in 1500 (Sambrook and Russell, 2001). This can lead to significant errors in the final deduced protein sequence after amino acid translation. Second, the use of two enzymes simultaneously allows for much longer amplification than with *Taq* alone, which can only extend to 3000 bp (Barnes, 1994; Cheng *et al.*, 1994).

RACE PCR mixes (25 µl in volume) were set up in 0.2 µl thin walled PCR tubes (Axygen Scientific, California, USA). For the primary RACE, 1.25 µl of the required template (5’ or 3’ cDNA from section 4.2.2 and 4.2.3) was pipetted into the bottom of the PCR tube. Then, 0.5 µl of the gene specific primer (10 µM stock for a final concentration of 0.5 µM) and 2.5 µM of the universal primer mix (UPM) (long primer 0.2 µM, short primer 1 µM) (supplied in the kit) was pipetted onto the inside walls of the tube. A master mix was then made with the remaining PCR components to ensure an accurate mix, as in section 3.2.3, containing 2.5 µl of 10x buffer (400 mM Tricine-KOH, pH 9.2; 150 mM KOAc; 35 mM Mg(OAc)₂; 37.5 µg/ml bovine serum albumin) (supplied with the Advantage cDNA polymerase), 0.5 µl dNTP mix (10 µM each nucleotide), 0.5 µl of cDNA polymerase mix and 17.25 µl of nuclease-free water for each individual reaction of 25 µl. Tubes were placed on ice and 21.75 µl of the master

mix was added to each. After brief centrifugation the tubes were placed into the thermal cycler (same as section 3.2.3) pre-heated to 94 °C.

Thermal cycling was performed in either a Touchdown or Sprint Thermal Cycler (Hybaid Ltd, Middlesex, UK). The only variables in the cycling conditions were the primer annealing temperatures, the number of cycles performed and the primer extension time. Annealing temperatures ranged from within 1 °C of the primer T_m (Table 4.1) to 10 °C below the T_m (Table 4.1). The cycle number began with 25 cycles and was increased up to 45 cycles for both the primary and nested RACE depending on the previous results from fewer cycles. Primer extension time was 4 min for the serine protease sequences to allow fragments over 2000 bp to be amplified and 6 min for the thiolester sequences to allow fragments over 3000 bp to be amplified. The basic cycling profile was;

Denaturation at 94 °C for 7 min

25- 45 cycles

Denaturation-94 °C for 30 sec

Primer annealing-variable °C for 30 sec

Primer extension-72 °C for 4-6 min

Elongation-72 °C for 10 min

Storage at -20 °C

If touchdown thermal cycling was performed the cycling profile was;

Denaturation at 94 °C for 7 min

5 cycles

Denaturation-94 °C for 30 sec

Primer annealing and extension-72 °C for 4-6 min

5 cycles

Denaturation-94 °C for 30 sec

Primer annealing -70 °C for 30 sec

Primer extension-72 °C for 4-6 min

25-40 cycles

Denaturation-94 °C for 30 sec

Primer annealing -68 °C for 30 sec

Primer extension-72 °C for 4-6 min

Elongation-72 °C for 10 min

Storage at -20 °C

The following control reactions were run in parallel:

1. Template made with RT with gene specific primer only

2. Template made with RT with RACE primer only
3. Template made with RT and no primers
4. Template made without RT with forward and reverse primers
5. No template with forward and reverse primers

The control reactions revealed which DNA fragments were amplified using both forward and reverse primers from cDNA rather than any contaminating genomic DNA from the RNA extraction or from primer-dimer interactions.

4.2.6 Nested RACE PCR

Using the primary RACE reaction, a nested template was then made by diluting 2.5 μ l of the first reaction and with 122.5 μ l of nuclease-free water. A reaction mix was then made as above (section 4.2.5) using 2.5 μ l of the diluted primary RACE as the template, the appropriate nested gene specific primer (Table 4.1) and in place of the universal primer mix the nested universal primer (NUP) 1(supplied in the kit) was used. This primes from a region on the shorter universal primer that is only be present on DNA amplified from correctly transcribed cDNA (section 4.2.5). Cycling was carried out as in section 4.2.5.

4.2.7 Analysis of Results

After RACE cycling, amplified DNA fragments were run on 1 % agarose gels as in section 3.2.6 and visualised under U.V light.

4.2.8 Cloning and Plasmid Extraction

RACE PCR products were excised from the agarose gel and cloned using a Topo® cloning kit (Invitrogen) as in Chapter 3 (section 3.2.7). Colony screening was undertaken with controls, as in section 3.2.7, to determine those colonies that contained an insert of the correct size that had the forward and reverse primers at either end. These control reactions contained the following primers:

1. M13 forward and reverse
2. Gene specific primer and RACE primer
3. Gene specific primer only
4. RACE primer only

Plasmid extraction was performed using the Wizard™ plus SV minipreps DNA purification system (Promega UK, Southampton, UK). This protocol was adapted from the kit manual. A single isolated colony was used to inoculate 5 ml of Luria-Bertani (LB) broth (Fluka-Sigma-Aldrich, Dorset, UK) containing $50 \mu\text{g ml}^{-1}$ ampicillin (the same concentration as the LB plates) and cultured in a shaking water bath overnight (12 – 16 h). Antibiotics were added to ensure only bacteria containing the plasmid of interest were cultured.

After incubation, the cells were pelleted at 10000 g for 5 min at room temperature and the supernatant removed. The protocol provided by the manufacturer was then followed. The cells were resuspended in 250 μl resuspension buffer (supplied in the kit) by pipetting and transferred to a sterile 1.5 ml Eppendorf tube. To this suspension 250 μl of

cell lysis buffer (supplied in the kit) was added and mixed by four inversions of the tube. This was incubated until the suspension cleared or for five minutes. Incubation for over five minutes can disrupt the plasmid DNA. To this, 10 μl of alkaline protease solution (supplied in the kit) was added, mixed by inversion four times, and incubated for no longer than five minutes. This was used to remove endonucleases and other proteins that can detrimentally affect the quality of the plasmid DNA. Neutralisation solution (350 μl) was added and immediately mixed by inversion four times. The resulting bacterial lysate was centrifuged at 14000 g for 10 min at room temperature and the supernatant containing the plasmid DNA was placed onto a spin column (provided in the kit). DNA was bound to the spin column matrix and wash several times before elution with 100 μl nuclease-free water. Further concentration of the DNA by precipitation was not necessary as a final concentration of plasmid DNA over 200 $\text{ng } \mu\text{l}^{-1}$ was consistently achieved.

4.2.9 Plasmid Quantification and Quality Control

Plasmid DNA quantification and quality control were carried out using agarose gel analysis against a known concentration of DNA markers (Lambda DNA *Hind* III digestion, Promega). The plasmid DNA was subjected to restriction enzyme digestion with *Eco*R1 (Promega). An *Eco*R1 site (G/AATTC) is situated either side of the pCR 2.1 – TOPO cloning site (Appendix 5) allowing the removal of the cloned PCR product from the vector. To a restriction mix of 10x buffer H (supplied with the restriction enzyme) (900 mM Tris-HCL (pH 7.5); 500 mM NaCl; 100 mM MgCl_2) and 0.2 μl acetylated bovine serum albumin (BSA) (10 $\mu\text{g } \text{ml}^{-1}$), 200 ng of plasmid DNA (approximately determined by reading the A_{260} value with an extinction coefficient of 50

using the same principle as section 2.2.6) was added. The mix was made up to a final volume of 19.5 μl and mixed gently by pipetting. Five units of *Eco*R1 (10 u μl^{-1}) restriction enzyme (Promega, Southampton, UK) was then added, bringing the final reaction volume to 20 μl before incubation for 3 h at 37 °C. The mix was then run on a 1 % agarose gel (as in section 3.2.5) and the concentration of the DNA and the size of the extracted insert was scrutinised.

The remaining plasmid DNA was stored at – 20 °C and could be used directly for sequencing.

4.2.10 Sequencing

All sequencing was performed using either MWG Biotech Sequencing Service (Ebersberg, Germany) or The Sequencing Service (University of Dundee, Dundee). The M13 forward and M13 reverse primers, corresponding to the insert flanking regions of the TOPO vector (Appendix 5), were used for the sequencing of each plasmid insert. Where sequencing of an insert had to be extended to obtain information about the entire length of the insert, sequencing extension primers were specifically designed based on the 3' end of the sequence so far determined. These were obtained from MWG Biotech and used these to perform a further round of sequencing. Each insert was sequenced a minimum of three times to confirm the correct nucleic acid at each position.

4.3 Results

4.3.1 SP1 RACE

Successful RACE cycling for the 5' end was performed using SP1-GSP1 (Table 4.1) at an annealing temperature of 54 °C with a 4 min extension time for 30 cycles, three faint bands of approximately 700, 1000 and 1100 bp were amplified from this primary RACE cycling following agarose gel analysis. Nested RACE, performed with SP1-NGSP1 (Table 4.1) at an annealing temperature of 54 °C with a 4 min extension time for 25 cycles, produced a single bright band of approximately 950 bp. Controls 1-5 (section 4.2.5) were all negative for the primary RACE and all controls were negative for the nested RACE apart from control 1 using SP1-NGSP1 (Table 4.1) alone, which produced a single bright band of approximately 1050 bp, 100 bp larger than the amplified DNA from the full reaction using SP1-NGSP1.

Successful RACE cycling for the 3' end, performed using SP1-GSP2 (Table 4.1) at an annealing temperature of 55 °C with a 4 min extension time for 30 cycles, followed by agarose gel analysis (section 4.2.6), revealed no visible amplified DNA fragments. Nested RACE, performed using SP1-NGSP2 (Table 4.1) at an annealing temperature of 55 °C with a 4 min extension time for 25 cycles, amplified two bands of approximately 600 bp and 900 bp. Controls 1-5 (section 4.2.5) were negative for the primary and nested RACE.

The 950 bp 5' fragment and both the 600 and 900 bp 3' fragments from the nested reactions above were successfully cloned and the plasmid purified before sequencing.

Colony controls revealed all the inserts had been amplified using the nested gene specific primer and the nested universal primer and was of the same size as the DNA extracted from the agarose gel.

Sequencing revealed that the 5' 950 bp DNA fragment had the correct primer sequences at both ends and the sequence corresponding to the SP1 RT-PCR fragment (section 3.3.1) was an identical match (Fig. 4.2). Both the 3' DNA fragments had the correct primer sequences at both ends, but only the larger 900 bp fragment contained the sequence corresponding to the SP1 RT-PCR fragment isolated in Chapter 3 (Fig. 4.2). Complete sequencing of both the 5' and 3' amplified fragments yielded a complete cDNA sequence of 2078 bp (Fig. 4.2) that encodes for a 519 amino acid protein.

4.3.2 SP2 RACE

Successful RACE cycling for the 5' end was performed using SP2-GSP1 (Table 4.1) at an annealing temperature of 54 °C with a 4 min extension time for 30 cycles. Agarose gel analysis revealed one faint amplified band of approximately 1000 bp. Nested RACE was performed with SP2-NGSP1 (Table 4.1) at an annealing temperature of 58 °C with a 4 min extension time for 25 cycles, producing a single bright band of approximately 750 bp. The 5' primary RACE controls 3, 4 and 5 (section 4.2.5) were negative but 1 and 2 had faint bands ranging from 500 to 1300 bp. Nested 5' RACE controls 1-5 (section 4.2.5) were negative.

Successful 3' RACE cycling, performed using SP2-GSP2 (Table 4.1) at an annealing temperature of 55 °C with a 4 min extension time for 30 cycles, amplified no visible

DNA fragments after agarose gel analysis. Nested RACE with SP2-NGSP2 (Table 4.1) at an annealing temperature of 55 °C with a 4 min extension time for 25 cycles amplified a single band of approximately 900 bp. Controls 1-5 (section 4.2.5) were negative for the 3' primary and nested RACE.

Both the 5' 750 bp fragment and the 3' 900 bp fragment from the 5' and 3' nested reactions above were successfully cloned and the plasmid purified before sequencing. Colony controls revealed the inserts had been amplified using the nested gene specific primer and the nested universal primer, and was of the same size as the DNA extracted from the agarose gel.

Sequencing revealed that the 5' 750 bp DNA fragment and the 3' 900 bp insert had the correct primer sequences at both ends and the sequence corresponding to the SP2 RT-PCR fragment (section 3.3.1) was an identical match (Fig. 4.3). Full sequencing of both the 5' and 3' amplified fragments yielded a complete cDNA sequence of 1832 bp (Fig. 4.3) that encodes for a 433 amino acid protein.

4.3.3 SP3 RACE

Primary 5' RACE cycling, using SP3-GSP1 (Table 4.1) at an annealing temperature of 63 °C with a 4 min extension time for 35 cycles, revealed no amplified DNA. Nested RACE with SP3-NGSP1 (Table 4.1) at an annealing temperature of 67 °C with a 4 min extension time for 30 cycles produced a single bright band of approximately 850 bp. Both the 5' primary and nested RACE controls 1-5 (section 4.2.5) were negative.

RACE cycling for the 3' end using SP3-GSP2 (Table 4.1) at an annealing temperature of 65 °C with a 4 min extension time for 35 cycles, revealed a single bright band was of approximately 1300 after agarose gel analysis. Nested RACE using SP3-NGSP2 (Table 4.1) at an annealing temperature of 65 °C with a 4 min extension time for 30 cycles produced a single bright band of approximately 1250 bp. Primary and nested RACE controls 1-5 (section 4.2.5) were negative.

Both the 850 bp 5' and the 1250 bp 3' fragments from the nested reactions above were successfully cloned and the plasmid purified before sequencing. Colony controls (section 3.2.7) revealed the inserts had been amplified using the nested gene specific primer and the nested universal primer and were of the same size as the DNA extracted from the agarose gel.

Sequencing revealed that the 5' and the 3' DNA fragments had the correct primer sequences at both ends and the sequence corresponding to the SP3 RT-PCR fragment (section 3.3.1) was an identical match (Fig. 4.4). Full sequencing of both the 5' and 3' amplified fragments yielded a complete cDNA sequence of 1907 bp that encodes for a 470 amino acid protein.

4.3.4 SP4 RACE

Primary 5' RACE cycling using SP4-GSP1 (Table 4.1), at an annealing temperature of 64 °C with a 4 min extension time for 40 cycles, amplified an approximately 700 bp DNA band. Nested RACE, performed with SP4-NGSP1 (Table 4.1) at an annealing temperature of 61 °C with a 4 min extension time for 30 cycles, amplified two bands of

approximately 670 bp and a fainter band of approximately 500 bp. The 5' RACE primary controls 1-5 (section 4.2.5) were negative. Nested RACE controls (section 4.2.5) were negative except for control 2 with the nested universal primer alone, which produced multiple banding from 1500 to 3000 bp. None of these bands were of the same size as the 670 and 500 bp bands from the nested RACE.

Successful RACE cycling for the 3' end was performed using SP4-GSP2 (Table 4.1) at an annealing temperature of 58 °C with a 4 min extension time for 35 cycles. Agarose gel analysis revealed a single faint amplified band of approximately 700 bp with smearing from 500 to 4000 bp. Nested RACE with SP4-NGSP2 (Table 4.1) at an annealing temperature of 58 °C with a 4 min extension time for 30 cycles produced multiple banding from 300 to 2300 bp with one distinct brighter band at 2400 bp. Controls 1-5 (section 4.2.5) were negative for the primary and nested RACE.

Both the 670 and 500 bp 5' bands and the bright 3' 2400 bp fragment from the nested reactions above were successfully cloned and the plasmid purified for sequencing. Colony controls (section 3.2.7) revealed only the 670 bp 5' insert had been amplified using the nested gene specific primer and the nested universal primer. Colonies containing the 500 bp fragment failed colony control 3 producing a 500 bp band thus indicating the 5' 500 bp fragment had been amplified using SP4-NGSP alone. Colony controls for the 3' 2400 bp fragment indicated this insert had been amplified using the SP4-NGSP2 and the nested universal primer and was of the same size as the DNA extracted from the agarose gel.

Sequencing revealed that the 5' 670 bp DNA fragment and the 3' 2400 bp insert had the correct primer sequences at both ends and the sequence corresponding to the SP4 RT-PCR fragment (section 3.3.1) was an identical match. Full sequencing of both the 5' and 3' amplified fragments produced a complete cDNA sequence of 3534 bp that encodes for a 1089 amino acid protein (Fig. 4.5).

4.3.5 SP5 RACE

Primary 5' RACE was performed using SP5-GSP1 (Table 4.1) at an annealing temperature of 63 °C with a 4 min extension time for 35 cycles. Agarose gel analysis revealed a faint 500 bp fragment was amplified. Nested RACE with SP5-NGSP1 at an annealing temperature of 58 °C with a 4 min extension time for 30 cycles produced a single bright band of approximately 450 bp. Controls 1-5 (section 4.2.5) were negative for the 5' primary and nested RACE.

RACE cycling using SP5-GSP2 for the 3' end of serine protease 4 at an annealing temperature of 56 °C with a 4 min extension time for 35 cycles amplified three faint fragments of approximately 1000, 650 and 600 bp. Nested RACE with SP5-NGSP2 at an annealing temperature of 58 °C with a 4 min extension time for 30 cycles produced a single bright band of approximately 600 bp. Controls 1-5 (section 4.2.5) were negative for the 3' primary and nested RACE.

Both the 450 bp 5' and 600 bp 3' fragments from the nested reactions above were successfully cloned and the plasmid purified before sequencing. Colony controls (section 3.2.7) revealed both the 5' and the 3' fragments had been amplified using the

nested gene specific primer and the nested universal primer, and were of the same size as the DNA extracted from the agarose gel.

Sequencing revealed that the 5' 450 bp insert and the 3' 600 bp insert had the correct primer sequences at both ends and the sequence corresponding to the SP5 RT-PCR fragment (section 3.3.1) was an identical match (Fig. 4.6). Full sequencing of both the 5' and 3' amplified fragments yielded a complete cDNA sequence of 1035 bp that encodes for a 225 amino acid protein.

4.3.6 SP6 RACE

RACE to isolate the 5' end of SP6 was unsuccessful. Several new primers were also synthesised for this purpose and used with the primers detailed in Table 4.1. Details of them all are not shown as they were unsuccessful. Although a band was consistently amplified using several of these primer combinations by primary and nested RACE, the size of these bands revealed that no amplification had taken place further than the 5' end of the original fragment isolated by degenerate PCR in Chapter 3. RNA and template integrity was checked and repeated and found to be of good quality. The 5' end of these gene found using RACE does not appear to have an initiation codon in the cDNA sequence before the serine protease domain and cannot, therefore, be a serine protease. The cDNA sequence of this protein does not correspond to the PCR fragment amplified in Chapter 3.

RACE cycling for the 3' end was performed using SP6-GSP2 at an annealing temperature of 54 °C with a 4 min extension time for 35 cycles. For the nested RACE,

2.5 μ l of the nested template was used with SP6-NGSP2 at an annealing temperature of 62 °C with a 4 min extension time for 30 cycles. Following agarose gel analysis (section 4.2.6), a faint DNA fragment was amplified of approximately 1450 bp from the primary RACE. Nested RACE using SP6-NGSP2 produced a single bright band of approximately 1400 bp.

Controls 1-5 (section 4.2.5) were negative for the 3' primary and nested RACE.

The 3' 1400 bp insert had the correct primer sequences at both ends and the sequence also corresponded to the SP6 RT-PCR fragment (section 3.3.1). Further sequencing was then performed to obtain the sequence information for the remaining 3' insert sequence. Figure 4.7 shows the full 1457 bp of this truncated sequence but the precise location of the coding region cannot be ascertained.

4.3.7 SP7 RACE

Primary RACE for the 5' end of SP7 using SP7-GSP1 at 63 °C for 35 cycles with a 3 min extension time amplified no visible DNA bands. Using this reaction as a template with the NUP and SP7-NGSP1 under the same cycling conditions at an annealing temperature 61 °C amplified a single DNA band of approximately 2600 bp. The controls 1-5 (section 4.2.5) were negative.

RACE to amplify the 3' end of this gene was performed using the 3' template with UPM and SP7-GSP2. Agarose gel analysis revealed that this reaction amplified a bright band of approximately 750 bp with faint bands and smearing ranging from 400 bp to 2000 bp.

Controls were negative apart from control 2 with the RACE UPM alone. This produced smearing also from 400 bp to 2000 bp with some faint banding visible. Using this diluted reaction as a template for nested RACE with the NUP and SP7NGSP2 amplified a single bright DNA band of approximately 700 bp with a single faint band of approximately 800 bp. All the controls were negative except for control 1 using SP7NGSP2 alone. This produced a faint band at approximately 720 bp.

Both the 5' 2600 bp band and the 3' bright 700 bp band from the nested reactions above were successfully cloned and the plasmids containing the insert were extracted and purified. Colony controls revealed that both the 5' and 3' inserts had been amplified using both gene specific primers and the SMART RACE primers.

Sequencing confirmed that these inserts were from the same gene as the SP7 RT-PCR fragment isolated in Chapter 3 (Fig. 4.8). The entire insert was sequenced and the full cDNA sequence was constructed by matching the overlapping fragments in the region of the RT-PCR fragment (Fig. 4.8). Full sequencing produced a cDNA sequence of 3943 bp encoding for a 1235 amino acid protein (Fig. 4.8).

4.3.8 Thioster 1 RACE

Primary 5' RACE cycling, using Thiol1-GSP1 (Table 4.1) at an annealing temperature of 62 °C with a 6 min extension time for 45 cycles, produced no visible DNA fragments after agarose gel analysis. Nested RACE, using Thiol1-NGSP1 (Table 4.1) at an annealing temperature of 62 °C with a 6 min extension time for 45 cycles, amplified one bright 3000 bp band. The 5' primary RACE controls 1-5 (section 4.2.5) were all

negative. Nested RACE control 2 (section 4.2.5) with the nested universal primer alone produced a single bright band of approximately 1800 bp, but all the other controls were negative.

The primer Thiol1-GSP2 (Table 4.1) for 3' RACE cycling at an annealing temperature of 62 °C with a 6 min extension time for 35 cycles amplified two bright DNA fragments of approximately 2200 and 2100 bp. Nested RACE, using Thiol1-NGSP2 (Table 4.1) at an annealing temperature of 62 °C with a 6 min extension time for 35 cycles, produced a faint band at 2170 bp and a smear from 3500 to 200 bp. Controls 1-5 (section 4.2.5) were negative for the 3' primary and nested RACE.

The 3000 bp 5' fragment from the nested reaction above was successfully cloned and the plasmid purified before sequencing. Both the fragments from the primary 3' RACE above were also successfully cloned. Colony controls (section 3.2.7) revealed that the 5' and the 3' inserts had been amplified using the nested gene specific primer and the nested universal primer, and was of the same size as the DNA extracted from the agarose gel.

Sequencing revealed that the 5' 3000 bp DNA fragment had the correct primer sequences at both ends and the sequence corresponding to the thiolester1 RT-PCR fragment (section 3.3.1) was an identical match (Fig. 4.7). The 3' 2200 bp insert had the correct primer sequences at both ends and the sequence also corresponded to the thiolester1 RT-PCR fragment (Fig. 4.7). The smaller 3' fragment of 2150 bp had the correct primer sequences at each end but the sequence did not correspond to the RT-PCR fragment in section 3.3.1. Full sequencing of both the 5' and 3' amplified fragments

produced a complete cDNA sequence of 5603 bp that encodes for an 1809 amino acid protein.

4.3.9 Thiolester 2 RACE

No amplification was observed with primers Thiol2-GSP2 or Thiol2-GSP2 at any of the annealing temperatures and cycling parameters tested. As this 3' amplification was unsuccessful no 5' amplification was attempted.

4.3.10 Summary

The complete cDNA sequences of all the serine protease RT-PCR fragments isolated in the present study (section 3.3) were found using RACE. Of the two thiolester sequences isolated in Chapter 3, only one was successfully extended by RACE. Different cycling parameters had to be optimised for each individual gene specific primer before the appropriate fragments were amplified. After primary RACE no amplification was observed for SP1 3', SP2 3' and SP3 5' RACE. However, after nested RACE bright and discrete bands were observed indicating a low level of gene expression. Faint bands were observed after primary RACE in SP1 5', SP2 5', SP4 5' SP5 5' and SP5 3' but at a very low intensity after 35 cycles, and often multiple banding was observed. The only reactions that produced a sufficient amount of DNA from the primary reaction were SP3 3' and SP4 3' after 35 cycles. This indicates a higher level of transcription of SP3 and SP4 than the other mRNA's. However, the 5' reaction for SP3 and SP4 did not produce the same level of amplification from the primary RACE. This may have been down to secondary structure with the template that can impede reverse transcription (Sambrook

and Russell, 2001), a lower intensity of 5' RACE product would be expected as fewer full length transcripts were synthesised.

Figure 4.2 Complete serine protease 1 cDNA sequence with amino acid translation determined from the 5' and 3' RACE fragments from nested RACE. Red marks the transcription initiation codon and blue marks the stop codon for a 519 amino acid protein. Bold text indicates the position of the RT-PCR fragment isolated in Chapter 3.

```

1      ACGCGGGGATATAATGTGTGTCCCCTCTCACCCCACCCTACTATATAATTTAAACGTTTA
1      T R G Y N V C P L S P H P T I * F K R L

61     TTAATATTTAGTACCAGAGAGGACGAACTGGGTGTTAACGCAAATACGTCAATACACATG
21     L I F S T R E D E L G V N A N T S I H M

121    ACGGAAATGCCCATTTCTGAAAGTATGTGCTCGAGTTCTACACAGCTACCAGCGGAAGTC
41     T E M P I S E S M C S S S T Q L P A E V

181    TTGCAATGTTTCATCCAGTTCCCGTGTTATCCGCATTGGAATAGATGGAGTCCATGGAAC
61     L Q C S S S S P C Y P H W N R W S P W N

241    CAATGTTCAAATTCCTTGTGGTGTGGAGTTTCAATAAAACGAAGAGTGTGTACATTAAGC
81     Q C S N S C G V G V S I K R R V C T L S

301    GGAAGATGCATGGGTGAATCGATCAAGTACAAAACATGCAGTTCGGCTCCTTGTGGAGT
101    G R C M G E S I K Y K T C S S A P C W S

361    GAATGGTCACCGTACAGTCCTTGCTCGACTTCTGTAATAGGGGAGTGAGAACAAGAGAC
121    E W S P Y S P C S T S C N R G V R T R D

421    AGGATTTGTTCTGCTGGTAATTCACATAGCACCTGCAATGGCAGTGCTCTCAAAGTAAC
141    R I C S A G N S H S T C N G S A L Q S N

481    GTCTGCAACACACAAGTTTGTCCGTTGTGGACTACGTGGACGAACTACGGCGAATGTTCA
161    V C N T Q V C P L W T T W T N Y G E C S

541    ACAACTTGTGGTAAAGGTTTTTCGACATCGAAGTAGGTCATGTTTACAAGGGAAGTGTGAT
181    T T C G K G F R H R S R S C L Q G N C D

601    AATAGATTAAGTTTGGAAAGCACATCATGTAACCTAAGGTATTTCTGCCAGCTTGGAGT
201    N R L S L E S T S C N L R Y F C P A W S

661    CCGTGGTCTGTATATTCCTGCTGCAGCGTCAGTTGTGGGATTGGTACACAAACTAGAAAA
221    P W S V Y S C C S V S C G I C T Q T R K

721    AGAACATGTTACCATGGTCAAGAAGGAGAAATAGGTTGCATCGGACCTTGAATGATACA
241    R T C Y H G Q E G E I G C I G P L N D T

781    ACTATTTGCAACATTGACTGCCACAACCAAACGCAACACGCAGTTAGCAGAAATATAGAG
261    T I C N I D C H N Q T Q H A V S R N I E

841    CAGTGCGGATTAAGAGTTGCTGCATCAAATAACAGAAGAAGTTCGATTATCTTAAAAATA
281    Q C G L R V A A S N N R R S S I I L K I

901    TTCGGTGGAAATATATCGCGGAGAAACAGCTGGCCATGGCAAGTGAAGTCTACAAGAATAC
301    F G G N I S R R N S W P W Q V S L Q E Y

961    TTTTATTCTCACCGCTTTAATTATAGCAATTGGATGCACTTTTGTGGTGGAAACAATTGTA
321    F Y S H R F N Y S N W M H F C G G T I V

1021   TCATCTCAATGGGTTATCACTGCCGCTCACTGTTTGCAACAAATAACGGAAAATGAGTAC
341    S S Q W V I T A A H C L Q Q I T E N E Y

1081   AGCATTCAAAGTTTTCGGCAGTATTTGGATTGTTTCGATTGAATTTGCAACACAACACA
361    S I H K F S A V F G L F R L N L Q H N T

```

1141 **CAGAGAATTGGTTTCAAAGAACATTTATTCATTTCGGATTTTCAAAGCGCACATTTAACT**
 381 Q R I G F K R T F I H S D F Q S A H L T

1201 **TTTAGAAACGATGTCGCATTGATACAATTAGATCGAAAAATACAATGGACAAGCAATATT**
 401 F R N D V A L I Q L D R K I Q W T S N I

1261 **CGCCCTGCCTGTTTGCCTGGTGGAGAGGAACCAATTGAAACAGAAAATTGTTACATCACA**
 421 R P A C L P G G E E P I E T E N C Y I T

1321 **GGGTGGGAAGAACAAGAATAAACTCGAGCGAACTCAGTAGTGAACCTTCGAGAATCAATC**
 441 G W G R T R I N S S E L S S E L R E S I

1381 **ATACCAATTCTGTCAAATAAGCAATGTCGACGATTGGGCAGCGGTTACAACACGATCAAT**
 461 I P I L S N K Q C R R L G S G Y N T I N

1441 **ATGACTTTGCACATATGCGCAGGTGACCCAGTGCGGGGGGGACGCGATACATGTCAGGGT**
 481 M T L H I C A G D P V R G G R D T C Q G

1501 **GATTCTGGTGGTCCGATCGTTTGTAAACAGGAGTGGTATCTGGTATATTGCTGGAGTTACT**
 501 D S G G P I V C N R S G I W Y I A G V T

1561 **TCTCATTCTCTTGCTTTTTGTGGTGTCTCGTAACAACGTTGGAATATACACGCGTACCACA**
 521 S H S L A F C G A R N N V G I Y T R T T

1621 **GCGTATGAAACTTGGATACATGATGTTATGACGAGGTACAATCGGCCGGGTTGCTGATGC**
 541 A Y E T W I H D V M T R Y N R P G C * C

1681 **AGTTGCAAAATTAACGACACATCATGTTTTCGATACAGTACTTAAACTAAATGGTTCCCA**
 561 S C K I N D T S C F R Y S T * T K W F P

1741 **CAGCCCACGCTACAACATTTTGTAAATTTCTAACCATGCAGCGAGGTAAATCAAGCTTTG**
 581 Q P T L Q H F V N F * P C S E V N Q A L

1801 **TTACAAACAGTGTTACAACCTTCCCTCCAATCACATTATGGGAGCACTGTATGCTTACAAT**
 601 L Q T V L Q P S S N H I M G A L Y A Y N

1861 **CGTATGTTGTTAATCATAACATCCTAAGACCACTGTGCGTTTTATTTCTCCCTCACATAGT**
 621 R M L L I I T S * D H C A F I S P S H S

1921 **TCATATCGTGTTACATACCTACATAGAGGTTTTTATTACCTATTTGACATCGTCACAACAT**
 641 S Y R V H T Y I E V F I T Y L T S S Q H

1981 **TTTTAATGTATATACTTCTTCTTTACCACCCGATTGCTGGTCACTATACATTCGTAATA**
 661 F * C I Y F F F T T R I A G H Y T F V L

2041 **TATATTGTCTTAAAAAAAAAAAAAAAAAAAAAAAAAAAAA**
 681 Y I V L K K K K K K K K K

Figure 4.3 Complete serine protease 2 cDNA sequence with amino acid translation determined from the 5' and 3' RACE fragments from nested RACE. Red marks the transcription initiation codon and blue marks the stop codon for a 433 amino acid protein. Bold text indicates the position of the RT-PCR fragment isolated in Chapter 3.

```

1      ACGCGGGGAGTTTAGTGAAGAGCAGTTTGGTTGTGTAGTTAAACGTCATGGAAAGCAAAA
1      A G S L V K S S L V V * L N V M E S K K

61     AAAATGTTCTTATTCTGCTTGAGTCATTCTTGCTATTTCTTATATTATCGAGTATGCAAG
21     N V L I L L E S F L L F L I L S S M Q G

121    GTGAAAGTTTAGTTTTTGAAACTGGAAGGTTTGAAGCTAAACGAATTGATCGAATGGAAAG
41     E S L V L K L E G L K L N E L I E W K G

181    GACAAATTTCCCGGAAGAACGCACCAGGTCAACTTCAGATTACAGCGAATACAAACAGAC
61     Q I S R K N A P G Q L Q I T A N T N R P

241    CAGGCACACAGTACACCATGATGCGATTGCAAATACAAGGGGCACGACACAGAATGATGT
81     G T Q Y T M M R L Q I Q G A R H R M M F

301    TTCGATATGGCATCAAGGGAAAGGAAAGGAGTACGGCGCCGAAACTTTGTCCATATGATC
101    R Y G I K G K E R S T A P K L C P Y D L

361    TTATCCAACCGGGTTGGGAGTTTCAGATTATTATTCATGTAACGACAGGATATCTAAACG
121    I Q P G W E F Q I I I H V T T G Y L N V

421    TATATTATGAAGGAAGATTAAAGTTTATTTTACCGATTTCCCTCTTACAAGCATTGAAA
141    Y Y E G R L K F I L P I S P L T S I E N

481    ACGCCGATCAATAAGAATCCATGGTTTCAGTGATTACAAAACGACTGGGATTACTTACCG
161    A V S I R I H G S V I T K R L G L L T G

541    GCGCAGATATTATAAGCGAACTTCAGGCACCAAACCTGCGGTAGGATCATAAGGGGAGGAA
181    A D I I S E L Q A P N C G R I I R G G N

601    ATGTGCCGAATTTTGGCGTGGTGGTGAGCAACAACGCATTGTTGGTGGAACAACCGCGC
201    V P Q F C R G G E Q Q R I V G G T T A R

661    GTCCGGGAAACTTTCTTGGCAAATATCTATTCGCAAGGTTAAAGCTTATTCAAATGGTT
221    P G N F P W Q I S I R K V K A Y S N G S

721    CCCCACGCTGTGTGGTGGGACACTTATAGCAGGACAGTGGGTGATTACTGCTGCTCACT
241    P H V C G G T L I A G Q W V I T A A H C

781    GCTTTACAAGTAGAGTAAAAAGAGAAAGAAACACTTTGTTTCGAGTTGGAGATTATT
261    F T S R V K R E R K K H F V R V G D Y F

841    TCAACCGAGATAACCTTCCTCATAGTCAGGATTCTATGGTTGAAGAGTCACATGATATAG
281    N R D N L P H S Q D S M V E E S H D I A

901    CAATTAGCCAAATTTATATTCATGAGGGTTTTACTCAGTACCCTGCAACAAGAAACGATA
301    I S Q I Y I H E G F T Q Y P A T R N D I

961    TTGCTTTAATTAACTAAGCGAACCGGTGTCGCTAACACGGTTTTGTTCAACCTGCTTGCC
321    A L I K L S E P V S L T R F V Q P A C L

1021   TTCCTACATCACCGGACCAGTTTACAGACGGAAACACGTGTGGCATATCTGGCTGGGGTG
341    P T S P D Q F T D G N T C G I S G W G A

1081   CTACCAATTTTACCAATTACGAGACGAATACCCTTTTTGTTTAAGAGCAGCAACCGTAC
361    T N F T Q L R D E Y P F C L R A A T V H

```

1141 ACACTTGGCCAGATAAAAAATTGTTCTCGATCTTATCCGAGAAGTTTTTCTAATGATAGTA
381 T W P D K N C S R S Y P R S F S N D S M

1201 TGCTATGTGCTGGGGATGAAGGTATTGACACGTGTCAAGGTGATAGTGGTGGGCCACTGA
401 L C A G D E G I D T C Q G D S G G P L T

1261 CGTGTCTCAGTAGGGATGGTAACATTACTCTGTGGGGAATTACAAGCTACGGGAAAGGAT
421 C L S R D G N I T L W G I T S Y G K G C

1321 GTGGGAATAAATCGCAGCCAGGTGTCTAGACAAAAGTGTCTGAGTTCGTTTTTGTGGGTA
441 G N K S Q P G V * T K V S E F V F V G I

1381 TACTTAAAAAATGAAATTACAAGATGCAAACCGGAGTGCCCCGACGCAGAAGAAGGTCGT
461 L K K * N Y K M Q T G V P R R R R R S F

1441 TTATTATACTGCCAACAATAGTTTTGACGGAAGCAACATCATAAATCAACAATATAAACTT
481 I I L P T R V * R K Q H H K S T I * T L

1501 TAATTTTTTGTAAACATTTGACATCTTCGTTCTTGTTTTAAAGTTTCATGGGACATGACAT
501 I F C K H L T S S F L F K V S W D M T *

1561 AATCACAACGTATTTTTAGCATTACATTTGTTGTATTTAAATCATTTTTTAGGCCAATCAT
521 S Q R I F S I T F V V F K S F L G Q S F

1621 TTGTTATAATATGTAGTATAACAACGAGTTCATTTAATTTGGTAAAAAGATAAACATCCA
541 V I I C S I Q R V H F N L V K R * T S S

1681 GTTTTAATGTCTTTATATAAAACCAAAGTCTGTTTTTAAACGATCAAAATGTTTTCAACG
561 F N V F I * T K L L F * N D Q N V F N V

1741 TATTCGGCCCAGCTTTATTATTGCGCTCCTATGTACCCTCTACGAAATAAACCCCCAGTT
581 F G P A L L L R S Y V P S T K * T P S C

1801 GTTGGGCAAAAAAAAAAAAAAAAAAAAAAAAAA
601 W A K K K K K K K K K

Figure 4.4 Complete serine protease 3 cDNA sequence with amino acid translation determined from the 5' and 3' RACE fragments from nested RACE. Red marks the transcription initiation codon and blue marks the stop codon for a 470 amino acid protein. Bold text indicates the position of the RT-PCR fragment isolated in Chapter 3.

```

1      ACGCGGGGATTCTATTTGAATAAGAATATTCTAAATTTAATTTCTTAATCTACTGTTATT
1      R G D S I * I R I F * I * F L N L L L L

61     GCTAACTCTGATCTCAACACAAACAAGACCAAGTCAGGATTTTCAATAAAAATGTCTCAA
21     L T L I S T Q T R P S Q D F Q * K C L N

121    TTTTTGTGCCAGAAGATCCAGTTCTTATAAAATCATAATTTAATCTGCCATATATACTTA
41     F C A R R S S S Y K I I I * S A I Y T Y

181    TATGCACTACATACTCATATAATATTTTAGATTACATTAAAAAATACAAATATTTTA
61     M H Y I H S Y N I L D Y I K K I Q I F Y

241    TTTAGATCCTTGTTTCATTGACAAATGGGGGTGTAACCAACTGTGCAACTGGACTGGTAA
81     L D P C S L T N G G C N Q L C N W T G N

301    TGCGGCAATCTGTGGTTGTCAGTCAGGATACCGACTCCAATCCGATAACAGAACTTGTGA
101    A A I C G C Q S G Y R L Q S D N R T C E

361    AGATATAGACGAATGTACTGAAGGCCCAAACCTTGTATTTTTCGCTTCCCTGCTTTCTG
121    D I D E C T E G P N P C Y F R F P A F C

421    TGTC AACACAATTGGTTCATATTCCTGCCAACCTACCGATGCAACGGCACCAATGAAAT
141    V N T I G S Y S C Q P Y R C N G T N E M

481    GAATTATTACAGGAGTGGTTCATGCTGTAAAGTTAGAAATGGTTCATGTGGTACAACAGC
161    N Y Y R S G S C C K V R N G S C G T T A

541    TAGTATAAGAAGCATGGTGGAGCCAGTTGTCCCAATAGAACTGAAAGGAGGGTTTTTCG
181    S I R S M V E P V V P I E T E R R V F R

601    TGGAATGGCCTCAGTGGTATCTGCTTGGCCCTGGATGGCACAAGTATTATACAGAAGTCA
201    G M A S V V S A W P W M A Q V L Y R S H

661    TCCTCACTGTGGAGCAACTTTAATATCAGATCGATGGTGGTTTCAGCTGCTCATTGTTTT
221    P H C G A T L I S D R W L V S A A H C F

721    CAGATCCGTATCTTACTCCGGTCTCCTTGTTTACCTCGGAACCACCAGGAGCTCTCATCT
241    R S V S Y S G L L V Y L G T T R S S H L

781    TACACATCTTGATACCACTAGGAGGCAACGAAGAGAGGTTGAACAGATCATAGTACACCC
261    T H L D T T R R Q R R E V E Q I I V H P

841    CGGTTTCACCGCTGAGTATTTGAACGACGTTGCATTAATAAAGCTGAGTCGCCCCGTTGT
281    G F T A E Y L N D V A L I K L S R P V V

901    GTTTAATGACATCATCACCCCTATTTGTCTCCCTTGTGGGGAAACACCTAGCCCCGGGGA
301    F N D I I T P I C L P C G E T P S P G D

961    TAAATGTTGGGTGACTGGGTTCCGGACGAACAGAAAACACCGGATACGATTCCTCACAAAC
321    K C W V T G F G R T E N T G Y D S S Q T

1021  CTTACAAGAAGTTGACGTCCCCATAGTCAATACAACCAATGTATGGAAGCTTATAGAGG
341    L Q E V D V P I V N T T Q C M E A Y R G

1081  AGTTCATGTTATTGATGAAAACATGATGATGTGTGCTGGGTATGAAGCTGGGGGGAAGGA
361    V H V I D E N M M M C A G Y E A G G K D

```

1141 TGCCTGTAATGGTGACTCGGGAGGACCGCTGGCATGCCAACGCGCTGACTCATGTGATTG
 381 A C N G D S G G P L A C Q R A D S C D W

1201 GTATTTATCGGGGGTGACATCATTGGTTCGGGGTTGTGGGTTAGCGAGGTACTACGGTGT
 401 Y L S G V T S F G R G C G L A R Y Y G V

1261 CTATGTTAACGTTGTTTCATTATGAGGGATGGATACGAACACAGATGGGCAATGACTCTAC
 421 Y V N V V H Y E G W I R T Q M G N D S T

1321 AGGGTTGTGTCCCCGCCAATACAATCCATGCAAAGGACTTGTAGACGCCACACTGATTG
 441 G L C P R Q Y N P C K G L V D A H T D C

1381 TGCTTCTAAGTTGGATAAATGCAGATCCTTCCCTTCTTACATGGCTACTAACTGTGCAAG
 461 A S K L D K C R S F P S Y M A T N C A R

1441 GTCGTGTTGTCAATTGAATAATGGAGAAATAACAAACTGCCAAGACAGCGCTGACTCAGC
 481 S C C Q L N N G E I T N C Q D S A D S A

1501 AGAAGCTTGTAAACTTTACGTCGGTTATTGTTCAAATCCTGCCATGTCATCATTATGCG
 501 E A C K L Y V G Y C S N P A M S S F M R

1561 GGAAAAATGTCGACGCACATGCGGATTTTGCTGAGTAATGGTTTCAATTATGATGCAGCA
 521 E K C R R T C G F C * V M V S I M M Q Q

1621 ATATTTCTGGACAAATTATGACTTACGTTTTATTTCCCAATTGAACTATTTTTTTTGTTT
 541 Y F W T N Y D L R F I S Q L N Y F F C S

1681 AAATTACTGTAAACCAACAGTAGTTCATTTTTCTGTTTAATTTATTGCGATTTTGGGTA
 561 N Y C * T N S S S F F C L I Y C D F G Y

1741 CTAGTGTAAGTCTTCCCCTGCGCCATAGTAAAATTGAAAGAAGGCAATATATTAAAAG
 581 * C R I F P L R H S K I E R R Q Y I K X

1801 NTAGCTCATAACACAGTGGAGGATTCTTCACTAGCATTCAATCTTGTACAAATTGCCATT
 601 * L I T Q W R I L H * H S I L Y K L P F

1861 TCTTTCTACTTTGCACATTTCTGTTATTTTTGCTTGAATAAAAAGTCA
 621 L S T L H I S V I F A * I K V

Figure 4.5 Complete serine protease 4 cDNA sequence with amino acid translation determined from the 5' and 3' RACE fragments from nested RACE. Red marks the transcription initiation codon and blue marks the stop codon for a 1089 amino acid protein. Bold text indicates the position of the RT-PCR fragment isolated in Chapter 3.

```

1      ACGCGGGGATTCTATTTGAATAAGGTTCAATTTCCACAGGTATATAATCAAAGTTAGTCT
1      T R G F Y L N K V Q F P Q V Y N Q S * S

61     TGTGTGAGGTTTAAATCAAAATTCGGAACAATGATTTTGCACGGAAGTTGGCTGCTGCTG
21     C V R F K S K F G T M I L H G S W L L L

121    ATATTTTGTGCTGCTCTAATACCTAATGTGTCTCAGGGACAGAGTAGCGTTTGTTCGACG
41     I F C A A L I P N V S Q G Q S S V C S T

181    AGCTTGGGTTGCGTAGATTGCTTTTCGTGGTGTCAAGCAAACGCTGCATCTTGACATCG
61     S L G C V D C F S W C Q A N A A S C T S

241    TCACCAGCTCTCATGGGAAGCTACTGCAAGAAAACCTGCAATCTCTGCGCATCAAATAGC
81     S P A L M G S Y C K K T C N L C A S N S

301    GCAGCTTGTATAGCGAAAGGATGCAACCACCGCTGCATTGAAACAACCGGATCGGAACCG
101    A A C I A K G C N H R C I E T T G S E P

361    GTTTGCGCCTGTTTTGAAGGTTTTCGCCTTGAAGCAAATGGTAGAACCTGTGTGCATATT
121    V C A C F E G F R L E A N G R T C V D I

421    GATGAATGTGCTGAAAATAGCACTTTGTGTTCTGATCAAATTTGCCGAATTGCAACAAC
141    D E C A E N S T L C S D P N L P N C N N

481    ACTCTTGGCCATTACGTTTGTACCGCATGTGGATCAACACCGAACAGACACGCTTACTAT
161    T L G H Y V C T A C G S T P N R H A Y Y

541    GAGAGGAATGAATGCTGCAAGATGTCCGGTGGCGCCTGTGGAAAAAGTTCAACCAACGGT
181    E R N E C C K M S G G A C G K S S T N G

601    GGACGAATAGTTGGCGGCAAACGTGGTCGTATTGCAAGGTGGCCTTGGATGGCGTATATT
201    G R I V G G K R G R I A R W P W M A Y I

661    GTAATTGGAAGAAATCTTTGCGGTGGAACCTTTTTATCGTCCGGTTGGGTGTTGACAGCA
221    V I G R N L C G G T L L S S G W V L T A

721    GCTCATTGCTTTGCATCTATAACAAACAACAACCCCAAGCACCATTAACGTCATATTGGGT
241    A H C F A S I T N N N P S T I N V I L G

781    GTTGTTGACACAATTGATTCAGGAAACATACATGAACAATCTTTTTCTGTTACAAGACTT
261    V V D T I D S G N I H E Q S F S V T R L

841    ATAATTCATCCAAACTACAATTTCCCAAACAACGACCTTGCATTGCTACAACCTGGACCAT
281    I I H P N Y N F P N N D L A L L Q L D H

901    GATGCTCTGATTGATGCGGCTTTTGTGAAACCTGTCTGTCTTCCAAATGGAGAGGAGCCA
301    D A L I D A A F V K P V C L P N G E E P

961    CCAGAAGGGGAGAAATGCTGGGCAACTGGATATGGAACGATAGCTTTTGGAGGAGTGGCC
321    P E G E K C W A T G Y G T I A F G G V A

1021   GCTAAATCACTTCAAGAAGTTGATTTGCCAATCGCTGACTTGGCGCACTGTGAGCGAATT
341    A K S L Q E V D L P I A D L A H C E R I

1081   TACGCAAATCTTACAAATCGAGTCAACAGAACAACAATGCTGTGTGCTGGATATATCACT
361    Y A N L T N R V N R T T M L C A G Y I T

```

1141 GGTCAAAGGATACATGTCAAGGAGATTCTGGGGGCCCGCTTGTGTGCCAACGATGCAAA
381 G Q K D T C Q G D S G G P L V C Q R C K
1201 AACTGTGACTGGTACTTGGCTGGTACAACATCTTTCGGTAGAGGATGCGCAAGACCTGGC
401 N C D W Y L A G T T S F G R G C A R P G
1261 TTCTTTGGAGTTTACACAAAAGTTTCCTTCTTTGAGCAGTGGATCTCATCTTACACCAGT
421 F F G V Y T K V S F F E Q W I S S Y T S
1321 ATTGCTATTAATCCAGGGCAGTGTGTAACCATCATGGACTACATGGGGTTCATGGACA
441 I A I N P G Q C V K P S W T T W G S W T
1381 CCGTGCCTCATGCTCAGGATCGTCATCACGGATCAGATTCTGTGCAAATGGTTACCT
461 P C A S C S G S S S R I R F C A N G S P
1441 GGGATCCTGGATGCGATGGGTTGCAGGAAGAATTCGACAATGTTCTACCGTCTGTACA
481 G D P G C D G L Q E E F R Q C S T V C T
1501 CAACCAACTGGGCCGAATACGGAGACTGGGGATCATGCTCAGTGACCTGCGGTGACGGA
501 Q P T W A E Y G D W G S C S V T C G D G
1561 TCAAGGTCCAGAAGTCAATCTGTAGGAACGAAACATTGGTGACCCTGGTTGCTCTACT
521 S R S R S R I C R N G N I G D P G C P T
1621 GGTGGTGAACACTGCAACAGAGGCATGCACAACACTGGAGTTCGTTGTCCAACCTGGTCAGCT
541 G G E T A T E A C T T G V R C P T W S A
1681 TGGTCTGGCTATGGAGTTTGTTCAGTCACATGTGGAGGTGGAACCTCAAGAGTCAACTCGA
561 W S G Y G V C S V T C G G G T Q E S T R
1741 ACTTGTAACAACCACGGACAAGCTGGAGTCACTTGTGATGGTCGAGATACACGATCTCAG
581 T C N N H G Q A G V T C D G R D T R S Q
1801 GCTTGTAATCCTCAGACATGCCCAGCACCAACGTGGGCAGCATAACGGTCTGGTCTGAT
601 A C N P Q T C P A P T W A A Y G A W S D
1861 TGTACACGGCAATGTGGGGGAGGTGAAAGAACACGAGTTCGAACCTGTCTCAACGGAGCA
621 C T R Q C G G G E R T R V R T C L N G A
1921 ATAGGTTCTCTGGCTGCCCGCTGCTGGGGTTTCTCAAACCTGAATCTTGTAACATTCAA
641 I G S S G C P A A G V S Q T E S C N I Q
1981 AGTTGCCAAGCAAATCCCCTTGGTTCAGCGTATGGTTCATGGTCTGGTTGCTCAGTAACT
661 S C Q A N P T W S A Y G S W S G C S V T
2041 TGTGCGTCTGGAACACGAACTCGCTCAAGAAGTTGTGTTGGTGGGAATATAGGAAATGTA
681 C A S G T R T R S R S C V G G N I G N V
2101 GGATGTGAATCTGGTGGTCAAACAGCCAGTGAGGCATGCACAACCTGGAGTCCAGTGCCCA
701 G C E S G G Q T A S E A C T T G V Q C P
2161 ACCTGGTTCAGCTTGGTCTGTCTACGGGGTTTGTTCAGTCACATGTGGAGGTGGAACCTCAA
721 T W S A W S V Y G V C S V T C G G G T Q
2221 GAGTCAACTCGAACTTGTAACAACCACGGACAAGTTGGAGTCACTTGTGATGGGCGAGAT
741 E S T R T C N N H G Q V G V T C D G R D
2281 ACAAGATCTCAGGCTTGTAATCCACAGGCTTGCCCAAGCTGGTCTGGATATGGAAGTTGG
761 T R S Q A C N P Q A C P S W S G Y G S W

2341 TCTGGGTGTAGTGAACTTGTGGTGATGGAACCAAACTAGAACCAAGGACTTGTAAATAAT
 781 S G C S E T C G D G T K T R T R T C N N

 2401 GGGCAAATTGGTGATAATGGTTGCAGTCCTGCTGCTGCTGCAACAGATTCAATGGCTTGT
 801 G Q I G D N G C S P A A A A T D S M A C

 2461 AGTGTGAGGAACTGCCACAATGGTCGAGCTGGGGTTCATGGGGTCAATGCTCTCTAACA
 821 S V R N C P Q W S S W G S W G Q C S L T

 2521 TGTGGCAGCGGAACGAGAAGTGCAGTGAGGCAATGCAACACATTTGGTGCCACCGGTGCA
 841 C G S G T R T A V R Q C N T F G A T G A

 2581 TCATGTGGCGCTGGTGCAACAAGTAAAAGTGAACCGTGCAACTTGGGTGCCTGTCCAGTG
 861 S C G A G A T S K S E P C N L G A C P V

 2641 TTTAGCGCCTGGAGTGGATGGAGCACTTGTAGTGCAGGTTGCGGTGGCGGCCAACAAACA
 881 F S A W S G W S T C S A G C G G G Q Q T

 2701 CGTACTCGCACATGCTCTAGTCCAGGCAACTGTGATCCTGATGCTTTTGGAACAGCTTTG
 901 R T R T C S S P G N C D P D A F G T A L

 2761 TCTGGTTCACAAGCTTGCAACACTGATGCCTGTATAGGGGAGTGGGGTGTGTGGGTGAAC
 921 S G S Q A C N T D A C I G E W G V W V N

 2821 AGTGAACATGTTCTGCTGCCTGTGGTCCCGGACAATTCAACAAACAAGAGAATGTATT
 941 S G T C S A A C G P G T I Q Q T R E C I

 2881 GGAGGAACAGCTGGGCAACCAACTGTGTTGGTTCTACACAGCAGACTGCTGCTTGTAAAT
 961 G G T A G Q P N C V G S T Q Q T A A C N

 2941 GTGGCTGCTTGTACATGGGGGAATGGGTTGCTTGGACAGCTTGTACAGTAACCTGTGGA
 981 V A A C T W G E W V A W T A C T V T C G

 3001 GCCGGAACCAAACCAGAAGCAGAACGTGCAGTGGTGAAGCAGGACGGTGCCCTGGTGGC
 1001 A G T Q T R S R T C S G E A G R C P G G

 3061 CAAAGTGCTGCCACTGAATCCAAGCGTGTGCAGCGAGTACATGTGCAAGTACTGTAAAC
 1021 Q S A A T E S Q A C A A S T C A S T V N

 3121 GATTGTTCAAATGACATAGACCTGGCACCAGCCATAACTTGCCGAGAATATGCAGTAGCT
 1041 D C S N D I D L A P A I T C R E Y A V A

 3181 GGATATTGTGAACAGTATAAAGACTACATGGATATAAACTGCATTAGGAGTTGCTGTGTG
 1061 G Y C E Q Y K D Y M D I N C I R S C C V

 3241 TTCGGAAGAAATCCATGTTCCATGTACAGAGATTCCTTATTTCAATGTCCTTCGTACCGT
 1081 F G R N P C S M Y R D S L F Q C P S Y R

 3301 CACTTGTGTACAAATACATTAGTAGAACCGCTTTGCAAGTACACATGCAACTGTGTATGA
 1101 H L C T N T L V E P L C K Y T C N C V *

 3361 GAATTAGAATGTGATCTTTTTGTAAGATAAAAAATCTCCAGAATTACCCATTTGTCGGGA
 1121 E L E C D L F V R * K F S R I T H L S G

 3421 ATTTTTTTCATTGTATCTAATTCATTTTAAAGTTTGGCTTCATTGTGATGTAAACGTGTA
 1141 I F F I V S N S F * S F A S L * C K R V

 3481 AATCTTATATTGCAATAAACTTAGCCCCGAAAAAAAAAAAAAAAAAAAAAAAAAAAA
 1161 N L I L Q * T * P E K K K K K K K K

Figure 4.6 Complete serine protease 5 cDNA sequence with amino acid translation determined from the 5' and 3' RACE fragments from nested RACE. Red marks the transcription initiation codon and blue marks the stop codon for a 225 amino acid protein. Bold text indicates the position of the RT-PCR fragment isolated in Chapter 3.

```

1      ACGCGGGGATTGGCCGTAGCGTGTTTTTAAGACTGTCGTTTTGTTGCTAATACACTTCTC
1      R G D W P * R V F K T V V L L L I H F S

61     TTCGTTGGTTTTAATTCGTTTTCTCTTAGTTATCGTTGTCGAAGATTGGCGCCGTGACACA
21     S L V * F V F S * L S L S K I G A V T Q

121    GTGGTTAATGCGGGCCCTCTATAACCCAGAGGTTACGGCTGCAACACTCGACGCTATTAT
41     W L M A G L Y N P E V T A A T L D A I I

181    CACTGTGGGCGGCGTAAGCGGGAACGAAATGGTTGATAAAACCACTCTCGGCTTTGGGAC
61     T V G G V S G N E M V D K T T L G F G T

241    TTCCCAGCTTTGGCGGCTGACCGCCTCTCGCACCTGCTGCGTGATAGTGACGTCACCGT
81     S Q L W R L T R L S H L L R D S D V T V

301    GACGTCAATGGATGGCGTCACTGGTGCGAGCGTGATCCGTTTAAAGTCGCATCTACAGCCA
101    T S M D G V T G A S V I R L S R I Y S H

361    CCCTACTTACGGCGAAAACCTGGATAGTGATATCGTATTGATCAAGGTTGCCGAGCCGAT
121    P T Y G E N L D S D I V L I K V A E P I

421    CACGTGGTCTCGGGTATTCCCCGTATGTCTGCCCTCGCCTGAATCCCTTCTTGATCA
141    T W S S R V F P V C L P S P E S L L D H

481    CGTGGGCCACGGTCGAGTGCACATACCAAAGCAATATTGTAAACTAGCTGGATGGGGAAG
161    V G H G R V H I P K Q Y C K L A G W G S

541    CTCTTCAGAATTGGGCGATTACGCGAGTGATTTGGCCAGTATAGAAATCCAGTGATGAG
181    S S E L G D Y A S D L A S I E I P V M S

601    CGACAGACATTGCGAGAGATCTTCAGCCAGTTTGTGTTGGTCGTAGAGTCAACATTCGGGC
201    D R H C E R S S A S L F G R R V N I R A

661    AACACTGTGCGCTGGTCATTTTCGACGGCACGAGACAAAGCCCTTGTAAGGCGACGATGG
221    T L C A G H F D G T R Q S P C K G D D G

721    GGGTGGTCTTACATGTTCTTGAACGGCAACCATTATTTAGTTGGTGTGGCTGGCGAGCA
241    G G L T C S W N G N H Y L V G V A G E Q

781    GTTCGGTGAATGCTTCGTTGCTTAACGTACCACGCTACTTTACACGCGTTTCCACTTTCG
261    F G E C F V A * R T T L L Y T R F H F R

841    TTAAGTGGATTGAGGATACGATTCGAAAATCAAATTCGGTTACGATTCGATTTTGGCAA
281    * L D * G Y D S K I K I R L R F V F C K

901    AGGGGGCCTAGGGAGTGGTGAATAATGTTTTAAATTGAAATAAATTTTAAGCCAGTTTGG
301    G G L G S G E N V F K L K * I L S Q F G

961    AGTAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
321    V K K K K K K K K K K K K K K K K K K K K K K K K K K K K K K K K K

1021   AAAAAAAAAAAAAAAAAA
341    K K K K

```

Figure 4.7 Complete serine protease 6 cDNA sequence with amino acid translation determined from the 5' and 3' RACE fragments from nested RACE. Bold text indicates the position of the RT-PCR fragment isolated in Chapter 3 although the 3' end does not extend beyond this and only part of the fragment is present.

```

1      ACGCGGGGGCGAACGCAAATACCAATGTTCCACGGCGGCCACAACTATTCTCAAAAATG
1      A G A N A N T N V P R R P Q T I L K N

61     AAATTGTTTACAGAACCGACAAAAGGATTGTCATACATCCAGAATTCGTTTTCCCGCATT
20     E I V Y R T D K R I V I H P E F V F P H

121    ATGACGTGCGGTTAATCGAAGTGGATCGCGCTTTTGACGTCACTGGCGTTTTTGTCAGGC
40     Y D V A L I E V D R A F D V T G V F V R

181    CGGTGTGTCTACCTAACGGTGAATACCCGGAAGCAGGAAAGCGTTGTTACACCACAGGCT
60     P V C L P N G E Y P E A G K R C Y T T G

241    TCGGAACATTGGAATATAAAGGAGATGTGTCGCCGTCCTCTACAACAAGTTGATCTCCCCA
80     F G T L E Y K G D V S P S L Q Q V D L P

301    TCATATCTCACAGCACTTGTTCAGTTGTATCGTAAAGTTGGTTGGAACCTTATAAATT
100    I I S H S T C S Q L Y R K V G W N L I N

361    ATCAGTTATGTGCAGGCAACCTCACTCATGGTGGTGTAGACTCTTGCCAGGGTGATAGTG
120    Y Q L C A G N L T H G G V D S C Q G D S

421    GTGGTCCACTGGTTTTGCCAACGTTGTTCAAACCTGCAACTGGTATCTAGCCGGTGTGACTT
140    G G P L V C Q R C S N C N W Y L A G V T

481    CGTTTGGACGTGGTTGTGCTCTTCCCGAATTCCTGGTGTTTACATGAGTGTAAAACACA
160    S F G R G C A L P E F P G V Y M S V K H

541    TTGAAAGATGGATTGAAACTATTACACAAATGTATGCCAGCAGCAACAAGACATGTCAGC
180    I E R W I E T I T Q M Y A S S N K T C Q

601    CAATTTTATAGAGTGGAAAGTGGTAGACACGCTACTTATAGTGGCAGCGTTGTTTACGAGGAT
200    P I L E W K W * T R Y L * W Q R C L R G

661    CTAAGTTGTTTTTACAAGGCTTACAGTGTGTGGGATGCTAATGATCCAACGATATGTGAA
220    S K L F L Q G L Q C V G C * * S N D M *

721    GTCTATTTCCATATTTCAAACCCAAGGCGTTGGGGTTCGATCGCACCAACGTCAATTCGAA
240    S L F P Y F K P K A L G F D R T N V I R

781    ACACCAGTATGGAGCGCCAGTACAGATAACTTAAATAACGCTGCAAGCATCGGCGGAAGC
260    N T S M E R Q Y R * L K * R C K H R R K

841    CCCGAGTCCGTGTTATTTGGTTACACATTGTTATCAAGAGATGACGATTTCCAGTCACTA
280    P R V R V I W L H I V I K R * R F P V T

901    ACAGTCACGTGAGGGCAGCATTGACACGAATTATGCAAGAAATCAATTCATTCACTCAGA
300    N S H V R A A L T R I M Q E I N S F T Q

961    TTGAAAACCTTCAACTGCAGCCGACTCAACGAGAGTTTGGTGGTTCCCGTCCATTTTGACG
320    I E N F N C S R L N E S L V V P V H F D

1021   TAGGATCCATTGAGGCTATGGCGTCACACATCCACCAGCGGTCAGGGAACGAGCCCCTC
340   V G S I E A M A S H I P P A V R E R A P

1081   CAGGTTTTATTCAATTTCCGAGCCTTCGACCCGAAAGGTGTTCCAAACGCACTGTGTCCTG

```

360 P G F I H F R A F D P K G V P N A L C P
1141 GTGTGAGATCTGATCCATGTCGACCTTCCTCCATCTGTGTTGGTGGAGTCAACACCATAC
380 G V R S D P C R P S S I C V G G V N T I
1201 CATCAAACACCAGGGCTTGTGGAGATTTCGCTGGTGGGATGGGTTACCGGTCGACCAAC
400 P S N T R A C G D F A G W D G L P V D Q
1261 CCACCGACACCGAACCTGTGGGTCATGCCAAGTCGAAGAATGACGTCGCTTCGTCTCTCT
420 P T D T E P V G H A K S K N D V A S S L
1321 TGTTATTTACTCGAAGCAGAGTTTCTTAACATCGTGTGTGATTGTTTGGTAATCGTGACG
440 L L F T R S R V S * H R V * L F G N R D
1381 TTCAATGTGTGGATTAATCAGTAATTGAATTCATTAATAACAATTCAATAACGAAAAAA
460 V Q C V D * S V I E F I K I Q F N N E K
1441 AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
480 K K K K K K K K K K K K

Figure 4.8 Complete serine protease 7 cDNA sequence with amino acid translation determined from the 5' and 3' RACE fragments from nested RACE. Red marks the transcription initiation codon and blue marks the stop codon for a 1235 amino acid protein. Bold text indicates the position of the RT-PCR fragment isolated in Chapter 3.

```

1      ACGCGGGGACAGTCCTCTAGTATTAAGGTATTCTAGAAATATTTCTTCATTAGCATGAAG
1      T R G Q S S S I K V F * K Y F F I S M K

61     TATATTTTCGTAGCATTCCCTAAGTATTCTGTGCTGTGCAAGTTCTTTTGATTACAAATGT
21     Y I F V A F L S I L C C A S S F D Y K C

121    TCACGTGCTGGATCAAGATGTCATTTTCCATTTTCTTCCAAGCACCAACCAACATAC
41     S R A G S R C H F P F F L P S T N Q T Y

181    CATGAATGTCCACCTTACCGACAAAGTGCATTATGGTGTGTTGTTAACAGGGATGGCCGT
61     H E C P P Y R Q S A L W C V V N R D G R

241    CTTGTACCAACAATTTGCGTCCCATGTTTAGGTGATGGAGAGTGTATGTAAAAGCAAAT
81     L V P T I C V P C L G D G E C Y V K A N

301    GATTTTCCATCTCAATTTACTTTGGAATGTCCCCGATTGTGTGCGTTTGTCCCGCTAAT
101    D F P S Q F T L E C P R L C A F V T A N

361    CTGTGGGGTACTAATGTTTATTCAAATAAATTCATTTGTCTGTTCTTCTGCCATACATGCT
121    L W G T N V Y S N N S F V C S S A I H A

421    GGTATCTATCCAGCAACTGTGGGTGGAACAATCAAAGAATAGATCGTCCAGCAAGCTAT
141    G I Y P A T V G G T I K R I D R P A S Y

481    ACTGGGTCGCCAAGAAATGCTCTTCGATCTAAAACCATTATTTCTCACATACTGCATTT
161    T G S P R N A L R S K T I I S S H T A F

541    CGCCCAACAAGAATCCCAACTCCTTCTTTTCTGGTTTAATTTACACTGTTGAAATAAA
181    R P T R I P T P S F P G L I Y T V G N K

601    ATTGAAGTTATTTCTTCAACAAGGCGACATGATAAGTTAACATTGGTATCAGAACCAAAT
201    I E V I S S T R R H D K L T L V S E P N

661    CGAATTATTTTCAGTTGATTTGGATACCAGGAGAAATTTTGTGTTTTGGATCATTCCAAAT
221    R I I S V D L D T R R N F V F W I I P N

721    ACAAGGCAGATTATGAAAGCAACTTTCAGTGATGATTATACATCAGTTACAGATACTTCT
241    T R Q I M K A T F S D D Y T S V T D T S

781    GTTCTACAAGGACCAACTTCGGTAAACAACCAATCCAGCTTTCATACGATTGGGTACAT
261    V L Q G P T S V N K P I Q L S Y D W V H

841    GAGGTAATATACTGGACAGACGCCACAGTGTCCGAGTTGCTATGACAACTAGTAATCAC
281    E V I Y W T D A H S V R V A M T T S N H

901    ATAACATTCCTAATTAACCGTGGCTCGCAATATCAACCGGATGCGATTCAAGTTGATCCA
301    I T F L I N R G S Q Y Q P D A I Q V D P

961    GAATCTGGGTATGTGTACATCAGCGATACTGGAAGTTCTCCCAAAATGAAAAATGTTTCG
321    E S G Y V Y I S D T G S S P K I E K C S

1021   ATGGGAAATCCAGATTCCCGTACATTGGTTGCGAGTGAAAATGTTCAACAACCAACAGCA
341   M G N P D S R T L V A S E N V Q Q P T A

1081   TTAACAATTGAATCATCTACAAGCAAAGTTTACTGGTTTGACAGTTCAACTAAAACCTTTA
361   L T I E S S T S K V Y W F D S S T K T L

```

1141 AATATGTGCCACAGTAGTGGTACAGATTGTACAGTCATTTTAAAGTTCAAATAAAATCATC
381 N M C H S S G T D C T V I L S S N K I I

1201 AATTTTCCGGTTGGAATGTTTTTGAACGATAATAAAGTTTATTGGATTGATGCTGGCGAC
401 N F P V G M F L N D N K V Y W I D A G D

1261 TTAACAATAAAATCTGTTAACCAACGTACTGGGGAAAGATTGCATCTATCAGCAGCAGGT
421 L T I K S V N Q R T G E R L H L S A A G

1321 TTGCATCGACCAAGTTCAATTA AAAAGTCTTGATCAACTTAATCAACCAATGGTCAGAAAG
441 L H R P S S I K S L D Q L N Q P M V R K

1381 CGTTGCCAGCATT CAGACTGTCCACACTTTTGTCTCCCTGCTGGTCGTGCATACAGATGT
461 R C Q H S D C P H F C L P A G R A Y R C

1441 GTGTGTCCTTACAACGTACCTTCGTGCAATCATACTTTCCAGCAATCCAATGTACAGATC
481 V C P Y N V P S C N H T F Q Q S N V Q I

1501 TTTATAGCTGATGATGATATTATAAGACGTTTAAATGTCAATATGTTAACAGGAAGCATT
501 F I A D D D I I R R L N V N M L T G S I

1561 ACAGGAGACGTTATAAAAACGTGGCTTAATAAATGCAGCTGATGTTGCATACAGTTCAATT
521 T G D V I K R G L I N A A D V A Y S S I

1621 ACAAGTAAATTACTGGTCCAATGAAACAAGTTCACAGTCTCGTATAACCAAGAAGGAA
541 T S K L Y W S N E T S S Q S R I T K K E

1681 GGAGTTGCAGTTGATTGGATTACCATAACTTATATTGGACAGATGCTACACATAACAAA
561 G V A V D W I H H N L Y W T D A T H N K

1741 GTCATGCTTGCTTTTGGTGATGAGGGAAACTTGGAAAATATTTCAACTTTGATTGAGAGA
581 V M L A F G D E G N L E N I S T L I E R

1801 AATTCTGTCTACAGGCCAAGGGCAATAGCATTAGATCCTCTAAAAGGCTACATGTATATA
601 N S V Y R P R A I A L D P L K G Y M Y I

1861 AGTGATATTGGAAGTAATCCTAAGATAGAAAAATGCTGGATGGATGGTGAACATTGTATG
621 S D I G S N P K I E K C W M D G E H C M

1921 ATTATTGTTGATGAAAATATTCAGCTTCCCAATGGTATAGCATTAGACTTTACAACACAA
641 I I V D E N I Q L P N G I A L D F T T Q

1981 AAGATGTTTTGGACTGACGGACGATTGAAAACCTTTGTCAATTTTCAAACCTTTGATGGTTCT
661 K M F W T D G R L K T L S F S N F D G S

2041 AATAGAACCATATTACTTGATGATTCTACTCTGATTGGTCAAGCTTATGGAATTGGAGTT
681 N R T I L L D D S T L I G Q A Y G I G V

2101 TTCTACAACAGAGTGTCTGGACAGACCTAACATCTAATGCTTTGTTTACAATCTCAAAG
701 F Y N R V F W T D L T S N A L F T I S K

2161 ACACCTCCTGTTAGAAGGCAAGCTATTATGACTGGCTTGGTTGAAGGAAAAGGAATTA
721 T P P V R R Q A I M T G L V E G K G I K

2221 CTTATTGCTCAATATAATCAACCACAAGGAGATAATGTTTGTGCCGAGAGTTCTGATTGT
741 L I A Q Y N Q P Q G D N V C A E S S D C

2281 TCAATATGTGTTCTGTGCCACATACCACTAATACAACCTAGATCATCATGTGTATGTCCT
761 S I C V P V P H T T N T T R S S C V C P

2341 GATCATTTAAGGGTTGCTCAATCCAATCGTCTGAATGCAGAAAACACACATTGACTTGT
 781 D H L R V A Q S N R P E C R K H T L T C

 2401 AGACGAGGCCTTCAACCAGATGCAAAC TATACAGGATGTGTGGATATTGATGAATGTGTA
 801 R R G L Q P D A N Y T G C V D I D E C V

 2461 ACTAATACACATCTATGTGAACAGATATGCTTTAATTACATGGGAAGTTACTTATGTATC
 821 T N T H L C E Q I C F N Y M G S Y L C I

 2521 TGTCGTGATGACTTTACACTAAACCTGGACGGTCGCTCTTGTTATGATGCGGGTTGTTCT
 841 C R D D F T L N L D G R S C Y D A G C S

 2581 AGCAGTCCGTGTATGAATGGTGGTTTATGTTT CAGATGTTGCCAATGGTTCATACTCCTAT
 861 S S P C M N G G L C S D V A N G S Y S Y

 2641 ACTTGTCAATGCTTGCAGGGATT CAGGGGGCGTTTATGTAATGAAATATACAGCAGTATG
 881 T C Q C L Q G F R G R L C N E I Y S S M

 2701 AATGTAGTATTAAGTGAAATGAGGTTTCATTTGAATGCATTGTTCCAAGAGTTGTTACA
 901 N V V L S G N E V S F E C I V P R V V T

 2761 TCAGTTATAAAGTGGTATCGGAATCGTGAGCGAATAACAAC TGAATGAGAAGCACATTT
 921 S V I K W Y R N R E R I T T A M R S T F

 2821 TTAACCTTTAATGGTCGTTTGTGAGAATTTTCTTCGTAAC T TATCTTGAAGCTGGGAAT
 941 L T F N G R L L R I F F V T Y L E A G N

 2881 TACAAATGTCAATTTGAGTACGGGGTTTGGAAATATGCTTCAACGC ACTTTCTTGAAGTA
 961 Y K C Q F E Y G G L E Y A S T H F L E V

 2941 CCAGTACAAATTT CAGCTGTTTGTGGTCAAGCACCTAGCATACTGATCGCATAGGTGGA
 981 P V Q I S A V C G Q A P S I P D R I G G

 3001 AGAATTACATCCGGTGTACCAACTGCAC TTTTGTGATGGTCCATTTATTGCTATGCTGGTA
 1001 R I T S G V P T A P F D G P F I A M L V

 3061 GAGGAAACTAATGAGGGAAGTGAACATTTTGTGGAGGTTCTATTGCTACAAGGAATAAAA
 1021 E E T N E G S E T F C G G S I A T R N K

 3121 ATAATCACAGCAGCACATTTGTTGCAAAA T GATGAAATAAATATAACATCAGTCCATGTG
 1041 I I T A A H C L Q N D E I N I T S V H V

 3181 TTTGTTGGGAAAGTCTTAACTGATGTTACAT T GATTGAACCATACCAACAACATTCTCTC
 1061 F V G K V L T D V T L I E P Y Q Q H S L

 3241 GTCTCCCATGTTGTATTT CATGAGAATTACGATCCCGATAATTTAAATTCAGATATCGCC
 1081 V S H V V F H E N Y D P D N L N S D I A

 3301 ATTCTTACGTTATCAACGCAAATAGTATTCACCAAAGCAGTAAAGCCGCTGTGTATCCCG
 1101 I L T L S T Q I V F T K A V K P L C I P

 3361 CTTCATACTGACACAAACCAAGATATCAAACCAAGGCCATACAGAGGCACATCTAAGATG
 1121 L H T D T N Q D I K P R P Y R G T S K M

 3421 GGGTTGGTGTAGGGTATGGAAGGACAAGTCATCGTGGTCCAGTTTCAACTCAATTACGT
 1141 G L V L G Y G R T S H R G P V S T Q L R

 3481 GAAGTTCTGGTTGAAATTCGCACACAACAATTTTGTACCCAAAGATACCGTACTGTAGAC
 1161 E V L V E I R T Q Q F C T Q R Y R T V D

 3541 AAAGAGGTGACTTCTGTCATGTTTGTG CAGGTGGTGGCGACAAGATGCTTGTAGTGG A

1181 K E V T S V M F C A G G G A Q D A C S G
3601 GATTCTGGCGGACCATTTGCCTTGTGGAGCAACAGAACACAGTCGTGGTGGTTGGCTGGT
1201 D S G G P F A L W S N R T Q S W W L A G
3661 ATTGTATCTTGGGGACCAAGAGGTTGCGGTGTGTCAAATTTACCTGGCGTTTACACAAGA
1221 I V S W G P R G C G V S N L P G V Y T R
3721 ATTGGCACTAGCATGCGACAGTGGATAACATAATCATATATAACTAGACAGCAGAAACACC
1241 I G T S M R Q W I H N H I * L D S R N T
3781 TAAAAAGCAATCGGAATGTTTAAGTGAGAATGTTATAACGATCGTTGTTTTACTCTTAAT
1261 * K A I G M F K * E C Y N D R C F T L N
3841 TCCGTTCTCCCCGTTTTATTTATGTTACTATGTTTTTCGTTATGGCATTAAAGAGATCA
1281 S V L P V L F M L L C F S L W H L K R S
3901 ATAAAGAGCCATATGATGTTGCTGAAAAAAAAAAAAAAAAAAAAA
1301 I K S H M M L L K K K K K K

Figure 4.9 Thiolester 1 cDNA sequence with amino acid translation determined from the 5' and 3' RACE fragments from nested RACE. Red marks the transcription initiation codon and blue marks the stop codon for a 1809 amino acid protein. Bold text indicates the position of the RT-PCR fragment isolated in Chapter 3.

```

1      ACGCGGGGACTAATAAAATATATAACAAGTATGAACCTACGTTGGAGGCCAGCGTGCTCG
1      T R G L I K Y I T S M N L R W R P A C S

61     CTGGGAACAATTTACCTTCTCGCCACGTTGTCGTCTCTTGCAACAGCCAGCAATGTCTAC
21     L G T I Y L L A T L S S L A T A S N V Y

121    AACATATACTTTCCCAAGCACATCAGACCTGGATTCAATATCTCATTACGGCTGCAATC
41     N I Y F P K H I R P G F N I S F T A A I

181    ATTGACAATCCAAATACCGTCCAAATCCACACTGCCTTTAGATCTATGGACAATTCTTTC
61     I D N P N T V Q I H T A F R S M D N S F

241    CATGTTGATTCCACTGATTCTGTCAACAGTGGTCTAGCTCAAGAATTTCAATGAATGGG
81     H V D S T D S V N S G S S S R I S M N G

301    TTGCCAATACACTACAGCGGAAGTCACGGCTTTGAGTTGAACATAACTGGCACAGACCTG
101    L P I H Y S G S H G F E L N I T G T D L

361    GTTACAGGTGCTCAGTTGTTTTTCAATTCATCAACAGACTTCCAGTTTCAAGCTAAATCC
121    V T G A Q L F F N S S T D F Q F Q A K S

421    ATCTCAATTCTAATTCAAACTGATAAAGCCATATACCAACCAGGACACACAGTCAAATTC
141    I S I L I Q T D K A I Y Q P G H T V K F

481    CGTGCCATTGCATTGAAACCTGACCTCAAGCCCCTCCAGGGAAATATCTCATATACATTC
161    R A I A L K P D L K P L Q G N I S Y T F

541    AAAGATCCAAGAGGTAATGTGGTGATGCTTGAACCAGAAGTACCACTTAACCATGGTGTG
181    K D P R G N V V M L E P E V P L N H G V

601    GCTGGTGGGCAGTTCTCACTTACTAAGGACGCAGTTGCTGGGATGTGGAAAGTGGAAATTC
201    A G G Q F S L T K D A V A G M W K V E F

661    ATGGCAGAGGGTTTCAAAGAAAGTTTATCAGTTGAAGTGAACGTTACAAGTTACCCAAG
221    M A E G F K E S L S V E V K R Y K L P K

721    TTTAAAGTTGAAGTCAAAGCACCTTCATACATCCACCCACAGTCTACAGGTCTCACCATC
241    F K V E V K A P S Y I H P Q S T G L T I

781    AAACTTGATGCAAAATATACATTCGGCAAAGGAGTTCAAGGCACGGGGCTTCTTGAAGTA
261    K L D A K Y T F G K G V Q G T G L L E V

841    GTTGGTGGATACCAATACCCTGTGTATCATGGATTTGGTGGTAGATTTGCTCCACGACCA
281    V G G Y Q Y P V Y H G F G G R F A P R P

901    CCAACACAAAATAAAATAACGCGGCGTTACCCGAATTTTGATGGAAGTGTGAATTGCTC
301    P T Q N K I T R R Y P N F D G T V E L L

961    ATCACTAATGATGAGATAAGAGAAGAAGTGGGTGGAATGGCGCAAGTGAATCTATTATC
321    I T N D E I R E E L G W N G A S E S I I

1021   ACAGTAACTGGGTCTGTTACTGAGGCCCTAACTCGAGAAGCATTCAACGACACACAGAGA
341    T V T G S V T E A L T R E A F N D T Q R

1081   ATTGATGCAAAAACAACGAACGTTAAAGTTGAAACTCTCGTCAAGCCATTAACCATCAAA
361    I D A K T T N V K V E T L V K P L T I K

```

1141 CCTGGACTTAAATACTCTGCTTATATCCAAATAACAGAAGTGGATGGGAAACCATTGCCA
381 P G L K Y S A Y I Q I T E V D G K P L P

1201 GAAGATGATCGTTTGGCAAATAATCTGCTACTTAATATAGAATACAGATACCCACGTGGA
401 E D D R L A N N L L L N I E Y R Y P R G

1261 GAGCCAGAGCCAGGCACCAACACAACAGTATCTACATGGTATGCATACAGATGGGAAGAA
421 E P E P G T N T T V S T W Y A Y R W E E

1321 ACACGGGTATTTGTCATCCCACCTTCTGGGATTGTTAAAGTCACGATTGATGCTCCATCT
441 T R V F V I P P S G I V K V T I D A P S

1381 GATACGTTTACTTCAATCAATTTTAGACCGTACACAAATGCAACAATGTCACAGCGTTGG
461 D T F T S I N F R P Y T N A T M S Q R W

1441 GCACTACAGTGGACGGCAGAGAGAGCAGATTACCTTCAAACCTCATATCTACAAATCACC
481 A L Q W T A E R A D S P S N S Y L Q I T

1501 ACTGAAGAGAACAGTGTGTGCGCCAGGCAATATGGCCACTGTGACCATTAGAACAACCTGAA
501 T E E N S V V P G N M A T V T I R T T E

1561 GCTGTTTCTGAATTCACTATATTGATTATATCTCGAGGAGAAATCTTTCCGAGCGAAAA
521 A V S E F T I L I I S R G E I L S E R K

1621 TTCCAAACTCTGTCCGGTGTACCAGAGAATTCATTTGTTTGAGTTCAGTGTCTGAATAT
541 F Q T L S G V P E N S H L F E F S V E Y

1681 GATATGATTCTCTGGGGTGCAGGTGCTTGCTTCTTACGTAAGGGATGATGGGGAGATAGTG
561 D M I P G V Q V L A S Y V R D D G E I V

1741 GCTGATTATATAAAGTTGACGGTCACTGCTGAACTGGAAAATCAGGTCTCCATCAGAGT
581 A D Y I K L T V T A E L E N Q V S I T S

1801 TCCAGCACCAATATTGACGCAGGAGAAGACGTTAGCATCCGTGTACAAACCTCATCATCT
601 S S T N I D A G E D V S I R V Q T S S S

1861 GGTGCTTATGTGGGAGCACGTGCCATTGATCAGAGCGTGTGTGCTACTTAAATCTGGCAAT
621 G A Y V G A R A I D Q S V L L L K S G N

1921 GATGTTTCCCAAGAAAGGATTGTCACGGACTTGAACAAATACAGTGTACCCAAGAATTG
641 D V S Q E R I V T D L N K Y S V T Q E L

1981 AACCACATGTGGAGGTGGTGGTGGTACCCTACCCCATCTGGTGCCAGTGATGCCAGT
661 N H M W R W W W W Y P T P S G A S D A S

2041 GATGTTTTTCAGGAAAGCTGGTATTCTAGTGTTCACTGATGCTCTTGTGTATCAAAGCCA
681 D V F R K A G I L V F T D A L V Y Q K P

2101 GAGGCTAGTATTTACCCTTTTCGGCCTATTGCGTTTTCCCTGAATGGGGGGTTTTGCTGAA
701 E A S I Y P F R P I A F S L N G G F A E

2161 CGCAATATAATAGCAACTGCCGAGTGGATACCTCAACCCCTGCCACCCCTACACGCACA
721 R N I I A T A A V D T S T P A T P T R T

2221 AGGACATTATTTCTGAAACTTGGTTATGGGATGAACAGATTTCTGGTGTGATGGATCA
741 R T L F P E T W L W D E Q I S G A D G S

2281 GCCACGTTCAACACAACAGCACCAGACACAATTACTTCTGGATCTTTAGTGCTTTCTCT
761 A T F N T T A P D T I T S W I F S A F S

2341 GTATCTGACCAACATGGTCTTGGTGTCTAGTGAGCAGCACAAGGTCACAGTATTCCGGAAC
 781 V S D Q H G L G V S E Q H K V T V F R N

 2401 TTCTTCATCACATTGAACCTCCCAGTTAGAGTAATTCGAGGTGAAGTACTGATCATTGTACAA
 801 F F I T L N L P V R V I R G E L I I V Q

 2461 GCAATCGTGTTCAACTATCTTAGTACTGAAGTTGACGCTGTGCTCACTTTGACCGAATCA
 821 A I V F N Y L S T E V D A V L T L T E S

 2521 AACAAATTCGTCTTCTCCGTCTGGCAACAACAGTGTGCCGTTGGTTTTTACGTGCGC
 841 N K F V L L R P G N N S A A V G F S R R

 2581 ATCACCATCCCTGCATCTGGATCAGTATCTGTTAAATTTCCCATCCGAATGGGAACACTG
 861 I T I P A S G S V S V K F P I R M G T L

 2641 GGTGAAATCCCAATCACCATGACTGCGATATCGGAAATTCATCTGACGCTCTGACAAGA
 881 G E I P I T M T A I S E I A S D A L T R

 2701 AAAGTTTTTGTCCAGCCTGAAGGTATTACCCAATGCACATCCGGGTGCGTTTTATTCCAA
 901 K V F V Q P E G I T Q C T S G S V L F Q

 2761 CGCATGGACGCTTCTGCCCTCCTGATGTTGAAAGTTTAAACATTCAAATTCAGCTGGA
 921 R M D A S A P P D V E S L N I Q I P A G

 2821 ATTGTACCGGATCAGAAAAAGTAAACTCTTAGTATACGGCGATATCCTTGGAAGTACT
 941 I V P G S E K V K L L V Y G D I L G S T

 2881 ATGAACAACCTCGGTAGCTTACTGAGGACCCCTAGTGGGTGCGGGGAGCAGAATATGCTC
 961 M N N L G S L L R T P S G C G E Q N M L

 2941 GGGTTTGCGCCAGATGTGTTGCTGACTCTCTACCTCCACTCGGCGGGCAAGCTCGACGCC
 981 G F A P D V F V T L Y L H S A G K L D A

 3001 GCAACGAGAGCAAAAGCTTTCAAACATTTCCAGACTGGTTACTCTAATGAACTAAACTAC
 1001 A T R A K A F K H F Q T G Y S N E L N Y

 3061 AAGCACAGAGATGGATCATTAGTGCATTCGGTGAAGGGGACGCCTCAGGCAGCACATGG
 1021 K H R D G S F S A F G E G D A S G S T W

 3121 CTCCTGCGTTCGCTGCTAAGTGCTTTATGTTGCTCGTGAATTGCGACCCACCCTTGTC
 1041 L T A F A A K C F M F A R E L R P T L V

 3181 AGTGCAAGTGTTATTGACCAAGCTCTCACTTTCCTGATCAACCAACAAAACACAACCGGA
 1061 S A S V I D Q A L T F L I N Q Q N T T G

 3241 ACTTTCAGAGAACCTGGTGTCTCTCACAAAGCTATGCAGGGTGGAGTGGACAGCCCT
 1081 T F R E P G R V S H K A M Q G G V D S P

 3301 ATCACAATGACTGCGTATGTTCTTATTACTTTGAAGGAGACAAATTATGCTGTGAAGAAC
 1101 I T M T A Y V L I T L K E T N Y A V K N

 3361 AGGGCTGTGCAAGAAGCTGCAGAAAATGCACGAATTTATCTTGAGAATCATCTCACATCA
 1121 R A V Q E A A E N A R I Y L E N H L T S

 3421 ATCAGTGACAACAAATACGCGCTTGCTATCGTTACTTATGCTCTACATGTAGCTGGTAGT
 1141 I S D N K Y A L A I V T Y A L H V A G S

 3481 TCAAGGGCCAATGAAGCGTTACTGGCTCTTGAGGCACTTGCAACTGTACAAGGTGGATTC
 1161 S R A N E A L L A L E A L A T V Q G G F

 3541 AAATTCTGGCACGATAACTCAGAATCACCTGACTCTTACTCTTCAAGATGGCGTCTTAT

1181 K F W H D N S E S P D S Y S S R W R P Y
 3601 TATTACAACCCACCCACCAATGATATAGAGATGTCCGCTTATGCATTGCTTACATATGTG
 1201 Y Y N P P T N D I E M S A Y A L L T Y V
 3661 AGGAGAAATGACTTAAATGCTGGGATACCTGTAATGAAGTGGTTGGCATCTAAAAGAAGC
 1221 R R N D L N A G I P V M K W L A S K R S
 3721 AGTCTTGGTGGATACTCTGGTACACAGGACACAGTAATAGCCATCCAGGCTTTATCAAAG
 1241 S L G G Y S G T Q D T V I A I Q A L S K
 3781 GTAGCTGGGTTGCTTGTGGGAAATACACAGAACCTTCAAATCAGTGCCAGTCATTCAAAT
 1261 V A G L L V G N T Q N L Q I S A S H S N
 3841 GATCCTTTCACTGCAAGTTATAACATTAACAGGGAAAATTCAATCGTGTTTAACTCTGTT
 1281 D P F T A S Y N I N R E N S I V F N S V
 3901 AACGTGCCTGCTGTGGATGGTACTGTACAAGTCACAGCAACAGGTGTAGGAGTAGCAGTA
 1301 N V P A V D G T V Q V T A T G V G V A V
 3961 GCACAGATATCTGTATGTTACAATACACCTAACCAACCTTATGAAATTGAGCCATTCCAA
 1321 A Q I S V C Y N T P N Q P Y E I E P F Q
 4021 TGCCTAACACTGTCGTTTCCACTGCTTTAAAGAAAGCTAAAGTCAACTGGTGTTCAGT
 1341 C T N T V V S T A L K K A K V N W C C S
 4081 TTGAGGCCTGGGGACAACGCAACAGGCATGTTCTTAATGGAAGTTAACCTACCAAGTGGA
 1361 L R P G D N A T G M F L M E V N L P S G
 4141 TACACAGTGAATATTGACAACGAACGTACGAGAAACCCATCAGCTAAGCTTGTGAGATT
 1381 Y T V N I D N E R T R N P S A K L V E I
 4201 GATGGAAATGGAGTGAATGTTTATTATGATGAGCTCGCACCTGGCAGAAGTGTATGTGCT
 1401 D G N G V N V Y Y D E L A P G R S V C A
 4261 GATATTGAGTTACTTAATCTTGGAAATGTTGGTGGGAGTAAAGCAAGGAAAGTAGCTGCA
 1421 D I E L L N L G N V G G S K A R K V A A
 4321 TCAGATTACTACCAACCAAAGGAAAGAGTTGAGGCGTTGTACCAAGTAGATGAAGCACCG
 1441 S D Y Y Q P K E R V E A L Y Q V D E A P
 4381 GTTGTGTTGTGATTCTTGTTC AACCGAAGATATTGCTGTCTGTT CAGTCTGTGCTGATTGC
 1461 V V C D S C S T E D I A V C S V C A D C
 4441 GTTGGTTGCCAGGTCCAGCCTTTACCCAATGGTCTGAATGGTCCGACTGTGCCTTCTGC
 1481 V G C P G P A F T Q W S E W S D C A F C
 4501 GGTCGTTCCACCTCGTT CAGAACCAGAGAATGCCGAAGTCCGTTCTCAGACAATCTAGCT
 1501 G R S T S F R T R E C R S P F S D N L A
 4561 GGCCATGTATGTGGAGGTGTTGACCGTGAATCTAGGCGTTGTGTCGCAACGTTTCCATGC
 1521 G H V C G G V D R E S R R C V A T F P C
 4621 CCAGATACTTTTGATGGTTTATGTTTCAAATATGCCGAGAAATTTCCATCCTCTAACAGC
 1541 P D T F D G L W F N M P R N F P S S N S
 4681 GTCCCTTTTTACGCACACCAGTGTAGGATGGAAAGAGGTTTCGAGACAGATCAGAGAACAG
 1561 V P F Y A H Q C R M E R G S R Q I R E Q
 4741 ATTCCAGGAATCGCACTATCTGGTTCGCAATACCTGACCTGCAATAACTACGACGTCAAT
 1581 I P G I A L S G S Q Y L T C N N Y D V N

4801 CCCAATAATAACTACACCTTCTCAATTCTGGTGAAGCCGAATAGATTCCGTTCTTCTGGA
 1601 P N N N Y T F S I L V K P N R F R S S G

4861 CCAACCACTATATTTTCTTACGGAATGGAACACAATTACGCCAGAGCACACTTGGAAAAA
 1621 P T T I F S Y G M E H N Y A R A H L E K

4921 GTTTGGTGGCGCAGTGAACCTTCGCTTTAAAGTAAGATCGGATACTGGAATGAGGGAAGTT
 1641 V W W R S E L R F K V R S D T G M R E V

4981 CGTGGAGTAAGCTCAAATCTCTTGAGAACAGACCAATGGAACCATATCGTAGTAGCTGTT
 1661 R G V S S N L L R T D Q W N H I V V A V

5041 CCTTCCGGAGATGGAGACGACATTCGCATGTTTGTAAATGGCAACGCTGTCGGAAGCACC
 1681 P S G D G D D I R M F V N G N A V G S T

5101 AAATCTTTCACTACCCGGTATTTTCGGTAAACACGGAAGAAACCGGTTCTTCTCGGGCAA
 1701 K S F T T R Y F G K H G R N R F F L G Q

5161 AACACACGAGGAAACGCTTGGGCTAGAGGTTACTTCCAAGGGGTTTGGCTGCTGTTGGG
 1721 N T R G N A W A R G Y F Q G G L A A V G

5221 ACTTGGCGTTCGGTGTTAACTGACCAACAAATCACTGCTTTGTACGAAGCCTACCGACCC
 1741 T W R S V L T D Q Q I T A L Y E A Y R P

5281 GCCATTGAGTCTAGCGATCCACTTTCTGTGAAACTTCTCCGCCATTTTTCGGTCCAGCAG
 1761 A I E S S D P L S V K L L R H F A V Q Q

5341 TTACTGCTATGCTTTCAATCGCCC GCCACCATCGAGGATCTCTACAGCAGATCAGCTGCG
 1781 L L L C F Q S P A T I E D L Y S R S A A

5401 CCTGTTACATGCCAACAGCCCCCATTAGCCCCTGATGCCTTTTCTGCCGATTTTATGA
 1801 P V T C P T A P I S P L M P F L P I L *

5461 GGCTAAAGCAACATAATAAAATAACGGTGCTACTGATATACTCTTTTTGCGCTCTATACG
 1821 G * S N I I K * R C Y * Y T L F A L Y T

5521 CTTGCTATTACCAACCGCATGAACTTGTATAATTTGATTTAAATACATGATGCAATTGAA
 1841 L A I T N R M N L Y N L I * I H D A I E

5581 AAAAAAAAAAAAAAAAAAAAA
 1861 K K K K K K K

4.4 Discussion

RACE PCR cycling relies on the same principles as PCR cycling (3.2.3) using a DNA polymerase enzyme and gene specific primers that anneal to a specific region on the cDNA of interest. There were several requirements for these RACE primers as well as the general needs in the design of any successful primer that include the guanine:cytosine content and the formation of secondary structures (McPherson and Møller, 2000). Touchdown PCR (Don *et al.*, 1991) was used where possible but required primers with an annealing temperature of over 70 °C. It was not always possible to design primers with an annealing temperature of ≤ 70 °C that met the other criterion for good primer design. Ultimately, touchdown PCR was only used for SP6 5' RACE which were all unsuccessful in amplifying sequence further than the RT-PCR fragment isolated in Chapter 3.

Hot start PCR was used to ensure that there was no non-specific annealing of primer-to-template or primer-to-primer at lower temperatures during the period when the sample was heated to 94 °C to denature the template. This was achieved by a modification of the physical separation of reagents technique (Daquila *et al.*, 1991) by using an inactive *Taq* polymerase, that provides automatic hot start (Kellogg *et al.*, 1994). An antibody (in the case of the advantage cDNA polymerase mix this antibody is Taqstart™) is bound to the enzyme until denaturation, rendering it inactive, so non-specific annealing cannot take place while the sample is being heated.

Several modifications to the PCR conditions were also used in an attempt to gain successful full-length amplification of the gene of interest, without amplifying non-

specific DNA or gene fragments that were not fully transcribed in cDNA synthesis. Touchdown PCR (Don *et al.*, 1991) was adapted for RACE as the NUP (supplied in the kit) has an annealing temperature of below 70 °C. If the gene specific primer has an annealing temperature above 70 °C, initial thermal cycling above 70 °C should enable a critical amount of only gene-specific template to be amplified by the DNA polymerase, thus providing more template for later cycling. Reducing the temperature in subsequent cycles then allows the NUP to anneal to the smart sequence, added to the end of all the cDNA during reverse transcription, and then amplification continues exponentially with both primers.

Step-out PCR (Matz *et al.*, 1999) and suppression PCR (Siebert *et al.*, 1995) techniques were adopted by using the UPM (supplied in the kit) containing a long and a short universal primer. The longer primer primes from the smart RACE sequence added onto the cDNA during reverse transcription, and has an additional non-annealing overhang that is incorporated into template DNA in the early rounds of PCR. The shorter primer primes only from this overhang. This overcomes the problem of cDNA's that have the smart sequence at both ends after template switching during reverse transcriptase, which are then amplified by the UPM alone during RACE PCR causing background smearing. Those cDNA's that have the smart sequence at both ends incorporate the universal primer at both ends. The longer of the universal primers has an inverted repeat region so intramolecular binding causes the universal primer inverted repeat sequences to bind together forming a pan-handle structure, which cannot be amplified further. The shorter primer primes only from the non-annealing overhang, which can only present in cDNA correctly transcribed.

Nested RACE proved necessary to distinguish between multiple bands from the primary RACE and/or to amplify enough DNA for all subsequent procedures except for thiolester 1. Nested RACE failed for thiolester 1 whereas primary RACE succeeded. As the correct sequence was amplified from the primary RACE, and therefore must have been present in the template for the nested RACE, it is likely there was a problem with the nested primer. A faint band of the correct size was amplified according to the results obtained from the primary RACE but heavy smearing was also observed. As the 5' reaction was successful, the RNA used for the RACE cDNA synthesis must have been of good quality and cannot be the reason for the smearing. Also, as the smearing is below and above the amplified band, it cannot represent degraded RNA as premature termination of the reverse transcription would only result in multiple banding, which can appear as a smear, below the amplified fragment. The problem must have been due to the NUP priming from several templates amplified in the primary RACE that had been non-specifically amplified.

Successful annealing temperatures varied from between 4.7 °C to 0.3 °C below the primer T_m as determined by MWG Biotech (Appendix 2), while the annealing temperatures of 5-10 °C below the T_m in all cases allowed non-specific primer annealing, producing multiple banding or smearing from non-specific annealing. The higher annealing temperatures in relation to the annealing temperature highlights the level of stringency needed when using gene specific primers for this kind of RACE amplification.

The complement homologues isolated from a number of lower vertebrate (Nonaka and Takahashi, 1992; Nonaka, 1994; Nonaka *et al.*, 1998; Nonaka and Smith, 2000) or

invertebrate species (Matsushita *et al.*, 1998b; Smith *et al.*, 1998; Gross *et al.*, 1999b; Nonaka and Azumi, 1999; Nonaka *et al.*, 1999), from which degenerate primers were designed in Chapter 3 (section 3.2.2), provide an estimate of the expected size of 5' and 3' RACE fragments. From the conserved regions, used as the basis for the design of the serine protease RT-PCR degenerate primers (section 3.2.2) (Appendix 3), Bf and MASP homologues should have between 1500 and 1900 coding bp upstream of the region corresponding to the 5' RACE fragment, and between 400 and 500 coding bp downstream of the region the anti-sense primers were designed from, corresponding to the 3' RACE fragment. None of the serine protease sequences 1-5 in the present study have a coding 5' end of equivalent length and are consequently unlikely to contain all the same domains. The coding 3' ends of the serine proteases SP1, SP2, SP3, and SP5 from the present study are of similar length and may contain all the domains. The conserved thiolester region of the C3 and a2m homologues (section 3.2.2), used for the design of degenerate primers, provides an estimate for a coding 5' end of between 3000 and 3200 bp and a coding 3' end of 2000-2400 bp. The thiolester 1 sequence from the present study has a 5' and 3' end corresponding to this predicated size.

RACE amplification failed for the 5' end of SP6, although 3' amplification was successful. Several attempts to amplify the 5' sequence were tried using many different conditions. However, the same fragment was always amplified leading to the conclusion that the 5' end of the fragment isolated in Chapter 3 by RT-PCR must be the result of a non-specific reaction in the PCR. The cDNA sequence for SP6 (Fig. 4.7) extended the appropriate 1400 bp corresponding to the serine protease sequences used in the design of the degenerate primers. However, the motif used to design the reverse serine protease primers is present but much of the DNA sequence upstream of this does not correspond

to the RT-PCR fragment. This again suggests that the SP6 fragment amplified in Chapter 3 is an artefact of the PCR reaction.

However, this mRNA sequence is present in *C. intestinalis* and appears to lack any complete domain or an initiation codon that produces a useful reading frame. For this reason it is possible that this sequence represents a non-processed pseudogene. These are genes that have arisen during evolution through duplication, but lose function at the transcription or the translation levels (<http://bioinfo.mbb.yale.edu/genome/pseudogene/#what>). This theory would explain the similarity to serine protease family but the lack of an initiation codon or complete serine protease domain.

RACE was unsuccessful for thiolester 2. Although only 3' RACE was attempted it was considered that as this was the simplest RACE amplification 5' RACE was not worth attempting. The most likely reason for this failure was a problem with the DNA fragment isolated in Chapter 3 against which the primers were designed. This sequence was only identified from one clone. As no consensus sequence could be elucidated from this, errors introduced by PCR may have been present that caused critical errors in the gene-specific primer design. This sequence may also be an artefact of PCR amplification and not represent a biologically significant mRNA fragment.

In summary, 5' and 3' RACE was successful for gene fragments SP1, SP2, SP3, SP4, SP5 and thiolester 1 isolated using RT-PCR with degenerate primers in Chapter 3. Complete cDNA sequences for each of these genes was assembled containing the 5' UTR, CDS and 3' UTR to the poly (A⁺) tail. The sequences SP1, SP2, SP3 and SP5

have a coding 3' end of similar length to serine protease complement homologues from lower vertebrate and invertebrates (Kuroda *et al.*, 1996; Nakao *et al.*, 1998; Smith *et al.*, 1998; Nonaka and Azumi, 1999) but none of the serine protease sequences have a similar sized 5' end. Thiolester 1 has a 5' and 3' end of a similar size to all other C3 and a2m homologues (Nonaka and Takahashi, 1992; Nonaka *et al.*, 1998; Gross *et al.*, 1999b; Nonaka *et al.*, 1999).

Chapter 5

Bioinformatic Analysis

5.1 Introduction

In order to investigate if the genes isolated in Chapter 4 are likely to be involved in a complement system, a bioinformatic analysis of each gene was performed and compared to the known structure and function of characterised complement homologues. A bioinformatic approach was chosen for several reasons. An array of bioinformatic tools has recently been developed to characterise the vast number of protein sequences that have been generated from the human genome project and from full-length cDNA projects (Edwards and Cottage, 2001). The number of these sequences is many orders of magnitude greater than experimentally deduced protein structures (Edwards and Cottage, 2001), so extra need has arisen to predict accurately the structure and function of a protein from its sequence. Use of these, mainly web based, bioinformatic tools allow the scrutiny of sequences against others that have been experimentally characterised. Although this technique is not as reliable as direct analysis of experimentally deduced structures, it is an accurate way to predict protein characteristics before experimental proof is obtained (Baxevanis and Ouellette, 1998). This is because a multitude of programmes are used to develop a consensus of the most likely protein structure.

Other studies into the evolution of complement have used a bioinformatic approach to characterise the sequence information that molecular studies have produced (Kaidoh and Gigli, 1989; Ishikawa *et al.*, 1990; Hanley *et al.*, 1992; Ji *et al.*, 1997; Matsushita *et al.*, 1998b; Nakao and Yano, 1998; Pahler *et al.*, 1998; Nonaka and Azumi, 1999; Nonaka *et al.*, 1999; Endo *et al.*, 2000; Nair *et al.*, 2000; Nonaka, 2001; Sekine *et al.*, 2001). Bioinformatic analysis now provides a reliable and up-to-date method to determine the likely function of a gene before embarking on experimental analyses of the findings.

When dealing with well-characterised protein domains with well-studied representatives of a gene family, such as all the complement genes, the results can be very reliable. However, it is very difficult, if not impossible, to assign probable functions to novel and less well understood genes that contain domains of unknown function or known domains in different arrangements.

The aim of this chapter is to scrutinise the complete amino acid sequences of the expressed genes isolated in Chapter 3 and fully sequenced in Chapter 4. The bioinformatic information obtained will allow putative functional assignments to be given to each of the sequences and a possible role in immunity or a complement system can be predicted.

5.2 Materials and Methods

Web based bioinformatic tools were chosen for the analysis of these sequences because of their free availability and accuracy.

5.2.1 Similarity Searching

Sequence databases used for similarity searches were EMBL (Baker *et al.*, 2000), SWISS-PROT and TrEMBL (Bairoch and Apweiler, 1998) all of which are maintained by the European Bioinformatics Institute; a member of the International DNA Databases. All publicly available sequence information is represented in one or more of these databases. Similarity searches, with the cDNA sequences cloned in Chapter 4 of this study, within these databases was performed using the programmes BLAST and PSI BLAST (Altschul *et al.*, 1990; Altschul *et al.*, 1997). These are available from the Japanese GenomNet (<http://gfit.genome.ad.jp>) and the Human Genome Mapping Project Resource Centre (HGMP-RC) (<http://www.hgmp.mrc.ac.uk>). By altering the searching parameters available at these sites stringency and effectiveness can be changed, allowing for more or less powerful similarity searching, tailoring the search for each specific sequence. Both BLOSUM (Henikoff and Henikoff, 1993) and PAM (Jones *et al.*, 1992) scoring matrices were used. BLOSUM produces a matrix of amino acid substitution scores from multiple alignments of related proteins (Henikoff and Henikoff, 1993), whereas PAM predicts the amount of expected substitution for a given amount of evolutionary time (Jones *et al.*, 1992). BLOSUM also allows for the most related sequences to be clustered, reducing the frequency of changes among the most common amino acids (Henikoff and Henikoff, 1993).

To determine if any high scoring alignment pairs (the pairwise alignment of a database sequence similar to the query sequence) are significantly similar, the Expect-value (E-value) and P-value are reported. The E-value is a probability range from 0 to 1 of the number of similarity hits that would have arisen by chance. The closer the E-value is to 0, the less likely this match is due to chance. E-values are dependent upon several factors, including the length of the query sequence, the size of the database and the scoring system used. They cannot be meaningful unless all this information is available. The P-value is the score in bits that allows a reliability value to be placed on the alignment of a high scoring pair. Bits are units of information representing the increase in reliability associated with a unit increase in the alignment score (Altschul *et al.*, 1997). This value, although in part variable between scoring systems and databases, is a more reliable statistical comparison between similarity searches. The higher the number of bits, the more reliable the information.

5.2.2 Protein Domains and Motifs

A protein domain is a structurally important and consequently stable globular region of a protein (Edwards and Cottage, 2001). The secondary structures of domains usually contain a hydrophobic core (http://smart.embl-heidelberg.de/help/smart_glossary.shtml). Due to its functional importance, characteristics of domains are often conserved between proteins spanning many different phyla (Edwards and Cottage, 2001). Motifs are usually smaller regions of amino acids or structure that are also functionally important conserved between proteins of similar function (<http://www.mblab.gla.ac.uk/dictionary/>).

Protein domains within the amino acid sequences from this study are compared to other proteins whose domains have been characterised. Domains are recognised using the same principles as similarity searches because important functional regions have been identified using multiple sequence alignments. Prodom (Jones *et al.*, 1992; Corpet *et al.*, 2000) (<http://www.protein.toulouse.inra.fr/prodom.html>) and Pfam (Bateman *et al.*, 2002) (<http://www.sanger.ac.uk/Software/Pfam/search.shtml>) contain many sequence alignments and hidden Markov models that cover many protein domains likely to have been conserved through their function.

Several software applications are freely available on the Internet for the prediction of protein domains. InterPro was used for a comprehensive analysis against all the known and characterised sequences using the different methods each search tool employs, (Apweiler *et al.*, 2001) (<http://www.ebi.ac.uk/interpro/scan.html>). This incorporates the signature databases PROSITE (Falquet *et al.*, 2002), PRINTS (Attwood *et al.*, 2002), Pfam (Bateman *et al.*, 2002), ProDom (Corpet *et al.*, 2000), SMART (Letunic *et al.*, 2002) and TIGRFAMs (Haft *et al.*, 2001). Thus, InterPro is an integrated layer on top of these databases creating a comprehensive, non-redundant characterisation of a given protein family, domain or functional site. (<http://www.ebi.ac.uk/interpro/README1.html#>).

Domain signatures have been built up from multiple alignments and are described in the form of consensus sequences in a number of database annotations, including PROSITE (<http://ca.expasy.org/tools/scanprosite/>). These have been built using a multiple alignment of all the known sequences containing a known domain to develop an amino

acid signature that fits all the different sequences. Such residues are likely to have been conserved through their functional importance.

Once a domain or likely domain structure has been recognised, information about it is available from several of the annotation notes e.g. PROSITE documentation (<http://ca.expasy.org/prosite/>). These notes describe the known and probable functions of each specific domain, as well as links to the domain profile and gene sequences that are known to contain it. Ultimately, once a domain profile has been built up and the information obtained, it is possible to assign a putative function to the gene.

Combining the information obtained from similarity and domain searching it is possible to ascertain if an amino acid sequence, deduced from the cloned cDNA sequences, is structurally and functionally similar to any proteins from other organisms. Thus, in the present study the domain and similarity profile for each of the genes isolated in Chapter 4 were compared to other protein sequences available on the databases and the functional domains they contain. This was achieved manually using sequences determined as significantly similar by BLAST (Altschul *et al.*, 1990; Altschul *et al.*, 1997; Adams *et al.*, 2000) and also by using ProfileScan (Swiss Institute for Experimental Cancer Research) available through SMART (<http://smart.embl-heidelberg.de/>) that searches SwissProt/ TrEmbl databases for proteins of a similar domain composition. All this information can be combined to elucidate whether or not a protein sequence from this study is a complement homologue. If it is not, then the information can be used to infer its likely function.

5.2.3 Further Analysis

Comparison with the known complement homologues from the invertebrates and lower vertebrates was performed using published information (Nonaka *et al.*, 1984; Fujii *et al.*, 1992; Hanley *et al.*, 1992; Nonaka and Takahashi, 1992; Al-Sharif *et al.*, 1997; Ji *et al.*, 1997; Ji *et al.*, 1998; Matsushita *et al.*, 1998b; Quesenberry *et al.*, 1998a; Smith *et al.*, 1998; Nonaka and Azumi, 1999; Nonaka *et al.*, 1999; Ji *et al.*, 2000; Zarkadis *et al.*, 2001a). For each gene family to which these proteins belong, several characteristics are key to their complement activity and have been described in detail (Jensen *et al.*, 1981; Bentley, 1988; Law and Dodds, 1990; Farries and Atkinson, 1991; Hanley *et al.*, 1992; Nonaka and Takahashi, 1992; Dodds and Day, 1993; Nonaka, 1994; Nonaka *et al.*, 1994; Dodds *et al.*, 1996; Kuroda *et al.*, 1996; Turner, 1996; Ji *et al.*, 1997; Law and Dodds, 1997; Armstrong *et al.*, 1998; Dodds and Law, 1998; Dodds *et al.*, 1998; Ji *et al.*, 1998; Matsushita *et al.*, 1998a; Nakao *et al.*, 1998; Nakao and Yano, 1998; Nonaka *et al.*, 1998; Smith *et al.*, 1998; Cross *et al.*, 1999; Gross *et al.*, 1999b; Lee, 1999; Nonaka and Azumi, 1999; Nonaka *et al.*, 1999; Azumi *et al.*, 2000; Clow *et al.*, 2000; Endo *et al.*, 2000; Gross *et al.*, 2000; Kuroda *et al.*, 2000; Nakao *et al.*, 2000; Dahl *et al.*, 2001; Nakao *et al.*, 2001; Nonaka, 2001; Zarkadis *et al.*, 2001b).

As the present study is concerned with looking for some of the most primitive complement molecules, it is possible that ancestral molecules may not precisely fit the model complement gene characteristics, most recently described by Nonaka (2001). Instead, they may have only some of the features possessed by the complement proteins of the more developed complement systems of vertebrates and be too different from the model for the bioinformatics applications to match. It was therefore necessary to obtain

the domain profile and determine by eye if a similar consensus pattern was present in the amino acid sequences. This is also why several bioinformatic resources were used. The results of the different searching methods could be compared to determine the most likely positive result in case of ambiguity.

Other bioinformatic analysis packages were used if similarity searching for homologous sequences and protein domains was insufficient to confidently determine the function and/or homology of a cDNA sequence. Web based software is available to elucidate information about a sequence that is not as dependent on similarity searching. These have been designed using the information available about the most characterised protein structures and their amino acid sequences, thereby enabling correlation between a protein feature and its amino acid sequence. Ultimately, certain specific features can now be searched for within an amino acid sequence and its secondary structure. Examples of these are cleavage sites (von Heijne, 1986), trans-membrane sites (Hofmann and Stoffel, 1993), theoretical molecular weight (Wilkins *et al.*, 1997), theoretical Pi (Bjellqvist *et al.*, 1993), hydrophobicity (ProtScale; <http://ca.expasy.org/cgi-bin/protscale.pl>), secondary structure folding pattern (King *et al.*, 1997; King and Sternberg, 2002; In Press) or post-translational modifications (PSORT; <http://psort.nibb.ac.jp/>).

5.2.4 Multiple Alignment and Phylogenetic Trees

Sequences found using BLAST (Altschul *et al.*, 1990) and PSI Blast (Altschul *et al.*, 1997) as well as known complement homologues (Nonaka *et al.*, 1984; Fujii *et al.*, 1992; Hanley *et al.*, 1992; Nonaka and Takahashi, 1992; Al-Sharif *et al.*, 1997; Ji *et al.*, 1997;

Ji *et al.*, 1998; Matsushita *et al.*, 1998b; Quesenberry *et al.*, 1998a; Smith *et al.*, 1998; Nonaka and Azumi, 1999; Nonaka *et al.*, 1999; Ji *et al.*, 2000; Zarkadis *et al.*, 2001a) were aligned together with the genes isolated in this study using CLUSTAL W (Thompson *et al.*, 1994) available from the European Bioinformatics Institute (<http://www2.ebi.ac.uk/clustalw/>) and from the Pole BioInformatique Lyonnais (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_clustalw.html). CLUSTAL W uses molecular data in the form of sequences to determine relatedness by aligning sequences based on their most similar regions (Thompson *et al.*, 1994). From these alignments phylogenetic trees can be produced to indicate the evolutionary distance between the sequences by the number of differences between them. These differences represent mutations, the frequency of which increases over time, indicating the level of divergence between species and sequences. In the present study phylogenetic trees were produced using CLUSTAL W using the multiple alignments mentioned above.

The most powerful alignments use only orthologous sequences that have diverged through speciation; paralogues have diverged through gene duplication. Homologous sequences share a common evolutionary ancestor and have arisen through gene duplication; orthologues and paralogues are both homologues (Edwards and Cottage, 2001). Alignment of such sequences is very powerful. Analagous sequences are non-homologous and share a similar function or tertiary structure thought to have arisen through convergent evolution (Edwards and Cottage, 2001) alignments including these sequences provides little information for evolutionary questions. If a protein has an identity of 30 % or higher over a region of 70 or more amino acids to another protein of known structure or function, it can be safely described as homologous (Edwards and Cottage, 2001).

Commercial software is also available for multiple alignment phylogenetic tree analysis. The DNAMAN programme version 5.2.0 (Lynnon BioSoft, Quebec Canada) was used in the present study alongside CLUSTAL W (Thompson *et al.*, 1994). DNAMAN allows alignments to be constructed using the same parameters as CLUSTAL W (Thompson *et al.*, 1994) with colours rather than symbols illustrating the more conserved regions. DNAMAN produces a more intuitive graphical image than CLUSTAL W (Thompson *et al.*, 1994), which has symbols rather than colours representing conserved regions. Phylogenetic trees can also be produced in DNAMAN using these multiple alignments. CLUSTAL W was used through the interface available at Pôle BioInformatique Lyonnais; Lyn-Gerland (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_clustalw.html)

5.3 Results

In the present study using all these methods, the significance values are not given as they cannot be directly compared with each other.

5.3.1 Serine Protease 1 (SP1)

Figure 5.1 Schematic diagram of the domain organisation of SP1. TSP1 is thrombospondin type 1 repeat domain.



BLAST similarity searches reveal this sequence is significantly similar to mosaic serine protease from *Homo sapiens* (accession number BAB39742). Significant similarity is also observed to hypothetical proteins based on genomic sequences from *Caenorhabditis elegans* (accession number T25061). A vast majority of the serine proteases significantly similar to SP1 are involved in coagulation pathways including fibrilin, plasminogen, kallikrien and coagulation factor IX. Of these, the most similar was human plasminogen (accession number P00747). Another serine protease significantly similar to SP1 was a corticle granule serine protease from the sea urchin, *Strongylocentrotus purpuratus* thought to be critical in blocking polyspermy (accession number AF149789). The BLAST results showed that significantly similar serine protease sequences were from both invertebrates including *C. elegans* as well as higher mammals including humans.

BLAST searches against the expressed sequence tagged sub-database that does not include human or mouse data revealed that SP1 is most like a larval *Ciona intestinalis* clone (accession number AL666970). However, this is the 5' end sequence from a cDNA library and there is no information about the total length of this cDNA. The only indication that it is a serine protease is because it aligns to SP1 at the 5' end. This is a similar sequence from the same species it is not the same sequence.

The protein domain profile built up using the range of bioinformatic tools detailed above was four thrombospondin type 1-like domains (TSP1) repeats from residues 35-79, 81-128, 137-177 and 181-228 (Fig. 5.1). This same number of domains and their positions was predicated by all the applications. Repeats are sequences known to form a globular region when in combination with one another, i.e. cannot form a domain alone. A trypsin serine protease domain (spanning residues 261-507) was identified by all the applications used, immediately after the TSP1-like repeats (Fig. 5.1). The trypsin family serine protease domain indicates this protein belongs to the chymotrypsin superfamily, and indeed the chymotrypsin signature was identified by PRINTS (Attwood *et al.*, 2002). This trypsin domain has both an active histidine region (consensus pattern [LIVM]-[ST]-A-[STAG]-H-C) and active serine region (consensus pattern [DNSTAGC]-[GSTAPIMVQH]-x(2)-G-[DE]-S-G-[GS]-[SAPHV]-[LIVMFYWH]-[LIVMFYSTANQH]). The catalytic activity of the trypsin family serine proteases is reliant upon both the active histidine and serine. An aspartic acid residue is hydrogen-bonded to the histidine which itself is hydrogen-bonded to the serine, producing a charge relay system (PROSITE documentation PDOC00124). No other sequences in the databases share this same domain structure.

Multiple alignment of SP1 with the most similar sequences reveals the best alignment is with the trypsin serine protease domains at the C-terminus (approximately from alignment point 620 on Appendix 6). All the aligned sequences share this region and the majority of the cysteine residues align in the same place (Appendix 6). The lowest identity with SP1 is with the hypothetical protein from *C. elegans* at 14.93 %. No serine protease domain is present in this protein but seven TSP1 domains are predicted and this produces significant identity scores using BLAST. The cortical granule protein from, *S. purpuratus*, contains one TSP1 domain before the serine protease domain from residues 261-308 and has a pair wise identity of 27.17 % with SP1. This level of identity is still below the level where proteins can be confirmed as homologous.

5.3.2 Serine Protease 2 (SP2)

Figure 5.2 Schematic diagram of the domain organisation of SP2.



Blast homology searches again revealed this sequence was a serine protease of the trypsin family. The significantly similar proteins were again coagulation serine proteases although the level of significance was lower than that of SP1. The most significant matches were with kallikrein precursor from the mouse, *Mus. musculus* (accession number P26262), a putative immune serine protease from the African malaria mosquito, *Anopheles gambiae* (accession number AF117751) and kallikrien precursor from the Norway rat, *Rattus norvegicus* (accession number O88780). These sequences

were significantly similar to SP2 based mainly on the 3' end alignment, the area containing the serine protease domain.

Only one domain was identified using both the bioinformatic software and manual searching for likely domains. This was the trypsin serine protease domain at the 3' end of the molecule from residue 278 to the final residue 433 (Fig. 5.2). This domain possesses the chymotrypsin protein family signatures described by PRINTS (Attwood *et al.*, 2002) as well as the catalytic histidine and serine residues. No domain profile was matched for the first 277 amino acids of the 5' end before the beginning of the serine protease domain at residue 228. A low complexity transmembrane helix region is predicted at the beginning of the sequence by TMHMM2 (Krogh *et al.*, 2001; Moller *et al.*, 2002). However, as this is a serine protease the most likely explanation for this is a low complexity signal peptide at the beginning of the N-terminus. This result was disregarded.

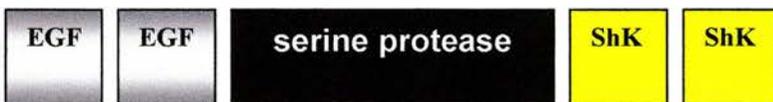
BLAST was then performed with just the 277 amino acids of the 5' end to determine if this region was similar to other protein sequences, although either no domain profile is known or this sequence is too far removed from any consensus pattern. Results showed that although the significance level was lower, significant matches were still found for coagulation serine proteases, the most significant of which was clotting factor IX from the domestic rabbit, *Oryctolagus cuniculus* (accession number P16292), sodium channel activating serine protease from the African clawed frog, *Xenopus laevis* (accession number AF029404) and plasminogen activator from the Korean centipede, *Scolopendra subspinipes* (accession number AAD00320).

Multiple alignments using CLUSTAL W were carried out with the sequences shown to be similar at both the N-terminus and the C-terminus (Appendix 7). It is clear that the alignment of the serine protease domain at the C-terminus of all these proteins provides the most powerful similarity (Appendix 7). The N-terminus similar sequences, identified using BLAST P, produce less significant alignments and do not show any conserved region between SP2 and any of these proteins. No domain can be predicted to be in SP2 from any of these sequences.

Pair wise alignment reveals that the sequence showing the least identity to SP2 is the serine protease from *A. gambiae* at 13.94 %. This is due to the much larger N-terminus region before the serine protease domain in comparison to SP1 and all the other serine proteases in this alignment. All the other sequences in this alignment show a similar level of identity between 21 and 24 %, the highest being plasminogen activator from *S. subspinipes*, at 24.04 %. At this level of identity homology between these sequences and SP2 cannot be presumed.

5.3.3 Serine Protease 3 (SP3)

Figure 5.3 Schematic diagram of the domain organisation of SP3.



Several serine proteases are significantly similar to SP3 from this study. The highest level of significance is with riken and riken fragments from *M. musculus* (accession numbers AK004939, BC024903). However, SP3 shows a similar significance level of

similarity to a range of serine proteases from many mammalian species, several of which are involved in coagulation pathways.

A domain profile of SP3 begins with at least one epidermal growth factor-like domain (EGF) (Fig. 5.3). Two types of EGF-like domain are predicted that both span 6 conserved cysteine residues that form disulfide bonds. Both types of domain are usually between thirty and forty amino acids in length. The first is an EGF-2 (consensus pattern x(4)-C-x(0,48)-C-x(3,12)-C-x(1,70)-C-x(1,6)-C-x(2)-G-a-x(0,21)-G-x(2)-C-x) and the second a calcium binding EGF domain (EGF-CA) (consensus pattern nxnnC-x(3,14)-C-x(3,7)-CxxbxxxxaxC-x(1,6)-C-x(8,13)-Cx). Pfam (Bateman *et al.*, 2002) places the domain between positions 24-59 a region spanning six conserved cysteine residues and prosite/motifs and SMART (Schultz *et al.*, 1998) recognises an EGF-2 domain signature between residues 44-59. Prosite (Falquet *et al.*, 2002) and SMART (Schultz *et al.*, 1998) identifies an EGF-CA functional signature from position 61 to 109/110 also spanning six cysteine residues.

A trypsin serine protease domain is identified from positions 138-371 confirming this is a serine protease sequence of the chymotrypsin superfamily containing a catalytic histidine and serine (Fig. 5.3). Two ShK toxin-like domains are predicted after the serine protease domain between residues 389-424 and 432-470 by SMART (Schultz *et al.*, 1998) (Fig. 5.5). This domain has been identified in metridin, a toxin from the sea anemone, *Metridium senile*, and several hypothetical proteins from *C. elegans* (http://smart.emblheidelberg.de/smart/do_annotation.pl?DOMAIN=ShKT&BLAST=DU MMY&LITERATURE>Show#Literature). No other sequences on the databases share the same domain profile as SP3.

Alignment with some of the most significantly similar protein sequences reveals, again, that the best alignment is obtained in the serine protease domain of these sequences (Appendix 8). The serine protease domain of SP3 aligns well with all the other sequences but shows the highest level of identity with mouse coagulation factor IX (accession number AF356627) at 20.91 %. Only the serine protease domain at the C-terminus is shared between SP3 and mouse coagulation factor XI (Appendix 8). None of these proteins showing similarity to SP3 share any other similar domains except the serine protease domain at the C-terminus. This level of identity is too low to represent any homology between these proteins.

5.3.4 Serine Protease 4 (SP4)

Figure 5.4 Schematic diagram of the domain organisation of SP4. ShK is ShK toxin domain, EGF is epidermal growth factor domain and TSP1 is thrombospondin type 1 domain.



Sequences determined as significantly similar using BLAST were mainly serine proteases. However, alignments with the highest significance were fibulin-6 (accession number CAC37630) and hemicentin (accession number AAK68690) from *H. sapiens*, neither of which are serine proteases. Among those sequences with high similarity were hypothetical proteins from *C. elegans* (accession numbers AAK68231 and T25061) and angiogenesis inhibitor homologue (accession number T18856) from *C. elegans*, as well

as semaphorin 5B precursor (accession number Q60519) and riken (accession number AK004939) from *M. musculus*.

SP4 is similar both to SP1 and SP3 with respect to the domains present (Fig. 5.4) but the arrangement is unique. Three ShK-like domains are present one at the extreme 5' end from residues 33-67 and two at the extreme 3' end from residues 1011-1050 and 1055-1089 (Fig. 5.4). After the initial ShK-like domain, two EGF-like domains predicted by SMART (Schultz *et al.*, 1998) are present from residues 72-108 (EGF-2) and 109-157 (EGF-CA) (Fig. 5.7). These signatures are also identified by Prosite/motifs (Falquet *et al.*, 2002). Pfam (Bateman *et al.*, 2002) identifies only the EGF-CA domain but manual observation confirms that two non-overlapping domain signatures are present each spanning six cysteine residues that are very close to the consensus pattern.

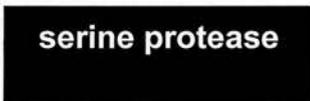
A trypsin serine protease domain is present from residues 172-405 (Fig. 5.4). It too contains a catalytic histidine and serine residue and confirms this protein is part of the chymotrypsin superfamily. Eleven TSP1-like repeat domains follow from residues 423-470, 474-526, 528-578, 582-633, 638-690, 692-742, 744-796, 798-849, 851-902, 905-955, 956-1006, 1011-1050, 1055-1089 (Fig. 5.4). Both Pfam (Bateman *et al.*, 2002) and SMART (Schultz *et al.*, 1998) identify the same number of domains in the same places with high levels of significance. No other proteins on the databases have a similar domain structure.

Those sequences that BLAST recognised as significantly similar to SP4 are very different in their overall length and domain structure making a meaningful multiple alignment very difficult. After several attempts at producing a multiple alignment, it is

clear that these proteins cannot not be aligned informatively due to the differences in the positions and types of domains present. Neither the *C. elegans* hypothetical proteins, fibulin-6 nor semaphorin B contain a serine protease domain although they have several TSP1 repeats. These are also the sequences that show the highest identity to SP4. Both the hypothetical proteins from *C. elegans* have identities of over 22.0 %; the highest being mouse semaphorin B at 25.59 % to SP4. Only mouse riken has a serine protease domain but does not have any TSP1 repeats, and consequently, has a lower identity value at 17.26 %. Fibulin-6 contains an EGF domain at its C-terminus but shows the lowest identity at 15.29 %. None of these identity values indicates that any of these sequences are homologous to SP4.

5.3.5 Serine Protease 5 (SP5)

Figure 5.5 Schematic diagram of the domain organisation of SP5.



SP5 is a shorter protein sequence (225 amino acids) than the previous four. Similarity searching reveals that this serine protease is significantly similar to a range of serine proteases. The most significant matches are to trypsin from the shrimp, *Litopenaeus vannamei* (accession number Y15041) and trypsinogen from the puffer fish, *Takifugu rubripes* (accession number U25747). Both are incomplete fragments of the coding sequence. Significant complete proteins identified by BLAST searches included a serine protease from *M. musculus* (accession number BC010970) similar to distal intestinal

serine proteases and a predicted protein from a genomic scaffold from the fruit fly, *Drosophila melanogaster* (accession number AE003455).

Pfam (Bateman *et al.*, 2002) predicts a trypsin serine protease domain from residue 34-225 (Fig. 5.8). The same domain from residue 1-220 is predicted by SMART (Schultz *et al.*, 1998)(Fig. 5.5). This information confirms the presence of this domain and SP5 as a member of the serine proteases from the chymotrypsin family. However, the precise location of this domain is not clear as the consensus pattern is fulfilled in both these areas. No catalytic residues are predicted.

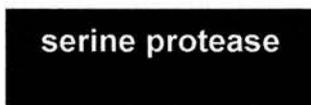
Alignment reveals that it is the serine protease domain in all these proteins that produces best alignment (Appendix 9). Alignment with all the significantly similar proteins produces an overall identity of over 20 %. The shrimp, *L. vannamei* trypsinogen fragment has the highest level of identity with SP5 at 40.74 %. Of the complete amino acid sequences, mouse distal serine protease (accession number BC010970) has the highest level of identity with a value of 31.38 %. Both the sequence fragments from *T. rubripes* and *L. vannamei* have only a serine protease domain. However, the complete sequences from *D. melanogaster* and *M. musculus* also have only a serine protease domain. The genomic scaffold from *D. melanogaster* has a lower level of identity at 20.62 %.

A level of identity over 30 % allows homology to be inferred between these two sequences. The distal intestinal serine protease from mouse is a complete protein sequence and has 31.38 % identity, so these proteins may have evolved from the same common ancestor. However, at such a level of identity this is a putative hypothesis.

The higher level of identity shown by the trypsin and trypsinogen fragments from an invertebrate and lower vertebrate indicates homology but due to the incomplete sequences this identity level may not reflect the identity of the whole protein.

5.3.6 Serine Protease 6 (SP6)

Figure 5.6 Schematic diagram of the domain organisation of SP6.



Blast similarity searching against expressed sequence tagged library clones (EST) shows that this sequence has been detected in *C. intestinalis* as an EST (accession number AV968564). This match is from residue 137-204 along the serine protease domain of SP6. This EST clone does not match SP6 along its entirety, as there are some discrepancies at both the extreme 5' and 3' end of the clone. The sequencing of the EST clone was only performed once and it is indicated that the sequence is the 5' end of the clone (accession number AV968564).

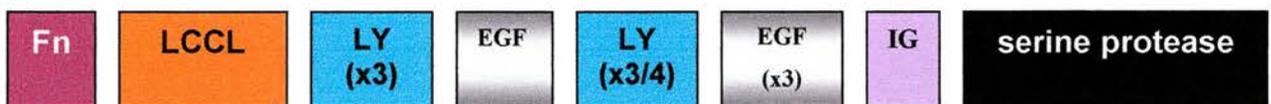
Sequences other than this identified as significantly similar by BLAST were enteropeptidase precursor from *H. sapiens* (accession number P98073), the pig, *Sus scufra* (accession number P98074) and *M. musculus* riken (accession number AK004939).

SP6 has a length of only 206 amino acids and the 5' does not extend the full distance of the PCR clone as discussed in Chapter 4 (section 4.4). Both SMART (Schultz *et al.*, 1998) and Pfam (Bateman *et al.*, 2002) predict a trypsin serine protease domain from residue 2/3 respectively to residue 184 (Fig. 5.6). Twenty-one amino acids remain at the C-terminus end of the sequence and no features are present.

Multiple alignment reveals that the serine protease domain of SP6 shares motifs with the serine protease domain of both enteropeptidase sequences and mouse riken. However, the N-terminus of all three sequences SP6 is aligned with show no similarity to SP6. This is reflected in the identity of these proteins to SP6, the highest being mouse riken at only 11.11 %. None of these sequences are homologous to SP6.

5.3.7 Serine Protease 7 (SP7)

Figure 5.7 Schematic diagram of the domain organisation of SP7. Fn is fibronectin domain. LCCL is LCCL domain, LY is low-density lipoprotein receptor type B domain, EGF is epidermal growth factor domain and IG is an immunoglobulin domain.



Similarity searching shows that this sequence is significantly similar to a number of low-density lipoprotein receptor proteins and related proteins. The matches that produce the most significance are low-density lipoprotein receptor protein from *M. musculus* (accession number AF247637) and MEGF7 (multiple epidermal growth factor) protein from humans (accession number BAA32468). Both these proteins are membrane bound. The majority of the most significantly matched sequences are lipoprotein receptor

proteins or related proteins based on their similarity to known lipoprotein receptors. Other sequences showing less, but still good significance are alpha2-macroglobulin receptor precursor from *M. musculus* (accession number Q69219), coagulation X precursor from chicken, *Gallus gallus* (accession number P25155) and mannan associated serine protease (MASP) from the common carp, *Cyprinus carpio* (accession number AB009073).

The 5' end of SP7, before the trypsin serine protease domain, is the largest of all the serine protease sequences from this study spanning 983 amino acids (Fig. 5.7). Within this region approximately thirteen domains are predicted (Fig. 5.7). SigCleave (available through the Human Genome Mapping Project; <http://hgmp.mrc.ac.uk>) (von Heijne, 1986) also predicts a signal peptide sequence between residues 1-18. SMART (Schultz *et al.*, 1998) then predicts a fibronectin-like domain from positions 22-68 (Fig. 5.7). This identification is on the threshold of significance but more confidence is gained from the position of the domain. The previous signal peptide finishes four amino acids before this fibronectin domain and the following domain after this fibronectin-like domain begins five residues downstream.

An LCCL-like domain beginning at position 73 to 153 is identified by SMART (Schultz *et al.*, 1998) followed by a region of sequence for which no domain was elucidated (Fig. 5.7). This unknown region spans from position 154 to 240. After residue 240 is a region comprising of several low-density lipoprotein receptor type B domain repeats (LY), as identified by both SMART and Pfam (Fig. 5.7). Differences occur in the number of domains that are predicted due to differences in the signature profile used by these

programmes. SMART (Schultz *et al.*, 1998) predicts seven LY repeats from residues 241-282, 284-328, 329-372, 373-415, 531-573, 577-621, and 622-644, whereas Pfam predicts six LY repeats from positions 263-303, 305-348, 350-398, 551-596, 598-640 and 642-685. Both SMART (Schultz *et al.*, 1998) and Pfam (Bateman *et al.*, 2002) identify a predicted EGF-like domain between these LY domains from residues 443-472 but this is below the threshold of significance.

After the final LY domain region three EGF-like domains are identified (Fig. 5.7). The first, from positions 735-776, incorporates six conserved cysteine residues from the domain signature, but positioning of the internal amino acids between these cysteines is not so conserved. This causes the significance value of this match to be low. Both SMART (Schultz *et al.*, 1998) and Pfam (Bateman *et al.*, 2002) confidently predict two EGF-CA domains from amino acids 797-837 and 841-876. SMART (Schultz *et al.*, 1998) and Pfam (Bateman *et al.*, 2002) predict an immunoglobulin-like domain before the serine protease region from residues 880-963 and 888-947 respectively (Fig. 5.7). The 3' serine protease domain begins at residue 983 ending at 1231 and includes a catalytic histidine and serine residues and motifs associated with proteins of the chymotrypsin superfamily of proteins.

Alignment of those sequences identified by BLAST as being significantly similar shows SP7 has most identity with human MEG7 at 22.40 % (Appendix 11). MASP from carp and chicken coagulation factor X are the only proteins that have a serine protease domain and show 19.04 and 14.64 % identity respectively. The best alignment with these sequences begins at the serine protease domain from alignment position 3082 (Appendix 11). Carp MASP shows the highest identity level of the serine proteases due

to the EGF-like domain present in both this protein and SP7. Identity levels of these proteins to SP7 are too low to consider homology. None of these proteins show identity levels consistent with homology.

Importantly, pair-wise alignment with LDLR related proteins highlights a conserved region in both SP7 and mouse alpha2-macroglobulin receptor at alignment position 1662-1780 (Appendix 11). This region spans three LY repeats in both proteins from residues 537- 650 in SP7. The identity between these regions is 35.90 %, indicating homology.

5.3.8 Thiolester 1 (Thiol1)

Figure 5.8 Schematic diagram of the domain organisation of thiol1. A2m n is alpha2-macroglobulin N-terminus domain, A2m is alpha2-macroglobulin domain, TSP1 is thrombospondin type 1 domain and pentraxin is pentraxin domain.



All the sequences with a significant level of similarity found with BLAST P have a thiolester motif. These include alpha2-macroglobulins, TEP proteins from arthropods, alpha1-inhibitors and pregnancy zone proteins as well as thiolester containing proteins of the complement system. The protein sequences that showed the highest level of significance were cell surface antigen CD109 from *Homo sapiens* (accession number AF410459), glycosylphosphatidylinositol-linked (GPI-linked) protein from *M. musculus* (accession number AY083458), TEP1 and TEP2 from *D. melanogaster* (accession

numbers AJ269538 and AJ269539), *Limulus polyphemus* alpha2-macroglobulin (accession number D83196), alpha1-inhibitor III from *R. norvegicus* (accession number P14046) and a hypothetical protein from *C. elegans* (accession number Z82090).

BLAST searching in expressed sequence tagged (EST) protein databases reveals that almost identical sequence fragments have been isolated from *C. intestinalis* both from the adult and from larvae. The 3' and 5' end of several clones from EST libraries have identities to thioll ranging from 79 % to 94 % (accession numbers AV837571, AV837399, AV678703, AV947947, AV955787, AV850887, BP018415, AL666319 and AL664921). These sequences have been entered on the database after one sequencing attempt and consequently contain errors. For this reason identity levels for the same gene sequences are below 100 % as some of the bases are incorrect.

Both human CD109 and mouse GPI-linked protein are alpha2-macroglobulin homologues that have a GPI-linked post-translational modification. Big-PI predictor (<http://www.mendel.impunivie.ac.at/gpi/>) uses consensus patterns in the amino acid composition around the site of post-translational modification at the C-terminus where GPI-linking occurs. Analysis of thioll, CD109 and mouse GPI using Big-PI predictor only highlighted a potential GPI modification site for mouse GPI showing that predicting a GPI-linkage in this manner is not reliable. Visual observations of the amino acid composition of thioll show that the C-terminus displays several characteristics of GPI-linkage. These characteristics are an unstructured linker region of about eleven residues followed by a region of small residues, including the proteolytic cleavage site for GPI attachment, followed by a spacer region of moderately polar residues, ending in a hydrophobic tail to the C-terminus (Eisenhaber *et al.*, 1999). Amino acid composition

is not conserved between GPI-linkage sites (Eisenhaber *et al.*, 1999). The C-terminus residues of thioll (residues 1777-1780) are small amino acids followed by seven polar residues before a chain of hydrophobic residues from position 1788 to the C-terminus at position 1809. A leader sequence before this domain is shorter than eleven amino acids.

Pfam (Bateman *et al.*, 2002) detects two alpha2-macroglobulin domains in thioll from residues 1-647 and 723-1444 (Fig. 5.8). The first is an alpha2-macroglobulin N domain that is associated with the N-terminus of alpha2-macroglobulin molecules and the second is an alpha2-macroglobulin domain associated with the internal sequence. SigPep (von Heijne, 1986) predicts a signal peptide from residues 1-27 that overlaps the first alpha2-macroglobulin domain. One TSP1-like domain, also found in serine proteases 1 and 4 (Fig. 5.11), is predicted between residues 1478-1531 at the C-terminus of the protein (Fig. 5.8). Pfam (Bateman *et al.*, 2002) identifies a pentraxin domain between residues 1577-1770 that is also located by SMART (Schultz *et al.*, 1998) from 1580-1774 (Fig. 5.8). Although both these programmes predict a pentraxin domain in the same area, the significance level for both matches is at the threshold level. As both Pfam (Bateman *et al.*, 2002) and SMART (Schultz *et al.*, 1998) show the same finding and as this domain does not overlap the GPI-modification site, more confidence in the presence of this domain allowed.

This sequence does not contain a post-translational processing signal to split the protein into alpha and beta chains as displayed by all C3 sequences (Nakao *et al.*, 2000). This confirms that thioll is a one-chain molecule, as is alpha2-macroglobulin. A functioning thiolester motif is present from residues 963-967 (GCGEQ).

Alignment of thiol1 with similar proteins identified by BLAST, shows common regions between all these proteins (Appendix 12). The domain structure of the GPI-linked proteins CD109 and mouse GPI as well as both TEP1 and TEP2, *L. polyphemus* alpha2-macroglobulin and rat alpha1 inhibitor III is almost exactly the same as thiol1 from *C. intestinalis*, the only difference being in the extra C-terminus TSP1-like and pentraxin-like domains after the second alpha2-macroglobulin domain in thiol1 (Appendix 12). This produces an overall identity for the multiple alignment between these proteins of 33.81 % (Appendix 12). Comparison of the specificity defining residues of these proteins in comparison to those outlined by Dodds and Law (1998) (Fig. 5.16) reveals that thiol1 and human CD109 are more like C3 than alpha2-macroglobulin. This alignment also shows that the *D. melangogaster* thiolester proteins (TEPs) lack a double glycine position at the end of this specificity-defining motif. The catalytic histidine responsible for catalysis of the cleavage and binding reaction for the thiolester C3 (Law and Dodds, 1990; 1996) is also present in *C. intestinalis* thiol1 (position 1080), CD109, mouse GPI, TEP1 and the hypothetical protein from *C. elegans* within this motif (Fig. 5.9). None of the most similar proteins have over 27 % identity to thiol1 from this study, so homology between any of the most similar proteins and thiol1 cannot be inferred. The highest identity is with human CD109 at 26.72 %. The lowest identity was with *D. melangogaster* TEP1 at 23.60 %.

A phylogenetic tree constructed from a multiple alignment of all a range of thiolester containing sequences confirms the BLAST results and the place of thiol1 between C3 and alpha2-macroglobulin (Fig. 5.10). This alignment shows that, when assuming a common ancestor for all these thiolester proteins, C3 and alpha2-macroglobulin species form separate branches (Fig 5.10). Thiol1 appears to be in this group but to have diverged from the alpha2-macroglobulin at a very early point. A distinct group is formed with the insect

TEP proteins from *A. gambiae* and *D. melanogaster* and the GPI anchored thiolester proteins from the mouse and human discussed above (Fig. 5.10).

Figure 5.9 Multiple alignment using CLUSTAL W of thioll1 with other thiolester proteins showing the thiolester region and comparing the specificity defining residues (**BOLD**) as described by Dodds and Day (1992). Asterisks below the sequence indicates positions where all the sequences share the same amino acid residue, two dots indicates conserved amino acid substitutions, one dot indicates semi-conserved amino acid substitutions. **Hum.a2m**; human alpha2-macroglobulin Acc No P01023. **Hum.PZP**; human pregnancy zone protein Acc No X54380. **Lim.a2m**; *Limulus* alpha2-macroglobulin Acc No D83196. **Mus.gpia2m**; *Mus musculus* GPI-linked protein Acc No AY083458. **Hum.Cd109**; human cell surface antigen CD109 Acc No AF410459. **Dros.tep1**; *Drosophila melangogaster* thiolester protein 1 Acc No AJ269538. **Dros.tep2**; *Drosophila melangogaster* thiolester containing protein 2 Acc No AJ269539. **Dros.tep3**; *Drosophila melangogaster* thiolester containing protein 2 Acc No AJ269540. **Dros.tep4**; *Drosophila melangogaster* thiolester containing protein 2 Acc No AJ269541. **Anga.Tep**; *Anopheles gambiae* thiolester containing protein Acc No AF291654. **Amphi.C3**; *Amphioxus* C3 Acc No AB050668. **Lamp.C3**; *Lampetra japonica* C3 Acc No D10087. **Hum.C3**; human C3 Acc No P01024. **SP.C3**; *Strongloccentrotus purpuratus* C3 Acc No AF025526. **As.C3**; *Halocynthia rorezi* C3 Acc No AB006964.

	10	20	30	40	50	60
Hum.A2m	GCGEQNMVLFAPNIYVLDYLN	NETQQLTPEVKSK	---	AIGYLNTGYQRQLNYKHYD	---	GS
Hum.PZP	GCGEQNMVLFAPNIYVLDYLN	NETQQLTQEI	KAK	---	AVGYLITGYQRQLNYKHQD	---
Lim.A2m	GCGEQNMVLFAPNIYVLDYLN	TATGSI	TDSIKEK	---	ALNNMRKGYARQQNYRHPD	---
Mus.GPIA2m	GCGEQNMIFAPNIYILDYLT	TKKQLTVNLKEK	---	ALSYMRQGYQRELLYQRED	---	GS
Hum.CD109	GCGEQNMIFAPNIYILDYLT	TKKQLTDNLKEK	---	ALSFMRQGYQRELLYQRED	---	GS
Thioll1	GCGEQNMVLFAPDVFVTLYL	LHSAGKLDAA	TRAK	---	AFKHFQGTGYSNELNYKHRD	---
Dros.Tep4	GCGEQTMVNFVLPNYLVRDYL	LKSIK	KLTPALDTR	---	IKRNLQDGYQHMLHYRHDD	---
Dros.Tep3	GCGEQTMVNFVLPNLI	LVRLRGLRQLTPEVELR	---	ATNNLAIGYQRILYYRHEN	---	GA
Dros.Tep2	GCGEQNMVNFVLPNLI	LVKYLEVTGRKLP	SVESK	---	ARKFLEIGYQRELYKHDD	---
Dros.Tep1	GCGEQNMVNFVPSILALS	YLKAKNRQDQEI	ENK	---	AKRYVETGYQIELNYKRND	---
Anga.Tep	GCGEQNMVNFVLPNLI	LDYLYATGSKEQHLIDK	---	ATNLLRQGYQNMRYRQTD	---	GS
Amphi.C3	GCGEQTMIKLAPNVYVLS	YLHCTDQITKDV	EELK	---	AYDFIRQGYNKQLSHRRPE	---
Lamp.C3	GCGEQNMIKMAPTTLT	LIYLDVSVQEW	EKIGLHRRE	EAIGFLKQGY	SRELSYRKAD	---
Hum.C3	GCGEQNMIGMTP	TVIAVHYLDETEQ	WEKFGLEKRQ	GALELIKKGYT	QQLAFRQPS	---
Sp.C3	GCGEQTMIIYLAPT	LFVYQYLI	AVGSDTAEQ	EAR	---	YDYIADGVARELTYRQDN
As.C3	GCGEQNMIRIAPVV	YIHA	YRSNLEAFTV	TDAQR	---	AQTLKYIQDGYAHELEYKTQVPQGWA
Homology	*****	.*	.*	*	:	. * .:
	70	80	90	100	110	120
Hum.A2m	YSTFGERYG	-----RNQGNTWLTAFV	LKTF	FAQARAY	--	IFIDEAHITQALIWL
Hum.PZP	YSTFGERYG	-----RNQGNTWLTAFV	LKTF	FAQARSY	--	IFIDEAHITQSLTWL
Lim.A2m	YSAFGNRD	-----KQGNLFLTAFV	YRSFAQ	AERF	--	ILINKNKLNETENWILNR
Mus.GPIA2m	FSAFGDID	-----SSGSTWLSAFV	LRCFLEAD	Y	--	IDIDQDVLHRTYTWLNAH
Hum.CD109	FSAFGNYD	-----PSGSTWLSAFV	LRCFLEAD	PY	--	IDIDQNVLHRTYTWLKG
Thioll1	FSAFGEGD	-----ASGSTWLTAF	AAKCFMFA	REL	RPTLV	SASVIDQALTF
Dros.Tep4	FSSF	GPTKWRQEDPVR	NGSTWLTAYV	LR	SFSKI	KDI
Dros.Tep3	FSAFG	-----LDIKRS	-STWLTAY	VARSLRQA	APF	--
Dros.Tep2	YSAFG	-----KSDASG	-STWLTAY	VMR	SFHQAGTY	--
Dros.Tep1	FSAWG	-----QHDALG	-STWLTAY	VIR	SFHQA	AKY
Anga.Tep	FGVWEK	-----SGSSVFLTAF	VATSMQ	TASKY	-MNDIDA	AMVEKALDWLASKQHSS
Amphi.C3	FVWGQNN	-----RYP	CSTWLTAFV	NKVF	CQAKKF	-VTSIDEEAVCKATEWLL
Lamp.C3	YAAFIKR	-----PSSTWLTAF	VV	KVYSLAKR	VII	-VDNQELCGPV
Hum.C3	FAAFVKR	-----APSTWLTAY	VV	KVFS	LAVNLIA	-IDSQVLCGAVKWLILEKQKPD
Sp.C3	YAAWKHR	-----PGSTWLTAY	VV	KVFS	QANR	FTR
As.C3	FAVWANN	-----PPSTWLTNG	FVSR	V	FASARKY	WPGMEVDRICQSV
Homology	:: :	.	.*	:	:

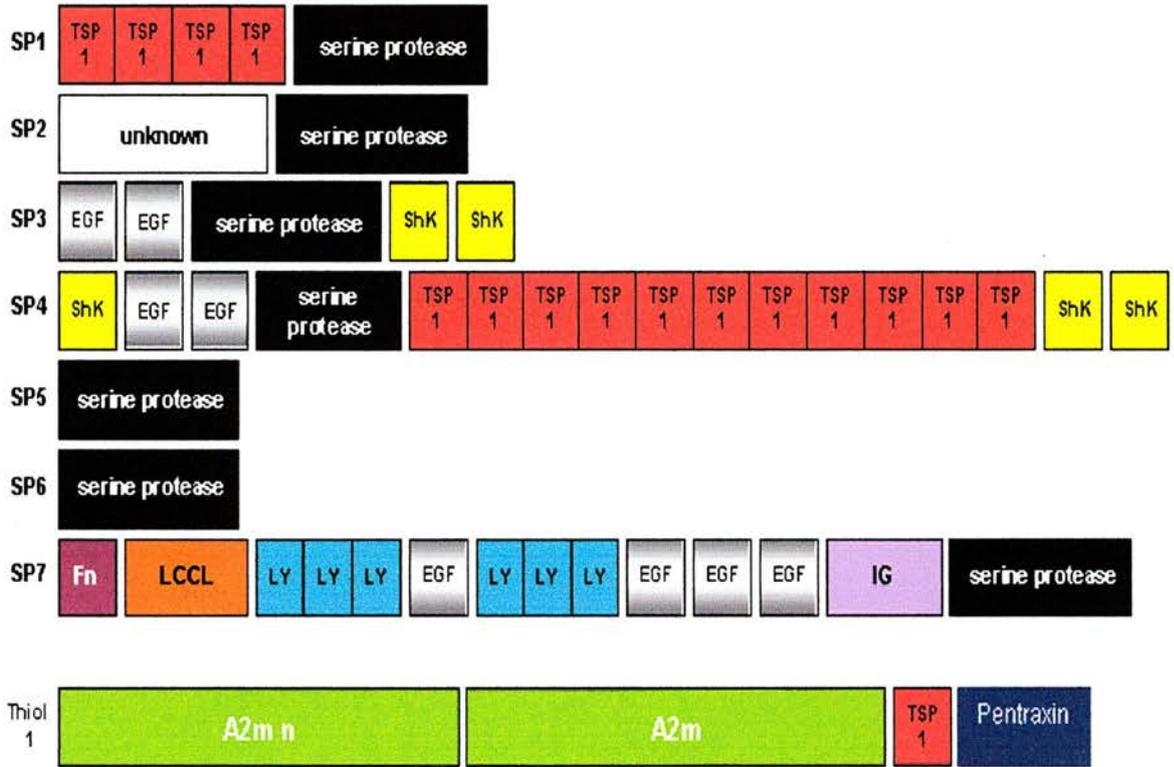
Hum.A2m	GCFRSSGSLLNNAIKGG
Hum.PZP	GCFRSSGSLLNNAIKGG
Lim.A2m	GCFRKIGKLFNSALKGG
Mus.GPIA2m	GEFWEPGRVIHSELQGG
Hum.CD109	GEFWDPGRVIHSELQGG
Thio11	GTFREPGRVSHKAMQGG
Dros.Tep4	GSFTEHGEYFYSSQRSL
Dros.Tep3	GGFEERGDVFERFGDDG
Dros.Tep2	GEFPEVGKLFNANQNP
Dros.Tep1	GKFKELG MVIHNSHGSP
Anga.Tep	GRFDETGKVWHKDMQGG
Amphi.C3	GAFKEVYKVH HREMTGG
Lamp.C3	GSYREDGPVIHREM QGG
Hum.C3	GVFQEDAPVIHQEMIGG
Sp.C3	GAFQESQQVIHQEMIGA
As.C3	GHFDEDDPVH HKEMDGQ
Homology	* :. .

Figure 5.10 Rooted phylogenetic tree produced by DNAMAN from a multiple alignment using the BLOSUM scoring matrix. An alphabetical legend of each branch name and accession number is given on the next page.



Legend	Species and Protein	Accession number
Amphi C3	<i>Branchiostoma belcheri</i> C3	AB050668
Anga TEP	<i>Anopheles gambiae</i> thiolester protein	AF291654
Carp A2m1	<i>Cyprinus carpio</i> alpha2-macroglobulin-1	AB026128
Carp A2m2	<i>Cyprinus carpio</i> alpha2-macroglobulin-2	AB026129
Carp A2m3	<i>Cyprinus carpio</i> alpha2-macroglobulin-3	AB026130
Carp C3-H1	<i>Cyprinus carpio</i> C3-H1	AB016210
Carp C3-H2	<i>Cyprinus carpio</i> C3-H2	AB016212
Carp C3-Q1	<i>Cyprinus carpio</i> C3-Q1	AB016214
Carp C3-Q2	<i>Cyprinus carpio</i> C3-Q2	AB016215
Carp C3-S	<i>Cyprinus carpio</i> C3-S	AB016213
Carp C4A	<i>Cyprinus carpio</i> C4A	AB037278
Carp C4B	<i>Cyprinus carpio</i> C4B	AB037279
Cd109 h.s.	<i>Homo sapiens</i> CD109	AF410459
Chicken C3	<i>Gallus gallus</i> C3	U16848
ChickenC4	<i>Gallus gallus</i> C4	AL023516
Chicken Ovo	<i>Gallus gallus</i> ovomacroglobulin	X78801
CiC31	<i>Ciona intestinalis</i> C3-1	AJ320542
CiC32	<i>Ciona intestinalis</i> C3-2	AJ320543
Cob C3	<i>Naja naja</i> C3	L02365
Drme TEP1	<i>Drosophila melanogaster</i> thiolester protein-1	AJ269538
Drme TEP2	<i>Drosophila melanogaster</i> thiolester protein-2	AJ269539
Drme TEP3	<i>Drosophila melanogaster</i> thiolester protein-3	AJ269540
Drme TEP4	<i>Drosophila melanogaster</i> thiolester protein-4	AJ269541
Gpia2m mouse	<i>Mus musculus</i> GPI anchored alpha2-macroglobulin	AY083458
Guinea A2m	<i>Cavia porcellus</i> alpha2-macroglobulin	D84338
Guinea C3	<i>Cavia porcellus</i> C3	M34054
Guinea Murino	<i>Cavia porcellus</i> Murinoglobulin	D84339
Halo C3	<i>Halocynthia rorezi</i> C3	AB006964
Human A2m	<i>Homo sapiens</i> alpha2-macroglobulin	P01023
Human C3	<i>Homo sapiens</i> C3	NM_000064
Human C4	<i>Homo sapiens</i> C4	AF019413
Human C5	<i>Homo sapiens</i> C5	M57729
Human PZP	<i>Homo sapiens</i> pregnancy zone protein	X54380
Lamp A2m	<i>Lampetra japonica</i> alpha2-macroglobulin	D13567
Lamp C3	<i>Lampetra japonica</i> C3	D10087
Limulus A2m	<i>Limulus sp.</i> Alpha2-macroglobulin	D83196
Medaka C31	<i>Oryzias latipes</i> C3-1	AB025575
Medaka C32	<i>Oryzias latipes</i> C3-2	AB025576
Medaka C4	<i>Oryzias latipes</i> C4	AB025577
Mouse A2m	<i>Mus musculus</i> alpha2-macroglobulin	M93264
Mouse C3	<i>Mus musculus</i> C3	K02782
Mouse C4	<i>Mus musculus</i> C4	AF049850
Mus murino1	<i>Mus musculus</i> murinoglobulin-1	P28665
Mus murino2	<i>Mus musculus</i> murinoglobulin-2	P28666
Paralich C3	<i>Paralichthys olivaceus</i> C3	AB021653
Rat A1I3	<i>Rattus norvegicus</i> alpha1-inhibitor 3	J03552
Rat A1m	<i>Rattus norvegicus</i> alpha1-macroglobulin	M84000; J05359
Rat A2m	<i>Rattus norvegicus</i> alpha2-macroglobulin	J02635
Rat C3	<i>Rattus norvegicus</i> C3	X52477
StpuC3	<i>Strongylocentrotus purpuratus</i> C3	AF025526
Thiol1	<i>Ciona intestinalis</i> C3 (thiol1 from this study)	AJ431688
Trout C3-1	<i>Oncorhynchus mykiss</i> C3-1	L24433
Trout C3-3	<i>Oncorhynchus mykiss</i> C3-2	U61753; AF271079
Trout C3-4	<i>Oncorhynchus mykiss</i> C3-4	AF271080
Xenopus C4	<i>Xenopus laevis</i> C4	D78003

Figure 5.11 Schematic representation of the domain structure of the molecules isolated in this study from *C. intestinalis*. TSP1; thrombospondin type 1 domain, EGF; epidermal growth factor domain, ShK; ShK toxin domain, Fn; fibronectin domain, LCCL; LCCL domain, LY; low density lipoprotein receptor type B domain, Ig; immunoglobulin domain, A2mn; alpha2-macroglobulin N-terminus domain, A2m; alpha2-macroglobulin domain.



5.4 Discussion

5.4.1 Serine Proteases

All the complete cDNA sequences isolated in Chapter 4 from the RT-PCR fragments in Chapter 3 and amplified using degenerate primers designed against motifs of conserved amino acids in the serine protease domain, were indeed all trypsin serine proteases of the chymotrypsin superfamily. The motifs that the degenerate primers exploited are not present in digestive serine proteases. Correspondingly, the serine proteases discovered in this study do not appear to be involved in digestion.

All of the serine proteases involved in the three known complement activation pathways (Bf, MASP, C1r, C1s and C2) also belong to the trypsin family. Several other proteins are known to belong to this family including many blood coagulation factors, cytotoxic cell proteases, trypsins and chymotrypsins, kallikreins, elastases and many others. A comprehensive list is available from Prosite (<http://ca.expasy.org/cgi-bin/prosite-search-de?search=PDOC00124>).

C. intestinalis and other ascidians are known to have a clotting mechanism (Shishikura *et al.*, 1997), inflammatory-like reaction (Parrinello and Patricolo, 1984; Parrinello *et al.*, 1984; Di Bella and De Leo, 2000), as well as phagocytic and cytotoxic mechanisms (Smith and Peddie, 1992; Peddie and Smith, 1993; 1994a; 1994b). Serine protease inhibitors are also known to inhibit phagocytosis (Smith and Peddie, 1992). Apart from the serine proteases of a theoretical complement system it is clear that several other

molecules involved in immune pathways have similar sequence motifs, especially in their serine protease domain.

5.4.1.1 Serine Protease 1 (SP1)

The structure of this protein appears to be unique. Although several proteins have been identified which have TSP1 repeats no serine protease has a similar structure with four repeats at the N-terminus. TSP1 domains have been found in complement proteins, including properdin and several terminal components, as well as extra-cellular matrix proteins (http://smart.embl-heidelberg.de/smart/do_annotation.pl?ACC=SM00209). The function of the TSP1 domain is thought to be cell-cell interaction binding both protein and glycoconjugate ligands and inhibition of angiogenesis apoptosis (<http://ca.expasy.org/cgi-bin/nicedoc.pl?PDOC50092>).

SP1 does not appear to be involved in a complement system for several reasons. No serine proteases of the innate activation pathways in other species are known to contain a TSP1 domain and all have several other domains that SP1 is lacking. As similarity searching indicates that this protein sequence is most like coagulation proteins and TSP1 domains are contained in several serine proteases of coagulation pathways, the most probable function of SP1 is involvement in coagulation.

5.4.1.2 Serine Protease 2 (SP2)

Although no domain can be identified as present in the C-terminus of SP2 the BLAST similarity searching with just the 5' end before the serine protease domain shows that

this protein could be involved in coagulation. However, alignment of the serine protease domain alone does not provide any conclusive answers as many serine proteases with varying function have this domain. Without any domain identified before the serine protease domain, sequences with a similar profile cannot be compared. Although the function of this protein is likely to be in coagulation, the sequence alone does not provide enough evidence to confidently say this.

5.4.1.3 Serine Protease 3 (SP3)

SP3 has a domain organisation that is unique. Epidermal growth factor, type 1, type 2 and calcium binding domains are found in a large number of mainly animal proteins including coagulation factors, blastula proteins and developmental proteins (<http://ca.expasy.org/cgi-bin/get-prodoc-entry?PDOC00021>) (<http://ca.expasy.org/cgi-bin/nicedoc.pl?PDOC00913>). EGF-1 and EGF-2 like domains are found in several complement components, including MASP, C1r, C1s and several of the terminal components. Calcium binding regions are also found in these complement factors, although not in the terminal components, as well as several calcium dependent proteins. The presence of an EGF calcium binding-like domain in SP3 indicates a possible reliance on calcium for a biological function.

The two Shk toxin domains at the C-terminus of this protein are a novel feature of a serine protease. ShK domains are cysteine-rich but clearly distinguishable from other cysteine-rich domains, such as EGF domains, by a characteristic three of six cysteines in the last eight residues. This domain has been identified in metridin (accession number P11495), a 36 amino acid peptide toxin from the sea anemone, *M. senile*, that targets

potassium channels. More recently this domain has also been found in astacin-like metalloproteinase from the jellyfish, *Podocoryne carnea*, (accession number O62558) that is involved in both development and digestion. Several hypothetical proteins from *C. elegans* are also predicted to have this domain (http://smart.embl-heidelberg.de/smart/do_annotation.pl?DOMAIN=ShKT&BLAST=DUMMY). Due to the presence of this domain in such a range of species and proteins it is unlikely that it has a primarily toxic role.

A functional assignment is very difficult for this protein sequence. No other database entry has a similar structure and the ShK domain has not been characterised at any depth. It would appear that SP3 requires calcium for its biological function but that function remains unclear. As metridin acts on potassium channels SP3 may be involved in ionic regulation and as Astacin-like metalloproteinase has a role in development and feeding, the function of SP3 may be similar (Pan *et al.*, 1998).

5.4.1.4 Serine Protease 4 (SP4)

The unique domain structure and multiple repeats of the TSP1 domains in SP4 causes poor alignments and similarity values to be exaggerated in BLAST searching. No serine protease on the sequence database has as many multiple TSP1 repeats after the serine protease domain at the C-terminus that SP4 contains. Additionally, no protein sequence on the databases has eleven TSP1 repeats.

Serine protease 4 shows some similarity to SP3. SP4 has two ShK-like domains at the C-terminus, as does SP3, but has an additional ShK-like domain at the extreme N-

terminus. Two EGF-like domains before the serine protease domain are also a shared feature between SP4 and SP3.

Again, due to the unique structure of this protein it is very difficult to assign a function using only bioinformatic tools. However, several factors allow for speculation as to the function of this protein. The nature of the proteins in which TSP1 repeats are found in (<http://ca.expasy.org/cgi-bin/nicedoc.pl?PDOC50092>), the likely ability of SP4 to bind calcium, the serine protease domain lacking the motif associated with digestive enzymes and the similarity to SP3 all indicate that this protein may be involved in coagulation. Similarity searching reveals that fibulin-6 and hemicentin are both significantly similar. Fibulin-6 has six TSP1 repeats that provide a significant level of similarity with SP4 and also has some EGF domains at the C-terminus. Hemicentin also has EGF domains at the C-terminus, but does not contain any TSP1 repeats. Neither fibulin-6 nor hemicentin are serine proteases. As the identity level between these sequences and SP4 is too low to infer any homology, the cell differentiating function of these proteins cannot be attributed to SP4 from *C. intestinalis*.

5.4.1.5 Serine Protease 5 (SP5)

The fact that the most similar proteins determined by BLAST are fragments is a consequence of having only a serine protease domain. Unusually, for a serine protease containing this kind of domain, it has a very short sequence so very little can be determined about any possible function. Notwithstanding, the significant level of identity between mouse distal intestinal serine protease and SP5 indicates homology. The high level of identity with the fragments of trypsin and trypsinogen from the shrimp,

L. vannamei and the teleost, *T. rubripes*, adds weight to the notion that SP5 may be involved in digestion, although it lacks the motif associated with the digestive serine proteases.

5.4.1.6 Serine Protease 6 (SP6)

As with SP5, SP6 has only a serine protease domain. This makes alignments and BLAST searches of limited value. The serine protease domain stretches the entire length of the sequence and appears to be truncated, as indicated by the differences between the RT-PCR results in Chapter 3 and the RACE results for SP6 in Chapter 4. A possible explanation for this is a failure in the original PCR to produce a true fragment of a transcribed gene. The RACE results indicate that the truncated form of SP6 is present in the cDNA while the non-truncated is not. Again, is impossible to determine a possible function for this gene and indeed it may be a pseudo gene.

5.4.1.7 Serine Protease 7 (SP7)

None of the most similar sequences identified by BLAST show similar overall domain structure. However, it is clear that these proteins do contain homologous domains. The identity levels are highest for proteins that contain several LY repeats, but neither the low-density lipoprotein receptor (LDLR) from mouse nor MEGF7 from human are serine proteases. Indeed, both these proteins are receptors and have transmembrane regions at their C-terminus. Carp MASP has LY repeats and an epidermal growth factor and is consequently the serine protease that shows the highest level of identity. SP7

does not share the same domain structure as any MASP molecule and is lacking several key domains associated with complement function.

Alignment shows little identity between SP7 and the most similar sequences. However, none of these proteins has a similar domain structure. One region of SP7 aligns well with mouse alpha2-macroglobulin receptor but less well with LDLR from mouse, indicating there may be some functional similarity of SP7 to alpha2-macroglobulin receptor. Low-density lipoprotein receptors are aligned to SP7 because of the LY repeats that are thought to form a beta-propeller structure (http://smart.embl-heidelberg.de/smart/do_annotation.pl?DOMAIN=LY&BLAST=DUMMY). However, none of these receptors have a serine protease domain as they have transmembrane regions at their C-terminus.

Importantly, serine protease 7 contains domains that are known to play a role in immune functions, although it has a unique structure. The fibronectin-2 domain was first characterised from the plasma protein fibronectin that binds cell surfaces and various compounds (Balian *et al.*, 1980). The fibronectin-2 domain repeat is associated with the collagen-binding region of fibronectin but one copy has been found in several other proteins (<http://ca.expasy.org/cgi-bin/nicedoc.pl?PDOC00022>). These include blood coagulation factor XII, mannose-6-phosphate receptor, mannose receptor of macrophages and hepatocyte growth factor (<http://ca.expasy.org/cgi-bin/prosite-search-de?search=pdoc00022>). The LCCL domain has been found in a range of metazoan proteins in association with complement B-type domains, such as CUB domains, von Willebrand type-A domains and C-type lectin domains (<http://ca.expasy.org/cgi-bin/prosite-search-de?search=pdoc50820>). *L. polyphemus* factor C contains one copy of

this domain and is a trypsin-like serine protease that is endotoxin LPS-sensitive involved in recognising LPS from bacteria (Muta *et al.*, 1991). This has led to proposals that the LCCL domain is involved in LPS binding of bacterial pathogens (Trexler *et al.*, 2000).

Two forms of low-density lipoprotein repeats (A and B) are usually found in low-density lipoprotein receptors and these molecules are the major cholesterol carrying proteins of plasma (<http://ca.expasy.org/cgi-bin/get-prodoc-entry?PDOC00929>). However, with *C. intestinalis*, SP7 only the type B domain is present. The predicted structure of LDLR repeats is a beta-propeller type structure; the same structure is also predicted for scavenger receptors (Springer, 1998). This structure relies on six LY repeats being in close proximity bound together with EGF domains (Springer, 1998). This is the precise structure of the LY region in *C. intestinalis* SP7. This highly structured domain brings other modules of the protein close together. In the protein, nidogen, the LDLR domain functions to bind laminin (Springer, 1998). Laminin is a membrane protein that provides a potent source of adhesion (<http://www.mblab.gla.ac.uk/~julian/dict2.cgi?3545>).

Immunoglobulin domains spanning approximately 100 amino acids have been found in many proteins, including all the constant chains of immunoglobulin molecules, and this region is thought to be involved in protein/protein interaction (Cushley and Owen, 1983). As this domain is also found in a range of proteins (Cushley and Owen, 1983) it is difficult to elucidate the exact function it may have in each molecule.

The complex domain structure of *C. intestinalis* SP7 provides information about its function. The majority of the domains present in SP7 seem to be involved in binding

but, as SP7 is a serine protease, it cannot be a cell surface receptor as many other proteins are that share homologous domains. Because SP7 has regions that may bind both bacterial surfaces and host surfaces as well as an Ig domain that regulates protein/protein interactions, it is possible that its function lies in the immune system and the recognition of pathogens.

5.4.2 Thiolester 1

Levels of identity between the most similar thiolester containing proteins are low because the majority of these proteins have only two main domains that span the majority of the sequence. The amino acids in this domain are not necessarily conserved but the important motifs within them are. Alignment of thiol1 with the C3 sequences from the invertebrates *S. purpuratus* (accession number AF025526) and *H. rorezi* (accession number AB006964) with *Amphioxus amphioxus* (accession number AB050668) and the lamprey, *Lampetra japonica* (accession number D10087) shows similar identity levels of between 21 and 25 %. Similarity searching is consequently of limited use when dealing with a novel protein sequence from this family. Important motifs have very subtle differences and small domain differences characterise the different functional groups in this family.

It further appears that *C. intestinalis* thiol1, although expressed constitutively in the present study, is up-regulated after immune challenge, as revealed by several EST clones of thiol1 isolated from both the larvae and adult that are fragments of thiol1 (accession numbers AV678703, AV947947, AV955787, AV850887, AL666319, AV837571, AV837399, BP018415).

Thiolester 1 is a one-chain structure and therefore resembles alpha2-macroglobulin rather than C3 sequences. However, as shown in the present study, several motifs within thiol1 more closely resemble C3. The substrate defining residues detailed by Dodds and Day (1993) (Fig. 5.9) show that *C. intestinalis* thiol1 contains a catalytic histidine and shares residues in common with the C3 sequences that alpha2-macroglobulin and pregnancy zone proteins do not. A histidine at the catalytic site is a key factor in the evolution of the complement system (Dodds and Day, 1993). Polymeric forms of alpha2-macroglobulin must have arisen from the monomeric forms, whose likely function was to transport substrate molecules in to the cell using pinocytic or phagocytic methods (Dodds and Day, 1993). In the process, these molecules acquired the ability to trap proteins by conformational change (Dodds and Day, 1993). The catalytic histidine residue seems to have arisen independently, allowing the polymeric molecule to bind to hydroxyl groups, with an explosive reaction confined in time and space by its increased reaction with water (Dodds and Day, 1993). An alpha2-macroglobulin molecule with this reactive histidine, such as *C. intestinalis* thiol1 in this study, could have been the ancestor to C3. Murinoglobulin, related to alpha2-macroglobulins, from mouse (accession number P28665) and guinea pig (accession number D84339) also have a histidine at the catalytic position but are found in few species and are therefore likely to have evolved separately from C3 sequences (Dodds and Day, 1993).

All alpha2-macroglobulin sequences have two alpha2-macroglobulin domains (<http://ca.expasy.org/cgi-bin/nicedoc.pl?PDOC00440>) spanning the entire length of the protein. All C3 sequences also contain these domains but have two additional domains. The first is an anaphylatoxin domain between the two large alpha2-macroglobulin

domains (<http://www.expasy.ch/cgi-bin/nicedoc.pl?PDOC00906>) that mediates a local inflammatory response, and the second is a C345C domain, a subfamily of the netrins which are thought to be involved in protein migration located at the C-terminus (<http://www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF01759>). *C. intestinalis* thioll lacks an anaphylatoxin domain, but as this protein is a one-chain structure that is not proteolytically cleaved it could not function in the same way a conventional C3. The extra domains predicted by both SMART and Pfam at the C-terminus of *C. intestinalis* thioll make this protein unique among the alpha2-macroglobulin one-chain proteins. Only C3 sequences have an extra domain after the final alpha2-macroglobulin domain. The predicted TSP1 domain in *C. intestinalis* thioll, as discussed for the serine proteases SP1 and SP2 from *C. intestinalis* above, is present in several terminal components of the complement pathway, properdin and several others proteins (<http://ca.expasy.org/cgi-bin/nicedoc.pl?PDOC50092>) but is commonly found in repeats of two or more. A single TSP1 domain is more unusual but is still thought to be involved in cell-to-cell interaction. This domain spans approximately half the residues of the C345C domain in C3 sequences.

The pentraxin domain predicted at the extreme C-terminus of *C. intestinalis* thioll is below the threshold value for Pfam and SMART domain searches, but manual observations confirm that a domain corresponding to the pentraxin profile is present. Pentraxins are a group of acute phase proteins including C-reactive protein that have a very different structure to thiolester1 (<http://ca.expasy.org/cgi-bin/nicedoc.pl?PDOC00261>). C-reactive proteins activate complement through calcium mediated binding to phosphorylcholine, an antigen determinant on eukaryotic membranes and several bacterial pathogens (Gould and Weiser, 2001), and enhance

phagocytosis. The horseshoe crab (*L. polyphemus*) expresses C-reactive proteins constitutively in the haemolymph (Muta *et al.*, 1991). A C-terminal pentraxin domain is also present in a few other proteins but its function is unknown (<http://ca.expasy.org/cgi-bin/nicedoc.pl?PDOC00261>). Electron microscopy has revealed that the pentraxin domain forms a discoid pentameric structure and is thought to demonstrate calcium-mediated ligand binding (Pepys and Baltz, 1983; Gewurz *et al.*, 1995).

It is still unclear if thiol1 from *C. intestinalis* is a GPI-anchored protein. Since the available software is unreliable in this case, manual observations show that there are some characteristic GPI primary structures at the C-terminus. However, there are fewer leader amino acids before the cleavage site and hydrophobic tail than would be expected. Functional studies are the only way to determine conclusively if thiol1 is GPI-anchored.

The phylogenetic tree confirms that thiol1 has similarities to C3 but has more identity to alpha2-macroglobulins (Fig. 5.10). The distinct subgroup that this protein forms a part of appears to have very ancient origins and may be close to the ancestral form that C3 and alpha2-macroglobulin diverged from. Several of the proteins within the same group as thiol1 (Fig. 5.10) have either substrate specific residues similar to C3 (Fig. 5.9) or have been shown to be opsonic or have a role in immunity (dos Remedies *et al.*, 1999; Levashina *et al.*, 2001). This is not the case for the majority of the alpha2-macroglobulins, several of which lack a catalytic histidine (Fig. 5.9) and have been shown not to be opsonic (Armstrong *et al.*, 1998).

Thiol 1 appears to have a larger role in the innate immune system of *C. intestinalis* than an alpha2-macroglobulin molecule. The catalytic histidine, the C3-like substrate

specificity and the presence of both TSP1 and pentraxin domains at the C-terminus provide evidence that thio11 has the potential to function as a C3-like molecule. An ancestral one-chain alpha2-macroglobulin/C3-like molecule would not have an anaphylatoxin domain but would be able to catalytically bind to hydroxyl groups bringing about a conformational change in its structure, exposing the thiolester site. As the pentraxin domain can bind pathogen surfaces and is recognised by phagocytes, with the cell-to-cell interaction provided by the TSP1 domain, thio11 could theoretically bind both pathogen and host cells to act as an opsonin. Although function can only be determined through experimentation, this molecule represents an important stage in the evolution of C3 from alpha2-macroglobulin-like proteins.

In conclusion, bioinformatic analyses have provided useful information about the protein sequences isolated in this study. However, the level of information is entirely dependent on the information characterisation of other proteins on the databases. As all the protein sequences isolated in this study are novel in their structure and contain domains of unknown function, it has been impossible in some cases to determine a role for each of the protein sequences. Using the information determined from these analyses has been able to confirm the families that these proteins come from and proteins that share similar function. In combination this evidence has allowed a putative assignment of function for most of the sequences discovered in this study.

Chapter 6

General Discussion

6.1 General Discussion

The overall aim of this study was to provide evidence for a functioning complement system in *Ciona intestinalis* by isolating and characterising mRNA sequences of complement-like genes. Several genes of the same gene families as the serine proteases and thiolester proteins of the complement system were isolated from different tissues in *C. intestinalis*. Several of these have domains and structures that indicate they are involved in the immune system. Several also share domains that are present in complement proteins. However, none of the cDNA sequences isolated in this study have the same domain organisation as the known complement proteins.

A functioning complement system in *C. intestinalis* has yet to be proven. This study has discovered protein sequences that provide significant evidence that *C. intestinalis* has protein pathways thought to be ancestral to complement and has molecules that are likely to have the ability to function as part of an ancestral complement system.

This study has shown that *C. intestinalis* has several novel proteins likely to be involved in clotting pathways and immunity. Several of the sequences found contain unique domain structures and provide evidence about the evolution of several vertebrate protein families. Seven novel serine protease sequences were isolated, two of which appear to be involved in coagulation and one of which is likely to have a role in immunity. Although several domains are shared between these serine proteases and those of the complement system, none have a similar domain structure. The thiolester containing sequence cloned in the present study shares similar domains, domain structure and

motifs within these domains to C3 sequences, representing an important stage in the evolution of C3 from an ancestral alpha2-macroglobulin type molecule.

The serine protease and thiolester sequences discovered here were all isolated from the hepatopancreas RNA, although the larval RNA also contained the thiolester sequence. MASP, Bf and C3 sequences from the Japanese ascidian, *H. rorezi*, were all also discovered from hepatopancreas RNA isolated from wild specimens that had not been artificially exposed to any immune stimulant (Nonaka and Azumi, 1999; Nonaka *et al.*, 1999; Li *et al.*, 2000). The discovery of transcribed immune genes from this organ in *H. rorezi* without LPS challenge illustrates its diverse nature. Being closely associated to the intestine, the hepatopancreas plays a key role in digestion, and the likely presence of a complement gene from non-stimulated specimens.

None of the serine proteases isolated in the present analysis from the hepatopancreas and larvae are homologues of complement proteins. One reason may be the tissue from which the RNA was isolated. The activation of any complement pathway is a very local response that is regulated by many factors to prevent any damage to the host animal (Dodds *et al.*, 1996; Sim and Dodds, 1997). Thus, complement proteins are likely to be expressed in the cells as well as other tissues so they are present at many points of possible infection. Cellular RNA, as shown in Chapter 2 of this study, is transcribed at very low levels making it impossible to use in this instance. This pool of RNA, however, may contain a higher proportion of RNA from immune genes, causing a higher chance of isolating clones from RT-PCR containing sequence fragments from these genes.

A second reason that no complement genes were isolated in *C. intestinalis* during this research is because no new genes were isolated from the hepatopancreas after stimulation with LPS compared to non-LPS stimulated animals. As the primary role of this organ is not in immunity, this again indicates that the up-regulation of immune genes does not occur in the hepatopancreas. Genes may be up-regulated in the cells. The complement homologues from the sea urchin, *S. purpuratus*, (SpC3 and SpBf) have both been amplified from exclusively cellular RNA (Al-Sharif *et al.*, 1997) or a mix of tissue including the blood cells. The expression of SpC3 is also up-regulated in the blood cells of this animal after stimulation with LPS (Clow *et al.*, 2000).

Recently a research group in Italy led by Dr Pinto (Cell Biology, Stazione Zoologica "A. Dohrn", Villa Comunale, Napoli, 80121, Italy) have successfully isolated RNA from the blood cells of *C. intestinalis* (pers. com.). Two novel thiolester sequences that were not identified in this study were amplified from this pool of RNA and have been submitted to the EMBL database (accession numbers AJ320542 (CiC31) and AJ320543 (CiC32)). These proteins both have similar specificity defining residues as thiol1 and C3 (Fig 6.1), but these genes appear to have both a two-chain structure and a C345C domain at the C-terminus (Fig 6.2). Although the proteins have a thiolester, only one (accession number AJ320543) has two alpha2-macroglobulin domains identified by the bioinformatic application used in Chapter 5 to characterise the sequences isolated in this study. Visual comparison with alpha2-macroglobulin domains and its consensus sequence confirms that an alpha2-macroglobulin domain is indeed present but does not completely fit the consensus pattern. This domain organisation of these proteins is more similar to C3 than

alpha2-macroglobulin. From the comparison of the motifs within these domains, including the specificity defining residues, and similarity searching it is likely that both these proteins are ancestral to C3.

The degenerate primers designed in Chapter 3 amplified seven serine proteases and two thiolester gene fragments. Other sequences may have been amplified in this pool of gene fragments but at a lower frequency. A comprehensive sequencing project would enable the serine protease repertoire of the hepatopancreas to be fully ascertained. For this study a limited number of clones could be sequenced and the seven novel serine proteases isolated are only a partial representation of this family of genes. Only two thiolester containing gene fragments were found. One of these was detected in one clone of the hundreds sequenced, and RACE failed to amplify this cDNA sequence. All the other clones sequenced contained another thiolester sequence that was designated thio11. If another protein shares similar motifs to thio11 in the pools of RNA used in this study then it is very rare in comparison.

As *C. intestinalis* is the most ancient species studied so far, it may have complement-like proteins that do not contain all the motifs of the known complement homologues from more recently evolved animals. The major weakness of RT-PCR in this study is the reliance upon sequence information from those complement homologues discovered so far. These have mainly been isolated from vertebrates. Those from invertebrates are from the sea urchin, a deuterostome belonging to a group with an unclear phylogeny. Alignment of these sequences highlights the most important conserved regions (Appendices 3 & 4) but these invertebrate species may have different motifs. *C. intestinalis* may have a more ancestral gene, similar to that which was carried into the

vertebrate lineage. The more developed ascidian and sea urchin species may have evolved genes with characteristics differing from the genes that the early vertebrates possessed. In addition, species within each phylum (Echinodermata and Urochordata) have continued to evolve from the point at which the groups split.

6.1.2 Thiolester Containing Genes

Similarity searching using BLAST (Altschul *et al.*, 1997) reveals the protein with the highest level of similarity is AsC3 from the ascidian *H. rorezi* (Nonaka *et al.*, 1999). The levels of identity to AsC3 of CiC31 and CiC32 are 29.31 % and 26.87 % respectively (Appendix 13). This falls below the 30 % level that is expected for homologous proteins. Identity levels with the other invertebrate C3 homologue, isolated from the sea urchin *S. purpuratus* (Al-Sharif *et al.*, 1997) are lower, with the highest being CiC31 at 22.35 %. Identity levels with thio11 also from *C. intestinalis* are lower still, the highest again being CiC31 at 16.68 % (Appendix 13).

These proteins show key differences within the motifs that the degenerate primers were designed against in Chapter 3. RT-PCR was used to isolate the cDNA sequence of these proteins but benefited from information supplied by the genome of *C. intestinalis* that is currently being sequenced in Japan. It is not clear if this protein is also expressed in the hepatopancreas. These proteins, although representing another key stage in the evolution of C3, do not show a level of similarity to alpha2-macroglobulin, or indeed level of identity to thio11, that provides evidence to the specific ancestor of this protein. The presence of thio11, CiC31 and CiC32 shows that all three proteins are likely to be functioning in the immune system of *C. intestinalis*. It also becomes more likely that a

functioning complement system is present. Several requirements, including the presence of specific serine proteases, have yet to be met to show that the mechanism for a complement system exists in this species.

If CiC31 and CiC32 are part of a functioning complement system, then the fact that two homologous sequences (an identity level of 33.82 % between CiC31 and CiC32) are present introduces speculation about the number of C3-like sequences that may be present in *H. rorezi* and *S. purpuratus*. As *C. intestinalis* is more ancestral than *H. rorezi* and thought to be closer to the vertebrate line than *S. purpuratus*, two C3-like molecules may have been part of the original complement system and one was later lost as the system became more complicated.

Thiol1 from this study has several of the same domains and similar structural organisation to known thiolester containing complement homologues. The differences between thiol1 and C3 genes allows speculation that thiol1 represents an intermediate stage between a more ancient alpha2-macroglobulin and C3 gene. Thiol1 shares more structural similarities with alpha2-macroglobulin because it appears to be a one-chain structure. All C3 sequences are known to be two-chain with the alpha chain being an anaphylatoxin only in the vertebrates. However, as C3 is thought to have evolved from the one chain alpha2-macroglobulin, it is likely that the first protein having C3-like function was also a one-chain structure. As the complement system became more complex through evolution, a secondary immune regulatory function of cleavage products may have evolved once the two-chain structure was present.

The specificity defining residues of thioll from *C. intestinalis* (Fig. 5.16) are more similar to C3 than any of the other thiolester containing proteins, including alpha2-macroglobulins and the thiolester containing proteins (TEPs) from *D. melanogaster* and *A. gambiae*. The TEP protein from *A. gambiae* is a one-chain structure containing a catalytic histidine that allows the binding of hydroxyl groups. This protein has been shown to act as an opsonin (Levashina *et al.*, 2001). Thioll from *C. intestinalis* shares the same structure as the TEP protein from *A. gambiae* as well as several other TEP and alpha2-macroglobulin proteins. Alpha2-macroglobulin sequences do not contain this catalytic histidine residue. By contrast, TEP1 from *D. melanogaster*, and the GPI-anchored thiolester proteins from mouse and human (Fig. 5.16) do have this histidine, but studies of their function have yet to be published.

Another important difference between the structure of alpha2-macroglobulin and C3 molecules is the extra domain at the C-terminus (Fig 6.2). No other alpha2-macroglobulin sequence has the extra domains found in *C. intestinalis* from this study (Fig. 6.1). All the C3 sequences share the same C345C C-terminus domain. Thioll lacks this domain, but has a pentraxin domain associated with acute phase proteins, complement activating proteins and opsonins. An ability to bind calcium through this domain and the associated EGF domain indicates that the activity of this protein may be mediated by calcium ions. The function of the C345C domain is unknown but the C-terminal domain is homologous to the C-terminal domains in other proteins so it is likely that its role is an interaction with metzincins (Banyai and Patthy, 1999). Metzincins are proteins that share a C-terminal netrin-like domain (netrins, secreted frizzled-related proteins and tissue inhibitors of metalloproteases (Banyai and Patthy, 1999) also appear

to have EGF domains located immediately before the netrin-like region. Thioll1 has an EGF domain before the pentraxin-like domain.

It is impossible to compare the function of the C-terminal domain of thioll1 from *C. intestinalis* to the C345C domain. Proteins having a C-terminal C345C domain may have various functions ranging from digestion to development (Banyai and Patthy, 1999), so its specific role in a complement system is unclear. The pentraxin domain has been more fully characterised and it is known to play a key role in immune proteins through calcium mediated ligation and binding. However, this is the only thiolester-containing protein known from any species that has an extra C-terminal domain after the final alpha2-macroglobulin domain other than C3 sequences.

Thioll1 from *C. intestinalis* shows the highest level of similarity to a GPI anchored protein (section 5.3.8). It is of key importance to determine if thioll1 is also a GPI-anchored protein. The mRNA sequence of thioll1 shows features of GPI anchorage but as little is known about the signals that cause this post-translational modification, experimental evidence is required to show this modification exists. The function of this protein will be key to its location. If membrane bound, the protein is basically acting as a receptor to foreign surfaces, mediated through the binding of the thiolester. If this protein is not membrane-bound, it may have an inhibitory and/or an opsonic effect on microorganisms in the plasma. Either scenario has implications for the evolution of C3.

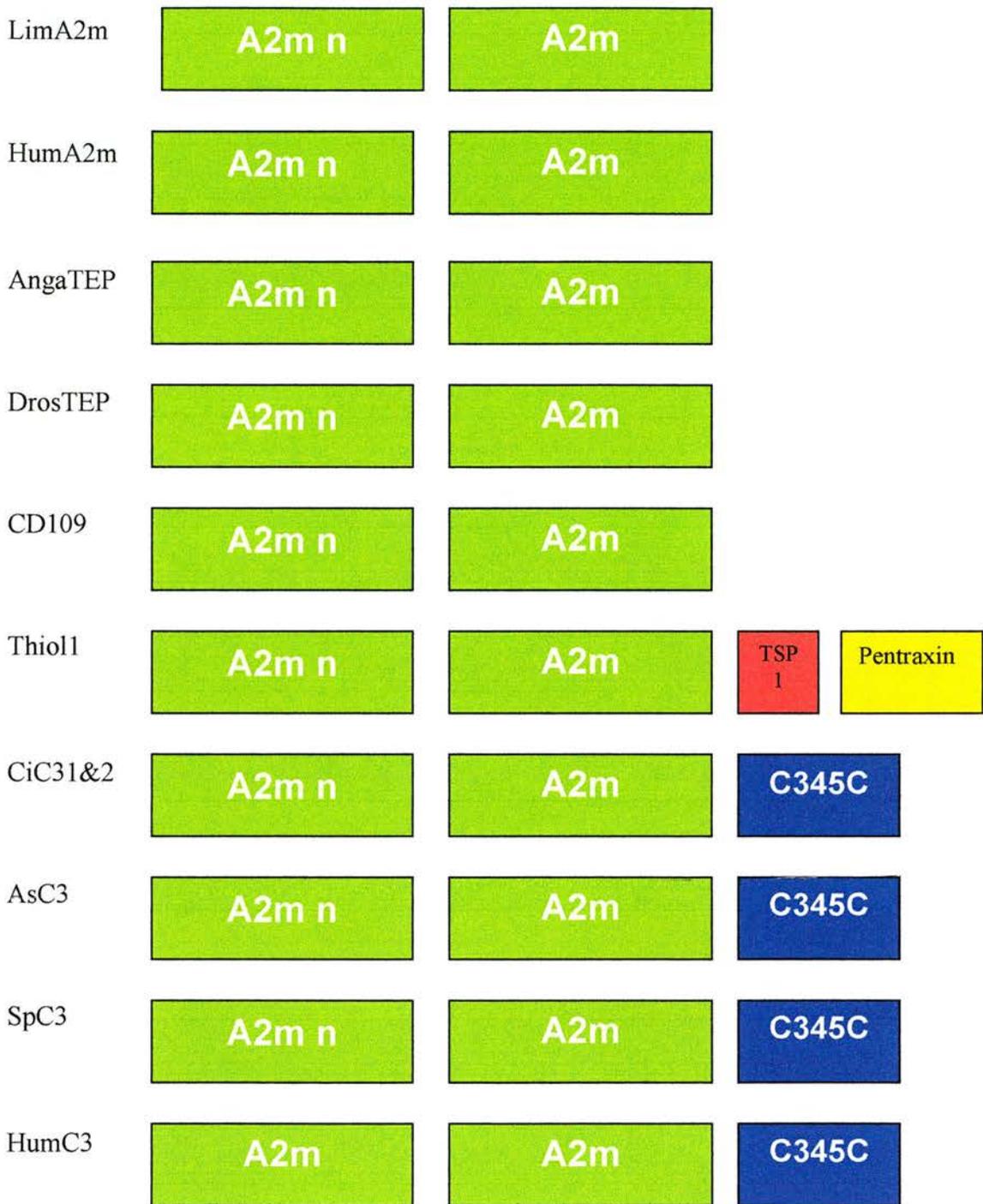
6.1.3 Serine Proteases

It is clear from the bioinformatic analysis of the domain organisation of the seven serine protease sequences in Chapter 5 that none have the structure of known complement genes. SP7 has a repertoire of domains, although unique in their organisation, that can provide the capability for immune function and interaction with a complement system. Domains contained within SP7 that indicate an immune function include the LCCL domain (also found in factor C from *L. polyphemus*). This is an important innate immune protein involved in recognising LPS on bacterial surfaces through the LCCL domain (Trexler *et al.*, 2000). The fibronectin-2 domain is known to be in proteins whose function is to bind to other cell surfaces (Balian *et al.*, 1980). The LDLR repeat structure is known to strongly bind laminin, a membrane protein (Springer, 1998). These domains provide the capacity for the binding of both pathogen surfaces and host cells, and in combination with an immunoglobulin domain thought to be involved in protein/protein interaction, this protein has the capacity to act as an opsonin.

LDLR domains are present in the mannan-associated serine proteases (MASPs) of the complement system and epidermal growth factors (EGF) are present in factor B (Bf) serine proteases of the complement system. Both MASP and Bf have the ability to cleave C3 and bind to other complement factors, as well as pathogen surfaces in complexes. Bf binds to C3 through a von Willebrand domain and a Short Consensus Repeat (SCR) domain. SP7 lacks these specific domains but the LDLR LY domain is also a repeat domain, and is present in alpha2-macroglobulin receptors. This could allow SP7 to bind to an alpha2-macroglobulin-like protein.

The novel proteins found in *C. intestinalis* by the present investigation indicate that protein pathways involved in host defence exist, some of which are thought to have been ancestral to the evolution of the complement system. A novel serine protease also exists that shares domains with serine proteases from the complement system that is likely to have an immune function. A novel thiolester containing protein isolated from both larvae and adult mRNA appears to have both characteristics of alpha2-macroglobulin and C3. Such a molecule will be a key protein involved in host defence and may be have similar functions to both alpha2-macroglobulin and C3. This molecule represents a key evolutionary stage in the evolution of C3. Two C3-like sequences have been isolated from cellular RNA by Dr Pinto (unpublished; accession numbers AJ320542; CiC31, and AJ320543; CiC32). These genes appear to have more characteristics of C3 than alpha2-macroglobulin and present the possibility that several thiolester containing proteins may be involved in host defence and the complement system.

Figure 6.1 Schematic representation of the domain structure of representatives of the thiolester protein family. A2m n is Alpha2-macroglobulin n-terminus domain, A2m is alpha2-macroglobulin domain, TSP1 is thrombospondin type1 domain, Pentraxin is pentraxin domain and C345C is C345C/netrin N-terminal domain.



6.2 Future work

The role of the novel proteins isolated from *C. intestinalis* in the present study can only be determined through functional studies. The serine protease SP7 is the most likely to have a role in immunity although the precise role of this protein cannot be deduced from the sequence data alone. Several methods could be employed to investigate where and when SP7 is produced and what immune mechanisms it affects. Antibodies could be raised against SP7 that could be purified from *C. intestinalis* or recombinantly expressed. This would enable the location of this protein to be determined in the tissues, cells or plasma. Relative protein levels after immune stimulus/suppression could then be measured with Western blotting, a technique that was employed for SpC3 (Clow *et al.*, 2000). Expression levels could also be measured after the same stimulus using quantitative real-time PCR to determine which cell type expresses this protein and how quickly the immune stimulus affects expression.

Immune assays could also be performed using the blood cells from *C. intestinalis in vitro*. The function of SP7 could be removed by specific antibody binding or its expression blocked by dsRNA inhibition. Tests of innate immune function such as phagocytosis, respiratory burst, cytotoxicity and others could provide information into the function of SP7 and consequently which immune pathways it involves. Several of these assays have been performed using the blood cells of *C. intestinalis* (Parrinello *et al.*, 1984; Smith and Peddie, 1992; Jackson and Smith, 1993; Peddie and Smith, 1994a; 1994b; Parrinello *et al.*, 1995; Peddie *et al.*, 1995) proving this as a useful technique to elucidate function.

Thiol1 from *C. intestinalis* contains a thiolester that could be exploited for protein purification. The thiolester site will bind methylamine that can be radioactively labelled. This can then be traced through fractionation with other thiolester proteins. As the precise size of thiol1 is known this protein would be easily identified. As with SP7, several tests of expression, location and immune function could be performed. However, as C3 acts as an opsonin (Nonaka, 2000), phagocytosis studies in which the function of thiol1 is removed would be of key importance. Again, dsRNA inhibition or specific antibody binding could block this phagocytic function enabling the binding of the thiolester to bacterial surfaces to be investigated. Some of these techniques have been successful in characterising the function of other novel thiolester-containing proteins (Levashina *et al.*, 2001).

In addition a determination of the serine protease inhibitory function of thiol1 would allow speculation as to its primary role. If indeed thiol1 has the functions of both alpha2-macroglobulin and C3, it could represent a hybrid ancestral molecule. Alpha2-macroglobulin can be detected using its unique ability to bind proteases. A protease entrapped by alpha2-macroglobulin is prevented from interacting with macromolecular substrates, but some substrates are small enough to diffuse through the alpha2-macroglobulin cage, binding to the active enzymatic site of the protease (Armstrong and Quigley, 1999). Alpha2-macroglobulin can thus be detected as it inhibits the proteolytic, but not the amidolytic, activity of an exogenously added protease.

In addition to these investigations the presence of CiC31 and CiC32 (accession numbers AJ320542 and AJ320543) could be determined in different tissues from *C. intestinalis* by northern blotting or *in situ* hybridisation. As the expression of CiC31 and CiC32

was only determined in the blood cells, it may be necessary to use specimens of *C. intestinalis* from a population not exposed to the high level of pollutants associated with marinas. This may allow for the isolation of high quality RNA from the blood cells not achieved during the present research. If CiC31 and CiC32 are detected in combination with thiol1 from this study, the functional characterisation studies performed on thiol1 could also be undertaken with these proteins that show structural similarities to C3. This could reveal which of these proteins could be involved in a functioning complement pathway.

A clue to the function of thiol1 is its location. Human CD109 is a GPI-anchored membrane bound protein with C3-like reactivity and binding potential (Lin *et al.*, 2002). Sequence comparison of CD109 with both alpha2-macroglobulin and complement thiolester containing proteins leads Lin (2002) to suggest that the function of this protein is a membrane bound cross linking reagent mediating cell-substrate, cell-matrix, or cell-cell interactions that play a role in haematopoiesis or innate immune responses (Lin *et al.*, 2002). Indeed, it has been reported that CD109 may have a role in T-cell-antigen-presenting cell or T-cell-B-cell interactions (Suciu-Foca *et al.*, 1985). Thiol1 contains many similarities to CD109 but until GPI-anchorage is proved or disproved, the functional relationship cannot be ascertained. Proving GPI-anchorage would define the subcellular location of thiol1 and limit its range of possible functions. To test for the GPI-anchor site a solubilisation test can be undertaken involving phospholipase cleavage of the GPI anchor (Eisenhaber *et al.*, 1999).

6.3 Conclusion

This study has shown that the origins of complement are likely to have arisen from a group of animals very close in evolutionary terms to *C. intestinalis*. This has been elucidated from the isolation of novel genes from the same families. Some of these are putatively involved in pathways thought to be ancestral to the complement system. One serine protease (SP7) shares some domains with serine protease of the complement system but has a unique structure. However, the domain organisation of this protein indicates a role in the innate immune system. One thiolester-containing protein was isolated that shares similarities to both alpha2-macroglobulin and C3. The domain organisation and the motifs within these domains provide evidence that this novel member of the thiolester family represents an important stage in the evolution of C3.

Appendices

Appendix 1 Table showing the degeneracy code and the multiple bases they represent.

Degeneracy code	Multiple bases
N	A, C, G, T
V	G, A, C
D	G, A, T
B	G, T, C
H	A, T, C
W	A, T
M	A, C
R	A, G
K	G, T
S	G, C
Y	C, T

Appendix 2 Formulae used to calculate the melting point (T_m) of oligonucleotide primers. GC is GC content and GC% is GC x 100

Up to 14 bases

$$T_m = GC \times \text{sequence length} \times 4 + (1 - GC) \times \text{sequence length} \times 2$$

Over 14 bases

$$T_m = 69.3 + 0.41 \times GC\% - 650 / \text{sequence length}$$

Appendix 3 Multiple alignment of the serine protease sequence regions (BOLD) that were used to design degenerate primers for factor B and MASP in Chapter 3. Asterisks below the sequence indicate positions where all the amino acid sequences share same amino acid residue, two dots indicate conserved amino acid substitutions and one dot indicates semi-conserved amino acid substitution. **AsMASPb** is ascidian MASPb from *Halocynthia rorezi*; accession number BAA19763, **SuBf** is sea urchin Bf from *Strongylocentrotus purpuratus*; accession number AAC79682, **Carp MASPa** from *Cyprinus carpio*; accession number BAA34706, **Pig Trypsin** from *Sus Scrofa*; accession number P00761, **Hu Hepsin** from *Homo sapiens*; accession number P05981, **LampBf** is Bf from the lamprey *Lampetra japonica*, **Carp BfA** is Bf from *Cyprinus carpio*; accession number BAA34706, **XenBf** is Bf from *Xenopus laevis*; accession number BAA06179, **AsMASPa** is MASPa from *Halocynthia rorezi*; accession number BAA19762, **AsBF** is factor B from *Halocynthia rorezi*; accession number AAK00631.

```

                10         20         30         40         50         60
                |         |         |         |         |         |
AsMASPb      ILTAAHCLYNTEYEG--NVRYPNATHAWLGVHNRLE-DRNIAKSQVINAKVESIVLHPQ
SuBf          ILTAAHCFS---GE---NTLSQNGTTVYLGLTHRNVN-DLNRPSVRCGIDYAPGLLQGL
CarpMASPa    VLTAAHVLRSHRRDFSVVPVASEHIRVHLGLTDIRD-KHLATNRS----VAKVILHPQF
PigTrypsin   VVSAAHCYKS-R-----IQVRLGEHNIDV-LEGNEQFIN---AAKIITHPNF
HuHepsin     VLTAAHCFPERNR---VLSR---WRVFAGAVAQAS-PHGLQLGVQA-VVYHGGYLPFR
LampBf       ILTAAHCFDEFAITDDEWWRG--SIDVVISSNKLK-GDKISPQK-IIIHEGYNRNPDA
CarpBfA      ILTAAHCFKE-----GDTHDKITVQLEK-DKPVKVKYVIHPYINLTAKQQ
XenBf        ILTAAHCFDL-----DDKTQKIHVKID--GKEYLVKDFYRHPKYDPI SKKD
AsMASPa      VITAAHCVELRNPS-----DITAWFGVDDRSI-NDNIVQKRD-ILEINIHQDYEN
AsBf         VLTAAHLFDRLKGG---EDNWHESVLVHLGISIKPTSEDDMISSIRMYIPGEIIHPRY
Homology     :::***

                70         80         90         100        110        120
                |         |         |         |         |         |
AsMASPb      YFKESPWDFDFGLIRVSEE-----IKMSNKTRPVCLPQTPNEFD--MVDDG--
SuBf          DGGE---HNDIALLRLDRE-----AELSPFVRTTVCLPPSDPQKVNWYVNP--
CarpMASPa    DPQNYNN--DIALIKLSQE-----VVL SALIQPVCLPPRPGVKGHTLMPLPN--
PigTrypsin   NGNTLDN--DIMLIKLSPP-----ATLNSRVATVSLPR---S--CAAAG----
HuHepsin     DPNSEENSNDIALVHLSSP-----LPLTEYIQPVCLPPAAG---QALVDG----
LampBf       HVQIENLDNDIALIKLSKR-----LTFGYTYRPICLPPCTKETNAILDLSANK
CarpBfA      MGIQEYEFYFDVALIQLEKP-----VDFSSTLRPICIPPCTKETNGALKLSESEG
XenBf        KGIKRAFDDYDVALLELQRNDK-----IEFSENARPICIPPCTQGT AQALKPQSGAP-
AsMASPa      KRHTTFFDSDI AVLKLDSP-----VTLTPVVRPICLPLTETEKQLPQKSQNPQ
AsBf         DKNTLKN--DVTLLILLKEYHRNMTSTYIERISYTFYIRPVCLPCMNSCLKESQLTDNDG
Homology     * . : : :                               . : : *

                130        140        150        160        170        180
                |         |         |         |         |         |
AsMASPb      -----AEGEVAGWGLYTTVSGSSYK-----LY
SuBf          -----RTAFVTGWG-HTLKGQTS PA-----LM
CarpMASPa    -----TLGIVAGWGINTANTSASTSGL----TSDLGTVSELLQ
PigTrypsin   -----TECLISGWGNTKSSGSSYPS-----LLQ
HuHepsin     -----KICTVTGWGNTQYYGQQAGV-----LQ
LampBf       -----DWTTL CNIHGKNLIDVKKNTSLTVTGFGLLEGDKKHAQQ LQ
CarpBfA      T-----CRKHEEILMSNELVEASFTSDMETD-----HSPKHIKN
XenBf        -----CSSHEKTL LSEEEVKAVFIAEES-----NKPMKEMH
AsMASPa      H-----NVNTWYKGVVTGWG--KTEVGTLSN-----HLL
AsBf         KSLTGPGQDRCDIEEKILLENNAKVVATGFGDTSRKNEPDRKK-----NIKLSKKLQ
Homology

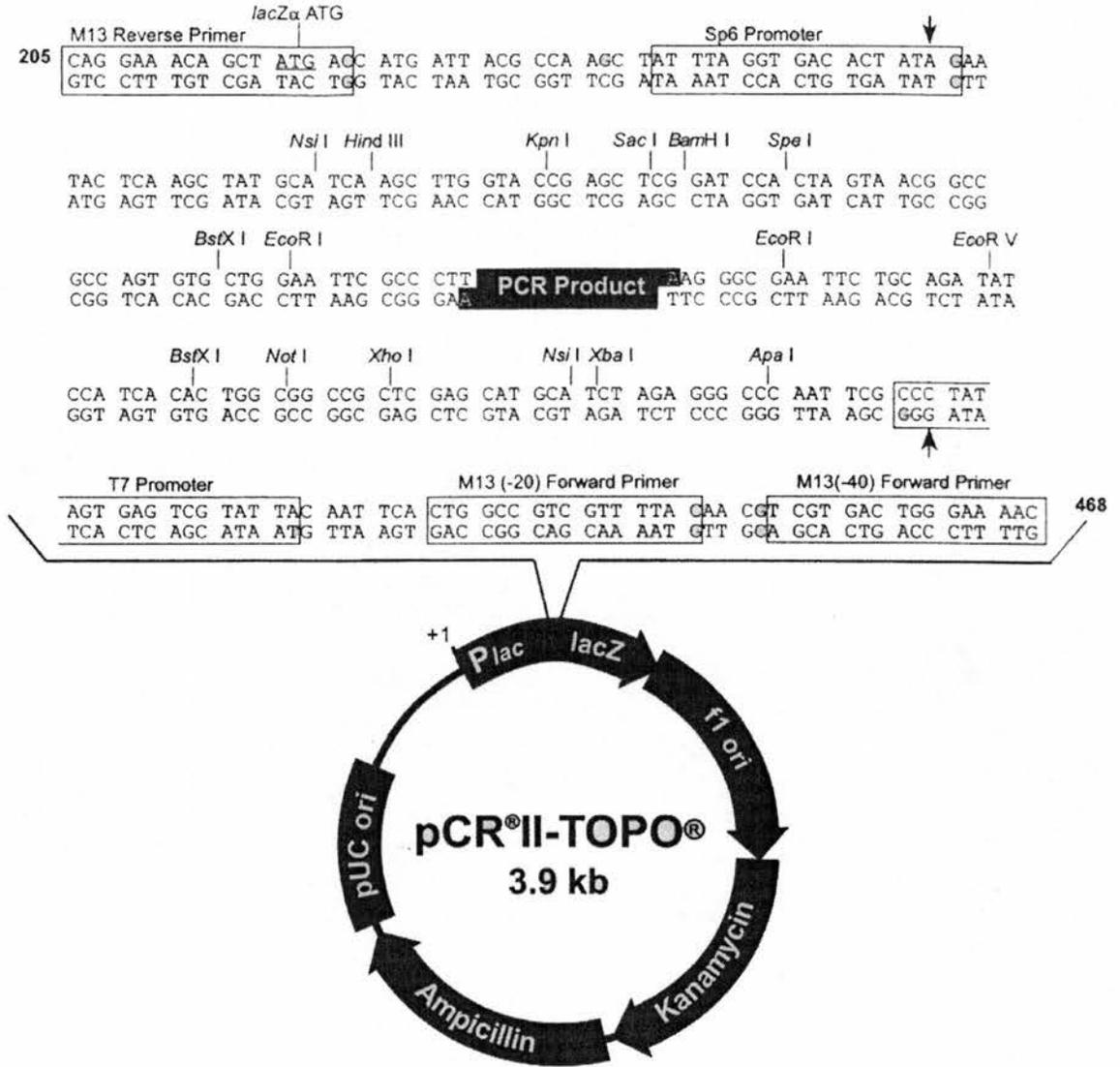
```

	190	200	210	220	230
AsMASPb	QAQFPIVSTQ	RCEDAFIELSRQLN	KTLNKGRLRITERM	FCAFEGD	SH-TTCEGDSGSP
SuBf	EIMIPVLDSSCS	IAMSAHGIAVD	TTT-----	ELCAGIERKD---	S-CQGDSGGP
CarpMASPa	YVKLPIVPQDECE	ASYASR----	SVNY----	NITSNMFCAGFYEGG	QD-T-CLGDSGGA
PigTrypsin	CLKAPVLS	DSSCKSSYPG-----	QITGNM	ICVGFLEGGKD-S-	CQGDSGGP
HuHepsin	EARVPIISNDVC	NGADFYG----	N-----	QIKPKMFCAGYPEGG	ID-A-CQGDSGGP
LampBf	QATVQYAKKEVCL	KDIMARFN--	VTEEKAEKHITENML	CAWNATADT----	CRGDSGGP
CarpBfA	IIFKLGKYRDAC	VEDAKKAKGIN--	VENAREAVTDN	FLCSGGIEPETDD	VACKGESGGA
XenBf	VLIKRGQKRSAC	LEAAKAPELKN-	VTNIEDAVTDQFL	CTGGIVPVADPPV	CKGDSGGP
AsMASPa	KVRLPFVSNEVC	QTYDELYEHIT-----	ITENMICAGYPGGHR--	DACKGDSGGP	
AsBf	QALLKIQDDQSC	QDAIKYINEKQR----	IKYSYNTTLFCC	CLDPHDNG-V	DTCQGDSGGP
Homology	. *			:*	* **:. .

Appendix 4 Multiple alignment of the thiolester containing sequence regions (**BOLD**) that were used to design degenerate primers 5, 6, 7 and 9 for C3 in Chapter 3. Primer 8 was designed against a different region in AsC3. Asterisks below the sequence indicate positions where all the amino acid sequences share same amino acid residue, two dots indicate conserved amino acid substitutions and one dot indicates semi-conserved amino acid substitution. **LimA2m** alpha2-macroglobulin from *Limulus*; accession number D83196, **LampA2m** is alpha2-macroglobulin from the lamprey *Lampetra japonica*; accession number T43166, **CarpC4** is C4 from *Cyprinus carpio*; accession number BAB02384, **HumC3** is C3 from *Homo sapiens*; accession number P01024, **XenC3** is C3 from *Xenopus laevis*; accession number AAB60608, **LampC3** is C3 from *Lampetra japonica*; accession number BAA00983, **SuC3** is C3 from *Strongylocentrotus purpuratus*; accession number AF025526, **AsC3** is C3 from *Halocynthia rorezi*; accession number AB006964.

	10	20	30	40	50	60	
LimA2m	GCGEQNM VKFPVNI	FVLDYLTAT---	GSITDSI	KEKALNN	MRKGYARQ	QNYR---HPDGS	
LampA2m	GCGEQNM VKFAPNI	YIQEYLQNS---	GQLTDAV	RDKALNF	LRVGYQR	QLTYK---RDDHS	
CarpC4	GCAEQTM VKMSPAI	HAMRYLDAT	NQWISL	KAERRDE	AQSMIQT	GYNTVLTYK---KVDGS	
HumC3	GCGEQNM IGMTP	TVI	IAVHYL	DETEQWE	KFGLEK	RQGALELIKGGYTQQLAFR---QPSSA	
XenC3	GCGEQNM I	STTPSVI	ATRYLD	ASGQW	ERVGVN	RRDQALKNMRQGYAQMAFR---KPDNS	
LampC3	GCGEQNM I	KMAPT	TTLTI	YLD	SVQEW	EKIGLHRREEAIGFLKQGY	SRELSYR---KADHS
SuC3	GCGEQTM I	YLAPT	LFVYQY--	LI	AVGSD	TAEQEAR-IYDYIADGVARELTYR---QDN	
AsC3	GCGEQNM I	R	IAPV	VYI	HAYRS	NLEAFTVTDAQRAQ-TLKYIQDGYAHELEYKTQVPQ	
Homology	**.*.*.*:	*	*		:	* : : . :	
		70					
LimA2m	YSAFGNR	DKQGNL	FLTAFV				
LampA2m	YSAFGK	SDDDG	NWTAFV				
CarpC4	YGAFLR	TP--	SSIWLTAFI				
HumC3	FAAFV	KRA--	PSTWLTAYV				
XenC3	YAAWK	DRP--	ASTWLTGYV				
LampC3	YAAFI	KRP--	SSTWLTAFV				
SuC3	YAAWK	HRP--	GSTWLTAFV				
AsC3	FAVW	ANNP--	PSTWLN				
Homology	:.:.:		.:*:.:.:				

Appendix 5 Map of pCR® 2.1 TOPO® vector sequence surrounding the cloning site. Taken from the Invitrogen instruction manual version J.



	310	320	330	340	350	360
S.p.corticle	SCDNGTRCVQEGEICDGT	----	QHCSDDLDESDELCSAGN	-----	-----	-----
SP1	PCS	-----	TSCNRGVTRDRICSAGN	-----	-----	-----
Homo.plas	SAQTPHHTNRTPENFPCKNLDENYCRNPDGKRAPWCHTTNSQVRWEYCKIPSCDSSPVST					
Homo.mosaic	ASPALASLSR	-----	SSSGRSSARSASVTTSP	-----	-----	-----
C.e.hypo	PPQQPQSFSGTHELHLQRQREQQQQQQQQQQQQQQQQQQQQNPOQQPQQTTQFGQSQIQLQSG					
Homology	.					

	370	380	390	400	410	420
S.p.corticle	-----	VKCFSCDG	----	GSKCLKWN	-WVCDEFADCSMDAD	-----
SP1	-----	SHSTCNGS	----	ALQSNVCNTQVCPLWTTWTNYG	-----	ECSTTC
Homo.plas	EQLAPTAPPELTPVVQDCYHGDGQSYRGTSSTTTTGKKCQSWSSMTPHRHQKTPENYPNA					
Homo.mosaic	-----	TRVYLVRATPVGAVPIRSSPARSAPATRATRESP	----	VQFWQGHT	-----	-----
C.e.hypo	PVPPQQHPPQQQPQQQPELERSPLDQHAQLYQQRMSQYRENFNRHPARPKADPCPGGFC					
Homology	.					

	430	440	450	460	470	480
S.p.corticle	GTVFQRCWKGAYLCGHTHFCVL	----	QRWRNNHDDCGDDTD	-----	-----	EEDCET
SP1	GKGR	-----	HRSRSLQGNCDNRLS	-----	-----	LESTSC
Homo.plas	GLTMNYCRNPDADKGPWCFTTDPVSRWEYCNLKKCSGTEASVVAPPPVLLPDVETPSEE					
Homo.mosaic	GIRYKEQRESCP	-----	KHAVRCDGVDCKLKSDELG	-----	-----	CVRFDW
C.e.hypo	APVPQAPQQERP	-TPPPVLAPVINTATQPPLPQPYPTRYRPAPPPPP	-----	ACDGQGC	-----	-----
Homology

	490	500	510	520	530	540
S.p.corticle	DFAWTGSYGWSSWGDWSECHPSCG	----	LGTRSRSRFCASPGR	-----	-----	CLGESQEEE
SP1	NLRYFCP	-AWSPWSVYSCSVSCG	----	IGTQTRKRTCYHQGEIGE	-----	CIGPLNDTT
Homo.plas	DCMFGNGKGYRGRATTVTGTPCQDWAQEPHRHSIFTPETNPRAGLEKNYCRNPDGDVVG					
Homo.mosaic	DKSLLLKIYSGSSHQWLPICSSNWNDSYSEKTRQLGFESAHR	-----	TTEVAHRDFA	-----	-----	-----
C.e.hypo	VNPPVVSQVWHDWSDWSTCSTCG	----	DGAKSRRRECSTNNCQG	-----	-----	ADYETPCNLG
Homology

	550	560	570	580	590	600
S.p.corticle	EC	-----	EQVPCVDENVIACGIKSHIHFR	----	DGGLALAERIVGGQPATA	----
SP1	ICNIDCHNQTQHAVSRNIEQCGLRVAASN	----	NRRSSIILKIFGGNISRR	----	NSWPWQ	----
Homo.plas	GPWCYTTNPRKLYDYCDVPQCAAPSFDGKQ	-QVEPKKCPGRVVGCVVHP	----	HSWPWQ	----	----
Homo.mosaic	NSFSLRYNSTIQESLHRSHCPSQRYISLQCSHCGLRAMTGRIVGGALASD	----	SKWPWQ	----	----	----
C.e.hypo	PCQT	----	WSEWCEWSTCSASCQSGQRERTRFCHLGTNRCEGKDYESEQCSAGPCPEWSQW			
Homology		*		:	.	..*

	610	620	630	640	650	660
S.p.corticle	A	----	QLFYRTRG	----	SWQLVCGGTLIDPQVVLTAAHCFMGPMMATSR	-WQVHLGKHSVDF
SP1	VSLQEFYFYSHRFNYSNWMHFCGGTIVSSQWVITAAHCLQQITENEYS	----	IHKFSAVFGLFR	----	VS	----
Homo.plas	VS	----	LRTRFG	----	MHFCGGTLISPEWVLTAAHCLEKSPRPS	----
Homo.mosaic	VS	----	LHFG	----	TTHICGGTLIDAQWVLTAAHCFVTRKLVLEGWKVYAGTNSLHQ	----
C.e.hypo	EDWGQCSVTCCQGVAVRQRTCLGGVFGDHLCCQGPKEQRACDGGPCSLWSPWQEWSTCSA					
Homology		*	*

	670	680	690	700	710	720
S.p.corticle	VPEAGSQHRL	-VREIFVHKKFG	-EHGGVGC	DIALLLILDEPVPQETGQINWACLDE	-GMPL	
SP1	LNQHQHTQRIGFKRTFIHSD	FQSAHLTFRNDVALIQLDRKIQWTS	-NIRPACLPG	-GEEP		
Homo.plas	VNLEPHVQEIEVSRLFLEP	TRK-----	DIALLLKSSPAVITD	-KVIPACL	PSPNYVV	
Homo.mosaic	LPEAASIAEIIINSNYTDEEDD	-----	YDIALMRLSKPLTL	SA-HIHPACL	PMHGQTF	
C.e.hypo	SCGSGMKRRQRVCQFGTDCQGP	-----	NEESQFCYGP	PPCAEWTEWCEW	SGCSSKCGPGQ	
Homology	.	.	.	:	:	.*

	730	740	750	760	770	780
S.p.corticle	NDRTECYISGWGVTEMGGNGPD	-VLHEARMPLIPRRICN	-YKKS	YNGKIEKTMLCAG	-HL	
SP1	IETENCYITGWGRTRINSSSEL	SSELRESIIPILSNKQCRRLG	SGYNTINMTLHICAGDPV			
Homo.plas	ADRTECFITGWGET-QGTFGAG	-LLKEAQLPV	ENKVCN-RYEF	LNGRVQSTELCAG	-HL	
Homo.mosaic	SLNETCWITGFGKTRETDDKT	SPFLREVQVNLIDFKCN	-DYL	VYDSYLT	PRMMCAG	-DL
C.e.hypo	RTRTRGCLGPNQEATTCQGPS	--IETT	LCE--GQSCCN--	WSEWCHWSMCDKECGGQV		
Homology	:	*	.	::	*	.*. * :

	790	800	810	820	830	840	
S.p.corticle	EGGIDACQGDSSGGLSCLGPDD	HVYVVGVT	SWG	HG-CAIANKPGVYTKVSSYL	DWIDEMI		
SP1	RGGRDTCQGDSSGP	IVCN	RS	G-IWYIAGVTS	SHSLAF	CGARNNVGIYTRTT-----	
Homo.plas	AGGTDS	CQGDSSG	PLVCFEKD	-KYILQGV	TSWGLG	-CARPNKPGVYVRVSRFVTWIEGVM	
Homo.mosaic	HGGRD	S	CQGDSSG	PLVCEQNN	-RWYLAGV	TSWGTG-CGQRN	KPGVYTKVTEVLPWIYSKM
C.e.hypo	RYIEYMFRTGCEWSPCSTQL	LACEVGVQ	RSR	RQCVG-ESG	CHCIGLAEES	QQCRGLTQCPP	
Homology	:	:	.*: .	

S.p.corticle	HHY LHHE--
SP1	-----
Homo.plas	RNN-----
Homo.mosaic	ESEVRF
C.e.hypo	KPPC-----
Homology	

Appendix 7 Multiple alignment using CLUSTAL W of SP2 with the most similar proteins identified by BLAST. Asterisks below the sequence indicates positions where all the sequences share the same amino acid residue, two dots indicates conserved amino acid substitutions, one dot indicates semi-conserved amino acid substitutions. **Mouse.pkall**; *Mus musculus* kallikrein precursor Acc No P26262. **Scolo.plas**; *Scolopendra subspinipes* plasminogen activator Acc No AAD00320. **Rab.fIX**; *Oryctolagus cuniculus* clotting factor IX Acc No P16292. **Xeno.sp**; *Xenopus laevis* channel activating serine protease Acc No AF029404. **Rat.pkall**; *Rattus norvegicus* prekallikrein precursor Acc No O88780. **Anoph.sp**; *Anopheles gambiae* putative immune serine protease Acc No AF117751.

```

                10         20         30         40         50         60
                |         |         |         |         |         |
mouse.pkall  -----
scolo.plas  -----
rab.fIX     -----
xeno.sp     -----
rat.pkall   -----
anoph.sp    MRQRIWNRPRLVLLALAVLIGGWCNMVVVGIYDPRTAPHSRHHVHMMPPEMHGAYSQVHHH
SP2        -----
Homology

```

```

                70         80         90         100        110        120
                |         |         |         |         |         |
mouse.pkall  -----
scolo.plas  -----
rab.fIX     -----
xeno.sp     -----
rat.pkall   -----
anoph.sp    RAQDPTPQQYIQTDQYQYAQPQRQHPSLVAGPQQQQQHQHQHGPSGPQYQPGVPLAPYPT
SP2        -----
Homology

```

```

                130        140        150        160        170        180
                |         |         |         |         |         |
mouse.pkall  -----
scolo.plas  -----
rab.fIX     -----
xeno.sp     -----
rat.pkall   -----
anoph.sp    ETQRSPAYGRSQAYTQQPAPVPLAPRFGYGEEDRLIGETAPAAKLIRQPVHTLLKDFNGL
SP2        -----
Homology

```

```

                190        200        210        220        230        240
                |         |         |         |         |         |
mouse.pkall  -----
scolo.plas  -----
rab.fIX     -----
xeno.sp     -----
rat.pkall   -----
anoph.sp    ECPEGRTGHFPYVMDCRQFLSCWKGRGFILNCAPGTLFNPNTRECDHPSKVSCLPVPSLN
SP2        -----
Homology

```

	250	260	270	280	290	300
mouse.pkall	-----					
scolo.plas	-----					
rab.fIX	-----					
xeno.sp	-----					
rat.pkall	-----					
anoph.sp	SVNEPANRAPPKLSYTDQRPPQFQQQQRQPQYLQPQSQRQQEELTCPPGVIGLRPHP					
SP2	-----					
Homology						

	310	320	330	340	350	360
mouse.pkall	-----					
scolo.plas	-----					
rab.fIX	-----					
xeno.sp	-----					
rat.pkall	-----					
anoph.sp	TDCRKFLNCNNGARFVQDCGPGTAFNPLILTCDHLRNVDCDKSENVIVDYDRPTSRPVAS					
SP2	-----					
Homology						

	370	380	390	400	410	420
mouse.pkall	-----					
scolo.plas	-----					
rab.fIX	-----					
xeno.sp	-----					
rat.pkall	-----					
anoph.sp	GPTSHYYPISHIPAGSQVPVAVVNPHQQSRPTIPAPQQQTPPRQPPATGDRAPAHDPVEQI					
SP2	-----					
Homology						

	430	440	450	460	470	480
mouse.pkall	-----					
scolo.plas	-----					
rab.fIX	-----					
xeno.sp	-----					
rat.pkall	-----					
anoph.sp	DPDHQPTESNFDEDYGEQPDADGEEPVDYDGFDLRSNFGAPEQVDRRRPKASRAQATTTAK					
SP2	-----					
Homology						

	490	500	510	520	530	540
mouse.pkall	-----					
scolo.plas	-----					
rab.fIX	-----					
xeno.sp	-----					
rat.pkall	-----					
anoph.sp	PYPVYIRPPSRQPESLHRDPDVVQSVQRPVYVALPLEQTTVPVPTSTTSRPLRTPFPVTRK					
SP2	-----					
Homology						

	550	560	570	580	590	600
mouse.pkall	-----					
scolo.plas	-----					
rab.fIX	-----					
xeno.sp	-----					
rat.pkall	-----					
anoph.sp	EDIEIQQLDALKLMLTPYMKEHKDTVALNNTTKLSTMMTTTTTTTTTEPPPIVQVIGLPAPT					
SP2	-----					
Homology						

	610	620	630	640	650	660
mouse.pkall	-----					
scolo.plas	-----					
rab.fIX	-----					
xeno.sp	-----					
rat.pkall	-----					
anoph.sp	PRNNYKPSSAAAAPYVLPRASEVNDFFYGASEPVPLASWPLPPPYITEPVEGPAKKEPES					
SP2	-----					
Homology						

	670	680	690	700	710	720
mouse.pkall	-----					
scolo.plas	-----					
rab.fIX	-----					
xeno.sp	-----					
rat.pkall	-----					
anoph.sp	VVYPIYRRTTPTTTTTTASPAPAPAIRSRFGDNRPSWRPLIVPHATTTKTPTTTPPATT					
SP2	-----					
Homology						

	730	740	750	760	770	780
mouse.pkall	-----					
scolo.plas	-----					
rab.fIX	-----					
xeno.sp	-----					
rat.pkall	-----					
anoph.sp	TSTTPRDPCYGFNCGNGVCIDEAEVCDGRDGCGRRADEQVCDHIGYELKLSKKAQGSVE					
SP2	-----					
Homology						

	790	800	810	820	830	840
mouse.pkall	-----					
scolo.plas	-----					
rab.fIX	-----					
xeno.sp	-----					
rat.pkall	-----					
anoph.sp	VRVYDRWGYVCDGFTLEAGNVVCRELGFAGGAIEIKSHSYFPPNGTDPDEPEKQHGPF					
SP2	-----MESKKNVLIL					
Homology						

```

      850      860      870      880      890      900
      |       |       |       |       |       |
mouse.pkall MILFNRVGYFVSLFATVSCGCM TQLYKNTFFRGGDLAAIYTPDAQYCQKMCTFHPRCLLF
scolo.plas -----
rab.fix -----
xeno.sp -----
rat.pkall -----
anoph.sp MMDAVRCQGNESLRECSFNGWGVSDCNREEVVGVCRT PVMSCPQDYWLCHASEECIPV
SP2 LESFLLFLILSSMQGESLV LKLEGLKLNELIEWKGQISRKNAPGQLQITANTNRPGTQYT
Homology

```

```

      910      920      930      940      950      960
      |       |       |       |       |       |
mouse.pkall SFLAVTPPKETNKRF GCFMKESITGTLPRIHRTG AISGHS LKQC GHI SACHRDIYKGLD
scolo.plas -----
rab.fix -----GVSVSHASKK ITR-----
xeno.sp -----
rat.pkall -----
anoph.sp QFLCDNVRDCADGSD ESDPHCKAPLAVRLVAGPTDREGRVE INYHGTWGTVCDDDFGVRE
SP2 MMRLQIQGARHRMMFR YGIKGERSTAPKLC PYDLIQPGWEFQ I I IHVTTGYLNVYYEGR
Homology

```

```

      970      980      990      1000      1010      1020
      |       |       |       |       |       |
mouse.pkall MRGSNFNISKTDNIEECQ KLC TNNFHCQFFTYATS AFYRPEYRKKCLLKHSASGTPTS I K
scolo.plas -----MNSFTILIVTYFSLAFG--S--RCGIKN-----
rab.fix -----ATTIFSNTEYENFTEAET--I--RGNVTQ-----
xeno.sp -----MEPLPLLSLFL LAVVHLE-----PSRSQE-----
rat.pkall -----MGRPPPCAIQTWILLFLLMG-----
anoph.sp ARVICRQLG-FNGTAEVRKSVYPPGVGQIWL DQVACNGTEPSIEDCVHWHWGEHNCAHTE
SP2 LKFILPISPLTSIENAVSIRIHGSVITKRLGLL TGADI ISELQAPNCGRIIR-----
Homology

```

```

      1030      1040      1050      1060      1070      1080
      |       |       |       |       |       |
mouse.pkall SADNLVSGFSLKSCALSEIGCPMDIFQHS AFADLNVSQVITPDAFVCR TICTFHPNCLFF
scolo.plas -----G-PM-----LDEFN-----
rab.fix -----RSQS-----SDDFT-----
xeno.sp -----G-----VQS-----
rat.pkall -----AWAGLTRAQ-----
anoph.sp DV-----G--VRCGVYVPTKARPARLRATRPNPR FDFVRSRKI-----
SP2 G-----G-----NVPQFCRGGE-----
Homology

```

```

      1090      1100      1110      1120      1130      1140
      |       |       |       |       |       |
mouse.pkall TFYTNEWETESQRNVCF LKTSKSGRPSPP I PQENAI SGYSLLTCRKTRPEPCHSKIYSGV
scolo.plas -----
rab.fix -----
xeno.sp -----
rat.pkall -----
anoph.sp -----
SP2 -----
Homology

```

```

                1150      1160      1170      1180      1190      1200
                |        |        |        |        |        |
mouse.pkall1 DFEGEELNVTFVQGADVCQETCTKTIRCQFFIYSLLPQDCKEKGCKCSLRLSTDGSPTRI
scolo.plas  -----
rab.fIX     -----
xeno.sp     -----
rat.pkall1  -----
anoph.sp    -----
SP2         -----
Homology    -----

```

```

                1210      1220      1230      1240      1250      1260
                |        |        |        |        |        |
mouse.pkall1 TYGMQGSSGYSRLRLCKLVDSPDCTTKINARIVGGTNASLGEWPWQVSLQVK-LV--SQTH
scolo.plas  -----RIVGGEEAAEPGEFPWQISLQVVSWEY--GSYH
rab.fIX     -----RIVGGENAKPGQFPWQVLLNGK-----VEA
xeno.sp     -----RIVGGENATPGKFPWQVSLRYN-----GRH
rat.pkall1  -----GSKILEGQECKPHSQPWQTALFQG-----ERL
anoph.sp    TY-----GARVVHGSETVYGHHPWQASLRVK-----TMH
SP2         -----QQRIVGGTTARPGNFPWQISIRKVKAYSNGSPH
Homology    ::: *                *** :
```

```

                1270      1280      1290      1300      1310      1320
                |        |        |        |        |        |
mouse.pkall1 LCGGSIIGRQWVLTAAHCFDG--IPYPDVWRIYGGILSLSE-----ITKETPSSRIK
scolo.plas  YCGGSILDES WVVTAAHCFVEG--MN-PSDLRILAGEHNFKK-----EDGTEQWQDVI
rab.fIX     FCGGSIINEKWVVTAAHCIPK--DD--NITVVAGEYNIQE-----TENTEQKRNI
xeno.sp     VCGASLISSNYILTAAHCFPS--DHLMSDYKVYLGVLQLEV-----PTSESQLLSLK
rat.pkall1  VCGGVLVGDWRVLTAAHCKK-----DKYSVRLGDHSLQK-----RDEPEQEIQVA
anoph.sp    WCGAVLITRYHVLTAAHCLIG--YP-KSTYRVRIGDYHTAA-----YDNAELDIFIE
SP2         VCGGTLIAGQWVITAAHCFTSRVKRERKKHFVRVGDYFNDRNLPHSQDSMVEESHDI AIS
Homology    **. ::      : :*****      : *                :
```

```

                1330      1340      1350      1360      1370      1380
                |        |        |        |        |        |
mouse.pkall1 ELIIHQEY---KVSEGNVDIALIKLQTPLNYTEFQK-PICLPSKADTN--TIYTNCWVTG
scolo.plas  DIIMHKDY---VYSTLENDIALLLKLAEPDLTPTAVGSICLPSQNNQ---EFGHCIVTG
rab.fIX     RIIPYHKYN-ATINKYNHDIALLELDKPLTLNSYVT-PICIANREYTNIFLNFSGSYVSG
xeno.sp     EIIIHPSY---SHDTSTGDVALAALDPPATFSNVVQ-PIPLPDENVQFP--IGMNCQVTG
rat.pkall1  RSIQHPCFNSSNPEDHSHDIMLIRLQNSANLGDKVK-PIELANLCP---KVGQKCIISG
anoph.sp    NTYIHEQFREG--HHMSNDIAVVVLKTPVRFNDYVQ-PICLPARDAPY--LPGQNCTISG
SP2         QIYIHEGFT--QYPATRNDIALIKLSEPVSLTRFVQ-PACLPTSPDQ--FTDGNTCGISG
Homology    : :                * : * .                . :.                : :*
```

```

                1390      1400      1410      1420      1430      1440
                |        |        |        |        |        |
mouse.pkall1 WGYTKEQGETQ---NILQKATIPLVPNEECQKKYRDY-----VINKQMICAGYKEGGT
scolo.plas  WGSVREGGNSP---NILQKVSVPLMTDEECSEYYN-----IVDTMLCAGYAEGGK
rab.fIX     WGRVFNRRGQA---SILQYLRVPFVDRATCLRSTK-F-----TIYNNMFCAGFDVGGK
xeno.sp     WGNIQQGVSLPGS-KTLQVGNVKIISRQTCNCLYHINPSSDSLGSVQQDMICAGSAAGSV
rat.pkall1  WGTVTSPQENF-P-NLNCFAEVKIYSQNK CER--AYP-----GKITEGMFCAGSSN-GA
anoph.sp    WGATEAGSKDS-S-YDLRAGTVPLLPDSVCRRPEVYG-----DSLIDGMFCAGTLEPGV
SP2         WGATNFTQLRDEYPFCLRAATVHTWPKNCSRSYPRS-----FSNDSMLCAGDEG--I
Homology    **                * .                : *                . * .***
```

	1450	1460	1470	1480	1490	1500
mouse.pkall	DACKGDSGGPLVCKHS-GRWQLVGITSWG-EGCGRKDQPGVYTKVSEYMDWILEKTQSSD					
scolo.plas	DACQGDSGGPLVCPNGDGTYSLAGIVSWG-IGCAQPRNPGVYTQVSKFLDWIRN-TNIDG					
rab.fIX	DSCGDSGGPHVTEVE-GTSFLTGIISWG-BECAIKGKYGVYTRVSWYVNW-----					
xeno.sp	DACQGDSGGPLTCTVN-NQPYLAAVVSWG-DECGAPNRPGVYILISLYSSWIRSIDPSAT					
rat.pkall	DTCQGDSGGPLVCNG-----VLQGITTWGS DPCGKPEKPGVYTKICRYTNWIKKTMGKRD					
anoph.sp	DSCDGDSSGGLVCPNSEGLHTLTGIVSWG-KHCGYANKPGVYLKVAHYRDWIEQKLNQSL					
SP2	DTCQGDSGGPLTCLSRDGNITLWGITSYG-KGCGNKSQPGV-----					
Homology	*:*.****** . * .: ::* *. . **					

	1510	1520	1530	1540	1550
mouse.pkall	VRALETSSA-----				
scolo.plas	SNVIEFII-----				
rab.fIX	-----				
xeno.sp	VQYFTVDIPSDPQNSGCVGADGQFYPNPNGASIFLVTFAALPFYWLTTYILSDF				
rat.pkall	-----				
anoph.sp	HQHGV-----				
SP2	-----				
Homology					

Appendix 8 Multiple alignment using CLUSTAL W of SP3 with the most similar proteins identified by BLAST. Asterisks below the sequence indicates positions where all the sequences share the same amino acid residue, two dots indicates conserved amino acid substitutions, one dot indicates semi-conserved amino acid substitutions. **Sus.kall**; *Sus scrofa* kallikrein Acc No BAA37147. **Mus.coagxi**; *Mus musculus* coagulation factor XI Acc No AAK40233. **Mus.rik**; *Mus musculus* riken Acc No AK004939.

```

          10          20          30          40          50          60
          |          |          |          |          |          |
SUS.KALL  -----
Mus.coagxi -----
SP3       -----
Mus.rik   MPTTEVVPQAADGQGDAGDGEEAAEPEGKFKPPKNTKRKNRDYVRF TPLLLVLAALVSAGV
Homology

          70          80          90          100         110         120
          |          |          |          |          |          |
SUS.KALL  -----MEVIVLFR IISFRQAVYFMCLFAAVS
Mus.coagxi -----MTSLHQVLYFIF FASVS
SP3       -----
Mus.rik   MLWYFLGYKAEVTVSQVYSGSLRVLNRHFSQDLGRRESIAFRSESAKAQKMLQELVASTR
Homology

          130         140         150         160         170         180
          |          |          |          |          |          |
SUS.KALL  CGCLPQLHKNTFFRGGDVSAMYTPSARHCQMMCTFHPRCLLFSFLPADSTSVTDKRFGCF
Mus.coagxi SECVTKVFKDISFQGGDLSTVFTPSATYCR LVCTHHPRCLLFTFMAESSDDPTKW FACI
SP3       -----
Mus.rik   LGTYYNSSSVYSFGEGPLTCFFWFILDIPEYQRLTLSPEVVRELLVDELLSNSSTLAS YK
Homology

          190         200         210         220         230         240
          |          |          |          |          |          |
SUS.KALL  LKDSVTGMLPRVLRENAISGHSLKQCGHQIRACHRD IYK GIDMRG-----
Mus.coagxi LKDSVTEILPMVNMTGAISGY SFKQCPQQLSTCSKDEYVNLDMKG-----
SP3       -----
Mus.rik   TEYEVDPEGLVILEASVNDIVVLNSTLGCYRYSYVNP GQVLP LKGPDQQTTSCLWHLQGP
Homology

          250         260         270         280         290         300
          |          |          |          |          |          |
SUS.KALL  --VNFNVSKVKTVEECQERCTNSIHCLFFTYATQAFNNAEYRNNCLLKHSPGGTPTS IKV
Mus.coagxi --MNYNSSVVKNARECQERCTDDAHCQFFTYATGYFPSVDHRKMCLLKYTRTGTPTTITK
SP3       -----MHYIHSYNILD
Mus.rik   EDLMIKVRLEWTRVDCRDRVAMYDAAGPLEKRLITSVYGCSRQEPVMEVLASGSVM AVVV
Homology

          310         320         330         340         350         360
          |          |          |          |          |          |
SUS.KALL  LANVESGFSLKPCADSEIGCHMDIFQH LAFSDVDVA----RVIAPDAFVCRTICTYHPNC
Mus.coagxi LNGVVSGFSLKSCGLSNLACIRDIFPNTVLADLNID----SVVAPDAFVCRRIC THHPTC
SP3       YIKKIQIFYLDPCSLTNGGCN-----QLCNWTGNA AICG-----
Mus.rik   KKGMSYYPFLLSVKSVAFQDCQVNLTL EGR LDTQGFLRTPYYPSPSTHCSWHLTV
Homology          . : . . . . . *

          370         380         390         400         410         420
          |          |          |          |          |          |
SUS.KALL  LFFTF-----YTNAWKIESQRNVCFLK TSHSGTSPSFPTQENAI S
Mus.coagxi LFFTF-----FSQAWPKESQRHLCLLKT SESGLPSTRITKIHALS
SP3       -----
Mus.rik   PSLDYGLALWFDAYALRRQKYNRLCTQGQWMIQN RRLCGFR TLQPYAERIPMVASDGVTI
Homology

```

	430	440	450	460	470	480
SUS.KALL						
Mus.coagxi	GYSLLTCKQTLPEPCHSKIYSEVD-----				FEGEELNVTFVQGANLCQETCT	
SP3	GFSLQHCRHSVPVVFCHPSFYNDTD-----				FLGEELDIDVVKGQETCQKTCT	
Mus.rik	-----CQSGYRLQSD-----				NRTCEDIDECTEGPNPCYFRFP	
Homology			: *		:	: * :
	490	500	510	520	530	540
SUS.KALL						
Mus.coagxi	KTIRCQFFTYSLHPEDCR-----				GEKCKCSLRLSSDGSPTKITHGMRA	
SP3	NNARCQFFTYPSHRLCNER-----				NRRGRCYLKLSSNGSPTRILHGRGG	
Mus.rik	AFCVNTIGSYSCQPYRCN-----				GTNEMNYRSGSCCKVRNGSCG	
Homology	:				. * :	:
	550	560	570	580	590	600
SUS.KALL						
Mus.coagxi	SSGYSRLRLCRSGDHSACATKANTRIVGGTDSFLGEWPWQVSLQAKLRAQNHLCGGSIIIGH					
SP3	LSGYSRLRLCKMDN--VCTTKINPRVVGGAASVHGEPWQVTLHIS---				QGHLCGGSIIIGN	
Mus.rik	TTASIRSMVEPVVP----				IETERRVFRGMASVVSAPWMAQVLYR---	
Homology	.		. * . * * . * * . * * . :			* * . . : * . .
	610	620	630	640	650	660
SUS.KALL						
Mus.coagxi	QWVLTAAHCFDGLSLPDIWRIYGGILNISEITKETPFSQVK--				EIIIHQNYKILESGHDI	
SP3	QWILTAAHCFSGIETPKKLRVYGGIVNQSEINEGTAFFREQ--				EMIIHDQYTTAESGYDI	
Mus.rik	RWLVSAAHCFRVSYSGLLVYLGTTTRSSHLTHLDTTRRQRREVEQIIIVHPGF				TAEYLNVDV	
Homology	: * : : * * * * :				. : . : . :	* :
	670	680	690	700	710	720
SUS.KALL						
Mus.coagxi	ALLKLETPLNYTDFQKPICLPSRDDTNVVTNCWVTGWGFTEEKGEIQN-				ILQKVNIPLV	
SP3	ALLKLESAMNYTDFQRPICLPSKGDNRNAVHTECWVTGWGYTALRGEVQS-				TLQKAKVPLV	
Mus.rik	ALIKLSRPVVFNDIITPICLP-CGETPSPGDKCWVTGFGR				TENTGYDSSQTLQEVDPVIV	
Homology	** : * . : : . :	* : * * * * * : * * : * * . *		* . . * * : . . : *
	730	740	750	760	770	780
SUS.KALL						
Mus.coagxi	SNEECQKSYRDHKISKQ--				MICAGYKEGGKDACKGESGGPLVCKYNG--	
SP3	SNEECQTRYRRHKITNK--				MICAGYKEGGKDTCKGDSGGPLSCKYNG--	
Mus.rik	NTTQCMEAYRGVHVIDENMMMCAGYEAGGKDACNGDSGGPLACQ				RADSCDWYLSGVTSTFG	
Homology	* * : : . :	* : * * * . *	* * * : * * : * * * * *	*		* . * * * . * : *
	790	800	810	820	830	840
SUS.KALL						
Mus.coagxi	EGCARREQPGVYTKVIEYMDWILEKTQDDDGQSWMK-----					
SP3	EGCGQKERPGVYTNVAKYVDWILEKTQTV-----					
Mus.rik	RGCGGLARYYGVVVNVVHYEGWIRTQMGNDSTGLCPRQYNPCKGLVDAHTDCASKL				DKCRS	
Homology	LGCGRPNFFGVYTRVTRVINWIQQVLT-----					
	** . . * * * . * . . **					

	850	860	870	880	890	900
SUS.KALL	-----					
Mus.coagxi	-----					
SP3	FPSYMATNCARSCCQLNNGEITNCQDSADSAEACKLYVGYCSNPAMSSFMREKCRRTCGF					
Mus.rik	-----					
Homology						

SUS.KALL	-
Mus.coagxi	-
SP3	C
Mus.rik	-
Homology	

Appendix 9 Multiple alignment using CLUSTAL W of SP5 with the most similar proteins identified by BLAST. Asterisks below the sequence indicates positions where all the sequences share the same amino acid residue, two dots indicates conserved amino acid substitutions, one dot indicates semi-conserved amino acid substitutions. **Vann.tryp**; *Litopenaeus vannamei* trypsin Acc No Y15041. **Fugu trypg**; *Takifugu rubripes* trypsinogen Acc No U25747. **Dros.sp**; *Drosophila melanogaster* hypothetical serine protease Acc No AE003455. **Mus.dist**; *Mus musculus* distal intestinal serine protease like protein Acc No BC010970.

```

          10          20          30          40          50          60
          |          |          |          |          |          |
vann.tryp -----SLVLCLLLLAGA
fugu.trypg -----LIAAAA
dros.sp      MCNFHLLLILATALGLACATPSLRASDPEKILNNLAQLRQSSFLDWIQSILGPEVPAE
mus.dist    -----MESRARCIIFLLLLQ
SP5         -----MAGL
Homology    :

          70          80          90          100         110         120
          |          |          |          |          |          |
vann.tryp   FAAPSRKPTFRR-----GLNKIVGGSDATPGELPYQLSFQDVSFGFAFHFCGAS IYNNEN
fugu.trypg  YAAPIDE-----DDKIVGGYECRKNSVAYQVSLNSG-----YHFCGGLSVNEN
dros.sp     WSSPAKRECAECSGGINTRHRIVGGQETEVHEYPWMIMLMWFG---NFYCGASLVNDQ
mus.dist    ILTRARGDILPSVCGHSRDAGKIVGGQDALEGQWPQVSLWITEDG---HICGGLIHEV
SP5        YNPEVTAAT-----LDAIITVGGVSGNEMVDKTTLGFG-----TSQL
Homology    .                *:                .                :                :

          130         140         150         160         170         180
          |          |          |          |          |          |
vann.tryp   WAICAGHCVQGEDMNNPDYLQVVAGEHNRDVDEGNEQTVVLSKIIQHEDYNGFTISN-DI
fugu.trypg  WVVSAAHCHYKSR-----VVVRLGEHNIRANEGTEQFISSSRVIRHPNYSSYNIDN-DI
dros.sp     YALTAAHCVNGFYHR---LITVRLLEHNQRDQSHVKIVDRRVSRLIHPKYSTRNFDS-DI
mus.dist    WVLTAAHCFRRSLN--PSFYHVKVGGLTSLLEPHSTLVAVRNIFVHPTYLWADASSGDI
SP5        WRLTRLSHLLRD-----SDVTVTSMDVVTGASVTRLRSRIYSHPTYGENLSD--I
Homology    : :                .                .: * *                .. *

          190         200         210         220         230         240
          |          |          |          |          |          |
vann.tryp   SLLQLSQPLSFNDFVAPIALPEAG-----HAASGDCIVSGWGTTSEGGSTPSVL
fugu.trypg  MLIKLSKPATLNQYVQVALPSSC-----AAAGTMCKVSGWGNTMSSTADRNLK
dros.sp     ALIRFNEPVRLGIDMHPVCMPTPSE-----NYAGQTAVVWTGWGALSEGGPISDTL
mus.dist    ALVQLDTPLRP-SQFTPVCLPAAQTP-----LTPGTVCWVTGWGATQER-DMASVL
SP5        VLIKVAEPITWSSRVFPVCLPSPESLLDHVGHGRVHIPKQYCKLAGWGSSSELGDYASDL
Homology    *::: *                . *:::*                .                .:***                . *

          250         260         270         280         290         300
          |          |          |          |          |          |
vann.tryp   QKVSVPIVSDDECRDAYGQN-----DIDDSMICAGMPE-GGKDSCQGDSSGGPLACSD
fugu.trypg  QCLNIPILSDRDCENSYPG-----MITDAMFCAGYLE-GGKDSCQGDSSGGPVVCNN
dros.sp     QEVEVPILSQEECRNSNYGES-----KITDNMICAGYVEQGGKDCQGDSSGGPMHVLG
mus.dist    QELAVPLLDSEDCCKMYHTQGSLSGERIIQSDMLCAGYVE-GQKDSCQGDSSGGPLVCSI
SP5        ASIEIPVMSDRHCERSASLFG---RRVNIRATLCAGHFDGTRQSPCKGDDGGGLTCSW
Homology    : :*:::.. *.                :*** :                .:..*:*:*:* :

          310         320         330         340         350         360
          |          |          |          |          |          |
vann.tryp   TG-STYLVGIVSWGYGICARPNYPGVYAEVSYHVDWIKANA-----
fugu.trypg  -----ELQGVVSWGYGICAEERDHPGVYAKVCLFNDWLESTMASY-----
dros.sp     SGDAYQLAGIVSWGEGCAKPNAPGVYTRVGSFNDWIAENTRDACSCAQPEAAGEPASPME
mus.dist    NS-SWTQVGITSWGIGICARPYRPGVYTRVPTYVDWIQRILAENHSDAYGYHSSASAAAYQM
SP5        NG-NHYLVGVAGEQFGECFVA-----
Homology    *::: *                .

```

	370	380
vann.tryp	-----	
fugu.trypg	-----	
dros.sp	TTEQGDQENTTANGAAEADPEVEEANKLI	
mus.dist	LLPVLLAVALPGSL-----	
SP5	-----	
Homology		

Appendix 10 Multiple alignment using CLUSTAL W of SP6 with the most similar proteins identified by BLAST. Asterisks below the sequence indicates positions where all the sequences share the same amino acid residue, two dots indicates conserved amino acid substitutions, one dot indicates semi-conserved amino acid substitutions. **Homo.entk**; human enteropeptidase precursor Acc No P98073. **Sus.entk**; *Sus scrofa* enteropeptidase precursor Acc No P98074. **Mus.rik**; *Mus musculus* riken Acc No AK004939.

```

          10      20      30      40      50      60
          |      |      |      |      |      |
homo.entk  MGSKRGISSRHHSLSSEYIMFAALFAILVVLCAGLIAVSLTIKESQRGAALGQSHEARA
sus.entk   MGSKRIIPSRHRSLSTYEVMTALFAILMVLCAGLIAVSWLTIKSEKDAALGKSHEARG
SP6
-----
mus.rik
-----
Homology

          70      80      90      100     110     120
          |      |      |      |      |      |
homo.entk  TFKITSGVTYNPNLQDKLSVDFKVLAFDLQQMIDEIFLSSNLKNEYKNSRVLQFENGSI
sus.entk   TMKITSGVTYNPNLQDKLSVDFKVLAFDIQQMIGEIFQSSNLKNEYKNSRVLQFENGSI
SP6
-----
mus.rik
-----
Homology

          130     140     150     160     170     180
          |      |      |      |      |      |
homo.entk  VVFDLFFAQVWSDQNVKEELIQGLEANKSSQLVTFHIDLNSVDILD-----
sus.entk   VIFDLLFAQVWSDENIKEELIQGIEANKSSQLVAFHIDVNSIDITESLENYSTTSPSTTS
SP6
-----
mus.rik
-----
Homology

          190     200     210     220     230     240
          |      |      |      |      |      |
homo.entk  -KLTTSHTLATPGNVSIECLPGSSPCTDALTCIKADLFCDEVNCPDGSDEDNKMCATVC
sus.entk   DKLTSSPPATPGNVSIECLPGRPCADALKCIAVDLFCDEGNCPDGSDEDSKICATAC
SP6
-----
mus.rik
-----MPTTEVPQAADGQGDAGDGEEAAEPEGKFKPPKN
Homology

          250     260     270     280     290     300
          |      |      |      |      |      |
homo.entk  DGRFLLTGSSGSFQATHYKPKSETSVVCQWIIRVNQGLSIKLSFDDFNTYYTDILDIEYEG
sus.entk   DGKFLLESSGSFDAAQYPKLSEASVVCQWIIRVNQGLSIELNFSYFNTYSMDVLNIEYEG
SP6
-----
mus.rik
-----TKRKNRDYVRFTPLLLVLAAVSAGVMLWYFLGYKAEVTVSQVYSGSLRVLNRHFSQDLG
Homology

          310     320     330     340     350     360
          |      |      |      |      |      |
homo.entk  VGSSKILRASIWETNPGTIRIFSNQVTATFLIESDESDYVGFNATYTAFNSSSELNNYEKI
sus.entk   VGSSKILRASLWLMNPGTIRIFSNQVTVTFLIESDENDYIGFNATYTAFNSTELNNDKI
SP6
-----
mus.rik
-----RRESIAFRSESAKAQKMLQELVASTRLGTYYNSSSVYSFGEGPLTCFFWFILDIPYQRL
Homology

          370     380     390     400     410     420
          |      |      |      |      |      |
homo.entk  NCNFEDGFCFWVQDLNDDNEWERIQQSTFSPFTGPNFDHTFGNASGFYISTPTGPGGRQE
sus.entk   NCNFEDGFCFWIQDLNDDNEWERIQQSTFPPFTGPNFDHTFGNASGFYISTPTGPGGRQE
SP6
-----
mus.rik
-----TLSPEVVRELLVDELLSNSSTLASYKTEYEVDPEGLVILEASVNDIVVLNSTLGCYRYSY
Homology

```

```

          430      440      450      460      470      480
          |        |        |        |        |        |
homo.entk RVGLLSLPLDPTLEPACLSFWYHMYGENVHKLSINISNDQNMEKTVFQKEGNYGDNWNYG
sus.entk  RVGLLSLPLEPTLEPVCLSFWYYMYGENVYKLSINISNDQNIKIIIFQKEGNYGENWNYG
SP6      -----
mus.rik   VNPGQVLPLKGPDQQTTSCLWHLQGPEDLMIKVRLEWTRVDCRDRVAMYDAAGPLEKRLI
Homology

```

```

          490      500      510      520      530      540
          |        |        |        |        |        |
homo.entk QVTLNETVKFKVAFNAFKNKILSDIALDDISLTYGICNGSLYPEPTLVPTPPPELPTDCG
sus.entk  QVTLNETVEFKVAFNAFKNQFLSDIALDDISLTYGICNVSLYPEPTLVPTSPPELPTDCG
SP6      -----
mus.rik   TSVYGCQRQEPVMEVLASGSVMAVVWKKGMHSYYDPFLLSVKSVAFQDCQVNLTLLEGRLD
Homology

```

```

          550      560      570      580      590      600
          |        |        |        |        |        |
homo.entk GPFELWEPNTTFSSTNFPNSYPNLAFCVWILNAQKGKNIQLHFQEFDLNINDVVEIRDG
sus.entk  GPFELWEPNTTFTSMNFPNNYPNQAFVWNLNAQKGKNIQLHFEEFDLENIADVVEIRDG
SP6      -----
mus.rik   TQGFRLTP----YPSYSPSTHCSWHLTVPSLDYGLALWFDAYALRRQKYNRLCTQGQW
Homology

```

```

          610      620      630      640      650      660
          |        |        |        |        |        |
homo.entk EEADSLLLAVYTGPGPVKDFVSTTNRMVLLITNDVLRGGFKANFTTGYHLGIPEPCKA
sus.entk  EEDSLLAVYTGPGVEDVSTTNRMVLFITNDALTKGGFKANFTTGYHLGIPEPCKE
SP6      -----
mus.rik   MIQNRRLCGFRTLQPYAERIPVASDGVITINFTSQISLTGPGVQVYYSLYNQ--SDPCPG
Homology

```

```

          670      680      690      700      710      720
          |        |        |        |        |        |
homo.entk DHFQCKNGECVPLVNLCDGHLHCEGDSDEADCVRFFNGTTNNNGLVRFRIQSIWHTACAE
sus.entk  DNFQCKNGECVLLVNLCDGFHCKDGSDEAHCVRFLNGTANNGLVQFRIQSIWHTACAE
SP6      -----
mus.rik   EFLCSVNGLCVP---ACDGIKDCPNGLDERNCVCR-----AMFQCQE
Homology

```

```

          730      740      750      760      770      780
          |        |        |        |        |        |
homo.entk NWTQISNDVCQLLGLGSGNSSKPIFSTDGGPFVKLNTAPDGHILITPSQQCLQDSLIRL
sus.entk  NWTQTSDDVCQLLGLGTGNSSMPFSSGGGPFVKLNTAPNGSLILTASEQCFEDSLILL
SP6      -----
mus.rik   DSTCISLPRVCDRQPDCLNGSDEEQCQEGVPCGTFTFQCEDRSCVKKPNPECDGQSDCRD
Homology

```

```

          790      800      810      820      830      840
          |        |        |        |        |        |
homo.entk QCNHKSCGKKLAAQDITPKIVGGSNAKEGAWPWVVGGLYGGRLLCGASLVSSDWLVSAAH
sus.entk  QCNHKSCGKKQVAQEVSPIVGGNDSREGAWPWVVALYNGQLLCGASLVSRDWLVSAAH
SP6      -----
mus.rik   GSDEQHCDG--LQGLSSRIVGGTVSSEGEWPWQASLQIRGRHICGGALIADRWVITAAH
Homology

```

:.*:

	850	860	870	880	890	900
homo.entk	CVYGRNLEPSKWTAILGLHMKSNLTSPQTVPRLIDEIVINPHYNRRRKDNDIAMMHLEFK					
sus.entk	CVYGRNLEPSKWKAILGLHMSTNLTSPQIVTRLIDEIVINPHYNRRRKDSDIAMMHLEFK					
SP6	ANTNVPRRPQ-----TILKNEIVYRTDKRIVIHPEFVFPHYDVALIEVDRAFD					
mus.rik	CFQEDSMASPKLWTVFLGKMRQNSRWPGEVVSFKVSRLFLHPYHEEDSHDYDVALQLDHP					
Homology	. . . * * : : .					
	910	920	930	940	950	960
homo.entk	VNYTDYIQPICLPEENQVFPPGRNCSIAGWGTVVYQGTTANILQEADVPLLSNERCQQQM					
sus.entk	VNYTDYIQPICLPEENQVFPPGRICSIAGWGKVIYQGGSPADILQEADVPLLSNEKCCQQM					
SP6	VTG-VFVRPVCLPN-GEYPEAGKRCYTTGFGTLEYKGDVSPSLQQVDLPIISHSTCSQLY					
mus.rik	VVYSATVRPVCLPARSHFFEPGQHCWITGWGAQREGGPVSNLQKVDVQLVPQDLCEAY					
Homology	* :*:*** .. .*: * :*: * : **:.*: :.:. *.:					
	970	980	990	1000	1010	1020
homo.entk	PEY--NITENMICAG-YEEGGIDSCQGDSGGPLMCQ--ENNRWFLAGVTSFGYKCALPNR					
sus.entk	PEY--NITENMMCAG-YEEGGIDSCQGDSGGPLMCL--ENNRWLLAGVTSFGYQCALPNR					
SP6	RKVGWNLINYLQLCAGNLTHGGVDSCQGDSGGPLVCQRCSNCNWYLAGVTSFGRGCALPEF					
mus.rik	RYQ---VSPRMLCAG-YRKGKKDACQGDSGGPLVCRE-PSGRWFLAGLVSWGLGCGRPNF					
Homology	: :*** . * *:*****:* . . * ***:.*: * . *:					
	1030	1040	1050			
homo.entk	PGVYARVSRFTEWISFLH-----					
sus.entk	PGVYARVPKFTEWISFLH-----					
SP6	PGVYMSVKHIERWIETITQMYASSNKTCQPILEWKW					
mus.rik	FGVYTRVTRVINWIIQQVLT-----					
Homology	*** * :. .**:					

Appendix 11 Multiple alignment using CLUSTAL W of SP7 with the most similar proteins identified by BLAST. Asterisks below the sequence indicates positions where all the sequences share the same amino acid residue, two dots indicates conserved amino acid substitutions, one dot indicates semi-conserved amino acid substitutions. **Human.meg7**; human multiple epidermal growth factor protein 7 Acc No BAA32468. **Mouse.LDLR**; *Mus musculus* low density lipoprotein receptor Acc No AF247637. **Mouse.a2mrec**; *Mus musculus* alpha2-macroglobulin receptor Acc No Q69219. **Chick.coagX**; *Gallus gallus* coagulation factor X Acc No P25155. **Carp.masp**; *Cyprinus carpio* mannan associated serine protease Acc No AB009073.

	10	20	30	40	50	60
human.MEGF7	-----					
mouse.LDLR	-----					
mouse.a2mrec	MLTPPLLLLVLPLLSALVSGATMDAPKTCSPKQFACRDQITCISKGWRCDCGERDCPDGSDE					
SP7	-----					
Chick.coagX	-----					
carp.masp	-----					
Homology	-----					
	70	80	90	100	110	120
human.MEGF7	-----					
mouse.LDLR	-----					
mouse.a2mrec	APEICPQSKAQRCPPEHNSCLGTELCVPMRLCNGIQDCMDGSDEGAHCRELRANCSRMG					
SP7	-----					
Chick.coagX	-----					
carp.masp	-----					
Homology	-----					
	130	140	150	160	170	180
human.MEGF7	-----					
mouse.LDLR	-----					
mouse.a2mrec	CQHHCVPTPSGPTCYCNSSFQLEADGKTCKDFDECSVYGTCSQLCTNTDGSFTCGCVEGY					
SP7	-----					
Chick.coagX	-----					
carp.masp	-----					
Homology	-----					
	190	200	210	220	230	240
human.MEGF7	-----					
mouse.LDLR	-----					
mouse.a2mrec	LLQPDNRSCAKNEPVD RPPVLLIAN SQNILATYLSGAQVSTITPTSTRQT TAMDFS YAN					
SP7	-----					
Chick.coagX	-----					
carp.masp	-----					
Homology	-----					
	250	260	270	280	290	300
human.MEGF7	-----					
mouse.LDLR	-----					
mouse.a2mrec	ETVCWVHVGD SAAQTQLKCARMPGLKGFVDEHTINISLSLHHVEQMAIDWLTGNFYFVDD					
SP7	-----					
Chick.coagX	-----					
carp.masp	-----					
Homology	-----					

```

          310      320      330      340      350      360
          |        |        |        |        |        |
human.MEGF7 -----
mouse.LDLR -----
mouse.a2mrec IDDRIFVFCNRNGDTCVTLLELYNPKGIALDPAMGKVFFTDYGGQIPKVERCDMDGQNR
SP7 -----
Chick.coagX -----
carp.masp -----
Homology -----

```

```

          370      380      390      400      410      420
          |        |        |        |        |        |
human.MEGF7 -----
mouse.LDLR -----
mouse.a2mrec KLVDSKIVFPHGITLDELVSRLVYWADAYLDYIEVVVDYEGKGRQTIIQGILIEHLYGLTVF
SP7 -----
Chick.coagX -----
carp.masp -----
Homology -----

```

```

          430      440      450      460      470      480
          |        |        |        |        |        |
human.MEGF7 -----
mouse.LDLR -----
mouse.a2mrec ENLYYATNSDNANTQQKTSVIRVNRFNSTHEYQVVTRVDKGGALHIYHQRROPVRS
SP7 -----
Chick.coagX -----
carp.masp -----
Homology -----

```

```

          490      500      510      520      530      540
          |        |        |        |        |        |
human.MEGF7 -----
mouse.LDLR -----
mouse.a2mrec NDQYGKPGGCSDICLLANSHKARTCRCRSGFSLGSDGKCKKPEHELFLVYGKGRPGIIR
SP7 -----
Chick.coagX -----
carp.masp -----
Homology -----

```

```

          550      560      570      580      590      600
          |        |        |        |        |        |
human.MEGF7 -----
mouse.LDLR -----
mouse.a2mrec GMDMGAKVPDEHMIPIENLMNPRALDFHAETGFIYFADTTSYLIGRQKIDGTERETILKD
SP7 -----
Chick.coagX -----
carp.masp -----
Homology -----

```

```

          610      620      630      640      650      660
          |        |        |        |        |        |
human.MEGF7 -----
mouse.LDLR -----
mouse.a2mrec GIHNVEGVAVDWMGDNLWYTDGPKKTIISVARLEKAAQTRKTLIEGKMTHPRAIVVDPLN
SP7 -----
Chick.coagX -----
carp.masp -----
Homology -----

```

	670	680	690	700	710	720
human.MEGF7	-----					
mouse.LDLR	-----					
mouse.a2mrec	GWMYWTDWEEDPKDSRRGRRLERAWMDGSHRDI FVTSKTVLWPNGLSLDIPAGRLYWVDAF					
SP7	-----					
Chick.coagX	-----					
carp.masp	-----					
Homology	-----					

	730	740	750	760	770	780
human.MEGF7	-----					
mouse.LDLR	-----MRRWWGALLLG					
mouse.a2mrec	YDRIETILLNGTDRKIVYEGPELNHAFGLCHHGNYLFWTEYRSGSVYRLERGVAGAPPTV					
SP7	-----					
Chick.coagX	-----					
carp.masp	-----					
Homology	-----					

	790	800	810	820	830	840
human.MEGF7	-----					
mouse.LDLR	ALLCAHG-----IASSLECACGRSHFTCAVSALGECTCIPAQWQCDGDND					
mouse.a2mrec	TLRSERPPIFEIRMYDAHEQQVGTNKCRVNNGGCSSLCLATPGSRQCACAEDQVLDTDG					
SP7	-----					
Chick.coagX	-----					
carp.masp	-----					
Homology	-----					

	850	860	870	880	890	900
human.MEGF7	-----					
mouse.LDLR	CGDHSDEDEGCTLPPTCSPLDFHCDNGKCI RRSWVCDGDND CEDDSDEQ--DCPPRECEEDE					
mouse.a2mrec	VTCLANPSYVPPPQCQPGQFACANNRCIQERWKCDGDNDCLDNSDEAPALCHQHTCPSDR					
SP7	-----					
Chick.coagX	-----					
carp.masp	-----					
Homology	-----					

	910	920	930	940	950	960
human.MEGF7	-----					
mouse.LDLR	FPCQNGYCI RSLWHCDGDND CGDNSDEQ---CDMRKCS DKEFRCS DGSCIAEHWYCDGDT					
mouse.a2mrec	FKCENNRCI PNRLCDGDND CGNSEDESNATCSARTCPPNQFSCASGRICIPISWTCDLDD					
SP7	-----					
Chick.coagX	-----					
carp.masp	-----					
Homology	-----					

	970	980	990	1000	1010	1020
human.MEGF7	-----					
mouse.LDLR	DCKDGSDEESCP SAVSPPCNLEEFQ CAYGR CILDIYHCDGDDDCGDWSDESDCSSHQPC					
mouse.a2mrec	DCGDRSDESAS--CAYPTCFPLTQFTCNNGRCININWRCDNDND CGDNSDEAGCS--HSC					
SP7	-----MKYIFVAFLSILCCASSFDYKCS-----					
Chick.coagX	-----					
carp.masp	-----					
Homology	-----					

```

                1030      1040      1050      1060      1070      1080
                |        |        |        |        |        |
human.MEGF7  -----
mouse.LDLR  RSGEFMCDSGLCINSGWRCDGDADCDQSDERNCTTS-----MCTAEQFRC-RSGRC
mouse.a2mrec SSTQFKCNSGRCIPEHWTCDGDNDGCDYSDETHANCTNQATRPPGGCHSDEFQCPLDGLC
SP7         -RAGSRCHFPPFLPS--TNQTYHECPPYRQSALWCVVN-----
Chick.coagX -----
carp.masp   -----
Homology

```

```

                1090      1100      1110      1120      1130      1140
                |        |        |        |        |        |
human.MEGF7  -----LCNGVNDCGDNS
mouse.LDLR  VRLSWRCDGEDDCADNSDEENCENTGSPQCASDQFLCWN-GRCIGQRKLCNGINDCGDSS
mouse.a2mrec IPLRWRCGDGTDGDCMDSSEKSCGVTHVCDPNVKFGCKDSARCISKAWVCDGSDCEDNS
SP7         -----RDGRLVPTICVP-----CLGDGECYVKA
Chick.coagX -----
carp.masp   -----
Homology

```

```

                1150      1160      1170      1180      1190      1200
                |        |        |        |        |        |
human.MEGF7  DE-----SPOQNCRPRTGE----ENCNVNNGGCAQK
mouse.LDLR  DE-----SPOQNCRPRTGE----ENCNVNNGGCAQK
mouse.a2mrec DEENCEALACRPPSHPCANNTSVCLPPDKLCDGKDDCGDGSDEGELCDQCSLNNGGCSHN
SP7         ND-----FPSQFTLECP-----LCAFVTANLWGT
Chick.coagX -----
carp.masp   -----
Homology

```

```

                1210      1220      1230      1240      1250      1260
                |        |        |        |        |        |
human.MEGF7  CQMVVG-AVQCTCHTGYRLTEDGHTCQDVNECAEEGYCSQGCTNSEGAFQCWCETGYELR
mouse.LDLR  CQMVVG-AVQCTRHTGYRLTEDGRTCQDVNECAEEGYCSQGCTNTEGAFQCWCCEAGYELR
mouse.a2mrec CSVAPGEGIVCSCPLGMELGSDNHTCQIQSYCAKHLKCSQKCDQNKFSVKCSCYEGWVLE
SP7         NVYSNN---SFVCSAIHAGIYPATVGGTIKRIDRPASYTGSPRNALRSKTI ISSHTAFR
Chick.coagX -----
carp.masp   -----
Homology

```

```

                1270      1280      1290      1300      1310      1320
                |        |        |        |        |        |
human.MEGF7  PDRRSCKALGP-EPVLLFANRIDIRQVLPHRSEYTLNLENAIALDFHHRRELVFWS
mouse.LDLR  PDRRSCKALGP-EPVLLFANRIDIRQVLPHRSEYTLNLENAIALDFHHRRELVFWS
mouse.a2mrec PDGETCRSLDPFKLFIIFSNRHEIRRIDLHKGDYSVLVPLRNTIALDFHLSQSALYWTD
SP7         PTRIPTPSFPG--LIYTVGNKIEVISSTRRHDKLTLVSE-PNRIISVDLDRNFVFWII
Chick.coagX -----
carp.masp   -----MELTRVIVILAQCWVWPLWTQVIHLT
Homology

```

```

                1330      1340      1350      1360      1370      1380
                |        |        |        |        |        |
human.MEGF7  VTLDRILRANLNG---SNVEEVVSTG---LESPGGLAVDWVHDKLYWTD SGTSRIEVAN
mouse.LDLR  VTLDRILRANLNG---SNVEEVVSTG---LESPGGLAVDWVHDKLYWTD SGTSRIEVAN
mouse.a2mrec AVEDKIYRGKLLDNGALTSFEVVIQYG---LATPEGLAVDWIAGNIYWVESNLDQIEVAK
SP7         PNTRQIMKATFSDDYT-SVTDTSVLQGPTSVNKPIQLSYDWVHEVIYWTD AHSVRVAMTT
Chick.coagX -----
carp.masp   DIYGTIKSPNFPES-----YPKEIDLQWN---ITVPDGYQIRLYFMH
Homology

```

```

                1390      1400      1410      1420      1430      1440
                |        |        |        |        |        |
human.MEGF7    LDGAHRKVLLWQNLEKPRALHPMEGTIYWTDWGN-TPRIEASSMDGSGRRIIADTH--
mouse.LDLR     LDGAHRKVLLWQSLEKPRALHPMEGTIYWTDWGN-TPRIEASSMDGSGRRIIADTH--
mouse.a2mrec   LDGTLRRTLLLAGDIEHPRALDPRDGILFWTDWDASLPRIEAASMSGAGRRTIHRETGS
SP7            SN-HITFLINRGSQYQPDIAIQVDPESGYVYISDTGS-SPKIEKCSMGNPDSRTLVASEN-
Chick.coagX    -----
carp.masp      FD-----IEPSYLCEYDYLVKVYSDSEELAVFCGKENTDTERVPADN--
Homology

```

```

                1450      1460      1470      1480      1490      1500
                |        |        |        |        |        |
human.MEGF7    LFWPNGLTIDYAGRRMYWVDAKHHVIERANLDGSHRKAVIS--QGLPHPPFAITVFEDSLY
mouse.LDLR     LFWPNGLTIDYAGRRMYWVDAKHHVIERANLDGSHRKAVIS--QGLPHPPFAITVFEDSLY
mouse.a2mrec   GGCANGLTVDYLEKRILWIDARSDAIYSARYDGSQHMEVLRGHEFLSHPPFAVTLYGGEVY
SP7            VQOPTALTIESSTSKVYWFDSSTKTLNMCHSSGTDCTVILSSNKIINFPPVGMFLNDNKVY
Chick.coagX    -----MAGRLLLLLLCAALPDELRAEG-----GVFIKK-----
carp.masp      -----VITS---PRNVLSVAFRSDFSNEERYSG-----FEAHFSAADVD
Homology
                .: . . . *

```

```

                1510      1520      1530      1540      1550      1560
                |        |        |        |        |        |
human.MEGF7    WTDWHTKSINSANKFTGKNQEIIRNKLHFPMDIHTLHPQRQPAGKNRCGDNNGG--CTHL
mouse.LDLR     WTDWHTKSINSANKFTGKNQEIIRNKLHFPMDIHTLHPQRQPAGKNRCGDNNGG--CTHL
mouse.a2mrec   WTDWRTNTLAKANKWTGHNVTVVQRTNTQPFDLQVYHPSRQFMAPNPFCEANGGRGFCSHL
SP7            WIDAGDLTIKSVNQRTGERLHLSAAGLHRPSSIKSLDQLNQPMVRKRCQHSDCP----HF
Chick.coagX    --ESADKFLERTKRAN-----SFLEEMKQCNIERECNEERCS-----
carp.masp      ECRDRNDHRQDLHFFCHNYIG----GFYCSCRYGFLHSDNRTCKVECNESMYT-----
Homology
                : : : *

```

```

                1570      1580      1590      1600      1610      1620
                |        |        |        |        |        |
human.MEGF7    CLP-SGQNYTCACPTGFRKISSHACAQSLDKFLLFARRMDIRRISFDTEDLSDDV-IPLA
mouse.LDLR     CLP-SGQNYTCACPTGFRKINSHACAQSLDKFLLFARRMDIRRISFDTEDLSDDV-IPLA
mouse.a2mrec   CLINYNRTVSWACPHLMKIHKDNNTCYEFKKFLLYARQMEIRGVLDLAPYYNYITISFTVP
SP7            CLP-AGRAYRCVCP--YNVPSCNHTFQOSNVQIFIADDDIIRRLNVNMLTGSITGDVIKR
Chick.coagX    -----KEEA-----R-----
carp.masp      -----ERSGE-----
Homology

```

```

                1630      1640      1650      1660      1670      1680
                |        |        |        |        |        |
human.MEGF7    DVRSVALDWDSDRDDHVYWTDVSTDTISRAKWDGTGQEVVVDTSLESPAGLAIDWVTKL
mouse.LDLR     DVRSVALDWDSDRDDHVYWTDVSTDTISRAKWDGTCQEVVVDTSLESPAGLAIDWVTKL
mouse.a2mrec   DIDNVTVLDYDAREQRVYWSDVRTQAIKRAFINGTGIVTVVSADLPNAHGLAVDWVSRNL
SP7            GLINAADVAYSSITSKLYWS----N-----ETSSQSRITKKEGVAVDWIHHNL
Chick.coagX    ----EAFEDN--EKTEEFWN-----
carp.masp      -ITSADFPEPYPKTSDCTYH-----
Homology
                :

```

```

                1690      1700      1710      1720      1730      1740
                |        |        |        |        |        |
human.MEGF7    YWTDAGTDR--IEVANTDGSMTVLIWEN--LDRPRDIVVEPMGGYMYWTDWGASPKIER
mouse.LDLR     YWTDAGTDR--IEVANTDGSMTVLIWEN--LDRPRDIVVEPMGGYMYWTDWGASPKIER
mouse.a2mrec   FWTSYDTNKKQINVARLDGSFKNAVVGQ---LEQPHGLVVHPLRGKLYWTDG---DNISM
SP7            YWTDATHNKVMLAFGDEGNLENISTLIERNVSVYRPRALDPLKGYMYISDIGSNPKIEK
Chick.coagX    -----
carp.masp      -----
Homology

```

```

                1750      1760      1770      1780      1790      1800
                |        |        |        |        |        |
human.MEGF7    AGMDASGRQVIISSNLTWPNGLAIDYGSQRLYWADAGMKTIEFAGLDGSKRKVL--IGSQ
mouse.LDLR    AGMDASSRQVIISSNLTWPNGLAIDYGSQRLYWADAGMKTIEFAGLDGSKRKVL--IGSQ
mouse.a2mrec  ANMDGSNHTLLFSG-QKGPVGLAIDFPESKLYWISSGNHTINRCNLDGSELEVIDTMRSQ
SP7           CWMDGEHCMIIVDENIQLPNGIALDFTTQKMFWDGRLKTLFSFNFDGSNRTILLDDSTL
Chick.coagX   -----IYVDGDQCSSNPCHYGGQCKDGL-----
carp.masp     -----IELEEGFQITLFDDTFDIEDHPEVTCYPYDFIKIHAGDK-----
Homology

```

```

                1810      1820      1830      1840      1850      1860
                |        |        |        |        |        |
human.MEGF7    LPHPFGLTLYGERIYWTDWQTKSIQSADRLTGLDRETLQENLENLMDIHVFHR---RRPP
mouse.LDLR    LPHPFGLTLYGQRIYWTDWQTKSIQSADRLTGLDRETLQENLENLMDIHVFHR---QRPP
mouse.a2mrec  LGKATALAIMGDKLWWADQVSEKMGTCNKADGSGSVVLRNSTTLMHMKVYDESIQLEHE
SP7           IGQAYGIGVFYNRVFWTDLTNSALFTISKTPPVRRAIMTGLVEGKGIKLIQYN--QPQ
Chick.coagX   -----G--SYTCSCLDG-----YQ-----GKNCEFVIP---KY-
carp.masp     ---VFG-----PFCGEQSPGKIQTGSNIVN----ILFHSdstgenlgwkltytstgse
Homology

```

```

                1870      1880      1890      1900      1910      1920
                |        |        |        |        |        |
human.MEGF7    VSTPCAMENGGCSHLCLRSPNPSG---FSCCTPTGINLLSDGKTCSPGMNSFLIFARRID
mouse.LDLR    VTTLCAVENGGCSIHLCLRSPNPSG----FSCCTPTGINLLRDGKTCSPGMNSFLIFARRID
mouse.a2mrec  GTNPCSvNNGDCSQLCLPTSETT---RSCMCTAGYSLRS-GQQACEGVGSFLLYSVHEG
SP7           GDNVCAESSD--CSICVPVPHTTNTTRSSCVCPDHLRVAQSNRPECRKHT---LTCRRG
Chick.coagX   ----CKINNGDCEQFCSIKKSvQKD--VVCSTSGYELAEDGKQCVSKVK-----
carp.masp     CSPLAAPLNGLHLEPLQSNYIFKDH---ITLTCDPGYSLRQ-GDKEFEHYQ----IECQRD
Homology      . . . : * : .

```

```

                1930      1940      1950      1960      1970      1980
                |        |        |        |        |        |
human.MEGF7    IRMVSLDIPYFADVVPINITMKNTIAVGVDPEQEGKVYWSDSLHRI SRANLDGSQHEDI
mouse.LDLR    IRMVSLDIPYFADVVPINMTMKNTIAIGVDPLEGKVYWSDSLHRI SRASLDGSQHEDI
mouse.a2mrec  IRGIPLDPNDKSDALVPVSGTS-LAVGIDFHAENDTIYVWDMGLSTISRAKRQDTWREDV
SP7           LQPDANYTGCVDIDECVTNTHLCEQICFNMGSYLICI RDDFTLNLDGRSCYDAGCSSSP
Chick.coagX   -----YPCGK---VLMKRIKRSVILPTNSNTNATSDQDV
carp.masp     GKWSSDVPLCKMVD CGPVDVVLGEVIFESFGNSTVFGSRIQYSCRDS PQVNNTYTCHQSG
Homology

```

```

                1990      2000      2010      2020      2030      2040
                |        |        |        |        |        |
human.MEGF7    ITTGLQTTDGLAVDAIGRKVYWTDGTNRIEVGNLDGSMR--KVLVWQNLDSPRAIVLYH
mouse.LDLR    ITTGLQTTDGLAVDAIGRKVYWTDGTNRIEVGNLDGSMR--KVLVWQNLDSPRAIVLYH
mouse.a2mrec  VTNGIGRVEGIAVDWIAGNIYWDQGFVIEVARLNGSFR--YVVISQGLDKPRAITVHP
SP7           CMN--G---GLCSDVANGSYSYTCQCLQGFRGLCNEIYSSMNVVLSGNEVSFEICVPRV
Chick.coagX   PST-----NGS-ILEEVFT-----TTTESPTPPP
carp.masp     EWVS-----EDGTPLPTCLPG-DFETTLS-----VNAESQLPTP
Homology

```

```

                2050      2060      2070      2080      2090      2100
                |        |        |        |        |        |
human.MEGF7    EMGFMYWTDWGENAKLERSGMDGSDRAVLINNNLGPWNGLTVDKASSQLLWADAHTERIE
mouse.LDLR    EMGFMYWTDWGENAKLERSGMDGSDRTVLINNNLGPWNGLTVDKTS SQLLWADAHTERIE
mouse.a2mrec  EKGYLFWTEWGHYPRIERSRLDGTERTVVLVNVSISWPNGISVDYQGGKLYWCDARMDKIE
SP7           VTSVIKWYRN-----RERITAMRSTFLTFNGRLLRIFFVTYLEAGNYKCFEYGGLE
Chick.coagX   -----R-----N-----GSS
carp.masp     LTS-----TP-----LA
Homology

```

	2110	2120	2130	2140	2150	2160
human.MEGF7	AADLN-GANRHTLVSP-----					
mouse.LDLR	VADLN-GANRHTLVSP-----					
mouse.a2mrec	RIDLETGENREVVLSSNNMDMFSVSVFEDFIYWSDRTHANGSIKRGCKDNATDSVPLRTG					
SP7	YASTH-----					
Chick.coagX	ITDPN-----					
carp.masp	CGEQS-----					
Homology	.					

	2170	2180	2190	2200	2210	2220
human.MEGF7	-----					
mouse.LDLR	-----					
mouse.a2mrec	IGVQLKDIKVFNRDRQKGTNVCAVANGGCQQLCLYRGGGQRACACAHGMLAEDGASCREY					
SP7	-----					
Chick.coagX	-----					
carp.masp	-----					
Homology	-----					

	2230	2240	2250	2260	2270	2280
human.MEGF7	-----					
mouse.LDLR	-----					
mouse.a2mrec	AGYLLYSERTILKSIHLSDERNLNAPVQPFEDPEHMKNVIALAFDYRAGTSPGTPNRIFF					
SP7	-----					
Chick.coagX	-----					
carp.masp	-----					
Homology	-----					

	2290	2300	2310	2320	2330	2340
human.MEGF7	-----					
mouse.LDLR	-----					
mouse.a2mrec	SDIHFGNIQQINDDGSGRTTIVENVGSVEGLAYHRGWDTLYWTSYTTSTITRHTVDQTRP					
SP7	-----					
Chick.coagX	-----					
carp.masp	-----					
Homology	-----					

	2350	2360	2370	2380	2390	2400
human.MEGF7	-----					
mouse.LDLR	-----					
mouse.a2mrec	GAFERETVITMSGDDHPRAFLVLEDCQNLMFWTNWNELHPSIMRAALSGANVLTLLIEKDIR					
SP7	-----					
Chick.coagX	-----					
carp.masp	-----					
Homology	-----					

	2410	2420	2430	2440	2450	2460
human.MEGF7	-----VQHPYGLTLLDSYIYWTDW					
mouse.LDLR	-----VQHPYGLTLLDSYIYWTDW					
mouse.a2mrec	TPNGLAIDHRAEKLYFSDATLDKIERCEYDGSRYVILKSEPVHFPGLAVYGEHIFWTDW					
SP7	-----FLEVPVQISAV					
Chick.coagX	-----					
carp.masp	-----					
Homology	-----					

	2470	2480	2490	2500	2510	2520
human.MEGF7	QTRSIHRADKGTGSNVILVRSNLPG--LMDMQAVDRAQPLGFNKCGRNNGGCSHLCLPRP					
mouse.LDLR	QTRSIHRADKSTGSNVILVRSNLPG--LMDIQAVDRAQPLGFNKCGRNNGGCSHLCLPRP					
mouse.a2mrec	VRRAVQRANKYVGSMDKLLRVDIPQQPMGIIAVANDTNSCELSPCRINNGGCQDLCLLTH					
SP7	CGQAPSIPDRIGG-R---ITSGVPT-----APFDGPFIAMLVEET					
Chick.coagX	-----VDTRIVGG-----D-----ECRPGECPWQAVLINEK					
carp.masp	QLFPAQQKRIVGG-----RTASPGLFPWQVLLSVED					
Homology		*				:

	2530	2540	2550	2560	2570	2580
human.MEGF7	S-----					
mouse.LDLR	S-----					
mouse.a2mrec	QGHVNCSCRGGRILQEDFTCRAVNSSCRAQDEFECANGECISFSLTCDGVSHCKDKSDEK					
SP7	N-----					
Chick.coagX	G-----					
carp.masp	VSR-----					
Homology						

	2590	2600	2610	2620	2630	2640
human.MEGF7	-----					
mouse.LDLR	-----					
mouse.a2mrec	PSYCNSRRCKKTFRQCNNGRCVSNMLWCNGVDYCGDGSDEI PCNK TACGVGEFRCDGSC					
SP7	-----					
Chick.coagX	-----					
carp.masp	-----					
Homology						

	2650	2660	2670	2680	2690	2700
human.MEGF7	-----					
mouse.LDLR	-----					
mouse.a2mrec	IGNSSRCNQFVDCEDASDEMNC SATDCSSYFRLGVKGVLFQPCERTSLCYAPSWVCDGAN					
SP7	-----					
Chick.coagX	-----					
carp.masp	-----					
Homology						

	2710	2720	2730	2740	2750	2760
human.MEGF7	-----					
mouse.LDLR	-----					
mouse.a2mrec	DCGDYSDERDCPGVKRPRCPLNYFACPSGRCI PMSWTCDKEDDCENGEDETHCNKFCSEA					
SP7	-----					
Chick.coagX	-----					
carp.masp	-----					
Homology						

	2770	2780	2790	2800	2810	2820
human.MEGF7	-----					
mouse.LDLR	-----					
mouse.a2mrec	QFECQNHRCISKQWLCDGSDDCGDGSDEAAHCEGKTCGPSSFSCPGTHVCVPERWLCGDG					
SP7	-----					
Chick.coagX	-----					
carp.masp	-----					
Homology						

```

                2830      2840      2850      2860      2870      2880
                |        |        |        |        |        |
human.MEGF7  -----
mouse.LDLR  -----
mouse.a2mrec KDCTDGADESVTAGCLYNSTCDDREFM CQNRLCIPKHFVCDHDRDCADGSDESPECEYPT
SP7        -----
Chick.coagX -----
carp.masp   -----
Homology    -----

```

```

                2890      2900      2910      2920      2930      2940
                |        |        |        |        |        |
human.MEGF7  -----
mouse.LDLR  -----
mouse.a2mrec CGPNEFR CANGRCLSSRQWECDGENDCHDHSDEAPKNPHCTSPEHKCNASSQFLCSSGRC
SP7        -----
Chick.coagX -----
carp.masp   -----
Homology    -----

```

```

                2950      2960      2970      2980      2990      3000
                |        |        |        |        |        |
human.MEGF7  -----GFSCACPTGIQLKGDG
mouse.LDLR  -----SFSCACPTGIQLKGDG
mouse.a2mrec VAEALLCNGQDDCGDGS DERGCHVNECLSRKLSGCSQDCEDLKIGFKRCRCPGFRLKDDG
SP7        -----EGSE
Chick.coagX -----E
carp.masp   -----VPED
Homology    -----

```

```

                3010      3020      3030      3040      3050      3060
                |        |        |        |        |        |
human.MEGF7  KTCD-----PSPETYLLFSSRG
mouse.LDLR  KTCD-----PSPETYLLFSSRG
mouse.a2mrec RTCADLDECSTTFPCSQLCINTHGSYKCLC VEGYAPRGGDPHSCKAVTDEEPFLIFANRY
SP7        -----
Chick.coagX EFCG-----
carp.masp   RWFG-----
Homology    -----

```

```

                3070      3080      3090      3100      3110      3120
                |        |        |        |        |        |
human.MEGF7  SIRRISLDTSDHTDVHVPPELNNVISLDYDSVDGKVYYTDVFLD--VIRRADLNGSNM-
mouse.LDLR  SIRRISLDTDDHTDVHVPV PGLNNVISLDYDSVHGKVYYTDVFLD--VIRRADLNGSNM-
mouse.a2mrec YLRKLNLDGSNYT---LLKQGLNNAVALAFDYREQMIYWTGVTTQGS MIRRMLHNGSNV-
SP7        -----G--SIATR-----NKIITA AHCLQNDEINITSVHVF---VGKVLTDVTLI-
Chick.coagX -----GTILN-----EDFILTA AHCINQSKE--IKVVVG---EVDREKEE----
carp.masp   -----SGALLS-----STWVLTAAHVLRSHRRDFSVV PVAS-EHIRVHLGLTDIR
Homology    .                :: .                *

```

```

                3130      3140      3150      3160      3170      3180
                |        |        |        |        |        |
human.MEGF7  -ETVIGRGLKTTDGLAVD WVARNL YWTD TGRNTIEASRLD GSCRKVLINN-----SLDEP
mouse.LDLR  -ETVIGHGLKTTDGLAVD WVARNL YWTD TGRNTIEASRLD GSCRKVLINN-----SLDEP
mouse.a2mrec -QVLHRTGLSNPDGLAVD WVGGNLYWCDKGRDTIEVSKLNGAYRTVLVSS-----GLREP
SP7        -EPYQQHSLVSHVVFHENYDPDNLNSDIA ILTSTQIVFTKAVKPLCIPL-----HTDTN
Chick.coagX --HSETHTHTAEKIFVHSKYIAETYDNDIA LIKLKEPIQFSEYVVPACL P-----QADFA
carp.masp   DKHLATNRSVAKVILHPQFDPQNYNNDIA LIKLSQEVVLSALI QPVCLPRPGVKGHTLMP
Homology    . . .                :                :

```

	3190	3200	3210	3220	3230	3240
human.MEGF7	RAIAVFPRKGYLFWTDWGHIAKIERANLDGSEKVLINTDLGWPNGLTLDYDTRRIYWVD					
mouse.LDLR	RAIAVFPRKGYLFWTDWGHIAKIERANLDGSEKVLINTDLGWPNGLTLDYDTRRIYWVD					
mouse.a2mrec	RALVVDVQNGYLYWTDWGDHSLIGRIGMDGSGRSIIVDTKITWPNGLTVDYVTERIYWAD					
SP7	QDIKPRPYRG---TSKMGLVLGYGRTSHRGPVSTQLREVLVEIR---TQQFCTQRYRTVD					
Chick.coagX	NEVLMNQKSG-----MVSGFGREFEAGRLSKRLKVLEVPYVD--R--ST--CKQSTN					
carp.masp	LPNTLGIVAGWGINTANTSASTSGLTSDLGTVSELLQYVKLPVQPQ-DECEASYASRSVN					
Homology	*		*	:	:	::

	3250	3260	3270	3280	3290	3300
human.MEGF7	AHLDRIESADLNGKLRQVLVG-HVSHPFALTQQDRWIYWTDWQTKSIQRVDKYSGRNKET					
mouse.LDLR	AHLDRIESADLNGKLRQVLVS-HVSHPFALTQQDRWIYWTDWQTKSIQRVDKYSGRNKET					
mouse.a2mrec	AREDYIEFASLDGNSNRHVLSQDIPHIFALTLFEDYVYWTDWETKINRAHKTTGANKTL					
SP7	KEVTSVMFCAGGG--AQDACSGDSGGPFALWSNRTQSWWLAGIVSWGPRGCGVS--NLPG					
Chick.coagX	FAITENMFCAGYETEQKDACQGDSSGPHVTRYKD--TYFVTGIVSWGEGCARKG---KYG					
carp.masp	YNITSNMFCAGFYEGGQDTCLGDSGGAFVTQDARSGRWVAQGLVSWGGPEECGS-QRVYV					
Homology	.	:	.	..	:	..

	3310	3320	3330	3340	3350	3360
human.MEGF7	VLANVEGLMDIIVVSPQRQTGTN--ACGVNNGGCTHLCFAR-----					
mouse.LDLR	VLANVEGLMDIIVVSPQRQTGTN--ACGVNNGGCTHLCFAR-----					
mouse.a2mrec	LISTLHRPMDLHVHALRQPDVPHPCVKVNNGGCSNLCLLSPGGGHKACPTNFYLGDDG					
SP7	VYTRIGTSMRQWIHNH-----					
Chick.coagX	VYTKLSRFLR-WVRTVMRQK-----					
carp.masp	VYTRVANYIH-WLHRHMDGEEVAKV-----					
Homology	:	:	:	:	:	:

	3370	3380	3390	3400	3410	3420
human.MEGF7	-----ASDFVC-----					
mouse.LDLR	-----ASDFVC-----					
mouse.a2mrec	RTCVSNCTASQFVCKNDKCI PFWWKCDTEDDCGDHSDEPPDCPEFKCRPGQFQCSTGICT					
SP7	-----					
Chick.coagX	-----					
carp.masp	-----					
Homology	-----					

	3430	3440	3450	3460	3470	3480
human.MEGF7	-----					
mouse.LDLR	-----					
mouse.a2mrec	NPAFICDGDNDQCQDNSDEANCDIHVCLPSQFKCTNTNRCIPGIFRCNGQDNCGDGEDERD					
SP7	-----					
Chick.coagX	-----					
carp.masp	-----					
Homology	-----					

	3490	3500	3510	3520	3530	3540
human.MEGF7	-----					
mouse.LDLR	-----					
mouse.a2mrec	CPEVTCAPNQFQCSITKRCIPRVWVCDRDNHCVDGSDEPANCTQMTCGVDEFRCCKDSGRC					
SP7	-----					
Chick.coagX	-----					
carp.masp	-----					
Homology	-----					

```

          3550      3560      3570      3580      3590      3600
          |        |        |        |        |        |
human.MEGF7 -----ACPDEPDS
mouse.LDLR -----ACPDEPDG
mouse.a2mrec IPARWKCDGEDDCGDGSDEPKKEECDERTCEPYQFRCKNNRCVPGRWQCDYDNDCGDNSDE
SP7 -----
Chick.coagX -----
carp.masp -----
Homology -----

```

```

          3610      3620      3630      3640      3650      3660
          |        |        |        |        |        |
human.MEGF7 QPCSLVPG-----
mouse.LDLR HPCSLVPG-----
mouse.a2mrec ESCTPRPCSESEFFCANGRCIAGRWKCDGDHDCADGSDEKDCTPRCDMDQFQCKSGHCIP
SP7 -----
Chick.coagX -----
carp.masp -----
Homology -----

```

```

          3670      3680      3690      3700      3710      3720
          |        |        |        |        |        |
human.MEGF7 -----
mouse.LDLR -----
mouse.a2mrec LRWPCDADADCMDGSDEEACGTGVRTCPLDEFQCNNTLCKPLAWKCDGEDDCGDNSDENP
SP7 -----
Chick.coagX -----
carp.masp -----
Homology -----

```

```

          3730      3740      3750      3760      3770      3780
          |        |        |        |        |        |
human.MEGF7 -----LVPPAPRATGMSEKSPVLP-----NTPPTTLYSSTTRTR---
mouse.LDLR -----LVPPAPRATSMNEKSPVLP-----NTLP TTLHSSTTKTR---
mouse.a2mrec EECARFICPPNRPFRCKNDRVCLWIGRQCDGVDNCGDGTDEEDCEPPTAQNPCHKDKKEF
SP7 -----
Chick.coagX -----
carp.masp -----
Homology -----

```

```

          3790      3800      3810      3820      3830      3840
          |        |        |        |        |        |
human.MEGF7 -----TSLEEVEGRCSERDARLG-----LCARSNDVAVPA
mouse.LDLR -----TSLEGAGGRCSERDAQLG-----LCAHSNEAVPA
mouse.a2mrec LCRNQRCLSSSLRCNMFDDCGDGSDEEDCSIDPKLTSCATNASMCGDEARCVRTEKAAYC
SP7 -----
Chick.coagX -----
carp.masp -----
Homology -----

```

```

          3850      3860      3870      3880      3890      3900
          |        |        |        |        |        |
human.MEGF7 APGEGLHISYAIGLLSILLILVVIAALMLYRHKKS KFTDPGMGNLTYSNPSYRTS----
mouse.LDLR APGEGLHVSYAIGLLSILLILLVIAALMLYRHRKSKFTDPGMGNLTYSNPSYRTS----
mouse.a2mrec ACRSGFHTVPGQPQCQDINECLRFGTCSQLWNKPKGGHLCSCARNFMKTHNTCKAEGSEY
SP7 -----
Chick.coagX -----
carp.masp -----
Homology -----

```

```

          3910      3920      3930      3940      3950      3960
          |         |         |         |         |         |
human.MEGF7 -----TQEVKIEAIPKPAMYNQLCYK-----
mouse.LDLR -----TQEVKLEAAPKPAVYNQLCYK-----
mouse.a2mrec QVLYIADDNEIRSLFPGHPHSAYEQTFQGDSEVRIDAMDVHVKAGR VYWTNWHTGTISYR
SP7 -----
Chick.coagX -----
carp.masp -----
Homology -----

```

```

          3970      3980      3990      4000      4010      4020
          |         |         |         |         |         |
human.MEGF7 --KEGGPDHNYTKEKIKIVEGICLLSGD-----DAEWDDLKQLRSSRGGLLRDHVCM
mouse.LDLR --KEGGPDHSYTKEKIKIVEGIRLLAGD-----DAEWGDLKQLRSSRGGLLRDHVCM
mouse.a2mrec SLPPAAPPTTSNRHRRQIDRGVTHLNI SGLKMPRGIAIDWVAGNVYWTDSGRDVIEVAQM
SP7 -----
Chick.coagX -----
carp.masp -----
Homology -----

```

```

          4030      4040      4050      4060      4070      4080
          |         |         |         |         |         |
human.MEGF7 KTDTVSIQAS-----SGSLDDTEMEQLLQEEQSEC
mouse.LDLR KTDTVSIQAS-----SGSLDDTETEQLLQEEQSEC
mouse.a2mrec KGENRKTLISGMIDEPHAI VVDPLRGTMYSWDWGNHPK IETAAMDGTLRETLVQDNIQWP
SP7 -----
Chick.coagX -----
carp.masp -----
Homology -----

```

```

          4090      4100      4110      4120      4130      4140
          |         |         |         |         |         |
human.MEGF7 SSVHTAATPERR-----GSLPDTGWKHERKLSSSESQV-----
mouse.LDLR SSVHTAATPERR-----GSLPDTGWKHERKLSSSESQV-----
mouse.a2mrec TGLAVDYHNERLYWADAKLSVIGSIRLNGTDP IVAADSKRGLSH PFSIDVFEDYIYGVTY
SP7 -----
Chick.coagX -----
carp.masp -----
Homology -----

```

```

          4150      4160      4170      4180      4190      4200
          |         |         |         |         |         |
human.MEGF7 -----
mouse.LDLR -----
mouse.a2mrec INN RVFKIHKFGHSPLYNL TGGLSHASDVVLYHQHKQPEVTN PCDRKKCEWLCLLSPSGP
SP7 -----
Chick.coagX -----
carp.masp -----
Homology -----

```

```

          4210      4220      4230      4240      4250      4260
          |         |         |         |         |         |
human.MEGF7 -----
mouse.LDLR -----
mouse.a2mrec VCTCPNGKRLDNGTCVPVSP TPPPDPAPRPGTCTLQCFNGG SCFLNARRQPKCRCQPRYT
SP7 -----
Chick.coagX -----
carp.masp -----
Homology -----

```

```

          4270      4280      4290      4300      4310      4320
          |        |        |        |        |        |
human.MEGF7 -----
mouse.LDLR -----
mouse.a2mrec SP7 GDKCELDQCWEYCHNGGTCAASPSGMPTCRCPTGFTGPKCTAQVCAGYCSNNSTCTVNQG
Chick.coagX -----
carp.masp -----
Homology -----

```

```

          4330      4340      4350      4360      4370      4380
          |        |        |        |        |        |
human.MEGF7 -----
mouse.LDLR -----
mouse.a2mrec SP7 NQPQCRCLPGFLGDRCQYRQCSGFCENFGTCQMAADGSRQCRCTVYFEGPRCEVNKCSRC
Chick.coagX -----
carp.masp -----
Homology -----

```

```

          4390      4400      4410      4420      4430      4440
          |        |        |        |        |        |
human.MEGF7 -----
mouse.LDLR -----
mouse.a2mrec SP7 LQGACVVNKQTGDVTCNCTDGRVAPSLTCLDHCSNGGSCTMNSKMMPECQCPPHMTGPR
Chick.coagX -----
carp.masp -----
Homology -----

```

```

          4450      4460      4470      4480      4490      4500
          |        |        |        |        |        |
human.MEGF7 -----
mouse.LDLR -----
mouse.a2mrec SP7 CQEQVVSQQQPGHMASILIPLLLLLLLLLVAGVVFYKRRVRGAKGFQHQRTNGAMNVE
Chick.coagX -----
carp.masp -----
Homology -----

```

```

          4510      4520      4530      4540      4550      4560
          |        |        |        |        |        |
human.MEGF7 -----
mouse.LDLR -----
mouse.a2mrec SP7 IGNPTYKMYEGGEPDDVGGLLDADFALDPDKPTNFTNPVYATLYMGHGSRHSLASTDEK
Chick.coagX -----
carp.masp -----
Homology -----

```

```

          4570
          |
human.MEGF7 -----
mouse.LDLR -----
mouse.a2mrec SP7 RELLGRRGPEDEIGDPLA
Chick.coagX -----
carp.masp -----
Homology -----

```

Appendix 12 Multiple alignment using CLUSTAL W of thioll with the most similar proteins identified by BLAST. Asterisks below the sequence indicates positions where all the sequences share the same amino acid residue, two dots indicates conserved amino acid substitutions, one dot indicates semi-conserved amino acid substitutions. **Dros.tep1**; *Drosophila melangogaster* thiolester protein 1 Acc No AJ269538. **Dros.tep2**; *Drosophila melangogaster* thiolester containing protein 2 Acc No AJ269539. **Homo.Cd109**; human cell surface antigen CD109 Acc No AF410459. **Mus.gpi**; *Mus musculus* GPI-linked protein Acc No AY083458. **Celeg.hypo**; *Caenorhabditis elegans* hypothetical protein Acc No Z82090. **Lim.a2m**; *Limulus* alpha2-macroglobulin Acc No D83196. **Rat.a1inIII**; *rattus norvegicus* alpha1-inhibitor III Acc No P14046.

	10	20	30	40	50	60
dros.tep1	-----MLWLILSSTILHCVLLSNANGLYSVLAPKTLRSNSAYNVVVAI					
dros.tep2	-----MFRIFLTGIILQYALLVNATGIYSVVGGPTLRSNSKYNVVVSV					
homo.cd109	-----MQGPPLLTAAHLLCVCTAALAVAPGPRFLVTAPGIIRPGNVITIGVEL					
mus.gpi	-----MRSRLLSAAHLLCCLCAVALAAP-GSRFLVTAPGIIRPGANVTIGVDL					
thioll1	-----MNLWRPACSLGTIYLLATLSSLATASNVYNIYFPKHIRPGFNISFTA					
celeg.hypo	MRLILNLNLFVWVQIHGVIGQSTNAAVVSTTAAVPVKPATYMLVAPAVRDPQPF					
lim.a2m	-----MEEIKWQKMSTLLFLLLLFTHDVYSKSGFILTAPKSLTPGKSNILNLHL					
rat.a1inIII	-----MKKDREAQLCLFSAALLAFLPFASLLNGNSKYMVLVPSQIYTETPEKTI					
Homology				:	:	* : . . :
	70	80	90	100	110	120
dros.tep1	HNTTRT-----TEVSVSLTGPSLNSRKYVDVQSMSSKSVRFDIPKLTEG-DYELKVM					
dros.tep2	HKADGP-----SQIKVSLNGPSYNETKQIELPPMSTQNVEFEVFKLATG-NYNLSAE					
homo.cd109	LEHCPSQVTVKAELLKTASNLTVSVLEAEGVFEKGSFKTLTLPPLNSADE-IYELRVT					
mus.gpi	LENSPPQVLVKAQVFKIASNKSRSILEAEGVFHRGHFKTLVLPALPLSSADK-IYELHIN					
thioll1	IDNPNT---VQIHTAFRSMDSFHVDSTDSVNSGSSSRISMNGLPIHYSGSH-GFELNIT					
celeg.hypo	LKQATDEDMIVRIEVRTERNETIAARVISNLKPGIAQTVSLSEMPAQSLTPRQSYKLYIR					
lim.a2m	FDIKTNG----FLRIGVKDQDDGNVVAETEVSFNKDNPSSSIQLTIPSGVEVKRPKLYAN					
rat.a1inIII	YHLNET---VTVTASLISQRGTRKLFDELVVDKDLFHCVSFTIPRLPSSEEEESLDINIE					
Homology						..
	130	140	150	160	170	180
dros.tep1	GSG-----GIEFQNSTKLSFAPDLNWLVIQSDKATYKPGDKIQFRVFLDKNTRPAVIDK					
dros.tep2	GVS-----GVVFNKSTKLNADKKPSVVFVQTDKATYKPADLVQFRILFLDENTRPAKIEK					
homo.cd109	GRTQD---EILFSNSTRLSFETKRISVFIQTDKALYKPKQEVKFRIVTLFSDFKPYKTS-					
mus.gpi	GQSEN---EIVFSNRTRLTFESKISVLIQTDKAFYKPKQEVKFRVLTLCSDLKPYRTS-					
thioll1	GTDLVTGAQLFFNSSTDFQFQAKSISILIQTDKAIYQPGHTVKFRAIALKPDLPKPLQGN-					
celeg.hypo	GETLN--AELIFENENELKYDQKALS VFIQTDRAIYRPAASLVRYRAIVVKSDLKPYVGN-					
lim.a2m	GSYSSP-SSNDFFFEKDINMHKDKLIVFVQTDKPLYKPGQTVKVRILPTTDLKLVPKET					
rat.a1inIII	GAK-----HKFSERRVVLVKNKESVVFVQTDKPMYKPGQSVKFRVVSMDKNLHPLNELF					
Homology	*	*	.	.	:	:*:*..*:* : : * : : :
	190	200	210	220	230	240
dros.tep1	PIKIEIRDGDQNLIKSWKDIKPAKGVYSGELQLSDRPFVLCNWTVTATVQDECKVTNVLVV					
dros.tep2	PISVIIIDGAQNRIKQLSDVKLTGKGVFSGELQLSEQPVLGTWKISVSVGDGNRETKEFEV					
homo.cd109	-LNILIKDPKSNLIQQWLSQQSDLGVISKTFQLSSHPILGDWSIQVQVNDQ-TYYQSFQV					
mus.gpi	-VDIFIKDPKSNVIQQWFSQKGLGVVSKTFQLSSNPIFGDWSIQVQVNDQ-QYYQSFQV					
thioll1	-ISYTFKDPGRNVVMLEPEVPLNHGVAGGQFSLTKDAVAGMVKVEFMAEGF-KESLSVEV					
celeg.hypo	-ATIKIFDPSRNLIQSQTIGVTLDRGVYSGELQLAEETLLGDWFIIEVETSNVGVQDKSSFTV					
lim.a2m	IGSFQIENPDGIVLGYWPMLSFAEGIAQFELALPDEPTYGMWRKIGNIEDT-EIYENFEV					
rat.a1inIII	-PLAYIEDPKMNRIMQWQDVKTENGLKQLSFLSAEPIQGPYKIVILKQSGVKEEHSFTV					
Homology	:	:	:	*	:	*. . * : : .. . *

	250	260	270	280	290	300
dros.tep1	DKYVVPKFEVVVLTAKNVAASAGYIRATIKARYTFKKPVKGHVVATIE-----					
dros.tep2	DKYVLPKFEVIVDTPKAVVIADKVIKATIRAKYTYGKPKVKGKATVSMERSYGYFGDLNA-					
homo.cd109	SEYVLPKFEVTLQTPLYCSMNSKHLNGTITAKYTYGKPKVKGDTVLTFLPLSFWG-----					
mus.gpi	LEYVLPKFEVTVQTPLYCSLKSQKLNQSVIAKYTYGKPKVKGSLSLTFLPLSFWG-----					
thiol1	KRYKLPKFKVEVKAPSYIHPQSTGLTIKLDKAYTFGKGVQGTGLLEVGGYQYPVYHGF					
celeg.hypo	DTYVLPKFEVNIKTSSFITIN-DDLSVFDKAYTYGKGVAGKAKVSLELPHWRWHAMVPT					
lim.a2m	KEYVLPKFEVKITPFCYLLTNADSITWKICAQYTYGQPVVECTFVAETNVVKYNWEKE---					
rat.alinIII	MEFVLPFRFGVDVKVPNAISVYDEIINVTACATYTYGKPVPGHVKISLCHGNPTFSSETKS					
Homology	: : * : * : .	:	:	* * * : : * *		

	310	320	330	340	350	360
dros.tep1	-----GSSTEQSLPIDGEVNVVEFPISATAKR-----					
dros.tep2	-----NGNKQEKTIIDVDGKGHVEFDIIHWAQRGQYLP-----					
homo.cd109	-----KKKNITKTFKINGSANFSFNDEEMKNVMDSS-----NGLSEYLDLSS					
mus.gpi	-----KKKNITKSFENGFANFSFDNYEMKVMNLKPIITDVSREGSYENVDPSP					
thiol1	GR---FAPRPPTQNKITRRYPNFDGTVELLITNDEIREELGWNG-----AS					
celeg.hypo	IIDENGVKKEEELMVERTVKLNRRQGEAAVVSNDDELKRHKLLHEWG-----					
lim.a2m	-----GVPVIHKEGLIDGLDVTVNSSALGFNEQRLSYR-----					
rat.alinIII	G-----CKEEDSRLDNNGCSTQEVNITEFQLKENYLKMH-----					
Homology		:	*	.		

	370	380	390	400	410	420
dros.tep1	--LLKITAIVTEELTDIKHNGTAYVTVHQHRKLEDLFWPT---HYRPGVSSEFKTVVRN					
dros.tep2	--PIKLFVAVTEELTGKQONATATVVLHQQRYSIEPYERPE---HFEANKSFIYQVVVKN					
homo.cd109	PGPVEILTTVTESVTGISRNVTNFFKQHDYIIIEFFDYTT---VLKPSLNFTATVKVTR					
mus.gpi	PGPAEITATVTEESLTGISRMASNTNFFKQHDYIIIEIFDYTT---VLKPSLNFTATVKVSR					
thiol1	ESIITVTGVSVEALTEAFNDTQRIDAKTTNVKVETLVKPL---TIKPLKYSAYIQITE					
celeg.hypo	GGSIKIVASVTEDITEIERNATHQISTFREEVKLDVEKQGD---TFKPLTYNVVVALKQ					
lim.a2m	--AVNMFVAVTEKGTGKMNATDSIYRTSNPLNIMYLEPTSGKGYLKPGLPFYKGLKVEK					
rat.alinIII	-QAFHVNATVTEEGTGFSEFSGSGRIEVERTRNKFLFLKADS---HFRHGIPPFVVKVRLVD					
Homology	:	* * *	*	:	:	:

	430	440	450	460	470	480
dros.tep1	LDGSPVMDSSKMVNFNVLCCQVSKN-----FSASLQNSIAT					
dros.tep2	VDGSPVTNSAKNVKIGFDKSYSYFHEPSP-----KTRINFEAPVNGIAT					
homo.cd109	ADCNQLTLEERRNNVVIIVTQRYNTEYWSGNS-----GNQKMEAVQKINYTVPQSGTFK					
mus.gpi	SDGNQLTPEEIEIENDLVTVVTRKNNHPES---Q-----RDQEMDYIQTVNYTIPQNGI IK					
thiol1	VDGKPLPEDDRLANNLLLNIEYRYPRGEPEPGTNTTVSTWYAYRWEETRVFVIPPSTGIVK					
celeg.hypo	MDDTPVKATLPKRQVSTFYNYPNHDTSE-----SLQEEKETKIVEVDAHGTSV					
lim.a2m	PDGTPAPGEQIELCRFADRERWNRKR-----WLEEKIRACKEFTSDEAGI IK					
rat.alinIII	IKGDPIPNEQVLIKARDAGYTNATTTDQHG-----LAKFSIDTNGISD					
Homology	..					.

	490	500	510	520	530	540
dros.tep1	EHIMLPET-CQSCLVSTSTFDT-----AENIERIYIYKLN-----KPLMIAINTKPK					
dros.tep2	FNVRLPDSDSRYRIFASFDG-----SENTIGSISKFEPTP--MSREPLKIQVNTKPK					
homo.cd109	IEFPILEDSSSELQKAYFLGS-----KSSMAVHSLFKSP-----SKTYIQLKTRDENI					
mus.gpi	IEFPVMSISGELQKAYFLDG-----TSSVTVHSMFTSP-----SKTYIQLKTRDEYI					
thiol1	VTIDAPSDTFTSINFRPYTNA-----TMSQRWALQWTAERADSPSNSYLQITTEENSV					
celeg.hypo	LTLQPPINCTSARIEAHYDIGG--KDNFTATPIYSSLYVEAAVSPTKSFLLQLLADNEGAV					
lim.a2m	FTVPPQTPDITSRFRKAKALQYKKGDKDNKLNQPHSFTVSSWSYSPSGSHLQLEPITEEI					
rat.alinIII	YSLNIKVYHKEESSCIHSSCT-----AERHAEAHHTAYAVYSLSKSYIYLDTEAGVL					
Homology	.					:

```

                550      560      570      580      590      600
                |        |        |        |        |        |
dros.tep1      QLRKLLKINIISDT-----YLPYFILTIVVARGNIVLSLFQEMKE--KKK-----
dros.tep2      RLGEQVSFDVVSIE-----DLPYFVYTIIVARGNVILSDYVDVPD--GQK-----
homo.cd109     KVGSPFELVVSIGNK-----RLKELSYMVVSARGQLVAVGKQNST-----
mus.gpi        KVGSPFDLMVSGNR-----QFKDLSYMVISKGQLVAAGKQSSR-----
thiol1        VPGNMATVTIRTTE-----AVSEFTILIIISERGEILSERKFQTLGSGVPEN-----
celeg.hypo     DVGKSLSFSLKATQ-----PLSTITYQVMSRSNIVVSQQMTVN---SE-----
lim.a2m       ECGKPLTVKFKYTTG-----EEKKQKFYYQIMARNFIVDTGSFEHEFLLEDKSGSLTDET
rat.alinIII    PCNQIHTVQAHFILKGQVLGVLQQIVFHVYLVMAQGSILQGTGNHHTHQVEPGESQVQG----
Homology      . . . : : : : : : :

```

```

                610      620      630      640      650      660
                |        |        |        |        |        |
dros.tep1      -----SQEIEFEPTFALVPQATIFVHYIIDG--VL
dros.tep2      -----TYTVKFTPTPTFSMVPKATIIYVYVYVNN--DL
homo.cd109     -----MFSLTPENSWTPKACVIVYIIEDDG--EI
mus.gpi        -----TFSLTPEASWAPKACIIAYYIAEDG--EI
thiol1        -----SHLFEFSVEYDMPVGVQVLSYVRDDG--EI
celeg.hypo     -----HATISFPATANMAPKSRILIVYAIIESSQEV
lim.a2m       YLPIDVTALSLNPPNEPEWENNVIIVPPHIGETSLTLIPSEFEMNPSAKILVFYVREDG--ET
rat.alinIII    -----NFALEIPVEFSMVPVAKMLIYITILPDG--EV
Homology      . : * : : .

```

```

                670      680      690      700      710      720
                |        |        |        |        |        |
dros.tep1      MSDEKTVDIERDFENTIEILTTNEAL-PRDEVSLKVKTN-PHSFVGLLGVDQSVLLLRSG
dros.tep2      QFEEKTIDFEKEFSNSIDVSAPTNAK-PSEEVKLRIKTD-ADSFVGLLGVDQSVLLLKSG
homo.cd109     ISDVLKIPVQLVFNKIKLYWSKVKAEPSEKVSRLISVTQPDSIVGIVAVDKSVNLMNAS
mus.gpi        INDILKIPVQLVFNKIKLYWSKVKAEPSEKVSRLISATQSDSLVGVAVDKSVTLMENS
thiol1        VADYIKLTVTAELNQSITSSSTNIDAGEDVSIQVQTSSSGAYVGARAIQSVLLLKSG
celeg.hypo     LVDALDFKVEGIFQNQVALSIDKQAVEPGQNVKFKVTSQD-KNSFVGLLVVDQSVLLLKSG
lim.a2m       VADSTKITVKKCLRNKVLKFGEEKVLPASSTLQLTAS-PYSICGIGAVDKSVHILSSD
rat.alinIII    IADSVKFKQVEKCLRNKVHLSFSPSQSLPASQTHMRVTAS-PQSLCGLRAVDQSVLLQKPE
Homology      : . . : * : : . . : : : : * : * : * :

```

```

                730      740      750      760      770      780
                |        |        |        |        |        |
dros.tep1      NDLNRDLIILNNLATYSTDLVILTANANINIYRSS-----GGCYTNP
dros.tep2      NDLSQDDIFNLSLNIYQTSTPMMNGYGRYPGQTSQ-----LVTLTNA
homo.cd109     NDITMENVVHELELYNTGYLGMFMNSFAVFQECG-----LWVLTDA
mus.gpi        NSITMETMVHELELYNTEYYLGMFMNSFAVFQECG-----LWVLTDA
thiol1        NDVSQERIVTDLNKYSVTQELNHMWRWWWYPTPS-----GASDASD
celeg.hypo     NDITREKVEQDLENYDSNNVGGGFGGPRPWEAIDRKKRSIWRP-----WWGIGGSDAQS
lim.a2m       NRITEEVFNKLGGHDYYWPKQATSQDYKYCEDYKFKQTEGEHEGSFSSGFTSTNYLDSIT
rat.alinIII    AELSPSLIYDLPQMDSNFIASSNDPFEDEDYCLMYQP-----IAREKDVYR
Homology      : . . : .

```

```

                790      800      810      820      830      840
                |        |        |        |        |        |
dros.tep1      GYTNTCTGSLIGRTMFKNE-----PTKNSGPVPIVGSTRAQASLP-----
dros.tep2      NYPYNTGPLVMSYVFEGRSRHP-----WITRPRYRVGIRGDSGDRISFLSQSLNDRNLKE
homo.cd109     NLT KDYIDGVYDNAEYAER-----FMEENEGHIVDIHD-FSLGSSP-----
mus.gpi        TLIRDSIDEVYDTEEYSER-----FAEENEANLVDFED-ASSVNNV-----
thiol1        VFRKAGILVFTDALVYQKPEA-----SIYPFRPIAFSLNGGFAERNIIATAAVDTSTP-
celeg.hypo     IFSNAGLVVLTDALYREYRQ-----EFMSVMMMDGAPGMAEAAFAAPPMGGSS-----
lim.a2m       AFDEAGLVVISDMELETRPCKPSGFEDGGRPCPYDVAFAPQAANRIGGGGEAGGFGGG
rat.alinIII    YVRETGLMAFTNLKIKLPTYCNTDYDMVPLAVPAVALDSSTDRGMYESLPVVAVKSPLPQ
Homology      .

```

```

                850      860      870      880      890      900
                |        |        |        |        |        |
dros.tep1      -----PVRKLFPEWFLSNITDVGANGYEI IKETVPTDITLTSWVITGFSLS
dros.tep2      ILLKQTPQRT-----TIRKEFPETWFFENVGEE---EFTLTKKIPDTITSWVVTGFSLN
homo.cd109     -----HVRKHFPEWFLWLDTNMG-YRIYQEFVETVPDSITSWVATGFVIS
mus.gpi        -----HVRKNFPETWIWLDAYMG-SKIYEEFEVETVPDSITSWVASAFVIS
thiol1         -----ATPT-----RTRTLFPETWLWDEQISG-ADGSATFNNTAPDTITSWIFSAFVS
celeg.hypo     -----PPPP-----TVRKFPHETWIWSDLNSTSG--EVEMEIEAPDTITSWVASTFAIN
lim.a2m        IRKKTNKPVV-----EIRTYFPETWLWELQNIQ-ATGELSLKRDIPHTITWVGSATICIS
rat.alinIII    EPPRKDPPPKDPVIETIRNYFPETWIWDLVTVN-SSGVTELEMTVPDTITWVKAGALCLS
Homology      * . **.**: : : *.:*. * : :

```

```

                910      920      930      940      950      960
                |        |        |        |        |        |
dros.tep1      PQSGLAVTRNPSRIRVFQPPFITNLPYSVKRGEVIAIPVIVFNLYGMDVKAKVLMDNSD
dros.tep2      PTSGIALTKNPSKIRVFQPPFVSTNLPYSVKRGEVIAIPVIVFNLYDKTLDADVMDNSD
homo.cd109     EDLGLGLTTPVELQAFQPPFI FLNLPYSVIRGEEFALEITIFNLYKDATEVKVIEKSD
mus.gpi        EDLGFGLTTPAELQAFQPPFFLFLNLPYSVIRGEEFALEIVNLYKDTIKVILIEESD
thiol1         DQHGLGVS-EQHKVTVFRNFFITLNLVPRVIRGELIIVQAVFNLYLSTEVDVAVLTLTESN
celeg.hypo     EENGLGVAPTTSKLRVFRPFFIQLNLPYAVRRGEKFA LLVLFVFNMEKEQDVTVTLKYDK
lim.a2m        EETGLGVS-EAATVKGFQPPFVSTLFPYSVIRGEKVP IIVTVFNLYSECLPIKLSLEQSD
rat.alinIII    NDTGLGLS-SVASFQAFQPPFVELTMPYSVIRGEAFTLKATVNLNLP TSLPMAVLLLEASP
Homology      *.:.: . *: **: .:* * *** . : :.***: : : .

```

```

                970      980      990      1000      1010      1020
                |        |        |        |        |        |
dros.tep1      GQYEFIEETNKNVSQYLR-GVRRKKT LWIPANTG---RGISFMIRPKKVGLTTLKITAIS
dros.tep2      QEYEFTEATNEVLEK AID-EVRRVKRVTIPANSG---KSVSFMIRPKNVGFTTLKITATS
homo.cd109     ---KFDILMTSN--EIN--ATGHQQTLLVPS EDG---ATVLFPIRPHLGEIPITVTALS
mus.gpi        ---SFDILMTSN--DTN--GTIYRKT VQVPRDNG---VTLVFPKPHLGEIPITVTAAS
thiol1         ---KFVLLRPGN---NSA-AVGFSRRITIPASGS---VSVKFPIRMGTLGEIPITMTAIS
celeg.hypo     DS-GYDLLKKDGTVVRD-EVGQQNVRIVSVAGG GTSKAVYFPIVPS SIGEIPVHISAIA
lim.a2m        KFEMQNDTNSYTSVCVCGG-KSDTTRWMIKPRSLG--QVNLTVYGASLPNEAICGNQDYST
rat.alinIII    DFTAVPVENNQDSYCLGANGRHTSSWLVT PKSLGNVNFVSVAEARQSPGPCGSEVATVPE
Homology      . . . :

```

```

                1030      1040      1050      1060      1070      1080
                |        |        |        |        |        |
dros.tep1      KYAGDRHLHQILKVEADGVQKYVNKAVLINVQRLNRRSLAPPEKTI IIEKADNVI EGSETV
dros.tep2      ALAGDAIHQKLVKVEPEGVTLFENRAVFINLK----DQPEMSQSLDADIPNEVVPQSEFI
homo.cd109     PTASDAVTQMILVKAEGIEKSYSQSILLDLTDN---RLQSTLKTLSFSFPNTVTGSERV
mus.gpi        PTASDAVTQTIVVKPEGIEKSYSKSVLLDLTDS---NVESKQQSMRFSFPDPVTIGSERV
thiol1         EIASDALTRKVFVQPEGITQCTSGSVLFQRM DAS---APPDVESLNIQIPAGIVPGSEKV
celeg.hypo     SQGGDAVEMNLRVDPQGYKVDRNIPFVIDLNNN---SSDFSKNLELIWPN DVVDGSQKA
lim.a2m        VTARDAATRQLLVEPEGFPKEDTWFSTFACPKDQ---NGKFTATSDLLL PEDLVEDSARG
rat.alinIII    TGRKDTVVKVLIVEPEGIKKEHTFSSLLCAS-----DAELSETLSLLL PPTVVKDSARA
Homology      * : *.:* . . . . : *

```

```

                1090      1100      1110      1120      1130      1140
                |        |        |        |        |        |
dros.tep1      EFEVCGTSQAPQLEHLDDLVLHPCGCGEQNMFNVP SILALSYLKAKNRQDQEIENKAKR
dros.tep2      EFSVVDLLGPTLQNLNDLVRMPYGCGEQNMVNFV PNILVLKYLEVTGRKLPSVESKARK
homo.cd109     QITAIGDVLGPSINGLASLIRMPYGCGEQNM INFAPNIYILDYLTKKKQLTDNLKEKALS
mus.gpi        QITAIGDILGSSINGLSSLIRMPYGCGEQNM IYFAPNIYILDYLTQKQLTVNLKEKALS
thiol1         KLLVYGDILGSTMNNLGSLLRTPSGCGEQNMLGFAPDVFV TLYLHLSAGKLDAA TRAKAFK
celeg.hypo     RLDVIGDMMGPVLNNAHKL VQMPYGCGEQNM LNLVNPILVVKYLRATNRNESQLETKA I K
lim.a2m        YV SITGDLMPA IKNLDHLVRLPTGCGEQNMVKFV PNI FVLDYLTATGSITDSI KEKALN
rat.alinIII    HFSVMGDILSSAIKNTQNLIQMPYGCGEQNM VLFAPNIYVLKYLNETQQLTEKI KSKALG
Homology      . * .. : : *.: * ***** . :.:. : ** . **

```

	1150	1160	1170	1180	1190	1200
dros.tep1	YVETGYQIELNYKRNDGSFSAWQHDALG	--STWLTAYVIRSFHQAAKY	--IDIDKNVLV			
dros.tep2	FLEIGYQRELTQYKHDDGSYSAFGKSDASG	--STWLTAYVMRSFHQAGTY	--TDIDPKVIT			
homo.cd109	FMRQGYQRELLYQREDGSFSAFGNYDPSG	--STWLSAFVLRFCLEADPY	--IDIDQNVLH			
mus.gpi	YMRQGYQRELLYQREDGSFSAFGDIDSSG	--STWLSAFVLRFCLEADYY	--IDIDQDVLH			
thiol1	HFQTGYSNELNYKHRDGSFSAFGECDASG	--STWLTAFAAKCFMFARELRPTLVASVID				
celeg.hypo	FIEQGIQRELTQYKRADNSFSAFGDSDKAG	--STWLTAFVVRSFHHAKQY	--AFVDPNVIS			
lim.a2m	NMRKGYARQQNYRHPDGSYSAFGNRDKQG	--NLFLTAFVYRSFAQAERF	--ILINKNKLN			
rat.alinIII	YLRAGYQRELNKYKDKGSYSAFGDHNGQGQNTWLTAFVLKSFQAQARAF	--IFIDESHIT				
Homology	.. *	: **: *	.*:***:.. :	* . :*:***:..*	* * . :. . :	

	1210	1220	1230	1240	1250	1260
dros.tep1	AGLDFLVSQRQSTDGKFKELGMVIHNSHGS	----PLALTSFVLLTFFENEEMPKYKHVID				
dros.tep2	AGLDFLVSQKQKESGEFPEVGKLFNANQN	----PLALTSFVLLAFFENHELIPKYQSAIK				
homo.cd109	RTYTWLKGHQKSNGEFWDPRVHSELQ	----GNKSPVTLTAYIVTSLLGKRYQPNID				
mus.gpi	RTYTWLNAHKKFNGEFWEPRVHSELQ	----GTKSPVTLTAYIVTSLVGKRYQPNID				
thiol1	QALTFLINQQNTTGTFRPEPRVHSELQ	----GQDGGVALTAFVLISILENGME				
celeg.hypo	RAVAFNLNSQMESGAFGERGEVHHKDMQG	-----GAQDGGVALTAFVLISILENGME				
lim.a2m	ETENWILNRQRSNGCFRKIGKLFNSALKGG	ISSNDETPAPLTAYVLISLLEAGYKNETV				
rat.alinIII	DAFTWLSKQQKDSGCFRSGSLLNNAMKGGVDEITLSAYITMALLESSLPDTDPVVSKA					
Homology	:: :	* * . * :	.. .			

	1270	1280	1290	1300	1310	1320
dros.tep1	RAVEFVVTEVHQSN	----EPYDLAIAALALSLARN	-RNAYKVLDKLKLATRRGDHKWWT			
dros.tep2	KAVRYVAEEADKTD	----DQYSLAIAAVALQLAKH	-PQSEKVIAKLESVARKENDRMWWS			
homo.cd109	VQESIHFLESEFSRGIS	-DNYTLALITYALSSVGS	-PKAKEALNMLTWRAEQEGGMQFVW			
mus.gpi	VQDSIKFLEFEFSRGIS	-DNYTLAIIISYALSTVGS	-PKAEEALNLLMQRSEKEGDTQFWL			
thiol1	AENARIYLENHLTISD	-NKYALAIIVTYALHVAGS	-SRANEALLALEALATVQGGFKFWH			
celeg.hypo	NGKAVTYLEKHLDEVSG	-NAYTMVVAYALQLAKS	-KQAGKAFENLKKHKIVEKSGDVKF			
lim.a2m	IDQGISCLEALSNP	----STYSLALFAYATSLAGH	-PSAKDYLAKLEERAITEGGKTFWK			
rat.alinIII	LSCLESSWENIEQGGNGSFVYTKALMAYAFALAGNQEKRNEILKSLDKKAIKEDNSIHWE					
Homology	*	* * * : * .	. : *	. .		

	1330	1340	1350	1360	1370	1380
dros.tep1	GSDKCK	-----SSEVETTSYVLLALLEHNISD	-----EPKPIVDWLISK			
dros.tep2	KATESTGEDGR	---VFHWKPRSNDEITSYVLLALLEKDPAE	-----KALPIIKWLISQ			
homo.cd109	SSEKLSDS	-----WQPRSLDIEVAAYALLSHFLQFQTS	-----EGIPIMRWLSRQ			
mus.gpi	SSGPAISGS	-----WQPRSDIEIAAYALLAHTLHHVS	-----EGIPVMRWLIQQ			
thiol1	DNSESPDSYSSRWRPYYNPPNTDIEMSAYALLTYVRRNDLN	-----AGIPVMKWLASK				
celeg.hypo	ASAQKKVEKLKESRAYMFQARPVDIETTSYAVLSYLAQNQTS	-----ESLSIIRWLVSQ				
lim.a2m	SPSSGR	-----YYWGNISIGVEIAGYAVLTLQHGASN	---LAKVTPPIIRWLAKQ			
rat.alinIII	RPQKPTKSEG	---YLYTPQASSAEVEMSAYVVLARLTAQPAPSPEDLALSMGTIKWLTKQ				
Homology		. : * : . * : * :			: * * :	

	1390	1400	1410	1420	1430	1440
dros.tep1	RNSNGGFVSSQDQTVVGMALTKYELQSHASTEADIEFWHLN	-EDKKHVRVTKENEFKVQ				
dros.tep2	RNSNGGFSSQDQTVIGLQALTKFAYKTGSGSGTMDIEFSSAG	-ESKNTIKVNPENSLVLQ				
homo.cd109	RNSLGGFASTQDQTVVVALKALSEFAALMNTERTNIQVTVTGPSSSPVKFLIDTHNRLLLQ					
mus.gpi	RNSLGGFVSTQDQTVVVALKALSEFSALVHKENTDIQLTVTGPPIPRSIHFRIDSQNLFLH					
thiol1	RSSLGGYSGTQDQTVIAIQALSQVAGLLVGNTQNLQISASHSNDPPTASYNINRENSIVFN					
celeg.hypo	RNELGGFTSTQDQTVMALQALSSYAAYTSDKHTSQVTILNGK	-HTHSFDINIRNAIVLQ				
lim.a2m	QNYRGGFYSTQDQTVIALQAMSKFATIIYKDELDLEVGVESSG	-FEKKIMLTKDNSILMQ				
rat.alinIII	QNSYGGFSSQDQTVVALDALSKYGAATFSKSKQTPSVTVQSSGFSQKQVQDKSNRLLLQ					
Homology	.. **:	:.***:..:	*:..:		: * : . :	

```

                1450      1460      1470      1480      1490      1500
                |        |        |        |        |        |
dros.tep1      THQLPENT-NEVKLLAKGQGGAQVQLTYRYNVATKEARPSFKLTTTVKSHKGRLLILGIC
dros.tep2      THDLPKST-RKVDFTAKGTGSAMVQLSYRYNLAEKEKKPSFKVTPTKDTPNQLLIVDVC
homo.cd109     TAELAVVQPMAVNISANGFGFAICQLNVVYNVKASGSSRRRRSIQNQEAFDLDVAVKENK
mus.gpi        QEELHALDPITVNVSAHGSGFAICQLNVDYNVKSGSGSKRRRSTENQEVFDLDVIVN-NE
thiol1         SVNVPADV-GTVQVTATGVGVAVAQISVCYNTPNQP-----YEIEPFQCTNTVSTA
celeg.hypo     SYQLSSLN-DAVSINANGTGVVFAQLSYSYRDSLN-----DDAPFFCSQEIKEIRA
lim.a2m        TFRLQTVP-SPVDFEATGSGCGLVQTSRLRYNVNTPPPRKGFHLEVTVKRGLYRDCINAH
rat.alinIII    QVSLPYIP-GNYTVSVSVEGCVYAQTTLRYNVPLEKQQPAPALKVQTVPLTCNNPKGQNS
Homology       :          . . * * * . *

```

```

                1510      1520      1530      1540      1550      1560
                |        |        |        |        |        |
dros.tep1      GTYTPIAASERNKT-----TNMALMQVQLPSGYVCDIEPFADIEAISDVKRVETKNEDE
dros.tep2      AEYVPLEDADKDKD-----SNMAVMEIALPSGFVGDSTSLGKIQAADRVRKRVETKNSDST
homo.cd109     DDLNHVLDLNVCTSFSGPGRSGMALMEVNLVSGFMVPSAISLSETVKKVEYDHGK-----
mus.gpi        DDISHLNLNVCTSHLGSERTGMVLMVNLVSGFSASSDSIPLSETLKKVEYDNGK-----
thiol1         LKKAKVNWCCSLRP-GDNATGMFLMEVNLVPSGYTVNIDNERTRNPSAKLVEIDGNG----
celeg.hypo     GNRLQLDLCCNYTR--PGKSNMALAEIDALSGYRFDQVHTLTSIEDLQRVEMEKDDTK
lim.a2m        ATCVKYDGGKGVSN-----MAVLEMKMGVSGWIPDEESIKNIVDREELNLRREYVDGNQ
rat.alinIII    FQISLEISYMGSRP----ASNMMVIADV KMLSGFIPLKPTVKKLERLGHVSRTEVTNN--
Homology       * : :: **:          :
```

```

                1570      1580      1590      1600      1610      1620
                |        |        |        |        |        |
dros.tep1      VHIYFEKLSPGDRKCLTLEAIYTHAVANLKPSWVRLYDYYATERSATEFY--HVDTSLCD
dros.tep2      VVVYFDSLTPGDVRCLEASKAHAVAKQKPAVSLSLYDYDTERKATEYY--QVKSSLCD
homo.cd109     LNLVLDLSDVNETQ-FCVNIPAVRNFKVSNTQDASVSIVDYYPQAVRSYNSEVKLSSCD
mus.gpi        LNLVLDLSDVNESQ-FCVNIPTVRDYKVSNI RDGVSVM DYEPQAVRSYNTQVKLSSCY
thiol1         VNVYYDELAPGRSVCADIELLNLGNVGGSKARKVAASDYYQPKERVEALYQVDEAPVVC
celeg.hypo     MNVYFNPLGGRP-VCLSLYSDVTYQVADQK PANFRLVDYDPEEQKMTYAAKQTRSLQE
lim.a2m        LNLVYFSELTDQN-LCFNFWLEQDIEVQETKPATIRLYDYEYELQEVVTSYSIDENCEKLP
rat.alinIII    VLLYLDQVTNQ-TLSFSFIQQDIPVKNLQPAIVKVYDYYETDEVAFAEYSSPCSSDDQN
Homology       : :* . : . : * : . *** . *
```

```

                1630      1640      1650      1660      1670      1680
                |        |        |        |        |        |
dros.tep1      ICHGNECGNMC-----
dros.tep2      ICEGADCGEGCKKD-----
homo.cd109     LCSDVQGCRCPCEDGASGSHHSSVIFIFCFKLLYFMELWL-----
mus.gpi        LSPDTN-CKSHTDGATDSLRRSSLLVFCVLLYFVQH-----
thiol1         SCSTEDIAVCSVCADCVGCPGPAFTQWSEWSDCAFCGRSTSFRTRECRSPFSDNLAGHVC
celeg.hypo     KCGEDCWPPISPSPLPFDESTVTGTSSGFGAKWCALIIAVLLIA-----
lim.a2m        PLP-----
rat.alinIII    V-----
Homology

```

```

                1690      1700      1710      1720      1730      1740
                |        |        |        |        |        |
dros.tep1      -----
dros.tep2      -----
homo.cd109     -----
mus.gpi        -----
thiol1         GGVDRESRRCVATFFPCPDTFDGLWFMNPRNFPSSNSVPFYAHQCRMERGSRQIREQIPGI
celeg.hypo     -----
lim.a2m        -----
rat.alinIII    -----
Homology

```

	1750	1760	1770	1780	1790	1800
dros.tep1	-----					
dros.tep2	-----					
homo.cd109	-----					
mus.gpi	-----					
thiol1	ALSGSQYLTCNNYDVNPNNNYTFSILVKPNRFRSSGPTTIFSYGMEHNYARAHLEKVVWR					
celeg.hypo	-----					
lim.a2m	-----					
rat.alinIII	-----					
Homology	-----					

	1810	1820	1830	1840	1850	1860
dros.tep1	-----					
dros.tep2	-----					
homo.cd109	-----					
mus.gpi	-----					
thiol1	SELRFKVRSDTGMREVRGVSSNLLRTDQWNHIVVAVPSGDGDDIRMFVNGNAVGSTKSFT					
celeg.hypo	-----					
lim.a2m	-----					
rat.alinIII	-----					
Homology	-----					

	1870	1880	1890	1900	1910	1920
dros.tep1	-----					
dros.tep2	-----					
homo.cd109	-----					
mus.gpi	-----					
thiol1	TRYFGKHGRNRFFLGQNTRGNAWARGYFQGGLAAVGTWRSVLTDQQITALYEAYRPAIES					
celeg.hypo	-----					
lim.a2m	-----					
rat.alinIII	-----					
Homology	-----					

	1930	1940	1950	1960	1970
dros.tep1	-----				
dros.tep2	-----				
homo.cd109	-----				
mus.gpi	-----				
thiol1	SDPLSVKLLRHFAVQQLLFCFQSPATIEDLYSRSAAPVTCPTAPISPLMPFLPIL				
celeg.hypo	-----				
lim.a2m	-----				
rat.alinIII	-----				
Homology	-----				

Appendix 13 Multiple alignment using CLUSTAL W of thioll with CiC31 (accession number AJ320542) and CiC32 (accession number AJ320543) from *C. intestinalis*. Asterisks below the sequence indicates positions where all the sequences share the same amino acid residue, two dots indicates conserved amino acid substitutions, one dot indicates semi-conserved amino acid substitutions. Specificity defining residues are indicated in bold. Levels of identity of thioll to CiC31 and CiC32 are 29.31% and 26.87% respectively.

```

          10      20      30      40      50      60
CiC31x1  MVWFSFSLLVTLAVATAFDHTVVVVPKALRVDADAKIIVNLHGYNRATITGYLQDLPLGLQ
CiC32x2  -----
Thioll   MNLRWRPACSLGTIYLLATLSSLATASNVYNIYFPKHIRPGFNISFTAAIIDNPNTVQIH
Homology

          70      80      90      100     110     120
CiC31x1  TFFSRTGQRVLTTPAQCNPIEMTFRVTRDPOGADGIASFGLTQKVRLTIQVTNSNSDFTE
CiC32x2  -----ISAQQSNTISIIYIVSKNIGRKVEVVISCARATFTFTK
Thioll   TAFRSMDNSFHVDSTDSVNSGSSSRISMNGLPIHYSGSHGFELNITGTDLVITGAQLFFNS
Homology          .          *  . : : . : : *..

          130     140     150     160     170     180
CiC31x1  NIDVLVSKQSGYIYVITDRPIYKPNDTVKISAFLLNQNMGHQTGVDAEITIQTPDGIGLV
CiC32x2  RVQVLIDRNSGYLQVQDRPIYRPNERVEIRTYPLQDMSPEKNAMVQVIVKTPDGIGVN
Thioll   STDFQFQAKSISILIQTDKAIYQPGHTVKFRAIALKPKDLKP-LQGNISYTFKDPKRVVVM
Homology  :. . . :*  : : **:.**:*.. *:: : *:::  . .: * *  :
```

```

          190     200     210     220     230     240
CiC31x1  RESFVLDLQSNRLNHEFAINENPIYGTWSIEVKFSSDGYTTSSSTTSFKIDKYVLPITFDVAL
CiC32x2  KVERQLPASGFIDTTFHVAEHPMFGTWVQAKYVSKAFTTTAEATFAVRKYVVPFTFNQVL
Thioll   LEPEVPLNHGVAGGQFSLTKDAVAGMWKVEFMAEG----FKESLSVEVKRYKLPKFKVEV
Homology  . . * : : : * * : : . . . : : * : * . * . *
```

```

          250     260     270     280     290     300
CiC31x1  QLAQSHILTS DPRITGTIYANYSYGEVNGVYLSATLQKLPGGPAIKFYQIPPRITRTA
CiC32x2  ELERNYILISDSHITGKIMANYSYGLPVSGNYFLSMKLRKQNGEPQEFYKMPGNITLKA
Thioll   KAPS-YIHPQSTGLTIKLDKAYTFGKGVQGTGLLEVGGYQYPVYHGFGRFAPRPPTQN
Homology  :  :*  ... :*  . : *:::* *.. * . : : : * . . *
```

```

          310     320     330     340     350     360
CiC31x1  LFRNGVKPFSIPLVELIGILQPGETLAQMSDQAVVSI FATVNGEADGMESAVISNIPI
CiC32x2  NFRNGIQRFNVSVNLLLNLNPFGETIADLAEMGATFCVEATVNSRADSIMESDVTADLQF
Thioll   KITRRYPNFDGTVLELLITNDEIREELGWNGASESITTVTGSVT-EALTRAEAFNITQRIIA
Homology  : . * . : : * : : * : . . : : : * . *
```

```

          370     380     390     400     410     420
CiC31x1  LRTPYRIDKSRALKYHTPGISYLLQVDVQDVVTHQNMANIPVRIEITGPNQVRVIRNSTA
CiC32x2  LKSPFIIDTSITSKYIIPPVAYTLQGIIVTDAISLTPKQDVIRIRISVASS-----TYTTTT
Thioll   KTTNVKVETLVKPLTIKPLKYSAYIQITEVDGKPLPEDDRLANLLLN-----IEY
Homology  :  :. . * : * : : . : : : . :
```

	430	440	450	460	470	480
CiC31x1						
CiC32x2						
Thio11						
Homology						
	NQNGQVNYPHNFDTTGTQFIKVT	TAKPNLGAANQATVNI	TVEAYNSPGQRFL	TVTPNPHT		
	NSNGKFVTAFLVGNQVVR	IQTLDPDISNEQQARV	NLTIQPYRSPTSSYL	QILASRHT		
	RYP-----R-----	-----GEPEPGTN--	T-TVSTWYAYRWEET	RVFVIPP	SGIV	
	.		. * : .	* . *	: : . .	.
	490	500	510	520	530	540
CiC31x1						
CiC32x2						
Thio11						
Homology						
	VDVGREHVITLAFNQPTPAE	IRFYVVS	RGSVVLVGRIVPPAN	SHNIVQIIQVTQAMV	PSM	
	VTVRRTFQLTFTFGSSRP	TDIRYVVARGGIVLSS	VVRLTPLNRQKTIN	VWPSQAMVPFA		
	KVTIDAPSDTFTSINFR	PYTNATMSQRWALQ	WTAERADSPSNSYL	QITTEENS	VVPGNMA	
	.	* : .	*	: .
	550	560	570	580	590	600
CiC31x1						
CiC32x2						
Thio11						
Homology						
	RVIAYFLHGAEVVNSAFV	NVNRCE	TETVTNTRNTVKPG	APITYTIEGAQNAD	VLLYG	
	RVVAYYFKDNEVVSGSL	WFDVVDQCKRELS	IEVAP-LVTPGATFP	ITIS-APHALVRL	SG	
	TVTIRTTEAVSEFTILI	ISRGEILSERKFQ	TLSGVPENSHLFE	FSVEYDMPGVQ	VLASY	
	* :	. . . :	. . .	* .
	610	620	630	640	650	660
CiC31x1						
CiC32x2						
Thio11						
Homology						
	VDRAAYFLYNGSRLTRNS	MFSDMAAYDQGC	VSNNGSDGPNVFF	GAGLTLTTP	THKPGALD	
	VDKAAAYLYNGSRLTRD	VMFKRMESH	DQGCVRNGGEDWN	HVFMGAGLS	LYTSEQNPTIV	
	VRDDGEIVADYIKLTV	TAELENQVSITSS	STNIDAGEDVSI	RVQTS	SSGAYVGARAI	DQS
	*	.	: : **	. .	:
	670	680	690	700	710	720
CiC31x1						
CiC32x2						
Thio11						
Homology						
	TLDC	TAAQSRKKRQIQVE	QELS-IEAKLHKCG	EDGKKKSFSEC-D	TCEMRDRVIY	TFD
	SLNCDRNSRNKR	NVELEDSVLLTSL	NRKLQCCRRD	GKRDAFVN--E	TCEMRTARCG	INYG
	VLLLKSGNDV	SQERIVTDLNKYS	VTQELNHMWR	WWWYPTPSGAS	DASDVFRKAG	ILVFT
	*	: : :	: : :	: : :	: : : : :	:
	730	740	750	760	770	780
CiC31x1						
CiC32x2						
Thio11						
Homology						
	DAIPGCSERFYAECIAL	ARLNSG-----	TRRQRVQGRS	IGVNGQVER-----		
	DQYPGCCEVFHQSCV	QASLQNSGNEGE	AASVAQKRAS	FQGDSSPF	AVEVEQPEG	AILSQ
	DALVYQKPEAS	IYFPRPIAFSLN	-----GGFA	ERNIIATA	AVDTS-----	
	*	.	.	.	* :	
	790	800	810	820	830	840
CiC31x1						
CiC32x2						
Thio11						
Homology						
	-----	-----	-----	-----	-----	-----
	VSRPAVQSLGQSLAP	TRLPVFAVQAARPP	FLAMAH	TFLGMAPQPM	AHRFV	PNPFILSSEP
	-----	-----	-----	-----	-----	-----
	850	860	870	880	890	900
CiC31x1						
CiC32x2						
Thio11						
Homology						
	-----	-----	-----	-----	-----	-----
	TPPAT	TTTTTTTTTTTTTTTT	TTTTTTTTTTTT	TMTT	TATLPPVANN	PGRERSNFQERLSWPTIRIRPNG
	-----TP-----	-----	-----	-----	-----	-----ATPTRTRTLFPETWLWDEQISGADG
	*

	910	920	930	940	950	960
CiC31x1						
CiC32x2						
Thiol1						
Homology						
	RRQITAKARDSITTYEIDAMASADTPDGFCIAPTNNVKVFKNVFVQVYTPYSLKKREQAL	HITSYKTARDSITTFVVGAVGMQDSPDGFCIAPTKEMKVKDFVQINLPYSIRKLEQAQ	SATFNTTAPDTITSWIFSAFSVSDQH-GLGVSEQHKVTVFRNFFITLNLVPRVIRGELII	. * * : * * : : . . * . . * * : : : . . * * : : * : : * : : *		
	970	980	990	1000	1010	1020
CiC31x1						
CiC32x2						
Thiol1						
Homology						
	IKLSVFNY-GDTLVTVDIMMRAHPVLCTHFRTDGSYDLVRTISVGPNSAGSASFVLPPLR	LKITIFNYNAQNNYTLRLHAKTDDTFCTTFKSG-TWAQLGTFNIEAGGFASAPLTVIPLL	VQAIVFNY--LSTEVDVAVLTLTESNKFVLLRPGNNSAAVGFSSRRITIPASGSVSVKFPPIR	:: : * * * . . : : . . : : . . : : . . : : . . : : * :		
	1030	1040	1050	1060	1070	1080
CiC31x1						
CiC32x2						
Thiol1						
Homology						
	IPAGDIPKATVEVYITDNRNRTYDSVKKEILIEDEGELKDIYETFPIDLKNR--VQQTQ	IPLTGRSPIQLKVVN-DQTNVIQDSIRRVLIEPAGEMKDTYNSYPIDLSTG--NQSIE	MGTLGEIPITMTAIS----EIASDALTRKVFVQPEGITQCTSGSVLFQRMASAPPDVES	: . : . : . : * : : : : : * : : : : . : : . : .		
	1090	1100	1110	1120	1130	1140
CiC31x1						
CiC32x2						
Thiol1						
Homology						
	INFTFPEQFVLGTRKCMLYAYMDFMGPAIEVDPVTQEANNVNSLIRQPYGCGEQTMIIYAG	IGLNFPEQINLESRKWIYAYASYMGPSIEVNQITQEPRSIASIFRQPYGCGEQNMLVTG	LNIQIPAGIVPGSEKVKLLVYGDILG-----STMNNLGSLLRTPSGCGEQNMLGFA	:: : * : : : * : : * . : * . : * . : * . : * . : * . : * . : *		
	1150	1160	1170	1180	1190	1200
CiC31x1						
CiC32x2						
Thiol1						
Homology						
	PTVFALQYLVTGTITPNSPEYNSAVNKIEAAFQREMYRTHHTNPRVWSVFTHYLPSTW	PNVYAHMFLVTTGKMAPGSIRYEQSTRMEDGFNQMRYSRLVGVKRAWSVFSHYRPSTW	PDVFTLYLHSAGKLDAAATR--AKAFKHFQGTGYSNELNYKHRDGSFSAFEGEDASG-STW	* * : . : * : : * : : . : . : : : : : * : . . . : * * * * *		
	1210	1220	1230	1240	1250	1260
CiC31x1						
CiC32x2						
Thiol1						
Homology						
	LNAFVDKVFYHGRRYD-TDMNVGPICNSLNFLIGEOMAGEHFRERRPPLHREMHGAVKGP	LNAYVDRVFIQAQVYY-TEMDLTPVCRSLQWLVGEGHQHEGYFLERSPVIHREMHGAVGGR	LTAFAAKCFMFARELRPTLVASVIDQALTFLINQQNTTGTFRPGRVSHKAMQGGVDSP	* . * : . : * : . : : : * : : * * * * * * * * * * * * * * * * * .		
	1270	1280	1290	1300	1310	1320
CiC31x1						
CiC32x2						
Thiol1						
Homology						
	MTLTAHVAISMGEINSICLPELNQRVIASRVSAMNYLEQHKDHATFQRPYPLSLLAYAAA	YSLTAYVLVTLIEAQRINCSGVNQQIQSRDKAINYLNRNRRNHPAFQRPYGLSILTYAMA	ITMTAYVLITLKETN---YAVKNRAVQEAENARIYLEN---HLTISISDNKYALAIVTYA	:: * * * * : : : * : . * : : : . * * * : * : : : : * :		
	1330	1340	1350	1360	1370	1380
CiC31x1						
CiC32x2						
Thiol1						
Homology						
	LHNPRSQLAIEMNARLMAMKQTSNGAYVFWRAKTLAEISGTNAHAYWYRTRPLALDIET	LHDQSSAFTIELNQRLGLFQQLDDN-SYVHWQAHSHSDIQGTDTHDYWYVRRPQAIDVET	LHVAGSSRANEALLALEALATVQGG--FKFVHDNSESPDSYSSRWRPYYN-PPTNDIEM	** * : * * * : . . . : . * : : : . : . : * * : * : * * *		

	1390	1400	1410	1420	1430	1440
CiC31x1						
CiC32x2						
Thio11						
Homology	:	:	:	:	:	:
	1450	1460	1470	1480	1490	1500
CiC31x1						
CiC32x2						
Thio11						
Homology	:	:	:	:	:	:
	1510	1520	1530	1540	1550	1560
CiC31x1						
CiC32x2						
Thio11						
Homology	:	:	:	:	:	:
	1570	1580	1590	1600	1610	1620
CiC31x1						
CiC32x2						
Thio11						
Homology	:	:	:	:	:	:
	1630	1640	1650	1660	1670	1680
CiC31x1						
CiC32x2						
Thio11						
Homology	:	:	:	:	:	:
	1690	1700	1710	1720	1730	1740
CiC31x1						
CiC32x2						
Thio11						
Homology	:	:	:	:	:	:
	1750	1760	1770	1780	1790	1800
CiC31x1						
CiC32x2						
Thio11						
Homology	:	:	:	:	:	:
	1810	1820	1830	1840	1850	1860
CiC31x1						
CiC32x2						
Thio11						
Homology	:	:	:	:	:	:

	1870	1880	1890	1900	1910	1920
CiC31x1	NNIPAKDTRLFKEGRVLLITGSLIERRQNRNRLQLTVY-QVDEQTTAERLVTDVACARAK					
CiC32x2	PDRSEESQRYLKPNMKILLMSNFLDFSMDSRSGHVRHDY-QMGEGTTVERIIPDSKCVQIR					
Thiol1	PSGDGDDIRMFVNGNAVGSTKSFTTRYFGKHGRNRFFLGQNTTRGNAWARGYFQGGGLAAVG					
Homology	.	.. *	:	.	:	::

	1930	1940	1950	1960	1970	1980
CiC31x1	S---VLVR--CA---GMRPPRQSKCVKSRELN-----AICENMKRLKESLN					
CiC32x2	AR-VALPKFQCENPNFHRPNRQAKCDKMMKMK-----TSCDNMDRLKNQVQ					
Thiol1	TWRSVLTDDQITALYEAYRPAIESSDPLSVKLLRHFAVQQLLLCFQSPATIEDLYSRSA					
Homology	:	.*	.	::	::	:

	1990
CiC31x1	DIGCD-----
CiC32x2	QG-CDK-----
Thiol1	PVTCPTAPISPLMPFLPIL
Homology	*

References

Adams, M. D. Celniker, S. E. Holt, R. A. Evans, C. A. Gocayne, J. D. Amanatides,
P. G. Scherer, S. E. Li, P. W. Hoskins, R. A. Galle, R. F. George, R. A. Lewis, S. E.
Richards, S. Ashburner, M. Henderson, S. N. Sutton, G. G. Wortman, J. R.
Yandell, M. D. Zhang, Q. Chen, L. X. Brandon, R. C. Rogers, Y. H. C. Blazej, R. G.
Champe, M. Pfeiffer, B. D. Wan, K. H. Doyle, C. Baxter, E. G. Helt, G. Nelson, C.
R. Miklos, G. L. G. Abril, J. F. Agbayani, A. An, H. J. Andrews-Pfannkoch, C.
Baldwin, D. Ballew, R. M. Basu, A. Baxendale, J. Bayraktaroglu, L. Beasley, E. M.
Beeson, K. Y. Benos, P. V. Berman, B. P. Bhandari, D. Bolshakov, S. Borkova, D.
Botchan, M. R. Bouck, J. Brokstein, P. Brottier, P. Burtis, K. C. Busam, D. A.
Butler, H. Cadieu, E. Center, A. Chandra, I. Cherry, J. M. Cawley, S. Dahlke, C.
Davenport, L. B. Davies, A. de Pablos, B. Delcher, A. Deng, Z. M. Mays, A. D. Dew,
I. Dietz, S. M. Dodson, K. Doup, L. E. Downes, M. Dugan-Rocha, S. Dunkov, B. C.
Dunn, P. Durbin, K. J. Evangelista, C. C. Ferraz, C. Ferriera, S. Fleischmann, W.
Fosler, C. Gabrielian, A. E. Garg, N. S. Gelbart, W. M. Glasser, K. Glodek, A.
Gong, F. C. Gorrell, J. H. Gu, Z. P. Guan, P. Harris, M. Harris, N. L. Harvey, D.
Heiman, T. J. Hernandez, J. R. Houck, J. Hostin, D. Houston, D. A. Howland, T. J.
Wei, M. H. Ibegwam, C. Jalali, M. Kalush, F. Karpen, G. H. Ke, Z. X. Kennison, J.
A. Ketchum, K. A. Kimmel, B. E. Kodira, C. D. Kraft, C. Kravitz, S. Kulp, D. Lai,
Z. W. Lasko, P. Lei, Y. D. Levitsky, A. A. Li, J. Y. Li, Z. Y. Liang, Y. Lin, X. Y.
Liu, X. J. Mattei, B. McIntosh, T. C. McLeod, M. P. McPherson, D. Merkulov, G.
Milshina, N. V. Mobarry, C. Morris, J. Moshrefi, A. Mount, S. M. Moy, M.
Murphy, B. Murphy, L. Muzny, D. M. Nelson, D. L. Nelson, D. R. Nelson, K. A.
Nixon, K. Nusskern, D. R. Pacleb, J. M. Palazzolo, M. Pittman, G. S. Pan, S.
Pollard, J. Puri, V. Reese, M. G. Reinert, K. Remington, K. Saunders, R. D. C.
Scheeler, F. Shen, H. Shue, B. C. Siden-Kiamos, I. Simpson, M. Skupski, M. P.

Smith, T. Spier, E. Spradling, A. C. Stapleton, M. Strong, R. Sun, E. Svirskas, R. Tector, C. Turner, R. Venter, E. Wang, A. H. H. Wang, X. Wang, Z. Y. Wassarman, D. A. Weinstock, G. M. Weissenbach, J. Williams, S. M. Woodage, T. Worley, K. C. Wu, D. Yang, S. Yao, Q. A. Ye, J. Yeh, R. F. Zaveri, J. S. Zhan, M. Zhang, G. G. Zhao, Q. Zheng, L. S. Zheng, X. Q. H. Zhong, F. N. Zhong, W. Y. Zhou, X. J. Zhu, S. P. Zhu, X. H. Smith, H. O. Gibbs, R. A. Myers, E. W. Rubin, G. M. and Venter, J. C. (2000). The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185-2195.

Alsenz, J., Avila, D., Huemer, H. P., Esparza, I., Becherer, J. D., Kinoshita, T., Wang, Y., Oppermann, S. and Lambris, J. D. (1992). Phylogeny of the 3rd component of complement, C3 - analysis of the conservation of human Cr 1, Human Cr 2, human H, and human- B sites, concanavalin-a binding-sites, and the thiolester bond in C3 from different species. *Developmental and Comparative Immunology* **16**, 63-76.

Al-Sharif, W. Z., Sunyer, J. O., Lambris, J. D. and Smith, L. C. (1997). Sea urchin coelomocytes specifically express a homologue of the complement component C3 *Journal of Immunology* **162**, 3105-3105.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W. and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389-3402.

Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, T., Corpet, F., Croning, M. D. R., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N. J., Oinn, T. M., Pagni, M., Servant, F., Sigrist, C. J. A. and Zdobnov, E. M. (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Research* **29**, 37-40.

Arason, G. J. (1996). Lectins as defence molecules in vertebrates and invertebrates. *Fish & Shellfish Immunology* **6**, 277-289.

Armstrong, P. B. and Quigley, J. P. (1999). Alpha(2)-macroglobulin: An evolutionarily conserved arm of the innate immune system. *Developmental and Comparative Immunology* **23**, 375-390.

Armstrong, P. B., Melchior, R., Swarnakar, S. and Quigley, J. P. (1998). Alpha2-macroglobulin does not function as a C3 homologue in the plasma hemolytic system of the American horseshoe crab, *Limulus*. *Molecular Immunology* **35**, 47-53.

Asokan, R., Armstrong, M. T. and Armstrong, P. B. (2000). Association of alpha2-macroglobulin with the coagulin clot in the American horseshoe crab, *Limulus polyphemus*: A potential role in stabilization from proteolysis. *Biological Bulletin* **199**, 190-192.

Attwood, T. K., Blythe, M. J., Flower, D. R., Gaulton, A., Mabey, J. E., Maudling, N., McGregor, L., Mitchell, A. L., Moulton, G., Paine, K. and Scordis, P. (2002).

PRINTS and PRINTS-S shed light on protein ancestry. *Nucleic Acids Research* **30**, 239-241.

Azumi, K., Ishimoto, R., Fujita, T., Nonaka, M. and Yokosawa, H. (2000). Opsonin-independent and dependent phagocytosis in the ascidian *Halocynthia roretzi*: Galactose-specific lectin and complement C3 function as target dependent opsonins. *Zoological Science* **17**, 625-632.

Bairoch, A. and Apweiler, R. (1998). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucleic Acids Research* **26**, 38-42.

Baker, W., van den Broek, A., Camon, E., Hingamp, P., Sterk, P., Stoesser, G. and Tuli, M. A. (2000). The EMBL nucleotide sequence database. *Nucleic Acids Research* **28**, 19-23.

Balian, G., Click, E. M. and Bornstein, P. (1980). Location of a collagen-binding domain in fibronectin. *Journal of Biological Chemistry* **255**, 3234-3236.

Ballarin, L., Cima, F. and Sabbadin, A. (1994). Phagocytosis in the colonial ascidian *Botryllus schlosseri*. *Developmental and Comparative Immunology* **18**, 467-481.

Ballarin, L., Tonello, C. and Sabbadin, A. (2000). Humoral opsonin from the colonial ascidian *Botryllus schlosseri* as a member of the galectin family. *Marine Biology* **136**, 823-827.

Ballarin, L., Tonello, C., Guidolin, L. and Sabbadin, A. (1999). Purification and characterization of a humoral opsonin, with specificity for D-galactose, in the colonial ascidian *Botryllus schlosseri*. *Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology* **123**, 115-123.

Banyai, L. and Patthy, L. (1999). The NTR module: Domains of netrins, secreted frizzled related proteins, and type I procollagen C-proteinase enhancer protein are homologous with tissue inhibitors of metalloproteases. *Protein Science* **8**, 1636-1642.

Barnes, R. S. K., Calow, P. and Olive, P. J. W. (1993). *The Invertebrates; A New Synthesis*. Oxford: Blackwell Scientific Publications.

Barnes, W. M. (1994). PCR amplification of up to 35 Kb DNA with high-fidelity and high-yield from lambda-B bacteriophage templates. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 2216-2220.

Barrington, E. J. W. (1965). *The Biology of the Hemichordata and Protochordata*. Edinburgh: Oliver and Boyd.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M. and Sonnhammer, E. L. L. (2002). The Pfam protein families database. *Nucleic Acids Research* **30**, 276-280.

Baxevanis, A. D. and Ouellette, B. F. (1998). *Bioinformatics : A Practical Guide to the Analysis of Genes and Proteins*. New York: Wiley-Interscience.

Bentley, D. R. (1988). Structural superfamilies of the complement system. *Experimental and Clinical Immunogenetics* **5**, 69-80.

Berril, N. J. (1936). Studies in tunicate development V. The evolution and classification of ascidians. *Philosophical Transactions. Royal Society of London. Series B. Biological Sciences B* **226**, 43-70.

Bjellqvist, B., Hughes, G. J., Pasquali, C., Paquet, N., Ravier, F., Sanchez, J. C., Frutiger, S. and Hochstrasser, D. (1993). The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis* **14**, 1023-1031.

Bohn, H. (1986). Hemolymph clotting in insects. In *Immunity in Invertebrates*, (ed. M. Brehélin), pp. 188-207. Berlin: Springer-Verlag.

Bullough, W. S. (1958). *Practical Invertebrate Anatomy*. London: MacMillan and Company.

Carroll, M. C. (1998). The role of complement and complement receptors in induction and regulation of immunity. *Annual Review of Immunology* **16**, 545-568.

Carroll, M. C. and Prodeus, A. P. (1998). Linkages of innate and adaptive immunity. *Current Opinion in Immunology* **10**, 36-40.

Cerenius, L. and Söderhäll, K. (1995). Crustacean immunity and complement - a premature comparison. *American Zoologist* **35**, 60-67.

- Chelly, J., Montarras, D., Pinset, C., Berwald-Netter, Y., Kaplan, J. C. and Kahn, A.** (1990). Quantitative estimation of minor mRNAs by cDNA-polymerase chain reaction. Application to dystrophin mRNA in cultured myogenic and brain cells. *European Journal of Biochemistry* **187**, 691-698.
- Cheng, S., Fockler, C., Barnes, W. M. and Higuchi, R.** (1994). Effective amplification of long targets from cloned inserts and human genomic DNA. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 5695-5699.
- Clow, L. A., Gross, P. S., Shih, C. S. and Smith, L. C.** (2000). Expression of SpC3, the sea urchin complement component, in response to lipopolysaccharide. *Immunogenetics* **51**, 1021-1033.
- Corpet, F., Servant, F., Gouzy, J. and Kahn, D.** (2000). ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Research* **28**, 267-269.
- Cross, P. S., Al-Sharif, W. Z., Clow, L. A. and Smith, L. C.** (1999). Echinoderm immunity and the evolution of the complement system. *Developmental and Comparative Immunology* **23**, 429-442.
- Cushley, W. and Owen, M. J.** (1983). Structural and genetic similarities between immunoglobulins and class-I histocompatibility antigens. *Immunology Today* **4**, 88-92.
- Dahl, M. R., Thiel, S., Matsushita, M., Fujita, T., Willis, A. C., Christensen, T., Vorup-Jensen, T. and Jensenius, J. C.** (2001). A new mannan-binding lectin

associated serine protease, MASP-3, and its association with distinct complexes of the MBL complement activation pathway. *Molecular Immunology* **38**, 19.

Daquila, R. T., Bechtel, L. J., Videler, J. A., Eron, J. J., Gorczyca, P. and Kaplan, J. C. (1991). Maximizing sensitivity and specificity of PCR by pre-amplification heating. *Nucleic Acids Research* **19**, 3749-3749.

DeLeo, G., Parrinello, N., Parrinello, D., Cassara, G. and diBella, M. A. (1996). Encapsulation response of *Ciona intestinalis* (Ascidiacea) to intratunical erythrocyte injection. *Journal of Invertebrate Pathology* **67**, 205-212.

DeLeo, G., Parrinello, N., Parrinello, D., Cassara, G., Russo, D. and DiBella, M. A. (1997). Encapsulation response of *Ciona intestinalis* (Ascidiacea) to intratunical erythrocyte injection .2. The outermost inflamed area. *Journal of Invertebrate Pathology* **69**, 14-23.

Di Bella, M. A. and De Leo, G. (2000). Hemocyte migration during inflammatory-like reaction of *Ciona intestinalis* (Tunicata, Ascidiacea). *Journal of Invertebrate Pathology* **76**, 105-111.

Dishaw, L., Smith, S. L. and Bigger, C. (2000). Sequence analysis of the partial cDNA clones from a primitive coral, encoding a thiolester-containing protein (abstract). *Developmental and Comparative Immunology* **24**, S23.

Dodds, A. W. and Day, A. J. (1993). The phylogeny and evolution of the complement system. In *Complement in Health and Disease.*, (ed. M. L. K. WHALEY, and J.M. WEILER.). Boston: Kluwer.

Dodds, A. W. and Law, S. K. A. (1998). The phylogeny and evolution of the thioester bond-containing proteins C3, C4 and alpha(2)-macroglobulin. *Immunological Reviews* **166**, 15-26.

Dodds, A. W., Ren, X. D., Willis, A. C. and Law, S. K. A. (1996). The reaction mechanism of the internal thioester in the human complement component C4. *Nature* **379**, 177-179.

Dodds, A. W., Smith, S. L., Levine, R. P. and Willis, A. C. (1998). Isolation and initial characterisation of complement components C3 and C4 of the nurse shark and the channel catfish. *Developmental and Comparative Immunology* **22**, 207-216.

Don, R. H., Cox, P. T., Wainwright, B. J., Baker, K. and Mattick, J. S. (1991). Touchdown PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Research* **19**, 4008-4008.

dos Remedies, N. J., Ramsland, P. A., Hook, J. W. and Raison, R. L. (1999). Identification of a homologue of CD59 in a cyclostome: Implications for the evolutionary development of the complement system. *Developmental and Comparative Immunology* **23**, 1-14.

Edwards, Y. J. K. and Cottage, A. (2001). Prediction of protein structure and function by using bioinformatics. In *Genomics Protocols*, vol. 175 eds. M. P. Starkey and R. Elaswarapu), pp. 341-375. Totowa: Humana Press inc.

Eisenhaber, B., Bork, P. and Eisenhaber, F. (1999). Prediction of potential GPI-modification sites in proprotein sequences. *Journal of Molecular Biology* **292**, 741-758.

Ellis, A. E. (1982). Differences between the immune mechanisms of fish and higher vertebrates. In *Microbial Diseases of Fish*, (ed. S. Roberts), pp. 1-29.

Ember, J. A. and Hugli, T. E. (1997). Complement factors and their receptors. *Immunopharmacology* **38**, 3-15.

Endo, Y., Takahashi, M., Nakao, M., Saiga, H., Sekine, H., Matsushita, M., Nonaka, M. and Fujita, T. (2000). Two lineages of mannose-binding lectin-associated serine protease (MASP) in vertebrates (vol 161, pg 4924, 1998). *Journal of Immunology* **164**, 5530-5530.

Ezekowitz, R. A. B. and Hoffmann, J. (1998). Innate immunity the blossoming of innate immunity - overview. *Current Opinion in Immunology* **10**, 9-11.

Ezekowitz, R. A. B. and Hoffmann, J. (2001). Innate immunity - still blossoming Editorial overview. *Current Opinion in Immunology* **13**, 53-54.

- Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J. A., Hofmann, K. and Bairoch, A.** (2002). The PROSITE database, its status in 2002. *Nucleic Acids Research* **30**, 235-238.
- Farrell, R. E. J.** (1998). RNA Methodologies - A Laboratory Guide for Isolation and Characterisation. California: Academic Press.
- Farries, T. C. and Atkinson, J. P.** (1991). Evolution of the complement system. *Immunology Today* **12**, 295-300.
- Fearon, D. T.** (1997a). Seeking wisdom in innate immunity. *Nature* **388**, 323-324.
- Fearon, D. T.** (1997b). Innate and acquired immunity. *Nature* **388**, 323-323.
- Fearon, D. T.** (1999). Innate immunity and the biological relevance of the acquired immune response. *Qjm-Monthly Journal of the Association of Physicians* **92**, 235-237.
- Fearon, D. T. and Locksley, R. M.** (1996). Elements of immunity - the instructive role of innate immunity in the acquired immune response. *Science* **272**, 50-54.
- Field, K. G., Olsen, G. J., Lane, D. J., Giovannoni, S. J., Ghiselin, M. T., Raff, E. C., Pace, N. R. and Raff, R. A.** (1988). Molecular phylogeny of the animal kingdom. *Science* **239**, 748-753.
- Findlay, C. and Smith, V. J.** (1995). Antimicrobial factors in solitary ascidians. *Fish & Shellfish Immunology* **5**, 645-658.

- Franchini, S., Zarkadis, I. K., Sfyroera, G., Sahu, A., Moore, W. T., Mastellos, D., LaPatra, S. E. and Lambris, J. D.** (2001). Cloning and purification of the rainbow trout fifth component of complement (C5). *Developmental and Comparative Immunology* **25**, 419-430.
- Frank, M. M.** (1979). The complement system in host defense and inflammation. *Review of Infectious diseases* **1**, 483-501.
- Frohman, M. A., Dush, M. K. and Martin, G. R.** (1988). Rapid production of full-length cDNAs from rare transcripts - amplification using a single gene-specific oligonucleotide primer. *Proceedings of the National Academy of Sciences of the United States of America* **85**, 8998-9002.
- Fujii, T., Nakamura, T., Sekizawa, A. and Tomonaga, S.** (1992). Isolation and Characterization of a Protein From Hagfish Serum That Is Homologous to the 3rd Component of the Mammalian Complement-System. *Journal of Immunology* **148**, 117-123.
- Garstang, W.** (1928). The morphology of the Tunicata and its bearing on the phylogeny of the Chordata. *Quarterly Journal of the Microscopy Society* **72**, 51-187.
- Gewurz, H., Zhang, X. H. and Lint, T. F.** (1995). Structure and function of the pentraxins. *Current Opinion in Immunology* **7**, 54-64.
- Gongora, R., Figueroa, F. and Klein, J.** (1998). Independent duplications of Bf and C3 complement genes in the zebrafish. *Scandinavian Journal of Immunology* **48**, 651-658.

Götz, P. (1986). Encapsulation in arthropods. In *Immunity in Invertebrates*, (ed. M. Brehélin), pp. 153-170. Berlin: Springer-Verlag.

Gould, J. M. and Weiser, J. N. (2001). Expression of C-reactive protein in the human respiratory tract. *Infection and Immunity* **69**, 1747-1754.

Gross, P. S., Clow, L. A. and Smith, L. C. (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.

Gross, P. S., Al-Sharif, W. Z., Clow, L. A. and Smith, L. C. (1999a). Echinoderm immunity and the evolution of the complement system. *Developmental and Comparative Immunology* **23**, 533-533.

Gross, P. S., Clow, L. A., Shih, C. S. and Smith, L. C. (1999b). Complement protein C3 (SpC3) from the purple sea urchin, *Strongylocentrotus purpuratus*, is expressed specifically in a subpopulation of the phagocytic coelomocytes. *FASEB Journal* **13**, A284-A284.

Haft, D. H., Loftus, B. J., Richardson, D. L., Yang, F., Eisen, J. A., Paulsen, I. T. and White, O. (2001). TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Research* **29**, 41-43.

Hammond, J. A. and Smith, V. J. (2002). Lipopolysaccharide induces DNA-synthesis in a sub-population of hemocytes from the swimming crab, *Liocarcinus depurator*. *Developmental and Comparative Immunology* **26**, 227-236.

- Hanley, P. J., Hook, J. W., Raftos, D. A., Gooley, A. A., Trent, R. and Raison, R. L.** (1992). Hagfish humoral defense protein exhibits structural and functional homology with mammalian complement components. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 7910-7914.
- Hardy, S. W., Fletcher, T. C. and Olafsen, J. A.** (1977). Aspects of cellular and humoral defence mechanisms in the Pacific oyster, *Crassostrea gigas*,. In *Developmental Immunobiology*, eds. J. D. Solomon and J. D. Horton), pp. 59-66. Amsterdam: Elsevier.
- Hayward, P. J. and Ryland, J. S.** (1995). Handbook of the Marine Fauna of North-West Europe. New York: Oxford University Press.
- Henikoff, S. and Henikoff, J. G.** (1993). Performance evaluation of amino acid substitution matrices. *Proteins-Structure Function and Genetics* **17**, 49-61.
- Hofmann, K. and Stoffel, W.** (1993). TMBASE - A database of membrane spanning protein segments. *Biological Chemistry* **374**, 166.
- Hughes, A. L.** (1994). Phylogeny of the C3/C4/C5 complement component gene family indicates that C5 diverged first. *Molecular Biology and Evolution* **11**, 417-425.
- Hughes, A. L. and Yeager, M.** (1997). Molecular evolution of the vertebrate immune system. *Bioessays* **19**, 777-786.

- Ishikawa, N., Nonaka, M., Wetsel, R. A. and Colten, H. R.** (1990). Murine complement-C2 and factor-B genomic and cDNA cloning reveals different mechanisms for multiple transcripts of C2 and factor B. *Journal of Biological Chemistry* **265**, 19040-19046.
- Iwanaga, S.** (1989). Molecular mechanism of hemolymph clotting system in invertebrate animals. *Thrombosis and Haemostasis* **62**, 457-457.
- Iwanaga, S., Kawabata, S. and Muta, T.** (1998). New types of clotting factors and defense molecules found in horseshoe crab hemolymph: Their structures and functions. *Journal of Biochemistry* **123**, 1-15.
- Jackson, A. D. and Smith, V. J.** (1993). LPS-sensitive protease activity in the blood cells of the solitary ascidian, *Ciona intestinalis* (L). *Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology* **106**, 505-512.
- Jackson, A. D., Smith, V. J. and Peddie, C. M.** (1993). *In vitro* phenoloxidase activity in the blood of *Ciona intestinalis* and other ascidians. *Developmental and Comparative Immunology* **17**, 97-108.
- Janeway, C. A. and Travers, P.** (1996). The humoral immune response. In *Immunobiology: The immune System in Health and Disease.*: Churchill Livingstone.
- Jefferies, R. P. S.** (1986). The Ancestry of the Vertebrates. Dorchester: British Museum (Natural History).

Jensen, J. A., Festa, E., Smith, D. S. and Cayer, M. (1981). The complement system of the nurse shark - hemolytic and comparative characteristics. *Science* **214**, 566-569.

Ji, X., Azumi, K., Sasaki, M. and Nonaka, M. (1997). Ancient origin of the complement lectin pathway revealed by molecular cloning of mannan binding: Protein-associated serine protease from a urochordate, the Japanese ascidian, *Halocynthia roretzi*. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 6340-6345.

Ji, X., Namikawa-Yamada, C., Nakanishi, M., Sasaki, M. and Nonaka, M. (2000). Molecular cloning of complement factor B from a solitary ascidian: unique combination of domains implicating ancient exon shufflings (abstract). *Immunopharmacology* **49**, 43.

Ji, X., Azumi, K., Nonaka, M., Namikawa-Yamada, C., Sasaki, M., Saiga, H., Dodds, A. W., Sekine, H., Homma, M., Matsushita, M., Endo, Y. and Fujita, T. (1998). Opsonic complement C3 in the solitary ascidian. *Molecular Immunology* **35**, 130.

Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences* **8**, 275-282.

Kaidoh, T. and Gigli, I. (1989). Phylogeny of regulatory proteins of the complement system - isolation and characterization of a C4b/C3b inhibitor and a co-factor from sand bass plasma. *Journal of Immunology* **142**, 1605-1613.

Kasahara, M., Nakaya, J., Satta, Y. and Takahata, N. (1997). Chromosomal duplication and the emergence of the adaptive immune system. *Trends in Genetics* **13**, 90-92.

Katagiri, T., Hirono, I. and Aoki, T. (1999). Molecular analysis of complement component C8 beta and C9 cDNAs of Japanese flounder, *Paralichthys olivaceus*. *Immunogenetics* **50**, 43-48.

Kellogg, D. E., Rybalkin, I., Chen, S., Mukhamedova, N., Vlasik, T., Siebert, P. D. and Chenchik, A. (1994). Taqstart antibody (TM) - hot start PCR facilitated by a neutralizing monoclonal antibody directed against Taq DNA polymerase. *Biotechniques* **16**, 1134-1137.

Kelly, K. L., Cooper, E. L. and Raftos, D. A. (1993a). A humoral opsonin from the solitary urochordate *Styela clava*. *Developmental and Comparative Immunology* **17**, 29-39.

Kelly, K. L., Cooper, E. L. and Raftos, D. A. (1993b). Cytokine-like activities of a humoral opsonin from the solitary Urochordate *Styela clava*. *Zoological Science* **10**, 57-64.

King, R. D. and Sternberg, M. J. E. (2002; In Press). Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Science*.

King, R. D., Saqi, M., Sayle, R. and Sternberg, M. J. E. (1997). DSC: Public domain protein secondary structure prediction. *Computer Applications in the Biosciences* **13**, 473-474.

Koppenheffer, T. L. (1987). Serum complement systems of ectothermic vertebrates. *Developmental and Comparative Immunology* **11**, 279-286.

Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E. L. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology* **305**, 567-580.

Kuhlman, M., Joiner, K. and Ezekowitz, R. A. B. (1989). The human mannose-binding protein functions as an opsonin. *Journal of Experimental Medicine* **169**, 1733-1745.

Kuroda, N., Naruse, K., Shima, A., Nonaka, M. and Sasaki, M. (2000). Molecular cloning and linkage analysis of complement C3 and C4 genes of the Japanese medaka fish. *Immunogenetics* **51**, 117-128.

Kuroda, N., Wada, H., Naruse, K., Simada, A., Shima, A., Sasaki, M. and Nonaka, M. (1996). Molecular cloning and linkage analysis of the Japanese medaka fish complement Bf/C2 gene. *Immunogenetics* **44**, 459-467.

Kustin, K., Robinson, W. E. and Smith, M. J. (1990). Tunichromes, vanadium and vacuolated blood cells in Tunicates. *Invertebrate Reproduction & Development* **17**, 129-139.

Lambris, J. D., Reid, K. B. M. and Volanakis, J. E. (1999). The evolution, structure, biology and pathophysiology of complement. *Immunology Today* **20**, 207-211.

Law, S. K. A. and Reid, K. B. M. (1988). Complement. Oxford: IRL Press Ltd.

Law, S. K. A. and Dodds, A. W. (1990). C3, C4 and C5 - the thioester site. *Biochemical Society Transactions* **18**, 1155-1159.

Law, S. K. A. and Dodds, A. W. (1996). Catalysed hydrolysis - the complement quickstep. *Immunology Today* **17**, 105-105.

Law, S. K. A. and Dodds, A. W. (1997). The internal thioester and the covalent binding properties of the complement proteins C3 and C4. *Protein Science* **6**, 263-274.

Lee, C. C. (1999). Sequence analysis of cDNAs encoding C3 in the nurse shark (*Ginglymostoma cirratum*). In *Medical Laboratory Sciences*. Florida.: Florida International University.

Letunic, I., Goodstadt, L., Dickens, N. J., Doerks, T., Schultz, J., Mott, R., Ciccarelli, F., Copley, R. R., Ponting, C. P. and Bork, P. (2002). Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Research* **30**, 242-244.

Levashina, E. A., Moita, L. F., Blandin, S., Vriend, G., Lagueux, M. and Kafatos, F. C. (2001). Conserved role of a complement-like protein in phagocytosis revealed by

dsRNA knockout in cultured cells of the mosquito, *Anopheles gambiae*. *Cell* **104**, 709-718.

Li, X., Namikawa-Yamada, C., Nakanishi, M., Sasaki, M. and Nonaka, M. (2000). Molecular cloning of complement factor B from a solitary ascidian: unique combination of donains and implication of ancient exon shufflings. *Immunopharmacology* **49**, 43.

Lin, M., Sutherland, R., Horsfall, W., Totty, N., Yeo, E., Nayar, R., Wu, X. F. and Schuh, A. C. (2002). Cell surface antigen CD109 is a novel member of the alpha(2) macroglobulin/C3, C4, C5 family of thioester-containing proteins. *Blood* **99**, 1683-1691.

Loos, M. (1982). The classical complement pathway - mechanism of activation of the 1st component by antigen-antibody complexes. *Progress in Allergy* **30**, 135-192.

Lu, J., Thiel, S., Wiedemann, H., Timpl, R. and Reid, K. B. M. (1990). Binding of the pentamer hexamer forms of mannan-binding protein to zymosan activates the proenzyme C1r2cls2 complex, of the classical pathway of complement, without involvement of C1q. *Journal of Immunology* **144**, 2287-2294.

Malhotra, R., Wormald, M. R., Rudd, P. M., Fischer, P. B., Dwek, R. A. and Sim, R. B. (1995). Glycosylation changes of IgG associated with rheumatoid arthritis can activate complement via the mannose-binding protein. *Nature Medicine* **1**, 237-243.

Matsushita, M. (1996). The lectin pathway of the complement system. *Microbiology and Immunology* **40**, 887-893.

Matsushita, M. and Fujita, T. (1992). Activation of the classical complement pathway by mannose-binding protein in association with a novel C1s-like serine protease. *Journal of Experimental Medicine* **176**, 1497-1502.

Matsushita, M. and Fujita, T. (1995). Cleavage of the 3rd component of complement (C3) by mannose-binding protein-associated serine protease (MASP) with subsequent complement activation. *Immunobiology* **194**, 443-448.

Matsushita, M., Endo, Y. and Fujita, T. (1998a). MASP1 (MBL-associated serine protease 1). *Immunobiology* **199**, 340-347.

Matsushita, M., Endo, Y., Nonaka, M. and Fujita, T. (1998b). Complement-related serine proteases in tunicates and vertebrates. *Current Opinion in Immunology* **10**, 29-35.

Matz, M., Shagin, D., Bogdanova, E., Britanova, O., Lukyanov, S., Diatchenko, L. and Chenchik, A. (1999). Amplification of cDNA ends based on template-switching effect and step-out PCR. *Nucleic Acids Research* **27**, 1558-1560.

McPherson, M. J. and Møller, S. G. (2000). PCR. Oxford: Bios scientific publishers limited.

Millar, R. H. (1971). The biology of ascidians. *Advances in marine biology* **9**, 1-100.

Miyazawa, S., Azumi, K. and Nonaka, M. (2001). Cloning and characterization of integrin alpha subunits from the solitary ascidian, *Halocynthia roretzi*. *Journal of Immunology* **166**, 1710-1715.

Moller, S., Croning, M. D. R. and Apweiler, R. (2002). Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* **18**, 218-218.

Morgan, B. P. and Gasque, P. (1996). Expression of complement in the brain: Role in health and disease. *Immunology Today* **17**, 461-466.

Morris, S. C. (1993). The fossil record and the early evolution of the Metazoa. *Nature* **361**, 219-225.

Muta, T., Miyata, T., Misumi, Y., Tokunaga, F., Nakamura, T., Toh, Y., Ikehara, Y. and Iwanaga, S. (1991). *Limulus* factor C. An endotoxin-sensitive serine protease zymogen with a mosaic structure of complement-like, epidermal growth factor-like, and lectin-like domains. *Journal of Biological Chemistry* **266**, 6554-6561.

Nagai, T., Mutsuro, J., Kimura, M., Kato, Y., Fujiki, K., Yano, T. and Nakao, M. (2000). A novel truncated isoform of the mannose-binding lectin- associated serine protease (MASP) from the common carp (*Cyprinus carpio*). *Immunogenetics* **51**, 193-200.

Nair, S. V., Pearce, S., Green, P. L., Mahajan, D., Newton, R. A. and Raftos, D. A. (2000). A collectin-like protein from tunicates. *Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology* **125**, 279-289.

Nakao, M. and Yano, T. (1998). Structural and functional identification of complement components of the bony fish, carp (*Cyprinus carpio*). *Immunological Reviews* **166**, 27-38.

Nakao, M., Fushitani, Y., Fujiki, K., Nonaka, M. and Yano, T. (1998). Two diverged complement factor B/C2-like cDNA sequences from a teleost, the common carp (*Cyprinus carpio*). *Journal of Immunology* **161**, 4811-4818.

Nakao, M., Osaka, K., Kato, Y., Fujiki, K. and Yano, T. (2001). Molecular cloning of the complement C1r/C1s/MASP2-like serine proteases from the common carp (*Cyprinus carpio*). *Immunogenetics* **52**, 255-263.

Nakao, M., Mutsuro, J., Obo, R., Fujiki, K., Nonaka, M. and Yano, T. (2000). Molecular cloning and protein analysis of divergent forms of the complement component C3 from a bony fish, the common carp (*Cyprinus carpio*): Presence of variants lacking the catalytic histidine. *European Journal of Immunology* **30**, 858-866.

Nicholson-Weller, A. and Klickstein, L. B. (1999). C1q-binding proteins and C1q receptors. *Current Opinion in Immunology* **11**, 42-46.

Nonaka, M. (1994). Molecular analysis of the lamprey complement system. *Fish & Shellfish Immunology* **4**, 437-446.

Nonaka, M. (1997). Non-human complement. In *Complement-A Practical Approach*, eds. A. W. Dodds and R. B. Sim), pp. 247-263. New York: Oxford University Press.

Nonaka, M. (2000). Origin and evolution of the complement system. In *Origin and Evolution of the Vertebrate Immune System*, vol. 248, pp. 37-50.

Nonaka, M. (2001). Evolution of the complement system. *Current Opinion in Immunology* **13**, 69-73.

Nonaka, M. and Takahashi, M. (1992). Complete complementary DNA sequence of the 3rd component of complement of lamprey - implication for the evolution of thioester containing proteins. *Journal of Immunology* **148**, 3290-3295.

Nonaka, M. and Azumi, K. (1999). Opsonic complement system of the solitary ascidian, *Halocynthia roretzi*. *Developmental and Comparative Immunology* **23**, 421-427.

Nonaka, M. and Smith, S. L. (2000). Complement system of bony and cartilaginous fish. *Fish & Shellfish Immunology* **10**, 215-228.

Nonaka, M., Natsuumesakai, S. and Takahashi, M. (1981a). The complement system in rainbow trout (*Salmo gairdneri*) .2. Purification and characterization of the 5th component (C5). *Journal of Immunology* **126**, 1495-1498.

Nonaka, M., Takahashi, M. and Sasaki, M. (1994). Molecular cloning of a lamprey homolog of the mammalian MHC class-III gene, complement factor-B. *Journal of Immunology* **152**, 2263-2269.

Nonaka, M., Yamaguchi, N., Natsuumesakai, S. and Takahashi, M. (1981b). The complement system of rainbow trout (*Salmo gairdneri*) .1. Identification of the serum lytic system homologous to mammalian complement. *Journal of Immunology* **126**, 1489-1494.

Nonaka, M., Kuroda, N., Naruse, K. and Shima, A. (1998). Molecular genetics of the complement C3 convertases in lower vertebrates. *Immunological Reviews* **166**, 59-65.

Nonaka, M., Fujii, T., Kaidoh, T., Natsuumesakai, S., Yamaguchi, N. and Takahashi, M. (1984). Purification of a lamprey complement protein homologous to the 3rd component of the mammalian complement system. *Journal of Immunology* **133**, 3242-3249.

Nonaka, M., Azumi, K., Ji, X., Namikawa-Yamada, C., Sasaki, M., Saiga, H., Dodds, A. W., Sekine, H., Homma, M. K., Matsushita, M., Endo, Y. and Fujita, T. (1999). Opsonic complement component C3 in the solitary ascidian, *Halocynthia roretzi*. *Journal of Immunology* **162**, 387-391.

Ohno, S. (1999). Gene duplication and the uniqueness of vertebrate genomes circa 1970-1999. *Seminars in Cell & Developmental Biology* **10**, 517-522.

Ohta, M., Okada, M., Yamashina, I. and Kawasaki, T. (1990). The mechanism of carbohydrate mediated complement activation by the serum mannan-binding protein. *Journal of Biological Chemistry* **265**, 1980-1984.

Ohtake, S. I., Abe, T., Shishikura, F. and Tanaka, K. (1994). The phagocytes in hemolymph of *Halocynthia roretzi* and their phagocytic activity. *Zoological Science* **11**, 681-691.

Olafsen, J. A. (1986). Invertebrate Lectins: Biochemical Heterogeneity as a Possible Key to their Biological Function. In *Immunity in Invertebrates*, (ed. M. Brehélin), pp. 94-107. Berlin: Springer-Verlag.

Opal, S. M. (2000). Phylogenetic and functional relationships between coagulation and the innate immune response. *Critical Care Medicine* **28**, S77-S80.

Pahler, S., Blumbach, B., Muller, I. and Muller, W. E. G. (1998). Putative multiadhesive protein from the marine sponge *Geodia cydonium*: Cloning of the cDNA encoding a fibronectin-, an SRCR-, and a complement control protein module. *Journal of Experimental Zoology* **282**, 332-343.

Pan, T. L., Groger, H., Schmid, V. and Spring, J. (1998). A toxin homology domain in an astacin-like metalloproteinase of the jellyfish *Podocoryne carnea* with a dual role in digestion and development. *Development Genes and Evolution* **208**, 259-266.

Parrinello, N. and Patricolo, E. (1984). Inflammatory-like reaction in the tunic of *Ciona intestinalis* (Tunicata) .2. Capsule components. *Biological Bulletin* **167**, 238-250.

Parrinello, N., Patricolo, E. and Canicatti, C. (1984). Inflammatory-like reaction in the tunic of *Ciona intestinalis* (Tunicata) .1. Encapsulation and tissue injury. *Biological Bulletin* **167**, 229-237.

Parrinello, N., Cammarata, M., Lipari, L. and Arizza, V. (1995). Sphingomyelin inhibition of *Ciona intestinalis* (Tunicata) cytotoxic hemocytes assayed against sheep erythrocytes. *Developmental and Comparative Immunology* **19**, 31-41.

Peddie, C. M. and Smith, V. J. (1993). *In vitro* spontaneous cytotoxic activity against mammalian target cells by the hemocytes of the solitary ascidian, *Ciona intestinalis*. *Journal of Experimental Zoology* **267**, 616-623.

Peddie, C. M. and Smith, V. J. (1994a). Mechanism of cytotoxic activity by hemocytes of the solitary ascidian, *Ciona intestinalis*. *Journal of Experimental Zoology* **270**, 335-342.

Peddie, C. M. and Smith, V. J. (1994b). Blood cell mediated cytotoxic activity in the solitary ascidian *Ciona intestinalis*. In *Primordial Immunity: Foundations For the Vertebrate Immune System*, vol. 712, pp. 332-334.

Peddie, C. M., Riches, A. C. and Smith, V. J. (1995). Proliferation of undifferentiated blood cells from the solitary ascidian, *Ciona intestinalis* *in vitro*. *Developmental and Comparative Immunology* **19**, 377-387.

Pepys, M. B. and Baltz, M. L. (1983). Acute phase proteins with special reference to C-reactive protein and related proteins (pentaxins) and serum amyloid A protein. *Advances in Immunology* **34**, 141-212.

Quesenberry, M. S., O'Leary, N., Ahmed, H., Bianchet, M., Amzel, M., Marsh, A. and Vasta, G. R. (1998a). The protochordate *Clavelina picta* has key components of innate immunity in mammals: MBP-like, MASP-like, and complement-like molecules. *FASEB Journal* **12**, 238.

Quesenberry, M. S., O'Leary, N., Ahmed, H., Bianchet, M., Amzel, M., Marsh, A. and Vasta, G. (1998b). MBP, MASP, and complement-like molecules present in the protochordate *Clavelina picta* indicate early origin of innate immunity. *Glycobiology* **8**, 110.

Rast, J. P. and Litman, G. W. (1994). T-cell receptor gene homologs are present in the most primitive jawed vertebrates. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 9248-9252.

Rogers, E. (1986). Protocordates: Vertebrate Relatives. In *Looking at Invertebrates: A Practical Guide to Vertebrate Adaptions*, pp. 1-9. London: Longman.

Rowley, A. F. (1981). The blood cells of the sea squirt, *Ciona intestinalis* - morphology, differential counts and *in vitro* phagocytic activity. *Journal of Invertebrate Pathology* **37**, 91-100.

Rowley, A. F. (1982). Ultrastructural and cytochemical studies on the blood cells of the sea squirt, *Ciona intestinalis* .1. Stem cells and amebocytes. *Cell and Tissue Research* **223**, 403-414.

Saltercid, L. and Flajnik, M. F. (1995). Evolution and developmental regulation of the Major Histocompatibility Complex. *Critical Reviews in Immunology* **15**, 31-75.

Sambrook, J. and Russell, D. W. (2001). *Molecular Cloning - A laboratory manual*. New York: Cold Spring Harbor Laboratory Press.

- Sato, T., Endo, Y., Matsushita, M. and Fujita, T.** (1994). Molecular characterization of a novel serine protease involved in activation of the complement system by mannose-binding protein. *International Immunology* **6**, 665-669.
- Satoh, N. and Jeffery, W. R.** (1995). Chasing tails in ascidians - developmental insights into the origin and evolution of Chordates. *Trends in Genetics* **11**, 354-359.
- Schluter, S. F., Bernstein, R. M. and Marchalonis, J. J.** (1997). Molecular origins and evolution of immunoglobulin heavy-chain genes of jawed vertebrates. *Immunology Today* **18**, 543-549.
- Schultz, J., Milpetz, F., Bork, P. and Ponting, C. P.** (1998). SMART, a simple modular architecture research tool: Identification of signaling domains. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 5857-5864.
- Sekine, H., Kenjo, A., Azumi, K., Ohi, G., Takahashi, M., Kasukawa, R., Ichikawa, N., Nakata, M., Mizuochi, T., Matsushita, M., Endo, Y. and Fujita, T.** (2001). An ancient lectin-dependent complement system in an ascidian: novel lectin isolated from the plasma of the solitary ascidian, *Halocynthia roretzi*. *Journal of Immunology* **167**, 4504-4510.
- Shintani, S., Terzic, J., Sato, A., Saraga-Babic, M., O'HUigin, C., Tichy, H. and Klein, J.** (2000). Do lampreys have lymphocytes? The Spi evidence. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 7417-7422.

Shishikura, F., Abe, T., Ohtake, S. I. and Tanaka, K. (1997). Purification and characterization of a 39,000-Da serine proteinase from the hemolymph of a solitary ascidian, *Halocynthia roretzi*. *Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology* **118**, 131-141.

Siebert, P. D., Chenchik, A., Kellogg, D. E., Lukyanov, K. A. and Lukyanov, S. A. (1995). An improved PCR method for walking in uncloned genomic DNA. *Nucleic Acids Research* **23**, 1087-1088.

Sim, R. B. and Dodds, A. W. (1997). The Complement System: An Introduction. In *Complement-A Practical Approach*, eds. A. W. Dodds and R. B. Sim), pp. 1-17. New York: Oxford University Press.

Sim, R. B. and Laich, A. (2000). Serine proteases of the complement system. *Biochemical Society Transactions* **28**, 545-550.

Smith, L. C. (2002). Thioester function is conserved in SPC3, the sea urchin homologue of the complement component C3. *Developmental and Comparative Immunology* **26**, 603-614.

Smith, L. C. and Davidson, E. H. (1992). The echinoid immune system and the phylogenetic occurrence of immune mechanisms in deuterostomes. *Immunology Today* **13**, 356-362.

Smith, L. C., Britten, R. J. and Davidson, E. H. (1995). Lipopolysaccharide activates the sea urchin immune system. *Developmental and Comparative Immunology* **19**, 217-224.

Smith, L. C., Shih, C. S. and Dachenhausen, S. G. (1998). Coelomocytes express SpBf, a homologue of factor B, the second component in the sea urchin complement system. *Journal of Immunology* **161**, 6784-6793.

Smith, L. C., Azumi, K. and Nonaka, M. (1999). Complement systems in invertebrates. The ancient alternative and lectin pathways. *Immunopharmacology* **42**, 107-120.

Smith, L. C., Chang, L., Britten, R. J. and Davidson, E. H. (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. *Journal of Immunology* **156**, 593-602.

Smith, S. H. and Jensen, J. A. (1986). The 2nd component (C2n) of the nurse shark complement system - purification, physicochemical characterization and functional comparison with guinea pig C4. *Developmental and Comparative Immunology* **10**, 191-206.

Smith, S. L. (1997). Nurse shark complement- in retrospect and prospect. *Developmental and Comparative Immunology* **21**, 144.

Smith, S. L. (1998). Shark complement: An assessment. *Immunological Reviews* **166**, 67-78.

Smith, S. L., Riesgo, M., Obenauf, S. D. and Woody, C. J. (1997). Anaphylactic and chemotactic response of mammalian cells to zymosan-activated shark serum. *Fish & Shellfish Immunology* **7**, 503-514.

Smith, V. J. (1996). The prophenoloxidase activating system; A common defense pathway for deuterostomes and protostomes? In *Invertebrate Immune Responses. Advances in comparative and environmental physiology.*, vol. **23** (ed. E. L. COOPER), pp. 73-114.

Smith, V. J. and Söderhäll, K. (1991). A comparison of phenoloxidase activity in the blood of marine invertebrates. *Developmental and Comparative Immunology* **15**, 251-261.

Smith, V. J. and Peddie, C. M. (1992). Cell cooperation during host defense in the solitary tunicate *Ciona intestinalis* (L). *Biological Bulletin* **183**, 211-219.

Söderhäll, K. (1982). Prophenoloxidase activating system and melanization - a recognition mechanism of arthropods - a review. *Developmental and Comparative Immunology* **6**, 601-611.

Söderhäll, K. and Hall, L. (1984). Lipopolysaccharide-induced activation of the prophenoloxidase activating system in crayfish *Hemocyte lysate*. *Biochimica Et Biophysica Acta* **797**, 99-104.

Song, W. C., Sarrias, M. R. and Lambris, J. D. (2000). Complement and innate immunity. *Immunopharmacology* **49**, 187-198.

Sottrup-Jensen, L., Hansen, H. F., Mortensen, S. B., Petersen, T. E. and Magnusson, S. (1981). Sequence location of the reactive thiolester in human alpha-2-macroglobulin. *Febs Letters* **123**, 145-148.

Sottrup-Jensen, L., Stepanik, T. M., Kristensen, T., Lonblad, P. B., Jones, C. M., Wierzbicki, D. M., Magnusson, S., Domdey, H., Wetsel, R. A., Lundwall, A., Tack, B. F. and Fey, G. H. (1985). Common evolutionary origin of alpha-2-macroglobulin and complement component C3 and component C4. *Proceedings of the National Academy of Sciences of the United States of America* **82**, 9-13.

Springer, T. A. (1998). An extracellular beta-propeller module predicted in lipoprotein and scavenger receptors, tyrosine kinases, epidermal growth factor precursor, and extracellular matrix components. *Journal of Molecular Biology* **283**, 837-862.

Sreedhara, A., Susa, N., Patwardhan, A. and Rao, C. P. (1996). One electron reduction of vanadate(V) to oxovanadium(IV) by low-molecular-weight biocomponents like saccharides and ascorbic acid: Effect of oxovanadium(IV) complexes on pUC18 DNA and on lipid peroxidation in isolated rat hepatocytes. *Biochemical and Biophysical Research Communications* **224**, 115-120.

Suciu-Foca, N., Reed, E., Rubinstein, P., Mackenzie, W., Ng, A. K. and King, D. W. (1985). A late-differentiation antigen associated with the helper inducer function of human T-cells. *Nature* **318**, 465-467.

Sundsmo, J. S. and Fair, D. S. (1983). Relationships among the complement, kinin, coagulation, and fibrinolytic systems. *Springer Seminars in Immunopathology* **6**, 231-258.

Sunyer, J. O. and Lambris, J. D. (1998). Evolution and diversity of the complement system of poikilothermic vertebrates. *Immunological Reviews* **166**, 39-57.

Sunyer, J. O., Tort, L. and Lambris, J. D. (1997). Structural C3 diversity in fish - Characterization of five forms of C3 in the diploid fish *Spans aurata*. *Journal of Immunology* **158**, 2813-2821.

Sunyer, J. O., Zarkadis, I., Sarrias, M. R., Hansen, J. D. and Lambris, J. D. (1998). Cloning, structure, and function of two rainbow trout Bf molecules. *Journal of Immunology* **161**, 4106-4114.

Svane, I. and Havenhand, J. N. (1993). Spawning and dispersal in *Ciona intestinalis* (L). *Marine Ecology-Pubblicazioni Della Stazione Zoologica Di Napoli I* **14**, 53-66.

Takahashi, M., Endo, Y., Fujita, T. and Matsushita, M. (1999). A truncated form of mannose-binding lectin-associated serine protease (MASP)-2 expressed by alternative polyadenylation is a component of the lectin complement pathway. *International Immunology* **11**, 859-863.

Takemoto, T., Smith, S., Terado, T., Kimura, H. and Nonaka, M. (2000). Molecular cloning of complement factor B/C2 and C3/C4 from a Japanese shark *Triakis*. *Developmental and Comparative Immunology* **21**, 144.

The *C. elegans* sequencing consortium. (1999). Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **283**, 35-35.

Thiel, S., VorupJensen, T., Stover, C. M., Schwaeble, W., Laursen, S. B., Poulsen, K., Willis, A. C., Eggleton, P., Hansen, S., Holmskov, U., Reid, K. B. and Jensenius, J. C. (1997). A second serine protease associated with mannan-binding lectin that activates complement. *Nature* **386**, 506-510.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994). Clustal-W - improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673-4680.

Tokioka, T. (1971). *Publications of Seto Marine Biological Laboratory* **14**, 43-63.

Tomlinson, S., Stanley, K. K. and Esser, A. F. (1993). Domain structure, functional activity and polymerization of trout complement protein C9. *Developmental and Comparative Immunology* **17**, 67-76.

Trexler, M., Banyai, L. and Patthy, L. (2000). The LCCL module. *European Journal of Biochemistry* **267**, 5751-5757.

Turner, M. W. (1996). Mannose-binding lectin: The pluripotent molecule of the innate immune system. *Immunology Today* **17**, 532-540.

Uemura, T., Yano, T., Shiraishi, H. and Nakao, M. (1996). Purification and characterization of the eighth and ninth components of carp complement. *Molecular Immunology* **33**, 925-932.

Vasta, G. R., Quesenberry, M., Ahmed, H. and O'Leary, N. (1999). C-type lectins and galectins mediate innate and adaptive immune functions: Their roles in the complement activation pathway. *Developmental and Comparative Immunology* **23**, 401-420.

von Heijne, G. (1986). A new method for predicting signal sequence cleavage sites. *Nucleic Acids Research* **14**, 4683-4690.

Wada, H. and Satoh, N. (1994). Details of the evolutionary history from invertebrates to vertebrates, as deduced from the sequences of 18s rDNA. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 1801-1804.

Walport, M. (1996). Complement. In *Immunology*, (ed. I. RIOTT, J. BROSTOFF, and D. MALE.): Mosby.

Wilkins, M. R., Lindskog, I., Gasteiger, E., Bairoch, A., Sanchez, J. C.,

Hochstrasser, D. F. and Appel, R. D. (1997). Detailed peptide characterization using PEPTIDEMASS - A World Wide Web accessible tool. *Electrophoresis* **18**, 403-408.

Yano, T. and Nakao, M. (1994). Isolation of a carp complement protein homologous to mammalian factor D. *Molecular Immunology* **31**, 337-342.

Zarkadis, I. K., Mastellos, D. and Lambris, J. D. (2001a). Phylogenetic aspects of the complement system. *Developmental and Comparative Immunology* **25**, 745-762.

Zarkadis, I. K., Sarrias, M. R., Sfyroera, G., Sunyer, J. O. and Lambris, J. D. (2001b). Cloning and structure of three rainbow trout C3 molecules: a plausible explanation for their functional diversity. *Developmental and Comparative Immunology* **25**, 11-24.