

Characterisation of Secreted Phosphoprotein 24

Thesis submitted for the degree of

Doctor of Philosophy

At the University of Leicester

by

Clare Suzanne Bennett BSc (UCL)

Department of Genetics

University of Leicester

December 2001

UMI Number: U485874

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI U485874

Published by ProQuest LLC 2013. Copyright in the Dissertation held by the Author.
Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code.



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

This thesis is dedicated to my husband Phil Wharton.

Phil - you have been my inspiration, my support, my strength and my best friend throughout the three years it has taken me to complete this. This is for you babe!

Acknowledgements

There is no doubt that I would never have got this far were it not for the help and support of many, many people. Right back at the beginning of the three years, there was Alex. Alex and I found our way around lab G24 together and spent many lunchtimes together over sausage and chips. I would like to thank Alex for her friendship during the first year of my PhD.

During the last couple of years in the Genetics department the people who have kept me sane with coffee breaks full of gossiping, moaning and laughter have been Hilda, Zoë and Emma (Physics). Thank you to all of them, especially Zoë who was so generous with her car and is such a friendly next door neighbour.

Now we move on to the technical help. I would like to do a general thank you to the whole department for their help and support over the last three years. Particular thanks go to Jon Clayton for his help and advice on the protein modelling work that I attempted, to Marion MacFarlane and Xiao-Ming Sun for their advice and generosity with the baculovirus system and to Mark Jobling, Alec Jeffreys, David White and David Hosking for DNA samples that they donated to me.

Of course the biggest source of help, advice and inspiration for this thesis has been my supervisor Raymond Dalglish. A very big thank you to Raymond for putting up with me, teaching me and speaking so highly of me.

After a day in the lab you go home, but unfortunately with a PhD the work goes home with you. It's always inside your head and to cope with that requires support from your family and friends. I would like to thank my parents for the upbringing they gave me that enabled me to achieve so much in my life and for always being there for me no matter what. Then there is Phil to whom this thesis is dedicated. The biggest thanks goes to you for your continual support, for picking me up when I am down and just for being you really!

Abstract

Secreted phosphoprotein 24 (spp24) is a novel 24-kDa non-collagenous protein that was originally isolated from the acid demineralised extract of bovine cortical bone (Hu *et al.* 1995). The presence of spp24 in bone immediately suggested a potential role for the protein in the processes that occurred there. The N-terminal segment of the protein is related in sequence to the cystatin family of thiol protease inhibitors. It was therefore suggested that spp24 might inhibit thiol protease activity during bone turnover (Hu *et al.* 1995).

Three million people in the UK suffer from osteoporosis (National Osteoporosis Society estimated figure) and their care and treatment costs the NHS and the taxpayer £ 942 million every year (Dolan and Torgerson 1998). Therefore, it is essential that we begin to understand the genetic basis and the factors that can predispose people, to osteoporosis and many other bone diseases. If spp24 has a functional role in the process of bone remodelling it is likely that it may influence the development or severity of osteoporosis.

This study determines the human *SPP2* gene, encoding the spp24 protein, to comprise 8 exons with apparently TATA-less promoter. The gene is shown to have multiple transcription initiation sites, which demonstrate some tissue specificity. An extensive expression study was carried out on the human and mouse gene encoding spp24, indicating that the gene has an expression pattern of a tissue-specific nature, being expressed predominantly in liver.

Theoretical studies and computational methods were used to analyse spp24 from several species and proteins showing homology to spp24. These studies gave a good indication of the areas of the protein and specific residues that are likely to be critical to the function of spp24. The results supported the speculation that spp24 does not act as a typical cystatin, but instead is likely to have a fetuin-like function or an antimicrobial function.

Abbreviations

BMP	Bone morphogenetic protein
CRP	Cystatin-related protein
<i>E. coli</i>	<i>Escherishia coli</i>
EST	Expressed sequence tag
ET	Evolutionary trace
HRG	Histidine-rich glycoprotein
IPTG	Isopropylthio-beta-D-galactosidase
LB	Luria Bertani
MGP	Matrix-Gla protein
OLB	Oligo labelling buffer
PAC	P1 artificial chromosome
PCR	Polymerase chain reaction
PTH	Parathyroid hormone
Pycno	Pycnodystosis
RACE	Rapid amplification of cDNA ends
RT	Reverse transcription
SDS	Sodium dodecyl sulphate
<i>SPP2</i>	Secreted phosphoprotein 2
<i>Spp2</i>	Secreted phosphoprotein 24
<i>Spp24</i>	Secreted phosphoprotein 24
TdT	Terminal deoxynucleotidyl transferase
X-gal	5-bromo-4-chloro-3-indolyl-beta-D-galactosidase

Table of contents

Abstract

Acknowledgements

1.0 Introduction.....	1-43
1.1 General Introduction.....	1-2
1.2 Bone.....	3-15
1.2.1 Bone structure.....	3-9
1.2.1.1 The extracellular matrix of bone.....	3-7
1.2.1.2 Bone mineral.....	8-8
1.2.1.3 The cells found in bone.....	8-9
1.2.1.4 Bone patterns and bone architecture.....	9-9
1.2.2 Bone remodelling.....	9-14
1.2.2.1 Bone formation.....	10-13
1.2.2.2 Bone resorption.....	13-14
1.2.3 Bone diseases	14-15
1.3 Cystatins.....	16-29
1.3.1 The cystatin superfamily.....	16-22
1.3.2 The mode of action of cystatin.....	22-25
1.3.3 The structure of cystatin.....	25-27
1.3.4 Proposed functions of cystatins.....	27-29
1.4 Cathepsin K.....	30-32
1.4.1 The cathepsin K gene.....	30-31
1.4.2 The cathepsin K protein.....	31-31
1.4.3 Cathepsin K in health and disease.....	31-32
1.5 Secreted phosphoprotein 24.....	33-44
1.5.1 The isolation of spp24 and the determination of its amino acid and cDNA sequence.....	33-34
1.5.2 The expression of spp24.....	34-34
1.5.3 The structure of the spp24 protein and homologies with known proteins.....	34-36
1.5.4 Speculated functions of spp24.....	36-40
1.6 Aims and objectives.....	41-43
2.0 Materials and methods.....	45-69
2.1 Centrifugation.....	45-45
2.2 Storage and handling of <i>Escherichia coli</i> (<i>E. coli</i>).....	45-48
2.2.1 Storage.....	45-45
2.2.2 Media.....	45-45
2.2.3 <i>E. coli</i> strains.....	45-46
2.2.4 Antibiotics.....	46-46
2.2.5 Preparing and transforming chemically competent <i>E. coli</i> cells.....	46-46
2.2.6 Preparing and electroporating electrocompetent <i>E. coli</i> cells.....	46-48
2.2.7 Selection for transformants.....	48-48
2.3 Use of restriction endonucleases.....	49-49
2.4 Agarose gel electrophoresis.....	49-49
2.5 Ethanol precipitation.....	49-50
2.6 Isolation of plasmid DNA from <i>E. coli</i>	50-52
2.6.1 Standard miniprep.....	50-50
2.6.2 Preparation of plasmid DNA using Qiagen kits.....	51-52

2.7 DNA extraction from human blood.....	52-52
2.8 Extraction of RNA from mammalian tissues.....	52-54
2.8.1 RNA extraction using the guanidinium-lithium chloride method.....	52-53
2.8.2 RNA extraction using the RNAzol B kit.....	53-54
2.9 Recovery of DNA from an agarose gel.....	54-55
2.9.1 Recovery of DNA from an agarose gel using phenol/ chloroform.....	54-54
2.9.2 Recovery of DNA from an agarose gel using the QIAquick gel extraction kit.....	55-55
2.10 Hybridisations.....	55-60
2.10.1 Southern blotting.....	55-58
2.10.1.1 Blotting the gel.....	56-56
2.10.1.2 Preparation of the probe.....	56-56
2.10.1.2.1 Preparation of oligo labelling buffer (OLB).....	56-57
2.10.1.3 Checking incorporation of the probe.....	57-57
2.10.1.4 Hybridisation of the probe.....	57-57
2.10.1.5 Post-hybridisation washes.....	57-57
2.10.1.6 Autoradiography.....	58-58
2.10.2 Colony hybridisations.....	58-58
2.10.3 Hybridisation to an RNA array.....	58-60
2.10.3.1 Probe preparation.....	58-58
2.10.3.2 Probe purification.....	58-59
2.10.3.3 Hybridisation of the probe.....	59-59
2.10.3.4 Post-hybridisation washes.....	59-59
2.10.3.5 Visualisation of MTE RNA array hybridisation results.....	60-60
2.11 Polymerase chain reaction (PCR).....	60-62
2.11.1 Standard PCR.....	60-60
2.11.2 Radioactive PCR.....	61-61
2.11.3 RT-PCR.....	61-62
2.11.4 Purification of PCR products using the QIAquick PCR purification kit.....	62-62
2.12 Polyacrylamide gels.....	62-64
2.12.1 Preparing the plates.....	62-63
2.12.2 Pouring the gel.....	63-63
2.12.3 Gel electrophoresis.....	63-63
2.12.4 Post electrophoresis.....	63-64
2.13 Cloning procedures.....	64-64
2.13.1 Dephosphorylation.....	64-64
2.13.2 Ligation.....	64-64
2.13.3 Cre- <i>loxP</i> reaction.....	64-64
2.14 Sequencing.....	65-66
2.14.1 Manual sequencing.....	65-66
2.14.1.1 Preparing double stranded DNA for sequencing....	65-65
2.14.1.2 The sequencing reaction.....	65-66
2.14.1.3 Gel electrophoresis.....	66-66
2.14.2 Automated sequencing.....	66-66
2.15 5'RACE.....	66-67
2.15.1 Reverse transcription.....	67-67
2.15.2 Purification and tailing.....	67-67
2.15.3 PCR of dA-tailed cDNA.....	67-67
2.16 Primer extension.....	67-68

2.17 Bioinformatics.....	68-69
2.17.1 Computing facilities used.....	68-68
2.17.2 Software used.....	68-68
2.17.3 GCG v.9.1 molecular biology package programs.....	68-69
2.17.4 Primer design.....	69-69
2.18 Safety issues.....	69-69
3.0 The structure of the human and mouse genes that encode the protein secreted phosphoprotein 24.....	70-115
3.1 Introduction.....	70-79
3.1.1 The human gene encoding secreted phosphoprotein 24.....	70-75
3.1.2 The mouse gene encoding secreted phosphoprotein 24.....	75-76
3.1.3 A possible insertion/deletion polymorphism in the human <i>SPP2</i> gene	76-79
3.2 Results.....	80-109
3.2.1 The determination of the exon/intron boundaries in the human <i>SPP2</i> gene.....	80-85
3.2.2 An extensive sequence analysis of the human <i>SPP2</i> gene.....	85-90
3.2.3 The determination of the start of transcription in the human <i>SPP2</i> gene.....	90-92
3.2.4 The determination of the mouse <i>Spp2</i> cDNA sequence.....	92-97
3.2.5 The determination of the exon/intron boundaries of the mouse <i>Spp2</i> gene.....	97-104
3.2.6 A comparison of the spp24 promoter region between human, mouse and chicken.....	104-107
3.2.7 Determination of the nature of the possible insertion/deletion polymorphism seen in the human <i>SPP2</i> gene that was originally reported by Gill and Dagleish	107-109
3.3 Discussion.....	110-115
4.0 The expression of the gene encoding secreted phosphoprotein 24.....	116-148
4.1 Introduction.....	116-119
4.1.1 The use of expressed sequence tags (ESTs) to obtain expression data.....	116-117
4.1.2 The use of northern blot analysis, ribonuclease protection assays and RT-PCR to obtain expression data.....	117-118
4.1.3 The use of microarrays to obtain expression data.....	118-119
4.2 Results.....	120-137
4.2.1 Expression data obtained from ESTs.....	120-122
4.2.2 The use of RT-PCR to carry out an expression study in mouse.....	122-128
4.2.3 Hybridisation of human <i>SPP2</i> cDNA to an RNA array.....	128-132
4.2.4 Expression data for spp24 in mouse from the RIKEN READ database.....	133-135
4.2.5 Microarray results from Incyte Genomics Inc. with respect to spp24 and osteoblasts.....	135-137
4.3 Discussion.....	138-147
5.0 A comparison of the spp24 protein between species.....	148-166
5.1 Introduction.....	148-148
5.2 Results.....	149-162
5.2.1 An anomaly observed in rat spp24 ESTs.....	149-152
5.2.2 A chicken hypothetical protein (Accession number Q91982) showing homology to spp24.....	152-159

5.2.3 Generation of the mouse and pig spp24 protein sequence.....	159-159
5.2.4 The alignment of the spp24 protein from six species.....	159-162
5.3 Discussion.....	163-166

6.0 Protein homologies and protein modelling.....167-195

6.1 Introduction.....	167-170
6.1.1 Proteins showing homology to spp24.....	167-167
6.1.2 Computer-based analysis of the spp24 protein.....	167-168
6.1.3 Constructing a protein model for spp24 using an evolutionary trace analysis technique.....	168-170
6.2 Results.....	171-185
6.2.1 Proteins showing homology to spp24.....	171-171
6.2.2 Computer-based analysis of the spp24 protein.....	171-176
6.2.3 Constructing a protein model for spp24 using an evolutionary trace analysis technique.....	176-185
6.3 Discussion.....	186-195

7.0 Concluding remarks and future work..... 196-201

7.1 Concluding remarks.....	196-199
7.2 Future work.....	199-201

Appendices

Bibliography	202-219
---------------------------	----------------

List of Figures

Figure 1.1	Bone remodelling	pg	11
Figure 1.2	A diagrammatic representation of the three types of cystatins	pg	17
Figure 1.3	A schematic representation of three proteins reported by Hu <i>et al.</i> to have some homology or similarity to spp24.	pg	19
Figure 1.4	Figure showing the evolutionary relationship between all known human type 1, type 2 and type 3 cystatin domains	pg	20
Figure 1.5	Figure showing the mode of action of a thiol protease	pg	23
Figure 1.6	The highly conserved residues of cystatin thought to be functionally important	pg	24
Figure 1.7	The structure of chicken egg white cystatin and its interaction with papain	pg	26
Figure 1.8	The bovine cDNA and amino acid sequence	pg	35
Figure 1.9	A schematic representation of the structure of spp24	pg	37
Figure 3.1	The <i>EcoRI</i> fragments of the <i>SPP2</i> gene that contain coding sequence and the human <i>SPP2</i> cDNA sequence	pg	72
Figure 3.1A	The pattern of bands seen on the autoradiographs when 5 different placental DNAs were digested with <i>BglII</i> , <i>SstI</i> and <i>HpaI</i> and the two different haplotypes observed	pg	77
Figure 3.1B	The insertion/deletion polymorphism theory postulated by Gill and Dalglish (unpublished) in relation to the restriction enzyme <i>BglII</i>	pg	79
Figure 3.2	The placement of some of the human <i>SPP2</i> <i>EcoRI</i> fragments against the human <i>SPP2</i> cDNA	pg	81
Figure 3.3	The exon/intron boundaries of the human <i>SPP2</i> gene as determined by the sequencing strategy described in section 3.1	pg	82
Figure 3.4	Digests of PAC clones containing <i>SPP2</i> and <i>EcoRI</i> fragments of the human <i>SPP2</i> gene, with <i>KpnI</i> and <i>SphI</i>	pg	83
Figure 3.5	The exon/intron structure of the human <i>SPP2</i> gene	pg	86
Figure 3.6	NIX analysis of AC006037 that contains the human <i>SPP2</i> gene	pg	88
Figure 3.7	The location of the primers used in the 5'RACE and the product generated	pg	91
Figure 3.8	The results of primer extension carried out on human total RNA from liver and kidney	pg	93
Figure 3.9	The consensus mouse <i>Spp2</i> cDNA determined by the alignment of mouse ESTs	pg	96
Figure 3.10	The open reading frames of the human, bovine, mouse, rat, pig and chick cDNA	pg	98
Figure 3.11	The three mouse <i>Spp2</i> cDNA probes used to identify the regions covered by the mouse positive clones	pg	101
Figure 3.12	The results of hybridisations to the mouse positives using three different regions of the mouse <i>Spp2</i> cDNA as a probe	pg	102
Figure 3.13	The nature of the RFLPs that lie within the human <i>SPP2</i> gene with respect to the genomic sequence AC006037	pg	108

Figure 4.1	The position of mouse RT-PCR primers with respect to the mouse cDNA sequence	pg	125
Figure 4.2	RT-PCR performed on RNA from adult mouse tissues	pg	126
Figure 4.3	RT-PCR performed on RNA from adult mouse tissues	pg	127
Figure 4.4	The autoradiographs of the human <i>SPP2</i> cDNA hybridised to the Clontech human MTE array	pg	131
Figure 4.5	Comparison of genes with similar expression profiles to <i>spp24</i> that were identified using READ	pg	147
Figure 5.1	The regions of the <i>spp24</i> rat protein that could be missing if the ESTs with anomalies were translated	pg	151
Figure 5.2	RT-PCR of Lyon Normotensive and Lyon Hypertensive rats	pg	153
Figure 5.3	The protein domains of <i>spp24</i> and Q91982 identified by ProDom 99.1 and the alignment of these two proteins with and without the translated 5'UTR of the gene encoding the hypothetical chicken protein (Q91982)	pg	154
Figure 5.4	An alignment of the original chicken sequence (Q91982), the amended chicken sequence and the human <i>spp24</i> sequence	pg	156
Figure 5.5	The composition of the published chicken GHRG-1 cDNA and promoter sequence and the correct chicken cDNA sequence with some exon/intron boundaries defined	pg	158
Figure 5.6	The chicken GHRG-1 promoter region	pg	160
Figure 5.7	The alignment of the <i>spp24</i> protein from six different species and the generation of a consensus sequence	pg	161
Figure 6.1	The consensus secondary structure of human <i>spp24</i> and a typical cystatin (chicken egg white cystatin) as determined by NPS@	pg	173
Figure 6.2	The charge distribution in the human <i>spp24</i> protein sequence	pg	175
Figure 6.3	The residues in the cystatin-like region of <i>spp24</i> that are identical between species and the residues that are predicted to be buried within the protein	pg	177
Figure 6.4	The evolutionary tree produced by the program 'Growtree' that is part of the GCG Molecular Biology Package	pg	179
Figure 6.5	A 'short' log file from the program 'BRUTUS' showing the analysis of the cystatin region of <i>spp24</i> and homologous proteins	pg	180
Figure 6.6	The images of '1cewi' from each PIC level where residues appear in the 'BRUTUS' program	pg	182
Figure 6.7	The structure of '1cewi' (chicken egg white cystatin) showing the residues that are absolutely conserved and the core residues of the clusters that are class-specific in the evolutionary trace analysis	pg	183
Figure 6.8	The six clusters in the human <i>spp24</i> amino acid sequence thought to be functionally important	pg	184
Figure 6.9	The position in which the non-cystatin-like region of <i>spp24</i> could lie on the cystatin-like domain	pg	191
Figure 6.10	The structure of '1cewi' showing the position of the 'N' residue involved in legumain inhibition in CAMP and the 'N' residue located close by in <i>spp24</i>	pg	193
Figure 6.11	The location of the highly conserved N-terminal region of the mature <i>spp24</i> protein	pg	194
Figure 7.1	<i>Spp24</i> as an evolutionary intermediate	pg	197

List of Tables

Table 1.1	Noncollagenous proteins present in bone	pg	5
Table 1.2	Hormones and other systemic factors influencing osteoblasts	pg	12
Table 1.3	The extent of phosphorylation in the cluster of serine residues seen in purified bovine spp24	pg	38
Table 2.1	Antibiotics	pg	47
Table 3.1	The programs used in the HGMP NIX analysis environment	pg	74
Table 3.2	The exon/intron boundaries of the human <i>SPP2</i> gene	pg	87
Table 3.3	Mouse ESTs that were aligned to generate the consensus mouse <i>Spp2</i> cDNA sequence	pg	95
Table 3.4	The layout of the 96-well microtitre plate containing the positives obtained from the screening of the mouse small insert library	pg	100
Table 3.5	Scoring of the mouse <i>Spp2</i> cDNA hybridisations	pg	103
Table 3.6	The sequences from the mouse trace archives (NCBI) that contain mouse <i>Spp2</i> exons	pg	105
Table 3.7	The exon/intron boundaries of the mouse <i>Spp2</i> gene	pg	106
Table 3.8	A comparison of exon size between the exons of a typical cystatin and those seen in the cystatin-like region of the human <i>SPP2</i> gene	pg	110a
Table 4.1	The number of mouse spp24 ESTs from various tissues	pg	121
Table 4.2	The number of human spp24 ESTs from various tissues	pg	123
Table 4.3	The layout of the human Clontech MTE array	pg	129
Table 4.4	The results of a phosphorimage from a 24-hour exposure of the Clontech human MTE array, hybridised with human <i>SPP2</i> cDNA	pg	132
Table 4.5	The microarray expression data obtained for mouse spp24 from READ	pg	134
Table 4.6	A comparison of the expression level of <i>SPP2</i> between osteoblast precursor cells and mature osteoblasts	pg	136
Table 4.7	A summary of the spp24 expression data obtained for humans	pg	139
Table 4.8	A summary of the spp24 expression data obtained for mouse	pg	140
Table 4.9	Spp24 expression data from human, mouse and bovine tissues	pg	143
Table 5.1	Rat ESTs from the UniGene cluster Rn.84	pg	150
Table 6.1	The protein programs used to analyse the spp24 protein	pg	169
Table 6.2	The proteins identified that have a significant level of homology with spp24 either at the amino acid sequence level or with respect to the structure of the domains of the protein	pg	172
Table 6.3	The subgroups at 25% PIC and the residues they have at the cluster around the first and third cysteine	pg	189

Chapter 1

Introduction

1.1 General Introduction

Secreted phosphoprotein 24 (spp24) is a novel 24-kDa non-collagenous protein that was originally isolated from the acid demineralised extract of bovine cortical bone (Hu *et al.* 1995). The spp24 protein is described in more detail in section 1.5.

The use of twin studies has shown that bone mineral density has significant genetic factors (Smith *et al.* 1973). These could act alone or in conjunction with environmental factors to predispose people to the development of osteoporosis. If spp24 has a functional role in the process of bone remodelling it is likely that it may influence the development or severity of osteoporosis.

If spp24 exhibits a cystatin-like function then it is likely that the thiol protease(s) it inhibits is also found in bone. Cathepsin K is a thiol protease for which no natural inhibitor has yet been identified. Cathepsin K has been shown to be expressed specifically in osteoclasts (Tezuka *et al.* 1994; Brömme *et al.* 1996; Drake *et al.* 1996; Rantakokko *et al.* 1996; Dodds *et al.* 1998). A mutation in the gene has been shown to give rise to pycnodysostosis (Gelb *et al.* 1996a), a metabolic bone disease. Cathepsin K has also been implicated in osteoporosis (reviewed by Zaidi *et al.* 2001) and synthetic cathepsin K inhibitors have been used as treatments of osteoporosis (Votta *et al.* 1997). Consequently, cathepsin K is a good candidate for interaction with spp24.

There are many other bone diseases for which the genetic basis has been identified, but that have varying degrees of severity in the phenotype presented. One example of this is osteogenesis imperfecta (Raghunath 1995). It would seem that there must be further 'modifying' factors that act in conjunction with the primary genetic defect to influence the severity of the phenotype. The presence of spp24 in bone makes this protein a good candidate for being one of the factors that may act to 'modify' the phenotypes seen in these cases.

The cystatins that are the most homologous to the N-terminal segment of the spp24 protein do not have typical cystatin activity. These proteins include the kininogens and fetuins (Hu *et al.* 1995) for which there has been much speculation on many possible functions (Takagaki *et al.* 1985; Brown *et al.* 1992) and bradykinin and neutrophil antibiotic peptides (Hu *et al.* 1995)

for which antimicrobial functions have been shown (Romeo *et al.* 1988). It is therefore possible that spp24 has an entirely different function to the cystatin activity speculated. It is quite possible that it has multiple functions.

The aim of this project was therefore to characterise the structure and expression of the human *SPP2* gene and to begin to elucidate some possible functions for the spp24 protein. The following sections describe each aspect discussed above in more detail.

1.2 Bone

Spp24 was originally isolated from bovine cortical bone (Hu *et al.* 1995). This was the only information available about the localisation of the protein. The presence of the protein in bone immediately suggested a possible role in the processes that occur there. It is therefore important to understand the structure, composition and remodelling of bone in order to speculate some of the possible functions of spp24.

Bone is a specialised support tissue in which the extracellular matrix is mineralised. This mineralisation gives bone its characteristic hard and rigid properties. The main functions of bone are to provide mechanical support, protect internal organs, bring about locomotion and absorb stress. In conjunction with other organs, bone is also involved in mineral homeostasis and provides a source of calcium and other inorganic ions. To accommodate all these functions, bone is in a constant dynamic state of growth and resorption.

There are numerous textbooks available covering the properties of bone and several of these have been used as a source of information for the general overview that follows in sections 1.2.1 to 1.2.3 (Dickson 1993; Schenk 1993). Individual bones have evolved to optimise their properties to specialised functions, but the basic properties apply to all bone.

1.2.1 Bone Structure

About 20 to 30% of cortical bone is organic extracellular matrix (osteoid), approximately 10% is water and the remainder is inorganic mineral salts (*e.g.* calcium).

1.2.1.1 The extracellular matrix of bone

The specialised organic extracellular matrix of bone is called osteoid. The osteoid is about 90% collagen, most of which is type I collagen (Herring 1972; Broek *et al.* 1985).

The type I collagen molecule is a heterotrimer, consisting of two $\alpha 1$ (I) chains and one $\alpha 2$ (I) chain. The α -chains are very similar and comprise about 1000 amino acids. The details of the collagen triple helix were established by X-ray diffraction techniques (Cowan *et al.* 1953) and it was found that through 95% of the α -chains a glycine occurred at every third residue. There was also a high proline and hydroxyproline content. The regular occurrence of these amino

acids gives rise to a polymer of tripeptide units and a helical conformation of individual α -chains. The three α -chains in type I collagen together form a triple helical conformation. The triple helix α -chains are then organised into fibrillar structures, which are then bundled together to form collagen fibres.

In contrast to soft tissues, the synthesis of type I collagen in bone is particularly susceptible to the effects of hormones and other factors, having consequences on the rate of bone formation (Raisz and Kream 1983). Also in contrast to soft tissues, changes in plasma calcium affects the degree of collagen lysine hydroxylation by osteoblasts. Consequently, bone formation can be influenced by changes in mineral metabolism such as that seen in hypocalcaemia or rickets (Dickson *et al.* 1979).

The organic collagen in bone is embedded in a glycosaminoglycan gel. This gel contains proteoglycans (protein-carbohydrate complexes) which consist of a small polypeptide core to which are attached glycosaminoglycan side chains. Most of the glycosaminoglycan present in bone is chondroitin-4-sulphate although it has been estimated that 12-14% in adult human compact bone are chondroitin-6-sulphate (Hjerpe *et al.* 1979).

Bone proteoglycans are present in two forms, biglycan and decorin (Fisher *et al.* 1983). Biglycan and decorin are not bone specific and have a core protein of about M_r 38,000 to which are attached two or one chondroitin sulphate chains respectively.

The proteoglycans account for about 10% of the non-collagenous proteins in bone and are thought to control the water content of bone and to regulate the formation of collagen fibres in a form appropriate for subsequent mineralisation.

Spp24 was isolated as a non-collagenous protein of bone (Hu *et al.* 1995) and consequently it is of interest to look at the roles of the other non-collagenous proteins known to localise to bone. Table 1.1 (Dickson 1993) summarises the details of these non-collagenous proteins.

Several plasma proteins have been identified in bone. Some are there purely due to the presence of blood vessels in bone and others are specifically enriched in bone. Albumin and α_2 HS-glycoprotein are the two most abundant plasma proteins found in bone (Dickson 1974). Albumin is thought to be there due to both of the reasons mentioned above (Owen and Triffitt 1976). Much more is known about α_2 HS-glycoprotein and its origin and role in bone.

Table 1.1. Noncollagenous proteins present in bone (adapted from Dickson 1993 and Young *et al.* 1991).

Name(s)	Relative molecular mass	Structural features	Potential function(s)	Sites of synthesis	% of total bone noncollagenous protein
Albumin	68,000	Presence due to blood vessels	Presence due to blood vessels	Liver	3
α_2 HS-glycoprotein	51,000	Cystatin domains	Many speculated such as role in inflammatory response and bone remodeling?	Liver	5
Osteocalcin/ Bone Gla-protein	5,800	γ carboxylation of glutamic acid	Bone turnover?	Bone Dentine	15
Matrix Gla-protein	14,000	γ carboxylation of glutamic acid	Unknown	Bone Dentine Cartilage	2
Proteoglycan I/ PGI/ PG-SI	118,000	Leucine repeat structure, two GAG chains near NH2 terminus	Cell-cell or cell-protein interactions	Bone Cartilage Aorta	10
Proteoglycan II/ PGII/PG-SII	78,000	Leucine repeat structure, one GAG chain near NH2 terminus	Binds collagen, regulates fibril formation	Bone Eye Tendon	
Osteonectin/ SPARC/BM-40	32,000	EF hand consensus, acidic NH2 terminus	Ca ²⁺ and hydroxyapatite binding, cell spreading	Bone Skin Tendon Ligament Platelets Basement membrane	15
Osteopontin/ Sialoprotein I	44,000	RGD amino acid sequence, phosphorylation	Cell attachment/HA binding	Bone Epithelium Placenta Decidua	10
Bone sialoprotein/ Sialoprotein II	75,000	RGD amino acid sequence, sulphation of tyrosines	Cell attachment	Bone	10

Note: There is considerable species variation in the amounts present of each noncollagenous protein. The values shown for percentage of total noncollagenous protein are representative ones based on analytical data for human and animal bone from a number of laboratories.

Alpha₂HS-glycoprotein is a 50 kDa fetuin protein that is synthesised in the liver (Triffit *et al.* 1976) and found as a minor component of plasma. It has been suggested that the protein has a role in bone resorption and remodelling. Fetuins will be discussed in more detail in Chapter 4 and 6 as this protein shows some homology to spp24 and the non-collagenous proteins found in bone that is of greatest interest.

The remaining non-collagenous organic material includes osteocalcin (Gla protein), which is involved in binding calcium, osteonectin which may serve some bridging function between collagen and the mineral component and osteopontin (a sialoprotein), which is high in sialic acid.

Osteocalcin or bone Gla protein contains residues of γ -carboxyglutamic acid (Gla), an amino acid originally found in prothrombin that binds calcium (Hauschka *et al.* 1975). Osteocalcin is present only in bone and dentine (deVries *et al.* 1988) and is one of the most abundant noncollagenous bone proteins of most species (Triffit 1987).

Osteocalcin contains three residues of glutamic acid that become posttranslationally modified to become γ -carboxylated, though, in humans it has been reported that the first glutamic acid does not become fully carboxylated (Poser *et al.* 1980) and there are lower levels of the protein present in human bone than in other species.

In the presence of millimolar quantities of calcium, osteocalcin adopts an α -helical conformation (Hauschka and Carr 1982). Two regions of α -helical conformation are separated by a β -turn, which is stabilised by a disulphide bond between two cysteine residues. The structure is necessary for binding hydroxyapatite (mineral salt) and the sequence of osteocalcin is highly conserved between species. Osteocalcin may be important in bone resorption as purified osteocalcin has been shown to influence the recruitment of osteoclast precursors (Malone *et al.* 1982).

Another Gla-containing protein that is present at much lower concentrations in bone than osteocalcin is matrix Gla protein (MGP). MGP is very similar to osteocalcin. In fact, it is thought that a gene duplication event from a common ancestor may have given rise to matrix Gla protein and osteocalcin. However, osteocalcin is very soluble in neutral aqueous solutions, whereas matrix Gla protein is only soluble in the presence of strong dissociating

solvents such as 4 M guanidinium chloride. This may be a factor important in ensuring the retention of MGP in the matrix.

Amongst the noncollagenous proteins of bone are several phosphoproteins. These are relevant to this thesis because spp24 has a region of serine residues that are thought to be phosphorylated (section 1.5). Phosphoproteins are present in bone and dentine, but unfortunately those present in bone are less well characterised.

Bone phosphoproteins can be phosphorylated on serine, threonine and aspartic acid residues and can be enriched in glutamic acid. Evidence suggests that they are synthesised locally (Glimcher *et al.* 1984). Sixteen phosphoproteins ranging in molecular weight from 4,000 to 150,000 were isolated from chicken bone (Uchiyama *et al.* 1986). Although it was shown that many of the low molecular weight proteins were probably derived from larger precursors (Yamazaki *et al.* 1988).

Osteopontin (also called sialoprotein I) is one of the most well characterised bone phosphoproteins. Osteopontin contains an Arg-Gly-Asp-Ser sequence (Oldberg *et al.* 1986) similar to the cell attachment domain of fibronectin. It also binds strongly to hydroxyapatite; consequently osteopontin is able to form a bridge between cells and mineralised matrix. Osteopontin is not bone specific and high levels were also found in placenta, epithelium and decidua (Nomura *et al.* 1988).

A further phosphoprotein found in bone is bone sialoprotein or sialoprotein II. This protein has a lower level of phosphorylation than osteopontin (sialoprotein I) although its function is not clear.

In several species the major phosphoprotein is osteonectin. This is a protein of approximately M_r 32,000 that is not specific to bone and has also been found in skin, tendon, ligaments, platelets and basement membrane (Wasi *et al.* 1984). Osteonectin may be important in osteogenesis imperfecta as in a bovine form of this disease osteonectin levels were down to 1.2% of normal whereas the sialoprotein content was 48.9% that of normal (Termine *et al.* 1984). However, it was not clear whether this was a cause or an effect.

Osteoid (extracellular matrix) is synthesised by osteoblast cells which are discussed further in sections 1.2.1.3

1.2.1.2 Bone Mineral

About 70% of mature compact bone is made up of inorganic mineral salts which, in the osteoid (extracellular matrix) are what give bone its hardness. The mineral salt is in the form of hydroxyapatite ($\text{Ca}_{10}(\text{PO}_4)_6(\text{OH})_2$) crystals and is a complex of calcium and phosphate hydroxides (Posner 1987). This complex conjugates to a small proportion of magnesium carbonate, sodium and phosphate ions and also has an affinity for heavy metal and radioactive environmental pollutants.

Concentrations of Ca^{2+} and PO_4^{3-} ions must be above a threshold value for mineralisation to occur and there are several factors that influence this (Boskey 1981). Osteocalcin has an inhibitory effect on hydroxyapatite formation (Boskey *et al.* 1985).

Alkaline phosphatase, an enzyme found in osteoblasts, increases local Ca^{2+} and PO_4^{3-} ion concentrations. Matrix vesicles, probably derived from the cell membrane, bud off from osteoblasts during osteoid formation. These vesicles are rich in alkaline phosphatase and also pyrophosphatase (which inhibits mineralisation), both of which can cleave PO_4^{3-} ions from larger molecules. It is thought that these vesicles are the sites for the initial precipitation of amorphous (non-crystalline) calcium phosphate into hydroxyapatite crystals. About 20% of the mineral component remains in the amorphous form, providing a readily available buffer in calcium homeostasis.

Under normal local concentrations of Ca^{2+} and PO_4^{3-} ions, mineralisation occurs shortly after the new osteoid has been formed. However, when there is high bone turnover, mineralisation can lag behind. This can be seen in foetal bones and also in the healing of fractures.

1.2.1.3 The cells found in bone

There are three main cell types found in bone, osteoblasts, osteocytes and osteoclasts. These cells are derived from two different cell lines. Osteoblasts and osteocytes are derived from osteogenic mesenchymal stem cells (Friedenstein 1973) and osteoclasts are derived from fused mononuclear haematopoietic stem cells (Takahashi *et al.* 1988).

Pre-osteoblasts and osteoblasts line the bone surface and their main functions on activation are to synthesise the osteoid and to regulate osteoclast access to the bone surface.

As osteoblasts form osteoid they become engulfed by it, losing size and organelles. They are then known as osteocytes. Osteocytes are connected to each other and to other bone lining cells via cytoplasmic extensions. These act as pathways for communication through the bone matrix (Palumbo *et al.* 1990). Mature osteoclasts are large, multinucleated cells that are found wherever bone is being removed (Kölliker 1873).

The functions of these cells and their involvement in bone remodelling are discussed in more detail in section 1.2.2.

1.2.1.4 Bone patterns and bone architecture

Bone exists in two forms called woven and lamellar. Woven is an immature form of bone that is formed when the osteoblasts are rapidly producing osteoid. It is characterised by a random, loose organisation of collagen fibres and it is mechanically weak. Woven bone is also characterised by more numerous and larger osteocytes. It also often has a high mineral content due to deposition of apatite in the interfibrillar spaces as well as within collagen fibrils. Woven bone is the main bone pattern of foetal bones, but as the bone matures it becomes substituted by lamellar bone. In adults, woven bone is only found when there is a rapid formation of new bone, such as repairing a fracture.

Lamellar bone is very strong and is characterised by a regular parallel alignment of collagen fibres into successive layers (Marotti and Muglia 1988). Virtually all bone in a healthy adult is lamellar. When lamellar bone is formed as a solid mass it is called compact bone and when it forms a more open structure it is referred to as cancellous bone.

Most bones are composed of an outer cortical zone of compact bone and an inner trabecular zone of cancellous bone. The outer cortical zone is rigid and provides protection and support. The inner trabecular zone provides strength. The spaces between the trabecular meshwork are occupied by bone marrow, the main site of haemopoiesis.

1.2.2 Bone Remodelling

Bone is a tissue that is constantly remodelling itself. This process is critical to heal the damage caused by infections and fractures. It also serves to maintain the bone morphology and mass at its optimum for the demands of the organism and to mobilise minerals as required. Spp24 was originally isolated from demineralised bone (rather than bone marrow)

and it was speculated that it may have a role in the processes of bone remodelling (Hu *et al.* 1995).

Bone remodelling can broadly be divided into four stages: activation, resorption, reversal and formation. Figure 1.1 summarises the processes of bone turnover (Raisz 1998).

1.2.2.1 Bone formation

Pre-osteoblasts and osteoblasts line the bone surface and their main functions on activation are to synthesise the extracellular matrix of bone (osteoid) and to regulate osteoclast access to the bone surface. The sequence of events during bone formation has been studied extensively using marrow stromal cells in diffusion chambers (Ashton *et al.* 1980) (Bab *et al.* 1986) (Mardon *et al.* 1987). Initially there is a formation of fibrous tissue that has a high collagen III expression.

During pre-osteoblast differentiation, the expression of proteins is low. Bone morphogenetic proteins (BMPs) activate the migration of mesenchymal cells and induce osteoblastic differentiation (Ogata *et al.* 1993). Another factor that affects osteoblast activity is parathyroid hormone (PTH) that is secreted from the parathyroid gland.

The main factor that stimulates the secretion of PTH is a fall in ionised plasma calcium levels, although other factors have been implicated such as catecholamines and 1,25-dihydroxyvitamin D₃, which inhibits secretion of PTH (Habener *et al.* 1984; Wong 1986). The effects of PTH are not fully understood, but osteoblasts have a membrane receptor and continuous treatment with PTH inhibits bone formation. However, intermittent exposure to PTH is thought to stimulate osteoblast proliferation and differentiation (Canalis *et al.* 1989).

The principal function of PTH is thought to be to help maintain plasma calcium homeostasis without disturbing the phosphate balance (Raisz and Kream 1983; Habener *et al.* 1984; Wong 1986). However, PTH is thought to indirectly affect osteoclast activity by inducing osteoblasts to secrete osteoclast-stimulating factors. Table 1.2 shows the hormones and other systemic factors that influence osteoblasts (adapted from Dickson 1993).

Osteoblasts also have sex hormone receptors. There is substantial evidence that oestrogen deficiency can lead to increased bone loss (Avioli 1983; Johnston 1985) and thus osteoporosis is common in postmenopausal women. Estrogen elevates osteoblast proliferation and also the

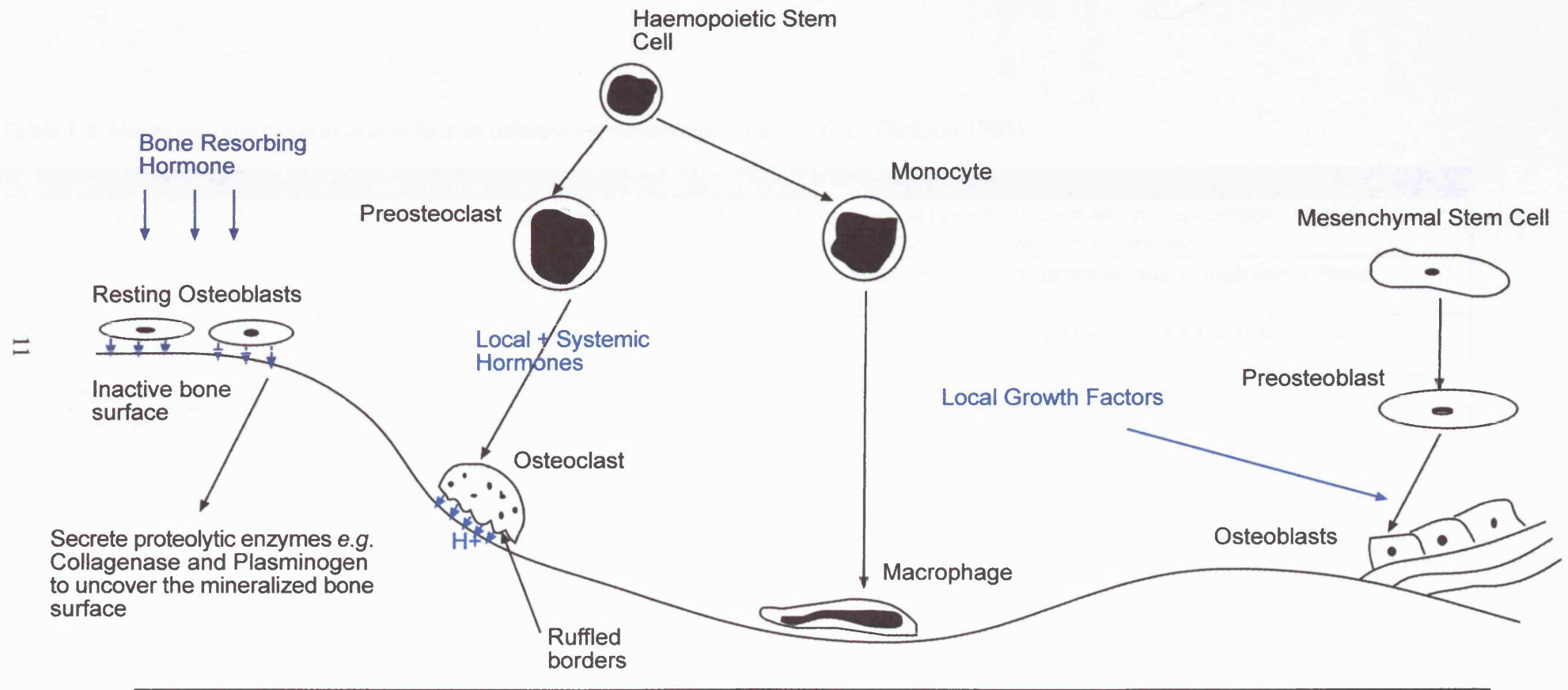


Figure 1.1. Bone remodelling.

This figure summarises the processes involved in bone remodelling. Mineralised bone surface is uncovered by resting osteoblasts, osteoclasts remove bone and osteoblasts form new bone. All cells are labelled accordingly. Blue indicates secretions or factors that act on cells.

Table 1.2. Hormones and other systemic factors influencing osteoblasts (adapted from Dickson 1993).

Name	Bone formation	Bone resorption	Effects <i>in vivo</i>
1,25-dihydroxyvitamin D ₃	Decrease	Increase	Necessary for normal bone formation and mineralisation. At high doses formation is decreased and resorption is increased
Parathyroid hormone	Decrease	Increase	At low doses, bone formation is increased and at high doses, bone resorption is increased
Cortisol	Increase and decrease	Decrease	At high doses, bone formation decreases and loss of bone occurs
17- β -estradiol	Increase	-	Estrogen deficiency leads to bone loss
Retinol-Retinoic acid	Decrease	Increase	At high doses, bone resorption increased
Insulin	Increase	-	Effects not well defined
Thyroxine/Triiodothyronine	Decrease	Increase	High doses lead to increased bone turnover and net loss of bone

Note: All of the factors mentioned above act on other target tissues besides bone. Their effects on bone formation and resorption result from the combination of both direct and indirect effects on osteoblasts and the pathways involved can be difficult to assess. The table indicates likely direct effects, based on data from studies *in vitro*, as well as the overall effect *in vivo*.

osteoblasts response to PTH. Estrogens also increase the expression of collagen genes and insulin-like growth factor 2 within the osteoblasts and may affect the production of lysosomal enzymes in osteoclasts (reviewed by Turner *et al.* 1994). Androgen receptors have also been found on osteoblasts (Colvard *et al.* 1989), but the exact effects and mechanisms of androgens are not clear.

Osteoblasts also secrete prostaglandins, PGE₂ being the most abundant (Rodan *et al.* 1981). PGE₂ is thought to stimulate bone formation and resorption. There is currently much speculation as to the exact roles of PGE₂, but it is thought that it can stimulate second messengers in osteoblasts or stimulate osteoclasts indirectly.

Other factors that can affect bone formation and hence osteoblasts include vitamin D, glucocorticoids, vitamin A, insulin and insulin-like growth factors and thyroid hormones. The mechanisms are complex and in some cases poorly understood.

As well as chemical factors there are also physical or mechanical factors that are by products of bone function. Long-term strain can promote bone formation.

1.2.2.2 Bone resorption

Haemopoietic stem cells differentiate into osteoclasts (reviewed by Nijweide *et al.* 1986). There are many cytokines, growth factors and hormones known to affect this differentiation, some of which have already been mentioned in previous sections.

The haemopoietic growth factors interleukin-3 and granulocyte-macrophage colony-stimulating factor, are important for the colony-forming unit for granulocytes and macrophages. Other local stimuli are required for further progression towards differentiated osteoclasts, such as 1,25(OH)₂D₃ (a derivative of vitamin D that influences both calcium levels and osteoclasts), PTH and tumour necrosis factor (TNF) (Suda *et al.* 1992). Some cytokines also have an effect such as interleukin-6 and interleukin-11 and may be significant in conditions such as postmenopausal osteoporosis (Jilka *et al.* 1992).

Mature osteoclasts are large, multinucleated cells that are found wherever bone is being removed. They have many mitochondria, vacuoles and lysosomes. Osteoclasts attach tightly to calcified matrix and towards the centre of the cell the membrane becomes folded to form the characteristic 'ruffled border' (Baron 1989). Under the ruffled border is a resorption pit

where an acidic environment is created by secretion of protons through a vacuolar proton pump (Blair *et al.* 1989) and lysosomal enzymes. The low pH in the resorption pit dissolves the mineral phase of the bone matrix and activates osteoclastic hydrolytic enzymes. The organic matrix is then dissolved by lysosomal enzymes, such as cathepsin B, and scalloped cavities known as Howship's lacunae are left in the surface of the bone.

Osteoclasts have calcitonin receptors and exposure to this hormone causes them to detach from the bone surface. Calcitonin is a calcium regulatory hormone. As calcium levels rise, calcitonin secretion follows.

Abnormalities in the osteoclastic resorption process that lead to a decrease in resorption can give rise to osteopetrosis and abnormalities leading to an increase in resorption can give rise to osteoporosis. There are several clinically important modulators of osteoclast activity. These include tamoxifen. Tamoxifen, an estrogen antagonist normally used in the treatment of breast cancer, has been shown to prevent bone loss (Love *et al.* 1992). Bisphosphonates are now also widely used to treat osteoporosis (reviewed by Reginster *et al.* 1997). They are incorporated in the place of phosphates into bone and they prevent osteoclast recruitment and activity (Rodan and Fleisch 1996).

The processes involved in bone turnover are complex and can be affected by other factors such as ageing, diet or the menopause. If spp24 does play a role in bone turnover then it could be involved in any of the processes described above. The similarity of spp24 to the cystatin (thiol protease inhibitor) family (discussed in section 1.5) led to initial speculation that spp24 may be secreted by osteoclasts as an enzyme to degrade components of the bone matrix.

1.2.3 Bone diseases

Of course there are many disorders and diseases that affect the bone, but the most common resulting bone conditions are osteoporosis, osteopetrosis and osteomalacia. Bone also has an involvement in cancer and maldevelopment.

Osteoporosis is a condition where both cortical and trabecular bone become thinned and are therefore more prone to fracture. There are many factors that are thought to increase the chances of osteoporosis; these include old age, diet, exercise, the menopause and some drugs. The most commonly affected area is the hip.

Osteopetrosis is a condition where bones become thicker and denser. It is less frequently associated with environmental factors such as those described for osteoporosis, but is usually associated with various bone disorders and conditions.

Osteomalacia is the failure of mineralisation in the osteoid. Mineralisation can only take place if there are sufficient Ca^{2+} and PO_4^{3-} ions. Low levels of Ca^{2+} ions can be due to inadequate dietary intake or malabsorption resulting from small intestinal disease. Less commonly, PO_4^{3-} ions can be low usually due to excessive loss in the urine. Patients with osteomalacia develop softening of the bone leading to an increased risk of fracture. Osteomalacia in children leads to the disease rickets. This results in permanently deformed bones.

Osteosarcoma is the most important tumour derived from bone cells. This is a malignant tumour of osteoblasts that is most common in children and usually involves the bone around the knee joint. The tumour cells produce osteoid, but in a haphazard, random way and do not mineralise normally. Osteosarcoma spreads extensively in the bloodstream and often produces metastatic tumours in the lungs.

Osteoid osteomas can occur which are benign tumours of the osteoblasts. As with osteosarcomas osteoid is produced, but this time there is more osteoid formation and an increase in the degree of mineralisation. Osteomas are benign and do not spread.

The bone marrow is a common site for metastasis of certain cancers, particularly the breast, bronchus, thyroid and kidney. Tumour cells proliferate and cause destruction of trabecular bone leading to an increased risk of fractures.

There are several diseases that arise as a result of impaired bone formation during early development. Many of these diseases are so severe that children die *in utero* or shortly after birth. One disease where the sufferers survive is achondroplasia. This is a form of dwarfism that is characterised by a normal sized trunk, but shortened limb bones.

Another disease which is a result of incorrect bone formation is sclerosteosis. This disease is very rare and has been described almost exclusively in Afrikaners of South Africa.

Sclerosteosis is characterized by gigantism, facial distortion and deafness caused by progressive bone overgrowth. The bones of an affected individual demonstrate excessive bone formation and a continual increase in bone mass, the opposite to osteoporosis.

1.3 Cystatins

1.3.1 The cystatin superfamily

Cystatins are also known as thiol protease inhibitors or cysteine proteinase inhibitors.

Cysteine proteinases are responsible for much of the intracellular proteolysis and are therefore abundant in the body (Kirschke *et al.* 1980). They are responsible for the processing of proenzymes and prohormones (Taugner *et al.* 1985; Marks *et al.* 1986), some aspects of bone resorption (Delaissé *et al.* 1984) and the breakdown of collagen (Etherington 1980).

Cysteine proteinases degrade proteins by cleaving peptide bonds and depend on a highly reactive thiol group of a cysteine residue at the catalytic site, for their catalytic activity (reviewed by Turk 1986; Barrett 1987). Cysteine proteinases have also been implicated in tumour invasion and metastasis (Sloane and Honn 1984) and in infection by microorganisms (Barrett *et al.* 1984).

Cystatins regulate the degradative actions of the cysteine proteinases and protect host tissues from destructive proteolysis by host, bacterial and viral cysteine proteinases (reviewed by Bobek and Levine 1992). The name cystatin was first applied to a protein isolated from chicken egg white (Barrett 1981) that was shown to inhibit papain (Fossum and Whitaker 1968) (a plant enzyme) and cathepsins B and C (Keilová and Tomášek 1974). Since this initial discovery, cystatin has been shown to be a potent inhibitor of many cysteine proteinases of the papain superfamily. Many proteins have been identified that are similar to chicken egg white cystatin in their structure and function and thus are members of the cystatin superfamily.

The cystatin superfamily was grouped into further subfamilies based on the size and complexity of the polypeptide chains (Barrett *et al.* 1986a,b); type 1 cystatins, type 2 and type 3. Figure 1.2 shows a diagrammatic representation of the three cystatin subfamilies (adapted from Barrett 1987).

Members of the type 1 family of cystatins are called stefins, which are usually about 100 amino acids (M_r 11,000) in length and have no disulphide bonds or carbohydrate groups. The type 1 cystatins are synthesised without signal peptides and are found primarily intracellularly. Two cystatins that are known to belong to this family are cystatin A and cystatin B.

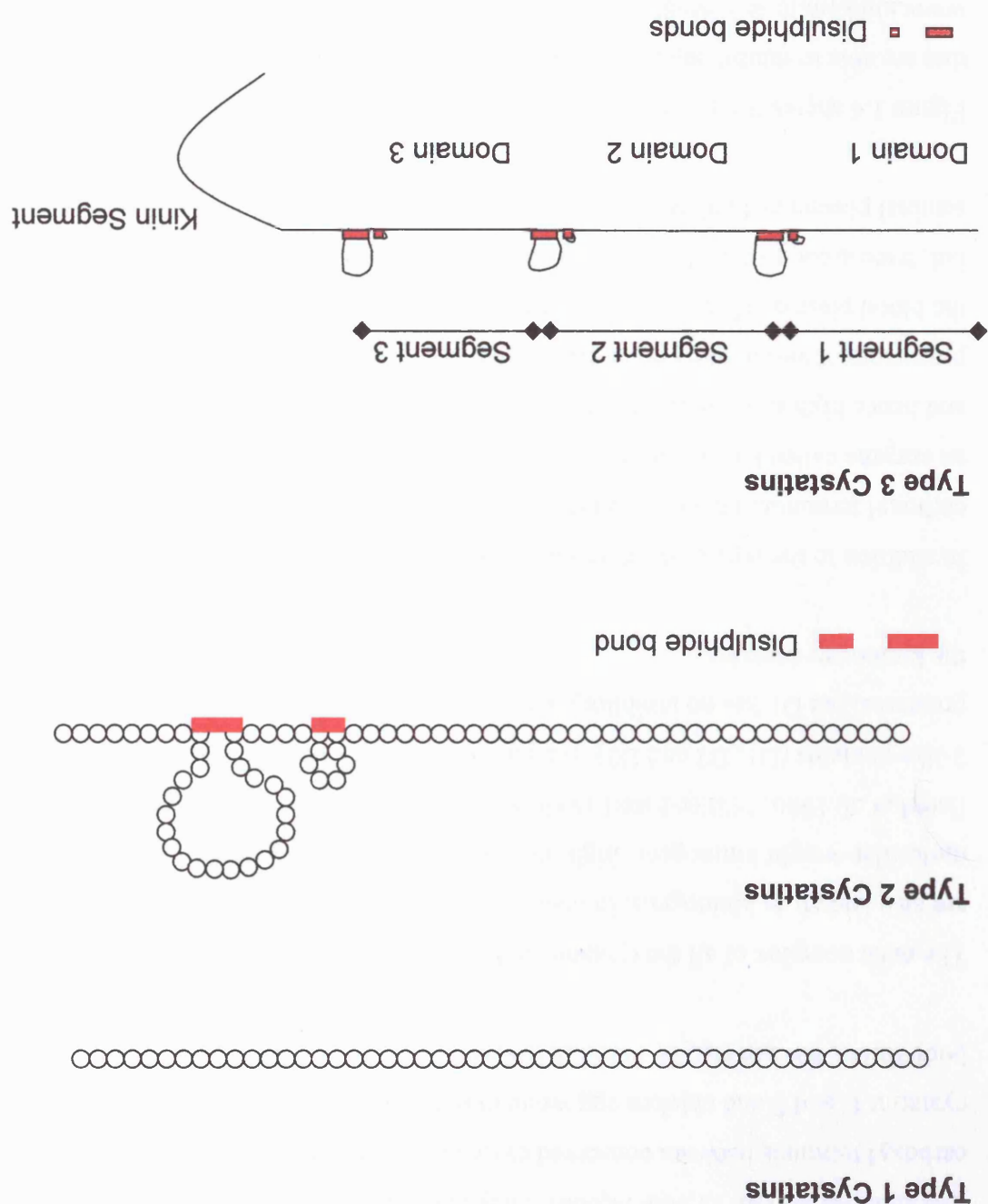


Figure 1.2. A diagrammatic representation of the three types of cystatins.

Type 1 cystatins or stefins, have no disulphide bonds and are represented here as a string of residues. Type 2 cystatins have a single cystatin domain that comprises two disulphide bonds. Type 3 cystatins or kininogens, have three cystatin domains that each contain two disulphide bonds. Type 3 cystatins also have a non-cystatin-like C-terminal segment (the kinin segment) that is released as a biologically active peptide. (Not drawn to scale).

More complex than the stefins are the secreted type 2 cystatins, which consist of about 115-120 amino acids (M_r 13,000-14,000). They contain at least two disulphide bonds towards the carboxyl terminus between conserved cysteine residues. Examples of type 2 cystatins are cystatins C and S and chicken egg white cystatin. Type 2 cystatins are found primarily in body fluids, but can also be found in tissues.

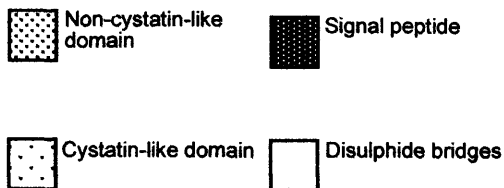
The most complex of all the cystatin subfamilies are the type 3 cystatins. The type 3 cystatins are also known as kininogens. In mammals there are three types of kininogens, low-molecular-weight kininogens, high-molecular-weight kininogens and T-kininogens (Müller-Esterl *et al.* 1986; Müller-Esterl 1989; Kato *et al.* 1981). Each kininogen contains three type 2-like domains (D1, D2 and D3). D2 and D3 are functionally active inhibitors of thiol proteases, but D1 has no inhibitory activity. Figure 1.3 shows a schematic representation of the kininogen domains.

In addition to the type 2-like domains, kininogens also have an unrelated polypeptide at the carboxyl terminus. This is called the bradykinin sequence and can be released by the action of an enzyme called kallikrein. Bradykinin plays a role in the intrinsic blood coagulation cascade and hence high and low molecular weight kininogens were first known as the biosynthetic precursors of vasoactive kinins. Kininogens are synthesised in the liver and then secreted into the blood plasma. They are found in the highest concentration in plasma and synovial fluid but, trace amounts are also found in other body fluids such as tears, cerebrospinal fluid, seminal plasma and colostrum (Abrahamson *et al.* 1986).

Figure 1.4 shows the evolutionary relationship between the known human cystatin domains that are able to inhibit papain-like cysteine proteinases (adapted from www.klinkem.lu.se/E/abrahamson/cystatin_text.html).

Since the late 1980s, when it was largely accepted that there were three types of cystatins in the cystatin superfamily, several new members have emerged. All the more recent additions to the superfamily are variations on the 'typical' cystatin structure described previously. Fetuin is a protein that was first discovered in its foetal bovine form in 1944 (Pederson 1944), but it was only much later that it was actually thought to be related to the cystatin superfamily (Elzanowski *et al.* 1988). As discussed in section 1.5, Hu *et al.* (1995) commented on the similarity of several properties between spp24 and fetuin. The human equivalent of fetuin was discovered in 1987 (Dziegielewska *et al.* 1987), human α_2 HS-glycoprotein. Fetuins have since been described for several other species.

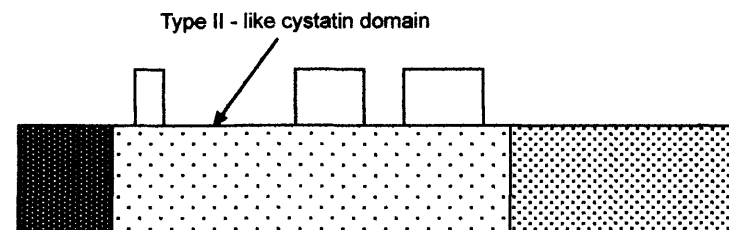
Key



N-terminus

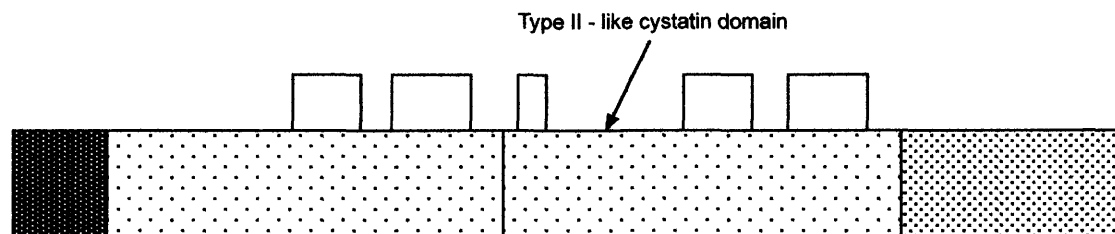
C-terminus

Cathelin



Type II - like cystatin domain

Fetuin



Kininogen

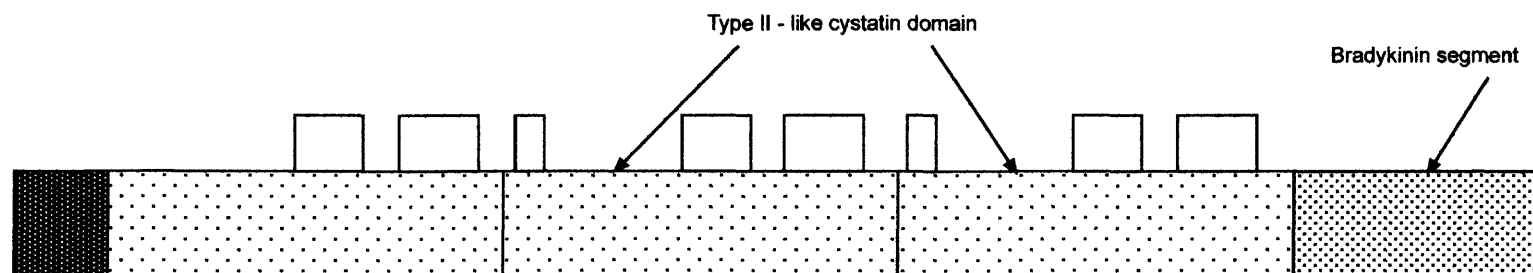


Figure 1.3. A schematic representation of three proteins reported by Hu *et al.* (1995) to have some homology or similarity to spp24.

The cathelins, although not actually part of the cystatin superfamily, have a single N-terminal domain containing two disulphide bonds that is cleaved to release the C-terminal antimicrobial domain as the mature peptide. The fetuins are variant cystatins that are part of the cystatin superfamily and have two cystatin domains at the N-terminal end of the protein. This is followed by a C-terminal domain. The fetuin protein is thought to circulate as a two-chain plasma protein with cleavage occurring somewhere in the C-terminal domain and the two chains being linked by a disulphide bond. Kininogen is a type 3 cystatin and has three cystatin domains at the N-terminal end of the protein. Only domains 2 and 3 have thiol protease inhibitory function. Kininogen is cleaved by kallikrein to release the C-terminal domain, bradykinin that is a biologically functional peptide involved in the blood coagulation cascade.

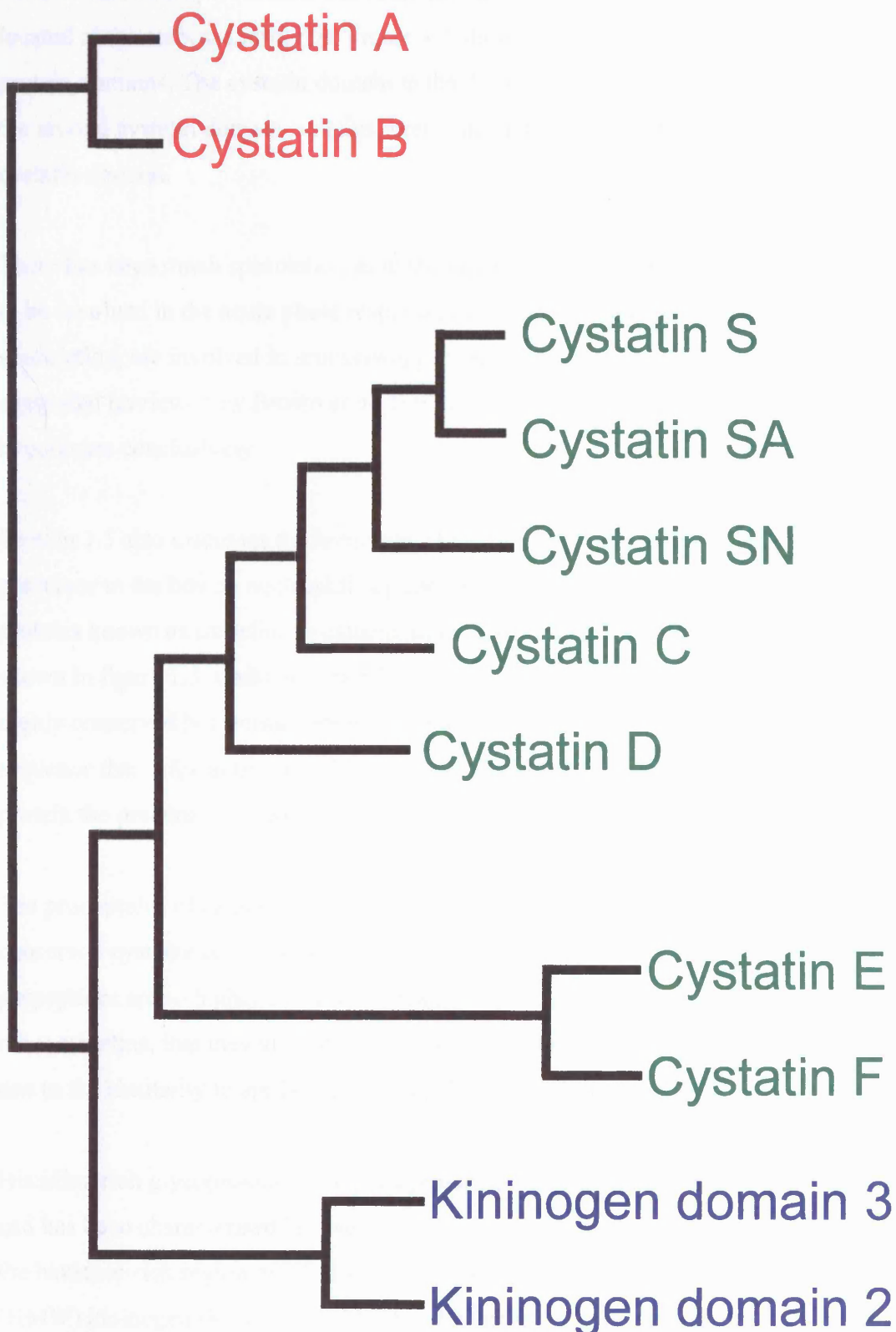


Figure 1.4. Figure showing the evolutionary relationship between all known human type 1, type 2 and type 3 cystatin domains.

(adapted from www.klinkem.lu.se/E/abrahamson/cystatin_text.html)

Type 1: Cystatin A, Cystatin B

Type 2: Cystatin S, Cystatin SA, Cystatin SN, Cystatin C, Cystatin D, Cystatin E, Cystatin F

Type 3: Kininogen domain 3, Kininogen domain 2

Fetuin has two cystatin domains (Elzanowski *et al.* 1988) followed by a non-related peptide located at the carboxyl terminus. Figure 1.3 shows a schematic representation of the fetuin protein domains. The cystatin domain at the 'N' terminus contains two disulphide bonds and the second cystatin domain contains three, one being narrower than the two found in the first cystatin domain.

There has been much speculation as to the function of the fetuins. They are proteins thought to be involved in the acute phase response, bind calcium, have a role in bone formation and modulation, are involved in immunosuppression and many other ideas have also been suggested (reviewed by Brown *et al.* 1992). However, there is little evidence to support any hypotheses conclusively.

Section 1.5 also discusses the homology Hu *et al.* (1995) reported between spp24 and the precursor to the bovine neutrophil peptide bactericidin. This protein falls into a group of proteins known as cathelins or cathelicidins. A schematic representation of the protein is shown in figure 1.3. Cathelins are a family of antimicrobial peptide precursors that have a highly conserved N-terminal preprosequence, followed by a highly variable C-terminal sequence that is the antibacterial peptide (reviewed by Zanetti *et al.* 1995). In the mature protein the precursor is cleaved to release the C-terminal antimicrobial peptide.

The propeptides of cathelins loosely resemble a cystatin domain in that they contain four conserved cysteine residues that are thought to form disulphide bonds. However, these propeptides are so highly conserved between each other, but with little sequence homology to other cystatins, that they are not included in the cystatin superfamily. They are discussed here due to the similarity to spp24 reported by Hu *et al.* (1995).

Histidine-rich glycoprotein (HRG) is a protein that is a member of the cystatin superfamily and has been characterised in several species (reviewed by Leung 1993). It was suggested that the histidine-rich region of HRG was related to human and bovine high molecular weight (HMW) kininogen (Koide *et al.* 1986), but its function is still unclear.

A further cystatin-like group of proteins was originally identified in rat called the cystatin-related proteins (CRPs) (Parker *et al.* 1978). All known cystatin domains to date are encoded in three exons of characteristic sizes (reviewed by Bobek and Levine 1992). However, CRPs were shown to have four exons, due to the duplication of the equivalent cystatin exon 2

(Devos *et al.* 1993). Consequently CRPs have two of the narrower disulphide bonds that are typically seen in cystatins, followed by the wider disulphide bond.

A more general class of proteins found to be cystatin-related are termed variant cystatins. This group of proteins includes divergent cystatins from the venom of the African puff adder (*Bitis arietans*), the flesh fly (*Sarcophaga peregrina*) and the fruit fly (*Drosophila melanogaster*). The class also contains the invariant chain (Ii chain) involved in the assembly of class II MHC molecules and the plant cystatins. All of these proteins have either incomplete sections of cystatin domains or, as in the case of the plant cystatins, differences in their genomic organisation.

1.3.2 The mode of action of cystatin

Figure 1.5 shows the hydrolysis of a peptide by a thiol protease (adapted from Baggio *et al.* 1996). The reaction takes place in two phases. During the first phase, the thiol protease enzyme interacts with the substrate peptide and cleaves a peptide bond. An acyl-enzyme complex is formed between the thiol protease and part of the cleaved substrate. The other part of the cleaved substrate is released as a leaving group. During the second phase of the reaction, H₂O reacts with the acyl-enzyme. This results in the thiol protease enzyme being regenerated and a second peptide being released.

On the thiol protease enzyme there is a defined area that binds the substrate acyl group and another that binds the substrate leaving group (Berger and Schechter 1970). Cystatins can inhibit thiol proteases by replacing the substrate peptide in the formation of the acyl-enzyme complex. The cystatin then hydrolyses very slowly, or possibly not at all. Consequently, whilst the reactive site of the thiol protease is occupied by cystatin, no further substrate molecules can be processed.

The amino acid sequences of the cystatins are highly conserved in three main regions thought to be involved in inhibitory activity of the enzyme. Figure 1.6 indicates these regions on chicken egg white cystatin.

The first conserved domain seen in cystatins is found at the N-terminal end of the protein. A glycine residue usually found at position 9 is thought to be important in the orientation of the N-terminal region towards the thiol proteinase (Hall *et al.* 1993). Many functional studies on chicken egg white cystatin and cystatin C have confirmed the importance of the N-terminal

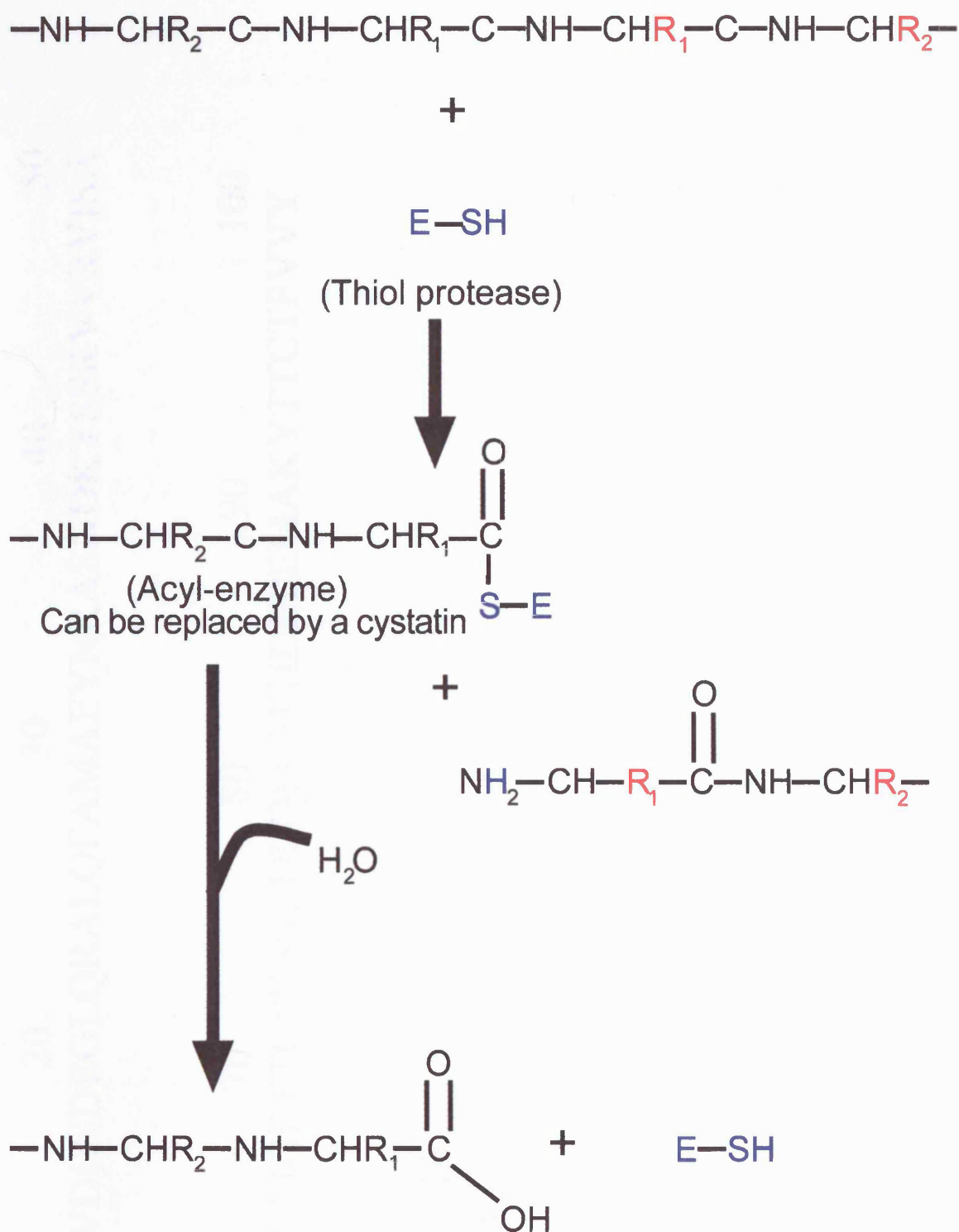


Figure 1.5. Figure showing the mode of action of a thiol protease. Adapted from (Baggio *et al.* 1996).

The cleavage of a peptide by a thiol protease takes place in two phases. During the first phase, the thiol protease enzyme reacts with the peptide and cleaves a peptide bond. An acyl enzyme is formed when a substrate acyl group is transferred to the -SH group of the thiol protease. A peptide with a free N-terminal amino group is released. During the second phase, H₂O, or some other nucleophile, reacts with the acyl enzyme. The thiol protease is regenerated and a peptide released with a free C-terminal carboxyl group. Cystatins act by forming an acyl enzyme complex that hydrolyses either slowly or not at all. While the binding site of the thiol protease is occupied, additional substrate molecules cannot be processed.

1 10 20 30 40 50
 SEDRSRL**L**GAPVPVDENDEGLQRALQFAMAEYNRASNDKYSSRVVRVISA

 60 70 80 90 100
 KR**QLVSG**IKYILQVEIGRTTCPKSSGDLQSCEFHDPEMAKYTTCTFVVY

 110 116
 SI**PWL**NQIKLLESKCQ

Figure 1.6. The highly conserved residues of cystatin thought to be functionally important.

The highly conserved regions of cystatin are indicated in red. The sequence shown is that of chicken egg white cystatin.

There is a conserved glycine residue at position 9, the conserved 'QxVxG' sequence at positions 53 to 57 which consists of a glutamine, valine and glycine with 'x' being any residue and finally the conserved proline and tryptophan residues at positions 103 and 104.

glycine for inhibitory function (Abrahamson *et al.* 1987a,b; Machleidt *et al.* 1989; Machleidt *et al.* 1991; Grubb *et al.* 1990; Abrahamson *et al.* 1991a,b; Genenger *et al.* 1991; Lindahl *et al.* 1992; Hall *et al.* 1992; Lalmanach *et al.* 1993). The second and third domains that are highly conserved in cystatins are residues found in the first and second hairpin loop of the cystatin structure (discussed in section 1.3.3).

The region found in the first hairpin loop is known as the 'QxVxG' sequence. The glutamine (Q), valine (V) and glycine (G) residues are found at positions 53 to 57 in all functional cystatins, with x being any residue. It has been shown that amino acid substitutions in this region of recombinant chicken egg white cystatin, reduce the efficiency of papain and cathepsin B inhibition by up to 1000-fold (Auerswald *et al.* 1992). However, cathepsin L inhibition was unaffected by substitutions in this region. This suggests that between closely related thiol proteinases there are differences in the proteinase-inhibitor interactions.

The highly conserved region found in the second hairpin loop of chicken egg white cystatin consists of a proline and tryptophan residue found at positions 103 and 104 ('PW'). It has been shown that modification of the tryptophan residue in chicken egg white cystatin reduces inhibition of papain (Lindahl *et al.* 1988).

The cystatin that has been the most extensively studied is chicken egg white cystatin. The X-ray crystal structure of chicken egg white cystatin has been determined (Bode *et al.* 1988) (section 1.3.3) and the three highly conserved regions of cystatins are thought to be directly involved in binding and docking.

1.3.3 The structure of cystatin

The crystal structure of chicken egg white cystatin has been determined by X-ray diffraction methods (Bode *et al.* 1988). Figure 1.7A-E shows the structure in several different orientations.

In terms of secondary structures of chicken egg white cystatin there are five extended strands that form an antiparallel, twisted β -pleated sheet. This is then partially wrapped around a long straight α -helix. There is a second α -helix that is aligned perpendicular to the β -strands and the first α -helix and lies away from the main body of the molecule.

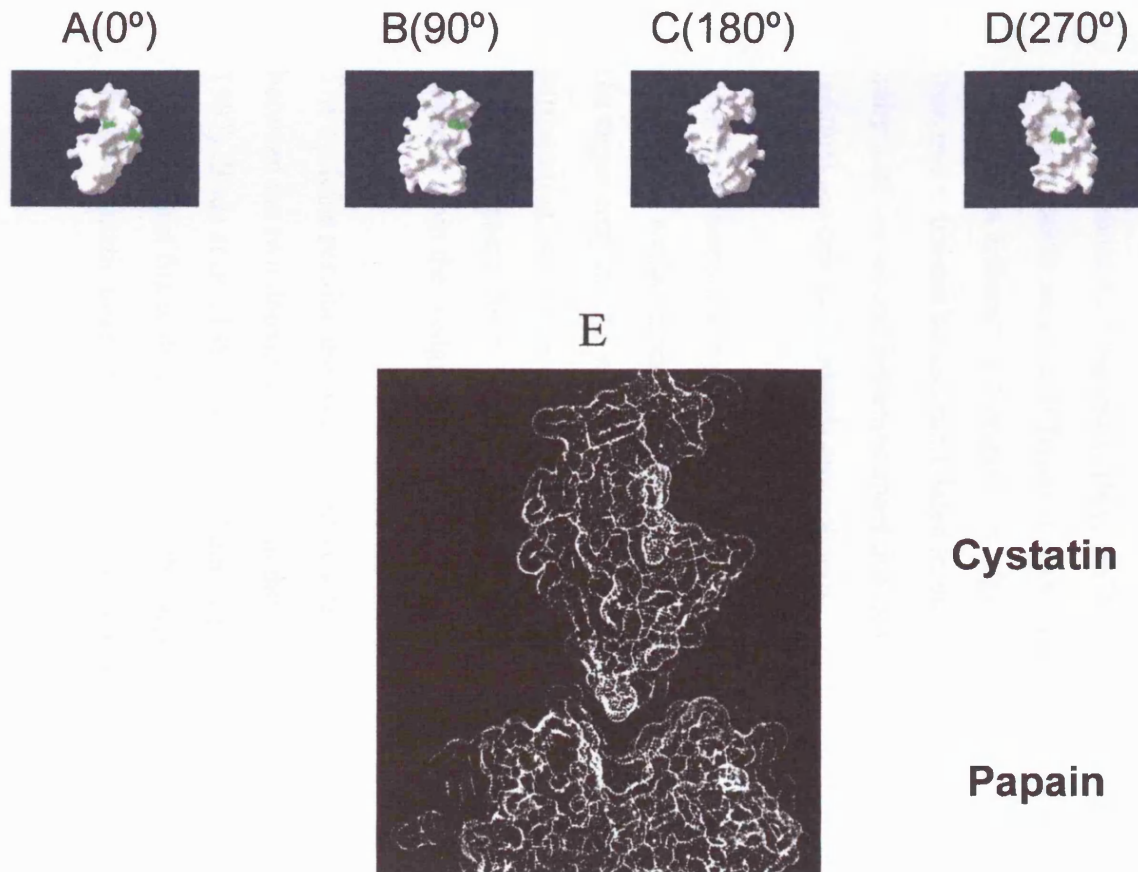


Figure 1.7. The structure of chicken egg white cystatin and its interaction with papain.

The 3-dimensional structure of the chicken egg white cystatin protein is shown in various orientations in A, B, C, D and E. The structure was determined by Bode *et al.* 1988 and the images presented here were produced in Swiss-PdbViewer v3.7b2 (<http://www.expasy.ch/spdbv>) (Guex and Peitsch 1997). The cysteine residues are indicated in green. The proposed interaction of chicken egg white cystatin with papain is shown in E. This image was taken from Bode *et al.* 1988, figure 4.

The characteristic disulphide bonds of the cystatins are buried in the molecule and serve to clamp the second α -helix and the carboxy terminus to the β -pleated sheet. The distribution of amino acids with charged side chains gives rise to a positive pole towards the loop formed by residues 53 to 59 and a negative pole towards the second α -helix.

Of particular importance are the highly conserved regions described in section 1.3.2. The glycine residue at position 9 is located at the extreme corner of the β -pleated sheet. It is suggested that residues 1 to 8 protrude from this into solution and are therefore accessible for proteolytic attack. The 'QxVxG' region is located at the hairpin of a β -strand, adjacent to the amino terminus. Consequently, these residues are exposed to any solvents and are thought to be able to easily adapt to different environments. The glycine residue at position 57 is actually buried. This is therefore thought to be highly conserved due to the fact that residues any larger than this could not be accommodated in this position. The 'PW' region is located at the hairpin of the second β -pleated sheet that is adjacent to the 'QxVxG' loop. Again, these residues are oriented towards any solvents.

Both the β -hairpin loops containing highly conserved residues and the amino terminus, form a contiguous wedge. Bode *et al.* (1988) suggested that this wedge was the contact region with the target enzyme. As well as all the highly conserved residues lying in this region, the surrounding residues are also relatively conserved between cystatins and could act as anchoring points for the interacting enzyme. The more variable residues are located some distance from the wedge.

The cysteine residue that is the reactive site of papain lies at the bottom of a cleft formed between the two domains that make up the papain structure (reviewed by Baker and Drenth 1987). Bode *et al.* (1988) performed docking experiments and demonstrated that chicken egg white cystatin fits with its wedge into the papain cleft and the result is shown in figure 1.7E. The electrostatic interactions between cystatin and papain were shown to be favourable in all models.

1.3.4 Proposed functions of cystatins

As well as the established function of inhibition of members of the papain superfamily, several other roles have been suggested for cystatins in health and disease which include tumours and metastasis, infections, neurological disorders and inflammatory disease.

Increased activity of some thiol proteases has been linked to tumour malignancy. The thiol proteases secreted by the cancer cell may facilitate metastasis by assisting the cell in penetrating through the stromal tissue and degrading basement membranes. An example of this is that plasma-membrane fraction of human tumours of breast, ovary, bladder and colon contain higher levels of cathepsin L mRNA and cathepsin B-like activity than the equivalent normal tissues (Sloane *et al.* 1987; Rozhin *et al.* 1989; Chauhan *et al.* 1991).

It was originally thought that either an over expression of cysteine proteinases or an underexpression of cystatin from the cancer cell leads to a facilitation of metastasis. However, it has also been suggested that the opposite could be true, that is cystatin could inhibit the proteolytic attack by cysteine proteinases on cancer cells by suppressing the inflammatory response (Collela *et al.* 1993). It has therefore been proposed that if there is an imbalance of cystatins and cysteine proteinases this could contribute to tumours and their metastasis. Impaired regulation of cystatins leading to an increased or decreased amount of cystatin could bring about this imbalance.

Cystatins are thought to have a role in defending the body against infection. For example, some viruses require the presence of cysteine proteinases in the cytoplasm of the infected cell to be able to replicate. Hence the presence of a cystatin could prevent the replication of the virus. Chicken egg white cystatin has been shown to cause a reduction in virus production of poliovirus-infected cells and also to cause absence of viral protein synthesis when cystatin exposure was prior to viral infection (Korant *et al.* 1985).

Cystatins have been implicated in neurological disorders. In particular, cystatin C which has been detected in cerebrospinal fluid (CSF) has been implicated with having a role in the aetiology of multiple sclerosis (MS). MS is a disease that involves demyelination, which may be a result of proteolytic enzymes. It was suggested that as macrophages predominate in MS lesions there could be a role for cysteine proteinases in the degradation of myelin (Bollengier 1987). Bollengier (1987) demonstrated a significantly lower than normal level of cystatin C in CSF in a large group of MS patients and also the absence of a correlation between age and cystatin C. Consequently, it was suggested that the decreased level of cystatin could lead to an increased level of cysteine proteinase activity and therefore a degradation of myelin.

It has also been suggested that cystatins may have a role in inflammatory diseases. Cystatin A, a type 1 cystatin, has been extensively studied in inflammatory skin diseases due to its epidermal origin. For example, increased amounts of cystatin A were demonstrated in

psoriatic epidermis and inflammatory skin samples (Järvinen *et al.* 1987) However, there has also been evidence to suggest that a cysteine proteinase inhibitor from psoriatic skin was less stable and less active towards papain than that found in normal cells (Othani *et al.* 1982) and so there is still much speculation.

Cysteine proteinases are thought to play a role in periodontal inflammatory diseases and consequently salivary cystatins have been speculated to play a protective role against cysteine proteinases both endogenous and exogenous in origin. There have been many studies that provide evidence for this. For example, in maximal cases of gingival inflammation, an induction of cystatin C secretion was observed (Henskens *et al.* 1994).

There may also be a role for cystatins in the destruction of cartilage and collagen. Patients suffering from rheumatoid arthritis were found to have very high levels of cystatin C in their synovial fluid (Lenarcic *et al.* 1988). In 1992, Lerner and Grubb (1992) analysed the parathyroid hormone stimulated release of ^{45}Ca and ^3H from prelabelled mouse calvarial bones. They showed that the use of recombinant cystatin C resulted in a significant reduction in the release of ^{45}Ca and ^3H and suggested a possible role for cystatin C in bone resorption (Lerner and Grubb 1992).

1.4 Cathepsin K

Hu *et al.* (1995) speculated that spp24 might have a role in bone turnover as several possible target cysteine proteinases are known to be expressed in bone. Cathepsin K is a protein that has, as yet, had no natural inhibitor identified and is known to be predominantly expressed by osteoclasts (Tezuka *et al.* 1994). For this reason cathepsin K is thought to be a candidate for a potential interactor with spp24.

Cathepsin K is also sometimes referred to as cathepsin O2, due to its original cloning and naming in rabbit as OC-2 (Tezuka *et al.* 1994).

1.4.1 The Cathepsin K gene

The human cathepsin K cDNA was originally cloned by Inaoka *et al.* (1995) and showed 94% homology to a previously cloned rabbit OC-2 cDNA, tentatively called cathepsin K and isolated from osteoclasts (Tezuka *et al.* 1994). The human cathepsin K gene was shown to be expressed at low levels in many tissues, but extremely high expression was seen in osteoclastoma and osteoarthritic hip bone suggesting that cathepsin K participates in bone remodelling (Inaoka *et al.* 1995). It was then confirmed by *in situ* hybridisation that cathepsin K was expressed selectively in human osteoclasts (Drake *et al.* 1996).

The human cathepsin K gene was localised to chromosome 1q21 by fluorescence *in situ* hybridisation and the gene shown to comprise 8 exons spanning 9 kb (Gelb *et al.* 1997). The promoter region of the gene lacked the canonical 'TATA' and 'CAAT' box sequences, but contained two AP1 sites and was not particularly GC rich (Gelb *et al.* 1997).

The mouse cathepsin K cDNA was cloned by Rantakokko *et al.* (1996). The cDNA showed 87% homology with the corresponding human and rabbit sequences and northern blot analysis revealed expression of the gene in bone, cartilage and skeletal muscle (Rantakokko *et al.* 1996). *In situ* hybridisation showed mouse cathepsin K mRNA was detected at high levels in osteoclasts and also in some hypertrophic chondrocytes of growth cartilages (Rantakokko *et al.* 1996). A developmental expression study investigated the expression of cathepsin K during foetal mouse development using *in situ* hybridisation and reported that cathepsin K expression during embryogenesis occurred only following the onset of osteoclast differentiation (Dodds *et al.* 1998).

The mouse cathepsin K gene was localised to 4.5 kb downstream of *Arnt* on mouse chromosome 3 at map position 47.9 (Rantakokko *et al.* 1999). The gene was shown to comprise 8 exons and span approximately 10.1 kb (Rantakokko *et al.* 1999). Rantakokko *et al.* (1999) aligned the promoter regions of the mouse and human cathepsin K gene and suggested the presence of a non-consensus 'TATA'-box ('AATAAAT') and a 'CAAT' box located 25-43 bp upstream of the transcription initiation sites. However, a second report stated that the mouse cathepsin K gene lacks canonical 'TATA' and 'CAAT' boxes (Li and Chen 1999). Both reports suggest the occurrence of two putative AP1 sites in the promoter region.

1.4.2 The cathepsin K protein

Human cathepsin K has been expressed in baculovirus-infected Sf21 cells and the recombinant soluble protein purified (Bossard *et al.* 1996). Cathepsin K has an inhibitory pro-leader sequence that is common to thiol proteases. Conditions were identified for removal of this pro-sequence and the release of the active mature enzyme (Bossard *et al.* 1996). The substrates that were identified for human cathepsin K were fluorogenic peptides, collagen and osteonectin (Bossard *et al.* 1996). Cathepsin K was shown to be inhibited by E-64 and leupeptin (Bossard *et al.* 1996), which is characteristic of thiol proteases. Mature cathepsin K was shown to be active at low pH (Bossard *et al.* 1996), consistent with the findings described in section 1.4.1 of cathepsin K being expressed at high levels in osteoclasts. Brömme *et al.* (1996) also expressed the human cathepsin K protein (although called it cathepsin O2) in a baculovirus system. It was shown that cathepsin K has a potent collagenolytic activity against type I collagen between pH 5 and 6 and elastinolytic activity against insoluble elastin at pH 7.0 (Brömme *et al.* 1996).

Cathepsin K has been shown to be the major proteolytic activity in osteoclast (Drake *et al.* 1996). It has also been demonstrated in rat osteoclast pit formation assays that specific inhibition of cathepsin K leads to a decrease in bone resorption (Xia *et al.* 1999). Cathepsin K is therefore thought to be one of the major thiol proteases responsible for bone resorption.

1.4.3 Cathepsin K in health and disease

Pyconodystosis (Pycno) is a rare, autosomal, recessive disease that was mapped to the same region as cathepsin K, human chromosome 1q21, (Gelb *et al.* 1996b). The disease is characterised by a short stature, osteosclerosis, bone fragility, clavicular dysplasia and skull deformities (Maroteaux and Lamy 1962).

In patients suffering from Pycno, the number of osteoclasts is normal, as are their ruffled borders, but the region of demineralised bone that surrounds each individual osteoclast is increased (Everts *et al.* 1985). The bone in this region is demineralised as normal, but the organic matrix is not adequately degraded.

Analysis of a large, consanguineous Israeli Arab family with 16 affected individuals revealed an A to G transition at cDNA nucleotide 1095 in individuals with Pycno (Shi *et al.* 1995; Brömme and Okamoto 1995; Tezuka *et al.* 1994; Inaoka *et al.* 1995; Li *et al.* 1995). This resulted in the substitution of the termination codon with a tryptophan residue and a 19 amino acid elongation of the C-terminus of the protein.

A further two cathepsin K mutations were found in unrelated families. A C to G transversion at nucleotide 541 (a glycine to arginine change) was seen in two affected Moroccan siblings and a C to T transition of a CpG dinucleotide at nucleotide 826 (nonsense mutation resulting in a loss of an arginine residue) was seen in an American Hispanic patient (reported in Gelb *et al.* 1996a). Cells transfected with cathepsin K constructs containing the mutations described above were shown to produce no detectable protein (Gelb *et al.* 1996a).

Cathepsin K has also been implicated in osteoporosis. Synthetic inhibitors of cathepsin K have been shown to reduce osteoclast resorption *in vitro* and *in vivo* and therefore may prove beneficial as therapeutic agents in the treatment of osteoclast-mediated bone loss in conditions such as osteoporosis. Many pharmaceutical companies are currently studying cathepsin K inhibitors with a view to developing osteoporosis treatments. As yet, the natural inhibitor of cathepsin K has not been identified.

Spp24 is a cystatin-like protein that was originally isolated from bone (Hu *et al.* 1995) (section 1.5). Cystatins inhibit thiol proteases (section 1.3). Since both spp24 and cathepsin K are both known to be found in bone and a natural inhibitor for cathepsin K has not yet been identified, spp24 is a potential candidate for a cathepsin K inhibitor and as such could have important implications in bone diseases such as osteoporosis.

1.5 Secreted phosphoprotein 24 (spp24)

There is currently a very limited knowledge regarding secreted phosphoprotein 24 (spp24). The protein was first reported as a novel non-collagenous protein purified from bovine cortical bone (Hu *et al.* 1995). Since its isolation, the only other published data reports the localisation of the human locus encoding spp24 (assigned the symbol *SPP2*) to chromosome band 2q37→qter by *in situ* hybridisation (Swallow *et al.* 1997).

1.5.1 The isolation of spp24 and the determination of its amino acid and cDNA sequence

Hu *et al.* (1995) described the demineralisation of ground calf bone with formic acid and the adsorption of the extracted proteins to a C¹⁸ matrix where bone mineral and neutral pH-soluble proteins were removed. Spp24 was found in the neutral pH-insoluble extract and was purified by application to a Sephacryl S-100 HR column followed by further purification using reverse phase HPLC. Spp24 co-isolated before the reverse phase HPLC with a known non-collagenous bone protein, matrix-Gla protein (MGP). This indicates the similar properties of spp24 and MGP (*i.e.* both are released from bovine bone by demineralisation with formic acid and both are insoluble at neutral pH).

The purified spp24 protein was shown to be homogenous by Hu *et al.* (1995) when electrophoresed on a SDS gel, showing a protein of 24 kDa molecular mass. The homogeneity of spp24 was also confirmed by N-terminal sequencing of the purified protein and internal peptides released by cleavage (Hu *et al.* 1995).

The cDNA sequence of bovine spp24 was determined by Hu *et al.* (1995) using a combination of RT-PCR, 3'RACE and nucleotide screening of a λ gt11 cDNA library. Based on the N-terminal amino acid sequence of the purified spp24 protein, degenerate primers were designed and RT-PCR performed on bovine bone periosteum or bovine liver preparations. The 380 bp fragment generated was cloned and sequenced and shown to be identical in both bone and liver. Part of this 380 bp fragment was used as a probe to screen a bovine liver λ gt11 cDNA library, which generated a clone covering the 5'-end of the cDNA. To determine the 3'-end of the cDNA, 3'RACE was performed on bone and liver RNA preparations. Identical 3'-end sequence was obtained from both tissues.

Figure 1.8 shows the bovine spp24 cDNA sequence and the deduced amino acid sequence of the protein, some of which was confirmed by N-terminal sequencing (as determined by Hu *et al.* 1995).

The cDNA sequence is 816 nucleotides in length. Translation is initiated by the 'ATG' codon at nucleotides 91-93. There is a 20-amino acid signal peptide with a potential cleavage site at amino acid residue 20. The N-terminal sequencing of the purified mature spp24 protein determined an N-terminus 20 amino acids downstream of the presumed initiation methionine. The open reading frame of spp24 encodes a 200-amino acid protein (including the signal peptide) and is terminated by a 'TGA' codon at nucleotides 691-693. A polyadenylation signal ('AATAAA') is seen at nucleotides 700-795.

1.5.2 The expression of spp24

Hu *et al.* (1995) performed northern blot analysis on total RNA from the bovine tissues bone, liver, heart, lung, kidney and spleen. Part of the 380 bp fragment generated by RT-PCR (as described in section 1.5.1) was used as a ³²P-labelled probe. Spp24 mRNA was detected in bone and liver as expected from the results discussed in section 1.5.1. A single transcript of 1000-1100 nucleotides was detected which agrees with the length of the determined cDNA. No spp24 mRNA was detected in bovine heart, lung, kidney or spleen. The northern blot results reported by Hu *et al.* (1995) suggest tissue-specific expression of spp24. Hu *et al.* (1995) suggested that the presence of spp24 in bone indicates a possible role in bone turnover.

1.5.3 The structure of the spp24 protein and homologies with known proteins

Hu *et al.* (1995) compared the complete 200-amino acid bovine spp24 sequence with known proteins in the non-redundant protein database of the NLM using the BLAST search program. The results of the BLAST search showed that the N-terminal region of the bovine spp24 protein had some homology with cystatin domain 3 of kininogen and the precursor to the bovine neutrophil antibiotic peptide batenecin, both of which are related to the cystatin superfamily (discussed in section 1.3).

Hu *et al.* (1995) aligned bovine spp24, bovine batenecin precursor, cystatin domains 1 and 3 of kininogen and two closely related sequences; porcine cathelin and chicken egg white cystatin. They demonstrated that the cathelin and batenecin precursor are more closely

```

1  ACAGTCTGAT CTGCCAAGTG CATTATACCA ATATCTCATT AATTCTCCCC
51  AAACCTCTGA ACGGAAATTG TTCTTCCCAT AATGGAGAAG ATGGCGATGA
    M A M
101 AGATGTTGGT GATATTTGTC CTTGGAATGA ACCACTGGAC TTGTACAGGT
    K M L V I F V L G M N H W T C T G
151 TTCCCGGTGT ATGACTATGA CCCGGCTTCC CTGAAGGAGG CTCTCAGCGC
    F P V Y D Y D P A S L K E A L S A
201 CTCTGTGGCA AAAGTGAATT CCCAGTCACT GAGCCCCTAT CTGTTTCGGG
    S V A K V N S Q S L S P Y L F R
251 CGTTTAGAAG CTCAGTTAAA AGAGTCAACG CCCTGGACGA GGACAGCTTG
    A F R S S V K R V N A L D E D S L
301 ACCATGGACT TAGAGTTCAG GATTCAAGAG ACGACGTGCA GGAGGGAATC
    T M D L E F R I Q E T T C R R E S
351 TGAGGCAGAC CCCGCCACCT GTGACTTCCA GAGGGGCTAC CACGTGCCCCG
    E A D P A T C D F Q R G Y H V P
401 TGGCCGTTTG CAGAAGCACC GTGCGGATGT CTGCTGAACA GGTGCAGAAC
    V A V C R S T V R M S A E Q V Q N
451 GTGTGGGTTC GCTGCCACTG GTCCTCCAGC TCTGGGTCCA GCAGCAGTGA
    V W V R C H W S S S S G S S S S E
501 AGAGATGTTT TTTGGGGATA TCTTGGGATC CTCTACATCA AGAAACAGTT
    E M F F G D I L G S S T S R N S
551 ACCTGCTTGG CCTCACTCCT GACAGATCCA GAGGTGAACC ACTTTATGAA
    Y L L G L T P D R S R G E P L Y E
601 CCATCACGTG AGATGAGAAG AAACTTTCCT CTTGGAAATA GAAGGTACTC
    P S R E M R R N F P L G N R R Y S
651 GAACCCGTGG CCCAGAGCAA GAGTAAACCC TGGCTTTGAG TGACAGCCTT
    N P W P R A R V N P G F E
701 AAGCAAAATG CACTGGAAGG AATAGAAGTT CCAATGAAGA AAGATACCTT
751 ATGAATTGTG TAATTTTCTT TTGATCAATT GCAGTCCCTA ATAAATGGCT
801 TACTTTTCCT CTTTCA

```

Figure 1.8. The bovine cDNA and amino acid sequence (Hu *et al.* 1995).

The bovine cDNA sequence for spp24 was deduced as described in section 1.5.1 by Hu *et al.* (1995). The cDNA sequence is shown here with the deduced amino acid sequence underneath. The signal peptide residues are shown in blue and the characteristic cysteine residues are shown in red.

related to spp24 than kininogen and chicken egg white cystatin and domains 1 and 3 of kininogen are more closely related to spp24 than to cathelin or the bactenecin precursor. Hu *et al.* (1995) therefore suggested that spp24 was an evolutionary intermediate between the cathelins and the bactenecin precursor and kininogen and the cystatin.

The homologies found to spp24 were only seen in the first approximately 107 residues at the N-terminal end of the mature protein. Residues 108-180 at the C-terminal end showed no homology to any known protein. Figure 1.9 shows a schematic representation of the spp24 protein. The cystatin-like region of bovine spp24 contains four cysteine residues, shown in red, that are likely to be involved in disulphide bonds as is seen in members of the cystatin superfamily (section 1.3).

Hu *et al.* (1995) determined the location and level of phosphorylation of phosphoserine residues in the bovine spp24 protein. Table 1.3 shows the phosphorylated serines and their degree of phosphorylation (adapted from Hu *et al.* 1995). This demonstrated a stretch of serine residues that are highly phosphorylated separating the cystatin-like and non-cystatin-like region (figure 1.9).

1.5.4 Speculated functions of spp24

Hu *et al.* (1995) speculated that the cystatin-like region of spp24 folds into a cystatin tertiary structure similar to that reported by Bode *et al.* (1988) for chicken egg white cystatin (section 1.3). It was suggested that spp24 might inhibit thiol proteases, as is a feature of most proteins with a cystatin domain. The presence of spp24 in bone implied that any target thiol proteinase must also be present in bone (Hu *et al.* 1995). There are several thiol proteases known to be expressed by osteoclasts to digest collagen and various non-collagenous bone proteins (Delaissé *et al.* 1980).

Of the proteins aligned with spp24 by Hu *et al.* (1995), cathelin, chicken egg-white cystatin and cystatin domain 3 of kininogen have been shown to possess thiol protease inhibitory activity (Salvesen *et al.* 1986). The bactenecin precursor has not been tested for cystatin function although the related neutrophil antibiotic peptide Bac 5 precursor has been shown to inhibit the cysteine proteinase cathepsin L (Zanetti *et al.* 1995).

Hu *et al.* (1995) also put forward a second suggestion for the function of spp24. Both the cystatin domain 3 of kininogen and bovine neutrophil antibiotic precursor, the two proteins

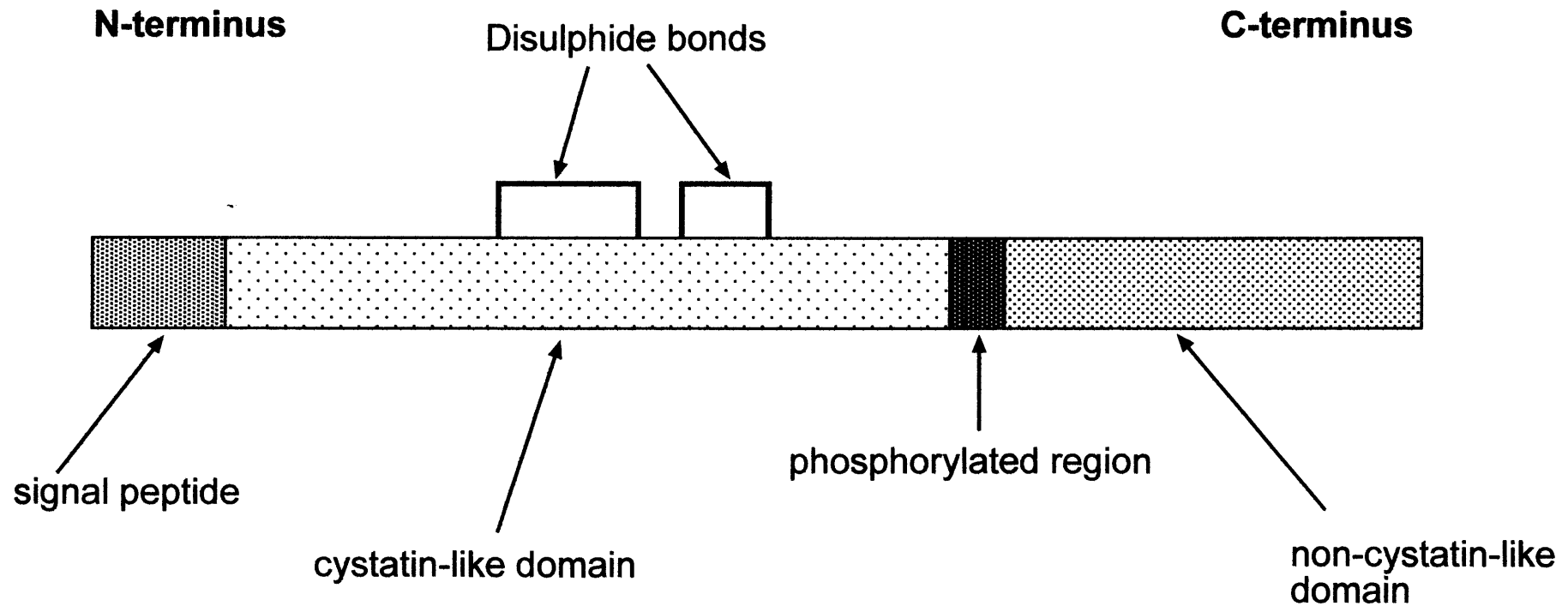


Figure 1.9. A schematic representation of the structure of spp24.

This figure is drawn to scale and shows the four regions that comprise spp24. The 20-amino acid signal peptide is shown at the N-terminal end, followed by the cystatin-like domain. Within the cystatin-like domain are four cysteine residues at positions 63, 74, 87 and 105 of the whole protein. These cysteine residues are thought to be involved in disulphide bonds as is seen in members of the cystatin superfamily. A spacer region of phosphorylated serine residues separates the cystatin-like domain from the C-terminal non-cystatin-like domain that shows no homology to any known protein.

Table 1.3. The extent of phosphorylation in the cluster of serine residues seen in purified bovine spp24 (adapted from Hu *et al.* 1995).

To determine the location of phosphorylated serine residues the spp24 protein was cleaved with BNPS-Skatole at tryptophan 127. The serine rich peptide was isolated by gel filtration over Sephacryl S-100 HR and then transferred to a poly(vinylidene fluoride) membrane. Reaction with ethanediol for 4 h at 60°C converted the phosphoserine residues to S-ethylcysteine. The percentage phosphorylation was then determined by the amount of phenylthiohydantoin (PTH)-S-ethylcysteine divided by PTH-S-ethylcysteine + PTH-serine at the indicated residue.

Amino acid and position in bovine spp24	Degree of phosphorylation (%)
Serine 128	5
Serine 129	63
Serine 130	70
Serine 131	81
Glycine 132	-
Serine 133	82
Serine 134	83
Serine 135	81
Serine 136	78
Glutamate 137	-
Glutamate 138	-

most closely related to spp24, have a cystatin domain that lies adjacent to a C-terminal peptide that is released by a protease action. In kininogen, the action of kallikrein releases bradykinin which is a potent vasodilator. In the precursor of bovine neutrophil bactenecin, cleavage by a protease releases antibiotic dodecapeptide bactenecin. Hu *et al.* (1995) suggested the C-terminal domain of spp24 could be released via the same mechanism, but as the peptides show no sequence homology they must have different target binding sites.

A third functional possibility was put forward by Hu *et al.* (1995). Spp24 is similar in its overall structure to fetuin. Fetuin has two cystatin domains as opposed to the one seen in spp24, (Elzanowski *et al.* 1988) but it also has the extended C-terminal domain following the last cystatin domain. Fetuin is a plasma protein synthesised in the liver as well as accumulating in the extracellular matrix of bone (Ohnishi *et al.* 1993). The human form of fetuin, α_2 HS-glycoprotein, circulates in the blood as a cleaved two-chain molecule (Lee *et al.* 1987). The cleavage is thought to occur in the C-terminal sequence following the second cystatin domain. Hu *et al.* (1995) suggested spp24 might act in a similar manner to fetuin. Fetuin is discussed in more detail in section 1.3.

The role of the phosphorylated serine residues clustered after the cystatin domain of spp24 is unclear. The phosphorylation of serine in spp24 is thought to follow the recognition motif that has been seen in other secreted phosphoproteins, that is Ser-X-Glu/Ser(P). All of the phosphorylated serines in spp24, with the exception of serine 130, have a glutamate or phosphoserine in the n+2 position. It has previously been noted that phosphoproteins secreted into the extracellular environment of cells tend to be partially phosphorylated at serine residues and those phosphoproteins that are secreted into milk and saliva are usually fully phosphorylated (Price *et al.* 1994).

Hu *et al.* (1995) speculated that the clustering of partially phosphorylated serine residues seen in spp24 could be responsible for regulating the extent of phosphorylation by a specific protein kinase or phosphatase and that the negative charge produced in this region could create sufficient repulsion to prevent the formation of secondary structure. In this way the serine residues could act as an anionic spacer region between the cystatin-like and non-cystatin-like region with the extent of phosphorylation regulating the separation of the two domains. This could modulate the susceptibility of spp24 to proteolytic cleavage or some other specific activity.

In summary Hu *et al.* (1995) presented three possible models for the likely function of spp24:

- thiol proteinase inhibitory activity
- cleavage of the C-terminal domain to release a biologically active peptide
- a fetuin-like plasma protein

It was suggested that spp24 had a function in bone and that the extent of phosphorylation in a series of serine residues may act to regulate its activity.

1.6 Aims and Objectives

With the recent completion of the first draft of the human genome project, more and more genes are now being identified that have an unknown function. The major challenges now facing researchers are those of functional genomics *i.e.* what a gene does, and those of proteomics *i.e.* what a protein does. In fact genetic research is now starting to reverse itself. Instead of taking a specific disease and looking for the responsible protein and gene, researchers are now taking a novel gene and the protein it encodes and trying to discover its function and its role in health and disease, particularly multifactorial diseases. The work presented in this thesis is an example of this approach.

A novel gene and the protein it encodes was identified from bovine cortical bone and named secreted phosphoprotein 24 (spp24) (Hu *et al.* 1995). The challenge is now to discover the structure of the gene and the function of the protein. This thesis describes the structural characterisation of the human and mouse genes encoding spp24 and highlights the difficulties of determining protein function.

The human gene encoding spp24 was sequenced by a group participating in the human genome project during the early stages of the work presented here. This enabled more detailed aims and objectives to be determined. The aims and objectives of the work presented in this thesis are as follows:

- **To determine the exon/intron structure of the human gene encoding spp24.**

The exon/intron structure of the human gene will enable a comparison to be made between the gene structure of spp24 and cystatins, which will provide evidence for its position in the cystatin superfamily. Knowledge of the exon/intron structure of the gene will also provide vital information for future functional studies and for the molecular basis of the gene. This work is presented in Chapter 3.

- **To determine the cDNA sequence for the mouse gene encoding spp24.**

The spp24 protein sequence is publicly available for humans, cattle and rat. However, no mouse sequence is currently available. The mouse cDNA sequence can be determined by alignment of ESTs to generate a consensus sequence. The protein sequence can then be deduced from the cDNA. It is important to have the mouse spp24 sequence not only for a more extensive comparison between species, but also as many functional studies are more easily performed in mice rather than humans. This work is presented in Chapter 3.

- **To determine the exon/intron structure of the human gene encoding spp24.**

Knowledge of the exon/intron structure of the mouse gene will enable a comparison to be made between the structure of the human and mouse gene. The structure of the mouse gene will also be essential for any functional studies that are performed in mouse. This work is presented in Chapter 3.

- **To perform an extensive sequence analysis of the human gene encoding spp24 and to determine the transcription initiation sites.**

An extensive sequence analysis of the complete genomic sequence of the human gene encoding spp24 will be performed using the HGMP Nix analysis environment. This will enable characterisation of many gene features and unusual features may be revealed that might provide clues as to the function of the gene. The transcription initiation sites of the gene will be determined by 5'RACE and primer extension. This work is presented in Chapter 3.

- **To investigate the nature of a suspected insertion/deletion polymorphism apparent with the restriction enzymes *BglII*, *HpaI* and *SstI*.**

Work completed prior to this thesis revealed a polymorphism in the human gene encoding spp24. This was thought to be an insertion/deletion polymorphism. The aim was to determine the nature of this polymorphism by subcloning fragments of different alleles and sequencing. This work is presented in Chapter 3.

- **To identify any polymorphic tandem repeats.**

Polymorphic tandem repeats in or near the human gene encoding spp24 will be useful as markers for future linkage or association studies if spp24 is suspected to be involved in a particular disease. This work is presented in Chapter 3.

- **To characterise the expression profiles of the human and mouse genes encoding spp24.**

The temporal and spatial expression profile of a gene can provide valuable information that may elucidate the potential functions of the protein. Expression data will be obtained in a variety of ways including information from ESTs, RT-PCRs, and microarrays. This work is presented in Chapter 4.

- **To compare the spp24 protein between species.**

Protein sequences for spp24 will be obtained from as many species as possible. These sequences will then be aligned and compared to identify the most conserved regions of the protein. This will highlight highly conserved residues that are likely to be critical to the function of the protein. This work is presented in Chapter 5.

- **To look for proteins showing significant homology to spp24 and attempt to model the protein.**

By identifying proteins that show significant homology to spp24, either in their sequence or in their domain complexity, it may be possible to determine some possible functions of spp24 based on similarity. Any proteins identified in this way will then be used to attempt to model the cystatin-like region of spp24 using an evolutionary trace (ET) analysis technique and identify residues that are likely to be important to the function of the protein. This work is presented in Chapter 6.

Chapter 2

Materials and Methods

2.1 Centrifugation

Unless otherwise stated, all volumes up to 1.5 ml were centrifuged in an MSE Micro Centaur centrifuge at 13,000 rpm. Larger volumes were centrifuged in either a Sorvall RC-5B Refrigerated Superspeed centrifuge (Du Pont Instruments) using a Sorvall SS-34 or GS-3 rotor up to 10,000 rpm or in a Sorvall RT 6000D (Du Pont Instruments) with free swinging rotor, type PN11053, up to 3,000 rpm.

2.2 Storage and Handling of *Escherichia coli* (*E. coli*)

2.2.1 Storage

Plate cultures were kept at 4°C and sealed with parafilm if long-term storage was required. Liquid cultures of *E. coli* for long-term storage were frozen in medium containing 1 × HMFM (3.6 mM K₂HPO₄, 1.3 mM KH₂PO₄, 2 mM sodium citrate, 1 mM MgSO₄, 4.4% (v/v) glycerol) at -70°C.

2.2.2 Media

All liquid cultures of *E. coli* were grown in Luria Bertani broth (LB) (10 g.l⁻¹ tryptone, 5 g.l⁻¹ yeast extract, 5 g.l⁻¹ NaCl, pH 7.5, autoclaved).

E. coli were streaked or plated onto LB agar plates (15 g.l⁻¹ of agar added to LB medium and autoclaved). All *E. coli*, unless otherwise stated, were grown at 37°C. Liquid cultures were shaken on a G10 Gyrotory shaker (New Brunswick Scientific) at 223 rpm.

2.2.3 *E. coli* strains

The *E. coli* strains DH5α and XL1-Blue MRF' were used. The genotypes of these strains are as follows: DH5α (Gibco BRL, Life Technologies); ϕ 80dlacZΔM15, *recA1*, *endA1*, *gyrA96*, *thi-1*, *hsdR17* (*r_k*⁻, *m_k*⁺), *supE44*, *relA1*, *deoR*, Δ(*lacZYA-argF*)U169. XL1-Blue

MRF' (Jerpseth *et al.* 1992); $\Delta(mcrA)183$, $\Delta(mcrCB-hsdSMR-mrr)173$, *endA1*, *supE44*, *thi-1*, *recA1*, *gyrA96*, *relA1*, *lac[F'proAB,lacIqZ Δ M15,Tn10(tet')]*^c.

2.2.4 Antibiotics

All antibiotics were made and stored at a 100 × concentration. The details of each antibiotic are shown in Table 2.1.

2.2.5 Preparing and transforming chemically competent *E. coli* cells

This procedure is based on the method of Hutchison and Halvorson (1980).

A 2.5 ml culture of *E. coli* host cells, with antibiotic if appropriate, was grown overnight at 37°C. To inoculate a larger culture, 1.5 ml of the overnight culture was added to 75 ml of fresh pre-warmed LB broth. The culture was grown to a cell density of 0.36-0.44 at 560 nm and then cooled on ice. The cells were harvested by centrifugation at 3,000 rpm for 5 minutes at 4°C in a Sorvall RT 6000D (Du Pont Instruments) with free swinging rotor, type PN11053, and then resuspended in 20 ml of cold 50 mM CaCl₂. The cells were incubated on ice for 15 minutes and then centrifuged as previously and resuspended in 5 ml of cold 50 mM CaCl₂, 5% (v/v) glycerol. The cells were split into 200 µl aliquots and frozen in microcentrifuge tubes in a dry ice/IMS (Industrial Methylated Spirits) bath. Aliquots of competent cells were stored at -70°C.

A 200 µl aliquot of competent cells was thawed on ice. The DNA to be transformed (approximately 10 ng) was diluted to 100 µl in 10 mM Tris-HCl pH 7.4, 10 mM MgCl₂, 10 mM CaCl₂. The diluted DNA and thawed competent cells were mixed and incubated on ice for 25 minutes. The mixture was then heat shocked at 37°C for 1.5 minutes and held at room temperature for 10 minutes. To allow the cells to recover, 1 ml of LB broth was added to the cells and they were incubated for 1 hour at 37°C. Aliquots of the culture were then plated onto the appropriate selective plates and incubated overnight at 37°C.

2.2.6 Preparing and electroporating electrocompetent *E. coli* cells

This procedure is based on the method from Dower *et al.* (1988) and Taketo (1988).

Table 2.1. Antibiotics.

This table gives details of all antibiotics used, the working concentration, the appropriate storage and the solvent in which they should be dissolved. All antibiotics should be filter sterilised after preparation.

Antibiotic	Solvent	Treatment	Working Conc.	Storage
Ampicillin	Water	Filter sterilise	50 $\mu\text{g.ml}^{-1}$	-20°C
Kanamycin	Water	Filter sterilise	25 $\mu\text{g.ml}^{-1}$	-20°C
Tetracycline	Ethanol/Water (50 % v/v)	Filter sterilise	12.5 $\mu\text{g.ml}^{-1}$	-20°C Protected from light
Gentamicin	Water	Filter sterilise	7 $\mu\text{g.ml}^{-1}$	4°C

A 2.5 ml culture of *E. coli* host cells, with antibiotic if appropriate, was grown overnight at 37°C. To inoculate a larger culture, 2 ml of the overnight culture were added to 1000 ml of fresh pre-warmed LB broth. The culture was grown to a cell density of approximately 0.45 at 560 nm and then cooled on ice.

The cells were harvested by centrifugation at 4,000 rpm for 15 minutes at 4°C in a Sorvall RC-5B Refrigerated Superspeed centrifuge (Du Pont Instruments) using a Sorvall SS-34 rotor and then resuspended in 1 litre of distilled water. The cells were centrifuged three more times as described above, being resuspended in 0.5 litres of distilled water, followed by 20 ml of 10% (v/v) glycerol and then 2 ml of 10% (v/v) glycerol. The cells were then split into 40 µl aliquots and frozen in microcentrifuge tubes in a dry ice/IMS bath. Aliquots of competent cells were stored at -70°C.

A 200 µl aliquot of competent cells was thawed on ice. The DNA to be transformed (approximately 10 ng) was in 1 to 5 µl of a low-conductivity medium, either 1 × TE or water. On ice, the DNA was mixed with 40 µl of electrocompetent cells and transferred to a chilled cuvette. The cuvette was placed in an electroporator (BioRad Genepulser) and a pulse delivered (1.5 kV, 25 µF). The cuvette was removed and 1 ml of SOC medium (2% (w/v) bacto-tryptone, 0.5% (w/v) yeast extract, 10 mM NaCl, 2.5 mM KCl, 20 mM MgCl₂, 20 mM MgSO₄, 20 mM glucose, Hanahan, (1983)) added to recover the cells. The electroporated cells in SOC medium were then shaken at 37°C for 1 hour. Aliquots of the culture were plated onto the appropriate selective plates and incubated overnight at 37°C.

2.2.7 Selection for transformants

Antibiotics were added to LB plates (as detailed in section 2.2.4) to select for the plasmid. When vectors capable of α-complementation were used, indication of plasmids containing inserts was done using the blue/white colour screening system (Horwitz *et al.* 1964; Ullman *et al.* 1967). X-gal (5-bromo-4-chloro-3-indolyl-beta-D-galactosidase; 40 µg.ml⁻¹ in N,N'-dimethylformamide) and IPTG (Isopropylthio-beta-D-galactosidase; 0.1 mM) were added to LB plates. Colonies containing a plasmid only gave blue colonies and colonies containing a plasmid with an insert gave white colonies.

2.3 Use of restriction endonucleases

Unless otherwise stated, all restriction endonucleases were from Gibco BRL, Life Technologies Ltd and were at a concentration of 10 units. μl^{-1} .

In a 20 μl total reaction volume, 1 μl of each required enzyme was used and 2 μl of the appropriate 10 \times REact buffer (Gibco BRL, Life Technologies Ltd). The reaction was incubated at 37°C ,or the temperature recommended by the manufacturer, for 60 or 90 minutes for single and double digests respectively.

2.4 Agarose gel electrophoresis

If DNA was to be recovered from a gel then SeaPlaque agarose (FMC BioProducts) was used with 1 \times TAE (40 mM Tris base, 40 mM acetic acid, 1 mM EDTA) as the gel solvent and gel running buffer. If DNA was not to be recovered then SeaKem LE agarose (FMC BioProducts) was used with 1 \times TBE (89 mM Tris base, 89 mM boric acid, 2 mM EDTA) as the gel solvent and gel running buffer.

Agarose was used at a concentration of between 0.8% and 1.5% (w/v). Ethidium bromide was added to the gel and the running buffer to a final concentration of 0.5 $\mu\text{g}.\mu\text{l}^{-1}$. Gels were cast at either a 40 ml or 100 ml volume in a transparent plastic tray with a plastic comb to create the wells. The gels were run at a maximum of 100 volts. The DNA on the gels was viewed using either AlphaImager v.3.24i on an AlphaImager 2000 system or Genesnap on a Gene Genius Bio Imaging system (Syngene). For analysis AlphaEase v.4.0 and GeneTools were used respectively.

2.5 Ethanol precipitation

Unless otherwise stated, all ethanol precipitations of DNA were carried out by adding 2.5 volumes of ethanol and 0.1 volumes of 3 M sodium acetate (pH 5.1). The tube was mixed well and then stored at -70°C for at least 15 minutes or until required. To recover the DNA, the solution was centrifuged at 13,000 rpm for 30 minutes in an MSE Micro Centaur centrifuge. The supernate was removed and the pellet washed in 70% (v/v) ethanol. The tube was then centrifuged again as previously for 15 minutes. The supernate was removed and the

pellet air dried for no more than 5 minutes to evaporate the last traces of ethanol. The pellet was then resuspended in the desired volume of 1 × TE (10 mM Tris-HCl, 1 mM EDTA, pH 7.5) or water.

2.6 Isolation of plasmid DNA from *E. coli*

2.6.1 Standard miniprep

This method is modified from Ish-Horowicz and Burke (1981).

A small culture of bacteria in LB broth (2.5 ml) was grown overnight with the appropriate antibiotic for plasmid selection.

To harvest the cells, 1.5 ml of the overnight culture was centrifuged at 13,000 rpm for 1 minute in an MSE Micro Centaur centrifuge. The broth was removed and the tubes were centrifuged again briefly. The last of the broth was pipetted off. The pellet was resuspended in 200 µl of Solution I (50 mM Glucose, 25 mM Tris-HCl pH 8.0, 10 mM EDTA) and held at room temperature for 5 minutes. Then, 200 µl of Solution II (0.2 N NaOH, 1% (w/v) SDS) was added and the tube mixed gently by inversion several times. The tube was then placed in ice for 5 minutes. Next, 200 µl of Solution III (5 M Potassium Acetate, pH 5.5) was added. The tube was again gently mixed by inversion and then replaced on ice for a further 5 minutes.

The tube was then centrifuged as previously for 1 minute and 500 µl of the clear supernate transferred to a fresh tube. To this tube, 1 ml of ethanol was added. The tube was mixed well and held at room temperature for at least 2 minutes. The DNA was then pelleted by centrifuging as previously for 1 minute and the supernate discarded. The pellet was washed in 70% (v/v) ethanol, centrifuged and the supernate again discarded. The tube was then centrifuged again briefly and the last of the ethanol removed with a pipette. The pellet was air dried for no more than 5 minutes before being resuspended in 50 µl of 1 × TE (10 mM Tris-HCl pH 7.6, 1 mM EDTA). DNA was stored at -20°C. When the was analysed by restriction enzyme digestion, as standard, 2 µl of DNA was used in a 20 µl reaction volume.

2.6.2 Preparation of plasmid DNA using Qiagen kits

This was carried out according to the manufacturer's protocol.

For large scale preps or cleaner preps than those produced by the standard miniprep method, Qiagen kits were used. The protocol below describes the Qiagen Midi prep. The procedure can be scaled up to a Maxi, Mega or Giga prep.

A single colony was picked from a freshly streaked plate and used to inoculate a 2 ml starter culture containing the appropriate antibiotic. The culture was incubated at 37°C overnight, with shaking. The starter culture was diluted 1 in 500 into 25 ml of selective LB broth. This culture was incubated at 37°C overnight, with shaking.

The bacterial cells were harvested by centrifuging at 6,000 rpm for 15 minutes at 4°C in a Sorvall RC-5B Refrigerated Superspeed centrifuge (Du Pont Instruments) using a Sorvall SS-34 rotor. The supernate was discarded and the pellet resuspended in 4 ml of buffer P1 (50 mM Tris-HCl pH 8.0, 10 mM EDTA, 100 µg.ml⁻¹ RNase A). Four ml of buffer P2 (200 mM NaOH, 1% (w/v) SDS) was added and the sample mixed gently by inverting 4 to 6 times. The mixture was incubated for 5 minutes at room temperature. After incubation, 4 ml of chilled buffer P3 (3 M potassium acetate, pH 5.5) was added. The sample was mixed gently by inverting 4 to 6 times and incubated on ice for 15 minutes.

Following incubation, the sample was filtered through plastic filter wool (Algarde) and the filtrate centrifuged as previously at 10,000 rpm for 15 minutes at 4°C. Meanwhile, a Qiagen-tip 100 was equilibrated by applying 4 ml of buffer QBT (750 mM NaCl, 50 mM MOPS pH 7.0, 15% (v/v) isopropanol, 0.15% (v/v) Triton X-100) and allowing the column to empty by gravity flow.

The supernate from the centrifugation was applied to the column to bind the DNA. The Qiagen-tip was then washed by applying 2 aliquots of 10 ml of buffer QC (1 M NaCl, 50 mM MOPS pH 7.0, 15% (v/v) isopropanol).

The DNA was eluted with 5 ml of buffer QF (1.25 M NaCl, 50 mM Tris-HCl pH 8.5, 15% (v/v) isopropanol) and then precipitated by the addition of 3.5 ml (0.7 volumes) of room temperature isopropanol. The sample was centrifuged as previously at 10,000 rpm for 30 minutes at 4°C. The supernate was decanted and the pellet washed with 2 ml of 70% (v/v)

ethanol and centrifuged as previously at 10,000 rpm for 10 minutes. The supernate was again decanted and the pellet air-dried for 5 to 10 minutes before being dissolved in a suitable volume of 1 × TE (10 mM Tris-HCl, 1 mM EDTA, pH 7.5).

2.7 DNA extraction from human blood

This is based on the method described by Sambrook *et al.* (1989).

Blood samples in 10 ml EDTA tubes, were thawed into 50 ml polypropylene, flat cap centrifuge tubes (Corning). Ice cold distilled water was added to make the volume up to 45 ml and the sample mixed well.

The tubes were then spun at 2,500 rpm for 20 minutes at 4°C in a Sorvall RT 6000D. The supernate was poured off and ice cold 0.1% (v/v) Nonidet P40 was added to the pellet to make the volume up to 35 ml. The sample was vortexed to break up the pellet and then centrifuged as previously for 20 minutes at 4°C. The supernate was discarded. To the pellet, 7 ml of filtered 6 M guanidinium hydrochloride was added and 0.5 ml of 7.5 M ammonium acetate. The tube was vortexed until the pellet had completely dispersed.

Next, 0.5 ml of 20% (w/v) sodium sarkosyl was added and 75 µl of proteinase K (20 mg.ml⁻¹). The sample was vortexed to mix and then incubated at 60°C for 90 minutes. Then, 17 ml of 96% (v/v) ethanol was added and the sample gently mixed. The DNA was then spooled out and redissolved overnight in 1 ml of 1 × TE (10 mM Tris-HCl pH 7.6, 1 mM EDTA). This was done in 10 ml tubes on a rotating wheel at 4°C. The DNA was then reprecipitated by adding 100 µl of 3 M sodium acetate, pH 5.5 and 2.5 ml of ice cold 96% (v/v) ethanol. The DNA was again spooled out and redissolved overnight, as previously, in 1ml of 1 × TE. DNA samples were stored at -70°C.

2.8 Extraction of RNA from mammalian tissues

2.8.1 RNA extraction using the guanidinium-lithium chloride method

This is based on a method from Wilkinson (1991).

Cells were lysed in GTEM buffer (5 M guanidinium thiocyanate, 50 mM Tris-HCl (pH 7.5), 10 mM EDTA, 1.12 M 2-mercaptoethanol). Lysis was taken to completion by homogenising tissue in the GTEM buffer with a Polytron homogeniser. An equal volume of chloroform:isoamyl alcohol (24:1) was then added to the homogenised tissue and the sample vortexed vigorously to mix. The sample was then centrifuged at 10,000 rpm for 10 minutes in a Sorvall RC-5B Refrigerated Superspeed centrifuge (Du Pont Instruments) using a Sorvall SS-34 rotor. The aqueous (upper) phase was then transferred to a fresh tube containing 1.4 volumes of 6 M lithium chloride. The sample was mixed gently by inversion and then incubated at 4°C for at least 15 hours.

After the incubation, the sample was centrifuged at 10,000 rpm for 30 minutes and the supernate was removed. The pellet was resuspended in PK buffer (50 mM Tris-HCl (pH 7.5), 5 mM EDTA, 0.5% (w/v) SDS, 200 µg.ml⁻¹ proteinase K) using half the volume that was used of GTEM buffer in the first stage of the protocol. The sample was then incubated at 45°C for 30 minutes. Next, a 0.1 volume of 3 M sodium chloride was added and the sample mixed. Then, three phenol:chloroform:isoamyl alcohol (25:24:1) extractions were carried out, each time using an equal volume of phenol:chloroform:isoamyl alcohol, centrifuging the sample as previously for 10 minutes and then removing the aqueous (upper) phase to a clean tube. After the third extraction, the RNA was precipitated by adding 2.5 volumes of ethanol and incubating at -20°C for at least 2 hours. The sample was then centrifuged as previously for 15 minutes at 4°C and then resuspended in a volume of DEPC-treated water appropriate for further procedures.

2.8.2 RNA extraction using the RNazol B kit

This method uses the RNazol B kit (AMS Biotechnology (Europe) Ltd) and was carried out according to the manufacturer's protocol.

The tissue sample was homogenised in RNazol B (2 ml per 100 mg of tissue) using a Polytron homogeniser. To every 2 ml of homogenate, 0.2 ml of chloroform was added and the sample shaken vigorously for 15 seconds. The sample was then incubated on ice for 5 minutes. Next, the sample was centrifuged at 10,000 rpm for 15 minutes in a Sorvall RC-5B Refrigerated Superspeed centrifuge (Du Pont Instruments) using a Sorvall SS-34 rotor. The aqueous (upper) phase was transferred to a clean tube and an equal volume of isopropanol added. The sample was incubated at 4°C for 15 minutes and then centrifuged as previously for

15 minutes at 4°C. The supernate was removed and the pellet washed in 0.8 ml of 75% (v/v) ethanol by vortexing and subsequent centrifugation as previously for 8 minutes at 4°C. The pellet was then resuspended in 0.5% (w/v) SDS or 1 mM EDTA, pH 7.0 (both solutions were treated with DEPC).

2.9 Recovery of DNA from an agarose gel

2.9.1 Recovery of DNA from an agarose gel using phenol/chloroform extraction

The phenol/chloroform extraction method was taken from www.bioproducts.com/technical/headers/tech_header9.shtml where it had been modified from Sambrook *et al.* (1989).

DNA was electrophoresed in SeaPlaque agarose (FMC BioProducts) prepared in 1 × TAE (40 mM Tris base, 40 mM acetic acid, 1 mM EDTA). The gel fragment containing the DNA was excised from the gel using a scalpel blade and placed in a microcentrifuge tube. The tube was weighed and if the gel slice was significantly more than 200 mg, then it was broken into smaller pieces and split between further tubes. The gel slice was then placed at 67°C for 10 minutes to melt the agarose.

The appropriate volume of prewarmed (67°C) 1 × TE (10 mM Tris-HCl pH 7.6, 1 mM EDTA) was added so that the final concentration of agarose was $\leq 0.5\%$. An equal volume of phenol was added to the sample and it was vortexed for 15 seconds to mix. The tube was then centrifuged at 13,000 rpm for 3 minutes in an MSE Micro Centaur centrifuge. The aqueous (upper) layer was transferred to a clean tube. The phenol extraction was repeated. A third extraction was performed with an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1). A final extraction was done using an equal volume of chloroform and the resulting aqueous phase chilled on ice for 15 minutes.

The sample was then centrifuged as previously for 15 minutes at 4°C. The supernate was decanted into a fresh tube and the DNA ethanol precipitated as detailed in section 2.5.

2.9.2 Recovery of DNA from an agarose gel using the QIAquick gel extraction kit

This method used the QIAquick gel extraction kit (Qiagen) and was carried out according to the manufacturer's protocol.

DNA was electrophoresed and excised as detailed in section 2.9.1. If the weight of agarose was greater than 400 mg then the gel slice was broken into smaller segments and split between further tubes. Three volumes of buffer QG were added to each volume of gel and the sample incubated at 50°C for 10 minutes or until the gel slice had completely dissolved. The tube was vortexed every 2 to 3 minutes during incubation to help dissolve the agarose. One gel volume of isopropanol was added to the tube and the sample mixed.

A QIAquick spin column was placed in a 2 ml collection tube and the sample applied into the column to bind the DNA. The column was then centrifuged at 13,000 rpm for 1 minute in an MSE Micro Centaur centrifuge. The flow-through was discarded and the column placed back in the same collection tube.

To wash the sample, 0.75 ml of buffer PE was applied to the column and the column centrifuged as previously for 1 minute. The flow-through was again discarded and the column centrifuged for an additional 1 minute at 13,000 rpm.

The QIAquick column was then placed in a fresh 1.5 ml microcentrifuge tube. To elute the DNA, 30 µl of buffer EB (10 mM Tris-HCl, pH 8.5) was added to the centre of the QIAquick membrane and the column left to stand for 1 minute. The column was then centrifuged as previously for 1 minute and the DNA collected was transferred to a fresh tube for storage at -20°C until required.

2.10 Hybridisations

2.10.1 Southern blotting

This is based on the method from Dalglish (1987) which is a modification of the original method by Southern (1975) (reprinted 1992).

2.10.1.1 Blotting the gel

Gels were prepared for blotting by washing for 7 minutes in depurinating solution (0.25 M HCl), rinsing in distilled water and washing for 30 minutes in denaturing solution (0.5 M NaOH, 1.5 M NaCl). Gels were then rinsed again in distilled water and soaked for a further 30 minutes in neutralizing solution (3 M NaCl, 0.5 M Tris-HCl, pH 7.4). Blotting apparatus was assembled as depicted in figure 4.2, Dalglish (1987). Clingfilm was used to mask off the 3MM paper surrounding the gel. All Whatman paper and the nylon membranes (Hybond-N; Amersham Pharmacia Biotech) were soaked in $3 \times \text{SSC}$ ($20 \times \text{SSC}$ is 3 M NaCl, 0.3 M sodium citrate, pH 7.2) before addition to the blotting apparatus. A glass pipette was used to roll out any bubbles between the gel, the membrane and Whatman papers. Blots were left overnight with wet paper towels being replaced with dry every 5 minutes during the first 30 minutes.

On completion of blotting, the apparatus was dismantled. The origin was marked on the membrane and the bottom right hand corner always cut off for orientation purposes. Membranes were left on 3MM paper to air dry and then crosslinked in a UV crosslinker. (Amersham Life Science model RPN 2500/2501, 10-15 seconds at $70,000 \mu\text{J}.\text{cm}^{-2}$).

2.10.1.2 Preparation of the probe

DNA to be labelled as a probe was boiled at 100°C for 2 minutes and then chilled on ice. The labelling reaction was assembled at room temperature by adding components in the following order to give a total reaction volume of 15 μl : 3 μl of oligo labelling buffer (OLB; section 2.10.1.2.1), 0.6 μl of BSA ($10 \text{ mg}.\text{ml}^{-1}$), 7 to 10 ng of DNA, 1 μl of $[\alpha\text{-P}^{32}] \text{dCTP}$ (NEN, $10 \text{ mCi}.\text{ml}^{-1}$) and 0.6 μl of Klenow DNA polymerase (USB Corporation, $1 \text{ unit}.\mu\text{l}^{-1}$). The labelling reaction was left to proceed overnight at room temperature. The reaction was stopped by adding 85 μl of oligo stop solution (20 mM NaCl, 20 mM Tris-HCl pH 7.5, 2 mM EDTA, 0.25% (w/v) SDS).

2.10.1.2.1 Preparation of oligo labelling buffer (OLB)

OLB comprises solutions A, B and C mixed together in the ratio 2:5:3. Solution A is made by assembling the following components: 625 μl 2 M Tris-HCl, pH 8.0, 25 μl MgCl_2 , 350 μl water, 18 μl 2-mercaptoethanol, 5 μl dATP, 5 μl dTTP and 5 μl dGTP. Solution B is 2 M

HEPES, pH 6.6 and solution C is random hexadeoxyribonucleotides (Amersham Pharmacia Biotech) at 90 OD units.ml⁻¹.

2.10.1.3 Checking incorporation of the probe

This is based on the method described by Sambrook *et al.* (1989).

From the stopped labelling reaction, 1 µl of the probe was taken and mixed with 11 µl of water. Five microlitres of the diluted probe was spotted onto each of two pieces of DE-81 paper (Whatman). One piece was labelled 'T' for total and the other 'P' for precipitable. Each piece was checked with the Geiger counter to ensure that it gave an equal number of counts as the other. The 'P' filter was then washed six times, for 5 minutes each time, in 0.5 M NaH₂PO₄. There were then a further two, 5-minute washes in water followed by two, 5-minute washes in IMS (Industrial Methylated Spirits). The filter was allowed to dry and then the counts of the 'P' and 'T' filters compared. The counts of 'P' divided by the counts of 'T', multiplied by 100 gives the percentage incorporation.

2.10.1.4 Hybridisation of the probe

Hybridisations were carried out in a Hybaid hybridisation oven. The filter to be probed was washed for 2 hours at 65°C with 15 ml of pre-hybridisation solution (1.5 × SSPE (0.27 M NaCl, 15 mM Na₂PO₄, 1.5 mM EDTA), 0.5% (w/v) dried milk (Marvel), 1% (w/v) SDS, 6% (w/v) polyethylene glycol 8000). The probe DNA was boiled at 100°C for 2 minutes and then snap cooled on ice. The probe was then added directly to the pre-hybridisation buffer and the hybridisation allowed to proceed at 65°C overnight.

2.10.1.5 Post-hybridisation washes

All these washes were carried out at 65°C. The hybridisation solution was discarded and the filter washed 3 times for 2 minutes each in 15 ml of 3 × SSC, 0.1% (w/v) SDS. Further washes were carried out if necessary.

Four more stringent washes were then done for 10 minutes each in 0.5 × SSC, 0.1% (w/v) SDS. The filters were then blotted dry and wrapped in Saran Wrap.

2.10.1.6 Autoradiography

Filters were placed in an X-ray cassette fitted with an intensifying screen. In a dark room, a piece of film (Kodak XAR) was placed over the filter within the cassette.

The cassette was then put at -70°C for an appropriate time and then the film was either processed automatically (Cronex CX-130, Du Pont) or manually (5 minutes in developer, 5 minutes in stop solution, 5 minutes in fixer, 10 minute wash in running water).

2.10.2 Colony hybridisations

This is based on the method from Sambrook *et al.* (1989).

Cultures, from which colonies were to be hybridised, were spotted onto Hybond-N (Amersham Pharmacia Biotech) filters from a 96-well plate using a metal 8×12 array device. The filter was then laid colony side up, on a LB agar plate containing the appropriate antibiotic. The plate was incubated at 37°C overnight. Following incubation, filters were removed from the agar plate and laid on 3 MM Whatman paper (colony side up) soaked in $2 \times \text{SSC}$, 5% (w/v) SDS for 3 minutes.

The filters were microwaved (modified since original protocol) (650 W) until dry. They were then laid on 3 MM Whatman paper soaked in $5 \times \text{SSC}$, 0.1% (w/v) SDS for 3 minutes, followed by 3 MM Whatman paper soaked in $2 \times \text{SSC}$ for 5 minutes. The filters were then allowed to dry at room temperature. The hybridisations then followed the protocols described in sections 2.10.1.2 to 2.10.1.6.

2.10.3 Hybridisation to an RNA array

2.10.3.1 Probe preparation

Twenty nanograms of DNA was used as a probe and labelled as described in section 2.10.1.2.

2.10.3.2 Probe purification

To the labelled probe, $9 \mu\text{l}$ of salmon sperm DNA (10 mg.ml^{-1}) was added followed by

20 µl of 3 M sodium acetate. The probe was then mixed, 570 µl of ethanol added and mixed again. Precipitated DNA was pelleted by centrifuging at 13,000 rpm for 15 minutes in a MSE Micro Centaur centrifuge. The supernate was removed into a beaker of soapy water and 500 µl of 70% (v/v) ethanol added to the pellet. The probe was centrifuged as previously for 15 minutes and the supernate was removed into the same beaker of soapy water as before. The pelleted DNA was resuspended in 100 µl of sterile water. The radioactivity of the resuspended probe DNA and the soapy water in the beaker was compared to ensure greater than 70% incorporation of the probe.

2.10.3.3 Hybridisation of the probe

Hybridisations were carried out in a Hybaid hybridisation oven, according to the recommendations of Clontech.

Fifteen millilitres of ExpressHyb solution (Clontech) were warmed to 65°C and 1.5 mg of herring sperm DNA was denatured at 95°C for 5 minutes and then chilled on ice. The denatured herring sperm DNA was then mixed with the warmed ExpressHyb solution. The MTE RNA array (Clontech) was then prehybridised in 10 ml of the ExpressHyb/herring sperm DNA mixture for 30 minutes at 65°C. The labelled cDNA probe was mixed with 150 µg of herring sperm DNA and 50 µl of 20 × SSC to a total volume of 200 µl.

The probe was denatured at 95-100°C for 5 minutes and then 68°C for 30 minutes. The probe mixture was then added to the remaining 5 ml of ExpressHyb/herring sperm DNA mixture and mixed thoroughly. The prehybridisation solution was poured off the MTE RNA array and replaced with the 5 ml of ExpressHyb containing the probe. Hybridisation was left to proceed at 65°C overnight.

2.10.3.4 Post-hybridisation washes

Post-hybridisation washes were carried out according to Clontech's recommendations. Five, 20 minute washes were performed at 65°C in 2 × SSC, 1% (w/v) SDS followed by two 20 minute washes at 55°C in 0.1 × SSC, 0.5% (w/v) SDS. The array was then blotted dry and wrapped in Saran wrap.

2.10.3.5 Visualisation of MTE RNA array hybridisation result

Results were visualised by autoradiography as described in section 2.10.1.5 and also by Phosphorimaging. Phosphorimaging was carried out using a Phosphorimager (Molecular Dynamics) according to the manufacturer's protocol. The results were analysed using the program ImageQuant.

2.11 Polymerase chain reaction (PCR)

2.11.1 Standard PCR

This is based on the method from Mullis and Faloona (1987).

All standard PCRs were set up in a total reaction volume of 10 μl . In each reaction 10 to 25 ng of template DNA were used and an optimised amount of primer (usually 0.5 to 1 μM). An 11.1 \times buffer (Jeffreys *et al.* 1990) was used, giving concentrations in the final reaction of 45 mM Tris-HCl pH 8.8, 11 mM ammonium sulphate, 4.5 mM MgCl_2 , 6.7 mM 2-mercaptoethanol, 4.4 μM EDTA, 113 $\mu\text{g}\cdot\text{ml}^{-1}$ BSA, 1 mM dATP, 1 mM dCTP, 1 mM dGTP and 1 mM dTTP. To each reaction, 1 unit of Taq DNA polymerase was added (AB gene).

A typical reaction would comprise 0.9 μl 11.1 \times buffer, 1 μl DNA (10 to 25 ng), 0.5 μl of each primer (from 10 μM stock), 5.9 μl water and 0.2 μl Taq DNA polymerase (AB gene, 5 units. μl^{-1}).

PCRs were each carried out with individually optimised cycling conditions. These are detailed in each chapter accordingly. All reactions were carried out using a PTC-200 peltier thermal cycler (MJ Research). PCRs were analysed typically by running 5 μl on an agarose gel.

It should be noted that where proofreading activity of the DNA polymerase was important, a mixture of Pfu DNA polymerase (Stratagene, 2.5 units. μl^{-1}) and Taq DNA polymerase was used at a unit ratio of 1:20 (Pfu:Taq).

2.11.2 Radioactive PCR

Synthetic primers for a radioactive PCR were labelled using [γ - ^{33}P] ATP (NEN, 10 mCi.ml $^{-1}$) and the enzyme T4 polynucleotide kinase (Gibco BRL, Life Technologies Ltd, 10 units. μl^{-1}). Enough primer for 10 PCR reactions was labelled in a 10 μl reaction. In each reaction the amount of primer optimised for the PCR was labelled in 1 \times REact 1 buffer (Gibco BRL, Life Technologies Ltd). A typical reaction would constitute 0.5 μl primer (from 10 μM stock), 0.1 μl 10 \times REact 1, 0.2 μl water, 0.1 μl T4 polynucleotide kinase and 0.1 μl [γ - ^{33}P] ATP. The labelling reaction was allowed to proceed overnight at room temperature. A PCR reaction was then carried out as described in section 2.11.1 using one labelled and one unlabelled primer.

To each completed PCR reaction 4 μl of stop solution (95% (v/v) formamide, 20 mM EDTA, 0.05% (w/v) bromophenol blue, 0.05% (w/v) xylene cyanol FF) were added and then 7 μl of the sample ran on a 6% (w/v) denaturing polyacrylamide gel (section 2.12).

2.11.3 RT-PCR

The reverse transcription method is based on that from Temin and Mizutani (1970). The PCR method is based on that from Mullis and Faloona (1987).

Four micrograms of total RNA and 2.5 μl of primer (1 pmol. μl^{-1}) were mixed on ice, then incubated at 70°C for 10 minutes to denature the RNA and then snap cooled on ice for 1 minute to anneal the primer.

Four microlitres of the appropriate 5 \times buffer (Gibco BRL, Life Technologies Ltd, or Promega) 2 μl of 0.1 M DTT, 1 μl of a 10mM dNTP mix and 0.25 μl of RNasin (Promega, 20-40 units.ml $^{-1}$) were then added to the RNA and primer to give a total reaction volume of 20 μl .

The sample was incubated at 42°C for 2 minutes and then 1 μl of Superscript II (Gibco BRL, Life Technologies Ltd, 200 units. μl^{-1}) or M-MLV reverse transcriptase RNase H minus (Promega, 100-200 units. μl^{-1}) added before incubation for a further 50 minutes at 42°C. The sample was heated at 70°C for 15 minutes to inactivate the enzyme and then chilled on ice.

Two microlitres of cDNA from the reverse transcription reaction were used as template DNA for a PCR. The PCR was then performed as described in section 2.11.1, but scaling up to total reaction volume of 100µl. Ten microlitres were then run on an agarose gel for analysis.

2.11.4 Purification of PCR products using the QIAquick PCR purification kit

This was carried out according to the manufacturer's protocol (Qiagen).

Five volumes of buffer PB were added to 1 volume of the PCR reaction and the sample mixed. The sample was then applied to a QIAquick spin column, placed in a 2 ml collection tube, and centrifuged at 13,000 rpm for 1 minute in an MSE Micro Centaur centrifuge. The flow-through was discarded and the column placed back in the same collection tube.

To wash the DNA, 0.75 ml of buffer PE was applied to the column and the sample centrifuged as previously for 1 minute. The flow-through was discarded and the column centrifuged for an additional 1 minute.

The column was then placed in a clean 1.5 ml microcentrifuge tube. To elute the DNA, 50 µl of buffer EB (10 mM Tris-HCl pH 8.5) were applied to the centre of the QIAquick membrane. The sample was allowed to stand for 1 minute. The column was then centrifuged as previously for 1 minute and the DNA collected was transferred to a fresh tube for storage at -20°C until required.

2.12 Polyacrylamide gels

Gels were prepared 1-20 hours prior to use.

2.12.1 Preparing the plates

Two glass plates (Gibco BRL, Life Technologies Ltd, 31 cm × 38.5 cm) were washed in diluted Decon 90 (dilution of approximately 1 in 5) and then rinsed in distilled water.

The plates were dried with paper towels, cleaned with ethanol and then left to dry. The shorter of the two plates was coated with Gel Slick (Flowgen) and left to air dry.

Plastic spacers were then inserted between the two plates at the edges. They were pushed down so the rubber pad on the top of the spacer was flush with the top edge of the shorter

plate. The plates were then inserted into a S2 casting boot (Gibco BRL, Life Technologies Ltd) ready to pour. Vinyl 0.4 mm spacers were used and a 28 cm Mylar sharktooth 62 point, 0.35 mm comb with a point to point tooth distance of 5 mm.

2.12.2 Pouring the gel

Sequagel solutions (National Diagnostics) were used to make a 6% polyacrylamide gel. Into a glass beaker, 14.4 ml of Sequagel concentrate (25%- 237.5 g.l⁻¹acrylamide, 12.5 g.l⁻¹ methylene bisacrylamide, 8.3 M urea), 39.6 ml of Sequagel diluent (8.3 M urea) and 6 ml of Sequagel buffer (50% urea (8.3M) in 1M Tris-Borate 20mM EDTA buffer) were added and mixed. To this mixture 60 µl of TEMED (Sigma) were added and 280 µl of 10% (w/v) ammonium persulphate.

A syringe was used to pour the gel between the two plates at such an angle as to avoid air bubbles. A comb was then inserted into the top of the gel (flat edge first) and bull dog clips clamped over the two plates. The gel was left to polymerise for at least one hour before use.

2.12.3 Gel electrophoresis

The gel was removed from the rubber casting boot and the comb removed from the top. The comb was then reinserted the other way around so that the teeth were about 2mm into the gel. The gel apparatus (Gibco BRL, Life Technologies Ltd, model S2) was assembled and 1 × TBE (89 mM Tris base, 89 mM boric acid, 2 mM EDTA) buffer added to the top and bottom chambers. The gel was prerun until the temperature of the front plate was about 50°C. Samples were then loaded and the gel run at a constant current of 55 mA.

2.12.4 Post-electrophoresis

The plates were removed from the gel running apparatus and the comb and spacers removed. The plates were then prised apart using a small spatula to leave the gel sticking to the plate that had not been treated with Gel Slick.

The gel was then soaked in 5% (v/v) acetic acid, 15% (v/v) methanol, to remove the urea, for a minimum of 5 minutes and a maximum of 10 minutes. The gel was then removed from the

glass plate onto a sheet of 3MM Whatman filter paper and dried at 80°C for 2 hours on a gel drier (BioRad, model 583). Autoradiography was carried out as described in section 2.10.1.6.

2.13 Cloning procedures

2.13.1 Dephosphorylation

This was carried out according to the manufacturer's protocol (Amersham Pharmacia Biotech).

The DNA to be dephosphorylated was incubated with shrimp alkaline phosphatase buffer (20 mM Tris-HCl pH 8.0, 10 mM MgCl₂) and 0.1 units per 1 pmol of 5'-protruding DNA termini of shrimp alkaline phosphatase (Amersham Pharmacia Biotech), for 1 hour at 37°C in a total reaction volume of 10 µl. The shrimp alkaline phosphatase was then inactivated by heating the sample at 65°C for 15 minutes.

2.13.2 Ligation

This method was based on that described by Sambrook *et al.* (1989).

DNA was ligated using T4 DNA ligase (Gibco BRL, Life Technologies Ltd) in the Gibco recommended buffer (50 mM Tris-HCl pH 7.6, 10 mM MgCl₂, 1 mM ATP, 1 mM DTT, 5% (w/v) polyethylene glycol-8000). A molar ratio of 1:3, vector:insert, was used in a total reaction volume of 10 µl. Ligation reactions were carried out at 16°C for 16 hours.

2.13.3 Cre-*loxP* recombination reaction

The cre-*loxP* reaction was carried out using the Clontech kit according to the manufacturer's protocols.

Two hundred nanograms of donor vector DNA was mixed with 200 ng of acceptor vector, 2 µl of 10 × Cre buffer, 2 µl of 10 × BSA (1 mg.ml⁻¹), 1 µl of Cre recombinase (100 ng. µl⁻¹) and water to give a total volume of 20 µl. The reactions were allowed to proceed at room temperature for 15 minutes before being stopped by heating at 70°C for 5 minutes.

Ten microlitres of the Cre-*loxP* reaction was then transformed into chemically competent *E. coli* cells as described in section 2.2.5.

2.14 Sequencing

2.14.1 Manual sequencing

Manual sequencing was carried out using the T7 Sequenase V2.0 kit (Amersham Pharmacia Biotech) and was based on the method by Sanger *et al.* (1977).

2.14.1.1 Preparing double stranded DNA for sequencing

For each sequencing reaction 3 to 5 µg of plasmid DNA was used that had been purified using a Qiagen kit (section 2.6.2).

The DNA was alkaline denatured by adding 0.1 volumes of 2 M NaOH, 2 mM EDTA and incubating at 37°C for 30 minutes. The mixture was neutralized by adding 0.1 volumes of 3 M sodium acetate (pH 4.5-5.5) and then precipitated with 2.5 volumes of ethanol and put at -70°C for 15 minutes. The DNA samples were then centrifuged at 13,000 rpm for 30 minutes in an MSE Micro Centaur centrifuge. The supernatant was discarded and the pellet washed in 70% (v/v) ethanol and then centrifuged as previously for 15 minutes. The pellet was then resuspended in 7 µl of distilled water.

2.14.1.2 The sequencing reaction

For each sequencing reaction, a single annealing reaction was set up in a total reaction volume of 10 µl. The reaction consisted of 3 to 5 µg of denatured DNA in 7 µl of water, 2 µl of reaction buffer (200 mM Tris-HCl pH 7.5, 100 mM MgCl₂, 250 mM NaCl) and 1 µl of primer (0.5 to 1.0 pm).

The reaction was heated for 2 minutes at 65°C and then cooled slowly to <35°C over 15-30 minutes. While the annealing mixture was cooling, 4 tubes were labelled, filled and capped with 2.5 µl of each termination mixture (ddG (80 µM dGTP, 80 µM dATP, 80 µM dTTP, 80 µM dCTP, 8 µM ddGTP and 50 mM NaCl), ddT (as ddG, but with 8 µM ddTTP instead of ddGTP), ddA (as ddG, but with 8 µM ddATP instead of ddGTP), ddC (as ddG, but with 8 µM ddCTP instead of ddGTP) and pre-warmed at 37°C. The labelling mix (7.5 µM dGTP, 7.5 µM dCTP, 7.5 µM dTTP) was diluted 5-fold to a working concentration.

Once the annealing mixture had cooled, it was centrifuged briefly at 13,000 rpm in an MSE Micro Centaur centrifuge and chilled on ice. To the ice-cold annealed DNA mixture (10 µl), 1 µl of DTT (0.1 M) was added, 2 µl of diluted labelling mix, 0.5 µl of [α -³⁵S] dATP (NEN, 12.5 mCi.ml⁻¹) and 2 µl of diluted Sequenase polymerase (diluted 1 in 8 with dilution buffer supplied in kit). The reaction was mixed and incubated at room temperature for 2-5 minutes.

To terminate the reaction, 3.5 µl of the above labelling reaction was transferred to each of the pre-warmed termination tubes. This was mixed and incubated at 37°C for 5 minutes.

The termination reactions were stopped by adding 4 µl of stop solution (95% (v/v) formamide, 20 mM EDTA, 0.05% (w/v) bromophenol blue, 0.05% (w/v) xylene cyanol FF). The contents of the tubes were mixed and either stored on ice or at -20°C until ready to load.

2.14.1.3 Gel electrophoresis

The completed sequencing reaction was heated to 75°C for 2 minutes and then snap cooled on ice. Three microlitres were then loaded in each lane. The gel was run as described in sections 2.12.1 to 2.12.4. When needed, a second loading was carried out sometime into the run to give a longer read.

2.14.2 Automated sequencing

Automated sequencing was carried out by PNACL DNA services at the University of Leicester. DNA purified using the Qiagen kits (section 2.6.2) was supplied to PNACL along with the relevant primer. PNACL carried out cycle sequencing reactions, cleaned them up using DyeEx columns and analysed the products using the 377 automated sequencer (Perkin-Elmer). Data were returned as text files and as traces in SCF format.

2.15 5'RACE

This is a variation on RT-PCR. There have been numerous variations on the original protocol; Schaefer (1995) reviews these.

2.15.1 Reverse transcription

Reverse transcription was performed as described in section 2.11.3 using a gene specific primer and liver polyA⁺ RNA (donated by Dalgleish, University of Leicester).

2.15.2 Purification and Tailing

The cDNA generated from the reverse transcription was purified using the QIAquick PCR purification system (described in section 2.11.4) and resuspended in 50 µl of water. Ten microlitres of this were then used in the tailing reaction.

The cDNA was tailed at the 3' end with dATP using Terminal deoxynucleotidyl Transferase (TdT) (Gibco BRL, Life Technologies Ltd, 15 units.µl⁻¹). A reaction was set up, in a volume of 24 µl, containing 2.5 µl of 2 mM dATP, 5 µl of 5 × tailing buffer (500 mM potassium cacodylate pH 7.2, 10 mM CoCl₂, 1 mM DTT) and 10 µl of cDNA and incubated at 94°C for 3 minutes. The sample was snap cooled on ice and then 1 µl of TdT added. The reaction was then incubated at 37°C for 10 minutes. To inactivate the enzyme, the reaction was then heated at 65°C for 10 minutes and then chilled on ice. The tailed DNA was ethanol precipitated as described in section 2.5 and resuspended in 25 µl of water.

2.15.3 PCR of dA-tailed cDNA

A PCR was performed, as described in section 2.11.1, using a gene specific primer and a dT-TAG primer (5' - GACTCGAGTCGACATCGA(T)₁₇ - 3') and ADAPT primer (5' - GACTCGAGTCGACATCG - 3').

The PCR conditions used were:

94°C 2 minutes [(94°C 30s, 55°C 30s, 72°C 1 min) × 30] 72°C 5 minutes

The product was purified as described in section 2.11.4 and cloned into a suitable vector using standard procedures. It was then sequenced by one of the methods described in section 2.14.

2.16 Primer extension

This method is based on that described by Sambrook *et al.* (1989).

A gene specific primer was end labelled with [γ -³³P] ATP as described in section 2.11.2. A reverse transcription was then carried out on 10 µg of total RNA as described in section 2.11.3 with the PCR on the resulting cDNA being omitted. The products were run on a 6% denaturing polyacrylamide gel as described in section 2.12, alongside a sequencing reaction of cloned genomic DNA using the same primer, as described in section 2.14.1. The results were visualised by autoradiography as described in section 2.10.1.5.

2.17 Bioinformatics

2.17.1 Computing facilities used

- IBM compatible microcomputer running Microsoft Windows NT4 Workstation
- AlphaImager 2000 IBM compatible microcomputer running Windows 95
- SGI origin running IRIX v.6.5
- Dell microcomputer running Microsoft Windows 98
- CanoScan N656U

2.17.2 Software used

- GCG v.9.1, v.10.0 and v.10.1 for IRIX
- Chromas v.1.44
- AlphaEase v.4.0
- Hummingbird eXceed v.6.1 and v.6.2
- Microsoft Office 97
- Freehand v.5.0
- Microsoft Picture It! Express v.2.0
- ArcSoft PhotoStudio 2000
- ScanGear CS-U 5.7
- EndNote v.3.0.1

2.17.3 GCG v.9.1 molecular biology package programs

For sequence comparisons the following programs in the GCG molecular biology package programs were used: Fasta, BLAST, Gap, Pileup, SeqLab and Clustalw.

For mapping sequences with respect to restriction endonuclease recognition sites the programs Map, Mapplot and Mapsort were used. For evolutionary analysis the programs Growtree and Distances were used. The programs Frames and Translate were used to identify open reading frames and translate a nucleotide sequence into a protein sequence respectively.

2.17.4 Primer Design

Primer design was always carried out using the program Primer 3 (Rozen and Skaletsky 1998, unpublished), Primer3. Code available at http://www.genome.wi.mit.edu/genome_software/other/primer3.html).

2.18 Safety Issues

All laboratory work was carried out observing good laboratory practice. Chemicals were handled in accordance with Control of Substances Hazardous to Health (COSHH) safety regulations. All genetic manipulations were carried out in compliance with the Genetically Modified Organisms Regulations and with the approval of the University Safety Office. All manipulations were at containment level 1 and were classified as 1A (group 1 organisms in a type A operation). All blood and tissues were handled and disposed of according to university health and safety regulations.

Chapter 3

The structure of the human and mouse genes that encode the protein secreted phosphoprotein 24

3.1 Introduction

Characterisation of the structure of a gene can provide information about the encoded protein, the expression of the gene and its regulation. It is also important to know the structure of a gene before functional studies are performed, *e.g.* expression of the protein or mouse knockouts.

The original report of spp24 (Hu *et al.* 1995) presented the bovine cDNA sequence and the deduced protein sequence. The structure of the bovine gene was not determined. This chapter presents a detailed analysis of the human gene and the exon/intron structure of the mouse gene. The structure of the human gene is obviously of importance in determining the role of spp24 in human health and disease. The structure of the mouse gene is necessary to enable gene knockouts in mice to be made in the future. It also enables speculation as to whether the gene encoding spp24 is conserved between species, providing evolutionary information.

3.1.1 The human gene encoding secreted phosphoprotein 24

This section begins by describing the work carried out at the University of Leicester prior to the start of this thesis.

The human gene encoding the spp24 protein has been assigned the symbol *SPP2* by the HUGO Gene Nomenclature Committee.

The bovine cDNA sequence (Accession number U03872) was used to search the human EST database (Dalglish, unpublished) and several ESTs were identified. One of these ESTs was used to screen the human male genomic PAC library RPC11 obtained from the HGMP UK Resource Centre (UK HGMP-RC) (Gill and Dalglish, unpublished). The RPC11 library contained approximately 120,000 clones, each containing an insert with an average size of 110 kb in the recombinant P1 vector pCYPAC-2. The screen identified 4 clones that contained most, if not all, of the *SPP2* gene. The clone numbers for the 4 positives were 14 E15, 37 E17, 137 C1 and 318 P19. The first number refers to the microtitre plate and the second number preceded by a letter refers to the location within that plate.

One of the PACs identified by the screen was used to localise the human *SPP2* gene. *SPP2* was assigned to chromosome band 2q37→qter by *in situ* hybridisation (Swallow *et al.* 1998). This confirmed the location of the gene previously mapped by Hudson (1996) using a radiation hybrid panel.

A further 2 PACs were then identified by screening a human chromosome 2 PAC library (Gingrich *et al.* 1996) (Dalglish, unpublished). The two strong positives had the clone numbers 3 N4 and 6 M9, the numbers and letters referring to the plate and well number as with the human genomic RPCI1 PAC library. Only one of these was shown to contain the entire gene.

In an attempt to begin to identify the exon/intron boundaries of the human *SPP2* gene, the four PACs identified in the screening of the male genomic RPCI1 PAC library were used to identify the *Eco*RI fragments of the gene that contained coding sequence (Merrison and Dalglish, unpublished). This was done by Southern blotting and hybridisation with a human *SPP2* EST originally identified by Dalglish (I.M.A.G.E. clone 204242). By using individual segments of the EST probe, the *Eco*RI fragments were ordered in relation to one another. The fragments, however, were not known to be contiguous. The order of the *Eco*RI fragments of the human *SPP2* gene that contain coding sequence is shown in figure 3.1. The sizes were estimated and then each fragment was assigned a label relating to its approximate size.

A preliminary human cDNA sequence determined by alignment of human ESTs (Dalglish, unpublished) was known to contain two *Eco*RI sites. For this reason, each of the five *Eco*RI fragments shown in figure 3.1 was cloned and sequenced from each end (Merrison and Dalglish, unpublished). The human *SPP2* cDNA sequence and the position of the *Eco*RI sites is shown in figure 3.1. One exon/intron boundary was found in this manner. This thesis describes the completion of the sequencing, which is discussed in section 3.2.1.

A further two enzymes, *Kpn*I and *Sph*I, were also shown to have recognition sites in the human *SPP2* cDNA (figure 3.3 for position). Section 3.2.1 of this chapter presents the results of identifying which of the individual *Eco*RI fragments contained the exons harbouring these restriction enzyme sites. The *Eco*RI fragments in question were digested, as appropriate, with either *Kpn*I or *Sph*I and cloned into plasmids. Each fragment was then sequenced from the *Kpn*I or *Sph*I end, which should lead straight into exon sequence and enable sequencing through the exon and into intron beyond, thus identifying an exon/intron boundary. Four of

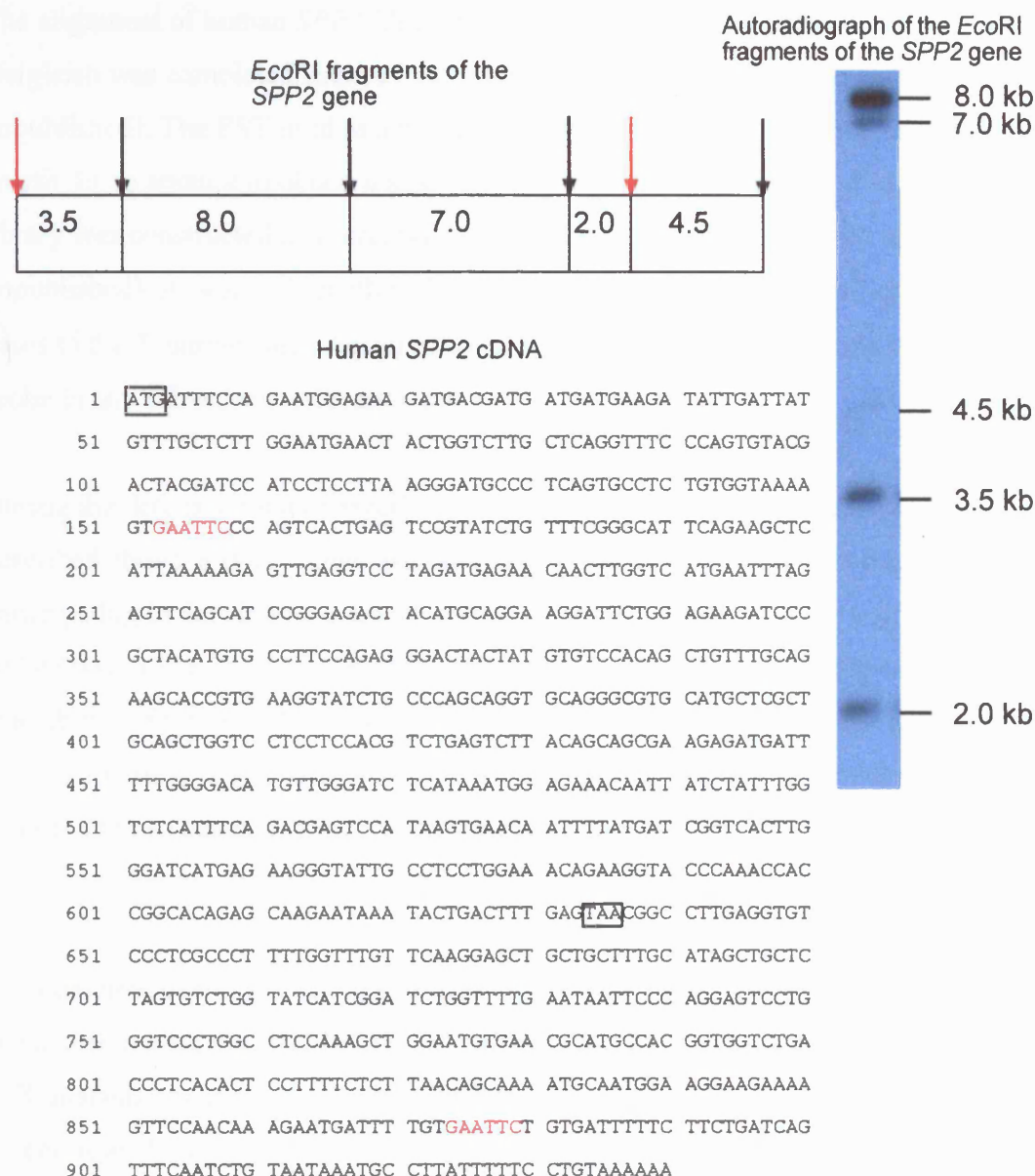


Figure 3.1. The *EcoRI* fragments of the *SPP2* gene that contain coding sequence and the human *SPP2* cDNA sequence (Merrison and Dagleish, unpublished).

This figure shows, as a block representation and on the autoradiograph, the fragments generated when the *SPP2* gene was digested with the restriction endonuclease *EcoRI* (Merrison and Dagleish, unpublished). The numbers in each block give the approximate size of the fragment in kilobases. The arrows indicate the *EcoRI* sites. The red arrows correspond to the *EcoRI* sites that are found in the human cDNA. The fragments are not known to be contiguous. The probe used to generate the autoradiograph ran from the left-most *EcoRI* site to the end of the cDNA at the 3' end. Beneath, the sequence of the human *SPP2* cDNA is given as determined by alignment of human ESTs (Kitchen and Dagleish, unpublished). The 'ATG' start codon and the 'TAA' termination codon are boxed. The *SPP2* *EcoRI* sites that are found in the coding DNA are indicated in red.

the exon/intron boundaries were identified using this cloning/sequencing strategy, which was a continuation of the work begun by Merrison and Dalgleish (unpublished).

The alignment of human *SPP2* ESTs to determine the cDNA sequence that was begun by Dalgleish was completed using a larger number of ESTs (Kitchen and Dalgleish, unpublished). The EST used as a probe in the work discussed above was known not to be full-length. In an attempt to obtain a near full-length human *SPP2* cDNA, a human liver cDNA library was constructed and screened with the incomplete EST (Kitchen and Dalgleish, unpublished). A near full-length cDNA clone was identified that added approximately 100 bases to the 5' untranslated region of the cDNA. This is the cDNA clone that was used as a probe in any subsequent relevant work.

During the determination of exon/intron boundaries using the cloning/sequencing strategy, as described above, a BAC clone containing the *SPP2* gene was sequenced by a group participating in the Human Genome Project. The BAC clone with the accession number AC006037 is 108,711 bp in length and contains the *SPP2* gene in the reverse orientation. The availability of the complete sequence of clone AC006037 meant that the determination of the exon/intron structure of the human *SPP2* could be completed more rapidly and the exons that were found by sequencing the *EcoRI* fragments could be confirmed. This chapter describes the completion of determination of the exon/intron structure of the human *SPP2* gene.

The complete genomic sequence of the human *SPP2* gene meant that the whole region could be analysed using the UK HGMP-RC NIX analysis environment (www.hgmp.mrc.ac.uk). The NIX analysis environment contains the programs shown in table 3.1 with the respective functions as described. A sub-group of the analyses use multiple programs that search for features conforming to a consensus and give a likelihood of that feature being real. In this way, the possibility of extra exons or alternative splicing can be investigated and identification of likely promoter regions, repetitive elements, polyadenylation signals and open reading frames is possible. The NIX analysis environment also contains programs that identify any ESTs or proteins that show homology to the analysed sequence. Using the NIX analysis environment it is therefore possible to perform an extensive sequence analysis using many different programs simultaneously.

The complete genomic sequence of *SPP2* also allowed an analysis of the sequence using the program Tandem Repeats Finder (Benson 1999) to find any tandem repeats lying within or

Table 3.1. The programs used in the HGMP NIX analysis environment (www.hgmp.mrc.ac.uk).

The NIX analysis environment can be ran from the HGMP website (www.hgmp.mrc.ac.uk). NIX is a world wide web tool that enables viewing of the results of running many DNA analysis programs simultaneously. In this way, it is possible to perform an extensive sequence analysis of a segment of DNA and search for features that conform to a consensus and give a likelihood of that feature being real.

¹Grail references : Uberbacher *et al.* (1991), Guan *et al.* (1991a), Uberbacher and Mural (1991), Einstein *et al.* (1991), Mural *et al.* (1991), Guan *et al.* (1991b), Guan *et al.* (1992), Uberbacher *et al.* (1992), Einstein *et al.* (1992), Xu *et al.* (1994a), Xu *et al.* (1994b), Mural *et al.* (1993), Xu *et al.* (1994c), Uberbacher (1994), Shah *et al.* (1995), Matis *et al.* (1996), Uberbacher *et al.* (1995a), Guan and Uberbacher (1996), Xu *et al.* (1995a), Uberbacher *et al.* (1995b), Xu *et al.* (1995b), Xu *et al.* (1995c), Mark *et al.* (1995), Uberbacher (1995), Xu and Uberbacher (1996a), Xu and Uberbacher (1996b), Shah (1996).

Table 3.1. The programs used in the HGMP NIX analysis environment
(www.hgmp.mrc.ac.uk).

Program found in NIX	Function of program	Reference
GRAIL/cpg	Predicts CpG islands	See legend ¹
GRAIL/pollIprom	Predicts promoters	See legend ¹
TSSW/Promoter	Predicts promoters	Solovyev and Salamov (1997)
GENESCAN/Prom	Predicts promoters	Burge and Karlin (1997a) Burge and Karlin (1997b) Burge (1997) Burset and Guigo (1996)
Fgenes/Prom	Predicts promoters	Solovyev (1995) Solovyev and Lawrence (1993)
Fex	Predicts exons	Solovyev <i>et al.</i> (1994a) Solovyev <i>et al.</i> (1994b)
Hexon	Predicts exons	Solovyev <i>et al.</i> (1994a) Solovyev <i>et al.</i> (1994b)
MZEF	Predicts exons	Zhang (1997)
Genemark	Predicts exons	Borodovsky and McIninch (1993)
GRAIL/exons	Predicts exons	See legend ¹
GRAIL/gap 2	Predicts genes	See legend ¹
Genefinder	Predicts genes	Green (unpublished)
FGene	Predicts genes	Solovyev (1995) Solovyev and Lawrence (1993)
GENSCAN	Predicts genes	Burge and Karlin (1997a) Burge and Karlin (1997b) Burge (1997) Burset and Guigo (1996)
FGenes	Predicts genes	Solovyev (1995) Solovyev and Lawrence (1993)
HMMGene	Predicts genes	Krogh (1997)
BLAST/trembl	Blasts against trembl database	Altschul <i>et al.</i> (1990) Altschul <i>et al.</i> (1997)
BLAST/swissprot	Blasts against swissprot database	Altschul <i>et al.</i> (1990) Altschul <i>et al.</i> (1997)
BLAST/EST	Blasts against EST database	Altschul <i>et al.</i> (1990) Altschul <i>et al.</i> (1997)
BLAST/Embl-	Blasts against EMBL database	Altschul <i>et al.</i> (1990) Altschul <i>et al.</i> (1997)
GENSCAN/polya	Predicts polyadenylation signals	Burge and Karlin (1997a) Burge and Karlin (1997b) Burge (1997) Burset and Guigo (1996)
FGenes/polya	Predicts polyadenylation signals	Solovyev (1995) Solovyev and Lawrence (1993)
GRAIL/polya	Predicts polyadenylation signals	See legend ¹
BLAST/gss	Predicts if frameshift errors are likely	Altschul <i>et al.</i> (1990) Altschul <i>et al.</i> (1997)
BLAST/sts	Blasts against STS database	Altschul <i>et al.</i> (1990) Altschul <i>et al.</i> (1997)
BLAST/ecoli	Blasts against <i>E. coli</i> database	Altschul <i>et al.</i> (1990) Altschul <i>et al.</i> (1997)
BLAST/vector	Blasts against vector database	Altschul <i>et al.</i> (1990) Altschul <i>et al.</i> (1997)
RepeatMasker	Predicts repetitive elements	Smit and Green (unpublished)
tRNAscan-RE	Scans for tRNA	Fichant and Burks (1991) Eddy and Durbin (1994) Pavesi <i>et al.</i> (1994) Lowe and Eddy (1997)

near the gene. Tandem repeats may be useful for future association studies in an attempt to link *spp24* with disease states.

From the ESTs and cDNA clones available it was possible to approximate the initiation point of transcription of the human *SPP2* gene. However, the methods used in the making of the majority of cDNA libraries means that the 5' end of the cDNA would never be complete. In an attempt to identify the exact start of transcription, 5'RACE and primer extension were performed and the results are reported in this chapter.

3.1.2 The mouse gene encoding secreted phosphoprotein 24

Following convention, it is proposed that the mouse gene encoding the *spp24* protein be assigned the symbol *Spp2*, though at the time of writing, this is unofficial.

The chapter describes the determination of the mouse *spp24* cDNA sequence from the alignment of mouse ESTs. The use of ESTs in this way enables the generation of a consensus cDNA sequence that is likely to be near full-length, limited only by the completeness of the 5' ends of the cDNAs.

Unfortunately, the complete genomic mouse *Spp2* sequence was not available and so the determination of the exon/intron structure was completed in a slightly different manner to that of the human gene.

A mouse genomic PAC library RPCI21 (Osoegawa *et al.* 2000) containing 254,217 clones with an insert of average size of 137 kb in the vector pPAC4 was screened for *Spp2* (Manship and Dalgleish, unpublished). One of the PAC clones known to contain the *Spp2* gene was then used to make a small-insert library (difficult to predict the average size, but due to the way in which the library was constructed inserts are probably in the region of 50 to 300 bp) (Swallow and Dalgleish, unpublished). This small-insert library provides a source of small genomic fragments from the mouse *Spp2* gene.

The length of an exon in the human *SPP2* gene ranges from 50 - 283 bp. It is probable that the mouse exons are similar in size. It is therefore likely that each small-insert in the library that hybridises to a *Spp2* cDNA probe will provide information regarding at least one exon/intron boundary. If only part of an exon is present, it should be possible to sequence through the exon and into the intron beyond it. Only if the intron is very small and the sequence read very

long will the next exon be found. If a whole exon is present in the insert then it may be possible to sequence through intron into the exon, through the exon, and into intron again, thus providing information about two exon/intron boundaries.

Another source of information of possible exon/intron boundaries is the mouse ESTs. The quality of ESTs is known to be suspect in some cases. ESTs constructed from partially spliced hnRNA or contaminating genomic DNA can result in some intron sequence occasionally being present. This can be readily detected when aligning ESTs.

Yet another source of information are the NCBI sequencing trace archives. The mouse genome is currently being sequenced and raw data are being deposited at the NCBI in the trace archives. It is possible to do a BLAST search against these genomic sequences, which may reveal information regarding the exon/intron boundaries of the mouse *Spp2* gene.

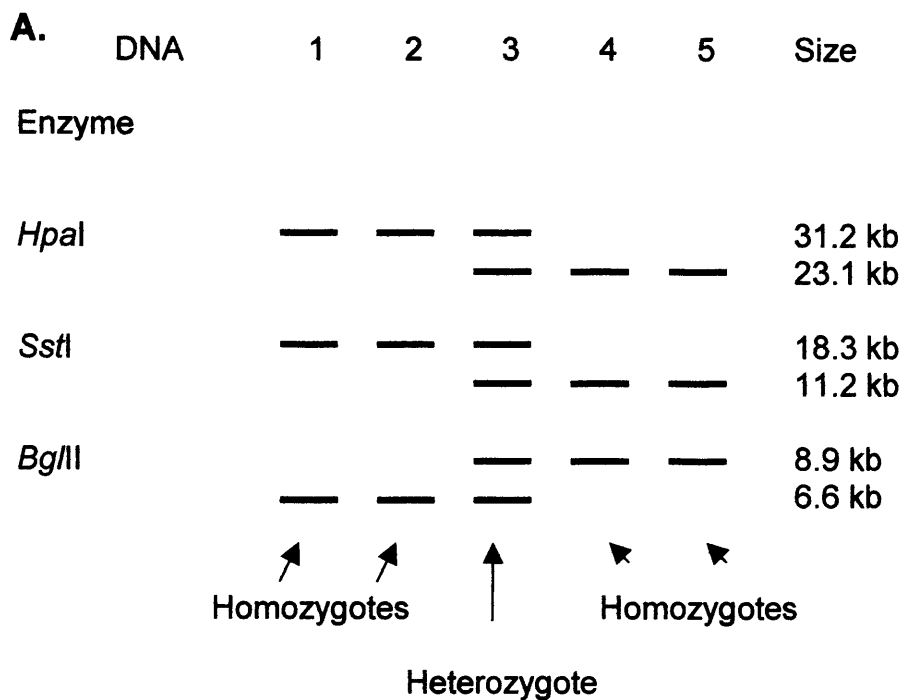
3.1.3 A possible insertion/deletion polymorphism in the human *SPP2* gene

The work described here was carried out by Gill and Dalgleish (unpublished) prior to the beginning of work for this thesis.

Five different placental DNAs were digested with 18 different restriction enzymes. The digested DNAs were then Southern blotted and probed with a human full length *SPP2* EST. Fifteen of the enzymes each gave the same pattern on the autoradiograph, but 3 of the enzymes (*Bgl*II, *Sst*I and *Hpa*I) showed a polymorphism. Figure 3.1A shows the pattern of bands seen on the autoradiographs and the haplotypes observed. The sizes of the alleles are approximate sizes in kilobases that were calculated from the original gel images.

The presence of 3 restriction enzyme dimorphisms means that there are 8 possible haplotypes. However, only the two haplotypes shown in figure 3.1A are seen in the homozygotes, while the haplotypes are unable to be determined in the heterozygote. It was thought that the chance of having 3 RFLPs in extreme disequilibrium or a single mutation that could simultaneously alter all 3 enzyme recognition sites was very small. Therefore, it was proposed that there was an insertion/deletion polymorphism.

The insertion/deletion polymorphism theory was supported by the fact that the difference between the larger and smaller alleles of *Sst*I and *Hpa*I was approximately the same for each enzyme (about 7.6 kb). The smaller allele of *Bgl*II was always associated with the larger allele



B.

Enzyme	Haplotype I	Haplotype II
<i>HpaI</i>	31.2 kb	23.1 kb
<i>SstI</i>	18.3 kb	11.2 kb
<i>BglII</i>	6.6 kb	8.9 kb

Figure 3.1A. The pattern of bands seen on the autoradiographs when 5 different placental DNAs were digested with *BglII*, *SstI* and *HpaI* and the two different haplotypes observed.

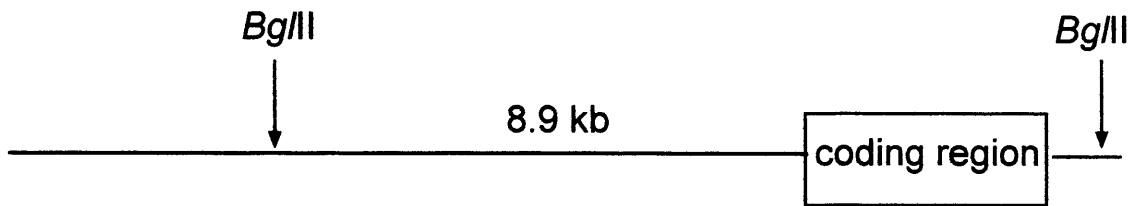
Figure 3.1A A shows the pattern of bands seen on the autoradiograph when different placental DNAs are digested with *BglII*, *SstI* and *HpaI*. The 5 different placental DNAs are labelled 1 to 5, horizontally, and the enzyme relating to each set of bands is indicated on the left hand side, vertically. The approximate size of each band is given on the right hand side in kilobases. The heterozygotes and homozygotes are indicated.

Figure 3.1A B shows the two haplotypes seen between the 5 placental DNAs. However, it is not possible to determine the haplotypes of the heterozygote. The size of each allele is given in kilobases.

of the other two enzymes and so it was thought that there was a *Bgl*III site within the postulated insert. This is shown diagrammatically in figure 3.1B.

This chapter presents the determination of the nature of this polymorphism by cloning and sequencing of the 18.3 kb and 11.2 kb fragments (larger and smaller alleles) generated by *Sst*I digestion of two different PACs containing the human *SPP2* gene. The fragments were cloned into the vector pCL1920 (Lerner and Inouye 1990) using the protocols described in section 2.13, Chapter 2 and sequenced using the method described in section 2.14.1, Chapter 2. These characterised polymorphisms will be a valuable resource for future association studies.

A.



B.

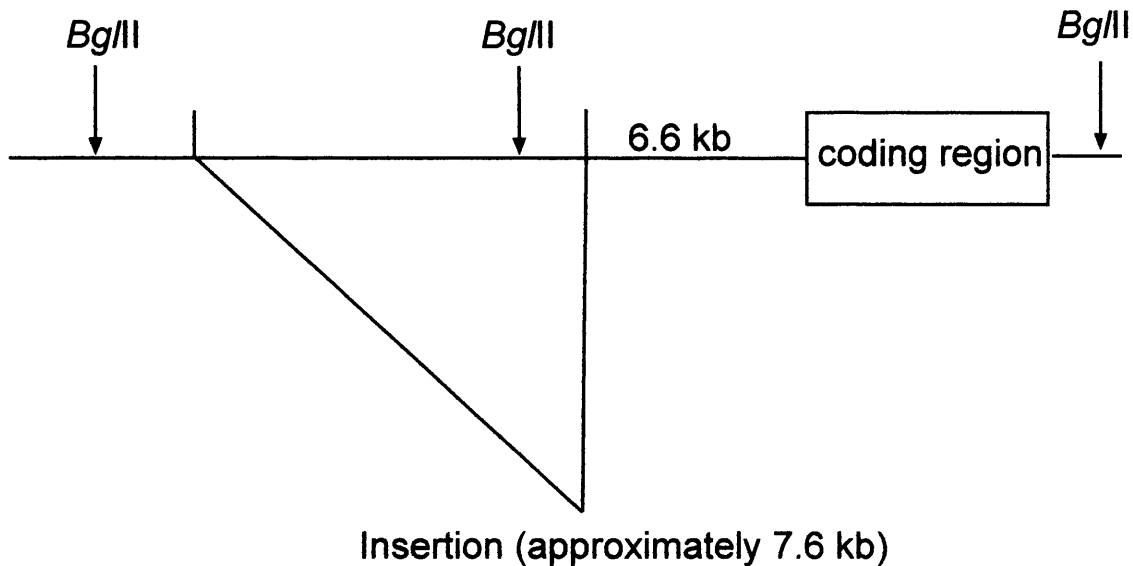


Figure 3.1B. The insertion/deletion polymorphism theory postulated by Gill and Dagleish (unpublished) in relation to the restriction enzyme *Bg*/II.

Figure 3.1B A shows the 8.9 kb fragment generated by digestion of placental DNA with *Bg*/II that would hybridise to a human *SPP2* cDNA probe. This is the allele that does not contain the insert.

Figure 3.1B B shows the allele that does contain the insert. The postulated insert is approximately 7.6 kb and thought to contain a *Bg*/II site. Consequently, the size of the *Bg*/II fragment that will hybridise to the human *SPP2* cDNA probe is reduced to 6.6 kb.

The position of the *Bg*/II sites are indicated by arrows.

3.2 Results

Identification of the exon/intron boundaries and an extensive sequence analysis of the human *SPP2* gene (presented in the following sections) meant that an annotated sequence could be produced of the region of sequence AC006037 that contained the gene. This was submitted to EMBL and allocated the accession number AJ272265. The annotations of AJ272265 are presented in appendix A.

3.2.1 The determination of the exon/intron boundaries in the human *SPP2* gene

All the work described in this section was carried out using the digestion, cloning, Southern blotting and sequencing protocols detailed in sections 2.3, 2.13, 2.10.1 and 2.14 of Chapter 2 respectively.

Merrison and Dalglish identified one exon/intron boundary by sequencing one end of the 3.5 kb *EcoRI* fragment. In this study, sequencing of the other *EcoRI* fragments of the *SPP2* gene that were identified by Merrison and Dalglish (unpublished) resulted in the determination of only one more exon/intron boundary, by sequencing the 2.0 kb *EcoRI* fragment. This meant that the two *EcoRI* sites located in the human *SPP2* cDNA could now be placed on the map of *EcoRI* fragments known to contain coding sequence. This is shown in figure 3.2.

The human *SPP2* cDNA was found to contain a *KpnI* site and a *SphI* site (figure 3.3). The fragments containing coding sequence that are generated from digesting a PAC clone with these enzymes are shown in figure 3.4A and B. To save sequencing every fragment generated from each end, as with *EcoRI*, the localisation of the *KpnI* and *SphI* site in the coding region was attempted.

From the placement of the cDNA against the *EcoRI* fragments (figure 3.2), it was expected that the region of cDNA containing the *KpnI* site lay either in the 2.0 kb or the 7.0 kb *EcoRI* fragment. Each of these *EcoRI* fragments were double digested with *EcoRI* and *KpnI*. Figure 3.4C shows that the 2.0 kb fragment was released intact from the vector (lane 1) indicating that there is no *KpnI* site present. However, the 7.0 kb fragment was released from the vector as two fragments (lane 2), one approximately 3.0 kb and one approximately 4.8 kb. This confirmed the presence of a *KpnI* site in the 7.0 kb *EcoRI* fragment.

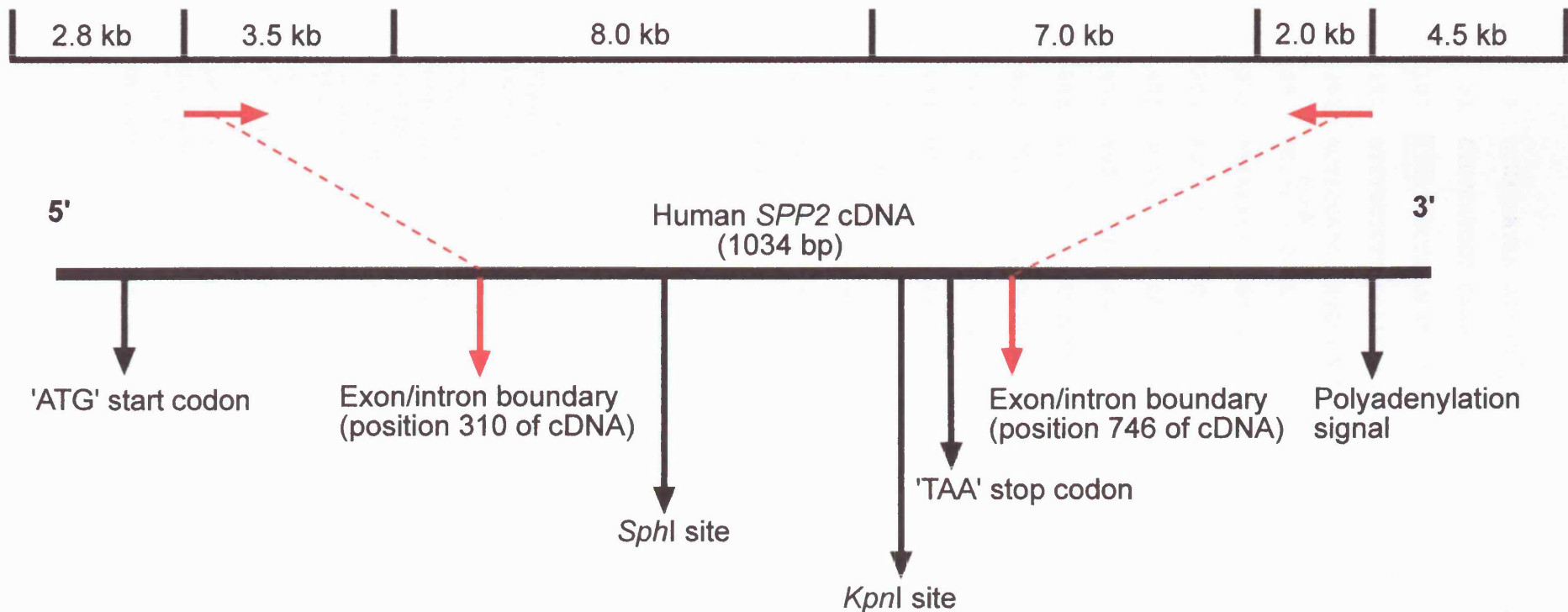


Figure 3.2. The placement of some of the human *SPP2* *EcoRI* fragments against the human *SPP2* cDNA.

This figure shows the *EcoRI* fragments of the human *SPP2* gene as determined by Merrison and Dalgleish (unpublished). Each *EcoRI* fragment was sequenced from each end. This was begun by Merrison and completed in the work reported in this thesis. Merrison and Dalgleish found an exon/intron boundary by sequencing from the end of the 3.5 kb fragment in the forward direction. This is indicated in red on the figure above. This placed the 3.5 kb *EcoRI* fragment at position 310 bp of the cDNA. This thesis reports the sequencing of the 2.0 kb fragment in the reverse direction (also shown in red) to place the 2.0 kb *EcoRI* fragment to the 746 bp region of the cDNA.

This enabled a prediction to be made as to which *EcoRI* fragments would contain the *KpnI* and *SphI* site. It was speculated that the *KpnI* site must be located in either the 7.0 kb or the 2.0 kb *EcoRI* fragment and the *SphI* site in either the 8.0 kb or the 7.0 kb *EcoRI* fragment.

```

1  GTCAAAATAA GCAGCCATG TTTGATAAAG ACAGCTCCTC TTAGGAAGAA
51  CTGTCATCCC CAAACACATA GAGAGACACT CTCTGTCTCT CGATTACATC
101 ATGATTTTCCA GAATGGAGAA GATGACGATG ATGATGAAGA TATTGATTAT
151  GTTTGCTCTT GGAATGAACT ACTGGTCTTG CTCAGGTTTC CCAGTGTACG
201  ACTACGATCC ATCCTCCTTA AGGGATGCCC TCAGTGCCTC TGTGGTAAAA
251  EcoRIGTGAATTCCC AGTCACTGAG TCCGTATCTG TTTCGGGCAT TCAGAAGCTC
301  ATTAAAAAAGA GTTGAGGTCC TAGATGAGAA CAACTTGGTC ATGAATTTAG
351  AGTTCAGCAT CCGGGAGACT ACATGCAGGA AGGATTCTGG AGAAGATCCC
401  GCTACATGTG CCTTCCAGAG GGACTACTAT GTGTCCACAG CTGTTTGCAG
451  AAGCACCGTG AAGGTATCTG CCCAGCAGGT GCAGGGCGTG SphICATGCTCGCT
501  GCAGCTGGTC CTCCTCCACG TCTGAGTCTT ACAGCAGCGA AGAGATGATT
551  TTTGGGGACA TGTTGGGATC TCATAAATGG AGAAACAATT ATCTATTTGG
601  TCTCATTTCA GACGAGTCCA TAAGTGAACA ATTTTATGAT CGGTCACTTG
651  GGATCATGAG AAGGGTATTG CCTCCTGGAA ACAGAAGGTA KpnICCCAAACCAC
701  CGGCACAGAG CAAGAATAAA TACTGACTTT GAGTAACGGC CTTGAGGTGT
751  CCCTCGCCCT TTTGGTTTGT TCAAGGAGCT GCTGCTTTGC ATAGCTGCTC
801  TAGTGTCTGG TATCATCGGA TCTGGTTTTG AATAATTCCC AGGAGTCCTG
851  GGTCCCTGGC CTCCAAAGCT GGAATGTGAA CGCATGCCAC GGTGGTCTGA
901  CCCTCACACT CCTTTTCTCT TAACAGCAAA ATGCAATGGA AGGAAGAAAA
951  GTTCCAACAA AGAATGATTT EcoRITGTGAATTCT GTGATTTTTC TTCTGATCAG
1001 TTTCAATCTG TAATAAATGC CTTATTTTTC CTGT

```

Figure 3.3. The exon/intron boundaries of the human *SPP2* gene as determined by the sequencing strategy described in section 3.1.

The figure shows the human *SPP2* cDNA. The 'ATG' start codon, the 'TAA' stop codon and the polyadenylation signal ('AATAAA') are boxed. Shown in red are the restriction enzyme sites for *EcoRI*, *KpnI* and *SphI* that are located within the cDNA and were used in the cloning/sequencing strategy described in section 3.1.

The exon/intron boundary determined by Merrison and Dalglish (unpublished) is shown in black and the exon/intron boundaries determined by the cloning/sequencing strategy presented in this chapter are shown in green. The exon/intron boundary found by comparing the human *SPP2* cDNA to the genomic clone AC006037 is shown in red. This comparison also confirmed the location of all the other exon/intron boundaries.

The first base of the cDNA is based on the 5'RACE data and the primary transcription initiation site in liver is marked in red and boxed.

Figure 3.4. Digests of PAC clones containing *SPP2* and *EcoRI* fragments of the human *SPP2* gene, with *KpnI* and *SphI*.

A typical pattern observed on the autoradiograph when a PAC clone containing *SPP2* that has been digested with *KpnI* and alternatively *SphI* is hybridised to the human *SPP2* cDNA is shown in A and B respectively. It is not possible to tell from this which fragment contains the *KpnI* or *SphI* site that is in the coding region. However, figure 3.2 shows that the *KpnI* site is likely to be in either the 2.0 kb or 7.0 kb *EcoRI* fragment and the *SphI* site is likely to be in either the 7.0 kb or the 8.0 kb *EcoRI* fragment.

The digestion of each of the appropriate *EcoRI* fragments with either *KpnI* or *SphI* and the corresponding autoradiograph after hybridisation to the human *SPP2* cDNA is shown in C and D respectively. The marker used was λ DNA cut with *HindIII*. The size of each relevant marker band is indicated in kb. It was concluded that the 7.0 kb *EcoRI* fragment contained the *KpnI* site and the 8.0 kb *EcoRI* fragment contained the *SphI* site that is present in coding sequence. Lane 1 contains the 2.0 kb fragment digested with *EcoRI* and *KpnI*, lane 2 contains the 7.0 kb fragment digested with *EcoRI* and *KpnI*, lane 3 contains the 7.0 kb fragment digested with *EcoRI* and *SphI* and lane 4 contains the 8.0 kb kb fragment digested with *EcoRI* and *SphI*.

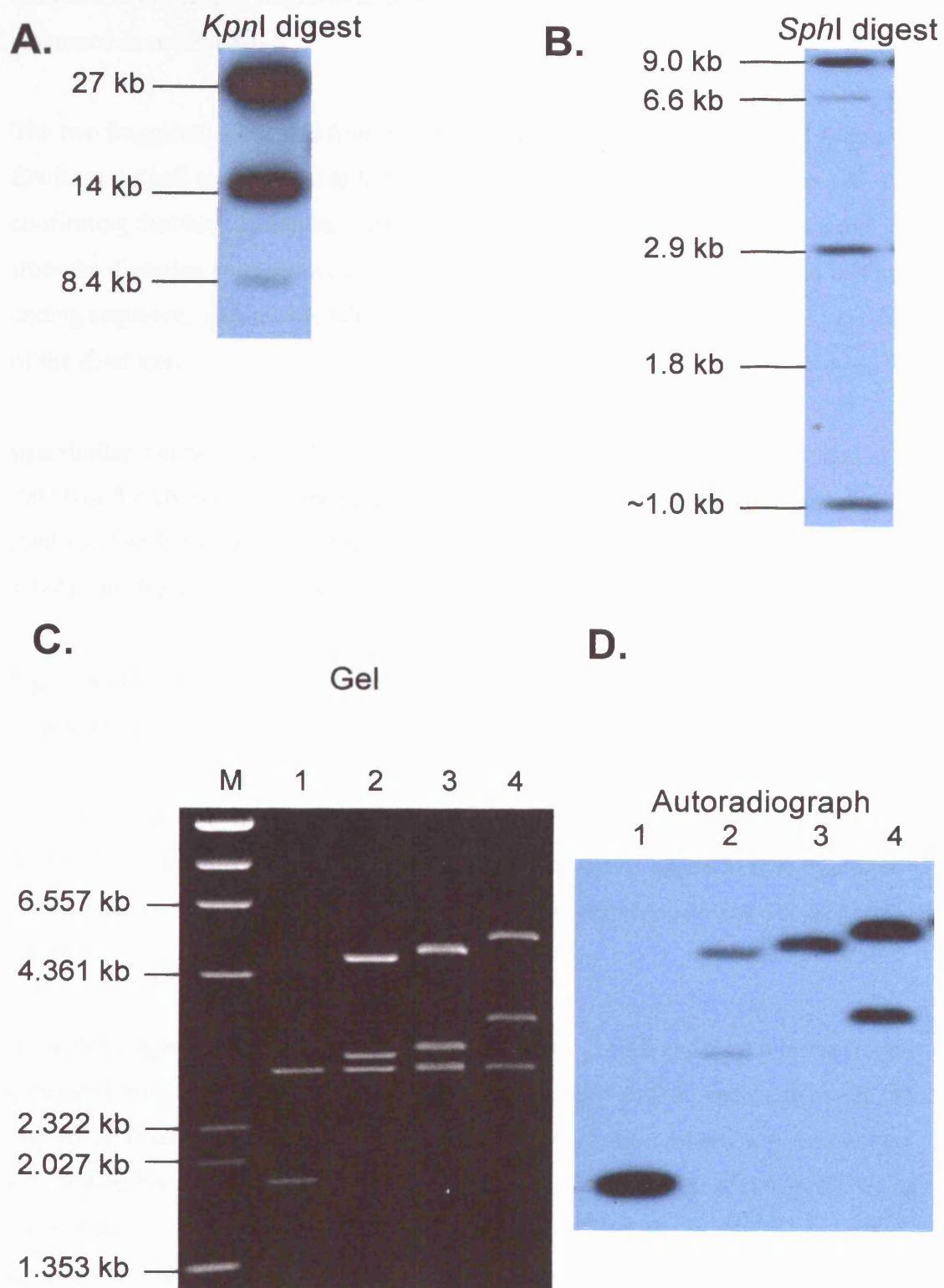


Figure 3.4. Digests of PAC clones containing *SPP2* and *EcoRI* fragments of the human *SPP2* gene, with *KpnI* and *SphI*.

The sizes of the *KpnI* fragments do not add up to exactly 7.0 kb as this was an initial size assigned to this *EcoRI* fragment as a convenient label based on estimates from initial analyses discussed in section 3.1.

The two fragments generated from the double digestion of the 7.0 kb *EcoRI* fragment with *EcoRI* and *KpnI*, were shown to hybridise to the human *SPP2* cDNA (figure 3.4D) confirming that they contained coding sequence. The 3.0 kb and 4.8 kb fragments generated from the digestion were cloned and sequenced from the *KpnI* site deduced to be located in coding sequence. This enabled determination of two exon/intron boundaries, one either side of the *KpnI* site.

In a similar manner, the *SphI* site found in the cDNA was expected to be located in either the 7.0 kb or 8.0 kb *EcoRI* fragment. Double digestion of each of these fragments with *EcoRI* and *SphI* resulted in the release of the insert in two pieces indicating that each fragment contained a *SphI* site (figure 3.4C, lanes 3 and 4).

Figure 3.4D shows the hybridisation of the *EcoRI/SphI* double digestions to the human *SPP2* cDNA. Of the two insert fragments generated from digestion of the 7.0 kb *EcoRI* fragment, only one hybridised to the cDNA. Consequently the *SphI* site that cleaves the insert must be located in an intron and the smaller fragment comprises entirely intron sequence. However, the two insert fragments from digestion of the 8.0 kb *EcoRI* fragment both hybridise to the cDNA indicating that the *SphI* site located in the cDNA is found in the 8.0 kb *EcoRI* fragment.

As with the *KpnI/EcoRI* fragments, the approximately 3.6 kb and 5.2 kb insert fragments generated from digestion of the 8.0 kb *EcoRI* fragment with *SphI* and *EcoRI* were cloned and sequenced from the *SphI* site known to be located in the coding region. A further two exon/intron boundaries were identified in this way, one either side of the *SphI* site. Again the sizes of the *SphI* fragments did not add up to exactly 8.0 kb as this was the estimated size assigned to this fragment as a convenient label in the original analyses discussed in section 3.1.

It was at this point that the complete sequence of the region of chromosome 2 in the vicinity of *SPP2* (accession number AC006037) became available. The identification of exon/intron boundaries was therefore completed by a simple comparison between the human *SPP2* cDNA

and the genomic sequence of AC006037. This was done using the Gap and FASTA programs within the GCG molecular biology package (section 2.21.3, Chapter 2).

Figure 3.3 shows the location of all the exon/intron boundaries in the human *SPP2* cDNA. The boundary identified by Merrison and Dalglish (unpublished) is shown in black. The boundaries shown in green indicate those found by the cloning/sequencing strategy described above. The boundaries shown in red indicate those that were found by alignment of the cDNA with the genomic sequence contained in the clone with accession number AC006037.

The human *SPP2* gene comprises 8 exons and 7 introns. The 'ATG' start codon is located in the first exon and the 'TAA' stop codon is located in the penultimate exon, the final exon containing exclusively 3' untranslated region. The gene spans approximately 26 kb and is shown schematically in figure 3.5.

Table 3.2 shows the sizes of each exon and intron and the sequence found at the boundaries. All boundaries show the consensus gt/ag sequences although not all junctions conform exactly to the consensus of 'GTRAGT' and 'YYTTYYYYYNCAG' for the donor and acceptor sites respectively (Senepathy *et al.* 1990). Junctions which are not identical to the consensus sequences are very similar to the consensus. The start of the first exon is defined by the primary transcription start site seen in primer extension performed on human liver (section 3.2.3) and not the 5'RACE (section 3.2.3) or the longest clone identified by the screening of the liver cDNA library (Kitchen and Dalglish, unpublished, section 3.1.1).

During the determination of exon/intron boundaries using the cloning/sequencing strategy, a trinucleotide tandem repeat was found that lay in intron 7 of the *SPP2* gene (figure 3.5). This was a 'GTT' repeat that in the sequenced sample (a sub-cloned fragment of a PAC clone) was repeated 8 times. This tandem repeat will be discussed further in Chapter 5.

3.2.2 An extensive sequence analysis of the human *SPP2* gene

The complete AC006037 sequence, which contained the human *SPP2* gene, was analysed using the NIX analysis environment at the HGMP website (www.hgmp.mrc.ac.uk).

The NIX analysis environment, which is described in section 3.1.1, enabled an analysis to be carried out using many programs simultaneously. The NIX results for AC006037 are shown in figure 3.6.

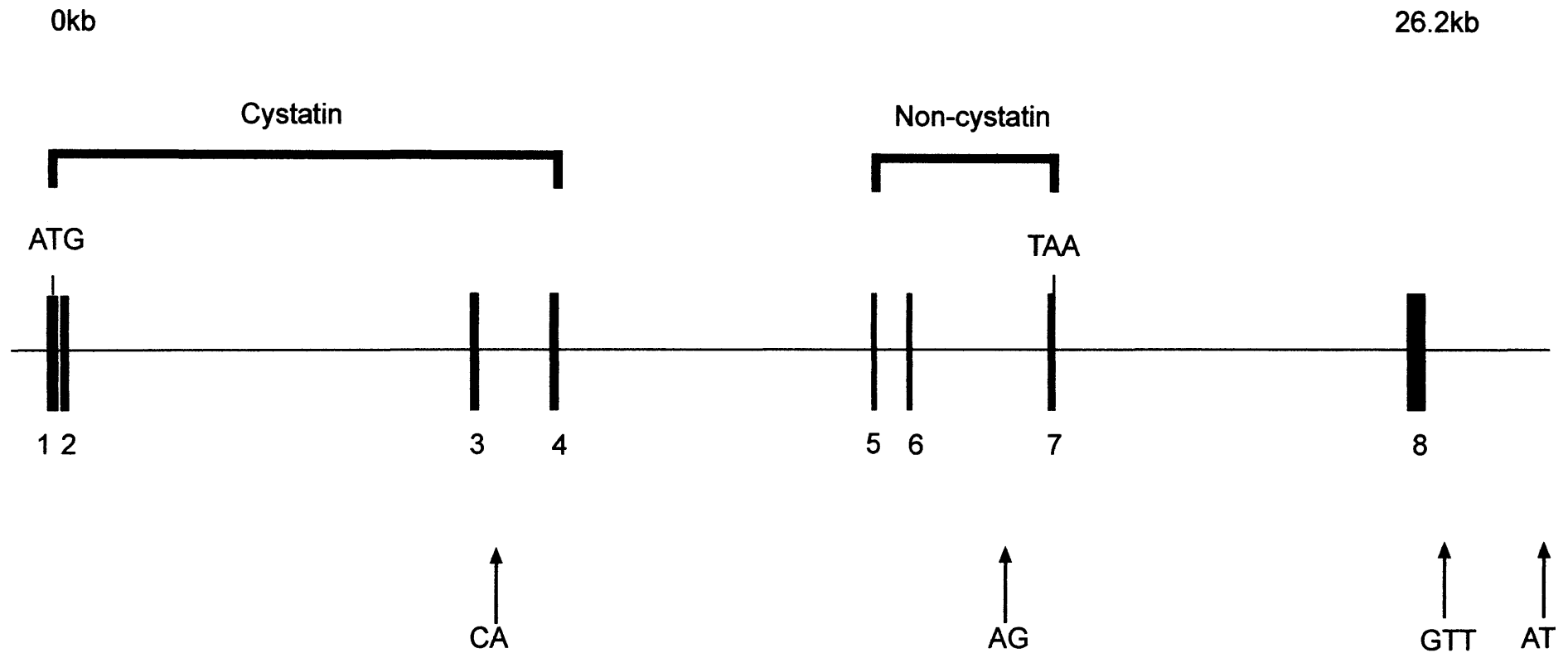


Figure 3.5. The exon/intron structure of the human *SPP2* gene.

This figure shows the 26.2 kb region of the genomic DNA segment (accession number AC006037) that contains the human *SPP2* gene. The figure is drawn approximately to scale with the exons labelled 1 to 8. The 'ATG' start codon in exon 1 and the 'TAA' stop codon in exon 7 are indicated. The exons that encode the cystatin-like region of the protein and the exons that encode the non-cystatin-like region of the protein are marked. Also shown on this figure are the positions of the tandem repeats found in the human *SPP2* gene (section 3.2.2).

Table 3.2. The exon/intron boundaries of the human *SPP2* gene.

The exon/intron boundaries of the human *SPP2* gene were determined either by the sequencing strategy outlined in section 3.1.1 or by comparing the human cDNA with the genomic sequence (accession number AC006037). This table shows the sizes of each exon and intron in base pairs as calculated from these results and the sequence found at each of the boundaries. Exon sequence is denoted by upper case letters and intron sequence by lower case letters. The consensus gt/ag sequences are shown in bold. A '*' indicates that this junction conforms exactly to the consensus of 'GTRAGT' and 'YYTTYYYYYYNCAG' for the donor and acceptor sites respectively (Senepathy *et al.* 1990). All other junctions are similar to these consensus sequences, but not identical.

The beginning of the first exon is defined by the primary start of transcription seen in the liver from the primer extension results (see section 3.2.3), not by the longest clone obtained in the 5'RACE or the screening of the liver cDNA library.

Exon	Position in cDNA	Size of exon	Size of intron	Sequence
1	17-185	170 bp	99 bp	AGT...CAG g taagg
2	186-310	125 bp	7740 bp	gcttgcggtgtcc ag GTT...AG A gtaagt *
3	311-433	123 bp	1410 bp	ttatTTTTgtga ag GTT...GT G gtaagt *
4	434-544	111 bp	6053 bp	tttgtcttttcc ag TCC...G A Ggtaatga
5	545-599	55 bp	636 bp	ctgaatttcttt ag ATG...TT G gtaagt *
6	600-650	51 bp	2653 bp	ttactgtgttac ag GTC...TT G gtaagt *
7	651-746	96 bp	6821 bp	ttttctatcttt ag GGA...G A Ggtaaga
8	747-1034	288 bp	-	* tcttcctcctgc ag GTG...TGT

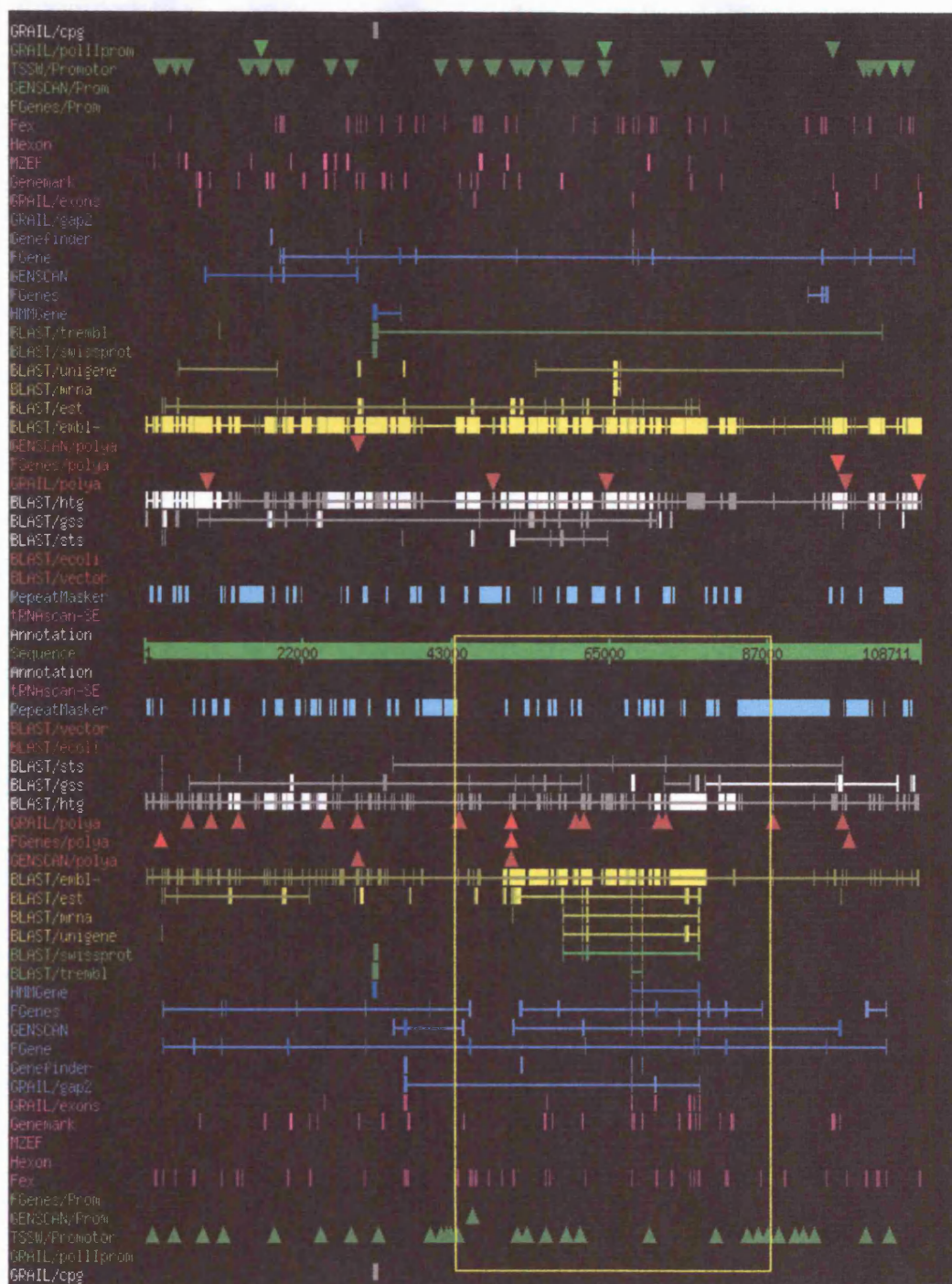


Figure 3.6. NIX analysis of AC006037 that contains the human *SPP2* gene.

The complete sequence of AC006037 was analysed using the HGMP NIX analysis environment (www.hgmp.mrc.ac.uk). The programs within the NIX analysis environment are described in table 3.1. The results of each program are shown graphically above. The top half of the figure shows the analysis of the forward strand and the lower half shows the analysis of the reverse strand.

The yellow box defines the region that contains the human *SPP2* gene.

The human *SPP2* gene is located between approximately positions 43,000 and 87,000 of AC006037, in the reverse orientation. No genes were found either side of *SPP2* and so the analysis was concentrated to the region of the *SPP2* gene.

The programs GAIL/gap2, Genefinder, FGene, GENSCAN, FGenes and HMMGene that are found in the NIX analysis environment, all predict a gene in the region of sequence occupied by *SPP2*.

The programs FEX, HEXON, MZEF, GENEMARK and GAIL/exons all predict the locations of exons. As expected, the predicted exons from each of these programs all clustered around the location of the exons as determined by the results presented in section 3.2.1. There was no obvious clustering in any other region, which would suggest that there are no unexpected exons.

The programs GAIL/poliIprom, TSSW/Promoter, GENSCAN/Prom and Fgenes/Prom all predict the location of any possible promoters. Unfortunately there is no clustering of predictions from these programs around the expected start of transcription of the *SPP2* gene. This suggests that *SPP2* has an unconventional promoter. A further promoter analysis is discussed in section 3.2.3 following the determination of the start of transcription from primer extension and 5'RACE analyses. The program GAIL/CpG did not predict any CpG islands preceding the *SPP2* gene.

The human *SPP2* gene has an obvious polyadenylation signal that is predicted by all three of the programs GENSCAN/polya, Fgenes/polya and GAIL/polya. This signal corresponds to the 'AATAAA' sequence seen at positions 1012 to 1017 of the human *SPP2* cDNA (figure 3.3).

As expected, BLAST searches revealed no homology to any vector or *E. coli* DNA. BLAST searches against the EMBL, EST, mRNA, UniGene, Swissprot and TREMBL databases showed homology to the expected spp24 sequences from either human, mouse, rat or bovine. The only non-spp24 protein that showed any significant homology was a hypothetical chick protein (accession number Q91982). This will be discussed further as a protein showing homology to spp24 in Chapter 6.

The program RepeatMasker in the NIX analysis environment predicted the location of many interspersed repetitive elements within the region of the *SPP2* gene. There are four distinct

families of interspersed repeats: SINEs, LINEs, LTR elements and DNA elements. Repeats from all of these families were found in both orientations within the *SPP2* gene. However, none were found in coding sequence. Details of the location of these interspersed repetitive elements with respect to the annotated sequence AJ272265 can be found in appendix A.

The *SPP2* gene was also searched for tandem repeats using the program TandemRepeatFinder (Benson 1999). This program identified a 'CA', an 'AG' and an 'AT' tandem repeat that are shown in relation to the *SPP2* gene in figure 3.5, along with the 'GTT' repeat found during sequencing of the *EcoRI* fragments. All the repeats lie in intron sequence. The 'CA' repeat lies in intron 3, the 'AG' repeat in intron 6 and the 'AT' and 'GTT' repeats lie just 3' of the gene. A preliminary investigation was carried out on these repeats (results not shown) and all were found to be polymorphic. These tandem repeats may therefore be useful in any future association studies.

3.2.3 The determination of the start of transcription in the human *SPP2* gene

5'RACE was performed on 1 µg of human liver poly A⁺ RNA (donated by Raymond Dalglish) as described in section 2.15, Chapter 2. The two gene specific primers (GSP1 and GSP2) had the following sequences:

GSP1 5' - GTTGTTCTCATCTAGGAC - 3'

GSP2 5' - CGAAACAGATACGGACTCAG - 3'

The location of these primers in the cDNA is shown in figure 3.7.

GSP1 was used to reverse transcribe liver poly A⁺ RNA and GSP2 was used along with the Anchor and Adapt primers to perform the PCR of the dA-tailed cDNA.

The Anchor and Adapt primers both contained a *XhoI* site and so the 5'RACE products were cloned using *XhoI* and *EcoRI* (located in the 5' region of the cDNA figure 3.1) into the vector pGEM-7Zf (Promega).

Figure 3.7, lane 1, shows the 5'RACE products before cloning. The product is heterogeneous in length due to the inability to control the number of residues that are added during the 'tailing' procedure as well as due to the efficiency of the reverse transcriptase. After cloning, the clone with the longest insert was sequenced as described in section 2.14.2, Chapter 2, and

A.

GTCAAATAAGCAGCCAGTGTGTTGATAAAGACAGCTCCTCTTAGGAAGAA
 CTGTCATCCCCAACACATAGAGAGACACTCTCTGTCTCTCGATTACAAT
 CATGATTTCCAGAATGGAGAAGATGACGATGATGATGAAGATATTGATTA
 TGTGCTCTTGGAATGAACTACTGGTCTTGCTCAGGTTTCCCAGTGTAC
 GACTACGATCCATCCTCCTTAAGGGATGCCCTCAGTGCCTCTGTGGTAAA
 AGTGAATTCCCAGTCACTGAGTCCGTATCTGTTTCGGGCATTGAGAAGCT
 GACTCAGGCATAGACAAAGC

GSP2

CATTAAAAGAGTTGAGGTCCTAGATGAGAACAACCTGGTCATGAATTTA
 CAGGATCTACTCTTGTTG

GSP1

GAGTTCAGCATCCGGGAGACTACATGCAGGAAGGATTCTGGAGAAGATCCCCG
 CTACATGTGCCTTCCAGAGGGACTACTATGTG

B.

Anchor primer

5'-GACTCGAGTCGACATCGAT₁₇-3'

Adapt primer

5'-GACTCGAGTCGACATCG-3'

C.

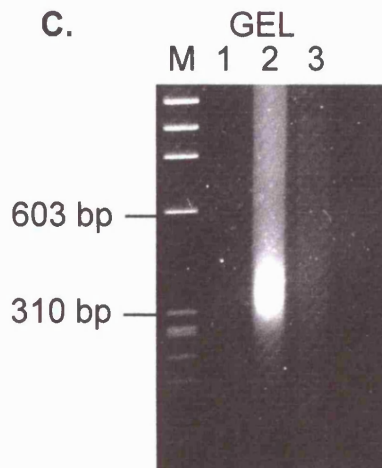


Figure 3.7. The location of the primers used in the 5'RACE and the product generated.

The location in the human *SPP2* cDNA of the gene specific primers is shown in A. Exon 1 is shown in red, exon 2 in green and exon 3 in blue. Boundaries are marked by a vertical line. The *EcoRI* site is underlined in the cDNA sequence.

The sequence of the Anchor and Adapt primers is shown in B. The *XhoI* site is underlined in these primers.

The gene specific primer 1 (GSP1) was used to reverse transcribe 1 µg of human liver polyA⁺ RNA. The resulting cDNA was then tailed at the 5' end with (A)_n. A PCR was then carried out using the gene specific primer 2 (GSP2) and the anchor and adapt primer. The gel image in C shows the 5'RACE product. Lane 2 is the 5'RACE product, lane 1 is the no terminal deoxynucleotidyl transferase (TdT) control and lane 3 is the PCR with water negative control. The sizes of the relevant marker bands are shown in base pairs. The marker is φX174 RF cut with *HaeIII*. The product was then cloned using the restriction enzyme sites *XhoI* and *EcoRI* and sequenced. The start of the sequence after the poly A tail at the 5' end determines the start of transcription. Exon 1 in the figure above starts at this position.

this is what is depicted in figure 3.7. However, this may only have been the longest due to the length of the tail added and not the cDNA itself. With hindsight, a selection of clones should have been sequenced. The clone that was sequenced gave exactly the same sequence as the longest clone obtained from the screening of the human liver cDNA library (Kitchen and Dalglish, unpublished, section 3.1.1). This suggests that the start of the 5'RACE clone may not represent the 5' most start of transcription as the cDNAs in the human liver cDNA library would be expected to be missing a small number of bases at the 5' end, due to the library being constructed by a method based upon Gubler and Hoffman (Gubler and Hoffman 1983).

In another attempt to determine the transcription initiation site, primer extension was performed on 10 µg of total RNA from liver and kidney with a gene specific primer (5' - GAGAGTGTCTCTCTATGTG - 3') using the method described in section 2.16, Chapter 2. A manual sequencing reaction was carried out on genomic DNA using the same gene specific primer. This sequencing reaction was then run alongside of the primer extension products on a polyacrylamide gel to identify the exact transcription initiation site.

Figure 3.8 shows the autoradiograph of the primer extension results and the corresponding positions in the human *SPP2* cDNA sequence. The position in the cDNA of the primer used is also indicated.

In both liver and kidney there are multiple start sites for transcription. However, the most frequent start site is different in both tissues. In liver there is a single primary transcription initiation site with several prominent sites giving rise to smaller transcripts and then many faint sites that give rise to larger transcripts. These larger transcripts must have yielded the 5'RACE and liver cDNA library clones. In kidney, the faint sites giving rise to larger transcripts are the same as in liver, but the primary transcription initiation sites in kidney are seen as a doublet band giving rise to two transcripts, one base different in size, that are both smaller than the primary transcript seen in liver. This suggests that the transcription initiation sites are tissue specific.

3.2.4 The determination of the mouse *Spp2* cDNA sequence

All of the work presented here was performed using the programs BLAST, Frames, Pileup and SeqLab in the GCG molecular biology package (section 2.21.3, Chapter 2).

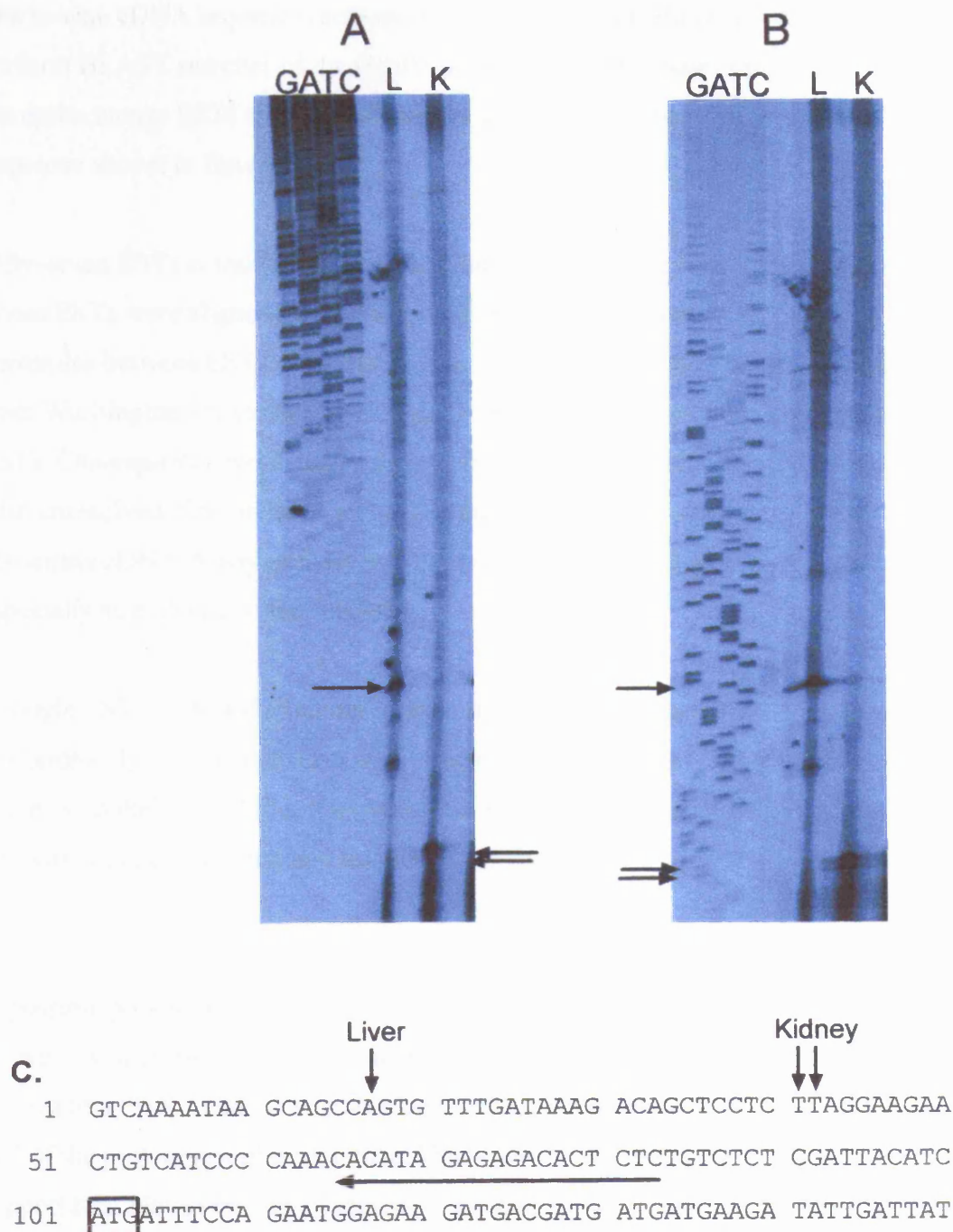


Figure 3.8. The results of primer extension carried out on human total RNA from liver and kidney.

Primer extension was carried out on 10 µg of total RNA from liver and kidney. The primer used was complementary to the sequence underlined in the cDNA sequence above.

The autoradiograph result of the primer extension after one week exposure are shown in A. The same primer that was used in the primer extension was used to carry out a sequencing reaction on cloned genomic DNA. This was loaded 'GATC' as indicated in A and B above. Lane L contains the liver sample and lane K the kidney. A repeat of the primer extension to confirm the result is shown in B. This is the same length exposure, but the sequencing reaction was a little more successful. The arrows indicate the primary transcription initiation sites in each tissue, most of which result in longer transcripts than the primary ones.

The bovine cDNA sequence (accession number U03872) (Hu *et al.* 1995) was used to perform BLAST searches of the GenBank mouse EST database. Table 3.3 gives the details of the entire mouse ESTs that were aligned to generate the consensus mouse *Spp2* cDNA sequence shown in figure 3.9.

Fifty-seven ESTs in total were identified that showed strong homology to the bovine cDNA. These ESTs were aligned using Pileup and then viewed and manually edited in SeqLab. All anomalies between ESTs were checked on the original sequence chromatograms, if available, from Washington University. The original chromatograms were not available for 14 of the 57 ESTs. Consequently anomalies in these sequences could not be checked and so sequences with unresolved differences to all of the others were not included in the generation of the consensus cDNA. Many of these sequences looked as though they were probably poor quality especially at each end of the sequence.

A single EST (AII874457) appeared to be missing the last part of exon 2 and most of exon 3. Unfortunately the original sequence chromatogram was not available to verify this. However, it seems unlikely that this is relevant as it is not a whole exon missing in which case it could be a case of exon skipping and also this phenomenon is seen in only 1 EST out of a total of 57.

A possible polymorphism was seen in ESTs AA839483, AI606606 and AI666747, which are all regions of sequence from the same clone. In these ESTs a 'TTC' codon was present instead of a 'TGC' codon. This would result in a change from a cysteine residue at position 105 of the protein to a phenylalanine. The sequence chromatograms when examined fully support this. However, this cysteine is a residue that is conserved between species (Chapter 6) and the change to a phenylalanine is not conservative. This makes the change seem unlikely and it is therefore probably not a true polymorphism. Also, there is only 1 clone in which this change is observed and so it is likely that this is simply a cloning artefact. A further 34 ESTs cover this region and all have a 'TGC' codon.

A second possible polymorphism was seen in ESTs AI6477723, AI790304 and AI788418. Again, all three ESTs are regions of sequence from the same clone. In these ESTs there is a 'ATG' codon present instead of a 'GTG' codon. This results in a change from a valine residue at position 109 of the protein to a methionine residue. The valine residue is highly conserved between species (Chapter 6) and so this polymorphism is also likely to simply be a cloning

Table 3.3. Mouse ESTs that were aligned to generate the consensus mouse *Spp2* cDNA sequence.

BLAST searches were performed on the GenBank mouse EST database using the bovine *spp24* cDNA sequence (accession number U03872) (Hu *et al.* 1995) as the query. A total of 57 ESTs were identified that showed significant homology. These were aligned to generate a consensus mouse *Spp2* cDNA. The GenBank accession number and the clone ID are given in the table. The source of the sequence is also given with 'W' representing Washington University School of Medicine, 'R' RIKEN (The Institute of Physical and Chemical Research) and 'D' the National Institute of Dental and Craniofacial Research. Most of these ESTs also featured in the UniGene EST cluster Mm.28247 and the TIGR Mouse Gene Index. A '*' indicates that the original sequence chromatogram was not available for that particular EST.

EST AI874457 is the EST that was missing the last part of exon 2 and most of exon 3.

Table 3.3. Mouse ESTs that were aligned to generate the consensus mouse *Spp2* cDNA sequence.

Accession number	Clone number	Sequence Source
AA209723	I.M.A.G.E. 676230	W
AI790304	I.M.A.G.E. 1973222	W
AI647723	I.M.A.G.E. 1971799	W
AA105900	I.M.A.G.E. 518426	W
AA208032	I.M.A.G.E. 662565	W
W41515	I.M.A.G.E. 351238	W
AA080459	I.M.A.G.E. 551108	W
W54491	I.M.A.G.E. 367893	W
AI043198	I.M.A.G.E. 1432233	W
AI876578	I.M.A.G.E. 1924341	W *
W66809	I.M.A.G.E. 387995	W
AA104930	I.M.A.G.E. 533453	W
AI048990	I.M.A.G.E. 1432414	W
AI874457	I.M.A.G.E. 2099494	W *
AA073515	I.M.A.G.E. 535038	W
AA208557	I.M.A.G.E. 662914	W
AA538031	I.M.A.G.E. 931063	W
AI049368	I.M.A.G.E. 1432285	W
AI042913	I.M.A.G.E. 1432414	W
AI666747	I.M.A.G.E. 1433655	W *
W96864	I.M.A.G.E. 422021	W
AA237477	I.M.A.G.E. 680711	W
AA107066	I.M.A.G.E. 519319	W
AA238068	I.M.A.G.E. 680589	W
AA208639	I.M.A.G.E. 661916	W
AI788418	I.M.A.G.E. 1973222	W
AW259020	I.M.A.G.E. 2301131	W
AI790944	I.M.A.G.E. 1970455	W
W17979	I.M.A.G.E. 335916	W
AA241152	I.M.A.G.E. 656221	W
AA286155	I.M.A.G.E. 732888	W
AA689117	I.M.A.G.E. 1105901	W
AI606606	I.M.A.G.E. 1433655	W *
AV037245	1600023D11	R *
AA110916	I.M.A.G.E. 520321	W
AI530291	I.M.A.G.E. 1889780	W
AV015392	1110059F13	R *
AW260313	I.M.A.G.E. 2301133	W
AA254931	I.M.A.G.E. 720110	W
W63973	I.M.A.G.E. 374204	W
W62618	I.M.A.G.E. 372600	W
AV004359	0610041K09	R *
AA268111	I.M.A.G.E. 733821	W
AK002814	0610038O04	R *
AV004017	0610038O04	R *
BB561875	0600002H21	R *
AV014362	1110054F24	R *
AA241994	I.M.A.G.E. 680768	W
AW681812	EGW1240	D *
AV038666	1600030N18	R *
AA209672	I.M.A.G.E. 676445	W
AV014095	1110051P06	R *
AV104788	2510009G14	R *
AI874906	I.M.A.G.E. 2099494	W *
AA104932	I.M.A.G.E. 533472	W
BB502892	D630041B16	R *
AA839483	I.M.A.G.E. 1433655	W

```

1  ACAAGAATAA GACAGCCACC CTCTGAAAGA GCTGTCATCC AGAAGCCTGG
51  AGAGAGGCCG TCTCCCTGAC TCTGGGTCGC CATCCTCTCA GTATGGAGCA
101  GGCAATGCTG AAGACGCTGG CTTTGTGTTG GCTGGGCATG CACTACTGGT
151  GTGCCACAGG TTTCCCGGTG TACGACTACG ACCCTTCCTC TCTGCAGGAA
201  GCTCTCAGTG CCTCAGTGGC AAAGGTGAAC TCGCAGTCCC TGAGTCCTTA
251  CCTGTTTCGG GCGACCCGGA GCTCCTTGAA GAGAGTCAAC GTCCTGGATG
301  AAGACACATT GGTCATGAAC TTAGAGTTCA GTGTTTCAGGA AACCACATGC
351  CTGAGAGATT CTGGTGATCC CTCCACCTGT GCCTTCCAAA GGGGCTACTC
401  TGTGCCAACA GCTGCTTGCA GGAGCACTGT GCAGATGTCC AAGGGACAGG
451  TAAAGGATGT GTGGGCTCAC TGCCGCTGGG CGTCCTCATC TGAGTCCAAC
501  AGCAGTGAGG AGATGATGTT TGGGGACATG GCAAGATCCC ACAGACGAAG
551  AAATGATTAT CTACTTGTTT TTCTTTCTGA TGAATCCAGA AGTGAACAAT
601  TCCGTGACCG GTCAC TTGAA ATCATGAGGA GGGGACAGCC TCCCGCCCAT
651  AGAAGGTTCC TGAACCTCCA TCGCAGAGCA AGAGTAAATT CTGGCTTTGA
701  GTGACATCCT GGAGATTTCA TGAAAGAAAG AGAAGCAGAA GCTGAAATGA
751  AGAAAGGCAT GGAGAATGGT GTCTTTTTTC TTTTATAAT CTCCACTCTG
801  CAATAAAGAT CTTTCCCTTC CTTT

```

Figure 3.9. The consensus mouse *Spp2* cDNA determined by the alignment of mouse ESTs.

Fifty-seven ESTs were identified that showed strong homology to the bovine *spp24* cDNA (accession number U03872). These ESTs were aligned using the program Pileup and then manually edited in the program SeqLab, both programs being part of the GCG Molecular Biology package.

The figure shows the consensus cDNA sequence generated by the 'pretty' function within SeqLab. The 'ATG' start codon and 'TGA' stop codon are boxed in black. The start and stop codons of the second ORF that is predicted are boxed in red.

This sequence has been submitted to EMBL and has the accession number AJ315513.

artefact. Only 1 clone shows this change and a further 32 ESTs cover this region and all have a 'GTG' codon.

Within the SeqLab program, the 'pretty' function was run, which generates a consensus sequence from an alignment. The consensus sequence was then run through the program 'Frames', which shows all possible open reading frames (ORFs) in both directions. The longest ORF when translated shows very high homology to the bovine and human protein. The 'ATG' start codon is at position 105 to 107 of the cDNA and the 'TGA' stop codon is at position 702 to 704.

A second ORF in the same reading frame is seen at position 720 to 806. A second ORF is also seen in the human (accession number AJ308099), bovine (U03872), rat (accession number U19485) and chicken (Chapter 6) cDNA. However, the length of the ORF and the sequence of the translated protein are different in every species. This could mean that it is unlikely to be significant or it could mean that there is a second short peptide important to the function of spp24 that is species specific. The longest ORF followed by the shorter second ORF for human, cattle, mouse, rat and chicken can be seen in figure 3.10. The 'ATG' codons of the second ORFs do not lie in a typical Kozak sequence. However, neither do the 'ATG' codons of the first longest ORFs.

The mouse *Spp2* cDNA sequence has been submitted to EMBL and has the accession number AJ315513.

3.2.5 The determination of the exon/intron boundaries of the mouse *Spp2* gene

The first approach that was taken to identify the exon/intron boundaries in the mouse *Spp2* genes involved the screening of a small insert library.

A small insert genomic library was constructed from a PAC clone containing the *Spp2* gene (Swallow and Dalglish, unpublished, described in section 3.1.2). The small size of the insert meant that the sequencing of any insert containing *Spp2* exons is likely to provide information regarding at least one exon/intron boundary.

The library had been stored as *E. coli* cultures containing individual clones in 96-well microtitre plates. The library was transferred onto a nylon membrane using the method described in section 2.10.2, Chapter 2. The library was then screened with the ³²P-labelled

Figure 3.10. The open reading frames of the human, bovine, mouse, rat, pig and chicken cDNA.

The cDNA sequences of the human (based on longest clone isolated by Kitchen and Dagleish, unpublished), cattle (based on U03872), mouse (based on cDNA presented in this chapter), rat (based on U19485), pig (based on AJ308100) and chicken (based on ammended cDNA sequence in Chapter 6) genes were put into the program 'Frames' in the GCG Molecular biology package (section 2.21.3, Chapter 2). The longest open reading frame (ORF) was identified. In each case a second smaller ORF in the same reading frame was also seen. The ORFs are depicted by boxes. The longest ORF of the rat cDNA does not have a sealed end as the sequence did not include the signal peptide and hence it does not include the 'ATG' start codon. The protein encoded by the longest ORF showed high homology between each species and is the protein known as spp24. The protein encoded by the smaller ORF differed in length, location and composition between each species. It is not known whether this is significant to spp24.

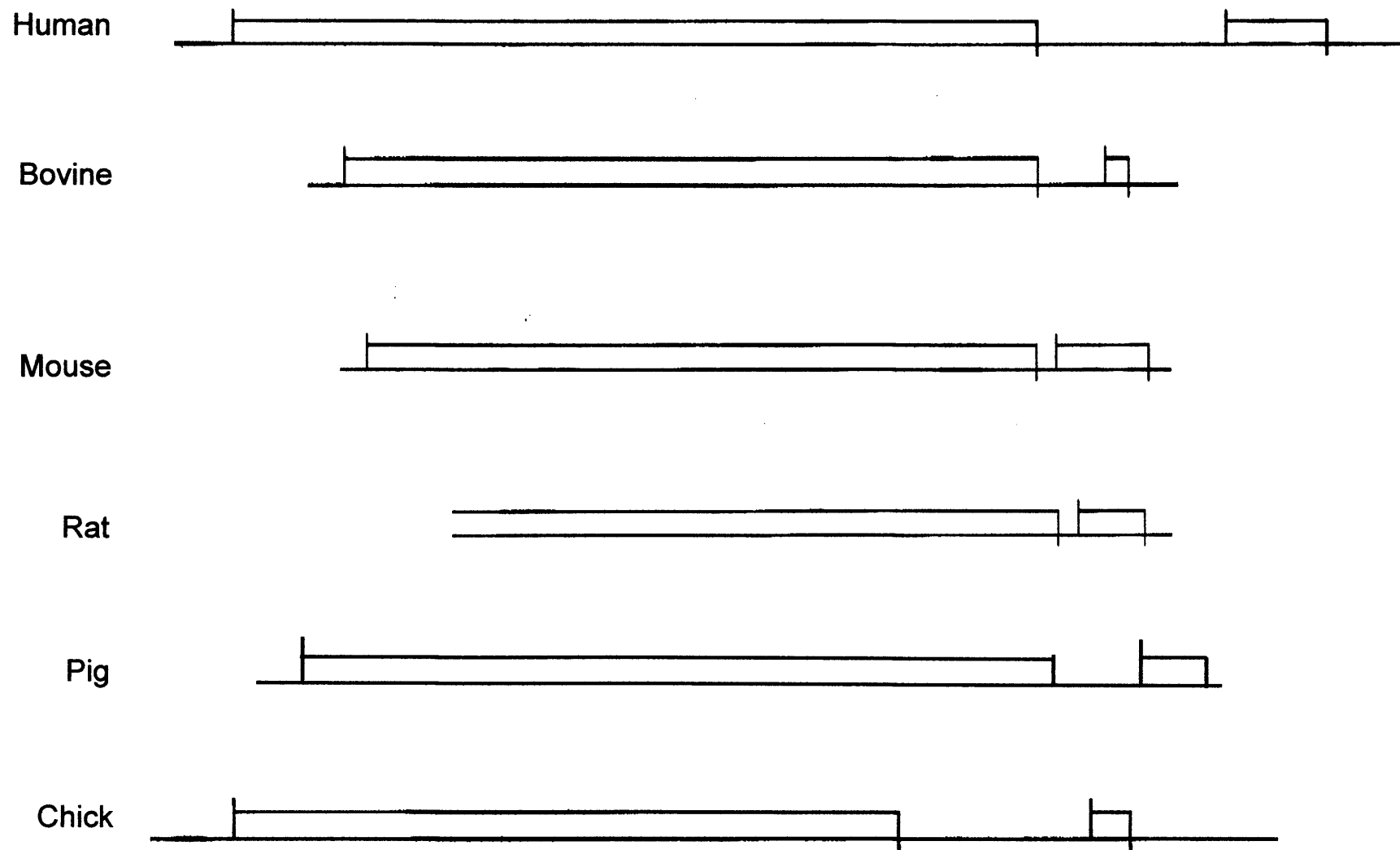


Figure 3.10. The open reading frames of the human, bovine, mouse, rat, pig and chick cDNA.

insert from the I.M.A.G.E. clone 335916 (accession number W17979). The probe was made as detailed in section 2.10.1.2, Chapter 2 and the hybridisation carried out as detailed in section 2.10.2, Chapter 2. The labelled cDNA was approximately 950 bp and was known to be near full-length.

Fifteen clones from a total of approximately 1500 gave a positive signal for hybridisation to the mouse *Spp2* cDNA probe. The positive clones were grown in duplicate in one half of a 96-well microtitre plate, along with a positive and negative control. The positive control was the I.M.A.G.E clone 335916 from which the cDNA probe originated and the negative was the vector pBluescript SK(+) containing no insert. The layout of the microtitre plate and thus the subsequent nylon membranes is shown in table 3.4.

The positive clones were then transferred from the microtitre plate to a nylon membrane in triplicate. The insert from the I.M.A.G.E. clone 335916 was cut using restriction enzymes and the appropriate fragment purified, as shown in figure 3.11, to generate three probes covering the 5', the 3' and the middle and 3' regions.

Each probe was labelled with ^{32}P as described in section 2.10.1.2, Chapter 2 and hybridised to a single filter as described in section 2.10.1.4 to 2.10.1.6, Chapter 2. The autoradiograph after 6 hours exposure with an intensifying screen is shown in figure 3.12 and the results are tabulated in table 3.5. The region covered by each clone is interpreted according to the combination of results with the three different probes. Due to the nature of the spotting onto the nylon membranes (*i.e.* by hedgehoging) it was impossible to apply exactly the same amount of culture to each spot, hence variable background signals sometimes made interpretation difficult.

Six of the originally positive clones were negative and so were deemed to be false positives. Another six had conflicting results with each probe and so a conclusion could not be drawn about the region of cDNA, if any, that they contained. Only three of the originally positive clones could be said to positively contain a defined region of mouse *Spp2* cDNA. Clones 11, 12 and 13 were sequenced as described in section 2.14.2, Chapter 2.

Clones 11 and 13 were shown to contain exon 2 and clone 12 contained exon 7. The complete exon was present in each clone and enabled the exon/intron boundary at each side to be defined.

Table 3.4. The layout of the 96-well microtitre plate containing the positives obtained from the screening of the mouse small insert library.

A small insert genomic library was constructed from a PAC clone containing the mouse *Spp2* gene (Swallow and Dagleish, unpublished). The library had been stored as *E. coli* cultures containing individual clones in 96-well microtitre plates. The library was transferred onto a nylon membrane using the method described in section 2.12.1, Chapter 2. The library was then screened with the ³²P-labelled insert from the I.M.A.G.E. clone 335916 (accession number W17979). The probe was made as detailed in section 2.12.1.2, Chapter 2 and the hybridisation carried out as detailed in section 2.12.1, Chapter 2. The labelled cDNA was approximately 950 bp and was known to be near full-length.

Fifteen clones gave a positive signal for hybridisation to the *Spp2* cDNA probe. The fifteen clones were each assigned a number and were then grown in duplicate in one half of a 96-well microtitre plate whose wells are depicted below as A1 to H6. The positive control was the I.M.A.G.E. clone 335916 from which the cDNA probe originated and the negative was the vector pBluescript SK (+) containing no insert.

	1	2	3	4	5	6
A	1	1	2	2	3	3
B	4	4	5	5	6	6
C	7	7	8	8	9	9
D	10	10	11	11	12	12
E	13	13	14	14	15	15
F	-ve	-ve	+ve	+ve		
G						
H						

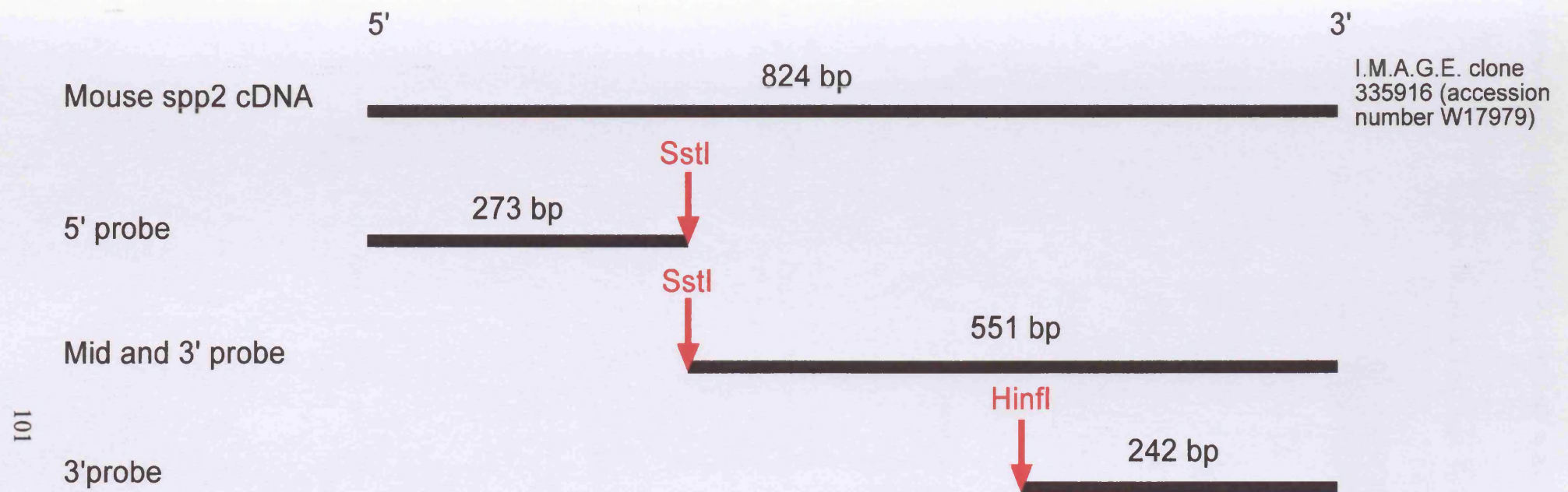


Figure 3.11. The three mouse *Spp2* cDNA probes used to identify the regions covered by the mouse positive clones.

The purified insert from the I.M.A.G.E. clone 335916 (accession number W17979) was digested with *Sst*I to generate two fragments. Each fragment was purified separately from an agarose gel as described in section 2.11.2, Chapter 2 to give the 5' and mid and 3' probes as shown above.

The purified insert from the I.M.A.G.E. clone 335916 was then digested with *Hinf*I. This generated seven fragments. The largest fragment was purified from an agarose gel as described in section 2.11.2, Chapter 2 to give the 3' probe.

The sizes of each fragment in bp are indicated on the fragments above. The restriction enzyme sites are indicated in red.

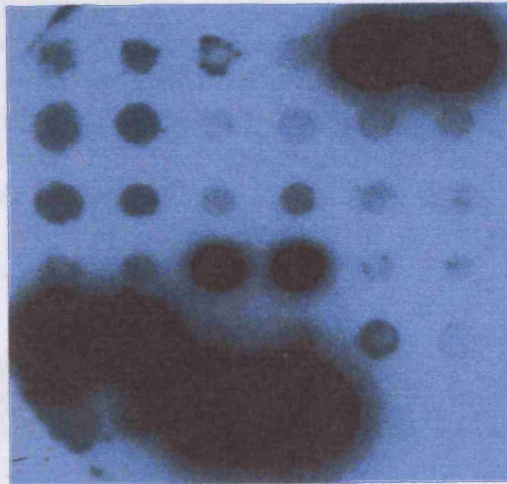
Figure 3.12. The results of hybridisations to the mouse positives using three different regions of the mouse *Spp2* cDNA as a probe.

The mouse positive clones are laid out on each filter as shown below:

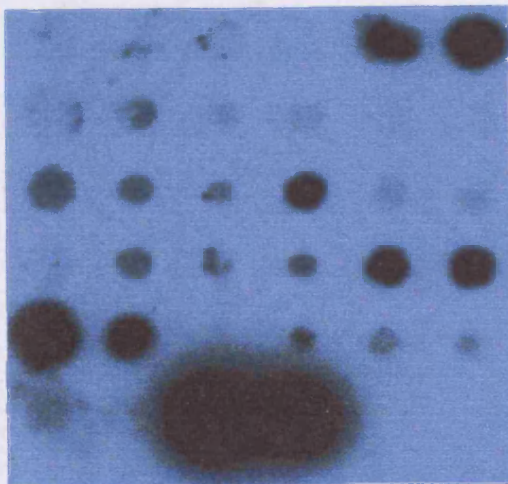
	1	2	3	4	5	6
A	1	1	2	2	3	3
B	4	4	5	5	6	6
C	7	7	8	8	9	9
D	10	10	11	11	12	12
E	13	13	14	14	15	15
F	-ve	-ve	+ve	+ve		
G						
H						

Each clone is spotted in duplicate. Each filter was probed with the ^{32}P -labelled probe indicated, each covering a different region of the mouse *Spp2* cDNA as shown in figure 3.10. The autoradiographs shown represent a 6-hour exposure with an intensifying screen. Table 3.6 indicates whether a positive or negative result is seen in each clone.

5' probe



Mid and 3' probe



3' probe

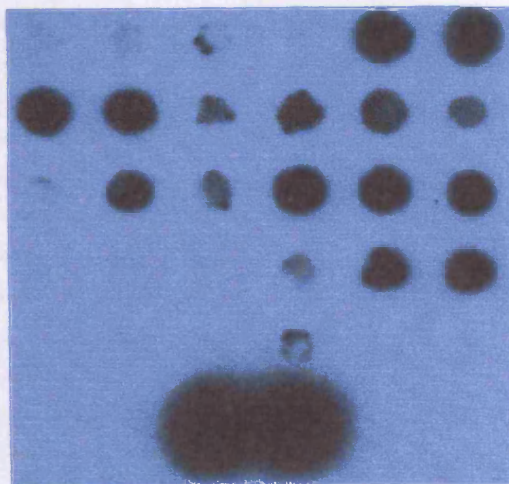


Figure 3.12. The results of hybridisations to the mouse positives using three different regions of the mouse *Spp2* cDNA as a probe.

Table 3.5. Scoring of the mouse *Spp2* cDNA hybridisations, results shown in figure 3.12.

The results of re-hybridising with three different regions of the mouse *Spp2* cDNA to the preliminary positive mouse genomic clones are shown as an autoradiograph in figure 3.12. This table presents the results of each mouse positive clone as a 'score' relative to the probes, which gave a positive signal. A '-' indicates a negative result. A '+' indicates a positive result and a '?' indicates that the result was not clear. A clone that covers the 5' region of the mouse *Spp2* cDNA is shown as '5'', a clone that is from the 3' region is shown as '3'' and a clone that covers the middle and 3' region is shown as 'M'. An 'F' indicates a false positive and a 'U' indicates that the result is uncertain.

Mouse clone	5' Probe	Mid and 3' Probe	3' Probe	'Region'
1	-	-	-	F
2	-	-	-	F
3	+	?	?	U
4	?	-	+	U
5	-	-	?	U
6	-	-	?	U
7	-	-	-	F
8	-	?	?	U
9	-	-	+	U
10	-	-	-	F
11	+	-	-	5'
12	-	+	+	3'
13	+	+	-	5'/M
14	-	-	-	F
15	-	-	-	F
-ve control	-	-	-	Negative for all
+ve control	+	+	+	Positive for all

The mouse 'trace' archive at the NCBI (www.ncbi.nlm.nih.gov/Traces/trace.cgi?) was then BLAST searched using the consensus mouse *Spp2* cDNA sequence that was generated in section 3.2.4. This is an archive that stores the raw sequence data from the mouse genome sequencing project that has not yet been processed. A total of twelve sequences were identified that contained some or all of a mouse *Spp2* exon. Table 3.6 gives the details of the sequences from the trace archives and table 3.7 shows the sizes and locations of the mouse *Spp2* exons. Unlike the human gene, the complete genomic sequence is not available and so it was not possible to determine the intron sizes. However, due to the small size of intron 1, this was contained in its entirety in trace number 13433728. The intron was 100 bp in size, only 1 bp difference compared to the human *SPP2* intron 1.

Traces 11634955, 17208091 and 17892499 all contained promoter sequence *i.e.* sequence preceeding exon 1. Trace 17892499 contained the longest region of sequence and so this sequence was used in the work described in section 3.2.6. A 'TATA' box sequence in the mouse promoter could not be found.

3.2.6 A comparison of the spp24 promoter region between human, mouse and chicken

The promoter regions for human, mouse and chicken (Chapter 6) were compared with one another using the Fasta program (2.21.3, Chapter 2) to see if there were any common motifs that may suggest regions important to the function of the promoter. A region of high homology between chicken and human spanning approximately 60 bp was seen at approximately -20 to -80 and -70 to -130 in human and chicken respectively. This region, although shorter, was also present in the mouse at approximately -50 to -90. This suggested a region that was important to the function of the promoter.

However, when the mouse and human promoter regions were compared they showed a very high level of homology over an unusually large region. High homology was seen over approximately 400 bp just prior to the start of transcription in each gene. Following this 400 bp region the homology completely broke down.

Analysis using the Frames program (2.21.3, Chapter 2) revealed an open reading frame in human and chicken in the opposite orientation to the gene encoding spp24. This reading frame was not seen in the mouse, although the mouse promoter sequence was taken from a trace sequence (*i.e.* is raw sequence that has not yet been edited and trimmed) and so there are

Table 3.6. The sequences from the mouse trace archives (NCBI) that contain mouse *Spp2* exons.

The mouse ‘trace archive’ at the NCBI was BLAST searched using the consensus mouse *Spp2* cDNA sequence. A total of 23 sequences were identified that contain part or all of a mouse *Spp2* exon. This table gives the details of the trace archive sequences and the mouse *Spp2* exons that they contain.

Trace archive ID	Mouse <i>Spp2</i> exon
gnl ti 11634955 ml2C-a84g10.q1c	1
gnl ti 17208091 G10P617462RB5.T0	1
gnl ti 17892499 G10P6299193RG12.T0	Part of 1
gnl ti 13433728 ml2B-a1366d06.q1c	Part of 1 and all of 2
gnl ti 29129761 jli48g09.b1	2
gnl ti 13290377 ml2B-a171e09.p1c	2
gnl ti 19824919 G10P636988RA11.T0	3
gnl ti 18512459 G10P634443FE7.T0	3
gnl ti 13224112 mk2A-a4827h09.p1c	3
gnl ti 3468936 G10P69389RG3.T0	3
gnl ti 19426564 G10P637862FC4.T0	4
gnl ti 16722548 jlf75e03.g1	4
gnl ti 13245908 mk2A-a4838d11.q1c	4
gnl ti 12044598 jil80e03.b1	5
gnl ti 18599363 G10P625878FB12.T0	5
gnl ti 21264180 jrr89d01.g1	Part of 5
gnl ti 4890613 G10P62463FH7.T0	6
gnl ti 18101168 G10P623962FA8.T0	7
gnl ti 10926951 G10P617847FD9.T0	7
gnl ti 13482307 ml2B-a1200d10.q1c	8
gnl ti 13773409 ml2C-a6796g09.q1c	8
gnl ti 1353022 ml1B-a961g10.p1c	8
gnl ti 20840990 G10P636931FD4.T0	8

Table 3.7 The exon/intron boundaries of the mouse *Spp2* gene.

The exon/intron boundaries of the mouse *Spp2* gene were determined either by sequencing of genomic clones from a mouse small insert library or by analysis of genomic sequences from the NCBI trace archive (see section 3.2.5).

The clones from the small insert library revealed the exon/intron boundaries at either end of exon 2 and exon 7. The exons contained in the sequences from the NCBI trace archive are detailed in table 3.6.

This table shows the exon/intron boundaries and the sizes of the exons as determined from a combination of the results discussed above. Exon sequence is shown in upper case and intron sequence in lower case. The consensus gt/ag sequences are shown in bold. A '*' indicates that this junction conforms exactly to the consensus of 'GTRAGT' and 'YYTTYYYYYYNCAG' for the donor and acceptor sites respectively (Senepathy *et al.* 1990). All other junctions are similar to these consensus sequences, but not identical. The size of exon 1 could not be determined as the site of transcription initiation has not been identified for the mouse *Spp2* gene.

Exon	Position in cDNA	Size of exon in bp	Sequence
1	1-159	?CAG g taaag
2	160-284	125 bp	tgtttgctgtct ag GTT.....AG A gtaagt *
3	285-404	120 bp	ttgcctttgtga ag GTC.....GT G gtaagt *
4	405-512	108 bp	tctgtcttttcc ag CCA.....G A Ggtaatga
5	513-567	55 bp	ttaaatttcttt ag ATG.....TT G gtaagt *
6	568-618	51 bp	ctaattgtgttac ag GTT.....TT G gtaagt *
7	619-714	96 bp	ccttccatcctt ag AAA.....G A Ggtaagg
8	715-824	110 bp	tctctcttgaat ag ATT.....TTT

likely to be sequencing errors. The human ORF was translated to give a peptide 90 amino acids in length. This peptide showed homology to the mouse promoter nucleotide sequence when compared using the Framealign program (2.21.3, Chapter 2) and allowing for mismatches. The human 90 residue peptide did not show homology to any known proteins or ESTs using BLAST searches (2.21.3, Chapter 2).

3.2.7 Determination of the nature of the possible insertion/deletion polymorphism seen in the human *SPP2* gene that was originally reported by Gill and Dalgleish

It was postulated that an insertion/deletion polymorphism lay within the human *SPP2* gene (Gill and Dalgleish). The identification of this polymorphism and its speculated basis is described in section 3.1.3.

Two different PACs (14 E15 from the PAC library RPCII, HGMP and 3 N4 from a human chromosome 2 PAC library (Gingrich *et al.* 1996), HGMP) thought to be homozygous for a different allele of the polymorphism were digested with *Sst*I. The polymorphic fragment from each PAC (18.3 kb and 11.2 kb, larger and smaller allele from 3 N4 and 14 E15 respectively) was purified and cloned into the low copy number vector pCL1920 (Lerner and Inouye 1990) using the protocols described in section 2.9.1 and 2.13, Chapter 2. The two cloned fragments were then sequenced from each end using the method described in section 2.14.1, Chapter 2.

If the polymorphism was indeed an insertion/deletion polymorphism as speculated by Gill and Dalgleish, then the expectation is that the ends of both cloned fragments would have the same sequence as it is unlikely that the insertion lies immediately adjacent to one of the ends of the fragment. However, the fragments were only identical at one end. Figure 3.13 depicts this.

Initial speculation was that a *Sst*I site must lie in the insert and that the sequence at the nonidentical ends must actually be the insert in the smaller allele. However, comparison of the unexpected sequence to the DNA segment AC006037 (containing the whole of the human *SPP2* gene) provided a match. This match is shown schematically in figure 3.13. In fact there is no insertion/ deletion, but instead an RFLP. The larger allele that was cloned exhibits the absence of a *Sst*I site that is present in the smaller allele. Unfortunately the extra *Sst*I site was too close to the primer to be accurately sequenced and so the exact sequence change that occurs to give the site is not known.

Figure 3.13. The nature of the RFLPs that lie within the human *SPP2* gene with respect to the genomic sequence AC006037.

A 50 kb region of the genomic sequence AC006037 that contains most of the human *SPP2* gene is shown (A). The sequence AC006037 has haplotype III (a previously unseen haplotype) with respect to three RFLPs. The three haplotypes are shown below with the exact allele sizes determined from sequence AC006037 where possible:

NB. Allele sizes for each RFLP are given in kilobases		Haplotype		
		I	II	III
RFLP	<i>HpaI</i>	29.38 kb	22.4 kb	22.4 kb
	<i>SstI</i>	16.9 kb	Approximately 11.2 kb	16.9 kb
	<i>BglII</i>	6.4 kb	8.5 kb	6.4 kb

The positions of the *HpaI* and *BglII* dimorphisms can be defined as the sites are present in the sequence AC006037 and so it is evident which site is absent in the other allele and the exact sizes can be calculated (B and C). However, with the *SstI* dimorphism the sequence AC006037 contains the larger allele where an *SstI* site is absent and so it is not possible to define exactly where the extra *SstI* site appears in the other allele.

The two *SstI* alleles were cloned and sequenced at each end and the sequences were found to match at one end (shown as a blue rectangle in E and the corresponding region in AC006037 (A)), but differ at the other ends (represented as green and yellow rectangles in E and the corresponding regions in AC006037 (A)). Unfortunately, the extra *SstI* site in the smaller allele was too close to the primer to be accurately sequenced and so the exact size of the smaller allele and the exact nature of the dimorphism could not be determined.

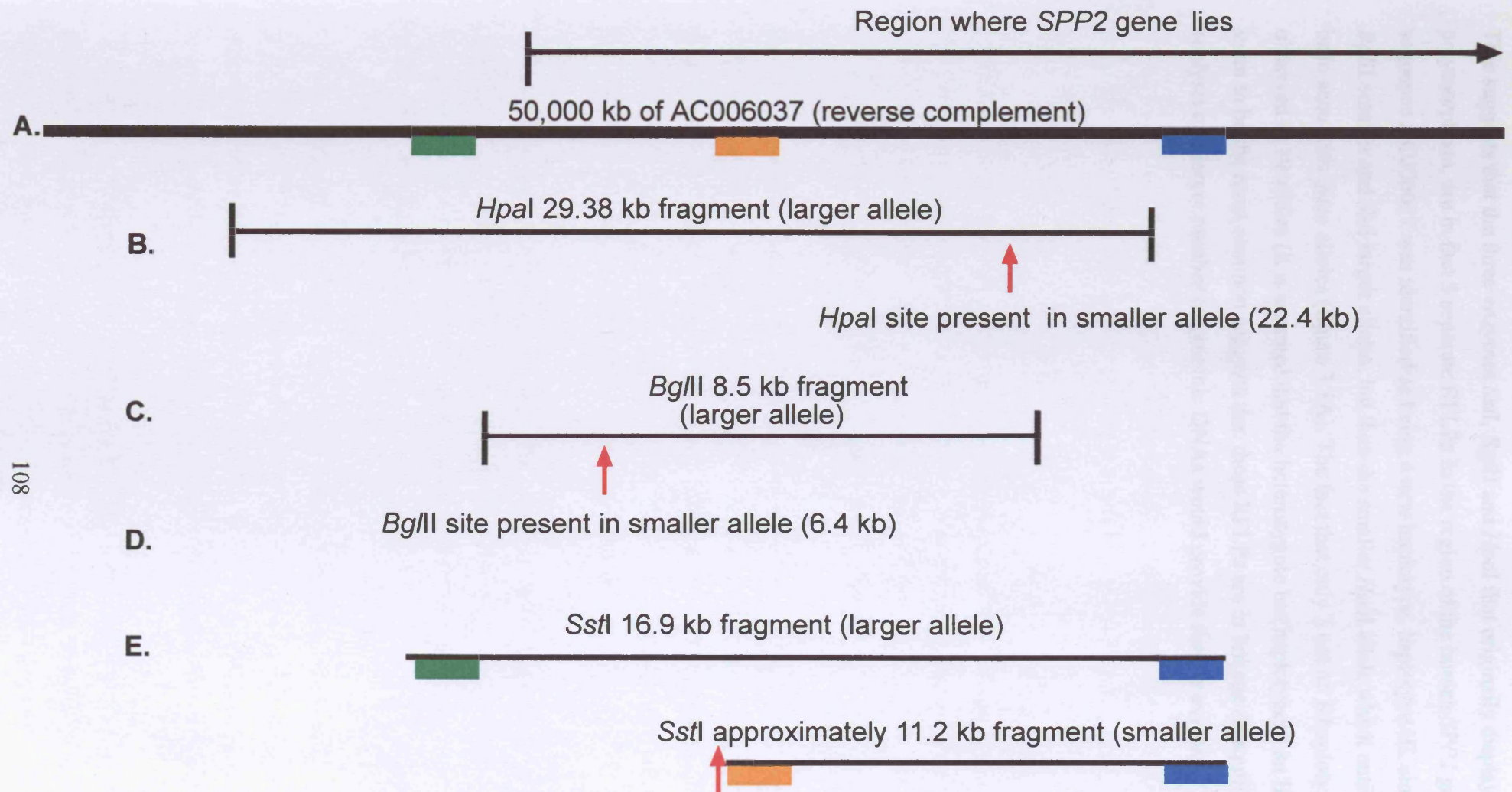


Figure 3.13. The nature of the RFLPs that lie within the human *SPP2* gene with respect to the genomic sequence AC006037.

This suggests that the three enzymes *Sst*I, *Bgl*II and *Hpa*I that originally displayed the polymorphism, are in fact 3 separate RFLPs in the region of the human *SPP2* gene. The sequence AC006037 was identified as being a new haplotype, haplotype III, since it has the *Bgl*II smaller and *Sst*I larger alleles, but then the smaller *Hpa*I allele which until now had not been seen with these alleles (figure 3.1A). The fact that only 3 out of 8 haplotypes were observed in 10 alleles (it is assumed that the heterozygote has haplotypes I and II) and I and II seem to be the most common suggests that these RFLPs are in linkage disequilibrium. An analysis of a larger number of genomic DNAs would provide further evidence.

3.3 Discussion

The spp24 protein can be thought of as two separate domains, the cystatin-like region and the non-cystatin-like region (Chapter 1). Figure 3.5 shows the exons that encode these regions of the protein. The human *SPP2* gene comprises 8 exons and 7 introns. Exon 1 encodes the 5' untranslated region and signal peptide, exons 2 to 4 encode the cystatin-like region, exons 5 to 7 the non-cystatin-like region and exon 8 comprises entirely 3' untranslated region.

A typical cystatin is encoded in 3 exons as opposed to the 4 seen in *SPP2*. Table 3.8 compares the sizes of the typical cystatin exons and the exons seen in the cystatin region of the human *SPP2* gene.

If spp24 is a member of the cystatin superfamily, it appears that the first exon seen in a 'typical' cystatin gene has split into two in *SPP2*. It is interesting that the exon boundary separates the signal peptide from the mature protein and that the size of the intron is only 99 bp. It is not clear what, if anything, the significance of this may be.

Exon 2 of a cystatin and exon 3 of *SPP2* look relatively equivalent, as do exon 3 of a cystatin and exon 4 of *SPP2*. However, exon 4 of *SPP2* also contains the phosphorylated serine region of spp24 and so the cystatin part of the protein found in exon 4 is actually smaller than that seen in exon 3 of a true cystatin. Intron 2 of the *SPP2* gene is much larger than the equivalent cystatin intron, but intron 3 is similar in size to its equivalent.

The only members of the cystatin superfamily known to have 4 exons are the CRPs (cystatin-related proteins). However, they have a specific exon 2 that appears not to be cystatin-related that forms the basis of the additional exon (Devos *et al.* 1993) and so are dissimilar to *SPP2* in the nature of their extra exon.

The similarity of the exon/intron structure of the spp24 cystatin-domain to the true cystatins provides further support for spp24 being a new member of the cystatin superfamily, supporting the original suggestions by Hu *et al.* (1995). However, the differences seen between spp24 and cystatin suggest that spp24 may only be a distant relative of cystatin.

Of the 8 exons encoding spp24, exons 3 and 4 in the human and mouse genes and exon 4 in the putative chicken gene (Chapter 6) encoding spp24 have the potential to be skipped and still maintain the reading frame. No evidence has been seen for this phenomenon in human

Table 3.8 A comparison of exon size between the exons of a typical cystatin and those seen in the cystatin-like region of the human *SPP2* gene.

The human cystatins SN,SA,S,C and D were used to calculate average exon and intron sizes (see review Bobek and Levine 1992). The size of each exon is given in base pairs and amino acids rounded to the nearest whole number. The size of each intron is given in base pairs. The exons are shown aligned against the corresponding exons of the *SPP2* gene. The signal peptides are included and in the case of *SPP2*, approximately 11 amino acids at the end of exon 4 correspond to the phosphorylated serine region of the protein (see Chapter 1).

Cystatins		<i>SPP2</i>	
Exon	Size in bp and amino acids	Exon	Size in bp and amino acids
1	232 bp 77 aa	1	170 bp 29 aa
	-	2	125 bp 41 aa
2	114 bp 38 aa	3	123 bp 41 aa
3	81 bp 27 aa	4	111 bp 29 aa
Intron	Size in bp	Intron	Size in bp
	-	1	99
1	1697	2	7740
2	1202	3	1410

and mouse, but evidence of exon 4 skipping and part of exon 6 has been seen in rat. This assumes that the exon/intron boundaries are the same in rat (Chapter 6) although they have not been formally characterised.

Within the *SPP2* gene 2 Alu Y, 1 Alu Sg and 2 Alu Sq elements were predicted. The presence of Alu elements also has other implications (reviewed by Mighell *et al.* 1997). For instance, the younger Alu elements may still retain the property of mobility. Younger Alus also have more CpG doublets and so have a higher mutation rate. One of the Alu Y elements within the human *SPP2* gene lies in intron 6, consequently the Alu would be present in heterogeneous nuclear RNA. The Alu could potentially affect gene expression as Alus contain several regions that differ only by one or two base pairs to consensus donor or acceptor splice sites. Therefore point mutations in this intronic Alu could potentially create splice sites and disrupt normal splicing (Makalowski *et al.* 1994).

The extensive sequence analysis of the human *SPP2* gene revealed a gene with no striking characteristics. It is rare to find the termination codon in the penultimate exon as seen with *SPP2*, however this is not uncommon in secreted proteins (Nagy and Maquat 1998). Spp24 is thought to be a secreted protein due to the presence of the signal peptide (Chapter 1).

The human *SPP2* gene does not have an obvious promoter. The promoter prediction programs in the NIX analysis environment could not determine a consensus promoter. A further search by eye of the region upstream of the primary transcription initiation sites determined by primer extension, could also find no TATA or CAAT box.

The promoter of a eukaryotic gene would normally be expected to comprise two major parts. The core promoter which lies in the region adjacent to the transcription start site and a more distant enhancer region (Roeder 1991; Tjian and Maniatis 1994). Within the core promoter the two key elements are the TATA box and an initiator sequence (Inr) (Breathnach and Chambon 1981; Smale and Baltimore 1989). It is possible for a core promoter to contain both of these elements, either one or the other of these elements, or neither of these elements.

The human *SPP2* gene does not have the TATA box, but there is the possibility of it having an Inr sequence. The consensus Inr sequence was reported as Py Py A⁺¹ N T/A Py Py, where Py is a pyrimidine (C or T) (Smale 1997). A more detailed analysis revealed that within this consensus sequence an A at +1, a T or an A at +3 and a pyrimidine at -1 are the most critical in determining the strength of the Inr sequence (Javahery *et al.* 1994; Lo and Smale 1996).

The primary transcription initiation site found in human liver (section 3.2.3) lies in the sequence 'GCCAGTGT', with the A being in the +1 position. This could be an initiation sequence. It matches the consensus at every position thought to be crucial in determining the Inr strength *i.e.* an A at +1, a T at +3 and a C at -1. It also has pyrimidines in several of the surrounding positions. The two primary transcription initiation sites seen in kidney lie in the sequences 'CTCTTAGG' and 'TCTTAGGA' with the second T and the third T being in the +1 position in each sequence respectively. The first sequence matches more closely to the consensus Inr sequence than the second, but neither are as good a match as the potential Inr sequence defining the liver transcription start point.

The Inr element of a promoter has been demonstrated as having the transcriptional responses necessary for lineage-specific gene expression and cannot be replaced by a TATA box (reviewed by Novina and Roy 1996). An example of this is seen in the *FcγR1b* gene promoter (TATA⁻Inr⁺), where the Inr element is required for myeloid-specific expression and selective interferon-γ (IFN-γ) responsiveness (Eichbaum 1994). The artificial introduction of a TATA box either in place of or in addition to the Inr element results in an increase in gene expression, but a loss in lineage specificity (Eichbaum 1994). Inr elements are also thought to be responsible for the temporal regulation of gene expression (reviewed by Novina and Roy 1996). An example of this is seen in the *Drosophila Adh* (alcohol dehydrogenase) gene. The Inr element mediates the molecular switch between a distal promoter preferentially used during embryonic development and adult developmental stages and the proximal promoter used at other times (Hansen and Tjian 1995).

The Inr element is also thought to control spatial expression. The *Drosophila Dpp* gene (decapentaplegic) encodes a protein related to transforming growth factor β (TGF-β), which is important in dorsal-ventral pattern formation. The promoter of the *Dpp* gene (TATA⁻Inr⁺) controls the spatial expression profile of the gene in the developing embryo and is resistant to ventral activation, thus preventing dorsalisation of the embryo (Schwyter *et al.* 1995).

The likelihood that the human *SPP2* gene core promoter is TATA⁻Inr⁺, therefore suggests that the gene has a lineage-specific expression and that the temporal and spatial expression of the gene is under tight control by the Inr element.

To try and identify any potential upstream regulatory regions of the human *SPP2* gene, the one kilobase of DNA sequence upstream of the primary transcription initiation site seen in liver, was searched using the program MatInspector V2.2 (www.gsf.de/biodv/index.html)

(Quandt *et al.* 1995). Search programs of this type can predict the presence of many transcription factor binding sites as they will pick up sites which only match loosely with consensus binding sites as it is very difficult to determine good consensus sequences for transcription factors. For this reason the programs should only be used as an indication of potential binding sites and should not replace experimental evidence.

MatInspector indicated the presence of 3 possible Sp1 sites, 6 AP1 sites and several C/EBP α , C/EBP β , HNF-1 and HNF-3 β sites in the region of the human *SPP2* gene searched, as well as many other transcription factor binding sites. Sp1 and AP1 (activator protein 1) are general transcription factors, but the C/EBPs and HNFs are more specific. The program also predicted 2 estrogen receptors and 1 glucocorticoid response element, but when these sequences were compared to a known consensus sequence for each of these elements they did not look genuine.

The C/EBPs (CCAAT/enhancer-binding proteins) are known to be found in liver, fat, lung and intestine (OMIM entries 116897 and 189965 for C/EBP α and C/EBP β respectively). The HNFs (hepatocyte nuclear factors) are known to be found in liver, kidney, lung and intestine (OMIM entries 142410 and 600288 for HNF-1 and HNF-3 β respectively).

Both these families of transcription factors are expressed in the liver and have a limited cellular distribution. The northern blots performed by Hu *et al.* (1995) (Chapter 1) suggest that the gene encoding bovine *spp24* is expressed in a tissue specific manner and is highly expressed in liver.

The study of the human *SPP2* gene promoter region, although a little inconclusive, does suggest, when combined with other evidence, that the *SPP2* gene is expressed in a tissue-specific manner. It may be expressed by cells of a specific lineage that are found in liver and at specific stages of embryonic and adult development.

The determination of the mouse *Spp2* gene exon/intron boundaries enables a comparison to be made between the structure of the mouse *Spp2* and the human *SPP2* gene. When the human and mouse *spp24* cDNAs were aligned using the program Gap in the GCG molecular biology package (section 2.21.3, Chapter 2), the exon boundaries were shown to be in essentially the same position in both species. All of the exon/intron boundaries in the human and the mouse gene conform to the gt/ag consensus.

Two of the exon/intron boundaries in the putative chicken spp24 have been identified (Agarwal *et al.* 1995). These are the boundaries between exons 3 & 4 and exons 4 & 5. These boundaries are in the same positions as in the human and mouse and also confirm that exon 4 could be skipped in chicken whilst still maintaining the reading frame (earlier in the discussion).

Generation of the mouse consensus cDNA for spp24 revealed the presence of a second shorter ORF that lies 3' to the major ORF. This is also seen in other species. None of the predicted 'ATG' codons for the major or minor ORFs lie in a typical Kozak sequence ('GCCGCCA/GCCAUGG' (Kozak 1989)). Approximately 5-10% of all vertebrate mRNAs do not have 'ATG' codons lying in a Kozak sequence (Kozak 1989). These mRNAs are thought to have a mechanism that ensures that the most 5' 'ATG' is not missed, but this codon may not be used exclusively. Therefore, it is possible that the second ORF is also translated, but the significance of this, if any, is not clear as the resulting peptide is different lengths and of different composition in all species. It is also possible that the signal peptides in each species do not start at the first 'ATG', but may start at subsequent ones. For example the human gene has 8 'ATG's in the signal peptide and the mouse gene 2.

The mouse promoter region (determined from trace sequences) does not reveal the presence of a 'TATA' box or Inr element. However, the sequence may contain sequence errors as it is from the trace archive.

A comparison between the human, mouse and chicken promoter regions revealed a high level of homology over an extensive region between mouse and human a small portion of which was also seen in chicken. An ORF was seen in human and chicken in this region. However, the lack of any ESTs or homologous proteins to this peptide suggest that the ORF is not expressed. This region is possibly the remnants of a pseudogene, hence the high level of homology seen between human and mouse, that lost its function before the divergence of chickens from mammals, hence only a small region is left in the chicken sequence. The close proximity of the sequence to the start of the gene encoding spp24 suggests that vital promoter elements for spp24 must lie within this sequence.

The nature of the polymorphism that was originally postulated to be an insertion/deletion polymorphism (Gill and Dalgleish) has been determined and found to be three RFLPs. These RFLPs are thought to be in extreme linkage disequilibrium as evidenced by the observation of

only 3 out of a possible 8 different haplotypes though much larger numbers of samples now need to be studied.

This linkage disequilibrium suggests that either this region is a fairly 'cold' spot in the genome (*i.e.* there is only very occasional, random recombination) and there has not been sufficient time for allelic association to have been disrupted or that the polymorphisms are a result of very recent mutations that actually lie in a region exhibiting normal levels of recombination, but because they are so recent they are still in linkage disequilibrium.

A high level of linkage disequilibrium across the region of the human *SPP2* gene would increase the power of association tests. This is due to the fact that particular microsatellite and RFLP alleles will tend to be associated more frequently with the mutation that is contributing to the disease being investigated than they would in regions of lesser linkage disequilibrium.

In summary, the exon structure of the cystatin-like region of the human *SPP2* gene suggests that *spp24* is a member of the cystatin superfamily. The gene could be relatively young in evolutionary terms due to the presence of a very small intron between exons 1 and 2 and an Alu Y element in intron 6.

The human *SPP2* gene appears to have a TATA⁻Inr⁺ promoter and several potential liver-specific transcription factor binding sites upstream of the primary transcription initiation site in liver. This suggests that gene expression is tissue-specific and possibly lineage-specific with tightly regulated temporal and spatial expression.

Both the human and mouse genes display the same structure, comprising 8 exons and 7 introns. The exon/intron boundaries correspond to the same cDNA position in both species and all conform to the gt/ag consensus. This suggests the structure of the gene encoding *spp24* is conserved between species.

Chapter 4

The expression of the gene encoding secreted phosphoprotein 24

4.1 Introduction

The expression profile of a gene provides information about when and where it is expressed. This can indicate possible protein functions and also provide evidence to support any previously speculated functions. In the case of spp24, where the function of the protein is unknown, it is essential to build up a detailed expression profile to try and provide some indication to what the function may be. This will aid the decision of which approach to take in future functional studies.

The spp24 protein was originally isolated from the demineralised extract of bovine cortical bone (Hu *et al.* 1995). This shows the localisation of the protein in this tissue, but it cannot be assumed that the gene is also expressed here. However, Hu *et al.* (1995) also reported the results of a northern blot analysis on bovine bone periosteum, heart, lung, kidney, spleen and liver. A single transcript corresponding to the size of the deduced cDNA sequence was seen in bovine bone periosteum and liver, but not in heart, lung, kidney or spleen. The highest level of expression was seen in liver. This provides limited expression data in the bovine species. This chapter presents a large amount of expression data for the mouse and human gene from a variety of sources, which are discussed below.

4.1.1 The use of expressed sequence tags (ESTs) to obtain expression data

Expressed sequence tags (ESTs) are available for many species from a growing number of EST databases. ESTs are short cDNA sequences that have been obtained from cDNA libraries and so are known to be expressed in the tissue from which the mRNA was obtained. ESTs can be used to build a contig of a full-length cDNA sequence (Chapter 3), but they can also provide some information regarding expression.

The quality of some ESTs is dubious and so ESTs should always be treated with caution. However, numerous ESTs for a particular gene in a particular tissue from a variety of sources are a good indication that the gene really is expressed there. Of course ESTs do not provide quantitative information. The number of ESTs from a particular tissue may simply be a reflection of the availability of the tissue and its cDNA libraries.

In this chapter, the source of spp24 ESTs from human and mouse are collated to provide some cautiously presented information for an expression profile of the human and mouse genes. The ESTs are from two main sources, the TIGR human and mouse gene indices (www.tigr.org/tdb) and the UniGene human and mouse databases (www.ncbi.nlm.nih.gov/UniGene/).

4.1.2 The use of northern blot analysis, ribonuclease protection assays and RT-PCR to obtain expression data

To obtain direct expression data for a gene there are several popular techniques, northern blot analysis, ribonuclease protection assays and RT-PCR.

Northern blot analysis is a relatively 'low tech' method and it requires very little enzymatic manipulation of RNA. It provides information about the size of a transcript and may indicate the presence of alternative splicing. Northern blot analysis also allows a direct comparison of mRNA abundance between tissues and the type of probe that can be used is very versatile. However, the technique is not without disadvantages. Northern blot analysis is intolerant of degradation and the RNA needs to be of a very high quality. Also, of the three techniques, northern blot analysis is the least sensitive.

Ribonuclease protection assays involve the hybridisation in solution of an antisense probe to an RNA sample. Unhybridised probe and RNA are then degraded by ribonucleases and the hybridised fragments are separated on a polyacrylamide gel. This technique is extremely sensitive, approximately 10 to 100 fold more sensitive than northern blot analysis, and is more tolerant of partially degraded RNA. Also, hybridisation in solution is more efficient than filter hybridisation.

Ribonuclease protection assays are quantitative and it is possible to carry out a multiprobe analysis. The drawbacks of the technique are the lack of information regarding size, as the protected fragment is determined by the length of the probe, and the fact that the probe must be RNA.

RT-PCR, reverse transcription followed by PCR, is the most sensitive of the three techniques described. In theory, a single copy of a transcript can be detected by this technique. RT-PCR is slightly tolerant of partially degraded RNA, but is intolerant to RNA contaminated with

DNA. The RNA samples used must be very pure and DNA free. RT-PCR can be used for quantitation, but the optimisation requirements are laborious.

In this chapter, the results of RT-PCR on various mouse RNA samples are presented. This technique was chosen due to its high sensitivity and its tolerance to partially degraded RNA as discussed above. The RT-PCR was carried out on mouse samples, rather than human, due to the ease of obtaining mouse tissues.

4.1.3 The use of microarrays to obtain expression data

The use of microarrays in expression studies is a relatively new phenomenon. With the advancement of automated robotics it is now possible to spot thousands of samples onto filters, plates or chips. The arrays are then probed, usually with a fluorescent probe, and each sample is then scanned to obtain their relative intensities.

Arrays can be made with DNA, RNA, or synthetic oligonucleotides, but currently the most popular technique is to make an array of cDNA clones. Individual cDNA clones are spotted onto filters, plates or chips, which are then probed with RNA from a particular tissue. In this way, the expression pattern of tissues can be compared to identify genes that are up- or down-regulated.

Quantitation using microarrays does not produce absolute values, but relative values. Hybridisations are always done in pairs, or more, to obtain figures relative to a reference. Microarrays have been used to compare gene expression in different tissues and also to investigate the effect on gene expression of different chemical treatments.

This chapter presents data obtained on the mouse gene encoding spp24 from the RIKEN cDNA Expression Array Database (READ) (Miki *et al.* 2001). Miki *et al.* (2001) arrayed approximately 19,000 cDNAs and characterised the gene expression profiles for a number of adult and developing mouse tissues. It was estimated that there were about 13,600 non-redundant genes in the array and all tissues were compared to pooled male and female 17.5-day embryos, which have a relatively complex RNA expression pattern and is easily reproducible.

As well as looking at the expression profile of individual genes or tissues, Miki *et al.* (2001) performed a cluster analysis and defined sets of genes that were expressed ubiquitously and

sets of genes that were expressed in similar groups of tissues. They also clustered the genes coding for known enzymes into 78 metabolic pathways. This revealed a co-ordination of expression within each pathway among different tissues, demonstrating how expression profiles can be useful in revealing possible functions for a protein.

Also presented in this chapter are data from a hybridisation carried out on a human RNA array. The array was prepared by Clontech Inc. and comprises poly A⁺ RNA from many different human tissues and cell lines. The array was probed with human spp24 cDNA to determine the expression profile of the *SPP2* gene in these tissues.

Finally, this chapter presents data purchased from Incyte Genomics Inc. from a hybridisation carried out on a human cDNA array. The experiment compares the expression of the *SPP2* gene in osteoblast precursor cells and osteoblast cells that have been stimulated to mature. These data were purchased because of speculation by Hu *et al.* (1995) that spp24 might have a role in the process of bone turnover.

4.2 Results

4.2.1 Expression data obtained from ESTs

As discussed in section 4.1.1, ESTs can be used as a source of expression data. The quality of some ESTs should be treated with caution, but the appearance of numerous ESTs from a particular tissue that have been submitted from several different sources are a good indication that a gene is indeed expressed in that tissue. The largest EST databases are those of mouse and human. For this reason these were expected to provide the most reliable expression data for spp24.

In total, 57 mouse spp24 ESTs were identified from the TIGR Mouse Gene Index (MGI) (www.tigr.org/tdb/mgi/) and the UniGene EST database (www.ncbi.nlm.nih.gov/UniGene). These EST sequences were submitted from either the National Institute of Dental and Craniofacial Research, RIKEN (The Institute of Physical and Chemical Research) or Washington University School of Medicine Table 4.1 shows the number of ESTs from each tissue.

From the mouse ESTs it was concluded that spp24 was expressed predominantly in kidney and liver. Thirty three percent of the mouse ESTs are from kidney (19 in total) and they have been submitted from several different research institutes. Only 7% of the ESTs are from liver (4 in total), however, they have been submitted from two different research institutes and the expression of spp24 in bovine liver has previously been reported (Hu *et al.* 1995).

Hu *et al.* (1995) reported that northern blot analysis on bovine tissues showed no spp24 expression in kidney. The mouse EST evidence contradicts these results. It may be that the expression of spp24 in bovine kidney is too low to be detected by northern blot analysis. Alternatively, this may represent a true difference between mice and cattle with respect to spp24 expression. The number of mouse ESTs seen from each tissue is not an indication of the level of expression, merely of the tissue bias in the availability of libraries.

The mouse ESTs also suggest that spp24 could be expressed in the uterus, placenta, macrophage, T-cell, proximal colon and diaphragm. However, the number of ESTs from these tissues is small and so it cannot be reliably concluded that spp24 is expressed there. Further evidence is needed to support these indications.

Table 4.1. The number of mouse spp24 ESTs from various tissues.

ESTs came from the TIGR Mouse Gene Index (MGI) (www.tigr.org/tdb/mgi) and the UniGene EST databases (www.ncbi.nlm.nih.gov/UniGene) cluster Mm.28247. These EST sequences were submitted to these databases from either the National Institute of Dental or Craniofacial Research, RIKEN (The Institute of Physical and Chemical Research), or Washington University School of Medicine. In total there are 57 mouse spp24 ESTs.

Source	Number of ESTs
Kidney	19
Soares mouse - strain NML	6
13.5-14.5 day total foetus	5
Liver	4
Kidney day 7	4
Uterus	3
18 day embryo	3
Placenta	2
Kidney day 0	2
Macrophage	2
19.5 day total foetus	2
T-cell	1
Proximal colon	1
Embryonic carcinoma	1
E8.5 mouse craniofacial subtraction cDNA library	1
Diaphragm	1
Total	57

The mouse ESTs also give some information regarding expression during various stages of development. For example it is likely that spp24 is expressed in some tissue at the 13.5, 14.5, 18 and 19.5 day of embryonic development. It can also be reliably concluded that as well as being expressed in the adult mouse kidney; spp24 is also expressed in the kidney of a 7-day-old mouse and a newborn mouse (day 0).

In total, 23 human spp24 ESTs were identified from the TIGR Human Gene Index (HGI) (www.tigr.org) and the UniGene EST database (www.ncbi.nlm.nih.gov/UniGene). These EST sequences were submitted from the Beijing Institute of Radiation Medicine, Washington University School of Medicine, the National Cancer Institute, Pohang Institute of Science and Technology or the University of California . Table 4.2 shows the number of ESTs from each tissue.

The human ESTs enable it to be reliably concluded that spp24 is expressed in liver. This is consistent with the results reported by Hu *et al.* (1995) and the mouse ESTs discussed above. Twenty two percent of the human ESTs are from adult liver and 13% from foetal liver. Forty three percent of the human ESTs are from a foetal liver and spleen library. The appearance of spp24 in this library would be expected, as we know spp24 to be expressed in foetal liver. However, it is impossible to say whether spp24 is expressed by spleen as all these ESTs could be from the liver.

There are 4 human ESTs from a foetal lung, testis and B-cell library. This suggests that spp24 is expressed in one of these tissues at some level. However, it is impossible to say from which tissue or tissues the ESTs actually came.

There is a single EST from human skeletal muscle and therefore a reliable conclusion cannot be drawn from this.

4.2.2 The use of RT-PCR to carry out an expression study in mouse

RNA was extracted from the tissues of an adult male mouse (13 weeks, supplied by Carole Yauk, University of Leicester) using the RNAsol method described in section 2.8.2. RT-PCRs were then carried out on 4 µg of total RNA using the method described in section 2.11.3. The PCR conditions used were: (96°C 30s, 67°C 30s, 72°C 30s) × 24.

Table 4.2. The number of human spp24 ESTs from various tissues.

ESTs came from the TIGR Human Gene Index (HGI) (www.tigr.org/tdb/hgi) and the UniGene EST database (www.ncbi.nlm.nih.gov/UniGene) cluster Hs.12230. These EST sequences were submitted to these databases from the Beijing Institute of Radiation Medicine, Washington University School of Medicine, the National Cancer Institute, Pohang Institute of Science and Technology or the University of California. In total there are 23 human spp24 ESTs.

Source	Number of ESTs
Foetal liver and spleen	10
Liver	5
Foetal lung, testis, B-cell	4
22 week foetal liver	2
Foetal liver	1
Muscle (skeletal)	1
Total	23

The primers used for the RT-PCR and their position in the mouse cDNA are shown in figure 4.1. The forward and reverse primers are 'tagged' with an *Eco*RI and *Bam*HI restriction enzyme site respectively. This was to enable cloning of any products if required. The primers span the region encoding the mature spp24 protein and the size of the expected RT-PCR product is 566 bp.

The RT-PCR products from each mouse tissue were electrophoresed on 1.3% agarose gels along with the corresponding no RT negative controls, a no RNA in the initial reverse transcription negative control and a PCR using water negative control. These gels are shown in figure 4.2. Ideally, this experiment needed a positive control for each RNA isolation with, for example, β actin. This would have verified the quality of the RNA and shown that the RT reaction had worked. It should also be noted that attempts were made to isolate RNA from samples of mouse bone, but failed.

RT-PCR products of approximately the expected size (566 bp) were seen in liver, brain, diaphragm and kidney, but not in heart, muscle, testis, eye, lung and stomach. These PCR products must be from template DNA that has been reverse transcribed from RNA as there are no products seen in the no RT controls. This confirms that the total RNA preparations are free from contaminating DNA that would otherwise have amplified in the PCR to give products. The size of the products also confirms that it is template DNA that has been reversed transcribed as the primers span several exon/intron boundaries (exon/intron boundaries defined in Chapter 3). The size of the product is consistent with there being no introns present, hence the template DNA must have been transcribed from mRNA.

The results support the tissue-specific expression that had already been seen. As expected from the results reported by Hu *et al.* (1995) and the human and mouse ESTs, RT-PCR products were seen in mouse liver and kidney. However, expression was also found in brain and diaphragm. Brain has not previously been analysed, but a single diaphragm EST was seen in the mouse ESTs. The RT-PCR result would suggest that this EST is reliable.

The RT-PCRs were repeated, but this time using only 1 μ g of total RNA, to check reproducibility. The products were again electrophoresed on a 1.3% agarose gel. The gel is shown in figure 4.3.

The expected 566 bp RT-PCR products are seen in the tissues expected from the first RT-PCR results. However, this time a second, smaller, but fainter band is seen in liver and brain.

```

1  ACAAGAATAA GACAGCCACC CTCTGAAAGA GCTGTCATCC AGAAGCCTGG
51  AGAGAGGCCG TCTCCCTGAC TCTGGGTCGC CATCCTCTCA GTATGGAGCA
101 GGCAATGCTG AAGACGCTGG CTTTGTGGT GCTGGGCATG CACTACTGGT
151 GTGCCACAGG TTTCCCGGTG TACGACTACG ACCCTTCCTC TCTGCAGGAA
201 GCTCTCAGTG CCTCAGTGGC AAAGGTGAAC TCGCAGTCCC TGAGTCCTTA
251 CCTGTTTCGG GCGACCCGGA GCTCCTTGAA GAGAGTCAAC GTCCTGGATG
301 AAGACACATT GGTCAATGAAC TTAGAGTTCA GTGTTTCAGGA AACCACATGC
351 CTGAGAGATT CTGGTGATCC CTCCACCTGT GCCTTCCAAA GGGGCTACTC
401 TGTGCCAACA GCTGCTTGCA GGAGCACTGT GCAGATGTCC AAGGGACAGG
451 TAAAGGATGT GTGGGCTCAC TGCCGCTGGG CGTCCTCATC TGAGTCCAAC
501 AGCAGTGAGG AGATGATGTT TGGGGACATG GCAAGATCCC ACAGACGAAG
551 AAATGATTAT CTAATTGGTT TTCTTTCTGA TGAATCCAGA AGTGAACAAT
601 TCCGTGACCG GTCACTTGAA ATCATGAGGA GGGGACAGCC TCCCGCCCAT
651 AGAAGGTTCC TGAACCTCCA TCGCAGAGCA AGAGTAAATT CTGGCTTTGA
701 GTGACATCCT GGAGATTTCA TGAAAGAAAG AGAAGCAGAA GCTGAAATGA
751 AGAAAGGCAT GGAGAATGGT GTCTTTTTCC TTTTATAAT CTCCACTCTG
801 CAATAAAGAT CTTTCCCTTC CTTT

```

Forward primer

5' CATAGAATTCCCGGTGTACGACTACG-3'

Reverse primer

5'-CATAGGATCCAGGATGTCACTCAAAG-3'

Figure 4.1. The position of mouse RT-PCR primers with respect to the mouse cDNA sequence.

The 824 nucleotide mouse consensus cDNA sequence is shown as determined in Chapter 3. The 'ATG' start codon at positions 105-107 and the 'TGA' stop codon at positions 702-704 are boxed.

The sequences of the RT-PCR primers are shown. The forward and reverse primer are tagged with an *Eco*RI and *Bam*HI restriction enzyme site respectively. These sites are underlined in the primer sequence. The directionality of the primer is indicated by an arrow in the cDNA sequence which also shows the nucleotides of the primer that hybridise to the cDNA. These nucleotides are shown in red in the primer sequence and the cDNA sequence. The size of the expected RT-PCR product is 566 bp.

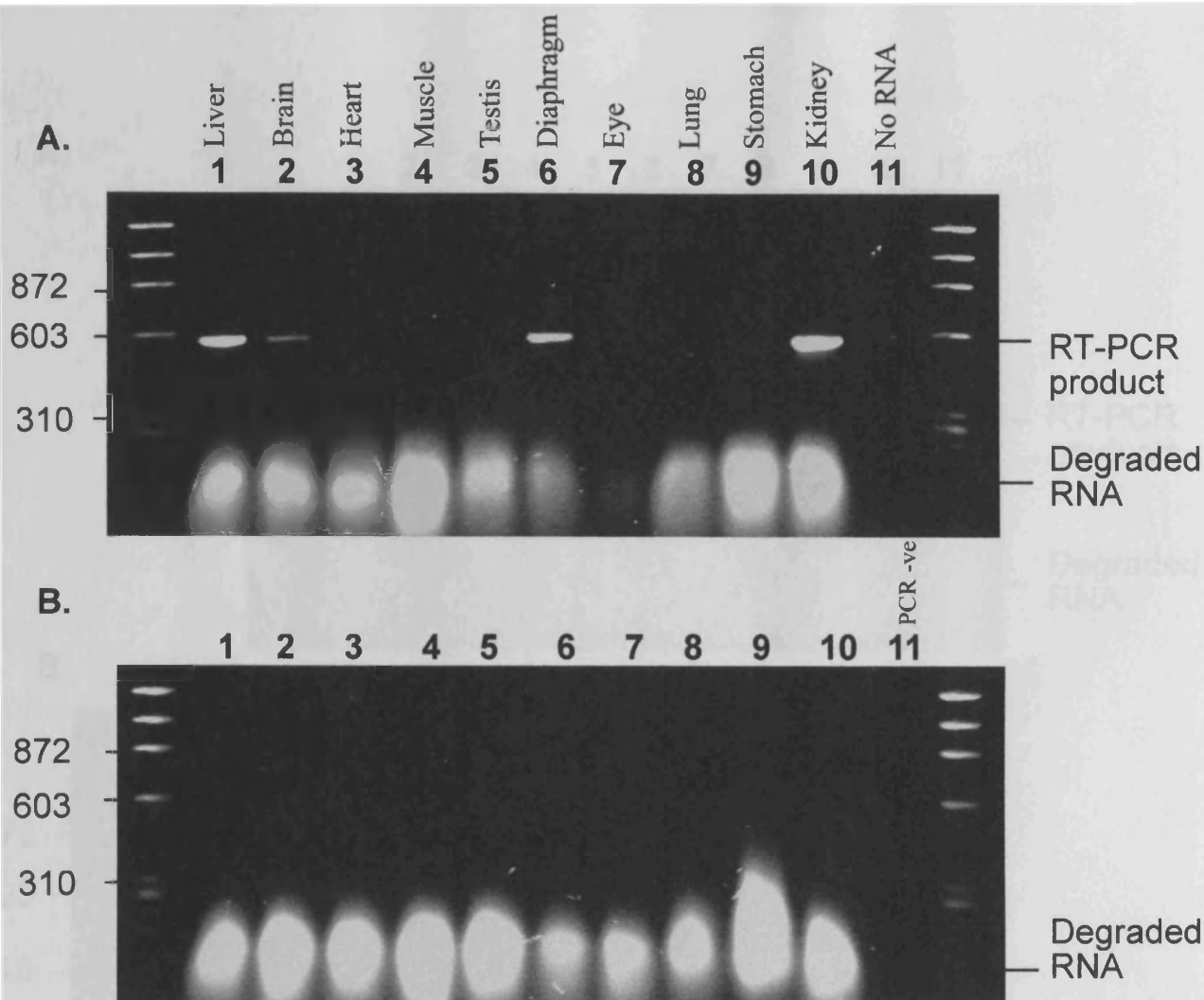


Figure 4.2. RT-PCR performed on RNA from adult mouse tissues.

RT-PCR was performed on 4 μ g of total RNA as described in section 2.13.3. The RNA had been extracted from adult mouse tissues (13 week mouse supplied by Carole Yauk, University of Leicester) using the method described in section 2.10.2.

The RT-PCR products were electrophoresed on 1.3% agarose gels. Gel A shows the RT-PCR products obtained in liver (lane 1), brain (lane 2), heart (lane 3), muscle (lane 4), testis (lane 5), diaphragm (lane 6), eye (lane 7), lung (lane 8), stomach (lane 9) and kidney (lane 10). Lane 11 on Gel A shows the no RNA control.

Gel B shows the no RT controls. Each lane has the same tissue as the corresponding lane in the top gel with the exception of lane 11. Lane 11 on Gel B contains the PCR negative.

RT-PCR products and degraded RNA are indicated on the right hand side of the gels. The sizes of three marker bands (ϕ X174 RF cut with *Hae*III) are indicated on the left hand side in basepairs. RT products of approximately the expected size (566 bp) are seen in liver, brain, diaphragm and kidney. All the negative controls are clear.

Ideally, this experiment needed a positive control for each RNA isolation with, for example, β actin. This would have verified the quality of the RNA and shown that the RT reaction had worked.

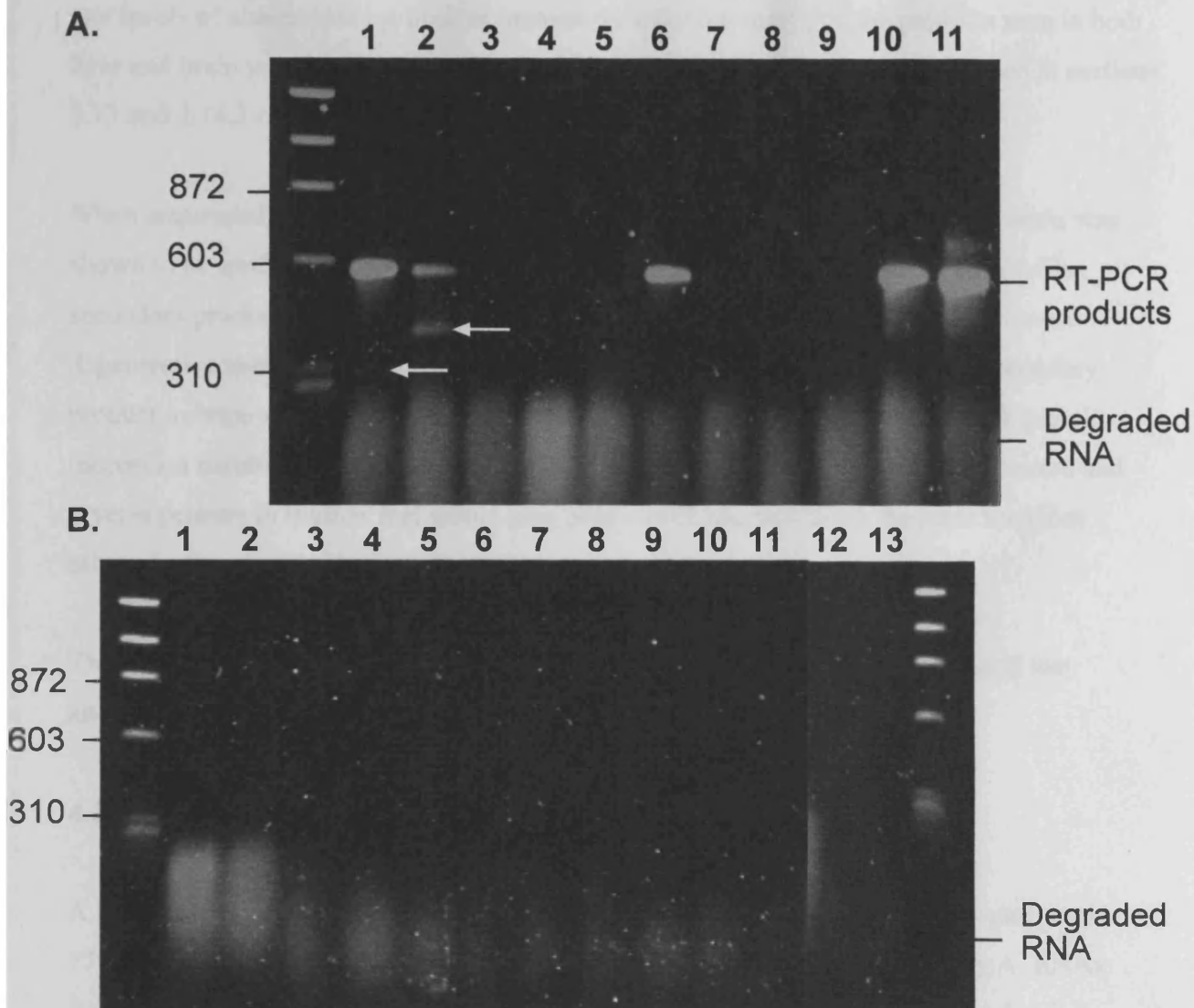


Figure 4.3. RT-PCR performed on RNA from adult mouse tissues.

RT-PCR was performed on 1 μ g of total RNA as described in section 2.13.3. The RNA had been extracted from adult mouse tissues (13 week mouse supplied by Carole Yauk, University of Leicester) using the method described in section 2.10.2.

The RT-PCR products were electrophoresed on a 1.3 % agarose gel. Gel A shows the RT-PCR products obtained in liver (lane 1), brain (lane 2), heart (lane 3), muscle (lane 4), testes (lane 5), diaphragm (lane 6), eye (lane 7), lung (lane 8), stomach (lane 9), kidney (lane 10) and a positive control (lane 11 - rat kidney).

Gel B shows the no RT negative controls; each lane has the same tissue as the corresponding lane in gel A. Lane 12 contains the no RNA negative control and lane 13 contains the PCR negative. All the negative controls are clear.

The sizes of three of the marker bands (ϕ X174 RF cut with *Hae*III) are indicated on the left hand side in kilobases.

The expected RT-PCR product (approximately 566 bp) and the degraded RNA are indicated on the right hand side of the gels. The white arrows on the top gel indicate the secondary products seen in liver and brain.

The band is a different size in both tissues. To investigate the possibility of the occurrence of low levels of alternatively spliced or incorrectly spliced transcripts, the products seen in both liver and brain were cloned into pGEM-7Zf (Promega) and sequenced as described in sections 2.13 and 2.14.2 respectively.

When sequenced, the RT-PCR product of the expected size (566 bp) in liver and brain was shown to be spp24. The secondary products seen in liver and brain were not spp24. The secondary product in liver showed significant homology to a protein similar to a mouse degenerative spermatocyte homologue (accession number AK002617) and the secondary product in brain showed significant homology to a protein similar to mouse SDP8 protein (accession number AK011257). Both proteins have high homology to both the forward and reverse primers in regions that would give products of approximately the sizes seen, but otherwise have no similarity with spp24.

The RT-PCR results suggest that in mouse there is a single transcript of spp24 and that alternative splicing does not occur.

4.2.3 Hybridisation of human *SPP2* cDNA to an RNA array

A multiple tissue expression (MTE) array was purchased from Clontech (catalogue number: 7775-1). The MTE array is a positively charged nylon membrane to which poly A⁺ RNAs from different human tissues, cancer cell lines and controls have been normalised and immobilised in separate dots. The poly A⁺ RNAs are guaranteed by Clontech to contain full-length transcripts, rare transcripts and be virtually free of contaminating genomic DNA. The array appears as a grid on the nylon membrane. Table 4.3 shows the layout of tissues in the squares of the MTE array grid.

The poly A⁺ RNAs on the MTE array have been normalised to the mRNA expression levels of eight housekeeping genes. This minimises the small tissue-specific variations in expression of any one housekeeping gene. It is therefore possible to quantitate the levels of expression relative to other tissues, but it is not possible to obtain an absolute value.

The human MTE array was chosen instead of a northern blot due to the much larger number of tissues it contained. The EST data and the RT-PCRs in mouse provided evidence for a single spp24 transcript. Therefore, the size of the transcript is already known and any small differences in length are unlikely to be resolved on a northern blot.

Table 4.3. The layout of the human Clontech MTE array

The MTE array is a positively charged nylon membrane to which poly A⁺ RNAs from different human tissues, cancer cell lines and controls have been normalised and immobilised in separate dots. The poly A⁺ RNAs on the MTE array have been normalised to the mRNA expression levels of eight housekeeping genes. The array appears as a grid on the nylon membrane. The table below shows the layout of tissues in each square of the grid. The blank squares are squares that have been left blank on the array for orientation.

	1	2	3	4	5	6	7	8	9	10	11	12
A	Whole brain	Cerebellum, left	Substantia nigra	Heart	Esophagus	Colon, transverse	Kidney	Lung	Liver	Leukemia, HL-60	Foetal brain	Yeast total RNA
B	Cerebral cortex	Cerebellum, right	Nucleus accumbens	Aorta	Stomach	Colon, descending	Skeletal muscle	Placenta	Pancreas	HeLa S3	Foetal heart	Yeast tRNA
C	Frontal lobe	Corpus callosum	Thalamus	Atrium, left	Duodenum	Rectum	Spleen	Bladder	Adrenal gland	Leukemia, K-562	Foetal kidney	<i>E. coli</i> rRNA
D	Parietal lobe	Amygdala	Pituitary gland	Atrium, right	Jejunum		Thymus	Uterus	Thyroid gland	Leukemia, MOLT-4	Foetal liver	<i>E. coli</i> DNA
E	Occipital lobe	Caudate nucleus	Spinal cord	Ventricle, left	Ileum		Peripheral blood leukocyte	Prostate	Salivary gland	Burkitt's lymphoma, Raji	Foetal spleen	Poly r(A)
F	Temporal lobe	Hippocampus		Ventricle, right	Ileocecum		Lymph node	Testis	Mammary gland	Burkitt's lymphoma, Daudi	Foetal thymus	Human C ₀ t-1 DNA
G	Paracentral gyrus of cerebral cortex	Medulla oblongata		Inter-ventricular septum	Appendix		Bone marrow	Ovary		Colorectal adenocarcinoma, SW480	Foetal lung	Human DNA 100 ng
H	Pons	Putamen		Apex of the heart	Colon, ascending		Trachea			Lung carcinoma, A549		Human DNA 500 ng

Twenty nanograms of the human cDNA encoding the mature protein was used to make a ³²P-labelled probe as described in section 2.10.1.2, Chapter 2 to 2.10.1.3, Chapter 2. The probe was then hybridised to the Clontech MTE array as described in section 2.10.3, Chapter 2.

Figure 4.4 shows the autoradiograph of the hybridised MTE array after a 24-hour and a 1-week exposure. Hybridisation to the MTE array confirmed the tissue-specific expression previously suggested. The MTE array demonstrated that spp24 is not expressed in many human tissues. In fact, the only tissues that were positive were liver, foetal liver and foetal kidney. Liver and foetal liver gave the strongest hybridisation signal and foetal kidney was relatively weak in comparison.

Weak signals were seen in the human DNA dots (G12 and H12), as expected with a cDNA probe. A signal comparable with that of the foetal kidney was also seen in *E. coli* DNA. This is probably due to contamination of the probe with small amounts of *E. coli* DNA as the probe was prepared from a recombinant plasmid propagated in *E. coli*. Likewise, there is a very faint signal seen in the *E. coli* rRNA.

To try and quantify the relative level of expression between tissues, a phosphorimage was taken with a 24-hour exposure (section 2.10.3.5, Chapter 2). The results of this are shown in table 4.4.

As expected, the lowest values are seen in the *E. coli* rRNA and the human DNA. The value seen in the *E. coli* DNA is quite high, approximately 10 times that of the *E. coli* rRNA. This indicates that there is substantial contamination of the probe with *E. coli* DNA. However, the contaminating DNA seems to be specific to *E. coli* as it does not cause hybridisation to the yeast control and does not interfere with the human RNAs, *i.e.* no non-specific background is seen.

Expression is seen in foetal kidney, but at a relatively low level comparable with the signal seen in the *E. coli* DNA. The highest expression is seen in liver and foetal liver. Expression is slightly higher in adult liver, but both values are approximately 11-fold higher than the expression seen in foetal kidney.

The human MTE array results suggest that human *SPP2* is expressed mainly in the liver, but is also expressed at a much lower level in the developing kidney.

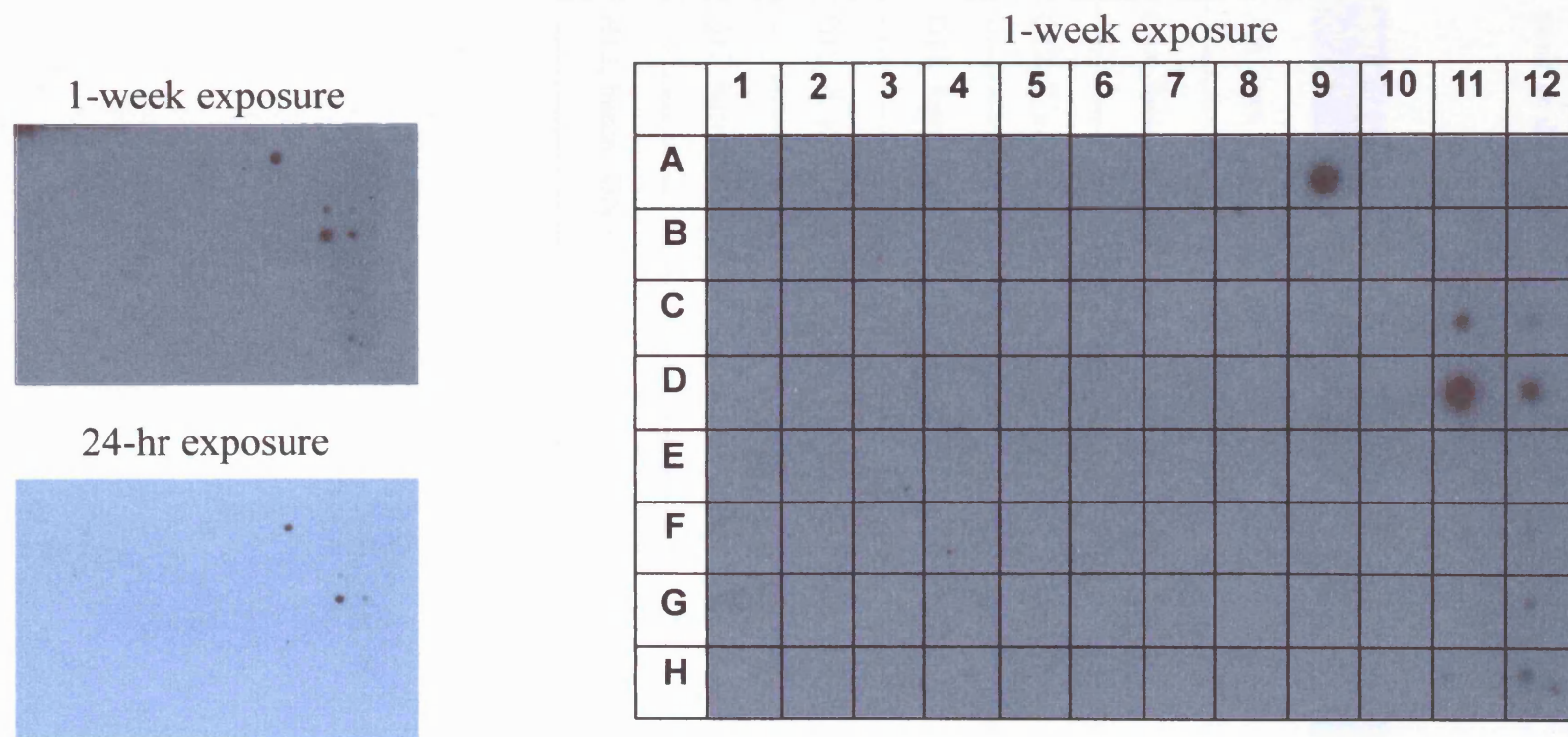


Figure 4.4. The autoradiographs of the human *SPP2* cDNA hybridised to the Clontech human MTE array.

Twenty nanograms of the human cDNA encoding the mature protein was used to make a 32 P-labelled probe as described in section 2.12.1.2 to 2.12.1.3. The probe was then hybridised to the Clontech MTE array as described in section 2.12.3.

The autoradiograph overlaid on the grid is the 1-week exposure. The two smaller images at the left hand side of the grid show the autoradiograph after a 24-hour and a 1-week exposure.

A positive signal appears in squares A9, C11, C12, D11, D12, G12 and H12. These correspond to liver, foetal kidney, *E. coli* rRNA, foetal liver, *E. coli* DNA, human DNA 100 ng and human DNA 500 ng respectively (see table 4.3).

Table 4.4. The results of a phosphorimage from a 24-hour exposure of the Clontech human MTE array, hybridised with human *SPP2* cDNA.

Positive signals were seen in A9, C11, D11, D12, G12 and H12, corresponding to liver, foetal kidney, foetal liver, *E. coli* DNA, human DNA 100 ng and human DNA 500 ng respectively. ImageQuant was used to compare the intensity of each signal to the local background. An arbitrary unit value was assigned to each spot enabling a comparison to be made between positive dots.

MTE array dot	Value compared to local background
A9, liver	41802
C11, foetal kidney	3703
C12, <i>E. coli</i> rRNA	455.4
D11, foetal liver	40766
D12, <i>E. coli</i> DNA	4825
G12, human DNA 100 ng	869.2
H12, human DNA 500 ng	982.5

4.2.4 Expression data for spp24 in mouse from the RIKEN READ database

Miki *et al.* (2001) recently reported the development of an expression database using microarray technology. The RIKEN mouse cDNA libraries (Carninci and Hayashizaki 1999), which were enriched for full-length cDNAs, were used to collect target cDNAs. A total of 18,816 unique cDNA clones were then arrayed and it was estimated that the array contained approximately 13,600 non-redundant genes. The expression profile with respect to these genes was determined for 49 adult and embryonic mouse tissues.

The expression level in each tissue was compared to the cDNA from pooled male and female 17.5-day embryos. This was chosen as a reference as the 17.5-day embryo has a relatively complex expression pattern and it was thought to be easily reproducible.

A web-based database search engine was developed by Miki *et al.* (2001) named READ (RIKEN cDNA Expression Array Database) (<http://genome.gsc.riken.go.jp/READ/>). This was searched for secreted phosphoprotein 24 and a clone was identified (RIKEN ID 1600023D11) that was highly similar to *Rattus norvegicus* spp24. The sequence of the EST was checked against the mouse spp24 cDNA sequence determined in chapter 3 and it was concluded that the RIKEN clone was mouse spp24. Table 4.5 shows the microarray expression data obtained for mouse spp24. All values are relative to a 17.5-day embryo and are given as a log-transformed (base 2) ratio value.

The RIKEN microarray results indicate that spp24 is expressed in kidney, liver, placenta, 10-day lactating mammary gland, thymus, 10-day neonate cerebellum, Sv40t, muscle, whole embryo days 10, 11 and 13 and the liver of a 13-day embryo. The results seen in liver and kidney were expected, but the positive results seen in cerebellum, placenta, lactating mammary gland, thymus and muscle were a little unexpected.

The Sv40t is a liver tumour sample from a transgenic mouse which harbours the SV40 virus under the control of the MUP (major urinary protein) promoter. The results therefore show that compared to 17.5-day embryo, the expression of spp24 is slightly up-regulated in a liver tumour sample, but is down-regulated relative to normal liver.

Table 4.5. The microarray expression data obtained for mouse spp24 from READ (RIKEN cDNA Expression Array Database) (<http://genome.gsc.riken.go.jp/READ/>).

READ was searched for secreted phosphoprotein 24. A clone was identified (RIKEN ID 1600023D11) that was highly similar to *Rattus norvegicus* spp24. The sequence of the EST was checked against the mouse spp24 cDNA sequence determined in chapter 3 and it was concluded that the RIKEN clone was mouse spp24. All values are relative to a 17.5-day embryo and are given as a log-transformed (base 2) ratio value.

The boxes giving the expression values are coloured to indicate whether the gene is up-regulated or down-regulated relative to the 17.5-day embryo reference and shading of the colour depicts the extent of divergence from the reference. A green box indicates that the gene is down regulated compared to the reference and a red box indicates that a gene is up regulated. Black boxes indicate no difference in the level of expression between this tissue and the 17.5-day embryo and white boxes indicate that for some reason there is no result for this tissue with respect to that clone.

Table 4.5. The microarray expression data obtained for mouse spp24 from READ (RIKEN cDNA Expression Array Database) (<http://genome.gsc.riken.go.jp/READ/>).

RIKEN CloneID	Kidney	Brain	Spleen	Heart	Lung	Liver	Cerebellum	Placenta	Testis	Pancreas	Small intestine	Stomach
1600023D11	0.955	-2.236	-2.359	-1.04	-2.155	2.315	-2.082	3.302	-0.427	-0.612	-2.227	-1.84

Tongue	Embryo 13 liver	Embryo 10	Embryo 11	Embryo 12 head	Embryo 13 head	Embryo 17 head	Embryo 13	Embryo 15 head	Embryo 16 head	Thymus preg 1 day	Embryo 14 liver	10 day lactating mammary gland
-2.451	0.592	0.724	0.325	-0.556			0.346	-1.519		-1.332	-2.141	2.813

Skin neonate 0 day	Skin neonate 10 day	Ovary uterus preg 11 days	Intestine neonate 10 day	Thymus	Embryo 11 head	Medulla oblongata	Olfactory brain	Cerebellum neonate 10 day	Embryo 12 wolffian duct	Eyeball	Cortex	Vesicular
	-0.646			1.45	-1.032	-1.719		2.248	-1.206	-2.33	-1.26	-0.591

Uterus	Embryo 16 lung	Colon	Cecum	Bone	Sv40t	Lung neonate 0 day	Muscle	Neonate 0 day whole head	Neonate 6 day whole head	Neonate 10 day whole head	Description
-1.521	-1.286	-1.864			0.776		1.267			-0.975	ESTs, Highly similar to Secreted phosphoprotein 24 [<i>R. norvegicus</i>]

The highest levels of expression were seen in liver, placenta, lactating mammary gland and 10-day neonate cerebellum. These values were approximately 2-fold to 8-fold higher than the values seen in the other tissues giving positive results.

4.2.5 Microarray results from Incyte Genomics Inc. with respect to spp24 and osteoblasts

Incyte Genomics Inc. is a company that has performed many experiments using human microarrays and is making the results available for purchase. The majority of experiments concentrate on comparing gene expression in normal and abnormal tissues and the effect on gene expression of drugs and chemical treatments.

The data are stored in an online database called LifeExpress Online (www.incyte.com/lifeexpress/). LifeExpress Online was searched for secreted phosphoprotein 24. Spp24 was located on microarray GEM-1 and a series of experiments was identified as being relevant to the *SPP2* gene. One of these was chosen for purchase.

The data purchased were from an experiment using RNA from osteoblasts as a probe. Due to the original isolation of spp24 in bone (Hu *et al.* 1995), this experiment was thought to be particularly relevant.

Osteoblasts are bone-forming cells derived from pluripotent mesenchymal stem cells. Osteogenic stimulation causes mesenchymal stem cells to differentiate into osteoblast precursor cells. Further differentiation then causes these precursor cells to develop into mature osteoblasts that secrete type I collagen and other non-collagenous bone matrix proteins.

The LifeExpress Online experiment compared the hybridisation to a human *SPP2* cDNA clone of RNA from osteoblast precursor cells isolated from long bones and RNA from osteoblasts that have begun to differentiate and secrete matrix proteins. The osteoblasts were stimulated to mature by a switch from osteoblast growth basal media into osteoblast differentiation media containing hydrocortisone and beta-glycerophosphate.

Table 4.6 shows the results that were purchased from Incyte Genomics Inc. Unfortunately, the results in the table of data that was purchased were 'grayed out'. This indicates that there is

Table 4.6. A comparison of the expression level of *SPP2* between osteoblast precursor cells and mature osteoblasts. Data purchased from Incyte Genomics Inc.

The LifeExpress Online experiment compared the hybridisation to *SPP2* of RNA from osteoblast precursor cells isolated from long bones and RNA from osteoblasts that have begun to differentiate and secrete matrix proteins. The osteoblasts were stimulated to mature by a switch from osteoblast growth basal media into osteoblast differentiation media containing hydrocortisone and beta-glycerophosphate.

The RNA from osteoblast precursor cells was labelled with Cy3 and the RNA from differentiated osteoblast cells was labelled with Cy5. The differential expression value is the value that indicates the fold difference between the Cy3 and Cy5 probes. A negative differential expression value indicates a down-regulation and a positive value an up-regulation.

The experiment was performed twice and both results are given in the table below.

Hybridisation name	Probe name	Differential expression value
NHStCells,t/ GrowthMedia/t/DiffM	Cy3: Human, NHSt Cells, t/Growth Media, 3d, Nrml	-1.01
	Cy5: Human, NHSt Cells, t/DiffM, 3d, Nrml	
NHStCells,t/ GrowthMedia/t/DiffM	Cy3: Human, NHSt Cells, t/Growth Media, 3d, Nrml	1.16
	Cy5: Human, NHSt Cells, t/DiffM, 3d, Nrml	

either low confidence in the quality of the data or that the gene is not expressed significantly in the sample that was used as a probe.

The quality of the hybridisation was checked with Incyte Genomics Inc. and was found to be of an adequate standard with respect to other clones. The *SPP2* clone was shown to be of a high standard with respect to other hybridisations and so it is likely that *SPP2* is simply not expressed in significant levels in either osteoblast precursor cells or mature osteoblasts.

4.3 Discussion

Table 4.7 summarises the expression data obtained for humans with respect to spp24 from three different sources. There are no conflicting results from the different sources and so it is possible to conclude that in humans spp24 is expressed in liver, foetal liver and foetal kidney, but not in all other the other tissues and cell types listed in table 4.7. A total of 77 different tissues were investigated and so it can be concluded that in humans, spp24 displays tissue specific expression.

Table 4.8 summarises the expression data obtained for the mouse with respect to spp24 from three different sources. However, the mouse data show some conflicting results. The RT-PCR results indicate that spp24 is expressed in the mouse adult brain but the READ microarray results give a negative result in this tissue. However, the READ microarray results show that spp24 is expressed in the cerebellum (posterior region of brain) in a 10-day neonate mouse, but not in an adult mouse. This suggests that spp24 is expressed in the cerebellar region of the brain at a specific stage of brain development in the infant mouse.

Fetuin is a protein, discussed as being similar to spp24 in Chapter 1, which is expressed mainly in the liver but also shows some expression in the brain. The human fetuin α_2 HS-glycoprotein is expressed in the cortical plate neurons of the neocortex in the developing embryonic brain (Dziegielewska *et al.* 1987). However, in the adult brain the protein can no longer be detected and it is thought that this is due to death of the cell population rather than loss of expression of the α_2 HS-glycoprotein (Saunders *et al.* 1992).

The expression data for spp24 in the mouse brain therefore suggest that in a similar way to α_2 HS-glycoprotein, spp24 may be expressed by a specific cell population that is formed at a particular developmental stage in the infant mouse cerebellum. If this cell population were to then die, spp24 would not be detected in this tissue in the adult mouse.

The detection of spp24 in the 13-week adult mouse brain by RT-PCR could be explained by the sensitivity of the technique. RT-PCR will detect minute amounts of a transcript and so if a small amount of the cell population was still present in the adult tissue, RT-PCR might detect spp24 expression when hybridisations would not.

Table 4.7. A summary of the spp24 expression data obtained for humans.

This table lists all of the tissues or cell types for which expression information was available with respect to spp24 in humans. The table states whether spp24 was expressed for that particular tissue or cell type in the human ESTs, the Clontech human MTE array and the Incyte Genomics Inc. results, with a simple yes or no. A dash indicates that data for this tissue or cell type were not available from that source.

Tissue/cell type	Human ESTs	Clontech human MTE array	Incyte Genomics Inc. data
Liver	Yes	Yes	-
Foetal liver	Yes	Yes	-
Foetal kidney	-	Yes	-
Muscle	-	No	-
Adult brain and individual regions	-	No	-
Adult heart and individual regions	-	No	-
Regions of adult digestive system	-	No	-
Kidney	-	No	-
Spleen	-	No	-
Thymus	-	No	-
Peripheral blood leukocyte	-	No	-
Lymph node	-	No	-
Bone marrow	-	No	-
Trachea	-	No	-
Lung	-	No	-
Placenta	-	No	-
Bladder	-	No	-
Uterus	-	No	-
Prostate	-	No	-
Testis	-	No	-
Ovary	-	No	-
Pancreas	-	No	-
Adrenal gland	-	No	-
Thyroid gland	-	No	-
Mammary gland	-	No	-
Salivary gland	-	No	-
Foetal brain	-	No	-
Foetal heart	-	No	-
Foetal spleen	-	No	-
Foetal thymus	-	No	-
Foetal lung	-	No	-
Leukaemia, HL-60 cell line	-	No	-
HeLa S3 cell line	-	No	-
Leukaemia K-562 cell line	-	No	-
Leukaemia MOLT-4 cell line	-	No	-
Burkitt's lymphoma Raji cell line	-	No	-
Burkitt's lymphoma Daudi cell line	-	No	-
Colorectal adenocarcinoma SW480	-	No	-
Lung carcinoma A549	-	No	-
Osteoblast cells	-	-	No

Table 4.8. A summary of the spp24 expression data obtained for the mouse.

This table lists all of the tissues for which expression information was available with respect to spp24 in the mouse. The table states whether spp24 was expressed for that particular tissue in the mouse ESTs, the RT-PCRs and the READ microarray results, with a simple yes or no. A dash indicates that data for this tissue or cell type was not available from that source. Suggested indicates a positive result that cannot be reliably concluded.

Tissue	Mouse ESTs	RT-PCRs	READ microarray results
Liver	Yes	Yes	Yes
Kidney	Yes	Yes	Yes
Brain	-	Yes	No
Spleen	-	-	No
Heart	-	No	No
Lung	-	No	No
Cerebellum	-	-	No
Placenta	Suggested	-	Yes
Testis	-	No	No
Pancreas	-	-	No
Small intestine	-	No	No
Stomach	-	-	No
Tongue	-	-	No
Embryo 13 day liver	-	-	Yes
Embryo 10 day	-	-	Yes
Embryo 11 day	-	-	Yes
Embryo 12 day head	-	-	No
Embryo 13 day	Suggested	-	Yes
Embryo 15 day head	-	-	No
Thymus pregnancy day 1	-	-	No
Embryo 14 day liver	-	-	No
Mammary gland lactate day 10	-	-	Yes
Skin neonate day 10	-	-	No
Thymus	-	-	Yes
Embryo day 11 head	-	-	No
Medulla oblongata	-	-	No
Cerebellum neonate day 10	-	-	Yes
Embryo day 12 wolffian duct	-	-	No
Eyeball	-	No	No
Cortex	-	-	No
Vesicular	-	-	No
Uterus	Suggested	-	No
Embryo day 16 lung	-	-	No
Colon	Suggested	-	No
Sv40t	-	-	Yes
Neonate day 10 whole head	-	-	No
Muscle	-	No	Yes
Diaphragm	Suggested	Yes	-
Macrophage	Suggested	-	-
T-cell	Suggested	-	-
19.5 day total foetus	Suggested	-	-
Embryonic carcinoma	Suggested	-	-

The mouse expression data for spp24 also shows some conflicting results for the uterus and the colon. A single EST was obtained from the uterus and the colon, but both these tissues gave a negative value in the READ microarray results. The ESTs must be used with caution as the quality can sometimes be dubious and the fact that there is only a single EST for each immediately suggests that spp24 may not actually be expressed in that tissue.

It has been suggested that many 'tissue-specific' genes may be expressed at a 'basal' rate in many cell types (Sarkar and Sommer 1989) (Linsk *et al.* 1989). This is often referred to as 'ectopic expression' and was suggested by Linsk *et al.* (1989) to be a mechanism by which T-cells become tolerant to 'self' tissue-specific proteins. The genes are thought to be expressed transiently and the translated protein rapidly catabolised to peptides.

Sommer and Sarkar (1989) suggested that other consequences of ectopic expression could be the predisposition of cells to neoplasia or metastasis or a certain rate of endogenous tissue injury due to expression of genes that may be deleterious to a particular cell. The spp24 ESTs seen in uterus and colon could be due to the phenomenon of ectopic expression and therefore these tissues are not a true expression site for spp24.

A further conflict of results is seen between the RT-PCR of muscle and the READ microarray result for muscle. The READ microarray gives a positive result for the expression of spp24 in muscle, but the RT-PCR gives a negative result. It is not possible to draw any conclusions from this. It could be that the RNA quality of the sample from muscle was poor in the RT-PCRs or that the READ microarray results are incorrect. The fact that spp24 expression is not seen in any of the other muscular tissues such as heart and tongue suggests that the READ microarray result may be incorrect.

From the mouse spp24 expression data it is therefore possible to conclude that spp24 is expressed in mouse liver, kidney, foetal liver, lactating mammary gland day 10, placenta, thymus, diaphragm and the cerebellum of an infant mouse. It may also be expressed in uterus, colon and muscle, but further evidence is required to resolve the conflicting results seen.

A small number of ESTs also suggest that spp24 may be expressed in macrophages and T-cells, but these cell types were not analysed by any other method and so the data cannot be considered reliable evidence on their own. However, a proportion of the chicken ESTs from what is thought to be the chicken spp24 orthologue (Chapter 6) also originated from

T-cell-enriched splenocytes. This suggests that spp24 may have some involvement in the immune system.

Table 4.9 compares the positive expression data obtained for spp24 from three species, human, mouse and bovine to see what can be deduced regarding an overall expression profile for spp24.

The only tissue that gives a positive result with respect to spp24 expression in all three species analysed is liver. The bovine northern blots performed by Hu *et al.* (1995) showed that spp24 was expressed at a high level in liver relative to bone, the tissue from which the protein was originally isolated. The human Clontech MTE microarray data also shows that spp24 is expressed at a high level relative to foetal liver and foetal kidney and the READ microarray results show that it is expressed at a high level relative to a 17.5-day embryo. It can therefore be concluded that in all species analysed, spp24 is expressed in the liver at high levels.

The human Clontech MTE microarray data and the READ microarray results show that spp24 is also expressed in foetal liver, but at lower levels than that seen in the adult tissues. Much of the expression seen in the whole embryos of the READ microarray data can probably be attributed to foetal liver (days 10 to 14). It is speculated that if the bovine northern blot had included foetal liver, a positive result would have been seen.

The data regarding spp24 expression in the kidney suggest a possible difference in expression between species. Spp24 expression was seen in the adult kidney of mouse, but not human and cattle. However, expression was seen in the human foetal kidney, which was unfortunately not analysed in mouse and cattle.

The timing of spp24 expression in the kidney could be different in humans and cattle compared with mouse or if spp24 is expressed in the foetal kidney of all three species, the expression of spp24 may persist into the mature mouse, but not the mature human or cattle.

It can be concluded from the READ microarray results that spp24 is expressed at a particular developmental stage of the mouse infant cerebellum. However, this tissue was not tested in the other species. It is speculated that a similar expression pattern would be seen. The Clontech human MTE array contained samples from adult brain, adult cerebellum and foetal brain (all of which gave a negative result), but unfortunately did not contain infant cerebellum.

Table 4.9. Spp24 expression data from human, mouse and bovine tissues.

The conclusions regarding spp24 expression for human, mouse and bovine tissues (from results in Table 4.7, Table 4.8 and by Hu *et al.* 1995) are summarised in this table. A dash indicates that there are no data regarding that tissue in that particular species. The word 'suggested' indicates that there was some evidence for expression in that tissue in that species, but that no firm conclusion could be drawn due to the quality of evidence or conflicting results. The words 'yes' or 'no' indicate expression or no expression respectively.

Tissue	Human	Mouse	Bovine
Liver	Yes	Yes	Yes
Foetal liver	Yes	Yes	-
Kidney	No	Yes	No
Foetal kidney	Yes	-	-
Brain- cerebellum (infant)	-	Yes	-
Bone	-	-	Yes
Diaphragm	-	Yes	-
Placenta	No	Yes	-
Thymus	No	Yes	-
Mammary gland lactate day 10	-	Yes	-
Uterus	No	Suggested	-
T-cell	-	Suggested	-
Macrophage	-	Suggested	-
Colon	No	Suggested	-
Muscle	No	Suggested	-

It is possible to say that spp24 is expressed in bone, as the protein was originally isolated from the bovine tissue (Hu *et al.* 1995) and a bovine northern blot gave a positive result. It is unfortunate that bone is a difficult tissue from which to obtain RNA and so the expression of spp24 in bone has not been tested in other species. It is speculated that a similar level of expression would be seen as the presence of the protein in bone suggests a role in bone processes.

The expression of spp24 was also seen in mouse diaphragm and mouse lactating mammary gland, but again these tissues were not tested in the other species. The Clontech human MTE array contained a sample from mammary gland, but not from lactating mammary gland. The expression of spp24 in mouse lactating mammary gland is interesting as it suggests that the protein may be found in milk.

Spp24 expression was also seen in mouse thymus and placenta at high levels, but a negative result was seen in the human tissues. This could be evidence of incorrect results or other potential species differences.

It is also possible that spp24 is expressed in uterus, T-cells, macrophages, colon and muscle, but there is not enough evidence to reach a firm conclusion.

In summary, spp24 is expressed in the liver of the foetus and the adult at high levels. It is also expressed in the kidney with a possible difference in the developmental timing of expression between species. Spp24 is expressed at a particular stage of development in the mouse infant cerebellum and is probably expressed in a similar way in the infant cerebellum of other species.

The spp24 mRNA and protein is found in bovine bone and so is likely to also be found in bone in other species. The same is true of mouse lactating mammary gland and mouse diaphragm. Spp24 is expressed in these mouse tissues and therefore expression is likely to be seen in this tissue in other species. There is a possible species difference in the expression of spp24 in thymus and placenta, or these results may be incorrect. It is possible that spp24 is expressed in several other tissues, but there is insufficient evidence to substantiate this.

From the general expression profile generated for spp24 it is possible to speculate on some potential functions of the spp24 protein. The high level of expression seen in both foetal and

adult liver suggests that the protein either has an essential role in liver function or that it is a plasma protein that is synthesised in the liver.

Fetuin, a protein with a similar overall structure to spp24, is expressed at high levels in the liver where it is synthesised before circulating in the plasma. Fetuin is thought to have a role in the acute phase response (reviewed by Brown *et al.* 1992). Many of the non-collagenous bone matrix proteins are also synthesised in the liver (Chapter 1, table 1).

The expression of spp24 in a specific developmental stage of the infant cerebellum suggests that spp24 may have a function in the formation of cells that arise at this stage in cerebellar development. A similar phenomenon is seen with fetuin, which is expressed in the developing neuronal cells in the cortical plate of the neocortex in the embryonic brain (Dziegielewska *et al.* 1987). As is seen with the human fetuin α_2 HS-glycoprotein (Saunders *et al.* 1992), death of the cell population expressing spp24 at this stage of cerebellar development could explain why no expression is seen in adult cerebellum.

Hu *et al.* (1995) speculated about a role for spp24 in the process of bone turnover due to the isolation of the protein from bovine bone and a positive bovine northern blot result. The human microarray results from Incyte Genomics Inc. show that spp24 is not expressed by osteoblasts at any significant level when they are precursor cells or when they are mature osteoblasts. This supports the speculation by Hu *et al.* (1995) that the spp24 in bone may be expressed by osteoclasts as this is where some possible thiol proteinase target proteins may be expressed.

However, Kobori *et al.* (1998) report the isolation of osteoclast-specific genes in the rabbit by the preparation of a subtracted cDNA library. A total of 424 novel cDNAs were identified and deposited in the DDJB/EMBL/GenBank data bank with the accession numbers C84253-C84676 (Kobori *et al.* 1998). Spp24 does not appear in these sequences neither does it appear in the known genes that they identified (Kobori, personal communication to R. Dalgleish). This suggests that the spp24 expression in bone reported by Hu *et al.* (1995) is from a cell type present in bone other than osteoblasts or osteoclasts.

The expression of spp24 in mouse placenta and lactating mammary gland suggests a possible antimicrobial function. Immunity can be passed from a mother to a foetus via the placenta and also through milk. This speculated function can be supported by the fact that spp24 shows some homology to the bovine neutrophil antibiotic peptide batenecin precursor (Hu *et al.*

1995) and that the C-terminal non-cystatin-like domain of spp24 is quite different between species (Chapter 3), unlike the highly conserved cystatin domain.

Using READ it was possible to search for genes with a similar expression profile to spp24. A search was performed for genes expressed in liver, kidney, placenta, lactating mammary gland and bone. Bone was included to narrow down the search as spp24 was shown to be expressed in bovine bone and the protein present by Hu *et al.* (1995). Thirty-six genes were identified that showed expression in all of these tissues. Many of these genes showed ubiquitous expression or expression through most of the major organs. However, 6 were identified that showed a similar tissue-specific expression pattern to spp24. A simplified comparison of each of their expression profiles to that of spp24 is shown in figure 4.5.

The clones identified as having a similar expression pattern to spp24 included, three ESTs of unknown identity (Riken ID 1700019K03, 2010110K18 and 2010009O05), one gene displaying some similarity to rat corticosteroid dehydrogenase (Riken ID 1600012F10), one gene identified as encoding the mouse biotinidase precursor (Riken ID 1600020N20) and finally a gene that is similar to the house mouse MAP kinase (Riken ID 2510027C03).

The similarity in their expression profiles to spp24 could simply be coincidence. Alternatively, it could be that the presence of these proteins is necessary for the function of mature spp24 *e.g.* a kinase could be present in the same tissues as spp24 needs to be phosphorylated. The clone with the most similar expression profile to spp24 was the biotinidase precursor. Biotinidase recycles biotin, which is a coenzyme for several carboxylases, but its significance, if any, with respect to spp24 expression is not clear.

In summary, the tissue-specific nature of spp24 expression and the diversity of the tissues it is expressed in suggest that spp24 has specific, multiple functions. These may include a role in liver function, a plasma protein function, a role in the immune response, a role in bone turnover and an antimicrobial function.

Figure 4.5. Comparison of genes with similar expression profiles to spp24 that were identified using READ.

The clones from READ are identified by their Riken ID, with the exception of spp24, which is simply named. The expression profile of each clone is shown against that of spp24. The tissues running from left to right are:

Kidney, brain, spleen, heart, lung, liver, cerebellum, placenta, testis, pancreas, small intestine, stomach, tongue, embryo 13-day liver, embryo 10-day, embryo 11-day, embryo 12-day head, embryo 13-day head, embryo 17-day head, embryo 13-day, embryo 15-day head, embryo 16-day head, thymus 1-day pregnancy, embryo 14-day liver, mammary gland lactate 10-day, skin neonate 0-day, skin neonate 10-day, ovary and uterus 11-day pregnancy, intestine neonate 10-day, thymus, embryo 11-day head, medulla oblongata, cerebellum neonate 10-day, embryo 12-day wolffian duct, eyeball, cortex, vesicular, uterus, embryo 16-day lung, colon, cecum, bone, sv40t, lung neonate 0-day, muscle, neonate day-0 whole head, neonate day-6 whole head and neonate day-10 whole head.

The exact expression values are not given. A red box containing a 1 indicates that expression was up-regulated in this tissue compared to 17.5-day embryo. A green box containing a 0 indicates that expression was down-regulated in this tissue compared to 17.5-day embryo. A white box indicates that either no difference in expression was seen compared to a 17.5-day embryo or that the hybridisation did not work.

Figure 4.5. Comparison of genes with similar expression profiles to spp24 that were identified using READ.

Spp24 and 1600020N20 (Biotinidase precursor)



Spp24 and 2510027C03 (House mouse mRNA for MAP kinase, kinase 3b)



Spp24 and 1600012F10 (EST)



Spp24 and 1700019K03



Spp24 and 2010110K18 (EST)



Spp24 and 2010009O05 (Similar to purine nucleoside phosphorylase (mouse))



Chapter 5

A comparison of the spp24 protein between species

5.1 Introduction

A comparison between species can highlight regions of a protein that are highly conserved and which are therefore likely to be crucial to its function, with residues that are identical between species obviously likely to be the most critical.

The spp24 proteins that exist in the Swissprot database are that of rat (accession number Q62740), bovine (accession number Q27967) and human (accession number Q13103). The bovine and human proteins include the signal peptide, but the rat protein does not.

The rat spp24 cDNA exists in GenBank (accession number U19485), but in an attempt to obtain a longer rat cDNA and therefore be able to determine the sequence of the signal peptide, rat ESTs were identified and aligned to generate a consensus cDNA. This chapter discusses an anomaly seen in rat ESTs that could be evidence of possible exon skipping or missplicing.

BLAST searches of the Swissprot database using the human spp24 protein revealed homology to a chicken hypothetical protein (accession number Q91982). This chapter presents evidence suggesting the published sequence of this protein is incorrect and in fact the protein could be the chicken counterpart of spp24.

This chapter also presents the determination of the mouse, pig and chicken spp24 protein sequences.

The work described in this chapter enabled six proteins in total to be aligned (rat, bovine, human, mouse, pig and chicken) that were all thought to be spp24 or very closely related to spp24. Consequently two protein representations could be produced; a consensus spp24 protein and a protein representation showing the residues that were identical between species. These representations may help elucidate the residues that are critical to the function of the protein.

5.2 Results

5.2.1 An anomaly observed in rat spp24 ESTs

Table 5.1 presents details of the rat ESTs identified by searching the rat UniGene database (www.ncbi.nlm.nih.gov/UniGene/) for spp24. The rat ESTs are part of the UniGene spp24 cluster, Rn.84. The TIGR rat gene index (www.tigr.org/tdb/rgi) was also searched, but was not found to contain any ESTs additional to those found in the UniGene database. The ESTs enabled a signal peptide to be deduced for rat spp24 and this is included in the protein alignment discussed in section 5.2.4.

Some of the rat ESTs identified appeared to be missing the whole of the region corresponding to exon 4 in the human and mouse genes (Chapter 3) and/or part of the region corresponding to human and mouse exon 6 (Chapter 3). The rat ESTs showing this anomaly are not all from the same source. Figure 5.1 shows the parts of the protein that would be missing if these ESTs were translated.

Original sequence chromatograms were not available for most of the ESTs. The only ESTs for which the chromatograms were obtained were AA858573 and AI043655 which were missing the exon 4 region plus part of 6 and just part of the exon 6 region respectively. The chromatograms appeared to support the EST sequences present in the database.

It was speculated that the missing regions might be strain specific. The ESTs with missing regions were either from the rat strain Sprague-Dawley or the source simply stated as *Rattus norvegicus*. To investigate this phenomenon in rat strains other than Sprague-Dawley, primers were designed for RT-PCR. The sequence of the primers is shown below:

Forward: 5'-TATGAATTCAGAGTCTGGTGATCCCTCCA-3'

Reverse: 5'-AATGGATCCTTGACTCTTGCTCTGCGTTG-3'

The primers lie either side of the regions that correspond to exon 4 and exon 6 in the human and mouse cDNA (Chapter 3). An *Eco*RI and a *Bam*HI restriction enzyme site was incorporated into the forward and reverse primers respectively for ease of cloning, should it be required. These sites are underlined in the primer sequences.

Table 5.1. Rat ESTs from the UniGene cluster Rn.84.

The rat UniGene spp24 cluster was identified (Rn.84) by searching the UniGene database (www.ncbi.nlm.nih.gov/UniGene/) with the keywords 'secreted phosphoprotein 24'. The cluster comprises 9 ESTs, the details of which are given in the table. A dash in the exon 4 or part of exon 6 column shows that the EST did not cover this region. 'Present' indicates that this region was present as expected and 'Missing' indicates that the region was absent.

Accession Number	Source and Tissue	Strain	Exon 4	Part of exon 6
BF550478	University of Iowa Embryonic	Sprague Dawley	-	-
AW862673	University of North Carolina Medical School Liver	Sprague Dawley	-	-
AW921840	The Institute for Genomic Research, Rockville Tissue mix	Only information given was <i>Rattus norvegicus</i>	Missing	Missing
BF282482	The Institute for Genomic Research, Rockville Tissue mix	Only information given was <i>Rattus norvegicus</i>	Present	Missing
AW916379	The Institute for Genomic Research, Rockville Tissue mix	Only information given was <i>Rattus norvegicus</i>	Present	Missing
AI043655	University of Iowa Tissue mix	Sprague Dawley	Present	Missing
AA891618	The Institute for Genomic Research, Rockville Kidney	Only information given was <i>Rattus norvegicus</i>	Present	Missing
AA858573	University of Iowa Embryonic	Sprague Dawley	Missing	Missing
AI233367	The Institute for Genomic Research, Rockville Kidney	Only information given was <i>Rattus norvegicus</i>	-	-

FPVYDYPSS LQEALSASVA KVNSQSLSPY LFRATRSSLK RVNVLDEDTL
 VMNLEFTVQE TTCLRESGDP STCAFQRGYS VPTAACRSTV QMSKGQVKDV
 WAHCRWRSTS ESNSSSEEMIF GDMARSHRRR NDYLLGFLYD EPKGEQFYDR
 SIEITRRGHP PAHRRFLNLQ RRARVNSGFE

Figure 5.1. The regions of the spp24 rat protein that could be missing if the ESTs with anomalies were translated.

The regions of the rat spp24 protein that correspond to the missing regions of exon seen in the ESTs are boxed. The exon 4 region is the longer of the two. The characteristic cysteine residues are shown in red. The signal peptide is omitted. The regions of the protein that are encoded in the human and mouse by exons 4 and 6 (Chapter 3) are shown in blue.

If all of the cDNA is present, as expected, then the size of the RT-PCR product should be 335 bp. If the region corresponding to exon 4 is missing then the size of the expected RT-PCR product is 225 bp. If the region corresponding to part of exon 6 is missing the expected size of the RT-PCR product is 308 bp and if both regions are missing the expected size is 198 bp.

RT-PCRs were performed as described in section 2.11.3, Chapter 2 on 4 µg of total RNA from the liver and kidney of 6 different rat strains (Milan Normotensive, Milan Hypertensive, Wistar Kyoto Normotensive, Wistar Kyoto Spontaneously Hypertensive, Lyon Normotensive, Lyon Hypertensive: RNAs donated by Nilesh Samani, University of Leicester). The PCR conditions were as follows: (96°C 30s, 63°C 30s, 70°C 15s) × 30 cycles.

The RT-PCR products from the entire collection of rat RNAs tested were of the normal expected size, 335 bp, indicating that there are no regions missing. The RT-PCR products from the Lyon Normotensive rat are shown in figure 5.2.

5.2.2 A chicken hypothetical protein showing homology to spp24

In a previous study (Chapter 1) a BLAST search revealed that spp24 had a comparable level of sequence identity to cystatin domains 1 and 3 of human kininogen and the precursor to the bovine neutrophil antibiotic peptide bactenecin (Hu *et al.* 1995). However, the same search now reveals a greater level of sequence identity between spp24 and a hypothetical chicken protein (accession number Q91982). This protein has part of a cystatin domain at the N-terminus, a serine-rich region and a C-terminal non-cystatin-like domain.

A BLAST search of the Swissprot database with the hypothetical chicken protein shows that the protein has the highest level of homology to spp24, with human spp24 being the most similar. The greatest homology is seen between the cystatin-like domains of the two proteins. The protein alignment is shown in figure 5.3B.

The hypothetical chicken protein was also identified by ProDom 99.1 (Altschul *et al.* 1997) (Gouzy *et al.* 1999) (Sonnhammer and Kahn 1994) which identifies homologous domains between proteins. Prodom uses a web interface (<http://protein.toulouse.inra.fr/prodom/doc/prodom.html>) to search a database compiled of consensus protein domains that have been assigned a number. Figure 5.3A shows the results of inputting the human, bovine, or rat spp24 proteins into ProDom.

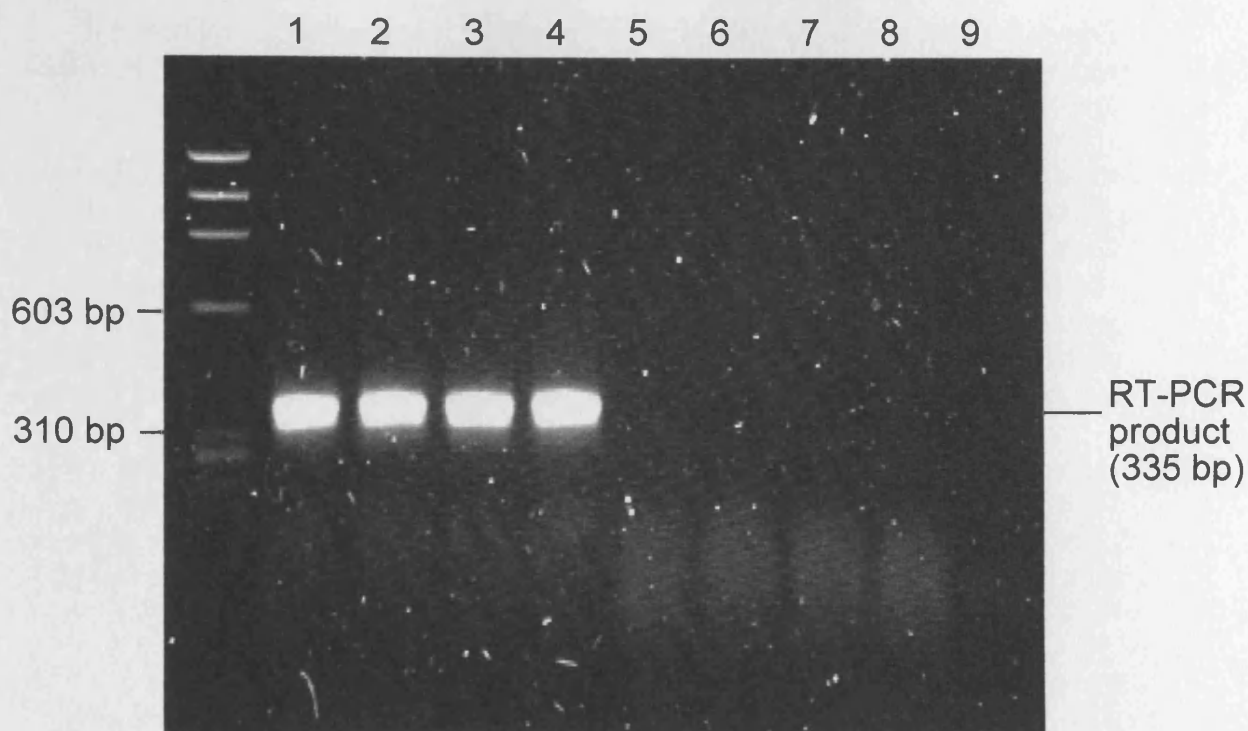


Figure 5.2. RT-PCR of Lyon Normotensive and Lyon Hypertensive rats.

RT-PCRs were performed on 4 g of total RNA from the liver and kidney of 6 different rat strains (RNAs donated by Nilesh Samani, University of Leicester).

The gel image above shows the results from the liver and kidney of two of the rat strains, the Lyon Normotensive and the Lyon Hypertensive.

Lanes 1 and 2 are Lyon Normotensive, lanes 3 and 4 are Lyon Hypertensive, lanes 1 and 3 are liver and lanes 2 and 4 are kidney. Lanes 5 to 8 are the corresponding no RT negative controls and lane 9 is the PCR negative control carried out using sterile water.

The sizes of the relevant bands of ϕ X174 RF cut with *Hae*III are indicated on the left hand side of the gel in base pairs. The RT-PCR product is indicated on the right hand side of the gel.

If all of the cDNA is present as expected then the size of the RT-PCR product should be 335 bp. If the region corresponding to exon 4 is missing then the size of the expected RT-PCR product is 225 bp. If the region corresponding to part of exon 6 is missing the RT-PCR product expected size is 308 bp and if both regions are missing the expected size is 198 bp. All of the RT-PCR products here are approximately 335 bp in size and so it is concluded that the transcripts in these rat tissues do not have exon 4 and part of exon 6 missing.

The results shown above were seen for all of the rat strains and tissues tested.

Figure 5.3. The protein domains of spp24 and Q91982 identified by ProDom 99.1 and the alignment of these two proteins with and without the translated 5' UTR of the Q91982 protein.

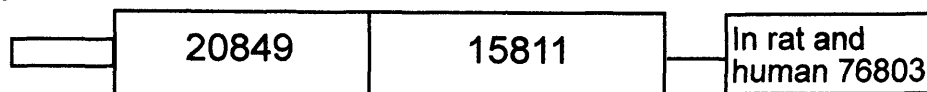
Figure 5.3A shows the protein domains identified by ProDom 99.1. The single line shown in both proteins represents the region of phosphorylated serine residues. The blank rectangle shared between both proteins represents the non-cystatin region. The first blank rectangle in the spp24 protein represents the signal peptide, which is not present in the hypothetical chicken protein (Q91982). The cystatin domain of spp24 is then split into two domains with the ProDom IDs 20849 and 15811. Only one of these domains (15811) is seen in the hypothetical chicken protein (Q91982).

Figure 5.3B shows the alignment using the Gap program in the GCG molecular biology package (see section 2.21.3, Chapter 2) of and the hypothetical chicken protein (Q91982) human spp24. The chicken protein is shown on the top line of the alignment (starting 'MWNS') and the human protein is shown on the bottom line of the alignment.

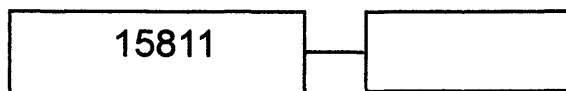
Figure 5.3C shows another Gap alignment, this time of human spp24 and the hypothetical chicken protein (Q91982) including the translation of the 5' UTR of the gene encoding the hypothetical chicken protein. Again the chicken protein is shown on the top line of the alignment and the human protein on the bottom line. The methionine of the chicken protein that was originally reported to be encoded by the 'ATG' start codon (Agarwal *et al.* 1995) is shown in bold.

A.

Spp24



Hypothetical
chick protein



B.

```

1 .....MWNSNDYRLELQLSIRETECTKASGRDP 28
51 VNSQSLSPYLFRAFRSSLKRVEVLDENNLVMNLEFSIRETTCKRDSGEDP 100
29 FTCGFKVGPFPVPTAVCKSVVEVSSEQIVNVIVRCHQSTFSSESMSSEEMT 78
101 ATCAFQRDYVYSTAVCRSTVKVSAQQVQGVHARCSWSSSTSESYSSEEMI 150
79 YMLMTDPRK.....RGSSRSEAFSSRGRGHSN.....GDWRKPDY 113
151 FGDMLGSHKWRNNYLFGLISDEISSEQFYDRSLGIMRRVLPNGNRRYPNH 200
114 TSPGKV....E
201 RHRARINTDFE

```

C.

```

1 .....HVQVRCSEFFSHFSYGLAKKNTL*WNCVSASGFPVYDYELPVT 45
1 MISRMEKMTMMKILIMFALGM...NY...WSC...SGFPVYDYDPSSLR 41
46 EALNASIARINSQTWGNLYGVVRSHVRHVD*WNSNDYRLELQLSIRETE 95
42 DALASVVKVNSQSLSPYLFRAFRSSLKRVEVLDENNLVMNLEFSIRETT 91
96 CTKASGRDPFTCGFKVGPFPVPTAVCKSVVEVSSEQIVNVIVRCHQSTFS 145
92 CRKDSGEDPATCAFQRDYVYSTAVCRSTVKVSAQQVQGVHARCSWSSSTS 141
146 ESMSSEEMTYMLMTDPRK.....RGSSRSEAFSSRGRGHSN..... 181
142 ESYSSEEMIFGDMLGSHKWRNNYLFGLISDEISSEQFYDRSLGIMRRVLP 191
182 .GDWRKPDYTSPGKV....E
192 PGNRRYPNHRHRARINTDFE

```

Figure 5.3. The protein domains of spp24 and Q91982 identified by ProDom 99.1 and the alignment of these two proteins with and without the translated 5' UTR of the gene encoding the hypothetical chicken protein (Q91982).

The hypothetical chicken protein (Q91982) is encoded by the gene GHRG-1 (Growth hormone regulated gene 1). This gene was identified in 1995 by a comparison of gene expression in normal and growth hormone receptor deficient dwarf chicken and was found to be expressed in the liver (Agarwal *et al.* 1995).

The 5' UTR of the GHRG-1 cDNA (U20160) was compared to human spp24 cDNA and demonstrated a high level of homology with exon 2 of *SPP2*. The 5'UTR was then translated and the hypothetical chicken protein (Q91982) including this region was compared to the human spp24 protein. A similar level of homology was seen in this region of the protein as in the rest of the protein. This is shown in figure 5.3C.

It is possible that the sequence of GHRG-1 cDNA that was reported (Agarwal *et al.* 1995) is incorrect and this protein is in fact the chicken counterpart of spp24. However, it is also possible that this is not the chicken counterpart of spp24, but a closely related protein in evolutionary terms that is a member of the same protein family. A third possibility is that this protein is the chicken counterpart of spp24, but that in the chicken the first part of the protein is lost indicating that this part of the protein is not essential for its function or that spp24 is functionally redundant in the chicken.

In an attempt to determine whether the chicken hypothetical protein Q91982 was incorrect, the chicken EST databases at the Roslin Institute (www.ri.bbsrc.ac.uk/cgi-bin/est-blast/) and the University of Delaware (www.chickest.udel.edu/chick.htm) were searched using BLAST with the human spp24 protein. A total of six ESTs were identified with the IDs pat.pK0042.c4.f, pat.pK0072.f10.f, pat.pK0053.d7.f, pat.pK0048.h2.f, pnl1s.pK003.h7 and pg11n.pK007.g13.

All of the chicken ESTs when translated in their longest ORF were identical to the hypothetical chicken protein (Q91982), but yielded additional N-terminal residues to those in the sequence reported by Agarwal *et al.* (1995).

The longest EST when translated in its longest ORF provided an additional 72 residues to the beginning of the hypothetical chicken protein (Q91982). Within these residues there are 2 possible 'ATG' start codons, neither of which lie in a classical Kozak sequence (Kozak 1989). Figure 5.4 shows the protein sequence determined from the chicken ESTs aligned with the hypothetical chicken protein (Q91982) and the human spp24 protein.

	1					50
Chicken A	MGKTPEDFER	HTMRSLIFVL	ALSVFTCSGF	PVYDYELPVT	EEALNASIAR	
Chicken B	-----	-----	-----	-----	-----	
Human	MISRMEKMTM	MMKILIMFAL	GMNYWSCSGF	PVYDYDPSSL	RDALSASVVK	
	51					100
Chicken A	INSQTWGPNL	YGVVRSHVRH	VDMWNSNDYR	LELQLSIRET	ECTKASGRDP	
Chicken B	-----	-----	--MWNSNDYR	LELQLSIRET	ECTKASGRDP	
Human	VNSQSLSPYL	FRAFRSSLKR	VEVLDENNLV	MNLEFSIRET	TCRKDSGEDP	
	101					150
Chicken A	FTCGFKVGPF	VPTAVCKSVV	EVSSEQIVNV	IVRCHQSTFS	SESMSSSEEMT	
Chicken B	FTCGFKVGPF	VPTAVCKSVV	EVSSEQIVNV	IVRCHQSTFS	SESMSSSEEMT	
Human	ATCAFQRDYY	VSTAVCRSTV	KVSAQQVQGV	HARCSWSSST	SESYSSEEMI	
	151					200
Chicken A	YMLMTDPRKR	GSSRSEAFSS	RGRGHSNGDW	RKPDYTSPGK	VE-----	
Chicken B	YMLMTDPRKR	GSSRSEAFSS	RGRGHSNGDW	RKPDYTSPGK	VE-----	
Human	FGDMLGSHKW	RNNYLFGLIS	DESISEQFYD	RSLGIMRRVL	PPGNRRYPNH	
	201	211				
Chicken A	-----	-				
Chicken B	-----	-				
Human	RHRARINTDF	E				

Figure 5.4. An alignment of the original chicken sequence (Q91982) (Agarwal *et al.* 1995), the amended chicken sequence and the human spp24 sequence.

The original hypothetical chicken protein (Q91982) (Agarwal *et al.* 1995) is shown as Chicken B, the amended chicken protein sequence is shown as chicken A and the human spp24 protein sequence as Human. A dash means that there are no corresponding residues in that protein and the conserved cysteine residues are shown in red. The signal peptides are shown in blue.

The existence of six chicken ESTs all showing additional residues at the N-terminal end compared with the hypothetical chicken protein (Q91982), suggests that the original sequence was incorrect and that Q91982 could be the chicken spp24 counterpart. The amended Q91982 protein sequence appears in the species alignment in figure 5.7.

Agarwal *et al.* (1995) also determined a promoter sequence for the GHRG-1 (accession number S75126). In an attempt to determine whether this sequence was also incorrect, S75126 was compared using the Fasta program (section 2.21.3, Chapter 2) to the corrected GHRG-1 cDNA. The last part of the promoter sequence was identical to exon 1, followed what appeared to be 10 bases of intron. These 10 bases were seen in the published cDNA sequence at position 11 to 20. The cDNA was therefore shown to contain the last 10 bases of exon 1, followed by the whole of intron 1 before then going into the correct cDNA sequence.

The promoter and cDNA sequences for GHRG-1 published by Agarwal *et al.* (1995) were therefore both shown to actually be genomic sequence containing both coding and non-coding regions. The composition of each sequence and the correct chicken cDNA sequence is shown in figure 5.5. Agarwal *et al.* (1995) performed primer extension to determine the transcription initiation site. The size of the transcript obtained is the same as the size that would be expected with the 'correct' cDNA sequence.

The GHRG-1 promoter sequence (S75126) was trimmed at the 3' end to remove exon 1 and the start of intron1. This was then the corrected chicken promoter sequence used in the work described in Chapter 3. The GHRG-1 cDNA (U20160) was corrected to include the whole of exon 1 and remove intron 1. This was the cDNA then used to determine the chicken spp24 protein presented in this chapter.

Agarwal *et al.* (1995) reported the location of three exon/intron boundaries in the GHRG-1 cDNA. An additional boundary was defined by one of the chicken ESTs that contained some intronic sequence and intron 1 was defined in its entirety as described above. All of the intron/exon boundaries defined (figure 5.5) were located in the regions corresponding to those seen in the human and mouse genes. Intron 1 in the chicken gene was 88 bp in length, comparable with the small intron 1 size seen in the human (99 bp) and mouse (100 bp) genes.

Agarwal *et al.* (1995) identified the GHRG-1 as a gene being regulated by growth hormone. They reported the location of a putative growth hormone response element (GHRE) by similarity to the GHRE in the *Sp1* 2.1 gene. However, this putative GHRE is now known to lie

Figure 5.5. The composition of the published chicken GHRG-1 cDNA and promoter sequence (Agarwal *et al.* 1995) and the correct chicken cDNA sequence with some exon/intron boundaries defined.

The published GHRG-1 promoter sequence (accession number S75126 (Agarwal *et al.* 1995) was shown to actually contain promoter sequence followed by exon1 and the beginning of intron 1 (section 5.2.2). This is depicted in A. A single black line represents the actual promoter sequence, a red rectangle represents exon sequence and a black rectangle represents intron sequence.

The composition of the published GHRG-1 cDNA sequence (accession number U20160 (Agarwal *et al.* 1995) is depicted in a similar manner in A. This sequence was shown to actually contain the last part of exon 1, the whole of intron 1 followed by the rest of the exons.

The correct chicken GHRG-1 cDNA sequence determined as described in section 5.2.2 is shown in B. The 'ATG' start codon and the 'TAA' termination codon are boxed, the exon/intron boundaries that have been defined are shown as a red line. All of the defined boundaries correspond with the exon/intron boundaries defined in the human *SPP2* gene and the mouse *Spp2* gene. Assuming the exon/intron structure is therefore the same in chicken GHRG-1, there are a further two exon/intron boundaries that remain undefined. These boundaries in their expected position are shown in green.

A.

GHRG-1 promoter
sequence (S75126)

exon 1 intron 1

GHRG-1 cDNA
sequence (U20160)

exon 1 intron 1

B.

```

1  ACATTCTGCGG AAAACACCAG AGGATTTTGA GAGGCACACT
51  ATGAGGAGCT TGATTTTGTG CCTCGCTCTG AGCGTTTCA CATGTTGAG
101  ATTTCCAGTG TACGATTATG AACTCCCTGT CACAGAAGAG GCTCTCAATG
151  CTTCTATTGC AAGGATCAAT TCTCAGACTT GGGGCCCAAA CCTGTATGGA
201  GTTGTGAGGA GCCACGTTAG ACACCTTGAC ATGTGGAACA GCAATGATTA
251  TAGACTAGAG CTGCAGCTCA GTATTCGTGA AACCGAATGC ACAAAGCTT
301  CAGGAAGAGA CCCATTACAG TGTGGCTTCA AAGTAGGGCC TTTGTGCCA
351  ACTGCTGTCT GCAAAAGTGT TGTAGAAGTC TCCAGTGAGC AGATTGTGAA
401  TGTTATTGTG CGATGCCATC AGAGCACATT CAGCTCTGAA TCGATGAGCA
451  GTGAGGAGAT GACGTATATG CTGATGACGG ACCCAAGGAA GCGAGGCAGC
501  AGTCGCTCCG AAGCCTTCTC ATCAAGGGGA AGAGGCCACA GCAATGGTGA
551  CTGGCGTAAA CCTGATTATA CTAGCCCTGG CAAGTTGAA TAAATGCAATT
601  TAGGAAAAAC TATTCTGTGA TGAAGTGAGT CTTTCCTTAA AATCACCTTC
651  TGCTTTACAG CCAAGTGGCC ATTGGATGAG TTCATCGGGT GCTGAATGGA
701  TGCACTGCTC AGTCAATAGT GTCCTGACAT ATTACAGCTC ATCGGAAGGA
751  CTGTCTCAGA CACCTAATGT AACTGTCTAG TATGCATTGT ACCATCTCAT
801  AGCAATGATA TTAAAGGATC AAAGGATTGC TCTTGC

```

Figure 6.5. The composition of the published chicken GHRG-1 cDNA and promoter sequence (Agarwal *et al.* 1995) and the correct chicken cDNA sequence with some exon/intron boundaries defined.

in exon 1 and therefore cannot be a GHRE. The corrected promoter sequence for GHRG-1 was searched for other possible GHREs (Dalglish, unpublished) with the consensus 'ANTTC C/T N A/G GAA A/T A/T' (Bergad *et al.* 1999). Two putative elements were found with only 2 mismatches from the consensus at positions -200 to -178 and -160 to -147 relative to the start of exon 1, taken to be the start of transcription. A potential 'TATA' box is also located at positions -24 to -21. These promoter elements are shown in figure 5.6.

5.2.3 Generation of the mouse and pig spp24 protein sequence

The mouse spp24 protein sequence was generated by the translation of the consensus cDNA sequence determined in Chapter 3 using the Translate program of the GCG Molecular Biology package (section 2.21.3, Chapter 2), in the longest ORF identified by the Frames program, also described in Chapter 3. The mouse spp24 protein sequence is presented in the protein alignment in figure 5.7.

A single pig EST was identified in the TIGR pig gene index (www.tigr.org/ssgi/) (accession number BE015092). This was translated in the longest reading frame, but found to be incomplete at the C-terminal end. The pig cDNA clone 127266 (57 E16) whose sequence is reported in BE015092 was obtained from Dr. Tim Smith of the U.S. Meat Animal Research Centre (MARC). The sequence of the insert was determined in its entirety (Dalglish, unpublished) and from this the complete protein sequence was derived. Corrections to the published sequence were also made. The complete pig *SPP2* cDNA has been deposited with the accession number AJ308100. The pig spp24 protein sequence is presented in the protein alignment in figure 5.7.

5.2.4 The alignment of the spp24 protein from six species

Figure 5.7 shows the alignment of the spp24 protein from human, bovine, mouse, rat, pig and chicken. A rat signal peptide sequence has been included that was determined from the rat ESTs in UniGene cluster Rn.84. Beneath the aligned sequences a consensus sequence is given and a sequence showing the residues that are identical between species.

It should be remembered that the amended chicken hypothetical protein (Q91982) has not been confirmed as being chicken spp24 and it may actually be a protein that is just closely related to spp24. The chicken protein is the most diverged from the consensus sequence but is most similar to human spp24. If the sequence showing residues that are identical between

```

1  GTGTCCAGGG TCTGCTCTGT GCCTTGTGTC TGGAGGATGT GGTCAGTGCT
51  TTGTCCTGAT GGGTCATGGT GGCAGCCATG TCTGGGTGCG TGGTTCCCAT
101 TCTGTTCTTA GGAGATGGGA GTTCCAGCTC CTAAGTGCTT TCTGTTGTGC
151 CTGGGACACA AAGGTGCTGG TAGGGGACTA AGGAAGGTGT TGCTTCACCA
201 GCTTTCACCT TAAAACTCAT GGCTAAAAGT GAAAAAATG ATGATGCTCT
251 GCCACTGCCT GTGGAGCTAT CAGGTATCAA GCACTTCTTT TTAATGACAG
301 TGCTAAGGAA AAATACAGCA GTCCATCTCT TATCAGAGAA CTGCTGCTAA
351 ATGGAGAAAA GCTGACAGCA AATATTTACT CTCAGATCAA TTCTGTTTAA
401 AGTGCAATGT TTGTAGCAGA GCTTGAACAC AGGAAGGTCA GATATAGTAT
451 CAAGGCTGCA TTTATATAAA GACAGGAATA TGCAGTGG

```

Diagrammatic representation of the sequence features:

- Two arrows above the sequence indicate the orientation of GHREs: one pointing right above the **TGCTAAGGAA** sequence and one pointing left above the **TATCAGAGAA** sequence.
- A box highlights the **TATA** sequence at position 451.
- An arrow at the end of the sequence (position 451) indicates the start of exon 1.

Figure 5.6. The chicken GHRG-1 promoter region.

The nucleotide sequence shown is the 'corrected' promoter sequence that has been trimmed at the 3' end to remove exon 1 and part of intron 1.

The putative growth hormone response elements (GHREs) are shown in red, their orientation is indicated by the direction of the arrow.

The putative 'TATA' box is shown in red and is boxed.

The start of exon 1 which is assumed to be the start of transcription is indicated by the arrow at the end of the sequence.

Figure 5.7. The alignment of the spp24 protein from six different species and the generation of a consensus sequence.

This figure shows the alignment of the spp24 protein from six different species. The original rat protein (accession number Q62740) did not include the signal peptide. However, the signal peptide was deduced from the rat ESTs and is shown in the figure above. All the signal peptides are shown in blue. The chick protein is thought to be spp24, but this has not yet been confirmed. It is possible that this is actually a very closely related protein from the same family as spp24.

A consensus sequence has been generated in an attempt to highlight the residues important in the function of the protein. Also shown is a sequence giving the residues that are identical between species with those residues that are absolutely conserved between all species being shown in black and those that are identical in all species with the exception of chicken shown in green. There are several instances where a residue is identical in all species except pig. As the pig protein was translated from a single EST it is possible that these residues may be different as a consequence of cDNA cloning errors.

	1				50
Mouse	-----	MLKTLALLVL	GMHYWCATGF	PVYDYDPSSL	QEALSASVAK
Rat	-----	MELA TMKTLVMLVL	GMHYWCA.SF	PVYDYDPSSL	QEALSASVAK
Human	MISRMEKMTM	MMKILIMFAL	GMNYWSCSGF	PVYDYDPSSL	RDALSASVVK
Bovine	-----	M AMKMLVIFVL	GMNHWCTCTGF	PVYDYDPASL	KEALSASVAK
Pig	-----	MEKR AMRMLAMFVL	GTSFWSCAGF	PVYDYDPSSL	REAVGASVAK
Chick	MGKTPEDFER	HTMRSLIFVL	ALSVFTCSGF	PVYDYELPVT	EEALNASIAR
Consensus	-----	M MMK-L--FVL	GMNYW-CTGF	PVYDYDPSSL	QEALSASVAK
Identical	-----	-----	-----	F PVYDYDP-SL	--AL-ASV-K
	51				100
Mouse	VNSQSLSPYL	FRATRSSLKR	VNVLDEDTLV	MNLEFSVQET	TCLRDSG.DP
Rat	VNSQSLSPYL	FRATRSSLKR	VNVLDEDTLV	MNLEFTVQET	TCLRRESG.DP
Human	VNSQSLSPYL	FRAFRSSLKR	VEVLDENNLV	MNLEFSIRET	TCKRDSGEDP
Bovine	VNSQSLSPYL	FRAFRSSVKR	VNALDEDSL	MDLEFRIQET	TCCRRESEADP
Pig	VNSQSLSPYL	FRAFRSSLKR	VNVLGEDSL	MDIEFGIRET	TCKRDSGEDP
Chick	INSQTWGPNI	YGVVRSHVRH	VDMWNSNDYR	LELQLSIRET	ECTKASGRDP
Consensus	VNSQSLSPYL	FRA-RSSLKR	VNVLDEDTLV	MNLEFS-QET	TC-R-SG-DP
Identical	VNSQSLSPYL	FRA-RSS-KR	V--L-E----	---EF---ET-TC	---S---DP
	101				150
Mouse	STCAFQRGYS	VPTAACRSTV	QMSKGQVKDV	WAHCRW.ASS	SESNSSEEMM
Rat	STCAFQRGYS	VPTAACRSTV	QMSKGQVKDV	WAHCRW.RST	SESNSSEEMI
Human	ATCAFQRDYY	VSTAVCRSTV	KVSAQQVQGV	HARCSWSSST	SESYSSEEMI
Bovine	ATCDFQRGYH	VPVAVCRSTV	RMSAEQVQNV	WVRCHW.SSS	SGSSSSEEMF
Pig	ATCDFQRGYF	TPSAICRSTV	QISAEKVQDV	WVRCRW.SSS	SESNSSEEMI
Chick	FTCGFKVGP	VPTAVCKSVV	EVSSEQIVNV	IVRCHQSTFS	SESMSEEMT
Consensus	-TCAFQRGYS	VPTA-CRSTV	QMS-GQV-DV	WA-CRW-SS-	SESNSSEEMI
Identical	-TC-F-R-Y-	---A-CRSTV	--S---V--V	---C-W--S-	S-S-SSEEM-
	151				200
Mouse	FGDMARSHRR	RNDYLLGFLS	DESRSEQFRD	RSLEIMRRGQ	PPAHRRLNL
Rat	FGDMARSHRR	RNDYLLGFLY	DEPKGEQFYD	RSIEITRRGH	PPAHRRLNL
Human	FGDMLGSHKW	RNNYLFGLIS	DESISEQFYD	RSLGIMRRVL	PPGNRRYPNH
Bovine	FGDILGSSTS	RNSYLLGLTP	DRSRGEPLYE	PSRE.MRRNF	PLGNRRYSNP
Pig	FGDILGSSTS	RNNYLRGLIP	DVSRTEPLYE	RSLETMRFP	PPGNRSFPNQ
Chick	YMLMTDPRKR	GSSRSEAFSS	RGRGHSNGDW	RKPDYTSPGK	VE
Consensus	FGDM--SHRR	RNDYLLG-LS	DESR-EQFYD	RSLEIMRRG-	PP--RR-LNL
Identical	FGD---S---	RN-YL-G---	D-----E----	-S-----RR--	P---R---N-
	201	211			
Mouse	HRRARVNSGF	E			
Rat	QRRARVNSGF	E			
Human	RHRARINTDF	E			
Bovine	WPRARVNPGF	E			
Pig	WPRARTNTGF	E			
Chick					
Consensus	-RRARVNSGF	E			
Identical	--RAR-N--F	E			

Figure 5.7. The alignment of the spp24 protein from six different species and the generation of a consensus sequence.

species is re-determined without using the chicken protein, more identical residues are observed. These residues tend to be expanded around the identical regions seen when the chicken protein is included.

5.3 Discussion

The anomaly seen in the rat ESTs cannot be resolved. It seems plausible to suggest that the rat ESTs seen in the UniGene cluster are either incorrect or a result of cloning artefacts as all of the RT-PCRs carried out could find no evidence of these regions being missing. However, the fact that these ESTs are from different sources (The Institute for Genomic Research, Rockville and University of Iowa) makes this unlikely.

It was speculated that the absence of the exon 4 and part of exon 6 regions might be strain specific. The ESTs with the anomalies were either from Sprague-Dawley rats or were simply described as *Rattus norvegicus* (the common Norway rat) which includes the Sprague-Dawley strain. However, the rat spp24 protein sequence (accession number Q62740) that was reported by Hu *et al.* in 1995 was from a Sprague-Dawley rat and this does not have any absent regions.

This appears to be a phenomenon that is seen only in rat. None of the ESTs from human (23 in total), mouse (57 in total) or chicken (6 in total), display any of the anomalies. If the anomaly did occur in other genera you would expect to see some of these 'deleted' ESTs due to the large numbers available, especially in mouse.

It could be that transcripts with errors are normally produced at very low levels and therefore do not interfere with the production of the normal spp24 protein. However, rat may be particularly susceptible to these errors, more specifically the Sprague-Dawley strain, and so the transcripts with errors are seen at higher levels. The Sprague-Dawley strain of rat is a laboratory strain that appears perfectly healthy and so if this were true it obviously does not affect production of normal spp24 or spp24 is not crucial to the health of the animal.

The rat exon/intron structure may not be identical to that of human and mouse and what looks as though it is part of exon 6 could in fact be a complete exon in the rat. The rat exon/intron structure needs to be determined to resolve this.

The splicing of hnRNA is a highly regulated process that normally results in the accurate and efficient removal of introns. However, there are instances where this process goes wrong and exons are skipped. For example, exon 9 of the cystic fibrosis transmembrane conductance regulator gene (CFTR) is frequently skipped due to sequence differences at the exon 9 splice branch/acceptor site (Chu *et al.* 1993). They showed that differences in the number of 'TG'

dinucleotide repeats and the poly-T tract of the exon 9 branch and acceptor sites resulted in exon 9 being completely skipped. The shorter the poly-T tract the higher the number of exon 9 (-) transcripts that were present. The major factor responsible for the skipping of exon 9 was found to be the (T)₅ allele. Interestingly, a greater relative amount of exon 9 (-) transcripts were found to be present in non-CF individuals. This was shown to be due to association between the CF Δ F508 mutation and the genotype (TG)₁₀T₉.

To establish whether a similar phenomenon was taking place in the rat gene encoding spp24, it would be necessary to obtain intronic sequence between exons 3 and 4 and between 5 and 6. However, the fact that only part of exon 6 is thought to be missing in the rat ESTs (assuming that exon 6 is in the equivalent positions to mouse and human) suggests that a cryptic donor splice site may be present in exon 6. Analysis, by eye, of the rat cDNA region in question does not reveal an obvious cryptic donor splice site. The nature of the rat anomaly is not yet understood.

The chicken ESTs identified in the Roslin Institute Chicken EST database suggest that the published sequence of the chicken hypothetical protein (Q91982) (Agarwal *et al.* 1995) is incorrect. The amended sequence has two possible 'ATG' start codons, neither of which lie in a classical Kozak sequence (Kozak 1989). It is therefore not possible to say which methionine is the true start of the protein.

The similarity of the chicken protein to spp24 suggests that it is likely to be the chicken counterpart. The conservation of intron/exon boundaries and the characteristic small size and location of the first intron provide further evidence to support this. However, of all the species aligned in figure 5.7, chicken is the most diverged. The non-cystatin-like region of the protein is also much shorter in chicken than the other species. It is possible that the chicken protein is just very closely related to spp24 and is from the same family of proteins. If this were the case it would be expected that cognate proteins would have been identified in the other species, but there have been no proteins, apart from spp24, showing a high degree of homology to the hypothetical chicken protein identified in any other species.

The chicken protein was deduced from chicken ESTs and the GHRG-1 cDNA (U20160). All of the sequences show some anomalies with respect to one another. It is therefore probable that there are sequencing errors present. Unfortunately, at present, the original sequence chromatograms are not available and so it is not possible to manually edit the sequences. It is therefore likely that some of the sequence divergence observed between the chicken protein

and spp24 in other species is simply a result of sequencing errors. Therefore, it is proposed that the chicken sequence shown in figure 5.7 (*i.e.* an amendment of Swissprot: Q91982) is the chicken counterpart of spp24.

The GHRG-1 cDNA that encodes the chicken protein was isolated from liver (Agarwal *et al.* 1995). This is consistent with the expression of spp24 seen in human, mouse and cattle (Chapter 4). The interesting thing about GHRG-1 is that it was identified as a gene regulated by growth hormone (Agarwal *et al.* 1995). If the protein encoded by this gene is the chicken counterpart of spp24 it suggests that spp24 may also be regulated by growth hormone in other species. Two putative growth hormone response elements (GHREs) have been identified in the corrected GHRG-1 promoter (Dalglish, unpublished).

The alignment of spp24 from different species as depicted in figure 5.7 shows the residues that may be crucial to the function of the protein. The non-cystatin-like region of the protein is the least conserved and contains fewer residues that are identical between all six species. This suggests that this region of the protein is not essential for the primary function of the protein or that it is adapted in a species-specific fashion.

The region of the spp24 protein containing phosphorylated serine residues shows a high number of residues that are identical between all species suggesting that this region of the protein is essential for its function. This may be because the region has a regulatory role, as speculated by Hu *et al.* (1995) that is dependent on the extent of phosphorylation.

The cystatin-like region of the spp24 protein also shows a high number of residues that are identical between species. These residues tend to be clustered, particularly in the first half of the cystatin-like domain. This suggests that it is the cystatin-like domain that is the functionally important domain and that the identical residues are those that are crucial to its function. Functionally important regions of the spp24 protein will be discussed further in Chapter 7.

In summary, the gene encoding spp24 in rat (or a particular strain of rat (Sprague-Dawley)) may be more prone to post-transcriptional processing errors than in other species. The basis of this and the consequences are not yet understood.

The hypothetical chicken protein (Q91982) identified by Agarwal *et al.* (1995) is likely to be the chicken counterpart of spp24 although the sequence originally reported is thought to be

incorrect. If this protein is the chicken counterpart then it is possible that the gene encoding spp24 is regulated by growth hormone in other species as well as in chicken.

The comparison of spp24 between six different species suggests that the cystatin-like region of the protein is the functional domain. There are regions of this domain that appear to be crucial for function, particularly in the first half of the cystatin-like domain. The region containing phosphorylated serine residues also appears to be an essential region. This is expected if the suggestion by Hu *et al.* (1995) that this region has a regulatory role is correct. The lack of identical residues in the non-cystatin-like domain suggests that it may not be functionally important, its function is not dependent on sequence conservation or that it may have evolved to have a species-specific role.

Chapter 6

Protein homologies and protein modelling

6.1 Introduction

The homology of a protein to other known proteins can provide clues to possible protein functions and indicate the residues that may be functionally important. Homologies between proteins can also be exploited to predict the 3D-structure of a protein based on the existing 3D-model of a homologue. Conservation of specific amino acids between species also indicates which are functionally important (Chapter 5).

In this way protein homologies can be used to predict function, functionally important residues and to determine a possible model for the protein and its mode of action. Whilst these comparisons cannot be considered conclusive evidence in themselves, they can provide a direction for functional investigations and may form the basis of a theoretical mechanism.

6.1.1 Proteins showing homology to spp24

Unfortunately, spp24 does not show strong homology across its whole structure to any known protein. As described in Chapter 1, Hu *et al.* (1995) reported the results of a BLAST search using bovine spp24 against the NLM non-redundant protein database. The results of this search revealed that the N-terminal region of spp24 showed homology to the cystatin domain 3 of kininogen and to the precursor of the bovine neutrophil antibiotic peptide batenecin. Both of these proteins are members of the cystatin superfamily and so it has been concluded that spp24 was also a member of this family. A cystatin-like function, a fetuin-like function or the release of a biologically active peptide were the functional possibilities for spp24. This will be built upon in this chapter, which presents further protein homologies that have been identified more recently as additions have been made to protein databases.

6.1.2 Computer-based analysis of the spp24 protein

There are currently many computer programs available to analyse a particular protein sequence. For example proteins can be searched for particular features, structures and post-translational modifications can be predicted and homologous proteins can be identified. Again, these programs are no substitute for experimental evidence, but they can provide good indications and a starting point for further investigations. All of these programs rely on

homology to already characterised proteins. Consequently, the programs that provided 'useful' information on spp24 were limited due to the lack of proteins showing a high degree of homology to spp24 and the non-existence of proteins showing homology to the non-cystatin-like region of spp24.

Table 6.1 details the programs that were used to analyse the spp24 protein. Unless otherwise stated, the human spp24 protein was analysed in each program. This chapter reports the results from some of these programs that contribute to building a picture of the structure and possible functions of the spp24 protein.

6.1.3 Constructing a protein model for spp24 using an evolutionary trace analysis technique

Spp24 is a member of the cystatin superfamily. It is possible to identify cystatin domains within this superfamily that are the most similar to the cystatin domain of spp24. These proteins can then be aligned to identify certain amino acids that are absolutely conserved and those that are different between subgroups. Where subgroups have evolved specific functions, the differing amino acids probably represent functionally important residues. Evolutionary trace (ET) techniques aim to identify these residues and build a hierarchy of protein relationships. The ultimate aim is then to determine the location of these residues on a proposed 3D-structure.

The ET method used for the work presented in this chapter identifies residues in aligned protein sequences whose variation can be linked to the development of different functional classes or subgroups (Lichtarge *et al.* 1996). An evolutionary tree is built for the protein group and residues are identified that are conserved within each subgroup, but variant (*i.e.* not conserved) between subgroups. Subgroups are selected at arbitrarily chosen points on the tree, which are expressed as a percentage and called a partition identity cut-off interval (PIC interval).

The PIC value reflects the degree of sequence homology between the subgroups at that point in the evolutionary tree. For example, at 10% PIC there may be two subgroups that show a low level of sequence homology within each subgroup, at 70% PIC there may now be ten subgroups but within each subgroup there is a higher level of sequence homology. Therefore, in order to obtain a high level of sequence identity the proteins need to be split into a greater number of subgroups to keep a high % identity in each subgroup.

Table 6.1. The protein programs used to analyse the spp24 protein.

Program	Comments	References
PredictProtein (http://dodo.cpmc.columbia.edu)	<p>PredictProtein is an analysis package that runs the following programs simultaneously:</p> <p>PROSITE motif search</p> <p>SEG low-complexity regions</p> <p>ProDom domain search</p> <p>MAXHOM alignment</p> <p>PHD information about accuracy</p> <p>PHD predictions</p> <p>GLOBE prediction of globularity</p>	<p>Rost (1996)</p> <p>Bairoch <i>et al.</i> (1997)</p> <p>Wootton and Federhen (1996)</p> <p>Sonnhammer and Kahn (1994)</p> <p>Sander and Schneider (1991)</p> <p>Rost and Sander (1994)</p> <p>Rost and Sander (1993); Rost (1996)</p> <p>Rost (1998)</p>
Network Protein Sequence Analysis (NPS@) (http://pbil.ibcp.fr/)	<p>NPS@ is an analysis package that runs the following secondary structure prediction programs simultaneously and generates a consensus:</p> <p>GOR 4</p> <p>HNN</p> <p>Predator</p> <p>SIMPA96</p> <p>SOPM</p>	<p>Combet <i>et al.</i> (2000)</p> <p>Garnier <i>et al.</i> (1996)</p> <p>No reference available</p> <p>Frishman and Argos (1996)</p> <p>Levin <i>et al.</i> (1986)</p> <p>Geourjon and Deleage (1994)</p>
Statistical Analysis of Protein Sequences (SAPS) (www.isrec.isb-sib.ch/software/SAPS_form.html)	SAPS is a program that evaluates a protein sequence by statistical criteria.	Brendel <i>et al.</i> (1992)
PSORT at the HGMP-RC (http://menu.hgmp.mrc.ac.uk/)	PSORT is an analysis package that runs 22 programs simultaneously to predict the cellular location of a protein.	Nakai and Horton (1999)

Table 6.1 continued. The protein programs used to analyse the spp24 protein.

Program	Comments	References
Nnpredict (www.cmpchem.ucsf.edu/cgi-bin/nnpredict.pl)	nnpredict is a program that predicts secondary structure	McClelland and Rumelhart (1988) Kneller <i>et al.</i> (1990)
NetPhos 2.0 (www.cbs.dtu.dk/services/NetPhos)	NetPhos predicts phosphorylated serine, threonine and tyrosine residues	Blom <i>et al.</i> (1999)
3D-PSSM (www.bmm.icnet.uk/servers/3dpssm/)	Searches for similar regions of secondary structure in a database of folds.	Kelley <i>et al.</i> (1999) Fischer <i>et al.</i> (1999) Kelley <i>et al.</i> (2000)

A program called 'BRUTUS' has been devised (unpublished, Jon Clayton, University of Leicester) that uses the evolutionary trace method to identify regions of a protein known to be important to its function and then models them onto the 3D-structure of a homologous protein. Obviously, the evolutionary trace method requires a group of proteins that show homology to one another. Consequently, only the N-terminal cystatin-like region of spp24 could be analysed in this way since the non-cystatin-like region shows no homology to any known protein.

This chapter presents the results of analysing the spp24 protein with the program 'BRUTUS' and proposes a model for the functionally important regions of the cystatin-like region of spp24.

6.2 Results

This work was carried out prior to the determination of the pig spp24 sequence, therefore the pig protein is not included in any of the analyses described here.

6.2.1 Proteins showing homology to spp24

Over the duration of this project many computer-based searches were carried out in an attempt to identify proteins showing homology to spp24. Some of these proteins have already been discussed (section 6.1.1). Once a protein was identified in a search, it was investigated more rigorously to reveal the nature and extent of homology. If the homology was thought to be significant, the function and structure of the homologous protein was investigated. Table 6.2 shows the proteins identified as having significant homology to human spp24 either in amino acid sequence or in the overall structure of domains within the protein.

As described in Chapter 1, spp24 has an N-terminal cystatin-like domain, a serine-rich region and a C-terminal non-cystatin-like domain. All of the proteins that show some homology to spp24 have either one, two or three cystatin-like domains followed by a non-cystatin-like domain with the exception of cystatin F (CMAP) which has a single typical cystatin domain. The homologues can be grouped into four families. These are typical cystatins, kininogens, cathelicidins (antimicrobial) and fetuins.

The only protein that has a function not already discussed is HSF. HSF is an antihemorrhagic factor isolated from the venom of the Japanese Habu snake, *Trimeresurus flavoviridis* (Yamakawa and Omori-Sato 1992). HSF shows significant sequence homology to bovine fetuin and human α_2 HS-glycoprotein and consequently is thought to be a snake venom fetuin. Although HSF has two cystatin-like domains it appears to lack the ability to inhibit thiol proteases (Yamakawa and Omori-Sato 1992), like many of the members of the cystatin superfamily with divergent functions. However, HSF has been shown to have the ability to inhibit metalloproteinases (Yamakawa and Omori-Sato 1992). It is possible that spp24 could also inhibit metalloproteinases.

6.2.2 Computer-based analysis of the spp24 protein

A consensus secondary structure was predicted for human spp24 and also for a typical cystatin (chicken egg white cystatin) using NPS@ (table 6.1) and is shown in figure 6.1. The

Table 6.2. The proteins identified that have a significant level of homology with spp24 either at the amino acid sequence level or with respect to the structure of the domains of the protein (not to scale).

Over the duration of this project many computer-based searches were carried out in an attempt to identify proteins showing homology to spp24. Some of these proteins have already been discussed (section 6.1.1).

Table 6.2 details the proteins identified in this way. The protein is identified by name and accession number if applicable. The percentage identity to human spp24 was determined using the Gap alignment program from the GCG Molecular Biology package (see Chapter 2).

The structure of each protein in terms of domains is shown by a block representation. These representations are not drawn to scale. A red block represents a cystatin-like domain. A white block represents a non-cystatin-like domain and a line represents a serine-rich region, which is only seen in spp24.

Table 6.2. The proteins identified that have a significant level of homology with spp24 either at the amino acid sequence level or with respect to the structure of the domains of the protein (not to scale).























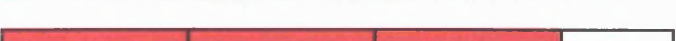

Protein	Accession number	% identity to human spp24	Structure of domains
CAMP/Cystatin F/Cystatin 7	O76096	38	
Human spp24	Q13103	100	
Bovine spp24	Q27967	71	
Rat spp24	Q62740	73	
Mouse spp24	see Chapter 6	73	
Chicken spp24	see Chapter 6	50	
Pig Protegrin 1	P32194	32	
Pig Protegrin 2	P32195	30	
Pig Protegrin 3	P32196	32	
Pig Protegrin 4	P49933	31	
Pig Protegrin 5	P49934	34	
Pig cathelin	P32195	36	

Table 6.2 continued. The proteins identified that have a significant level of homology with spp24 either at the amino acid sequence level or with respect to the structure of the domains of the protein (not to scale).

Protein	Accession number	% identity to human spp24	Structure of domains
Bovine antibacterial protein BMAP-28	P54229	31	
Bovine cyclic dodecapeptide precursor	P22226	32	
Pig antibacterial protein PMAP-23	P49930	31	
Pig antibacterial protein PR-39	P80054	34	
Mouse cathelin-related antimicrobial peptide CRAM	P51437	31	
HSF	P29695	31	
Human kininogen LMW	P01043	28	
Human kininogen HMW	P01042	28	
Bovine kininogen LMW I	P01046	25	
Bovine kininogen LMW II	P01047	25	
Bovine kininogen HMW I	P01044	25	
Bovine kininogen HMW II	P01045	25	

Human spp24

..FPVYDYPSSLRDALSAVVKVNSQSLPYLFRAFRRSLKRVEVLDENNLVMNLEFSIRETTCKRDSGEDPATCAFORDY
 ..CCCCCCCCCHHHHHH?EEEECCCCCHHHHHHHHHHEEEEECCCCHE??HHH?CCCCCCCCCCCCC??HHC?CE

CCCCCCCCCH.H.HHHHHHHHHHHCCCCCCCCCHHEEEEEHHH.HHHHCC?EEEE?CCCCCCCCCCCCCCCCCCCC
 GAPVPVDEND.E.GLQRALQFAMAEYNRASNDKYSSRVVRVISAK.RQLVSGIKYILQVEIGRTTCPKSSGDLQSCFEFHDEPE

Chicken egg white cystatin

Human spp24

YVSTAVCRSTVKVSAQQVQGVHAR..CSWSSSTSESYSSSEEMIFGDMGLGSHKWRNNYLFGLISDE
 EEEEECCCCCEEECHH?CC?CEE..EECCCCCCCCCHH?E?CHCCCCCCCCC?EEEECCCC

CCCCEEEEEEEECCCCCHHHHH?CCC
 MAKYTTCTFVYISIPWLNQIKLLESKCQ

Chicken egg white cystatin

Human spp24

SISEQFYDRSLGIMRRVLPNGNRRYPNHRHRARINTDFE
 CHHHHHHHC?CEE?EECCCCCCCCCCCCC??CCCCC

Figure 6.1. The consensus secondary structure of human spp24 and a typical cystatin (chicken egg white cystatin) as determined by NPS@ (table 6.1).

The consensus secondary structure for human spp24 and a typical cystatin (human cystatin C) generated by Network Protein Sequence Analysis (NPS@, table 7.1). A '?' indicates that a consensus could not be determined at this position. A 'C' represents a random coil. A 'H' represents an alpha helix and an 'E' represents a beta extended strand. In the human spp24 amino acid sequence, an orange coloured residue indicates that the program NetPhos predicted this residue would be phosphorylated.

consensus secondary structure for the cystatin-like region of spp24 and cystatin are similar suggesting that the cystatin-like region of spp24 may fold in a similar manner to a typical cystatin.

PHD, which is another secondary structure prediction program (table 6.1), predicted that human spp24 could be classed as a protein with a 'mixed' secondary structure. The program predicted that 36% of the protein was alpha helix and 15% was beta extended strand. This agrees with NPS@ with respect to the beta extended strands, but predicts that about 10% more of the protein is alpha helix than is predicted by NPS@.

The phosphorylation prediction program NetPhos (table 6.1) predicted 13 phosphorylated serine residues, 4 phosphorylated threonine residues and 3 phosphorylated tyrosine residues in the human spp24 protein. These residues are indicated in figure 6.1. Many of these predictions are also seen using the PROSITE program (table 6.1), which identifies some of these residues as being potential cAMP- and cGMP- dependent, protein kinase C and casein kinase II phosphorylation sites.

Most of these predicted phosphorylated residues lie in the cystatin-like region of spp24 or the serine-rich region. The prediction of phosphorylation in the serine-rich region is supported by the degree of phosphorylation shown experimentally in the same region of the bovine protein (Hu *et al.* 1995).

The phosphorylated residues predicted in the non-cystatin-like region of human spp24 are not residues that are conserved between species and so are unlikely to be significant to the function of the protein. However, 7 of the 9 residues predicted to be phosphorylated in the cystatin-like region of human spp24 are conserved between human, bovine, mouse, rat and pig. This suggests that they are functionally important residues and thus the extent phosphorylation could also be critical to the protein function.

The SEG low-complexity regions program (table 6.1) predicted, as expected, a low-complexity region of serine residues between the cystatin-like and non-cystatin-like regions of human spp24.

The SAPS program (table 6.1) provided some information about the charge distribution of the human spp24 protein. The charge distribution as predicted by SAPS is shown in figure 6.2. The charge distribution across each region of the protein is fairly even. There are no charge

		Overall charge of whole protein: including signal peptide +4 excluding signal peptide +2
	Cystatin-like region	000+0-+00000+00000000000000000000-0-0000+-0000000+00000000000+00+00 MISRMEKMTMMMILIMFALGMNYWSCSGFPVYDYDPSSLRDALSASVVKVNSQSLSPYLFRAFRSS 0++0-00--0000000-000+-000++-00--0000000+-00000000+000+000000000+0+000 LKRVEVLDENNLVMNLEFSIRETTCKRDSGEDPATCAFQRDYYVSTAVCRSTVKVSAQQVQGVHARCSW Overall charge: including signal peptide +3 excluding signal peptide +1
	Serine-rich region	00000-0000 SSSTSESYSS Overall charge: -1
	Non-cystatin-like region	--0000-00000+0+000000000--000-000-+00000++000000++0000+0+0+000-0- EEMI FGDMLGSHKWRNNYLFGLISDESISEQFYDRSLGIMRRVLPPGNRRYPNHRHRARINTDFE Overall charge: +1

Figure 6.2. The charge distribution in the human spp24 protein sequence.

An analysis of charge distribution in the human spp24 protein was carried out using the SAPS program (table 6.1). This program assumes the residues arginine (R), histidine (H) and lysine (K) to be positive and aspartic acid (D) and glutamic acid (E) to be negative. The charges are shown for each of the three regions of the protein: the cystatin-like, the serine-rich and the non-cystatin-like. A neutral residue is shown as a '0', a residue with a positive charge as a '+' and a residue with a negative charge as a '-'. The overall charges of each region are indicated as is the overall charge of the whole protein.

clusters predicted. Overall the protein has a slightly positive charge. The SAPS program took the amino acids with basic side chains to be positive (lysine, arginine and histidine) and those with acidic side chains to be negative (aspartic acid and glutamic acid).

The PSORT package (table 6.1) predicted a cleavage site for removal of the signal peptide between residues 29 and 30 (a G and F residue). This corresponds to the signal peptide cleavage site in the bovine protein reported by Hu *et al.* (1995). The PSORT package also predicted, as expected, that the human spp24 protein was cytoplasmic.

To begin to look at ways in which the protein may fold (*i.e.* its tertiary structure) the program 3D-PSSM was used (table 6.1). This program predicts the secondary structure of a protein and then searches a database of ‘folds’ (*i.e.* characterised 3-dimensional structures) for regions that are similar in their secondary structure. In this way it is possible to predict how segments of the protein may fold by comparison with homologues. The only protein that was identified by 3D-PSSM as being significantly similar to human spp24 was ‘1cewi’ (PDB identification), chicken egg-white cystatin, the structure of which has been well characterised (Chapter 1).

It was therefore only possible to analyse the cystatin-like region of human spp24 using 3D-PSSM. The program predicted, on the basis of comparison with ‘1cewi’, whether residues were likely to be buried or exposed on the surface of the molecule. Two clusters of residues that are highly conserved between species were predicted to be buried in the protein. These regions are indicated in figure 6.3, which also shows the residues of spp24 that are identical between species. It is possible that these residues are critical to an interaction with a target protein that takes place within a ‘pocket’ in the spp24 protein.

6.2.3 Constructing a protein model for spp24 using an evolutionary trace analysis technique

The evolutionary trace method requires proteins showing homology to the target protein both at the highest level possible and also with more remote homology to ensure that there is no bias and a sufficient level of variation. For this reason all of the proteins showing homology to spp24 (table 6.2) were included in this analysis.

The evolutionary trace method is based on homology to known proteins. Therefore, only the cystatin-like region of spp24 was modelled. Consequently, all signal peptides and non-cystatin-like regions were trimmed from protein sequences. Where proteins contained more

FPVDY-----AL-AS-----NSQ---P-L---RS---V-----
 ---L-----ET-C---S---DP-TC-F-----V--A-C-S-V---S-----
 -V---C-----S-S-SSEEM

Figure 6.3. The residues in the cystatin-like region of spp24 that are identical between species and the residues that are predicted to be buried within the protein.

The program 3D-PSSM (table 6.1) identified a single protein (PDB identification 1cewi) that showed significant homology to human spp24. The 3D-structure of this protein has been well characterised and therefore, by comparison, the program was able to predict which residues in the cystatin-like region of spp24 were likely to be buried and which were likely to be exposed on the surface of the molecule.

This figure shows, by single letter code, the residues of the cystatin-like region that are identical between human, bovine, mouse, rat, pig and chicken. The signal peptide is not included. A dash indicates that this residue was not identical between species.

The regions that were predicted by 3D-PSSM to be buried are shown in red and are underlined. The arrows indicate the residue that is predicted to be most deeply buried in any particular stretch.

than one cystatin-like domain these domains were split and treated separately. The cystatin-like domains in a protein were numbered in ascending order from the N-terminal end of the protein.

All of the cystatin-like regions of the proteins described in table 6.2 were then aligned using the multiple sequence alignment program CLUSTAL W (section 2.21.3, Chapter 2). The '*.msf' file generated from CLUSTAL W was then used as an input file in the program Distances which is part of the GCG Molecular biology Package (section 2.21.3, Chapter 2). Distances computes a 'distance' between each protein in evolutionary terms based on sequence identity.

The '*.distances' file produced from the Distances program was then used as an input file in the program 'Growtree', again part of the GCG Molecular biology Package (section 2.21.3, Chapter 2). 'Growtree' produces an evolutionary tree based on the distances computed by the Distances program. The tree produced is shown in figure 6.4. This evolutionary tree simply shows how the cystatin domains of all the homologous proteins analysed are related to one another. This tree forms the basis of the 'BRUTUS' analysis to build a structural model of spp24.

The 'BRUTUS' program (Jon Clayton, unpublished) uses the original '*.msf' file from CLUSTAL W and the '*.nex' NEXUS output file from the 'Growtree' program as input files. 'BRUTUS' then works up the evolutionary tree from the base (or root) up to the ends of each branch. At 5% PIC intervals (section 6.1.3) each residue is compared within and between subgroups. For example, at 0% PIC (*i.e.* the root of the tree where all proteins are in the same group and there is a low level of sequence homology) there are four cysteine residues (C) and an asparagine residue (N) that are present in all of the proteins. This is shown in the 'short' log file 'BRUTUS' output in figure 6.5. A copy of the 'long' log file from 'BRUTUS' can be found in Appendix B.

The conservation of the four cysteine residues throughout all of the proteins is not unexpected as this is characteristic of a cystatin domain. However, the asparagine is a surprise and this could be an indication of a residue that is critical to a general function linking all of the proteins. This asparagine does appear to be present in all type 2 and type 3 cystatins, but has not been identified as one of the residues that is critical to the interaction of a typical cystatin with papain (Chapter 1). It is therefore unclear as to why this residue is so well conserved.

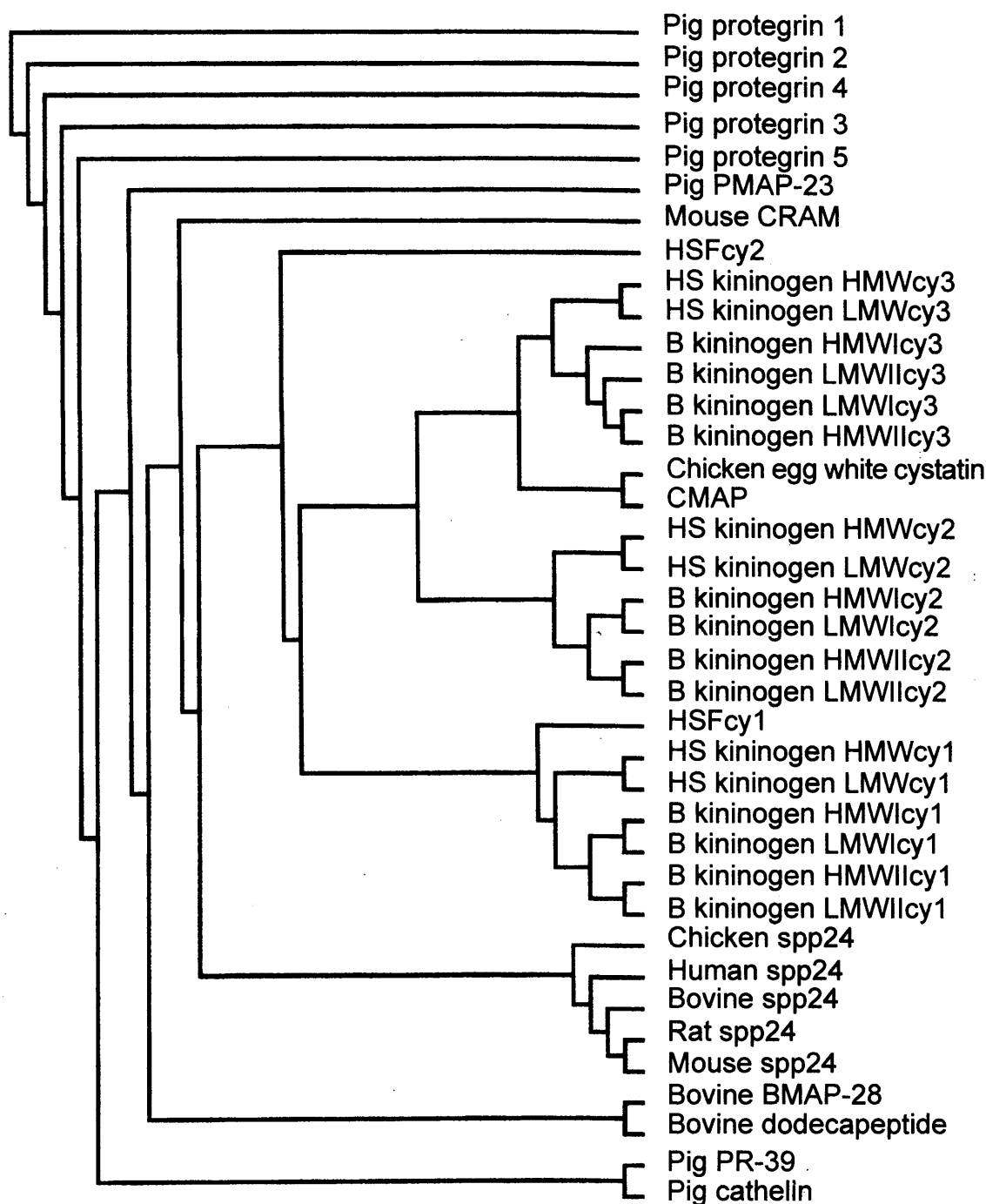


Figure 6.4. The evolutionary tree produced by the program 'Growtree' that is part of the GCG Molecular Biology Package.

Spp24 from 5 different species and all the proteins showing homology to spp24 (table 6.2) were aligned and evolutionary distances calculated. Only the cystatin-like regions were included in the analysis.

The program 'Growtree' was used to produce an evolutionary tree based on the calculated distances. The relationship between each protein is shown using lines. Each protein is identified on the right hand side by a short description. Cy1, cy2 and cy3 refer to the cystatin-like domains where there are more than one in a protein. The cystatin domains in each protein are numbered in ascending order from the N-terminal end of the protein.

Figure 6.5. A 'short' log file from the program 'BRUTUS' showing the analysis of the cystatin region of spp24 and homologous proteins.

The upper part of the figure shows the multiple sequence alignment that was generated using the program CLUSTAL W (see Chapter 2). The evolutionary trace is shown in the lower part of the figure and enables the regions of clustered residues in the proteins to be identified. Each protein sequence is identified on the left-hand side either by its accession number or a short self explanatory comment.

The lower part of the figure shows a summary of the evolutionary trace generated by the program 'BRUTUS' (Jon Clayton, unpublished). The residues that are absolutely conserved between all protein sequences are shown by the letter that represents the appropriate amino acid. The residues that are class-specific (*i.e.* variant between subgroups, but invariant within subgroups) are represented by the letter 'X'. The PIC level is indicated on the left-hand side in percent.

P01042cy3PPTKICVG.....	CPRDIPTNSP.....	ELEETLTHITIKLNAENNAATFYKIDNVKKAR.....	QVVVAGKKYFIDFVARETTCSKESNEELTESCETKKLG.QS.....	LDCNAEYVYVP.....	WEKKIYPTVN.CQPLGMSILMK.....	P01042cy3												
P01042cy2TAQYDCLG.....	CVHPISTQSP.....	DLEPILRHGIQYFNNNTQHSSFLMNEVKRAQ.....	QRVVAGNLFRTYSIVQTNCSKENFLFTPDCSKLSWG.DT.....	GECTDNAYIDI.....	QLRIASFQSN.CDIYPGKDFVQ.....	P01042cy2												
P01042cy1QESQ.....	SEEDICNDK.....	DLFKAVDAALKKYNQSNQNNQFVLYRTEAT.....	KTGVSDDTFYSFYKEIKEGDCPVQSG.TKWDQCEYKDA.....	KAATGECTATVKKRS.....	STKFSVAIQT.CLITPAEGPVV.....	P01042cy1												
P01043cy3PPTKICVG.....	CPRDIPTNSP.....	ELEETLTHITIKLNAENNAATFYKIDNVKKAR.....	QVVVAGKKYFIDFVARETTCSKESNEELTESCETKKLG.QS.....	LDCNAEYVYVP.....	WEKKIYPTVN.CQPLGMSILMK.....	P01043cy3												
P01043cy2TAQYDCLG.....	CVHPISTQSP.....	DLEPILRHGIQYFNNNTQHSSFLMNEVKRAQ.....	QRVVAGNLFRTYSIVQTNCSKENFLFTPDCSKLSWG.DT.....	GECTDNAYIDI.....	QLRIASFQSN.CDIYPGKDFVQ.....	P01043cy2												
P01043cy1QESQ.....	SEEDICNDK.....	DLFKAVDAALKKYNQSNQNNQFVLYRTEAT.....	KTGVSDDTFYSFYKEIKEGDCPVQSG.TKWDQCEYKDA.....	KAATGECTATVKKRS.....	STKFSVAIQT.CLITPAEGPVV.....	P01043cy1												
P01044cy3PPTRLCAG.....	CPKPIPVDS.....	DLEELPSHSIAKLNAEHDAFGFYKIDTVKKAT.....	QVVVAGLKYSVIFARETTCSKGSNEELTKSCEINIH.QI.....	LHCDANVYVP.....	WEEKVYPTVN.CDPLPGQTSLM.....	P01044cy3												
P01044cy2TAQYECGL.....	CVHPISTKSP.....	DLEPVLRYIAIQYFNNNTSHSLFDLKEVKRAQ.....	QKQVSGWNVYVNSIAQTNCSKEEFSFLTPDCSKLSWG.DT.....	GECTDKAHVDV.....	KLRISSFSQK.CDLYPVKDFVQ.....	P01044cy2												
P01044cy1QES.....	SEQIDICNDQ.....	DVFKAVDAALTKYNSKESGQNFVLYRTEVA.....	RMDNPDTFYSYLKYQIKEGDCPPQSN.TKWDQCDYKDSA.....	QAATGECTATVARRG.....	NMKFSVAIQT.CLITPAEGPVV.....	P01044cy1												
P01046cy3PMWCVG.....	CPKPIPVDS.....	DLEELNHSIAKLNAEHDGTFYKIDTVKKAT.....	QVVVGGLYKSVIFARETTCSKGSNEELTKSCEINIH.QI.....	LHCDANVYVP.....	WEEKVYPTVN.CQPLGQTSLM.....	P01046cy3												
P01046cy2TAQYECGL.....	CVHPISTKSP.....	DLEPVLRYIAIQYFNNNTSHSLFDLKEVKRAQ.....	QKQVSGWNVYVNSIAQTNCSKEEFSFLTPDCSKLSWG.DT.....	GECTDKAHVDV.....	KLRISSFSQK.CDLYPVKDFVQ.....	P01046cy2												
P01046cy1QES.....	SEQIDICNDQ.....	DVFKAVDAALTKYNSKESGQNFVLYRTEVA.....	RMDNPDTFYSYLKYQIKEGDCPPQSN.TKWDQCDYKDSA.....	QAATGECTATVARRG.....	NMKFSVAIQT.CLITPAEGPVV.....	P01046cy1												
P01045cy3PMWCVG.....	CPKPIPVDS.....	DLEELNHSIAKLNAEHDGTFYKIDTVKKAT.....	QVVVGGLYKSVIFARETTCSKGSNEELTKSCEINIH.QI.....	LHCDANVYVP.....	WEEKVYPTVN.CDPLPGQTSLM.....	P01045cy3												
P01045cy2TAQYECGL.....	CVHPISTKSP.....	DLEPVLRYIAIQYFNNNTSHSLFDLKEVKRAQ.....	QKQVSGWNVYVNSIAQTNCSKEEFSFLTPDCSKLSWG.DT.....	GECTDKAHVDV.....	KLRISSFSQK.CDLYPGEDFL.....	P01045cy2												
P01045cy1QES.....	SEQIDICNDQ.....	DVFKAVDAALTKYNSKESGQNFVLYRTEVA.....	RMDNPDTFYSYLKYQIKEGDCPPQSN.TKWDQCDYKDSA.....	QAATGECTATVARRG.....	NMKFSVAIQT.CLITPAEGPVV.....	P01045cy1												
P01047cy3PMWCVG.....	CPKPIPVDS.....	DLEELNHSIAKLNAEHDGTFYKIDTVKKAT.....	QVVVGGLYKSVIFARETTCSKGSNEELTKSCEINIH.QI.....	LHCDANVYVP.....	WEEKVYPTVN.CQPLGQTSLM.....	P01047cy3												
P01047cy2TAQYECGL.....	CVHPISTKSP.....	DLEPVLRYIAIQYFNNNTSHSLFDLKEVKRAQ.....	QKQVSGWNVYVNSIAQTNCSKEEFSFLTPDCSKLSWG.DT.....	GECTDKAHVDV.....	KLRISSFSQK.CDLYPGEDFL.....	P01047cy2												
P01047cy1QES.....	SEQIDICNDQ.....	DVFKAVDAALTKYNSKESGQNFVLYRTEVA.....	RMDNPDTFYSYLKYQIKEGDCPPQSN.TKWDQCDYKDSA.....	QAATGECTATVARRG.....	NMKFSVAIQT.CLITPAEGPVV.....	P01047cy1												
HSFcY2HSCFY.....NCSK.....CPILLPPNN.....HVDVSVEYVLNKH.....EKLSGHIYEVLEISRGQ.....HKYEPAYLEFVIVEINCTAQEADHHQCHQPYTAG.....	EDHIAFCRSTVFRSHASLEKPKDFKEDSCVDLVYKNGHAHSH	HSFcY2											
HSFcY1DQ.....VRGLEDCKDK.....EAKNWDAVRYINEHKLGHKQALNKLCPVVPNNQDVAVFLELNLLETCHVLD.....TPHVEKTVQRQHNHVAEMDCDAKIMFVN.....ETFKRDVEVK.....CHSTPDSVENVRR.....	HSFcY1											
HSSPP24FPVYDYDS.....SLRDALSASVVKVNSQSLSPYLFRFRSSLKRVEVLDDENNLMNLEFSIRETTCRKDSG.....EDPATCAFQRDY.....YVSTAVCRSTVKVSA.....QQQGVHAR.....CSW.....	HSSPP24											
BTSP24FPVYDYDPA.....SLKEALSASVAKVNSQSLSPYLFRFRSSVRKVNALDESDTLMEFRITQCTCRRESE.....ADPATCDFQRY.....HVPAVCACTVRMSA.....EQGVNVWR.....CHW.....	BTSP24											
RNSP24FPVYDYDS.....FLSQALSASVAKVNSQSLSPYLFRATRSSLRKVNLDDEITLVNLEFTVQETTLRES.....GDPSTCAFQRY.....SVPTAACRSTVQMSK.....GOVKDVAH.....CRWR.....	RNSP24											
MMSPP24FPVYDYDS.....FLSQALSASVAKVNSQSLSPYLFRATRSSLRKVNLDDEITLVNLEFSVQETTLRES.....GDPSTCAFQRY.....SVPTAACRSTVQMSK.....GOVKDVAH.....CRWA.....	MMSPP24											
CHICK SP24FPVYDYELP.....VTEALNASIARINSQTSQPNLYGVVRSHVRHVDMMNSNRYLELOLSIRETCTKASG.....RDPTCGKFGY.....FVPTAVKSVVVEYS.....EQIVNVIR.....CHQSTF.....	CHICK SP24											
P80054	METQRASLC	GRWSLW	LLLLGLV	VPSASAQALS	YREAVLR	RAVDR	RLNEQ	SEANLYRLLELDQPP.....	KADEDGPTGPKPVSFVTVKETVCPRPTR.....	QPPEL	CDFKENG.....	RV.....	KQCVG	VTVLN.....	PSI	HLDIS.....	CNEIQSV	RRR.....	P80054
P49934	METQRASLC	GRWSLW	LLLLGLV	VPSASAQALS	YREAVLR	RAVDR	RLNEQ	SEANLYRLLELDQPP.....	KADEDGPTGPKPVSFVTVKETVCPRPTR.....	QPPEL	CDFKENG.....	RV.....	KQCVG	VTVLN.....	QIK	DPLDIT.....	CNEVQGV	RRGR.....	P49934
P49933	METQRASLC	GRWSLW	LLLLGLV	VPSASAQALS	YREAVLR	RAVDR	RLNEQ	SEANLYRLLELDQPP.....	KADEDGPTGPKPVSFVTVKETVCPRPTR.....	QPPEL	CDFKENG.....	RV.....	KQCVG	VTVLN.....	QIK	DPLDIT.....	CNEVQGV	RRGR.....	P49933
P32196	METQRASLC	GRWSLW	LLLLGLV	VPSASAQALS	YREAVLR	RAVDR	RLNEQ	SEANLYRLLELDQPP.....	KADEDGPTGPKPVSFVTVKETVCPRPTR.....	QPPEL	CDFKENG.....	RV.....	KQCVG	VTVLN.....	QIK	DPLDIT.....	CNEVQGV	RRGR.....	P32196
P32195																			

Evolutionary Trace

0	PIC ETN.....C.....C.....C.....	0	PIC ET
5	PIC ETN.....C.....C.....C.....	5	PIC ET
10	PIC ETN.....C.....C.....C.....	10	PIC ET
15	PIC ETN.....C.....C.....C.....	15	PIC ET
20	PIC ETN.....C.....C.....C.....	20	PIC ET
25	PIC ETN.....X.C.....C.....C.....	25	PIC ET
30	PIC ETN.....X.C.....C.....C.....	30	PIC ET
35	PIC ETX.....N.....X.....XX.C.....X.C.....C.....	35	PIC ET
40	PIC ETX.....N.....X.....XX.C.....X.C.....C.....	40	PIC ET
45	PIC ETX.....X.....N.....X.X.XX.....X.....XXC.....X.C.....X.C.X.....C.....	45	PIC ET
50	PIC ETX.....X.....N.....X.X.XX.....X.....XXC.....X.C.....X.C.X.....C.....	50	PIC ET
55	PIC ETX.....X.XN.....X.X.XX.....X.....XXC.....X.C.....X.C.X.....C.....	55	PIC ET
60	PIC ETX.....X.XN.....X.X.XX.....X.....XXC.....X.C.....X.C.X.....C.....	60	PIC ET
65	PIC ETX.....X.XN.....X.X.XX.....X.....XXC.....X.C.....X.C.X.....C.....	65	PIC ET
70	PIC ETX.....X.XN.....X.X.XX.....X.....XXC.....X.C.....X.C.X.....C.....	70	PIC ET
75	PIC ETXX.X.XXX.XXN.....X.X.X.XXXXXXX.....XX.XX.XXXXXC.....XXC.XXX.....X.C.XXX.X.....X.....C.....	75	PIC ET
80	PIC ETXXX.X.XXXXXXXN.....XXX.X.XXXXXXX.XX.....XX.....XX.XXX.XXXXXC.X.....XXCXXXX.....X.C.XXXXX.....XX.....C.....	80	PIC ET
85	PIC ETXXX.X.XXXXXXXN.XXXX.XXXXXXXXXXXXXX.....XXX.XXXXXXXX.XXXXXCXX.X.....XXXXCXXXX.....X.CXXXXXX.....XX.....CX.....	85	PIC ET
90	PIC ETXXX.X.XXXXXXXN.XXXX.XXXXXXXXXXXXXX.....XXX.XXXXXXXX.XXXXXCXX.X.....XXXXCXXXX.....X.CXXXXXX.....X.....XXX.....CX.....	90	PIC ET
95	PIC ETXXXXXXXXXXXXXXN.XXXXXXXXXXXXXXXXXX.....XXXXXXXXXXXXX.XXXXXCXX.X.....XXXXCXXXX.....X.CXXXXXX.....XXXXXXXXXX.....CX.....	95	PIC ET

Figure 6.5. A ‘short’ log file from the program ‘BRUTUS’ showing the analysis of the cystatin region of spp24 and homologous proteins.

As the PIC level increases (*i.e.* the sequence identity within each subgroup increases) and more subgroups are defined, more and more residues are identified that are conserved within each subgroup, but vary between subgroups. These are shown as 'Xs' in figure 6.5. Clusters of residues begin to develop that indicate regions of the proteins that are functionally important and therefore change as the subgroups become more specific.

Scripts were written for 'BRUTUS' (by Jon Clayton, University of Leicester) that could map the residues appearing at each PIC level onto a known 3D-structure. The 3D-structure for chicken egg white cystatin (PDB identification 1cewi, the structure with the closest homology to spp24) was viewed in the Swiss-PdbViewer v3.7b2 (<http://www.expasy.ch/spdbv/>) (Guex and Peitsch 1997). The 'BRUTUS' scripts were then run and a sequence of images stored at each PIC interval where new residues appeared. The images are presented in figure 6.6.

In figure 6.6 the residues that are completely conserved in all sequences are coloured green and can be seen at all PIC levels. Class-specific residues are then coloured in red as they appear at each PIC level. Once a PIC level of approximately 80% is exceeded, there is too much 'background noise' and nearly the entire model becomes red. Most of the residues appear between 35 and 80% PIC. At 35% PIC there are 9 subgroups and by 80% PIC there are 18 subgroups comprising a total of 38 proteins.

There are two striking clusters (depicted in red) that appear on the 3D-structure of '1cewi'. These can be seen most clearly in the B(90°) or C(180°) images for one cluster beginning at 35% PIC, which appears to form strip in the groove around the middle of the structure and the D(270°) images for the other cluster beginning again at 35% PIC, which appears to form a strip going up the back of the structure and across the top.

Most of the clustering is seen around the conserved cysteine residues and appears to be in the upper half of the structure. Much of the bottom of the cystatin structure remains white suggesting this region is not crucial to function. The regions of clustered residues indicate regions of the protein that are critical to function and that change to become specific to the function of each subgroup.

Figure 6.6 shows the clusterings in a single colour of red. However, to relate this to regions of the spp24 protein sequence, figure 6.7 shows the core residues of each cluster coloured according to the region of spp24 protein sequence that they are from, depicted in figure 6.8.

FACE

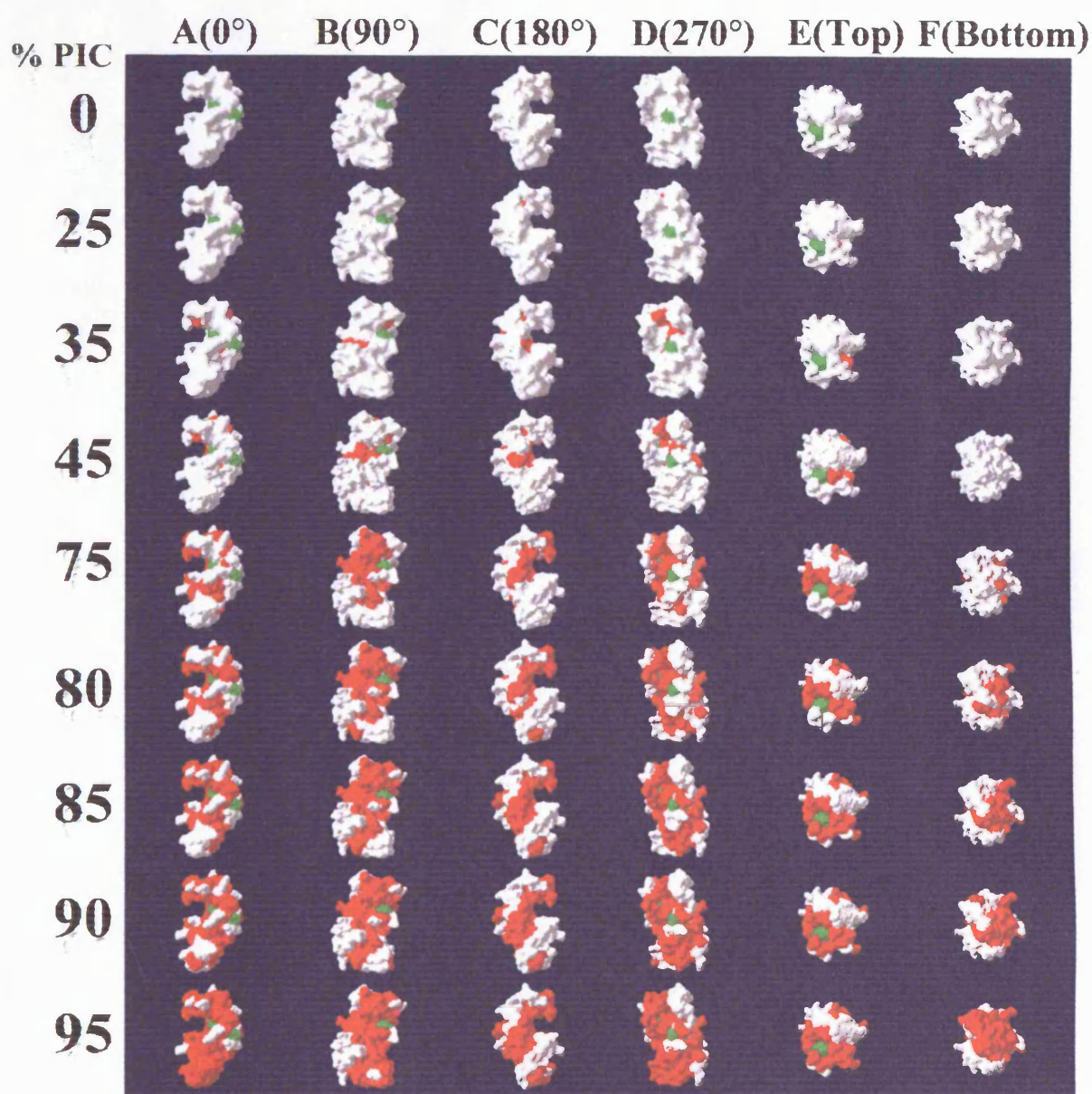


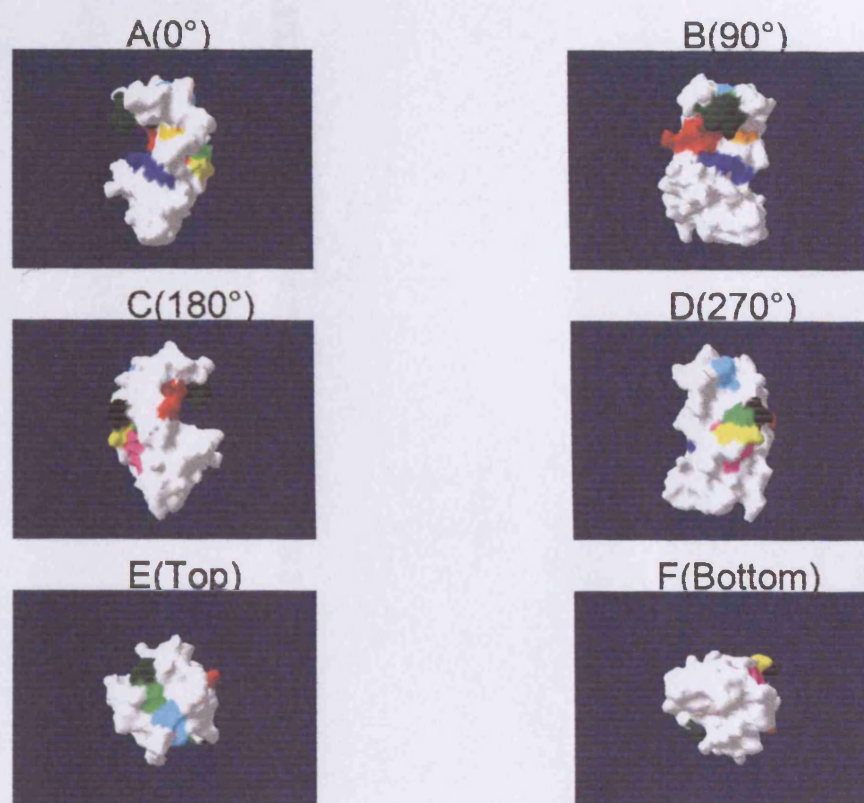
Figure 6.6. The images of '1cewi' from each PIC level where residues appear in the 'BRUTUS' program.

The structure of '1cewi' (chicken egg white cystatin) (Bode *et al.* 1988) is shown in four orientations about the vertical axis (A to D) and two about the horizontal axis (E and F). The residues identified as being absolutely conserved in all sequences are indicated in green, the class-specific residues are indicated in red and the neutral residues are indicated in white. The percent PIC interval is indicated at the left-hand side of the images. The green residues are the absolutely conserved cysteine residues. The red residues are the residues that are conserved within subgroups but not conserved between subgroups.

Images produced in Swiss-PdbViewer v3.7b2 (<http://www.expasy.ch/spdbv/>) (Guex and Peitsch 1997).

Figure 6.7. The structure of '1cewi' (chicken egg white cystatin) (Bode *et al.* 1988) showing the residues that are absolutely conserved and the core residues of the clusters that are class-specific in the evolutionary trace analysis.

The structure of '1cewi' (chicken egg white cystatin) (Bode *et al.* 1988) is shown in four orientations about the vertical (A to D) and two about the horizontal (E and F). The region of sequence corresponding to each colour is detailed in the key. The core residues from each of the six clusters seen in the 'BRUTUS' short log output (figure 6.5) were highlighted in different colours. Each of these regions of sequence are placed on the '1cewi' model so that they begin to form the two main clusters that are observed in figure 6.6. To relate these regions of amino acids to their location in the spp24 sequence see figure 6.8.



Key

Light green	conserved cysteine residues
Brown	conserved asparagine residue
Light blue	region around first cysteine 'ETTC'
Dark green	region around the second cysteine
Pink	region around the third cysteine 'CRSTV'
Yellow	region around the last cysteine
Dark blue	cluster around the 'ALSASV' region of spp24
Red	cluster around the 'FRAFRSS' region of spp24
Black	the C-terminal end of the cystatin domain

Figure 6.7. The structure of '1cewi' (chicken egg white cystatin) (Bode *et al.* 1988) showing the residues that are absolutely conserved and the core residues of the clusters that are class-specific in the evolutionary trace analysis. Images produced in Swiss-PdbViewer v3.7b2 (<http://www.expasy.ch/spdbv/>) (Guex and Peitsch 1997).

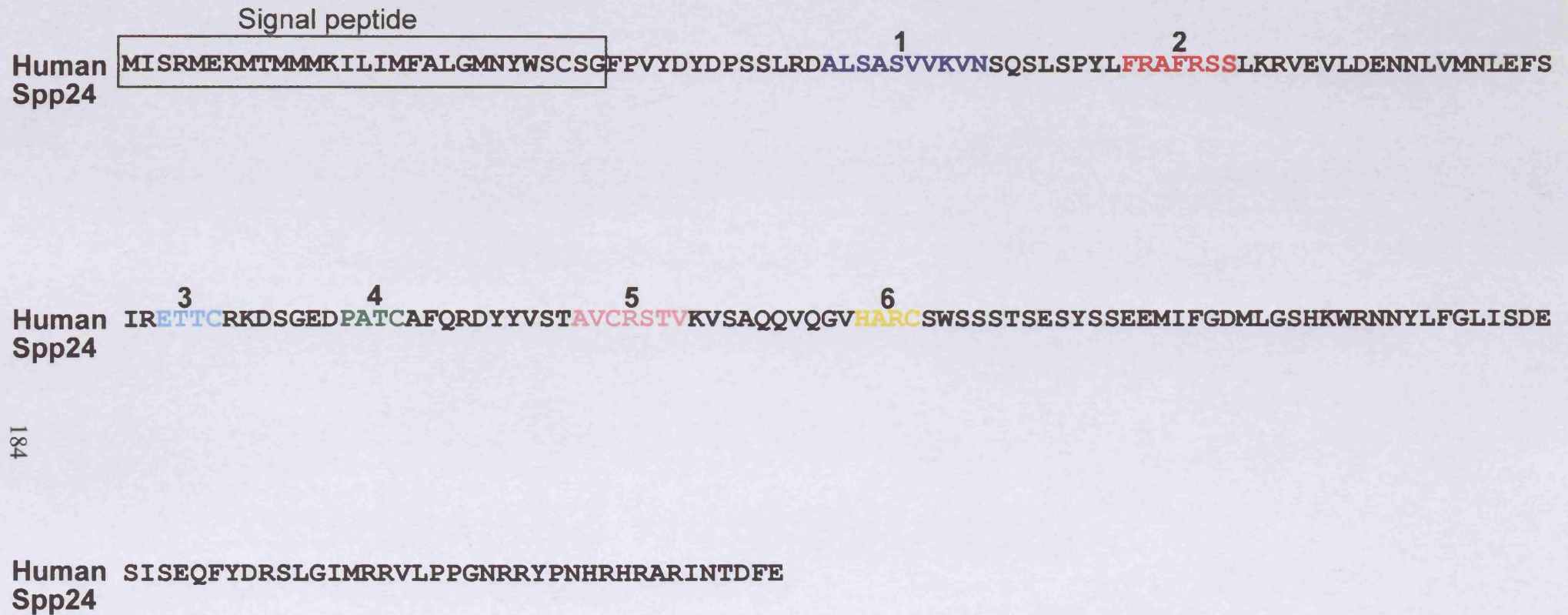


Figure 6.8. The six clusters in the human spp24 amino acid sequence thought to be functionally important.

Analysis with the 'BRUTUS' program (Jon Clayton, unpublished) revealed six clusters in the aligned protein sequences that were variant between subgroups, but invariant within subgroups. The regions of human spp24 corresponding to the core of each of these clusters are coloured in the same colours that were used for each region in figure 6.7. The signal peptide is boxed.

This figure also shows the absolutely conserved cysteine residues, the absolutely conserved asparagine residue and the C-terminal end of the cystatin domain.

6.3 Discussion

The search for proteins showing homology to spp24 has identified proteins from the kininogen, cathelicidin and fetuin families. All of these families are members of the cystatin superfamily and proteins from all of these families were originally reported by Hu *et al.* (1995). This chapter reports an expansion of proteins within these groups.

HSF is an antihemorrhagic factor isolated from the venom of the Japanese Habu snake, *Trimeresurus flavoviridis* (Yamakawa and Omori-Satoh 1992) that falls into the fetuin family, but is the only protein that displays a function not already discussed. HSF has been shown to have the ability to inhibit metalloproteinases (Yamakawa and Omori-Satoh 1992).

The cystatin-like region of spp24 is most similar to the second HSF cystatin-like domain, which is the domain thought to be responsible for the metalloproteinase inhibitory activity. However, the regions of this domain thought to be responsible do not show any homology to spp24. Spp24 shows no homology to any other protein known to have metalloproteinase inhibitory function. It is therefore unlikely that this is a possible function of spp24.

All of the proteins that show homology to spp24 (table 6.2) have one or more cystatin-like domains followed by a non-cystatin-like domain. None of the non-cystatin-like domains show any homology to spp24 and in fact are themselves quite divergent within groups of related proteins. It could therefore be speculated that the non-cystatin-like regions are not crucial to the function of the protein due to their lack of conservation but, are important in the 'fine-tuning' of protein function resulting in specificity within groups of proteins.

The only protein with just a cystatin domain that shows the highest level of homology to spp24 out of all the typical cystatins is CMAP (Cystatin-like Metastasis-Associated Protein). This is also known as cystatin F, cystatin 7 or leukocystatin (OMIM entry 603253). CMAP was identified as a metastasis-associated protein involved in liver metastasis (Morita *et al.* 1999) although the exact mechanism of its involvement is not yet understood. It is possible that spp24 could be associated with metastasis.

A company called Incyte Genomics Inc. (from which results were purchased in Chapter 4) make available for purchase human expression data from microarray studies carried out in-house. When their expression database (LifeExpress Online, www.incyte.com/lifeexpress/) was searched for *SPP2*, several results were available for purchase. Many of these results

were a comparison between normal tissue and tissue from a tumour. Tumour comparison results were available for lung, breast epithelial, colon, ovary, uterus and prostate. Many of these results were from secondary tumours originating in bone.

It was not practical to buy all of these results and consequently only an osteoblast result thought to be the most relevant was purchased (Chapter 4). However, the fact that results for these tumour tissues were available for purchase suggests that in these tumour tissues a change in the level of *SPP2* expression was seen. However, this should be cautiously assumed as the results may also show no alterations in levels as seen with the osteoblast results (Chapter 4). This, along with the similarity to CMAP, is the first hint that spp24 may be involved in metastasis. This is a possibility that should be kept in mind for future studies.

The computer-based analysis of the spp24 protein did not reveal anything unusual or unexpected. The protein is predicted to be phosphorylated as expected. The region of low complexity (serine-rich) is highly phosphorylated and also highly conserved between species. It is therefore likely that this region is critical to the function or regulation of the spp24 protein. There are also other residues, mainly in the cystatin-like region, that are predicted to be phosphorylated and are highly conserved between species. These residues could also be critical to the function or regulation of the protein.

The protein has a fairly even charge distribution with no obvious clusters of charge. Had there been any clusters of charge, this could have indicated regions of the protein that may have an increased affinity for a positively or negatively charged ion. Or, as is the case in some cathelicidins, there may have been a short region of the non-cystatin-like domain that is highly charged to aid in penetration of the membrane of bacteria (Wu *et al.* 1999). However, regions of high charge will only really be clear when the structure of the protein has been determined to see if charged regions that appear to be separated in the primary structure are forced together at any point in the tertiary structure to form a cluster.

In an attempt to model the cystatin-like region of the spp24 protein, an evolutionary trace method was employed. The 3D-structure chosen ('1cewi', chicken egg white cystatin) was thought to be the most appropriate for several reasons. First, the consensus secondary structure of the cystatin-like domain of spp24 determined by NPS@ was found to be fairly similar to that determined for '1cewi'. Therefore, it was thought likely that this region of spp24 folded in a similar manner to '1cewi'. Secondly, the most similar 3D-structure

identified by the program 3D-PSSM was '1cewi' and finally all of the proteins selected to be analysed were cystatin-like domains and '1cewi' is considered a 'typical' cystatin domain.

The evolutionary trace analysis of residues variant between subgroups, but invariant within subgroups when mapped onto the '1cewi' structure revealed two main clusters (figure 6.6). These clusters are made up of the regions of amino acid sequence that form six clusters seen in the 'short' log from the 'BRUTUS' program (figure 6.5 and 6.7). These are the regions of spp24 that are thought to be responsible for giving the protein its functional specificity. These regions in the amino acid sequence of human spp24 are shown in figure 6.8.

The residues that are known to be critical to cystatin anti-protease activity (*i.e.* the interaction of cystatin with papain, Chapter 1) are located at the bottom of the '1cewi' structure (view F in figures 6.6 and 6.7). There are no clusters of residues coloured red seen in this region suggesting that a cystatin anti-protease activity is not a function that differentiates the functional properties of the subgroups of proteins.

In theory, it should be possible to see which regions of the protein change as you move up the tree (*i.e.* through increasing PIC levels) to smaller subgroups and functions are lost. Consequently, determining the regions of the protein that are responsible for which functions. However, in practice it is very difficult. Many of the proteins used in this study are considered multifunctional and there is still some disagreement concerning what those functions are.

For example, class-specific residues only begin to appear at 25% PIC (*i.e.* when the proteins within each subgroup are at least 25% similar) ('long' log from 'BRUTUS', Appendix B). The first class-specific residues that appear are in the cluster around the first cysteine and the third cysteine, 'ETTC' and 'CRSTV' respectively. At 25% PIC there are 7 different subgroups. Table 6.3 shows the 7 subgroups and the class-specific residues they have at the two clusters around the first and third cysteine.

It should be possible to compare the functions common to different groups with the residues that are common to different groups. However, as can be seen from table 6.3, it is not possible to assign a common function to all subgroups due to the lack of knowledge of the individual protein functions especially in terms of which cystatin domain is responsible for which function in each protein.

Table 6.3. The subgroups at 25% PIC and the residues they have at the cluster around the first and third cysteine.

Class-specific residues only begin to appear at 25% PIC (*i.e.* when the proteins within each subgroup are at least 25% similar) ('long' log from 'BRUTUS', Appendix C). The first class-specific residues that appear are in the cluster around the first cysteine and the third cysteine, 'ETTC' and 'CRSTV' respectively in spp24 (the class-specific residues at 25% PIC are shown in bold).

At 25% PIC there are 7 different subgroups. The table shows which residue was at each position for each subgroup. A function common to that subgroup was assigned if possible. For each of the subgroups that showed the same residues at each or one position, it was not possible to assign a common function.

Subgroup	Class-specific residue at cluster around first cysteine	Class-specific residue at cluster around third cysteine	Common function to the subgroup	Common function
Cystatin domains 3 Typical cystatins (8 members)	'T'	'V'	Some cystatin	?
Spp24 (5 members)	'T'	'V'	?	
Cathelicidins (11 members)	'T'	'V'	Antibacterial	
Cystatin domains 2 (6 members)	'T'	'A'	Some cystatin	?
HSF cy1 (1 member)	'T'	'I'	?	
Cystatin domains 1 (6 members)	'G'	'V'		?
HSF cy2 (1 member)	'I'	'V'	Metalloproteinase inhibitor	

From the evolutionary trace analysis it is therefore not possible to give a detailed analysis of which residues are responsible for which changes in function. However, it is possible to say that there are two main regions on the cystatin-like domain structure that are likely to be important for the function of the protein and the corresponding regions of the spp24 protein sequence can be identified (figure 6.8).

The C-terminal end of the cystatin-like domain is located near the two main clusters. The serine-rich and non-cystatin-like region of spp24 would be attached here. It is therefore speculated that the non-cystatin like region may lie across one of these two regions and the degree of phosphorylation at the serine-rich region could regulate where exactly it lies.

It is possible that spp24 is self-inhibitory, with the non-cystatin-like region able to block one of the functionally important regions. These possibilities are shown in figure 6.9.

Alternatively, the non-cystatin-like region could be cleaved off at the end of the cystatin domain, which is exposed on the surface of the molecule, or within the phosphorylated region to release a biologically active peptide or to convert a pro-protein to its active form by releasing the blocking peptide. The extent of phosphorylation in the serine-rich region could be a means of regulating one of these possibilities.

A further possible function for spp24 emerged that was not highlighted by looking at protein homologies but, was deduced from a report by Alvarez-Fernandez *et al.* (1999). It was reported that some cystatins are able to inhibit the asparaginyl endopeptidase legumain due to a second novel reactive site (Alvarez-Fernandez *et al.* 1999). Cystatins C, E and F (CMAP), but not A, B, D or kininogen domains 2 and 3 were shown to be able to inhibit mammalian legumain. Mammalian legumain is an endopeptidase that hydrolyses asparaginyl bonds (Chen *et al.* 1997). Some cystatins inhibit legumain by providing an alternative substrate and thus taking up the active site of the enzyme. In cystatin C, E and F (CAMP) the site thought to be responsible for legumain inhibition is 'SNDM', 'SNSI' and 'TNDM' respectively, with cleavage at the asparaginyl bond.

Spp24 does not have an 'N' at the corresponding position, but it does have an 'N' close by. It is therefore possible that spp24 can also inhibit legumain. It is interesting that spp24 has a speculated role in bone turnover and that legumain has been identified as being an inhibitor of osteoclast formation and bone resorption (Choi *et al.* 1999).

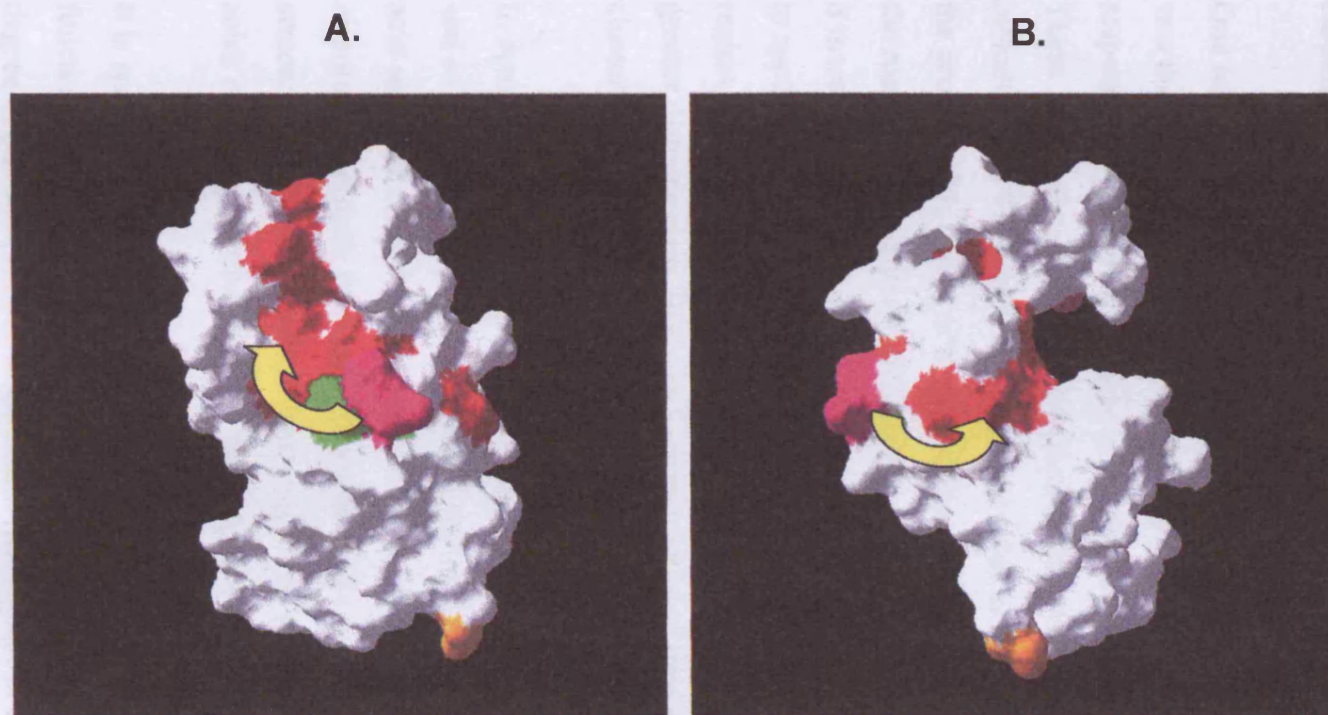


Figure 6.9. The position in which the non-cystatin-like region of spp24 could lie on the cystatin-like domain.

Two views of the '1cewi' structure are shown. The residues shown in red are the core residues of the two main clusters thought to be of functional importance in the cystatin-like region of the spp24 protein. The green residue is the third cysteine, the pink residue marks the C-terminal end of the cystatin domain and the orange residue marks the N-terminal end of the cystatin domain.

The arrow indicates the two possible directions in which it is speculated the non-cystatin-like region of spp24 could fold. It may fold up over the back of the molecule to lie across one cluster (as seen in A) or it may fold round into the groove in the middle of the structure to lie across the other cluster (as seen in B).

Images produced in Swiss-PdbViewer v3.7b2 (<http://www.expasy.ch/spdbv/>) (Guex and Peitsch 1997).

However, the 'N' that lies close by in spp24 is the 'N' that is absolutely conserved throughout all of the proteins involved in the evolutionary trace analysis. It is unlikely that all of these proteins can inhibit legumain and so it is probable that this 'N' is critical for some other function. Also, the absolutely conserved 'N' is likely to be less well exposed on the surface of the molecule than the 'N' seen in cystatin F (CAMP) (figure 6.10) and consequently is probably less susceptible to cleavage by legumain. Therefore, it was concluded that spp24 is unlikely to inhibit legumain.

One region of the spp24 protein that did not appear in any cluster identified in the ET analysis was the first five residues of the mature protein, 'FPVDY'. The location of this region with respect to the proposed structure of the spp24 cystatin domain is shown in red in figure 6.11. These five amino acids of spp24 are absolutely conserved in all species studied so far, including pig. Therefore, it seems reasonable to assume that they are crucial to the function of the protein. This region of the protein is not conserved within other subgroups and so this is the reason it has not been identified in the ET analysis. If this region is important for function it must therefore be specific to spp24 and unrelated to any of the proteins that show homology to spp24. These residues form a 'strip' that could be a possible site for the non-cystatin-like region to lie as discussed earlier on in this section. However, this 'strip' of residues is a greater distance away from the C-terminal end of the cystatin-like domain than the two clusters identified in the ET analysis.

In summary, only the cystatin-like region of spp24 shows homology to any known proteins and so consequently this is the only region of the protein that can be modelled. At the amino acid sequence level there are six clusters of residues that look as though they may be important in protein function. The cystatin-like region of the spp24 protein is likely to form a structure similar to '1cewi' (chicken egg white cystatin) in which these six clusters of amino acids come together to form two main regions that are deemed functionally important.

It is speculated that the non-cystatin region of spp24 folds back to lie across one of the two functionally important regions. The way in which it folds back could be regulated by the degree of phosphorylation in the serine-rich region. In this way spp24 could be self-inhibitory. Alternatively, the non-cystatin-like region could be released as a biologically active peptide. A further region thought to be functionally important was identified that did not appear in the ET analysis. The first five residues of the mature spp24 protein are absolutely conserved between all species studied and are therefore likely to be crucial for the function of the protein.

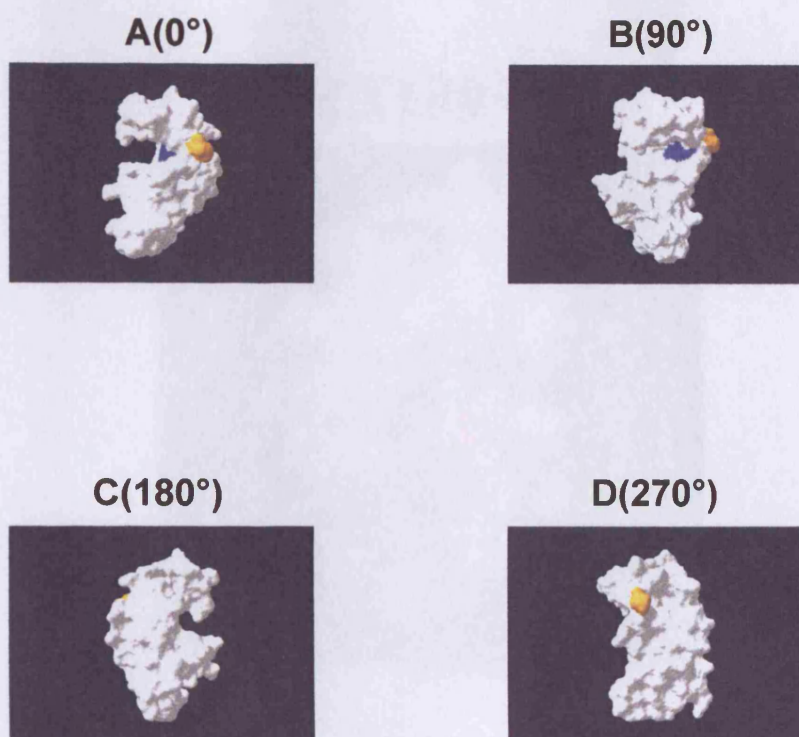


Figure 6.10. The structure of '1cewi' showing the position of the 'N' residue involved in legumain inhibition in CAMP and the 'N' residue located close by in spp24.

A second reactive site was identified in cystatin C, E and F (CAMP) that is thought to be involved in legumain inhibition. The 'N' where cleavage with legumain occurs in these proteins is shown in yellow.

Spp24 does not have an 'N' in the equivalent position, but it does have an 'N' close by. This 'N' is shown in dark blue and appears to be less accessible than the 'N' shown in orange.

Images produced in Swiss-PdbViewer v3.7b2 (<http://www.expasy.ch/spdbv/>) (Guex and Peitsch 1997).

C(180°)

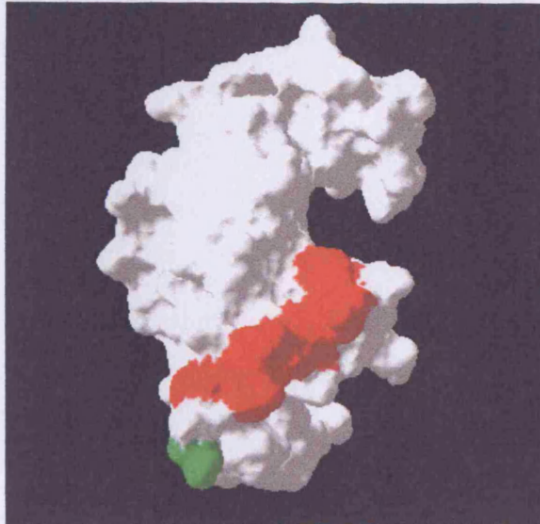


Figure 6.11. The location of the highly conserved N-terminal region of the mature spp24 protein.

The first five residues of the mature spp24 protein, 'FPVDY', are absolutely conserved between all species. The residues mapped onto the corresponding residues of the '1cewi' structure are shown in red. This is a region that does not form part of a cluster in the ET analysis.

Image produced in Swiss-Pdb Viewer v3.7b2 (<http://www.expasy.ch/spdbv/>) (Guex and Peitsch 1997).

Spp24 could have a cystatin-like function, a fetuin-like function or an antimicrobial function as discussed in Chapter 1. A further possibility is that it has an involvement in metastasis. It is thought unlikely that spp24 can inhibit legumain.

None of the work presented in this chapter results in firm conclusions, but it does identify regions of residues that can be targeted in mutation analysis once the functions of spp24 have been determined.

Chapter 7

Concluding remarks and future work

Unfortunately, this thesis seems to have raised more questions than it has answered. However, the structure and expression of the gene encoding spp24 has been successfully characterised and a direction for future functional work has been determined.

7.1 Concluding remarks

There does not seem to be any doubt now that spp24 is a new member of the cystatin superfamily. This was originally suggested by Hu *et al.* (1995) and all of the work presented in this thesis supports this. In terms of the number of cystatin-like domains and non-cystatin-like domains, spp24 fits into the cystatin superfamily with the cathelicidins (Chapter 5) coming after type II cystatins and before fetuins and kininogens in terms of domain complexity. This is depicted in figure 7.1. However, in terms of sequence homology to cystatin domains, spp24 is most closely related to domains 1 and 3 of kininogen and to the bovine neutrophil antibiotic peptide bactenecin (Hu *et al.* 1995). Hu *et al.* (1995) suggested that spp24 was an evolutionary intermediate between these two proteins. This is supported by the evolutionary trace (ET) analysis carried out in Chapter 5, where figure 5.4 presents an evolutionary tree and clearly shows spp24 to be at an intermediate level between the cathelicidins and the fetuins, kininogens and type II cystatins.

Although spp24 has been deemed a member of the cystatin superfamily it is unlikely to exhibit a typical cystatin function by inhibiting thiol proteases from the papain superfamily. Spp24 does not contain the residues identified as being crucial to the cystatin-papain interaction (Chapter 1) although its ability to inhibit papain should be tested experimentally before it is totally dismissed.

The structure of the spp24 gene in both human and mouse is unlike any other member of the cystatin superfamily in that it has an additional, very small, intron splitting what would otherwise be the first exon. This immediately makes it a unique member.

The expression pattern of the spp24 gene is similar to that of fetuin as it is expressed predominantly in liver. This suggests that spp24 is a plasma protein or that it has a role in processes that take place in the liver. It also seems that, like fetuin, spp24 has a role in the

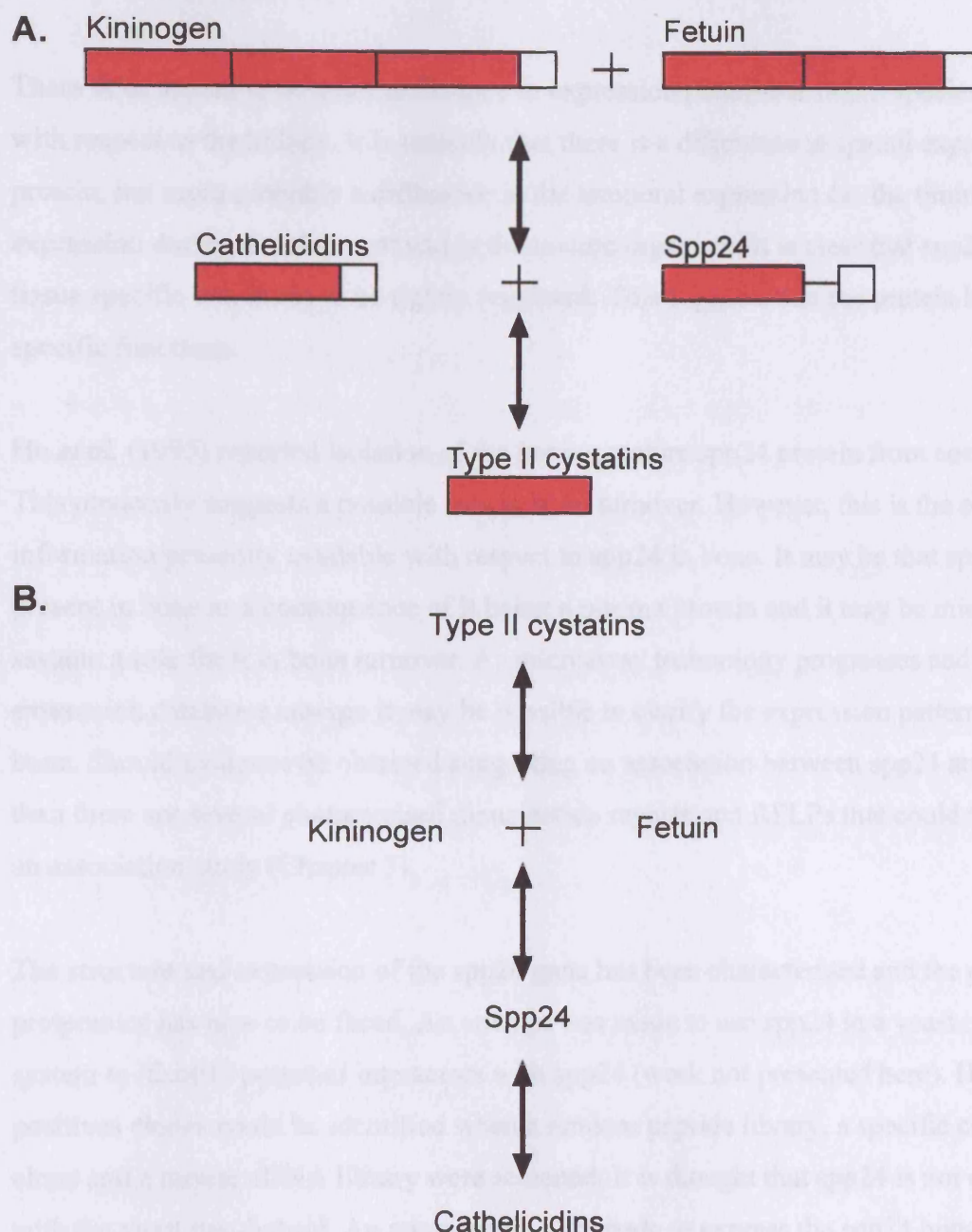


Figure 7.1. Spp24 as an evolutionary intermediate.

Spp24 is shown as an evolutionary intermediate, with respect to domains, between kininogen and type II cystatins in A. A red box depicts a cystatin-like domain, a white box a non-cystatin-like domain and a single line a serine-rich region. In B, spp24 is shown as an evolutionary intermediate, with respect to sequence homology of the cystatin-like domain, between kininogen and cathelicidins.

development of a particular population of cells in the cerebellum at a specific stage of development (Chapter 4).

There does appear to be some difference in expression patterns between species, particularly with respect to the kidney. It is unlikely that there is a difference in spatial expression of the protein, but more probably a difference in the temporal expression *i.e.* the timing of expression during development and in the mature organism. It is clear that spp24 is highly tissue specific and likely to be tightly regulated. This suggests that the protein has very specific functions.

Hu *et al.* (1995) reported isolation of the bovine mature spp24 protein from cortical bone. This obviously suggests a possible role in bone turnover. However, this is the only information presently available with respect to spp24 in bone. It may be that spp24 is simply present in bone as a consequence of it being a plasma protein and it may be misleading to assume a role for it in bone turnover. As microarray technology progresses and more precise expression databases emerge it may be possible to clarify the expression pattern of spp24 in bone. Should evidence be obtained suggesting an association between spp24 and a disease then there are several characterised dinucleotide repeats and RFLPs that could be used in such an association study (Chapter 3).

The structure and expression of the spp24 gene has been characterised and the problem of proteomics has now to be faced. An attempt was made to use spp24 in a yeast two-hybrid system to identify potential interactors with spp24 (work not presented here). However, no positives clones could be identified when a random peptide library, a specific cathepsin K clone and a mouse cDNA library were screened. It is thought that spp24 is not compatible with the yeast two-hybrid. An attempt was also made to express the spp24 human protein in a baculovirus protein expression system (work not presented here). However, only small amounts of insoluble protein were expressed. It is thought that the lack of a signal peptide may have posed a problem or that the protein required post-translational modifications that simply could not be achieved in this system.

The work presented in this thesis has led to the following speculations about the function(s) of the spp24 protein:

- The protein may have a cystatin-like function although this is thought unlikely due to the lack of residues thought to be critical for an interaction with papain (Chapter 1).

- The protein may be circulated as a plasma protein like fetuin and therefore have a role in processes such as inflammation, coagulation, the immune response and mineralisation (Chapters 4).
- The protein may have an antimicrobial function like the cathelins. The non-cystatin-like domain may be responsible for this activity, as it is the most divergent domain between species (Chapter 6).
- Spp24 may have the ability to inhibit legumain like several other cystatins although this is thought unlikely as the critical asparagine is not thought to be very accessible on the surface of the molecule (Chapter 6).
- The non-cystatin-like region of spp24 could be released as a biologically active peptide with the cystatin-like domain acting as a carrier (Chapter 6).
- Spp24 could be self-inhibitory with the non-cystatin-like domain folding back to block functionally important residues on the cystatin-like domain (Chapter 6).
- The serine-rich region is thought likely to be a regulatory region with the extent of phosphorylation determining the functional state of the protein (Chapter 6).
- The spp24 protein may have a role in cancer and metastasis like the cystatin CMAP. It is tempting to think that results available from Incyte Genomics (Chapter 6) suggest this, but these results could in fact all be of no significance. It would only be possible to determine this by purchasing all of the available results.

In support of the speculation of an antimicrobial role, the mouse *Spp2* gene has recently been mapped to chromosome 1 adjacent to a susceptibility locus for tuberculosis (TB) (Khorram Khorshid and Dalgleish, unpublished).

7.2 Future work

It seems that to progress with the functional studies of spp24, a purified protein is required. Expression in insect or mammalian cells is likely to achieve the correct post-translational modifications and the signal peptide should be included to help achieve this. It is important to study the individual domains of the protein as well as the protein as a whole and so three constructs should be made, although all three should have a signal peptide included. Extensive optimisation of expression and purification conditions may be required. Inclusion of the signal peptide may mean that the proteins receive the correct post-translational modifications and are secreted into the supernate of the culture as has been seen with other cystatins. This would make the isolation of the protein easier.

Once a purified, soluble protein has been obtained several biochemical tests can be performed to assess the possible functional properties of the protein. Papain is commercially available and relatively inexpensive and so the ability of spp24 to inhibit papain could be easily determined. To test for an interaction with cathepsin K or inhibition of legumain would require expression of these proteins or donation by a collaborator. It would be relatively straightforward to test for antimicrobial properties by adding preparations of spp24 to agar plates containing various bacterial cultures. Of course all three spp24 constructs should be tested so that if a function is found, the responsible protein domain can be identified.

Investigating the role, if any, of spp24 in more complex processes such as inflammation, mineralisation and metastasis would require extensive expertise and therefore would probably involve collaborations.

The expression studies presented in Chapter 4 suggest that the gene encoding spp24 may be expressed in lactating mammary gland cells. If this is the case then it is possible that spp24 is secreted in milk. An investigation into the possible levels of spp24 in milk may reveal a source of the spp24 protein from which it may be easier to isolate than from bone.

An *in situ* hybridisation study is currently being performed by Hamid Khorram Khorshid (University of Leicester) in an attempt to determine the exact temporal and spatial expression of the mouse *Spp2* gene during mouse development. This should also be done with respect to the spp24 protein to determine its localisation, which is not necessarily identical to the expression of the gene. This would require the raising of antibodies against the spp24 protein. Antibodies against the protein would be useful in any of the functional studies so that the presence of spp24, and not just a similar sized protein, could always be confirmed.

The creation of a spp24 knockout mouse may be one way of highlighting a potential function of spp24. The structure of the mouse *Spp2* gene is now known (Chapter 3) enabling the creation of a knockout or transgenic mouse. A collaborator has been approached regarding this work. Of course there is always the possibility that the knockout mouse will be perfectly normal and have no obvious defects. However, the mouse model would still be available to test for other possible abnormalities. For example, should spp24 have antimicrobial properties and given the mapping of the mouse *Spp2* gene to the same region as a possible TB susceptibility locus it would be possible to investigate the susceptibility of the knockout mouse to TB and other infections compared to a normal mouse.

Once the function of the spp24 protein has been identified it will be useful to investigate the regulation of the protein. Promoter studies should be carried out to determine the exact region required for transcription and to determine whether the gene is regulated by growth hormone (Chapter 5). It is also possible that the action of the spp24 protein is regulated by the extent of phosphorylation in the serine-rich region of the protein. However, it would not be possible to alter the degree of phosphorylation without denaturing the protein so this would be very difficult to investigate.

If it is suspected from spp24 functional studies that spp24 could have a potential role in a multifactorial disease, then an association study could be performed. This would almost certainly require collaborations and association studies should be well thought out. There are three characterised RFLPs and three characterised tandem repeats in the human *SPP2* gene region (Chapter 3) that could be used in any future association studies.

Once a function has been identified for spp24, mutation analysis could be performed to identify exact regions of the protein that are responsible. The ET analysis performed in Chapter 6 may provide target residues for these studies although determination of the 3D-structure of the spp24 protein by NMR or X-ray diffraction would obviously be more accurate.

It is easy to see that the groundwork for investigations of spp24 has been done, but the emphasis for the future is shifting from the gene to the protein. In the advent of the human genome project this is not uncommon and expertise is now required in the field of proteomics.

Appendix A

LOCUS HSA272265 30000 bp DNA PRI 16-FEB-2000
DEFINITION Homo sapiens SPP2 gene for secreted phosphoprotein 24 precursor,
exons 1-8.
ACCESSION AJ272265
VERSION AJ272265.1 GI:6996452
KEYWORDS secreted phosphoprotein 24 precursor; SPP2 gene.
SOURCE human.
ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE 1 (bases 1 to 30000)
AUTHORS Dalglish,R.W.M. and Bennett,C.S.
TITLE Human SPP2
JOURNAL Unpublished
FEATURES 2 (bases 1 to 30000)
AUTHORS Dalglish,R.W.M.
TITLE Direct Submission
JOURNAL Submitted (14-FEB-2000) Dalglish R.W.M., Department of Genetics,
University of Leicester, University Road, Leicester LE1 7RH, United Kingdom
COMMENT related sequences AC006037, U20530.
FEATURES Location/Qualifiers
source 1..30000
/organism="Homo sapiens"
/db_xref="taxon:9606"
LTR 1..249
/note="retroviral"
/evidence=not_experimental
repeat_region complement(250..616)
/note="L1, L1PA7"
/rpt_family="LINE"
/evidence=not_experimental
repeat_region 616..1464
/note="L1, 2L1PA22"
/rpt_family="LINE"
/evidence=not_experimental
LTR 1469..1532
/note="retroviral"
/evidence=not_experimental
repeat_region complement(1556..1658)
/note="MIR"
/rpt_family="SINE"
/evidence=not_experimental
exon 2320..2489
/gene="SPP2"
/number=1
mRNA join(2320..2489,2589..2713,10454..10576,11987..12097,18151..18205,18842..18892,21546..21641,28463..28750)
/gene="SPP2"
/product="secreted phosphoprotein 24 precursor"
5'UTR 2320..2405
/gene="SPP2"
/evidence=experimental
gene 2320..28750
/gene="SPP2"
sig_peptide join(2405..2489,2589..2590)
/gene="SPP2"
CDS join(2405..2489,2589..2713,10454..10576,11987..12097,18151..18205,18842..18892,21546..21631)
/gene="SPP2"
/codon_start=1

Appendix A

```
/product="secreted phosphoprotein 24 precursor"  
/protein_id="CAB75571.1"  
/db_xref="GI:6996453"
```

```
/translation="MISRMEKMTMMMILIMFALGMNYWSCSGFPVYDYPSSLRDAL
```

```
SASVVKVNSQSLSPYLFRAFRSSLKRVEVLDENNLVMNLEFSIRETTCTRKDSGEDPAT
```

```
CAFQRDYYVSTAVCRSTVKVSAQQVQGVHARCSWSSSTSESYSSEEMIFGDMLGSHKW
```

```
      RNNYLFGLISDEISIQFYDRSLGIMRRVLPPGNRRYPNHRHRARINTDFE"  
intron      2490..2588  
            /gene="SPP2"  
            /number=1  
exon        2589..2713  
            /gene="SPP2"  
            /number=2  
mat_peptide  
join(2591..2713,10454..10576,11987..12097,18151..18205,  
      18842..18892,21546..21633)  
            /gene="SPP2"  
intron      2714..10453  
            /gene="SPP2"  
            /number=2  
repeat_region 4657..4756  
            /note="L1, L1MA2"  
            /rpt_family="LINE"  
            /evidence=not_experimental  
repeat_region 5061..5229  
            /note="MER5B"  
            /rpt_family="MER1"  
            /evidence=not_experimental  
repeat_region complement(5503..5586)  
            /note="HERVFB21"  
            /evidence=not_experimental  
LTR         complement(6036..6361)  
            /note="LTR/MaLR"  
            /evidence=not_experimental  
LTR         complement(6685..7154)  
            /note="LTR/Retroviral (MLT2FB)"  
            /evidence=not_experimental  
repeat_region complement(7529..7673)  
            /note="L1, L1PA4"  
            /rpt_family="LINE"  
            /evidence=not_experimental  
repeat_region 7674..7986  
            /note="L1, L1P2"  
            /rpt_family="LINE"  
            /evidence=not_experimental  
repeat_region 8724..9226  
            /note="L2"  
            /rpt_family="LINE"  
            /evidence=not_experimental  
repeat_region 9527..10176  
            /note="MER82"  
            /rpt_family="MER2"  
            /evidence=not_experimental  
repeat_region 10182..10289  
            /note="MIR"  
            /rpt_family="SINE"  
            /evidence=not_experimental  
exon        10454..10576  
            /gene="SPP2"  
            /number=3  
intron      10577..11986  
            /gene="SPP2"
```

Appendix A

```

repeat_region    /number=3
                  10770..10836
                  /note="MIR"
                  /rpt_family="SINE"
                  /evidence=not_experimental
LTR              complement(11061..11419)
                  /note="Retroviral"
                  /evidence=not_experimental
exon             11987..12097
                  /gene="SPP2"
                  /number=4
intron           12098..18150
                  /gene="SPP2"
                  /number=4
repeat_region    12310..12333
                  /note="(CT)n repeats"
                  /evidence=experimental
repeat_region    12334..12412
                  /note="(CA)n repeats"
                  /evidence=experimental
repeat_region    12551..12731
                  /note="MIR"
                  /rpt_family="SINE"
                  /evidence=not_experimental
repeat_region    12820..12971
                  /note="L2"
                  /rpt_family="LINE"
                  /evidence=not_experimental
repeat_region    complement(13719..14088)
                  /note="L2"
                  /rpt_family="LINE"
                  /evidence=not_experimental
repeat_region    15757..15843
                  /note="HERVFB21"
repeat_region    15861..15890
                  /note="(TGTCTC)n repeats"
repeat_region    complement(16187..16900)
                  /note="L1, L1MB7"
                  /rpt_family="LINE"
                  /evidence=not_experimental
repeat_region    complement(16901..17196)
                  /note="AluSx"
                  /rpt_family="Alu"
                  /evidence=not_experimental
repeat_region    complement(17338..17428)
                  /note="L1, L1P_MA2"
                  /rpt_family="LINE"
                  /evidence=not_experimental
repeat_region    complement(17430..17595)
                  /note="L1, L1ME_ORF2"
                  /rpt_family="LINE"
                  /evidence=not_experimental
exon             18151..18205
                  /gene="SPP2"
                  /number=5
intron           18206..18841
                  /gene="SPP2"
                  /number=5
exon             18842..18892
                  /gene="SPP2"
                  /number=6
intron           18893..21545
                  /gene="SPP2"
                  /number=6
repeat_region    19114..19425

```

Appendix A

```

/feature="AluY"
/feature="Alu"
/feature="not_experimental"
repeat_region complement(19558..19644)
/feature="L1, L1MA3"
/feature="LINE"
/feature="not_experimental"
repeat_region 19838..19978
/feature="MIR"
/feature="SINE"
/feature="not_experimental"
repeat_region complement(20216..20443)
/feature="MARNA"
/feature="not_experimental"
LTR complement(20444..20964)
/feature="LTR/MaLR (MLT1F)"
/feature="not_experimental"
repeat_region 20804..20833
/feature="(AG)n repeats"
/feature="experimental"
repeat_region complement(20931..21006)
/feature="MLT1F"
/feature="not_experimental"
exon 21546..21641
/feature="SPP2"
/number=7
3'UTR join(21632..21641,28463..28750)
/feature="SPP2"
intron 21642..28462
/feature="SPP2"
/number=7
repeat_region complement(22087..22379)
/feature="AluSq"
/feature="Alu"
/feature="not_experimental"
repeat_region 22418..22804
/feature="L2"
/feature="LINE"
/feature="not_experimental"
repeat_region 23278..23722
/feature="L2"
/feature="LINE"
/feature="not_experimental"
repeat_region 25025..25113
/feature="L2"
/feature="LINE"
/feature="not_experimental"
repeat_region complement(25507..25561)
/feature="L2"
/feature="LINE"
/feature="not_experimental"
repeat_region 26232..26896
/feature="MER82"
/feature="MER2"
/feature="not_experimental"
repeat_region 27085..27415
/feature="L2"
/feature="LINE"
/feature="not_experimental"
exon 28463..28750
/feature="SPP2"
/number=8
polyA_signal 28728..28733
/feature="SPP2"
polyA_site 28750

```

Appendix A

```
repeat_region /gene="SPP2"  
/evidence=experimental  
28774..28800  
/note="trinucleotide; (GTT)n repeats"  
repeat_region /evidence=experimental  
complement(28993..29290)  
/note="AluY"  
/rpt_family="Alu"  
repeat_region /evidence=not_experimental  
29323..29593  
/note="L1,L1MC2"  
/rpt_family="LINE"  
/evidence=not_experimental
```

SPECIAL NOTE

**This item is tightly bound
and while every effort has
been made to reproduce the
centres force would result
in damage.**

[illegible]

Appendix B

```
1 2 members P01042cy3, P01043cy3
2 2 members P01042cy2, P01043cy2
3 2 members P01042cy1, P01043cy1
4 1 member P01044cy3
5 4 members P01044cy2, P01046cy2, P01045cy2, P01047cy2
6 4 members P01044cy1, P01046cy1, P01045cy1, P01047cy1
7 3 members P01046cy3, P01045cy3, P01047cy3
8 1 member H9FCY2
9 1 member H9FCY1
10 1 member H9SPF24_ID
11 1 member H9SPF24_ID
12 2 members H9SPF24_ID, H9SPF24_TR
13 1 member CHICK_my_c
14 2 members P80054_ID, P15175_ID
15 5 members P49934_ID, P49935_ID, P32196_ID, P32195_ID, P32194_ID
16 1 member P54229_ID
17 1 member P49930_ID
18 1 member P22226_ID
19 1 member P51437_ID
20 1 member icawi
21 1 member CMAP_also
...PPTKICVG...CPRDIPNSP.....ELEETLTHTITKLAEMNATFYFKIDNVQOAR..VQVVGAKKYFIDFVARETTSCSKESENEELTESCETHKLG.QS..LDCNAEYVVP...WEKKIYPTVN.CQPLGMSIMK... Group 1
...TAGTDCIG...CVHPISTQSP.....DLEPILRGIQTFNNNTQSSSLPMLNEVQOAR..RQVVGAGLNFRTITSIVGTNCSENFILTPDCKSLWNG.DT..GECTDNAYIDI...QLRIASFQW.CDIYFGKDPVQ... Group 2
...QESG.....SEKIDCKDK.....DLEKQVDAALAKYNGQSGSNQVFLYRIEAT..KTVGSDTFYSFYEIKGGDCPQSG.KTWQCKEYDAA.KAATSECTATVGHSS..STKFSVAYQT.CQITPAEGPVV... Group 3
...PPTALCAG...CKPKPIVDSP.....DLEEPILSHSIAKLNARHOGAFYFKIDTVKQAT..VQVVGAKKYSIYFIARETTSCSKESENEELTKSCENING.QI..LBCDANVYVP...WEKKIYPTVN.CQPLGQTSIM... Group 4
...TAQYECIG...CVHPISTKSP.....DLEEPILRYAIQYFNNNTSHSHFLDKCEVQAR..QVVGWNYEVNYSIAQTNCSENEELTTPDCKSLSSG.DT..GECTDKAJVDV...KLAISSFSQK.CDLTP..DF... Group 5
...QES.....SGEIDCNDQ.....DVFKAVIDAALTKYNSENKSGNQVFLYRIEVA..RQNDPDTFYSIKYQIKEGDCFFQSN.KTWQCKDYKDSA.QAATG.CTATVARG...NMKFSVAIQT.CLIITPAEGPVV... Group 6
...FPMQVQ...CKPKPIVDSP.....DLEELMHSIAKLNARHODTTTFKIDTVKQAT..VQVVGAKKYSIYFIARETTSCSKESENEELTKSCENING.QI..LBCDANVYVP...WEKKIYPTVN.CQPLGQTSIM... Group 7
...NCSR...CPILLPPNP.....HVVDSVEYVLKQDN.EKLSGHIYEVLZISRGQ..RKYPEEATYLEFVIVELNCTAQEADNDHQBCHPYTAG.EDNIAFCRSTVFRSHASLEKPKDEKTESDCVILDVNGHNSH... Group 8
...DQ.....VRGDLCCDK.....EAKNGADDAVRYINERKLGHKQALNVINONICVPMNGDLVAVFLEMLNLETECHVLPD.TFVEKCTVRQQNHAVENDCDAKIMFN..RTFKRDVFPK..CHSTPDSVENVR... Group 9
...FPVYDYP...SLRDALSASVVKVNSQSLSPYLFRAPRSSIKRVEVLDENNLVONLEFSIRETTCKRDSG.EDPATCAPQDY.YVSTAVCRSTVKVSA...QVQGVHAR..CSW... Group 10
...FPVYDYP...SLKEALSASVVKVNSQSLSPYLFRAPRSSIKRVEVLDENNLVONLEFSIRETTCKRDSG.EDPATCAPQDY.YVSTAVCRSTVKVSA...QVQGVHAR..CSW... Group 11
...FPVYDYP...SLKEALSASVVKVNSQSLSPYLFRAPRSSIKRVEVLDENNLVONLEFSIRETTCKRDSG.EDPATCAPQDY.YVSTAVCRSTVKVSA...QVQGVHAR..CSW... Group 12
...FPVYDYP...SLKEALSASVVKVNSQSLSPYLFRAPRSSIKRVEVLDENNLVONLEFSIRETTCKRDSG.EDPATCAPQDY.YVSTAVCRSTVKVSA...QVQGVHAR..CSW... Group 13
...L.YREAVLRVDRINEQSSSEANLYRLELDQPP.KADEDPOTPKRVSFTVKETVCPRPTR.QPPELCCDFKE.....KQCVSTVTLN...PSIHSLDIS..CNEIQSV... Group 14
METQASISLGRWSLWLLLLGVVPSASAQALSREAVLRVDRINEQSSSEANLYRLELDQPP.KADEDPOTPKRVSFTVKETVCPRPTR.QPPELCCDFKE.....KQCVSTVTLN...PSIHSLDIS..CNEIQSV... Group 15
METQASISLGRWSLWLLLLGVVPSASAQALSREAVLRVDRINEQSSSEANLYRLELDQPP.KADEDPOTPKRVSFTVKETVCPRPTR.QPPELCCDFKE.....KQCVSTVTLN...PSIHSLDIS..CNEIQSV... Group 16
METQASISLGRWSLWLLLLGVVPSASAQALSREAVLRVDRINEQSSSEANLYRLELDQPP.KADEDPOTPKRVSFTVKETVCPRPTR.QPPELCCDFKE.....KQCVSTVTLN...PSIHSLDIS..CNEIQSV... Group 17
METQASISLGRWSLWLLLLGVVPSASAQALSREAVLRVDRINEQSSSEANLYRLELDQPP.KADEDPOTPKRVSFTVKETVCPRPTR.QPPELCCDFKE.....KQCVSTVTLN...PSIHSLDIS..CNEIQSV... Group 18
METQASISLGRWSLWLLLLGVVPSASAQALSREAVLRVDRINEQSSSEANLYRLELDQPP.KADEDPOTPKRVSFTVKETVCPRPTR.QPPELCCDFKE.....KQCVSTVTLN...PSIHSLDIS..CNEIQSV... Group 19
...G...APVVDXND...GLQALQFANGLEYRASNDKYSVNVVISA..RLVSGIKYILQVIGRTCPKSSG..DLQCEFHDEPENAVTTCTVYVSTP..WLNQIKLESFQ... Group 20
...G...FPKTIKTND...GVLQARYSVEKFNCTNDMLFKESRITRAL..VQIVNGLYKMLEVEIGRTCKKQSB.LRLDDCDFQINHTLQTLSCISEVWVP...WLQHFVFPVLRCH... Group 21
C ET .....XXXXXXXXXX.XXXXXXXXXXXXXXXXXX...XXXXXXXXXX.XXXXXXX.X...XXXXCXXXX...X.CXXXXXXX...XXXXXXXXXX.CX..... 95 PIC ET
```

Bibliography

- Abrahamson M, Salvesen G, Barrett AJ, Grubb A (1986) Isolation of six proteinase inhibitors from human urine. Their physicochemical and enzyme kinetic properties and concentrations in biological fluids. *J. Biol. Chem.* 261: 11282-11289
- Abrahamson M, Grubb A, Olafsson I, Lundwall Å (1987a) Identification of the probable inhibitory reactive sites of the cysteine proteinase inhibitors human cystatin C and chicken cystatin. *J. Biol. Chem.* 262: 9688-9694
- Abrahamson M, Grubb A, Olafsson I, Lundwall Å (1987b) Molecular cloning and sequence analysis of cDNA coding for the precursor of the human cysteine proteinase inhibitor cystatin C. *FEBS Lett.* 216: 229-233
- Abrahamson M, Mason RW, Hansson H, Buttler D, Grubb A, Ohlsson K (1991a) Human cystatin C. Role of the N-terminal segment in the inhibition of human cysteine proteinases and its inactivation by leucocyte elastase. *Biochem. J.* 273: 621-626
- Abrahamson M, Buttler DJ, Mason RW, Hansson H, Grubb A, Lija H, Ohlsson K (1991b) Regulation of cystatin C activity by serine proteinases. *Biomed. Biochim. Acta* 50: 587-593
- Agarwal SK, Cogburn LA, Burnside J (1995) Comparison of gene expression in normal and growth hormone receptor-deficient dwarf chickens reveals a novel growth hormone regulated gene. *Biochem. Biophys. Res. Commun.* 206: 153-160
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J. Mol. Biol.* 215:403-410
- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* 25: 3389-3402
- Alvarez-Fernandez M, Barrett AJ, Gerhartz B, Dando PM, Ni J, Abrahamson M (1999) Inhibition of mammalian legumain by some cystatins is due to a novel second reactive site. *J. Biol. Chem.* 274: 19195-19203
- Ashton BA, Allen TD, Howlett CR, Eaglesom CC, Hattori A, Owen M (1980) Formation of bone and cartilage by marrow stromal cells in diffusion chambers *in vivo*. *Clin. Orthop.* 151: 294-307
- Auerswald EA, Genenger G, Assfalg-Machleidt I, Machleidt W, Engh RA, Fritz H (1992) Recombinant chicken egg white cystatins variants of the QLVSG region. *Eur. J. Biochem.* 209: 837-845
- Avioli L (1983) Osteoporosis. In: Peck WA (ed) *Bone and Mineral Research*, vol 1. Elsevier, Amsterdam, pp 280-318
- Bab I, Ashton BA, Gazit D, Marx G, Williamson MC, Owen ME (1986) Kinetics and differentiation of marrow stromal cells in diffusion chambers *in vivo*. *J. Cell. Sci.* 84: 139-151
- Baggio R, Shi Y, Wu Y, Abeles R H (1996) From poor substrates to good inhibitors: Design of inhibitors for serine and thiol proteases. *Biochemistry* 35:3351-3353

- Bairoch A, Bucher P, Hofmann K (1997) The PROSITE database, its status in 1997. *Nucl. Acids Res.* 25: 217-221
- Baron R (1989) Molecular mechanisms of bone resorption by the osteoclast. *Anat. Rec.* 224: 317-324
- Baker EN, Drenth J (1987) In: *Biological macromolecules and assemblies*. Jurnak F, McPherson A (eds) Vol 3, John Wiley and Sons, New York, pp 313-368
- Barrett AJ (1981) Cystatin, the egg white inhibitor of cysteine proteinases. *Meth. Enzymol.* 80: 771-778
- Barrett AJ, Davies M, Grubb A (1984) The place of human gamma-trace (cystatin C) amongst the cysteine proteinase inhibitors. *Biochem. Biophys. Res. Commun.* 120: 631-636
- Barrett AJ, Rawlings N, Davies M, Machleidt W, Salvesen G, Turk V (1986a) Cysteine proteinase inhibitors of the cystatin superfamily. In: *Proteinase inhibitors*. Barrett A, Salvesen G (Eds) Elsevier, Amsterdam, pp515-569
- Barrett AJ, Fritz H, Grubb A, Isemura S, Järvinen M, Katunuma N, Machleidt W, Müller-Esterl W, Sasaki M, Turk V (1986b) Nomenclature and classification of the proteins homologous with the cysteine-proteinase inhibitor chicken cystatin. *Biochem. J.* 236: 312-312
- Barrett AJ (1987) The cystatins: A new class of peptidase inhibitors. *Trends Biochem. Sci.* 12: 193-196
- Benson G (1999) Tandem repeats finder: A program to analyze DNA sequences. *Nucl. Acids Res.* 27: 573-580
- Bergad PL, Towle HC, Berry SA (1999) Definition of a high affinity growth hormone DNA response element. *Mol. Cell Endocrinol.* 150: 151-159
- Berger A, Schechter I (1970) Mapping the active site of papain with the aid of peptide substrates and inhibitors. *Philos. Trans. R. Soc. London* 257: 249-264
- Blair HC, Teitelbaum SL, Ghiselli R, Gluck S (1989) Osteoclastic bone resorption by a polarized vacuolar pump. *Science* 245: 855-857
- Blom N, Gammeltoft S, Brunak S (1999) Sequence and structure based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.* 294: 1351-1362
- Bobek LA, Levine MJ (1992) Cystatins-Inhibitors of cysteine proteinases. *Crit. Rev. Oral Biol. Med.* 3: 307-332
- Bode W, Engh R, Musil D, Thiele U, Huber R, Karshikov A, Brzin J, Kos J, Turk V (1988) The 2.0 Å X-ray crystal structure of chicken egg white cystatin and its possible mode of interaction with cysteine proteinases. *EMBO J.* 7: 2593-2599
- Bollengier F (1987) Cystatin C, alias post- γ -globulin: A marker for multiple sclerosis. *J. Clin. Chem. Biochem.* 25: 589-593
- Borodovsky M, McIninch JD (1993) GeneMark: Parallel gene recognition for both DNA strands. *Comp. Chem.* 17: 123-133

- Boskey AL (1981) Current concepts of the physiology and biochemistry of calcification. Clin. Orthop. 157: 225-257
- Boskey AL, Wians FHJ, Hauschka PV (1985) The effect of osteocalcin on *in vitro* lipid-induced hydroxyapatite formation and seeded hydroxyapatite growth. Calcif. Tissue Int. 37: 57-62
- Bossard MJ, Tomaszek TA, Thompson SK, Amegadzie BY, Hanning CR, Jones C, Kurdyla JT, McNulty DE, Drak FH, Gowen M, Levy MA (1996) Proteolytic activity of human osteoclast cathepsin K. J. Biol. Chem. 271: 12517-12524
- Breathnach R, Chambon P (1981) Organization and expression of eukaryotic split genes-coding for proteins. Annu. Rev. Biochem. 50: 349-383
- Brendel V, Bucher P, Nourbakhsh I, Blaisdell BE, Karlin S (1992) Methods and algorithms for statistical analysis of protein sequences. Proc. Natl. Acad. Sci. USA 89: 2002-2006
- Broek D, Madri J, Eikenberry E, Brodsky B (1985) Characterization of the tissue form of type V collagen from chick bone. J. Biol. Chem. 260: 555-562
- Brömme D, Okamoto K (1995) Human cathepsin O2 a novel cysteine protease highly expressed in osteoclastomas and ovary molecular cloning, sequencing and tissue distribution. Biol. Chem. 376: 379-384
- Brömme D, Okamoto K, Wang BB, Biroc S (1996) Human cathepsin O2, a matrix protein-degrading cysteine protease expressed in osteoclasts. J. Biol. Chem. 271: 2126-2132
- Brown WM, Saunders NR, Møllgård K, Dziegielewska KM (1992) Fetuin-an old friend revisited. Bioessays 14: 749-755
- Burge C (1997) Identification of genes in human genomic DNA. PhD thesis, Stanford University, Stanford, CA, USA
- Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268: 78-94
- Burset M, Guigo R (1996) Evaluation of gene structure prediction programs. Genomics 34: 353-367
- Burset M, Seledtsov IA, Solovyev VV (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucl. Acids Res. 28: 4364-4375
- Canalis E, McCarthy TL, Centrella M (1989) The role of growth factors in skeletal remodeling. Endocrinol. Metab. Clin. North Am. 18: 903-918
- Carninci P, Hayashizaki Y (1999) High-efficiency full-length cDNA cloning. Meth. Enzymol. 303: 19-44
- Chauhan SS, Golstein LJ, Gottesman MM (1991) Expression of cathepsin L in human tumors. Cancer Res. 51: 1478-1481

Chen J-M, Dando PM, Rawlings ND, Brown MA, Young NE, Stevens RA, Hewitt E, Watts C, Barrett AJ (1997) Cloning, isolation and characterization of mammalian legumain, an asparaginyl endopeptidase. *J. Biol. Chem.* 272: 8090-8098

Choi SJ, Reddy SV, Devlin RD, Menaa C, Chung H, Boyce BF, Roodman GD (1999) Identification of human asparaginyl endopeptidase (legumain) as an inhibitor of osteoclast formation and bone resorption. *J. Biol. Chem.* 274: 27747-27753

Chu CS, Trapnell BC, Curristin S, Cutting GR, Crystal RG (1993) Genetic basis of variable exon 9 skipping in cystic fibrosis transmembrane conductance regulator mRNA. *Nat. Genet.* 3: 151-156

Collela R, Chambers AF, Denhardt DT (1993) Anticarcinogenic activities of naturally occurring cysteine proteinase inhibitors. In: Toll W, Kennedy AR (eds) *Proteinase inhibitors as cancer chemopreventive agents*. New York, Plenum Press, pp 199-216

Colvard DS, Eriksen EF, Keeting PE, Wilson EM, Lubahn DB, French FS, Riggs BL, Spelsberg TC (1989) Identification of androgen receptors in normal human osteoblast-like cells. *Proc. Natl. Acad. Sci. USA* 86: 854-857

Combet C, Blanchet C, Geourjon C, Deléage G (2000) NPS@: Network protein sequence analysis. *Trends Biochem. Sci.* 25: 147-150

Cowan P, North A, Randall J (1953) High-angle X-ray diffraction of collagen fibres. In: Randall J (ed) *Nature and Structure of Collagen*. Butterworth, London, pp 241-249

Craven MW, Mural RJ, Hauser LJ, Uberbacher EC (1995) Predicting protein folding classes without overly relying on homology. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 3: 98-106

Dalgleish R (1987) Southern blotting. In: Boulnois GJ (ed) *Gene cloning and analysis*. Blackwell Scientific Publications, Oxford, pp 45-60

Delaissé J-M, Eeckhout Y, Vaes G (1980) Inhibition of bone resorption in culture by inhibitors of thiol proteinases. *Biochem. J.* 192: 365-368

Delaissé J-M, Eeckhout Y, Vaes G (1984) *In vivo* and *in vitro* evidence for the involvement of cysteine proteinases in bone resorption. *Biochem. Biophys. Res. Commun.* 125: 441-447

Devos I, de Clercq ND, Vercaeren I, Heyns W, Rombauts W, Peeters B (1993) Structure of rat genes encoding androgen-regulated cystatin-related proteins (CRPs): a new member of the cystatin superfamily. *Gene* 125: 159-167

de Vries IG, Coomans D, Wisse E (1988) Immunocytochemical localization of osteocalcin in human and bovine teeth. *Calcif. Tissue Int.* 43: 128-130

Dickson I (1974) The composition and antigenicity of sheep cortical bone matrix proteins. *Calcif. Tissue Res.* 16: 321-333

Dickson I (1993) Bone. In: Royce P, Steinmann B (eds) *Connective tissue and its heritable disorders, molecular, genetic and medical aspects*. Wiley-Liss Inc, New York, pp 249-285

- Dickson I, Eyre D, Kodicek E (1979) Influence of plasma calcium and vitamin D on bone collagen: Effects on lysine hydroxylation and crosslink formation. *Biochim. Biophys. Acta.* 588: 169-173
- Dodds RA, Connor JR, Drake F, Feild J, Gowen M (1998) Cathepsin K mRNA detection is restricted to osteoclasts during fetal mouse development. *J. Bone. Miner. Res.* 13: 673-682
- Dolan P, Torgerson DJ (1998) The cost of treating osteoporosis fractures in the United Kingdom female population. *Osteoporos Int.* 8: 611-617
- Dower WJ, Miller JF, Ragsdale WW (1988) High efficiency transformation of *E. coli* by high voltage electroporation. *Nucl. Acids Res.* 16: 6127-6145
- Drake FH, Dodds RA, James IE, Connor JR, Debouck C, Richardson S, Lee-Rykaczewski E, Coleman L, Rieman D, Barthlow R, Hastings G, Gowen M (1996) Cathepsin K, but not cathepsin B, L, or S, is abundantly expressed in human osteoclasts. *J. Biol. Chem.* 271: 12511-12516
- Dziegielewska KM, Møllgård K, Reynolds ML, Saunders NR (1987) A fetuin-related glycoprotein (α_2 HS) in human embryonic and fetal development. *Cell Tissue Res.* 248: 33-41
- Eddy SR, Durbin R (1994) RNA sequence analysis using covariance models. *Nucl. Acids Res.* 22: 2079-2088
- Eichbaum QG, Iyer R, Raveh DP, Mathieu C, Ezekowitz RA (1994) Restriction of interferon gamma responsiveness and basal expression of the myeloid human Fc gamma R1b gene is mediated by a functional PU.1 site and a transcription initiator consensus. *J. Exp. Med.* 179: 1985-1996
- Einstein JR, Uberbacher EC, Guan X, Mural RJ, Mann RC (1991) GAP - A computer program for gene assembly. ORNL/TM 11924
- Einstein JR, Mural RJ, Guan X, Uberbacher EC (1992) Computer-based construction of gene models using the GRAIL gene assembly program. ORNL/TM-12174
- Elzanowski A, Barxer WC, Hunt LT, Seibel-Ross E (1988) Cystatin domains in alpha-2-HS-glycoprotein and fetuin. *FEBS Lett.* 227: 167-170
- Etherington DJ (1980) Proteinases in connective tissue breakdown. *Ciba. Found. Symp.* 75: 87-100
- Everts V, Aronson DC, Beertsen W (1985) Phagocytosis of bone collagen by osteoclasts in two cases of pycnodysostosis. *Calcif. Tissue Int.* 37: 25-31
- Fichant GA, Burks C (1991) Identifying potential tRNA genes in genomic DNA sequences. *J. Mol. Biol.* 220: 659-671
- Fischer D, Barret C, Bryson K, Eloffson A, Godzik A, Jones D, Karplus KJ, Kelley LA, Maccallum RM, Pawowski K, Rost B, Rychlewski L, Sternberg MJ (1999) CAFASP-1: critical assessment of fully automated structure prediction methods. *Proteins Suppl.* 3: 209-17
- Fisher L, Termine J, Dejter SJ, Whitson S, Yanagishita M, Kimura J, Hascall V, Kleinman H, Hassell J, Nilsson B (1983) Proteoglycans of developing bone. *J. Biol. Chem.* 258: 6588-6594

- Fossum K, Whitaker JR (1968) Ficin and papain inhibitor from chicken egg white. *Arch. Biochem. Biophys.* 125: 367-375
- Friedenstein A (1973) Determined and inducible osteogenic precursor cells. *Ciba. Found. Symp.* 11: 169-185
- Frishman D, Argos P (1996) Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Eng.* 9: 133-142
- Garnier J, Gibrat JF, Robson B (1996) GOR secondary structure prediction method version IV. *Meth. Enzymol.* 266: 540-553
- Gelb BD, Shi G-P, Chapman HA, Desnick RJ (1996a) Pycnodysostosis, a lysosomal disease caused by cathepsin K deficiency. *Science* 273: 1236-1238
- Gelb BD, Spencer E, Obad S, Edelson GJ, Faire S, Weissenbach J, Desnick RJ (1996b) Pycnodysostosis: refined linkage and radiation hybrid analyses reduce the critical region to 2 cM at 1q21 and map two candidate genes. *Hum. Genet.* 98: 141-144
- Gelb BD, Shi G-P, Heller M, Weremowicz S, Morton C, Desnick RJ, Chapman HA (1997) Structure and chromosomal assignment of the human cathepsin K gene. *Genomics* 41: 258-262
- Genenger G, Lenzen S, Mentele R, Assfalg-Machleidt I, Auerswald EA (1991) Recombinant Q53E- and Q53N- chicken egg white cystatin variants inhibit papain, actinidin and cathepsin B. *Biomed. Biochim. Acta* 50: 621-625
- Geourjon C, Deleage G (1994) SOPM: a self-optimized method for protein secondary structure prediction. *Protein Eng.* 7: 157-164
- Gietz D, St. Jean A, Woods RA, Schiestl RH (1992) Improved method for high efficiency transformation of intact yeast cells. *Nucl. Acids. Res.* 20: 1425-1425
- Gingrich JC, Boehrer DM, Garnes JA, Johnson W, Wong BS, Bergmann A, Eveleth GG, Langlois RG, Carrano AV (1996) Construction and characterization of human chromosome 2-specific cosmid, fosmid and PAC clone libraries. *Genomics* 32: 65-74
- Glimcher MJ, Kossiva D, Brickley-Parsons D (1984) Phosphoproteins of chicken bone matrix. Proof of synthesis in bone tissue. *J. Biol. Chem.* 259: 290-293
- Gouzy J, Corpet F, Kahn D (1999) Whole genome protein domain analysis using a new method for domain clustering. *Comput. Chem.* 23: 333-340
- Grubb A, Abrahamson M, Olafsson I, Trojnar J, Kasprzykowska R, Kasprzykowska F, Grzonka Z (1990) Synthesis of cysteine proteinase inhibitors structurally based on the proteinase interacting N-terminal region of human cystatin C. *Biol. Chem.* 371: S137-S144
- Guan X, Mann RC, Mural RJ, Uberbacher EC (1991a) On parallel search of DNA sequence databases. *Proceedings of the 5th SIAM Conference on Parallel Processing for Scientific Computing*, pp 332-337
- Guan X, Mural RJ, Mann RC, Uberbacher EC (1991b) Searching consensus patterns on hypercube. *Sixth Distributed Memory Computing Conference*, Portland, OR, pp 470-472

- Guan X, Mural RJ, Einstein JR, Mann RC, Uberbacher EC (1992) GRAIL: An integrated artificial intelligence system for gene recognition and interpretation. Eighth IEEE Conference on AI Applications, IEEE Computer Society Press, Monterey, CA, March 2-6, pp 9-13
- Guan X, Uberbacher EC (1996) A fast look-up algorithm for detecting repetitive DNA sequences. Abstract in Proceedings of The First Pacific Symposium on Biocomputing, January 3-6, pp 718-719
- Gubler U, Hoffman BJ (1983) A simple and effective method for generating cDNA libraries. *Gene* 25: 263-269
- Guez N, Peitsch M C (1997) SWISS-MODEL and the Swiss-Pdb Viewer: An environment for comparative protein modeling. *Electrophoresis* 18: 2714-2723
- Habener JF, Rosenblatt M, Potts JTJ (1984) Parathyroid hormone: Biochemical aspects of biosynthesis, secretion, action and metabolism. *Physiol. Rev.* 64: 985-1053
- Hall A, Abrahamson M, Grubb A, Trojnar J, Kania P, Kasprzykowska R, Kasprzykowska F (1992) Cystatin C based peptidyl diazomethanes as cysteine proteinase inhibitors, influence of the peptidyl chain length. *J. Enzyme Inhib.* 6: 113-123
- Hall A, Dalbøge H, Grubb A, Abrahamson M (1993) Importance of the evolutionarily conserved glycine residue in the N-terminal region of the human cystatin C (Gly-11) for cysteine endopeptidase inhibition. *Biochem. J.* 291: 123-129
- Hanahan D (1983) Studies on transformation of *Escherichia coli* with plasmids. *J. Mol. Biol.* 166: 557-588
- Hansen SK, Tjian R (1995) TAFs and TFIIA mediate differential utilization of the tandem *Adh* promoters. *Cell* 82: 565-575
- Hauschka P, Lian J, Gallop P (1975) Direct identification of the calcium-binding amino acid γ -carboxyglutamic acid in mineralized tissue. *Proc. Natl. Acad. Sci. USA* 72: 3925-3929
- Hauschka PV, Carr SA (1982) Calcium-dependent α -helical structure in osteocalcin. *Biochemistry* 21: 2538-2547
- Henskens YMC, Veerman ECI, Mantel MS, Van der Velden U, Nieuw Amerongen AV (1994) Cystatins S and C in human whole saliva and glandular salivas in periodontal health and disease. *J. Dent. Res.* 73: 1606-1614
- Herring GM (1972) The organic matrix of bone. In: Bourne GH (ed) *The biochemistry and physiology of bone*. 2nd edition, vol 1, Academic Press, New York, pp 127-189
- Hill J, Donald KA, Griffiths DE (1991) DMSO-enhanced whole cell yeast transformation. *Nucl. Acids Res.* 19: 5791-5791
- Hjerpe A, Reinholt F, Engfeldt B (1979) The occurrence of chondroitin-6-sulphate in adult human compact bone tissue. *Calcif. Tissue Int.* 29: 169-171
- Horwitz JP, Cluna J, Curby RJ, Tomson AJ, DaRooge MA, Fisher BE, Mauricio J, Klundt I (1964) Substrates for cytochemical demonstration of enzyme activity I. Some substituted 3-indolyl- β -D-glycopyranosides. *J. Med. Chem.* 7: 574-574

- Hu B, Coulson L, Moyer B, Price PA (1995) Isolation and molecular cloning of a novel bone phosphoprotein related in sequence to the cystatin family of thiol protease inhibitors. *J. Biol. Chem.* 270: 431-436
- Hutchinson KW, Halvorson HO (1980) Cloning of randomly sheared DNA fragments from a ϕ 10S lysogen of *Bacillus subtilis*: Identification of prophage-containing clones. *Gene* 8: 267-278
- Inaoka T, Bilbe G, Ishibashi O, Tezuka K-i, Kumegawa M, Kokubo T (1995) Molecular cloning of human cDNA for cathepsin K: novel cysteine proteinase predominantly expressed in bone. *Biochem. Biophys. Res. Commun.* 206: 89-96
- Ish-Horowicz D, Burke JF (1981) Rapid and efficient cosmid cloning. *Nucl. Acids Res.* 9: 2989-2998
- Ito H, Fukada Y, Murata K, Kimura A (1983) Transformation of intact yeast cells treated with alkali cations. *J. Bacteriol.* 153: 163-168
- Järvinen M, Rinne A, Hopsu-Havu VK (1987) Human cystatins in normal and diseased tissues - A review. *Acta Histochem.* 82: 5-18
- Javahery R, Khachi A, Zenzie-Gregory B, Smale ST (1994) DNA sequence requirements for transcriptional initiator activity in mammalian cells. *Mol. Cell Biol.* 14: 116-127
- Jeffreys AJ, Neumann R, Wilson V (1990) Repeat unit sequence variation in minisatellites: A novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. *Cell* 60: 473-485
- Jilka RL, Hangoc G, Girasole G, Passeri G, Williams DC, Abrams JS, Boyce B, Broxmeyer H, Manolagas SC (1992) Increased osteoclast development after estrogen loss: mediation by interleukin-6. *Science* 257: 88-91
- Johnston CCJ (1985) Studies on prevention of age related bone loss. In: Peck WA (ed) *Bone and Mineral Research*, vol 3. Elsevier, Amsterdam, pp 233-257
- Kato H, Nagasawa S, Iwanaga S (1981) HMW and LMW kininogen. *Meth. Enzymol.* 80: 172-198
- Keilová H, Tomášek V (1974) Effect of papain inhibitor from chicken egg white on cathepsin B. *Biochem. Biophys. Acta.* 334: 179-186
- Kelley LA, Maccallum R, Sternberg MJE (1999) RECOMB 99. In: Istrail S, Pevzner P and Waterman M (eds) *Proceedings of the Third Annual Conference on Computational Molecular Biology*, The Association for Computing Machinery, New York, pp 218-225
- Kelley LA, MacCallum RM, Sternberg MJE (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* 299: 499-520
- Kirschke H, Langner J, Riemann S, Wiederanders B, Ansorge S, Bohley P (1980) *Ciba. Found Symp.* 75: 15-35
- Kneller DG, Cohen FE, Langridge R (1990) Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.* 214: 171-182

- Kobori M, Ikeda Y, Nara H, Kato M, Kumegawa M, Nojima H, Kawashima H (1998) Large scale isolation of osteoclast-specific genes by an improved method involving the preparation of a subtracted cDNA library. *Genes Cells* 3: 459-475
- Koide T, Foster D, Yoshitake S, Davie EW (1986) Amino acid sequence of human histidine-rich glycoprotein derived from the nucleotide sequence of its cDNA. *Biochemistry* 25: 2220-2225
- Kölliker (1873) Die normale resorption des knorpelgewebes und ihre bedeutung für die entstehung der typischen knochenformen. FCW Vogel, Leipzig
- Korant BD, Brzin J, Turk V (1985) Cystatin, a protein inhibitor of cysteine proteinases alters viral protein cleavages in infected human cells. *Biochem. Biophys. Res. Commun.* 127: 1072-1076
- Kozak M (1989) A scanning model for translation: an update. *J. Cell. Biol.* 108: 229-241
- Krogh A (1997) Two methods for improving performance of an HMM and their application for gene finding. *Proc. Fifth Int. Conf. Intell. Syst. Mol. Biol.* 179-186
- Lalmanach G, Hoebeke J, Moreau T, Brillard-Bourdet M, Ferrer-Di Martino M, Borrás-Cuesta F, Gauthier F (1993) Interaction between cystatin-derived peptides and papain. *J. Protein Chem.* 12: 23-31
- Lee C-C, Bowman BH, Yang F (1987) Human α_2 -HS-glycoprotein: the A and B chains with a connecting sequence are encoded by a single mRNA transcript. *Proc. Natl. Acad. Sci. USA* 84: 4403-4407
- Lenarcic B, Gabrijelcic D, Rozman B, Drobic-Kosorok M, Turk V (1988) Human cathepsin B and cysteine proteinase inhibitors (CPIs) in inflammatory and metabolic joint diseases. *Biol. Chem.* 369: S257-S261
- Lerner CG, Inouye M (1990) Low copy number plasmids for regulated low-level expression of cloned genes in *Escherichia coli* with blue/white insert screening capability. *Nucl. Acids Res.* 18: 4631-4631
- Lerner UH, Grubb A (1992) Human cystatin C, a cysteine proteinase inhibitor, inhibits bone resorption *in vitro* stimulated by parathyroid hormone and parathyroid hormone-related peptide of malignancy. *J. Bone Min. Res.* 7: 433-440
- Levin JM, Robson B, Garnier J (1986) An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Lett.* 205: 303-308
- Leung L (1993) Histidine-rich glycoprotein: An abundant plasma protein in search of a function. *J. Lab. Clin. Med.* 121: 630-631
- Li Y-P, Alexander M, Wucherpfenning AL, Yelick P, Chen W, Stashenko P (1995) Cloning and complete coding sequence of a novel human cathepsin expressed in giant cells of osteoclastomas. *J. Bone. Miner. Res.* 10: 1197-1202
- Li Y-P, Chen W (1999) Characterization of mouse cathepsin K gene, the gene promoter and the gene expression. *J. Bone. Miner. Res.* 14: 487-499

- Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* 257: 342-358
- Lindahl P, Alriksson E, Jörnwall, Björk I (1988) Interaction of the cysteine proteinase inhibitor chicken cystatin with papain. *Biochemistry* 27: 5074-5082
- Lindahl P, Abrahamson M, Björk I (1992) Interaction of recombinant human cystatin C with cysteine proteinases papain and actinidin. *Biochem. J.* 281: 49-55
- Ling M, Merante F, Robinson BH (1995) A rapid and reliable DNA preparation method for screening a large number of yeast clones by polymerase chain reaction. *Nucl. Acids Res.* 23: 4924-4925
- Linsk R, Gottesman M, Pernis B (1989) Are tissues a patch quilt of ectopic gene expression? *Science* 246: 261-261
- Lo K, Smale ST (1996) Generality of a functional initiator consensus sequence. *Gene* 182: 13-22
- Love RR, Mazess RB, Barden HS, Epstein S, Newcomb PA, Jordan VC, Carbone PP, DeMets DL (1992) Effects of tamoxifen on bone mineral density in postmenopausal women with breast cancer. *N. Engl. J. Med.* 326: 852-856
- Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucl. Acids Res.* 25: 955-964
- Machleidt W, Thiele U, Laber B, Assfalg-Machleidt I, Esterl A, Wiegand G, Kos J, Turk V, Bode W (1989) Mechanism of inhibition of papain by chicken egg white cystatin. *FEBS Lett.* 243: 234-238
- Machleidt W, Thiele U, Assfalg-Machleidt I, Förger D, Auerswald EA (1991) Molecular mechanism of inhibition of cysteine proteinases by their protein inhibitors: Kinetic studies with natural and recombinant variants of cystatins and stefins. *Biomed. Biochim. Acta.* 50: 613-620
- Makalowski W, Mitchell GA, Labuda D (1994) Alu sequences in the coding regions of mRNA: A source of protein variability. *Trends Genet.* 10: 188-193
- Malone JD, Teitelbaum SL, Griffin GL, Senior RM, Kahn AJ (1982) Recruitment of osteoclast precursors by purified bone matrix constituents. *J. Cell Biol.* 92: 227-230
- Mardon HJ, Bee J, vonderMark K, Owen ME (1987) Development of osteogenic tissue in diffusion chambers from early precursor cells in bone marrow of adult rats. *Cell Tissue Res.* 250: 157-165
- Marks N, Berg M, Benuck M (1986) Preferential action of rat brain cathepsin B as a peptidyl dipeptidase converting pro-opoid oligopeptides. *Arch. Biochem. Biophys.* 249: 489-499
- Maroteaux P, Lamy M (1962) *Presse Med.* 70: 999
- Marotti G, Muglia MA (1988) A scanning electron microscope study of human bony lamellae. Proposal for a new model of collagen lamellar organization. *Arch. Ital. Anat. Embriol.* 93: 163-175

Matis S, Xu Y, Shah M, Guan X, Einstein JR, Mural RJ, Uberbacher EC (1996) Detection of RNA polymerase II promoters and polyadenylation sites in human DNA sequence. *Comput. Chem.* 20: 135-140

McClelland JL, Rumelhart DE (1988) *Explorations in parallel distributed processing*. Vol 3, MIT Press, Cambridge MA, pp 318-362

Mighell AJ, Markham AF, Robinson PA (1997) Alu sequences. *FEBS Lett.* 417: 1-5

Miki R, Kadota K, Bono H, Mizuno Y, Tomaru Y, Carninci P, Itoh M, Shibata K, Kawai J, Konno H, Watanabe S, Sato K, Tokusumi Y, Kikuchi N, Ishii Y, Hamaguchi Y, Nishizuka I, Goto H, Nitanda H, Satomi S, Yoshiki A, Kusakabe M, DeRisi JL, Eisen MB, Iyer VR, Brown PO, Muramatsu M, Shimada H, Okazaki Y, Hayashizaki Y (2001) Delineating developmental and metabolic pathways *in vivo* by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays. *Proc. Natl. Acad. Sci. USA* 98: 2199-2204

Morita M, Yoshiuchi N, Arakawa H, Nishimura S (1999) CMAP: a novel cystatin-like gene involved in liver metastasis. *Cancer Res.* 59: 151-158

Müller-Esterl W, Iwanga S, Nakanishi S (1986) Kininogens revisited. *Trends Biochem. Sci.* 11: 336-339

Müller-Esterl W (1989) Kininogens, kinins and kinships. *Thromb. Haemostasis* 61: 2-6

Mullis KB, Faloona FA (1987) Specific synthesis of DNA *in vitro* via a polymerase-catalysed chain reaction. *Meth. Enzymol.* 155: 335-350

Mural RJ, Einstein JR, Guan X, Mann RC, Uberbacher EC (1991) An artificial intelligence approach to DNA sequence feature recognition. *Trends Biotech.* 10: 66-69

Mural RJ, Guan X, Uberbacher EC (1993) Computational methods for locating biological features in DNA sequences. *Current Protocols in Human Genetics*, Unit 6.5, Supplement 6
Nagy E, Maquat LE (1998) A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem. Sci.* 23: 198-199

Nakai K, Horton P (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.* 24: 34-36

Nijweide P, Burger EH, Feyen JHM (1986) Cells of bone: Proliferation, differentiation, and hormonal regulation. *Physiol. Rev.* 66: 855-886

Nomura S, Wills AJ, Edwards DR, Heath JK, Hogan BLM (1988) Developmental expression of 2ar (osteopontin) and SPARC (osteonectin) RNA as revealed by *in situ* hybridization. *J. Cell. Biol.* 106: 441-450

Novina CD, Roy AL (1996) Core promoters and transcriptional control. *Trends Genet.* 12: 351-355

Ogata T, Wozney JM, Benazra R, Noda M (1993) Bone morphogenetic protein 2 transiently enhances expression of a gene, Id (inhibitor of differentiation), encoding a helix-loop-helix molecule in osteoblast-like cells. *Proc. Natl. Acad. Sci. USA* 90: 9219-22

- Ohnishi T, Nakamura O, Ozawa M, Arakaki N, Muramatsu T, Daikuhara Y (1993) Molecular cloning and sequence analysis of cDNA for a 59 kDa bone sialoprotein of the rat: Demonstration that it is a counterpart of human α_2 -HS glycoprotein and bovine fetuin. *J. Bone Miner. Res.* 8: 367-377
- Oldberg Å, Franzén A, Heinegård D (1986) Cloning and sequence analysis of rat bone sialoprotein (osteopontin) cDNA reveals an Arg-Gly-Asp cell-binding sequence. *Proc. Natl. Acad. Sci. USA* 83: 8819-8823
- Osoegawa K, Tateno M, Woon PY, Frengen E, Mammoser AG, Catanese JJ, Hayashizaki Y, de Jong PJ (2000) Bacterial artificial chromosome libraries for mouse sequencing and functional analysis. *Genome Res.* 10: 116-128
- Othani O, Fukuyama K, Epstein WL (1982) Biochemical properties of thiol proteinase inhibitors purified from psoriatic scales. *J. Invest. Dermatol.* 82: 280-284
- Owen M, Triffitt J (1976) Extravascular albumin in bone tissue. *J. Physiol. (London)* 257: 293-307
- Palumbo C, Palazzini S, Zaffe D, Marotti G (1990) Osteocyte differentiation in the tibia of newborn rabbit: An ultrastructural study of the formation of cytoplasmic processes. *Acta. Anat. (Basel)* 137: 350-358
- Parker MG, Scrace GT, Mainwaring WIA (1978) Testosterone regulates the synthesis of major proteins in rat ventral prostate. *Biochem. J.* 170: 115-121
- Pavesi A, Conterio F, Bolchi A, Dieci G, Ottonello S (1994) Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix analysis of transcriptional control regions. *Nucl. Acids Res.* 22: 1247-1256
- Pederson KO (1944) Fetuin, a new globulin isolated from serum. *Nature* 154: 575-575
- Poser JW, Esch FS, Ling NC, Price PA (1980) Isolation and sequence of the vitamin K-dependent protein from human bone. Undercarboxylation of the first glutamic acid residue. *J. Biol. Chem.* 255: 8685-8691
- Posner AS (1987) Bone mineral and the mineralization process. In: Peck WA (ed) *Bone and mineral research*, Vol 5, Elsevier, Amsterdam, pp 65-116
- Price PA, Rice JS, Williamson MK (1994) Conserved phosphorylation of serines in the Ser-X-Glu/Ser (P) sequences of the vitamin K-dependent matrix Gla protein from shark, lamb, rat, cow and human. *Protein Sci.* 3: 822-830
- Quandt K, Frech K, Karas H, Wingender E, Werner T (1995) MatInd and MatInspector-new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucl. Acids Res.* 23: 4878-4884
- Raghunath M, Mackay K, Dalgleish R, Steinmann B (1995) Genetic counselling on brittle grounds: Recurring osteogenesis imperfecta due to parental mosaicism for a dominant mutation. *Eur. J. Pediatr.* 154: 123-129
- Raisz L, Kream B (1983) Regulation of bone formation. *N. Engl. J. Med.* 309: 29-35, 83-89

- Raisz L (1988) Local and systemic factors in the pathogenesis of osteoporosis. *N. Engl. J. Med.* 318: 818-828
- Rantakokko J, Aro HT, Savontaus M, Vuorio E (1996) Mouse cathepsin K: cDNA cloning and predominant expression of the gene in osteoclasts, and in some hypertrophying chondrocytes during mouse development. *FEBS Lett.* 393: 307-313
- Rantakokko J, Kiviranta R, Eerola R, Aro HT, Vuorio E (1999) Complete genomic structure of the mouse cathepsin K gene (*Ctsk*) and its localization next to the *Arnt* gene on mouse chromosome 3. *Matrix Biol.* 18: 155-161
- Reginster JYL, Halkin V, Gosset C, Deroisy R (1997) The role of biphosphonates in the treatment of osteoporosis. *Drugs Today* 33: 563-570
- Rodan G, Fleisch HA (1996) Biphosphonates: Mechanisms of action. *J.Clin. Invest.* 97: 2692-2696
- Rodan SB, Rodan GA, Simmons HA, Walenga RW, Feinstein MB, Raisz LG (1981) Bone resorptive factor produced by osteosarcoma cells with osteoblastic features is PGE₂. *Biochem. Biophys. Res. Commun.* 102: 1358-1365
- Roeder RG (1991) The complexities of eukaryotic transcription initiation: Regulation of preinitiation complex assembly. *Trends Biochem. Sci.* 16: 402-408
- Romeo D, Skerlavaj B, Bolognesi M, Gennaro R (1988) Structure and bactericidal activity of an antibiotic dodecapeptide purified from bovine neutrophils. *J. Biol. Chem.* 263: 9573-9575
- Rost B, Sander C (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* 232: 584-599
- Rost B, Sander C (1994) Conservation and prediction of solvent accessibility in protein families. *Proteins* 20: 216-226
- Rost B (1996) PHD: Predicting one-dimensional protein structure by profile-based neural networks. *Meth. Enzymol.* 266: 525-539
- Rozhin J, Wade R, Honn KV, Sloane BF (1989) Membrane associated cathepsin L, a role in metastasis. *Biochem. Biophys. Res. Commun.* 164: 556-561
- Salvesen G, Parkes C, Abrahamson M, Grubb A, Barrett AJ (1986) Human low-M_r kininogen contains three copies of a cystatin sequence that are divergent in structure and in inhibitory activity for cysteine proteinases. *Biochem. J.* 234: 429-434
- Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular cloning: A laboratory manual*. 2nd Edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor
- Sander C, Schneider R (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9: 56-68
- Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* 74: 5463-5467

- Sarkar G, Sommer SS (1989) Access to messenger RNA sequence or its protein product is not limited by tissue or species specificity. *Science* 244: 331-334
- Saunders NR, Reynolds ML, Habgood MD, Ward RA (1992) Origin and fate of fetuin-containing neurons in the developing neocortex of the fetal sheep. *Anat. Embryol.* 186: 477-486
- Schaefer BC (1995) Revolutions in rapid amplification of cDNA ends: New strategies for polymerase chain reaction cloning of full-length cDNA ends. *Anal. Biochem.* 277: 255-273
- Schenk RK, Felix R, Hofstetter W (1993) Morphology of connective tissue: Bone. In: *Connective tissue and its heritable disorders*. Wiley-Liss Inc, pp 85-101
- Schiestl RH, Gietz RD (1989) High efficiency transformation of intact cells using single stranded nucleic acids as a carrier. *Curr. Genet.* 16: 339-346
- Schwyter DH, Huang J, Dubincoff T, Courey AJ (1995) The decapentaplegic core promoter region plays an integral role in the spatial control of transcription. *Mol. Cell Biol.* 15: 3960-3968
- Sen LC, Whitaker JR (1973) Some properties of a ficin-papain inhibitor from avian egg white. *Arch. Biochem. Biophys.* 158: 623-632
- Senepathy P, Shapiro MB, Harris NL (1990) Splice junctions, branch point sites and exons: Sequence statistics, identification and applications to genome project. *Meth. Enzymol.* 183: 252-278
- Shah MB, Xu Y, Einstein JR, Guan X, Hauser LJ, Matis SA, Lee RW, Mural RJ, Uberbacher EC (1995) Gene discovery and sequence annotation in GRAIL 1.3. The Hilton Head DNA Sequence Conference, Hilton Head, S.C., September 16-20
- Shah MB, Guan X, Einstein JR, Matis S, Xu Y, Mural RJ, Uberbacher EC (1996) User's guide to GRAIL and GENQUEST (Sequence analysis, gene assembly and sequence comparison systems) E-mail servers and XGRAIL (Version 1.3c), GRAILCLNT (Version 1.3) command line interface and XGENQUEST (Version 1.1) Client-server systems. Available by anonymous ftp from arthur.epm.ornl.gov (128.219.9.76) from directory pub/xgrail or pub/xgenQuest or pub/grailclnt as file Manual.grail1.3-genquest.
- Shi GP, Chapman HA, Bhairi SM, DeLeeuw C, Reddy VY, Weiss SJ (1995) Molecular cloning of human cathepsin O, a novel endoproteinase and homologue of rabbit OC2. *FEBS Lett.* 357: 129-134
- Sloane B, Honn K (1984) Cysteine proteinases and metastasis. *Cancer Metastasis Rev.* 3: 249-263
- Sloane BF, Rozhin J, Hatfield JS, Crissman JD, Honn KV (1987) Plasma membrane-associated cysteine proteinases in human and animal tumours. *Exp. Cell Biol.* 55: 209-224
- Smale ST (1997) Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes. *Biochim. Biophys. Acta.* 1351: 73-88
- Smale ST, Baltimore D (1989) The "initiator" as a transcription control element. *Cell* 57: 103-113

- Smith DM, Nance WE, Kang KW, Christian JC (1973) Genetic factors in determining bone mass. *J. Clin. Invest.* 52: 2800-2808
- Solovyev VV, Lawrence CB (1993) Prediction of human gene structure using dynamic programming and oligonucleotide composition. In : Abstracts of the 4th annual Keck symposium, Pittsburgh, pp 47-47
- Solovyev VV, Salamov AA, Lawrence CB (1994a) The prediction of human exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. In: Altman R, Brutlag D, Karp R, Latrop R, Searls D (eds) *The Second International conference on Intelligent systems for Molecular Biology*, AAAI Press, Menlo Park, CA, pp 354-362
- Solovyev VV, Salamov AA, Lawrence CB (1994b) Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucl. Acids Res.* 22: 5156-5163
- Solovyev VV, Salamov AA, Lawrence CB (1995) Identification of human gene structure using linear discriminant functions and dynamic programming. *Proc. Third Int. Conf. Intell. Syst. Mol. Biol.* pp 367-375
- Solovyev VV, Salamov AA (1997) The Gene-Finder computer tools for analysis of human and model organisms genome sequences. In: *Proceedings of the Fifth International conference on Intelligent systems for Molecular Biology*. AAAI Press, Menlo Park.
- Sonnhammer EL, Kahn D (1994) Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci.* 3: 482-492
- Southern EM, Anand R, Brown WRA, Fletcher DS (1987) A model for the separation of large DNA molecules by crossed field gel electrophoresis. *Nucl. Acids Res.* 15: 5925-5943
- Suda T, Takahashi N, Martin TJ (1992) Modulation of osteoclast differentiation. *Endocr. Rev.* 13: 66-80
- Swallow JE, Merrison WK, Gill PK, Harris S, Dalgleish R (1997) Assignment of secreted phosphoprotein 24 (*SPP2*) to human chromosome band 2q37→qter by *in situ* hybridization. *Cytogenet. Cell Genet.* 79: 142-142
- Takagaki Y, Kitamura N, Nakanishi S (1985) Cloning and sequence analysis of cDNAs for human high molecular weight and low molecular weight prekininogens. *J. Biol. Chem.* 260: 8601-8609
- Takahashi N, Akatsu T, Udagawa N, Sasaki T, Yamaguchi A, Moseley JM, Martin TJ, Suda T (1988) Osteoblastic cells are involved in osteoclast formation. *Endocrinology* 123: 2600-2602
- Taketo A (1988) DNA transfection of *Escherichia coli* by electroporation. *Biochim. Biophys. Acta.* 949: 318-324
- Taugner R, Buhrlé C, Nobiling R, Kirschke H (1985) Coexistence of renin and cathepsin B in epithelioid cell secretory granules. *Histochemistry* 88: 102-108
- Temin HM, Mizutani S (1970) RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature* 226: 1211-1213

Termine JD, Gehron-Robey P, Fisher LW, Shimokawa H, Drum MA, Conn KM, Hawkins GR, Cruz JB, Thompson KG (1984) Osteonectin, bone proteoglycan, and phosphoryn defects in a form of bovine osteogenesis imperfecta. *Proc. Natl. Acad. Sci. USA* 81: 2213-2217

Teti A, Blair HC, Schlesinger P, Grano M, Zamboni-Zallone A, Kahn AJ, Teitelbaum SL, Hruska KA (1989) Extracellular protons acidify osteoclasts, reduce cytosolic calcium and promote expression of cell-matrix attachment structures. *J. Clin. Invest.* 84: 773-780

Tezuka K-i, Tezuka Y, Maejima A, Sato T, Nemoto K, Kamioka H, Hakeda Y, Kumegawa M (1994) Molecular cloning of a possible cysteine proteinase predominantly expressed in osteoclasts. *J. Biol. Chem.* 269: 1106-1109

Tjian R, Maniatis T (1994) Transcriptional activation: A complex puzzle with few easy pieces. *Cell* 77: 5-8

Triffit J, Gebauer U, Ashton B, Owen M, Reynolds J (1976) Origin of plasma α_2 HS-glycoprotein and its accumulation in bone. *Nature* 262: 226-227

Triffit JT (1987) The special proteins of bone tissue. *Clin. Sci.* 72: 399-408

Turk V (1986) Cysteine proteinases and their inhibitors. In: Turk, V (ed), *Proceedings of the International Symposium*. Walter de Gruyter Berlin, Porotoroz-Yugoslavia

Turner RT, Riggs BL, Spelsberg TC (1994) Skeletal effects of estrogen. *Endocr. Rev.* 15: 275-300

Uberbacher EC, Mann RC, Hand RC, Mural RJ (1991) A neural network-multiple sensor based method for recognition of gene coding segments in human DNA sequence data. ORNL/TM-11741

Uberbacher EC, Mural RJ (1991) Locating protein coding regions in human DNA sequences using a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. USA* 88: 11261-11265

Uberbacher EC, Einstein JR, Guan X, Mural RJ (1992) Gene recognition and assembly in the GRAIL system: Progress and challenges. *The Second International Conference on Bioinformatics, Supercomputing, and Complex Genome Analysis*, pp 465-476

Uberbacher EC (1994) ORNL Announces genQuest and X-GRAIL. *Hum. Genome News* 5: 8-9

Uberbacher EC, Xu Y, Mural RJ (1995a) Discovering and understanding genes in human DNA sequence using GRAIL. *Computer Methods for Macromolecular Sequence Analysis*, September

Uberbacher EC, Xu Y, Shah M, Matis S, Guan X, Mural RJ (1995b) DNA sequence pattern recognition methods in GRAIL. Presentation to be published as full article in DIMACS Workshop on Gene-Finding and Gene Structure Prediction, Philadelphia, PA, October 13-14

Uberbacher EC (1995) Discovering the intelligence in molecular biology. *Trends Biotech.* 13: 497-500

- Uchiyama A, Suzuki M, Lefteriou B, Glimcher M (1986) Isolation and chemical characterization of the phosphoproteins of chicken bone matrix: Heterogeneity in molecular weight and composition. *Biochemistry* 25: 7572-7583
- Ullmann A, Jacob F, Monod J (1967) Characterization by *in vitro* complementation of a peptide corresponding to an operator-proximal segment of the β -galactosidase structural gene of *Escherichia coli*. *J. Mol. Biol.* 24: 339-343
- Votta BJ, Levy MA, Badger A, Bradbeer J, Dodds RA, James IE, Thompson S, Bossard MJ, Carr T, Connor JR, Tomaszek TA, Szewczuk L, Drake FH, Veber DF, Gowen M (1997) Peptide aldehyde inhibitors of cathepsin K inhibit bone resorption both *in vitro* and *in vivo*. *J. Bone. Miner. Res.* 12: 1396-1406
- Wasi S, Otsuka K, Yao K-L, Tung PS, Aubin JE, Sodek J, Termine JD (1984) An osteonectin-like protein in porcine periodontal ligament and its synthesis by periodontal ligament fibroblasts. *Can. J. Biochem. Cell Biol.* 62: 470-478
- Weinbaums S, Cowin S, Zeng Y (1992) Fluid shear stress excitation of osteocytes. *Adv. Bioeng.* 22: 25-28
- Wilkinson M (1991) Isolation of total cellular RNA using the guanidinium-lithium chloride method. In: Brown TA (ed), *Essential molecular biology: A practical approach*. IRL Press, Oxford, pp 28-29
- Wong GL (1986) Skeletal effects of parathyroid hormone. In: Pecks WA (ed) *Bone and Mineral Research*, vol 4. Elsevier, Amsterdam, pp 103-129
- Wootton JC, Federhen S (1996) Analysis of compositionally biased regions in sequence databases. *Meth. Enzymol.* 266: 554-571
- Wu M, Maier E, Benz R, Hancock REW (1999) Mechanism of interaction of different classes of cationic antimicrobial peptides with planar bilayers and with the cytoplasmic membrane of *Escherichia coli*. *Biochemistry* 38: 7235-7242
- Xia LH, Kilb J, Wex H, Li ZQ, Lipyansky A, Breuil V, Stein L, Palmer JT, Dempster DW, Bromme D (1999) Localization of rat cathepsin K in osteoclasts and resorption pits: Inhibition of bone resorption and cathepsin K-activity by peptidyl vinyl sulfones. *Biol. Chem.* 380: 679-687
- Xu Y, Mural RJ, Shah M, Uberbacher EC (1994a) Recognizing exons in genomic sequence using GRAIL II. *Genetic Engineering, Principles and Methods*, Plenum Press, Vol. 15
- Xu Y, Einstein JR, Mural RJ, Shah M, Uberbacher EC (1994b) An improved system for exon recognition and gene modeling in human DNA sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2: 376-84
- Xu Y, Mural RJ, Uberbacher EC (1994c) Constructing gene models from accurately-predicted exons: An application of dynamic programming. *Comput. Appl. Biosci.* 10: 613-623
- Xu Y, Mural RJ, Uberbacher EC (1995a) An iterative algorithm for correcting sequencing errors in DNA coding regions. Presentation to be published as full article in DIMACS Workshop on Gene-Finding and Gene Structure Prediction, Philadelphia, PA, October 13-14

- Xu Y, Mural RJ, Uberbacher EC (1995b) Correcting sequencing errors in DNA coding regions using a dynamic programming approach. *Comput. Appl. Biosci.* 11: 117-124
- Xu Y, Mural RJ, Uberbacher EC (1995c) An iterative algorithm for correcting sequencing errors in DNA coding regions. *J. Comput. Biol.* 3: 333-344
- Xu Y, Uberbacher EC (1996a) Gene prediction by pattern recognition and homology search. *The Fourth International Conference on Intelligent Systems for Molecular Biology*, St. Louis, MO, June 13-15
- Yamakawa Y, Omori-Satoh T (1992) Primary structure of the antihemorrhagic factor in serum of the Japanese Habu snake: a snake venom metalloproteinase inhibitor with a double-headed cystatin domain. *J. Biochem.* 112: 583-589
- Yamazaki K, Suzuki M, Mikuni-Takagaki Y, Hiraiwa K, Lefteriou B, Glimcher MJ (1988) Preparation of monoclonal antibodies to chicken bone phosphoproteins. *Calcif. Tissue Int.* 43: 41-43
- Young MF, Kerr JM, Ibaraki K, Heegaard A, Robey PG (1991) Structure, expression, and regulation of the major noncollagenous matrix proteins of bone. *Clin. Orthop.* 281: 275-294
- Zaidi M, Troen B, Moonga BS, Abe E (2001) Cathepsin K, osteoclastic resorption and osteoporosis therapy. *J. Bone Miner. Res.* 16: 1747-1749
- Zanetti M, Gennaro R, Romeo D (1995) Cathelicidins: a novel protein family with a common proregion and a variable C-terminal antimicrobial domain. *FEBS Lett.* 374: 1-5
- Zhang MQ (1997) Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl. Acad. Sci. USA.* 94: 565-568